Different methods of selecting animals for genotyping to maximize the amount of genetic information known in the population

M. L. Spangler,*[1] R. L. Sapp,†[2] J. K. Bertrand,* M. D. MacNeil,† and R. Rekaya*‡§[3]

*Animal and Dairy Science Department, University of Georgia, Athens 30602-2771;
†USDA-ARS, Fort Keogh Livestock and Range Research Laboratory, Miles City, MT 59301;
‡Department of Statistics, and §Institute of Bioinformatics, University of Georgia, Athens 30602-2771

ABSTRACT: It is possible to predict genotypes of some individuals based on genotypes of relatives. Different methods of sampling individuals to be genotyped from populations were evaluated using simulation. Simulated pedigrees included 5,000 animals and were assigned genotypes based on assumed allelic frequencies for a SNP (favorable/unfavorable) of 0.3/0.7, 0.5/0.5, and 0.8/0.2. A field data pedigree (29,101 animals) and a research pedigree (8,688 animals) were used to test selected methods using simulated genotypes with allelic frequencies of 0.3/0.7 and 0.5/0.5. For the simulated pedigrees, known and unknown allelic frequencies were assumed. The methods used included random sampling, selection of males, and selection of both sexes based on the diagonal element of the inverse of the relationship matrix (A^{-1}) and absorption of either the A or A^{-1} matrix. For random sampling, scenarios included selection of 5 and 15% of the animals, and all other methods presented concentrated on the selection of 5% of the animals for genotyping. The methods were evaluated based on the percentage of alleles correctly assigned after peeling (AK_P), the probability of assigning true alleles (AK_G), and the average probability of correctly assigning the true genotype. As expected, random sampling was the least desirable method. The most desirable method in the simulated pedigrees was selecting both males and females based on their diagonal element of A^{-1}. Increases in AK_P and AK_G ranged from 26.58 to 29.11% and 2.76 to 6.08%, respectively, when males and females (equal to 5% of all animals) were selected based on their diagonal element of A^{-1} compared with selecting 15% of the animals at random. In the case of a real beef cattle pedigree, selection of males only or males and females yielded similar results and both selection methods were superior to random selection.

Key words: genotype sampling, marker-assisted selection, simulation

J. Anim. Sci. 2008. 86:2471–2479
doi:10.2527/jas.2007-0492

INTRODUCTION

Interest in identifying QTL and genes of economic importance for marker-assisted selection in livestock populations has increased greatly in the past decade. Yet it may not be viable to genotype each animal. Dekkers and Hospital (2002) state that economic considerations are one of the main limitations to marker- or gene-assisted selection. This raises the question of which animals to genotype. A method that would allow for a selected sample (e.g., 5%) of the population to be genotyped and at the same time inferring with high probability genotypes for the remaining animals in the population could be beneficial. By using such a method, fewer animals in a population would be needed for genotyping, which would decrease the time and cost of genotyping. Theoretically, the problem at hand is simple to solve. If it were possible to evaluate every subset of animals equal to the desired size, then the optimal solution could be found. However, this is computationally impossible at the current time and a more feasible solution is needed.

Once the selected animals are genotyped, several methods have been applied for the assignment of alleles to other animals in the population via allelic peeling (Wang et al., 1996; Thallman et al., 2001) or Gibbs sampling (Fernandez et al., 2001). The problem of calculating genotypic probabilities for nongenotyped animals in the presence of sparsely recorded genotypes, as is the case for genetic disorders, is complex and has been addressed in Henshall et al. (2001). However, it could be possible to infer genotypes of all other animals in the population with relatively high accuracy. Therefore, the objectives of the current study were to inves-

[1]Current address: University of Nebraska, Lincoln 68583.
[2]Current address: Aviagen, Huntsville, AL 35805.
[3]Corresponding author: rrekaya@uga.edu
Received August 3, 2007.
Accepted April 16, 2008.

tigate sampling techniques for genotyping a selection of animals and to determine the impact of estimating allele frequencies of selected animals using simulated pedigrees and genotypes. Selected procedures were tested using actual beef cattle pedigrees with simulated genotypes.

MATERIALS AND METHODS

All genotypes were simulated and all pedigrees were either simulated or obtained from field data. Consequently, Animal Care and Use Committee approval was not obtained for this study.

Selecting Animals for Genotyping

Random Sample. To determine the animals for genotyping, a random sample from the population was taken. It was assumed that either 5 or 15% of the population would be randomly selected for genotyping. The random selection scenario was utilized as a method for comparing different selection scenarios based on the relationship between animals in the population.

Relationship Matrix. The inverse of the relationship matrix ($\mathbf{A}^{-1}$) was used for selecting animals for genotyping. Once $\mathbf{A}^{-1}$ was computed, males and females were separated and sorted by their diagonal element of $\mathbf{A}^{-1}$ and the number of progeny. Females were additionally sorted by their number of mates. In the current study, it was assumed that 5% of the population would be selected for genotyping using the relationship between animals, the number of progeny, and the number of mates (females only).

For the scenario in which males and females were selected for genotyping, an equal number of each sex was selected. In other words, 5% of the population, half being males and half being females, were selected for genotyping. Within sexes, animals were ranked by their corresponding diagonal element of $\mathbf{A}^{-1}$ and tied ranks were broken using numbers of progeny (males and females) and number of mates (females only). This was done to maximize the number of alleles known through half-sib relationships. When the number of females within a diagonal element-number of progeny-number of mates group exceeded the number to be selected, females were then selected randomly within that group. Similarly, males were randomly selected within a diagonal element-number of progeny group when the number of males in that group exceeded the number of males to be selected.

When only males were selected, the method of selecting males as described previously was used. For this scenario, 5% of the population selected for genotyping consisted of only males (males with the highest diagonal elements). In other words, the top 5% of males based on their diagonal element of $\mathbf{A}^{-1}$ and number of progeny were selected for genotyping.

Absorption. Selection of animals was based on the diagonal element of either the relationship matrix, **A**, or the inverse of the relationship matrix, $\mathbf{A}^{-1}$. Animals were selected based on their diagonal element. Further, only one animal was selected in the iterative process. The iterative sampling process was run until a total of n animals were selected. The n animals selected were based on genotyping 5% of the animals in the population. In situations where more than one animal had the largest diagonal element, an animal was randomly selected by calling a uniform distribution, $U[0,1]$.

The absorption procedure used in the current study is described below.

$$\mathbf{P} = \mathbf{C} \cdot \left(1 / a^{ii} \right) \cdot \mathbf{R},$$

where $\mathbf{P}$ was a matrix with dimension $n \times n$ ($n = 1,\ldots,n_a$), n_a was the total number of animals in the population, a^{ii} was the diagonal element of animal i in $\mathbf{A}^{-1}$, and $\mathbf{C}$ and $\mathbf{R}$ were column and row vectors, respectively, of the selected animal i. Further,

$$\mathbf{C}_{n \times i} = \begin{bmatrix} a^{1i} \\ \vdots \\ a^{ii} \\ \vdots \\ a^{ni} \end{bmatrix}$$

and

$$\mathbf{R}_{i \times n} = \begin{bmatrix} a^{i1} & \cdots & a^{ii} & \cdots & a^{in} \end{bmatrix}.$$

After absorption of animal i, the new $\mathbf{A}^{-1}$, $\mathbf{N}_{A^{-1}}$, was computed as follows:

$$\mathbf{N}_{A^{-1}} = \mathbf{A}^{-1} - \mathbf{P}.$$

The equations presented here were for selection of animals using $\mathbf{A}^{-1}$. The procedure could be easily converted to selection of animals based on the diagonal element of **A**. However, forming **A** (or inverting $\mathbf{A}^{-1}$ to get **A**) could be time consuming, depending on the structure and size of the pedigree.

Peeling

Given that genotypes in this study were assigned at random from the parental genotypes in the population, it is possible to extract additional genotypic information from the pedigree. Animals with missing genotypic information can be assigned one or both alleles given parental, progeny, or mate information. Given this trio of information sources and following an algorithm similar to Qian and Beckmann (2002) and Tapadar et al. (2000), imputations on missing genotypes were made and additional genotypic information was gar-

nered. The peeling process used in the current study to determine known alleles in the population given the genotypes of animals selected was implemented in 3 steps. For the current study, it was assumed that there were no errors in the recorded pedigree, resulting in all animals having known paternity and maternity. Whenever possible, maternal and paternal alleles were identified based on inheritance. For the purpose of this study, the first allele was inherited from the sire and the second allele was inherited from the dam. If the parental origin of an allele was unclear, then the known allele was arbitrarily assigned as either the paternal or maternal allele.

Statistical Analysis and Computation

After selection of animals for genotyping, the number of animals with 1 or 2 alleles known was computed. This was done by simply counting the number of animals that were assigned either 1 or 2 alleles based on the peeling procedure described above. The percentage of alleles known based on the peeling procedure (**AK$_P$**) was computed as follows:

$$\mathrm{AK_P} = \left(\frac{(n_1 \times 2) + n_2}{n_a \times 2} \right) \times 100, \qquad [1]$$

where n_1 and n_2 were the number of animals with 2 and 1 allele(s) known and n_a was the total number of animals in the population. Furthermore, n_1 and n_a were multiplied by 2, because each animal has 2 alleles.

In this step, an animal with either 1 or 2 allele(s) known was not penalized if the position of the allele(s) was incorrectly assigned. For example, animal *i* was genotyped as *bb* and no information was available about the parent's genotype. Given that each parent had to have passed allele *b* to their progeny, animal *i*, the parent's genotype could then be assigned as _*b* or *b*_, where _ was the unknown allele, *b* was the known allele, _*b* indicated that allele *b* was inherited from the dam, and *b*_ indicated that allele *b* was inherited from the sire. If animal *i*'s sire's true genotype was *b*_ but was assigned as _*b*, then animal *i*'s sire was included in the computations of the number of animals with 1 or 2 alleles known and AK$_P$.

Gibbs Sampling. After the known alleles were determined by the peeling process described above, these alleles were used as prior information in the Gibbs sampler (Wang et al., 1993; Sorenson et al., 1994; Sheehan, 2000; Fernandez et al., 2001) to assign genotypes to the remaining animals in the population. For the base population animals, the unknown allele(s) were randomly sampled given the frequency of alleles in the population and the assumption of Hardy-Weinberg equilibrium. Unknown alleles for nonbase population animals were randomly sampled from the parent's genotypes according to Mendelian rules. An equal weight was assumed for inheriting either the first or second allele from a parent. For a nonbase population animal that had only one unknown allele, the unknown allele was sampled approximately half of the time from the sire's genotype and the remaining time from the dam's genotype. This was to compensate for incorrect assignment of the known allele as illustrated in the above example.

At the end of the sampling process, a benefit function that described the total number of alleles known in the population was computed. This function was computed from a combination of known alleles and the probability of unknown alleles assigned during the sampling process. To be included in the benefit function, an allele in a particular position had to be equal to the true allele of the same position (i.e., *Bb* and *bB* were not equal). The probability of allele $a_{i,j}$ (j = 1 or 2) being assigned as the true allele j for animal i was calculated as:

$$p(a_{i,j}) = \frac{\text{no. of times } \mathrm{a_{i,j}} \text{ was assigned}}{\text{no. of iterations}}. \qquad [2]$$

Using $p(a_{i,j})$ and the number of known alleles, the benefit function was then computed as

$$Benefit = n_1 \times 2 + \sum_{i=1}^{n_2} [1 + p(a_{i,j})] + \sum_{i=1}^{n_3} [p(a_{i,1}) + p(a_{i,2})], \qquad [3]$$

where n_1, n_2, and n_3 were the number of animals with 2, 1, or 0 alleles known, respectively, and $p(a_{i,j})$ as previously defined. The percentage of alleles known after the Gibbs sampling process (**AK$_G$**) was such that

$$\mathrm{AK_G} = \left(\frac{benefit}{n_a \times 2} \right) \times 100, \qquad [4]$$

where *benefit* was the benefit function computed above and n_a was the total number of animals in the population.

During each round of the sampling process, only one genotype for any given animal was assigned as the true genotype. Thus, at the end of the sampling process every animal had a probability of having the true genotype, PTG_{ig}, assigned as

$$PTG_{ig} = \frac{\text{no. of times genotype } g \text{ was assigned}}{\text{total no. of samples}}, \qquad [5]$$

where genotype g was the true genotype of animal i. The average probability of the true genotype being identified for every animal in the population (**APTG**) was computed using the following:

$$\text{APTG} = \frac{\sum_{i=1}^{n_a} \text{PTG}_{ig}}{n_a}, \quad [6]$$

where PTG_{ig} was defined as above and n_a was the total number of animals in the population. In contrast to the benefit function, APTG only required that the animal have the correct genotype—*Bb* was considered the same genotype as *bB*—and therefore was able to compensate for the incorrect allele position and sampling the correct unknown allele.

Simulation

A pedigree with 4 overlapping generations was simulated. The base population included 500 unrelated animals and subsequent generations consisted of 1,500 animals with a total of 5,000 animals generated. Approximately 10% of the animals were sires with approximately 8 progeny per sire and 42% dams with approximately 1.9 progeny per dam. One SNP with 2 alleles was simulated for every animal in the pedigree file. Genotypes of the base population animals were assigned based on allele frequencies. For the 3 subsequent generations, genotypes were randomly assigned using the parent's genotype, where an equal chance of passing either the first or second allele was assumed. Five replicates of the simulated data were generated.

Three different frequencies for the favorable allele were used in the simulation and analyses. The frequencies were 0.30, 0.50, and 0.80. Allele frequencies used in the analyses were either the true frequency (equal to the allele frequency used in the simulation) or estimated from the animals that were selected for genotyping. For the analyses using Gibbs sampling, a total chain length of 25,000 iterations of the Gibbs sampler was run, where the first 5,000 iterations were discarded as burn-in.

Two real beef cattle pedigrees were used to validate the selection scenarios using simulated genotypes. The first pedigree was obtained from a Gelbvieh field data set and was similar, but slightly smaller than, the pedigree used by Sapp et al. (2003) and consisted of 29,101 animals of which approximately 16.4 and 54.8% were sires and dams, respectively. There were approximately 5.7 offspring per sire and 1.7 offspring per dam. The second pedigree was a smaller research pedigree obtained from the USDA-ARS research station at Ft. Keogh (Montana) from the Line 1 Hereford selection project started in 1934 (Kealey et al., 2006) and consisted of 8,688 animals. It comprised approximately 6.6% sires and 33.0% dams. Each sire had 14.6 offspring on average, and each dam had 2.9 offspring on average. For the 2 beef cattle pedigrees, all animals with both parents unknown were assumed to comprise the base population. For these animals, genotypes were assigned based on allele frequencies. For all other animals, genotypes were randomly assigned using the parent's genotype, where there was an equal chance of passing either the first or second allele. Frequencies for the favorable allele were assumed to be either 0.3 or 0.5. The case in which the frequency of the favorable allele was 0.8 was omitted in the field data pedigrees due to the similarity of results in the simulated pedigrees between assuming a frequency of 0.3 or 0.8 for the favorable allele. The same Gibbs sampling procedure mentioned above was used.

RESULTS AND DISCUSSION

General

For all selection scenarios and allele frequencies, estimated allele frequencies were similar to their corresponding true frequencies. The number of animals with either 1 or 2 alleles known and AK_P (percentage of alleles known before Gibbs sampling) were identical when the true or estimated allele frequencies were used. This was because these parameters were estimated before the Gibbs sampling procedure and thereby depend only on the allele frequency used in the simulation. Across the 3 allele frequencies, the parameters that depend on allele frequency—benefit function, AK_G, and APTG—presented very small differences between the true and estimated allele frequency used in the analysis, suggesting that the estimated allele frequency did not have a significant impact on population parameters when different sampling strategies were implemented. Therefore, the results of the current study will be reported using estimated allele frequencies. Given that the estimated frequencies were similar to the true frequencies in all pedigrees, allele frequency will be referred to as the true frequency (i.e., estimated frequency of 0.79 will be referred to as 0.80). Because genotypes were randomly assigned in the base population and as such are not linked to any trait, they are not influenced by selection. In practice, one would expect larger differences between estimated and known allele frequencies if selection pressure has been applied to the trait for which the marker is associated. As the magnitude of this difference increases, the measures of AK_G and APTG would be adversely affected. The correct allele frequency in a population that has undergone artificial selection would be dependent on the amount and duration of selection pressure applied, the magnitude of the association between the marker and trait under selection, and the effect of the marker on fitness traits.

Based on the results of the current study, the allele frequency had an effect on population parameters regardless of the method of selecting animals for genotyping. For all selection scenarios, estimates of all parameters tended to be lowest when an allele frequency of 0.50 was used. Similarly, results indicated that estimates of parameters tended to be greatest when using

Table 1. Number of animals with 1 or 2 alleles known, percentage of alleles known (SD), and probability of assigning the true genotype (SD) when 5% of animals in the population were randomly selected for genotyping[1]

	Estimated allele frequency		
Parameter[2]	0.29 (0.04)	0.49 (0.05)	0.78 (0.04)
No. of animals with			
2 alleles known	258	260	258
1 allele known	528	486	577
Benefit function	6,738	6,018	7,310
AK_P	10.44 (0.007)	10.05 (0.007)	10.94 (0.008)
AK_G	67.38 (2.19)	60.18 (0.67)	73.10 (2.88)
APTG	0.51 (0.02)	0.44 (0.006)	0.58 (0.03)

[1]Results were based on the average of 5 replicates.
[2]Full descriptions of the parameters can be found in Eq. 1–6; AK_P = percentage of alleles correctly assigned after peeling, AK_G = percentage of alleles known after Gibbs sampling, and APTG = average probability of correctly assigning the true genotype.

an allele frequency of 0.80. Further, the results suggest that genotyping strategy depends on the structure of the pedigree and the relative influence of males and females in a particular pedigree.

Random Sample

Five Percent Selected. A description of the number of animals with 1 or 2 alleles known, percentage of alleles known, benefit function, and APTG based on randomly selecting 5% of the population for genotyping is presented in Table 1. Based on the number of animals with 1 or 2 alleles known, the percentage of alleles known before the Gibbs sampling procedure (AK_P) ranged from 10.05 to 10.94. The percentage of alleles known after Gibbs sampling (AK_G) ranged from 60.18 to 73.10, suggesting that 60 to 73% of the alleles in the population were known when the probability that the true allele j (j = 1, 2) of animal i was assigned [$p(a_{i,j})$]. This result suggests that the Gibbs sampler in conjunction with the peeling process was able to identify a larger number of alleles in the population than the peeling process alone. To determine the (dis)advantage of using the Gibbs sampling and peeling procedure (AK_G) compared with using the peeling procedure alone (AK_P) a percentage difference was computed as [(AK_G − AK_P)/AK_P] × 100. Using the percentage difference computed above, the Gibbs sampling procedure increased the percentage of alleles known in the population by over 500% across allele frequencies when compared with using the peeling procedure alone.

In contrast to AK_P, the benefit function used $p(a_{i,j})$ in cases where one or no alleles were known to determine the proportion of alleles known in the pedigree. Furthermore, the benefit function required that alleles not only be equal to the true allele, but also to be haplotype specific (knowing from which parent the allele was inherited), suggesting that the alleles known in the population were inherited from the correct parent. Using the benefit function and AK_G, more alleles were known in the population, and inheritance of alleles was more accurately known, when compared with AK_P.

The average probability of the true genotype being identified for every animal in the population, APTG, ranged from 0.44 to 0.58 for the 3 allele frequencies used in the current study. This result indicates that 44 to 58% of the animals in the population had their true genotype assigned after the peeling and Gibbs sampling processes. The parameter APTG is greatly affected by the number of animals with either one or no alleles known. If there are a large proportion of animals with no alleles known, then APTG would be expected to be lower.

Fifteen Percent Selected. A description of the number of animals with 1 or 2 alleles known, percentage of alleles known, benefit function, and APTG when 15% of the population was randomly selected for genotyping is presented in Table 2. Randomly sampling an additional 10% of the population increased the number of animals with 1 or 2 alleles known compared with only sampling 5% of the population. The parameter AK_P was increased by 172.32, 171.74, and 166.91% for allele frequency 0.30, 0.50, and 0.80, respectively, when 15% of the animals were genotyped compared with sampling 5% of the population for genotyping.

The increase in AK_G due to sampling an additional 10% of the population ranged from 9.02 to 17.53%. When 15% of the animals in the population were selected for genotyping, an increase between AK_G and AK_P ranged from 159 to 173%, using 3 different allele frequencies, further indicating that the benefit function was able to determine more of the alleles in the population.

Approximately a 15 to 27% increase in APTG was observed when 17% of animals in the population were randomly selected compared with randomly selecting 5%. Thus, 56 to 68% of the animals in the population had their true genotype assigned. This result indicates that more animals were assigned, with high probability, their true genotype than when only 5% were randomly selected.

Table 2. Number of animals with 1 or 2 alleles known, percentage of alleles known (SD), and probability of assigning the true genotype (SD) when 15% of animals in the population were randomly selected for genotyping[1]

	Estimated allele frequency		
Parameter[2]	0.29 (0.03)	0.50 (0.04)	0.78 (0.04)
No. of animals with			
2 alleles known	813	813	815
1 allele known	1,218	1,104	1,290
Benefit function	7,611	7,073	7,969
AK_P	28.43 (0.009)	27.31 (0.007)	29.20 (0.005)
AK_G	76.11 (1.26)	70.73 (0.26)	79.69 (1.91)
APTG	0.62 (0.02)	0.56 (0.004)	0.68 (0.02)

[1]Results were based on the average of 5 replicates.
[2]Full descriptions of the parameters can be found in Eq. 1–6; AK_P = percentage of alleles correctly assigned after peeling, AK_G = percentage of alleles known after Gibbs sampling, and APTG = average probability of correctly assigning the true genotype.

Relationship Matrix

Selection of Males and Females. A description of the number of animals with 1 or 2 alleles known, percentage of alleles known, benefit function, and APTG based on selecting 2.5% of males and 2.5% of females in the population using $\mathbf{A}^{-1}$ is presented in Table 3. Because of the large number of animals with 1 or 2 alleles known, AK_P ranged from 34.57 to 37.70 across the 3 allele frequencies used in the current study.

Similarly, AK_G ranged from 75.18 to 82.12% when 2.5% of males and 2.5% of females were selected based on the diagonal element of $\mathbf{A}^{-1}$. An increase of approximately 117.47 to 122.68% was achieved by using the Gibbs sampling procedure over the peeling process alone (AK_G vs. AK_P), further indicating that Gibbs sampling in conjunction with the peeling process was able to assign a large number of alleles in the population.

The average probability of assigning the true genotype for every animal in the population, APTG, was 0.62, 0.56, and 0.68 for frequencies of 0.30, 0.50, and 0.80, respectively, suggesting that 56 to 68% of the animals in the population had their true genotype assigned depending on the allele frequency.

When compared with randomly sampling 5% of the population for genotyping, selection of 2.5% of males and 2.5% of females based on the diagonal element of $\mathbf{A}^{-1}$ and the number of progeny or mates increased AK_P by 243.98 to 245.11% depending on the allele frequency. Likewise, AK_G was increased by 12.34 to 28.93% across the 3 allele frequencies when $\mathbf{A}^{-1}$ was used instead of randomly sampling 5% of the population. When animals were selected based on $\mathbf{A}^{-1}$, APTG was increased by 21.57, 27.27, and 17.24% for allele frequencies of 0.30, 0.50, and 0.80, respectively, compared with randomly selecting 5% of the population.

When compared with randomly sampling 15% of the population for genotyping, selection of 2.5% of males and 2.5% of females based on the diagonal element of $\mathbf{A}^{-1}$ increased AK_P by 26.58 to 29.11% depending on the allele frequency. Likewise, AK_G was increased by 3.05 to 6.29% across the 3 allele frequencies when $\mathbf{A}^{-1}$ was used instead of randomly sampling 15% of the population. When 2.5% of males and 2.5% of females were

Table 3. Number of animals with 1 or 2 alleles known, percentage of alleles known (SD), and probability of assigning the true genotype (SD) when 2.5% of males and 2.5% of females in the population were selected for genotyping using the inverse of the relationship matrix[1]

	Estimated allele frequency		
Parameter[2]	0.29 (0.03)	0.50 (0.05)	0.79 (0.03)
No. of animals with			
2 alleles known	670	652	683
1 allele known	2,263	2,153	2,404
Benefit function	8,023	7,518	8,212
AK_P	36.03(0.007)	34.57 (0.009)	37.70 (0.006)
AK_G	80.23 (1.28)	75.18 (0.56)	82.12 (1.62)
APTG	0.62 (0.02)	0.56 (0.002)	0.68 (0.03)

[1]Results were based on the average of 5 replicates.
[2]Full descriptions of the parameters can be found in Eq. 1–6; AK_P = percentage of alleles correctly assigned after peeling, AK_G = percentage of alleles known after Gibbs samplin, and APTG = average probability of correctly assigning the true genotype.

Table 4. Number of animals with 1 or 2 alleles known, percentage of alleles known (SD), and probability of assigning the true genotype (SD) when 5% of males in the population were selected for genotyping using the inverse of the relationship matrix[1]

	Estimated allele frequency		
Parameter[2]	0.30(0.04)	0.51(0.05)	0.78 (0.04)
No. of animals with			
2 alleles known	250	251	251
1 allele known	2,940	2,793	3,115
Benefit function	7,941	7,402	8,121
AK_P	34.40 (0.005)	32.94 (0.005)	36.17 (0.01)
AK_G	79.41(1.68)	74.02 (0.54)	81.21(1.72)
APTG	0.59 (0.02)	0.52 (0.003)	0.66 (0.03)

[1]Results were based on the average of 5 replicates.

[2]Full descriptions of the parameters can be found in Eq. 1–6; AK_P = percentage of alleles correctly assigned after peeling, AK_G = percentage of alleles known after Gibbs sampling, and APTG = average probability of correctly assigning the true genotype.

selected based on $\mathbf{A}^{-1}$, APTG was virtually identical compared with randomly selecting 15% of the population. The results comparing a relationship-based selection scheme versus random sampling should not be surprising. Kinghorn (1999) described the advantages of selection based on average numerator relationship as being superior to that of random selection using a much smaller pedigree (1,260 animals). The results from Kinghorn (1999) did not show the magnitude of separation between random sampling and the use of connectedness as the current study presumably due to differences in pedigrees, particularly size.

Selection of Males. A description of the number of animals with 1 or 2 alleles known, percentage of alleles known, benefit function, and APTG when 5% of males in the population were selected for genotyping using $\mathbf{A}^{-1}$ is presented in Table 4. Because only males were selected for genotyping, the number of animals with 2 alleles known was approximately 250 across the 3 allele frequencies used. Yet, the number of animals with 1 allele known ranged from 2,793 to 3,115, which was higher than with any of the other selection scenarios using the simulated pedigrees. However, due to the method of selecting both males and females having over twice the number of animals with both alleles known, the method of selecting equal numbers of both sexes yielded greater values for AK_P, AK_G, and APTG. For the measures of AK_P, and AK_G in particular, this method is still more desirable than selecting 5 or even 15% of the animals at random.

Absorption

Inverse of the Relationship Matrix. A description of the parameters estimated when 5% of the population was selected for genotyping using absorption of $\mathbf{A}^{-1}$ is presented in Table 5. The method of absorption was only performed on the simulated pedigrees. The results are similar when the allele frequency is known compared with when it is estimated from the selected animals. This was due to the fact that the estimated allele frequencies are close to the true values. The scenario when the allelic frequencies are 0.8/0.2 gives the most desirable results. Although the differences in the number of animals with both alleles known are negligible across allele frequencies, differences in the number of animals with one allele known are more prominent. Consequently, there are not observable differences in the benefit function, AK_P, AK_G, and APTG across allele frequencies. From these results it appears that, in situations with more extreme allele frequencies (0.8/0.2), it is easier to infer unknown genotypes.

The method of absorption using $\mathbf{A}^{-1}$ is superior to the case when absorption is performed using $\mathbf{A}$ both before and after the Gibbs procedure. However, the advantages of this method compared with the selection of animals based on diagonal elements of $\mathbf{A}^{-1}$ (Tables 3 and 4) varied. When compared with the case of selecting only males (Table 4), the current method had slight advantages in the number of animals with 2 alleles assigned, AK_P, AK_G, and APTG. The method of selecting both males and females was superior because of the much larger number of animals with both alleles assigned before the Gibbs method.

Relationship Matrix. The results of animals selected based on the absorption of $\mathbf{A}$ are not reported. This was due to the observation of similar trends for those reported using absorption of $\mathbf{A}^{-1}$ across the 3 allele frequencies. Selection of animals based on absorption of $\mathbf{A}$ was inferior to both selection methods of animals based on their diagonal elements of A^{-1} (Tables 3 and 4). The absorption of $\mathbf{A}$ still has advantages, albeit slight, in regard to AK_P over the method of selecting 5% of the animals at random.

Real Beef Cattle Pedigrees

The results using a field data pedigree of 29,101 animals are presented in Table 6. Similar patterns to the results using the simulated pedigrees were observed. Selecting candidates for genotyping (5% of population) using random selection, selection of males with

Table 5. Number of animals with 1 or 2 alleles known, percentage of alleles known (SD), and probability of assigning the true genotype (SD) when 5% of the population were selected for genotyping using absorption of the inverse of the relationship matrix[1]

	Estimated allele frequency		
Parameter[2]	0.30 (0.04)	0.51 (0.05)	0.78 (0.04)
No. of animals with			
2 alleles known	288	284	287
1 allele known	2,906	2,753	3,074
Benefit function	7,954	7,415	8,129
AK_P	34.83 (0.007)	33.21 (0.006)	36.48 (0.01)
AK_G	79.54 (1.73)	74.15 (0.60)	81.29 (1.73)
APTG	0.60 (0.02)	0.53 (0.03)	0.66 (0.03)

[1]Results were based on the average of 5 replicates.

[2]Full descriptions of the parameters can be found in Eq. 1–6; AK_P = percentage of alleles correctly assigned after peeling, AK_G = percentage of alleles known after Gibbs sampling, and APTG = average probability of correctly assigning the true genotype.

the greatest diagonal element of $\mathbf{A}^{-1}$ and selecting both males and females from their diagonal element of $\mathbf{A}^{-1}$ were compared. As expected from the simulation results, selection of candidates based on the relationship matrix yielded more desirable results compared with random selection. The advantages in AK_P for selection of both males and females based on their diagonal element of $\mathbf{A}^{-1}$ over random selection were 163.6 and 160.4% for allele frequencies 0.3/0.7 and 0.5/0.5, respectively. Similarly, AK_G increased by 65.3 and 69.1% and APTG increased by 12.8 and 14.3% for the more extreme (0.3/0.7) and intermediate (0.5/0.5) allele frequencies, respectively.

Selection of males appeared to be the most desirable selection method. This method showed increases in AK_P of 166.5 and 163.3% and increases in AK_G of 69.6 and 73.5% over random selection for allele frequencies of 0.3/0.7 and 0.5/0.5, respectively. Likewise, advantages in APTG were 11.4% for the intermediate allele frequency and 12.8% for the more extreme frequency.

Compared with the simulated pedigrees, the beef cattle pedigree used here appears to be best suited for selection based on animals with the greatest diagonal element of $\mathbf{A}^{-1}$ as opposed to selection of equal proportions of males and females. This can be explained by the fact that in the field data pedigree, numerous females had a small number of mates and offspring. This is in agreement with Koudande et al. (1999) who determined that when the reproductive rate of males is sufficiently high compared with that of females, genotyping costs can be reduced by genotyping males only. Although the results in Table 6 show that the differences between selecting both males and females or just males are slight, it does show that pedigrees with varying levels of complexities (or livestock species) might respond differently to these selection methods.

Table 7 displays the results of selection using a research pedigree. The results show that selection of males based on their diagonal element of $\mathbf{A}^{-1}$ led to increases in AK_P, AK_G, and APTG of 213.7, 45.1, and 18.6% for allele frequency of 0.5 and 265.5, 45.2, and 38.0% for allele frequency of 0.3 when compared with randomly selecting 5% of the population. Selection of both males and females based on their diagonal element of $\mathbf{A}^{-1}$ was also superior to randomly selecting 5%, showing increases in AKP, AKG, and APTG of

Table 6. Number of animals with 1 or 2 alleles known, percentage of alleles known, and probability of assigning the true genotype using a field data pedigree[1]

	Random		Males		Males and females	
Parameter[2]	(0.30)	(0.50)	(0.30)	(0.50)	(0.30)	(0.50)
No. of animals with						
2 alleles known	1,505	1,501	1,473	1,470	2,086	1,999
1 allele known	2,508	2,144	11,756	10,607	10,376	9,398
Benefit function	20,569	18,609	34,877	32,282	34,005	31,456
AK_P	9.48	8.84	25.26	23.28	24.99	23.02
AK_G	35.34	31.97	59.92	55.47	58.43	54.05
APTG	0.39	0.35	0.44	0.39	0.44	0.40

[1]Random = 5% selected at random; Males = 5% of males selected from their diagonal element of A^{-1}; Males and females = 2.5% males and 2.5% females selected from their diagonal element of A^{-1}. Numbers in parentheses are the true allele frequencies used in the simulation.

[2]Full descriptions of the parameters can be found in Eq. 1–6; AK_P = percentage of alleles correctly assigned after peeling, AK_G = percentage of alleles known after Gibbs sampling, and APTG = average probability of correctly assigning the true genotype.

Table 7. Number of animals with 1 or 2 alleles known, percentage of alleles known, and probability of assigning the true genotype using a research pedigree[1]

	Random		Males		Males and females	
Parameter[2]	(0.30)	(0.50)	(0.30)	(0.50)	(0.30)	(0.50)
No. of animals with						
2 alleles known	452	458	438	439	1,082	751
1 allele known	847	682	5,525	4,132	4,747	3,768
Benefit function	9,719	8,284	14,113	12,018	13,743	11,848
AK_P	10.08	9.19	36.84	28.83	39.77	30.33
AK_G	55.94	47.68	81.22	69.16	79.09	68.19
APTG	0.50	0.43	0.69	0.51	0.68	0.52

[1]Random = 5% selected at random; Males = 5% of males selected from their diagonal element of A^{-1}; Males and females = 2.5% males and 2.5% females selected from their diagonal element of A^{-1}. Numbers in parentheses are the true allele frequencies used in the simulation.

[2]Full descriptions of the parameters can be found in Eq. 1–6; AK_P = percentage of alleles correctly assigned after peeling, AK_G = percentage of alleles known after Gibbs sampling, and APTG = average probability of correctly assigning the true genotype.

230.0, 43.0, and 20.9% for allele frequency of 0.5 and increases of 294.5, 41.4, and 36.0% for 0.3 allele frequency. The methods of selecting only males or both males and females from their diagonal element were similar in performance, with the selection of both sexes having an advantage before the Gibbs method and the selection of males having a slight advantage after the Gibbs method.

Information concerning which animals to genotype in a given pedigree is obviously critical to the viability of marker- or gene-assisted selection. The results of the current study show that random selection is not desirable and numerous alternatives exist. Further, these results show that similar outcomes can be achieved regardless of whether the allele frequencies are known or estimated. Of the alternatives presented here, selection of animals based on their diagonal element of the inverse of the relationship matrix appears to be the most desirable solution. The proportion of males and females selected may depend on the particular pedigree. Other alternatives may exist and further investigation is warranted to explore other possibilities. It should also be noted that every pedigree will offer its own challenges due to its intrinsic structure, and the application of the methods presented here are limited to the pedigrees used in the current study.

LITERATURE CITED

Dekkers, J. C., and F. Hospital. 2002. The use of molecular genetics in the improvement of agricultural populations. Nat. Rev. Genet. 3:22–32.

Fernandez, S. A., R. L. Fernando, B. Guldbrandsten, L. R. Totir, and A. L. Carriquiry. 2001. Sampling genotypes in large pedigrees with loops. Genet. Sel. Evol. 33:337–367.

Henshall, J. M., B. Tier, and R. J. Kerr. 2001. Estimating genotypes with independently sampled descent graphs. Genet. Res. 78:281–288.

Kealey, C. G., M. D. MacNeil, M. W. Tess, T. W. Geary, and R. A. Bellows. 2006. Genetic parameter estimates for scrotal circumference and semen characteristics of Line 1 Hereford bulls. J. Anim. Sci. 84:283–290.

Kinghorn, B. P. 1999. Use of segregation analysis to reduce genotyping costs. J. Anim. Breed. Genet. 116:175–180.

Koudande, O. D., P. C. Thomson, and J. A. M. van Arendonk. 1999. A model for population growth of laboratory animals subjected to marker-assisted introgression: How many animals do we need? Heredity 82:16–24.

Qian, D., and L. Beckmann. 2002. Minimum-recombinant haplotyping in pedigrees. Am. J. Hum. Genet. 70:1434–1445.

Sapp, R. L., R. Rekaya, and J. K. Bertrand. 2003. Simulation study of teat score in first-parity Gelbvieh cows: Parameter estimation. J. Anim. Sci. 81:2959–2963.

Sheehan, N. A. 2000. On the application of Markov Chain Monte Carlo methods to genetic analysis of complex pedigrees. Int. Stat. Rev. 68:83–110.

Sorenson, D., C. S. Wang, J. Jensen, and D. Gianola. 1994. Bayesian analysis of genetic change due to selection using Gibbs sampling. Genet. Sel. Evol. 26:229–249.

Tapadar, P., S. Ghosh, and P. P. Majumder. 2000. Haplotyping in pedigrees via a genetic algorithm. Hum. Hered. 50:43–56.

Thallman, R. M., G. L. Bennett, J. W. Keele, and S. M. Kappes. 2001. Efficient computation of genotype probabilities for loci with many alleles: I. Allelic peeling. J. Anim. Sci. 79:26–33.

Wang, C. S., J. J. Rutledge, and D. Gianola. 1993. Marginal inferences about variance components in a mixed linear model using Gibbs sampling. Genet. Sel. Evol. 25:41–62.

Wang, T., R. L. Fernando, C. Stricker, and R. C. Elston. 1996. An approximation to the Likelihood for a pedigree with loops. Theor. Appl. Genet. 93:1299–1309.