Animal (2007), **1**: Pages 21–28 © The Animal Consortium 2007
doi: 10.1017/S1751731107392628

animal

A simple method to approximate gene content in large pedigree populations: application to the myostatin gene in dual-purpose Belgian Blue cattle

N. Gengler[1,2†], P. Mayeres[1,3] and M. Szydlowski[1]

[1]*Animal Science Unit, Gembloux Agricultural University, B-5030 Gembloux, Belgium;* [2]*National Fund for Scientific Research, B-1000 Brussels, Belgium;* [3]*Department of Research and Development, Walloon Breeding Association, B-5590 Ciney, Belgium*

(Received 13 July 2006; Accepted 2 November 2006)

Gene content is the number of copies of a particular allele in a genotype of an animal. Gene content can be used to study additive gene action of candidate gene. Usually genotype data are available only for a part of population and for the rest gene contents have to be calculated based on typed relatives. Methods to calculate expected gene content for animals on large complex pedigrees are relatively complex. In this paper we proposed a practical method to calculate gene content using a linear regression. The method does not estimate genotype probabilities but these can be approximated from gene content assuming Hardy-Weinberg proportions. The approach was compared with other methods on multiple simulated data sets for real bovine pedigrees of 1 082 and 907 903 animals. Different allelic frequencies (0.4 and 0.2) and proportions of the missing genotypes (90, 70, and 50%) were considered in simulation. The simulation showed that the proposed method has similar capability to predict gene content as the iterative peeling method, however it requires less time and can be more practical for large pedigrees. The method was also applied to real data on the bovine myostatin locus on a large dual-purpose Belgian Blue pedigree of 235 133 animals. It was demonstrated that the proposed method can be easily adapted for particular pedigrees.

Keywords: Belgian Blue, cattle, gene content, myostatin

Introduction

Gene content is the number of copies of a particular allele in a genotype of an animal (0, 1 or 2). Information on gene content can be used to estimate the effect of a candidate gene, in evaluation of the total genetic merit including polygenic and major gene effects or for identifying favourable selection candidates based on disease gene status. Gene content is known for genotyped animals but usually only a fraction of animals has their genotype known. For ungenotyped animals gene content can be predicted from the records on their typed relatives. Methods for exact calculation of gene content are not practical for large pedigrees (Lauritzen and Sheehan, 2003). Approximate methods have also been proposed but they are complex, require custom computer programs and may be slow when applied to pedigrees of millions of animals (Kong, 1991; Thallman *et al*., 2001).

The double-muscling condition is known in several breeds. The first description was made in Shorthorn cattle and it seems that the Belgian Blue breed (BBB) has inherited the condition from them. After 1950 the selection of double-muscled animals in the BBB was increased strongly and selection for milk production discontinued. In 1974 two strains were defined: meat and dual-purpose. The milking cows are mostly dual-purpose registered (DP-BBB), however some meat strain registered animals (M-BBB) are still milked. The double-muscling condition is under the control of a major gene but modified by other genes. Charlier *et al.* (1995) identified the major gene in BBB as being a mutation inactivating myostatin. The M-BBB animals are homozygous for the mutated allele called muscular hypertrophy (mh). In DP-BBB three genotypes are encountered: +/+, mh/+ and mh/mh. All DP-BBB sires have to be genotyped before they can be registered and over the years top breeders genotyped their cows too. Breeders use the genotype at the myostatin locus as a criterion in selection.

In this paper we present a practical method to calculate gene content in large pedigrees. In simulation study we compare the proposed method with other approximate techniques: Markov chain Monte Carlo (MCMC), iterative

† E-mail: gengler.n@fsagx.ac.be

peeling and the procedure of Israel and Weller (1998). We consider the calculated gene content for further calculation of approximate genotype probabilities. Finally, we exemplify the use of the method in a study of the mh mutation for the myostatin locus in DP-BBB cows under milk recording.

Material and methods

Method

To present the method we consider inheritance at single biallelic locus with hypothetical alleles *B* and *b*. Let q be the number of copies of allele *B* carried by an individual (gene content). For an animal that has been genotyped q is 0, 1 or 2. For an individual that has not been genotyped but is the progeny of genotyped parents the expected value for q is $E(q_p) = 0.5(q_s + q_d)$, with q_s, q_d and q_p being the content of allele *B* for sire, dam and progeny. We can replace q by its deviation from population mean, $d = q - \mu$. Then, expected value for a progeny of genotyped parents is $E(q_p) = \mu(d_s + d_d)$. If mother has not been genotyped her expected value has to be replaced by population mean, $q_d = \mu$. Further, if a relative with genotype known is not a parent or offspring expected vales for q can be derived using a recursive approach. For example, if maternal grand-sire (mgs) has been genotyped, the expected value of q for mother is $E(q_d|d_{mgs}) = \mu + 0.5d_{mgs}$, and expectation for progeny is $E(q_s|d_s,d_{mgs}) = \mu + 0.5d_s + 0.25d_{mgs}$. Up to this point the method is similar to the approach by Israel and Weller (1998). Their method, however, ignores information on descendents that usually are the main source of information. To use information from all genotyped relatives we apply known techniques developed for continuous traits. Further, we treat gene content as a continuous variable and assume that the relationship between gene contents is linear, and that the covariance between gene contents is proportional to the additive relationship between animals. Although the assumptions are violated we will show in simulation study that the method is robust enough to enable practical calculations. Under these assumptions the conditional expectation of gene contents for all ungenotyped animals given molecular and pedigree data is

$$\mathbf{q}_x = \begin{pmatrix} \mathbf{1} & \mathbf{A}_{xy}\mathbf{A}_y^{-1} \end{pmatrix} \begin{pmatrix} \mu \\ \mathbf{q}_y - \mathbf{1}\mu \end{pmatrix}$$

where $\mathbf{q}_x$ is an unknown vector of gene contents for animals that have not been genotyped, $\mathbf{1}$ is a vector of one's, $\mathbf{q}_y$ is a known vector of gene contents for animals that have been genotyped, $\mathbf{A}_{xy}$ is an additive relationship matrix between individuals with unknown genotype and their genotyped relatives, $\mathbf{A}_y$ is the additive relationship matrix for genotyped individuals. Note that the mean value (μ) can be easily calculated if we assume that animals are typed without error.

A more practical way of computation is derived under an incomplete penetrance model. Here, the incomplete penetrance model incorporates the probability of errors in marker phenotypes and is more accurate for real marker data. Assuming small error variance component (σ_e^2) in the variability of $\mathbf{q}$, we use mixed model equations that provides at the same time the solution for μ:

$$\begin{pmatrix} \mathbf{1}'\mathbf{1} & \mathbf{1}'\mathbf{M} \\ \mathbf{M}'\mathbf{1} & \mathbf{M}'\mathbf{M} + \mathbf{A}^{-1}\varepsilon \end{pmatrix} \begin{pmatrix} \hat{\mu} \\ \hat{\mathbf{d}}_y \\ \hat{\mathbf{d}}_x \end{pmatrix} = \begin{pmatrix} \mathbf{1}'\mathbf{q}_y \\ \mathbf{M}'\mathbf{q}_y \end{pmatrix}$$

where $\mathbf{d}_y$ is a vector of gene content deviations for animals with genotype records, $\mathbf{d}_x$ is a vector of gene content deviations for unobserved animals, $\mathbf{M}$ is incidence matrix linking $\mathbf{q}_y$ to $\begin{pmatrix} \mathbf{d}_y \\ \mathbf{d}_x \end{pmatrix}$ that can be rewritten as $\begin{pmatrix} \mathbf{I}_y & \mathbf{0}_x \end{pmatrix}$, $\mathbf{A}$ is the additive relationship matrix of the structure $\mathbf{A} = \begin{pmatrix} \mathbf{A}_{yy} & \mathbf{A}_{yx} \\ \mathbf{A}_{xy} & \mathbf{A}_{xx} \end{pmatrix}$, $\varepsilon = \sigma_e^2/\sigma_d^2$. The order of animals was chosen to be genotyped before ungenotyped, in practice however, the order is of user choice. For this model the assumptions of BLUP methodology are applied. Some small error variance is required to solve the system of equations. In practice, for the model of almost complete penetrance (small probability of mistyping) the ε can be set equal to 0.01 or smaller. A large value for ε would mean that probability of mistyping or pedigree error is high. Such data would be useless.

This method has some important advantages: standard genetic evaluation software can be used and gene contents can be easily computed for the whole population; population mean gene content will be computed at the same time; the model can be adapted to account for genetic groups due to different origins of animals; ratio ε between variances can be small but still different from zero, therefore the method allows for rare but not impossible genotyping errors, ε may also account for pedigree errors, that could be in certain cases up to over 20% of the animals (Banos *et al*., 2001).

Simulation study

The proposed method has been compared with three other methods: MCMC approach, iterative peeling and the method developed by Israel and Weller (1998). All three methods approximate genotype probabilities for each animal in a pedigree. Exact genotype probabilities are marginal probabilities obtained by summing over all possible genotype configurations which are consistent with observations, weighting the configurations by their probability of occurrence as calculated under the assumed genetic model. On a small pedigree exact genotype probabilities can be calculated using one of the graphical models and reverse or simultaneous peeling (Thompson, 1981). The problem considered in the simulation study was too

https://doi.org/10.1017/S1751731107392628 Published online by Cambridge University Press

complex for exact calculation, therefore only approximate methods were applied. The approximate probabilities computed by MCMC, iterative peeling and the method of Israel and Weller (1998) were used to calculate expected gene content of an animal applying the following formula $E(q) = 2P(BB) + P(Bb)$. These results were used to compare with values obtained from the proposed method. The proposed method was implemented with BLUPF90 (Misztal, 1999) and all other methods with our own software. To compare the methods 200 data sets have been randomly generated for a real bovine pedigree and the data sets have been analysed by all four methods.

The pedigree used in the simulations was extracted from the DP-BBB pedigree data. All cows of a single herd and their ancestors were represented. The pedigree consisted of 1082 individuals: 190 founders, 892 non-founders. Within the pedigree 716 nuclear families (pairs of parents with their siblings) and 531 loops were identified. On a pedigree graph, a loop occurs when an animal can be connected to itself through parents or progeny.

The milking BBB population is partially genotyped at the myostatin locus (recent actual DP-BBB animals). The genotype records were available for 129 individuals born from 1969 through 2001 (Table 1). The structure of this real data set was used in the simulation but real records were replaced by simulated values. Simulation of a complete data set for all pedigree members enables a comparison of the calculated gene content for an animal to its simulated value. A complete data set was generated for biallelic polymorphic system by the use of random process of gene dropping. To make the structure of the simulated data set similar to real data set, the complete simulated data set was reduced to 129 records for the animals, for which real records exist (Table 1), and the reduced data set was analysed by all four methods. Two frequencies of *B* allele were considered: 0.4 and 0.2. One hundred data sets were generated and analysed under each of the two gene frequencies.

The accuracy of the proposed method for different proportions of missing genotypes was tested on huge bovine pedigree of 907 903 animals (see below). Genotypes at a single biallelic locus were simulated under equal allelic frequency by the gene dropping method. Three proportions of missing genotypes were considered: 90%, 70% and 50%.

MCMC method

In the MCMC method genotype probabilities are calculated from multiple random samples of animals' genotypes which are consistent with pedigree and molecular information. To sample the genotype configuration over the pedigree we used the whole locus sampler (Kong, 1991). The sampler uses a peeling algorithm to sample genotypes of all animals jointly from the desired distribution. The pedigree was too complex to peel the pedigree considering the total genotype space. Thus, the genotype space was reduced using the concept of set-recoding and fuzzy inheritance (O'Connell and Weeks, 1995). In brief, two possible alleles of an individual were replaced with a set $s = \{B,b\}$ if they were not observed on the individual nor on his descendents. For example, if within a family of sire, dam and progeny only the sire has its genotype known, say *BB*, their recoded ordered genotypes were $\{B\}|\{B\}$, $\{B,b\}|\{B,b\}$ and $\{B\}|\{B,b\}$, respectively. Considering fuzzy inheritance, if a parent has genotype $s_1|s_2$ and its child has allele set s_3, then s_3 is inherited from the parent if $s_1 \subseteq s_3$ or $s_2 \subseteq s_3$. To sample genotypes we used the following scheme in each iteration: (1) a random ordered genotype configuration on reduced genotype space was obtained by random propagation after peeling the pedigree toward a root, the ordered genotypes were considered to differentiate *Bb* and *bB* heterozygotes, (2) segregation indicators (SI) were set based on current ordered genotypes or sampled randomly if both values were possible, (3) the gene frequencies were sampled based on current allelic types (not recoded) in founders, (4) each recoded allele of a founder was replaced by random allele based on current allelic frequencies, and, starting from the top of the pedigree, each recoded allele of a non-founder was replaced by parental allele according to its SI value. Note, for these parts of the pedigree, for which no data was available, the procedure was similar to simple gene dropping. This sampling procedure guarantees irreducibility and has good mixing properties. When experimenting with small pedigrees it provided solutions almost equal to exact values calculated by the use of a peeling algorithm (Elston and Stewart, 1971) and PAP package (Hasstedt, 2002). For each of the 200 simulated data sets we ran 100 000 iterations collecting samples from all iterations. The final solutions for genotype probabilities were calculated as averages from 100 000 samples. Note, as the whole pedigree was sampled jointly the samples were correlated only through gene frequency and the autocorrelation of the samples was very low.

Table 1 *Number of animals with genotype records within DP-BBB pedigree of 1082 individuals used in simulation study*

Year of birth	Bulls	Cows
1969–1989	27	–
1990–1995	6	28
1996–2001	4	64

Iterative peeling

The iterative peeling exploits the fact that given genotype probabilities of an animal's neighbours (parents, offspring and mates) the genotype of that animal depends only on neighbours and it is independent from the rest of the pedigree. Starting from some initial values genotype probabilities are updated individual by individual until convergence (Van Arendonk *et al.*, 1989; Janss *et al.*, 1995; Wang *et al.*, 1996, Thallman *et al.*, 2001). It was demonstrated on small pedigrees that the iterative peeling may lead to genotype probabilities close to exact calculations (Fernandez *et al.*, 2001), however, the behaviour of iterative peeling on large

https://doi.org/10.1017/S1751731107392628 Published online by Cambridge University Press

complex pedigrees is unknown. We used the method for comparison with the proposed approach on simulated and real data sets. In simulation study the gene frequency of mh was assumed to be known at the simulated value, and for analysis of real data on the myostatin locus the frequency was set equal to 0.05.

Analysis of myostatin locus

Sparse molecular data on the myostatin locus was used to exemplify the utilisation of the method on a huge bovine pedigree for which exact method is impossible and MCMC methods difficult. The population consisted of 907 903 animals born between 1960 and 2005 derived from Belgian dairy and dual-purpose cattle data base. Within the population 34% individuals were Holstein Friesian, 26% were BBB under milk recording including both M-BBB and DP-BBB strains, 10% were Red and White, 1% of other origin and 29% were crosses of various breeds. Within the population 1865 animals have been genotyped for the mh mutation at the myostatin locus. Additionally, recent M-BBB sires were considered homozygous for the mh gene and all dairy breeds were assumed free of the mh gene. In total, there were 381 742 records on the mh gene content and the mh gene content was estimated for the rest (58%) of the population. Including dairy breeds was necessary because ancestors could come from those breeds. The data set was also analysed with modified model linking progeny of unknown parents to unknown-parent groups (Westell *et al*., 1988).

Results

Simulation study

The study showed that calculation of gene content when only 12% of animals have genotype records may be difficult. Correlation coefficients between simulated and calculated gene contents under gene frequency 0.4 and 0.2 were 0.50 and 0.47, respectively. Similar correlations were calculated when gene contents were calculated by the use of MCMC approach (0.52 and 0.42) and iterative peeling (0.52 and 0.40). The proposed method showed significant improvement over the method of Israel and Weller (1998), for which the correlation coefficients were 0.38 and 0.38.

Considering MCMC solutions and allelic frequency of 0.4, the mean of calculated gene content for animals with simulated genotype *bb* (simulated $q = 0$) was 0.59, for alternative homozygote *BB* (simulated $q = 2$) only 1.14. The range between these means was only 28% of the true range of 2 (Table 2). For lower degree of polymorphism (allelic frequency of 0.2) the range between means was 26% of the true range. In general, the bias increases for rare genotype. Standard deviation for rare genotype *BB* was smaller than for more common genotypes, a consequence of the strong bias of the mean for this genotype. Iterative peeling provided solutions close to MCMC values. The proposed method behaves slightly worse. For population allelic frequency of 0.4, the range between means for alternative homozygotes was 25% of the true range, and for frequency of 0.2 the range decreased to 21%. In the later case the bias decreased for the common homozygote and increased for the rare homozygote. We considered the errors defined as the difference between simulated and calculated gene contents. Compared with the MCMC method the mean square error (results not shown) was smaller in case of the more common genotype (*bb*) and higher for the rare genotype (*BB*). For allelic frequency of 0.2 the total mean square error calculated across three genotypes was considerably lower (.222) than for MCMC method (.288).

The behaviour of the proposed method for different proportions of missing genotypes is presented in Table 3.

Table 2 *Results of simulation study: means and standard deviations (s.d.) of predicted gene B content for data sets simulated under two different gene frequencies 0.4 and 0.2 (for comparison solutions obtained by MCMC approach, iterative peeling and the method by Israel and Weller (1998) are presented)*

	MCMC		Iterative peeling		Israel and Weller (1998) method		Proposed method	
Simulated genotype	Mean	s.d.	Mean	s.d.	Mean	s.d.	Mean	s.d.
P(*B*) = 0.4								
bb ($q = 0$)	0.59	0.33	0.59	0.33	0.65	0.23	0.59	0.30
Bb ($q = 1$)	0.89	0.30	0.89	0.30	0.80	0.25	0.85	0.28
BB ($q = 2$)	1.14	0.22	1.14	0.26	0.95	0.27	1.09	0.28
P(*B*) = 0.2								
bb ($q = 0$)	0.46	0.31	0.50	0.35	0.33	0.17	0.26	0.19
Bb ($q = 1$)	0.73	0.29	0.78	0.30	0.48	0.24	0.49	0.26
BB ($q = 2$)	0.97	0.27	1.02	0.27	0.61	0.31	0.68	0.34

Table 3 *Results of simulation study: means and standard deviations (s.d.) of predicted gene B content for data sets simulated with different proportions of missing genotypes. The statistics were calculated for N animals with no records (for comparison solutions obtained by the iterative peeling method are presented)*

Proportion of missing genotypes	N	Iterative peeling		Proposed method	
		Mean	s.d.	Mean	s.d.
90%					
bb ($q = 0$)	284 953	0.46	0.28	0.62	0.33
Bb ($q = 1$)	434 830	0.74	0.37	0.92	0.32
BB ($q = 2$)	178 063	1.00	0.44	1.20	0.30
70%					
bb ($q = 0$)	217 773	0.40	0.29	0.55	0.34
Bb ($q = 1$)	334 305	0.79	0.39	0.93	0.34
BB ($q = 2$)	137 002	1.17	0.47	1.31	0.33
50%					
bb ($q = 0$)	150 018	0.36	0.29	0.50	0.34
Bb ($q = 1$)	249 611	0.84	0.33	0.94	0.36
BB ($q = 2$)	96 562	1.29	0.46	1.38	0.33

https://doi.org/10.1017/S1751731107392628 Published online by Cambridge University Press

For comparison, the data sets were also analysed by the use of the iterative peeling method. For the proposed method, the correlation between simulated and estimated gene content was 0.54 when 90% of animals had no genotype records, and increased to 0.62 and, 0.67 for the data sets with 70 and 50% of missing genotypes in the pedigree. The corresponding coefficients calculated for the iterative peeling method were slightly lower: 0.48, 0.58 and 0.65. The differences between means for alternative homozygotes were comparable between the two methods, only for the most dense data set (50% of missing genotypes) did the iterative peeling algorithm slightly surpassed the proposed method. The amount of time required to complete the calculation by the use of the proposed method was much shorter than that needed for the iterative peeling algorithm. In the case of 90% missing genotypes the proposed method took 4 min (by the preconditioned conjugate gradient algorithm) and the iterative peeling method took almost 4 h (32 iterations). For the proposed method the amount of time increased with the number of records but rapidly decreased in the case of the iterative peeling. For the dense data set (50% of missing genotypes) the proposed method was still three times faster than the iterative peeling algorithm.

Analysis of myostatin locus

Content of the mh gene was estimated for 526 161 animals, including 235 133 animals of DP-BBB breed. Two models were used for the proposed method: the standard model as presented above and a model with unknown parent groups as described by Westell *et al.* (1988). The model with unknown parent groups was used to account for selection for mh allele in BBB animals. Cows were assigned to six groups: five for BBB born in different time periods and one for all non-BBB animals. Similar groups were created for sires. For DP-BBB animals mean mh content was calculated for each year of birth (Figure 1). Results from the two models showed rapid increase in the frequency of mh allele. The standard model indicates that the frequency of mh gene increased since 1970, while the model with unknown parent groups shows that a significant increase of the gene started 10 years later. It is unclear which of the two models better fits the history of the DP-BBB breed.

For comparison with the proposed method, the content of mh gene was also calculated by the use of the iterative peeling algorithm (Figure 1). If only one genetic group was assumed, the calculated mean mh content for old animals was very high. This result should be considered incorrect because it is known that in the early years the frequency of the mh gene was rather low. An additional analysis was performed by the use of the iterative peeling and multiple genetic groups. All founders were assigned to the same 12 groups and frequency of mh gene in each group was assumed to be known. The assumed values for the mh frequency in each group were taken from the results obtained by the proposed method. Two analyses by the

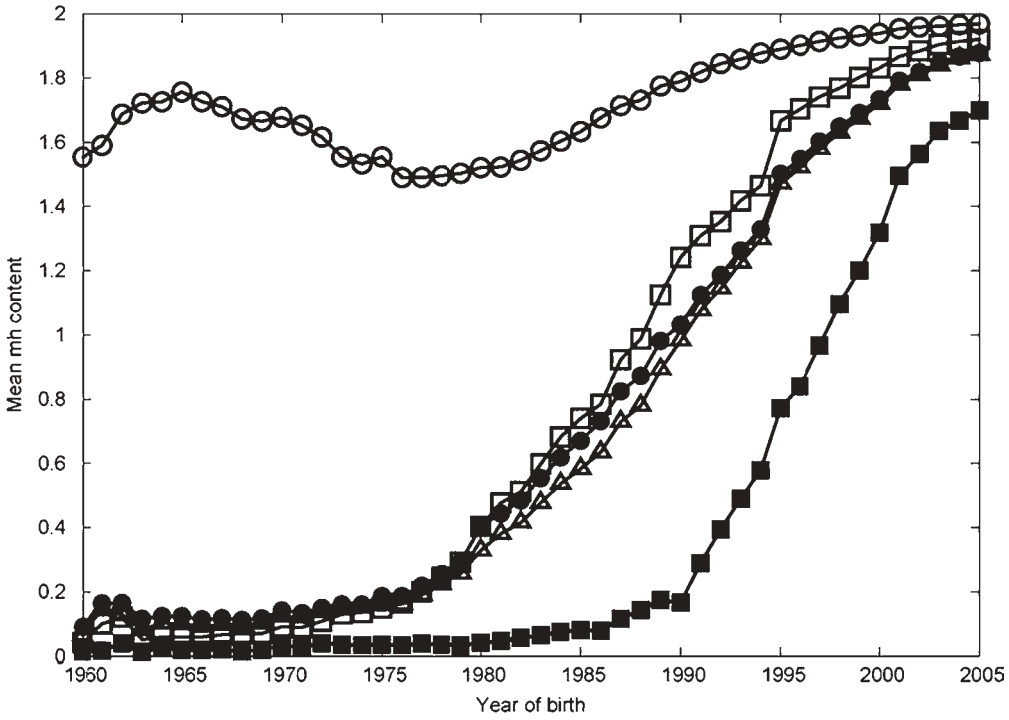

Figure 1 Mean content of mh gene by years of birth for DP-BBB cattle calculated by the use of the proposed method and the iterative peeling method: empty square – the proposed method under the standard model; solid square – the proposed method under the model with unknown parent groups (Westell *et al.* (1988) model); empty circle – the iterative peeling method with single genetic group; solid circle – the iterative peeling method with multiple genetic groups and the allelic frequency calculated under the standard model, empty triangle - the iterative peeling method with multiple genetic groups and the allelic frequency calculated under the Westell *et al.* (1988) model.

https://doi.org/10.1017/S1751731107392628 Published online by Cambridge University Press

use of the iterative peeling method under multiple genetic groups model were performed, one with the mh gene frequency based on the results from the proposed method under standard model, and the other under the Westell *et al.* (1988) model. The results of these two analyses were similar and little influenced by the assumed mh frequency. The results from the iterative peeling were similar to the results calculated by the proposed method under the standard model, even if the mh frequency was based on the Westell *et al.* (1988) model.

Discussion

There are two approaches to approximate gene content for the ungenotyped animals in huge pedigrees using the information from all their genotyped relatives: the Monte Carlo approach and the iterative peeling approach. The first is more flexible and therefore can be more useful in practice. The problem with these methods is that it is difficult to prove that they provide good solutions in large pedigrees unless the pedigree is peelable and exact values can be calculated. There are number of possible Monte Carlo samplers, however, all of them are Markov chain samplers unless again the pedigree is peelable. Advanced samplers use the peeling or the iterative peeling computation to sample from desired probability distribution (Kong, 1991; Fernandez *et al*., 2001). MCMC approaches are very useful in the analysis of genetic data on pedigrees, however, the construction of efficient Markov chain and monitoring its convergence requires extensive additional experience from the user of these methods. Moreover, both advanced MCMC samplers and the iterative peeling require relatively sophisticated computer programs and availability of such algorithms is limited. In this paper we propose a practical method to approximate the gene content in large animal pedigrees. All computations can be done using one of a few trusted statistical packages for genetic evaluation.

The time needed to complete computation depends on the quality of the software and the convergence criteria. In case of the MCMC method the time also depends on the mixing behaviour and the required precision of estimation. In practice, limited amount of time can be allocated to a method to do computation. Under time limit, some MCMC methods may provide inaccurate solutions (Totir *et al*., 2003). The proposed method uses a simplified genetic model and the simplification drastically reduces the amount of time required to complete calculation. In contrast to the alternative methods, the amount of time does not increase with the number of missing genotypes. The capability of the model to calculate unknown gene content is comparable with the models usually used by the MCMC and iterative peeling approaches. The proposed method can be considered practical for large animal pedigrees with a high proportion of missing genotypes.

Some genetic problems would require genotype probabilities. In contrast to other methods, the proposed method does not provide genotype probabilities. Approximate genotype probabilities can be derived from the calculated gene content. For this we treat half the gene content as probability of *B* allele and calculate the genotype probability from Hardy-Weinberg proportions. Note that the crucial assumption, which should hold for this calculation is that of equal viability of all gametes and zygotes, and this assumption has already been made when deriving the expected value for *q* from the mean of parental values. We believe that this assumption will hold in most cases unless a gene causes death at early age. An additional assumption required is that of random mating in base population. We compared genotype probabilities calculated in this way to genotype probabilities obtained directly from the iterative peeling method applied to a model with multiple genetic groups. For this comparison a population of DP-BBB cattle (235 133 animals) was considered. Note, within the population only 1284 animals had their genotype records, however, many DP-BBB animals have ancestors in dairy breeds, which were assumed free of the mh allele, and some were related to M-BBB animals which were considered all mh/mh. This fact was also taken into account during computations. The results of the comparison are presented in Table 4. The standard deviation of the absolute difference between alternative calculations was less than 0.12, however, for a few animals large dissimilarity between results have occurred. The highest difference between two alternative solutions was 0.95. The high dissimilarity occurred usually for an animal that had no genotype record but its only possible genotype is fixed based on typed relatives.

The proposed approach does not detect inconsistency between genotype and pedigree records. However efficient algorithms to detect non-mendelian records exist and they can be applied before calculation. One of possible options is to use a genotype elimination algorithm as proposed by Lange and Goradia (1987). The algorithm works iteratively on the nuclear families and has enough power to detect most of the errors that would result in null likelihood under a complete penetrance model. Application of the genotype elimination algorithm may add some data by concluding the only possible genotype for an animal.

The proposed method is similar to the best linear unbiased prediction applied to discrete data. Application of BLUP methodology to discrete variables has been considered in the context of breeding value estimation

Table 4 *Comparison between genotype probabilities at myostatin locus calculated for 235 133 DP-BBB animals by the use of proposed method and the iterative peeling: means and standard deviations (s.d.) of the absolute differences between alternative solutions*

Genotype	Mean	s.d.
+/+	0.046	0.061
+/mh	0.110	0.119
mh/mh	0.077	0.100

https://doi.org/10.1017/S1751731107392628 Published online by Cambridge University Press

(Gianola, 1980; Hoeschele, 1986). The main problem in the analysis of a discrete trait is that the restriction that the sum of response probabilities across all categories must total 1 is not taken into consideration. Numerous generalised linear mixed models have been proposed to study underlying continues variables from discrete data. These models may improve precision of the proposed method. Unlike the mixed model equations, however, the estimation equations for generalised linear mixed model must be solved iteratively. Such a method would not be more practical than the iterative peeling method. Moreover, the assumption of the linear relationship between gene contents would remain violated.

In the analysis of mh gene we considered multiple genetic groups to demonstrate that a genetic model can be easily adapted for a particular pedigree. The proposed method uses the general methodology of the linear mixed models and the trusted computer applications covering a wide range of various linear models are easily available. Adaptations of the model, as separate means of gene content (μ) or, as shown in this paper Westell groups can be easily specified for different breeds or birth years of animals (Westell *et al*., 1988). Although, a model used by the alternative approaches can also be adapted to fit particular data, the available computer programs for these algorithms are usually hard to modify.

The proposed method does not require that the frequency of alleles among founders are known. In the analysis of mh gene it was demonstrated that this assumption is important for the iterative peeling approach. Certainly, the effect of incorrect assumption on gene frequency increases with the proportion of the missing genotypes. Again, this indicates that the proposed method will be especially practical for large pedigree with sparse data.

The proposed method has a property that may be considered useful when the analysed data contain hidden errors. The example is a sire with hundred daughters, all *BB* but one *bb*, which suggests a typing or pedigree error but yields positive likelihood under complete penetrance model. The exact genotype probability for the sire is P(*Bb*) = 1 and consequently $q = 1$. The presented algorithm is robust against this potential error in the sense that it calculates a value depending on the number of *BB* progeny, here $q = 1.98$. This property is useful in large animal pedigrees because in most cases they are not subjected to close scrutiny. Accounting for errors in the iterative peeling approach would require relaxation of the assumed genetic model.

In future, genetic evaluation will more often be based on the identified polymorphisms affecting directly quantitative traits of interest. It can be anticipated that usually biallelic genes as single nucleotide polymorphisms (SNP) will be identified. The proposed method is well fitted to this kind of polymorphism. In case of a multiallelic polymorphism, however, the application of the proposed method is still possible by estimating the content of each allele separately. The method cannot be easily modified to use information from linked markers.

It was also shown that sparse data provides little information on true gene content of an ungenotyped animal, and consequently, such data would not contribute much to identify favorable selection candidates. Also estimation of candidate gene effects will be difficult. Accurate prediction of gene content would require typing more individuals. The way the animals are chosen for genotyping may influence the estimate of candidate gene effects (Israel and Weller, 1998). Phenotypes under the influence of the typed gene could contribute additional information on gene content.

The proposed method can also be useful in initialisation of MCMC genotype sampling for solving some more complex problem, e.g. a multilocus problem. Single site and blocked samplers require initial genotype configuration consistent with observed data. A consistent starting configuration may be obtained by 'gene-dropping', drawing the genes for the founders from some assumed base population and sampling subsequent generation according to Mendelian segregation law, rejecting configurations which are inconsistent with data. For large pedigrees, however, gene-dropping has low acceptance ratio. The presented method can improve the acceptance ratio of gene-dropping algorithm by including approximate gene contents in calculation for each sampled genotype.

In this paper we presented a practical method to calculate gene content for ungenotyped animals of large and complex pedigrees. The method is development out of the approach by Israel and Weller (1998) and can be used as alternative to more advanced approaches. The method can be easily modified to correctly use different base populations. Approximate genotype probabilities can also be calculated. The calculated values can be used to estimate the additive effect of a candidate gene or to support decision in marker-assisted selection.

Acknowledgements

Nicolas Gengler, who is Research Associate of the National Fund for Scientific Research (Brussels, Belgium), acknowledges this support. Additional support was provided through grant 2.4507.02F (2) of the National Fund for Scientific Research. Manuscript review by Dr George R. Wiggans (Animal Improvement Programs Laboratory – ARS – USDA, Beltsville; USA) and the support of the Walloon Breeding Association (AWE) and the Walloon Regional Ministry of Agriculture (project D31-1112) are gratefully acknowledged.

References

Banos G, Wiggans GR and Powell RL 2001. Impact of paternity errors in cow identification on genetic evaluations and international comparisons. Journal of Dairy Science 84, 2523-2529.

Charlier C, Coppieters W, Farnir F, Grobet L, Leroy PL, Michaux C, Mni M, Schwers A, Vanmanshoven P, Hanset R and Georges M 1995. The mh gene causing double-muscling in cattle maps to bovine chromosome 2. Mammalian Genome 6, 788-792.

https://doi.org/10.1017/S1751731107392628 Published online by Cambridge University Press

Elston RC and Stewart J 1971. A general model for the genetic analysis of pedigree data. Human Heredity 21, 523-542.

Fernandez SA, Fernando RL, Guldbrandtsen B, Totir LR and Carriquiry AL 2001. Sampling genotypes in large pedigrees with loops. Genetics Selection Evolution 33, 337-367.

Gianola D 1980. A method of sire evaluation for dichotomies. Journal of Animal Science 51, 1266-1271.

Hasstedt SJ 2002. Pedigree analysis package, revision 5.0 edition. In Department of Human Genetics. University of Utah, Salt Lake City, UT.

Hoeschele I 1986. Estimation of breeding values and variance components with quasi-continuos data. Ph.D. dissertation, Universitat Hohenheim, Germany.

Israel C and Weller JI 1998. Estimation of candidate gene effects in dairy cattle populations. Journal of Dairy Science 81, 1653-1662.

Janss LLG, Van Arendonk JAM and Van der Werf JHJ 1995. Computing approximate monogenic model likelihoods in large pedigrees with loops. Genetics Selection Evolution 27, 567-579.

Kong A 1991. Analysis of pedigree data using methods combining peeling abd Gibbs sampling. In Computer Science and Statistics: Proceedings of the 23rd symposium on the Interface (eds EM Keramidas and SM Kaufman), pp. 379-385. Interface Foundation of North America, Fairfax Station, VA.

Lange K and Goradia TM 1987. An algorithm for automatic genotype elimination. American Journal of Human Genetics 40, 250-256.

Lauritzen SL and Sheehan NA 2003. Graphical models for genetic analyses. Statistical Science 18, 489-514.

Misztal I 1999. Complex models, more data: simpler programming? Proceedings of the Interbull Workshop Computers and Cattle Breeds, Tuusala, Finland. Interbull Bulletin 20, 33-42.

O'Connell JR and Weeks DE 1995. The VITESSE algorithm for rapid exact multilocus linkage analysis via genotype set-recoding and fuzzy inheritance. Nature Genetics 11, 402-408.

Thallman RM, Bennett GL, Keele JW and Kappes SM 2001. Efficient computation of genotype probabilities for loci with many alleles: II. Iterative method for large, complex pedigrees. Journal of Animal Science 79, 34-44.

Totir LR, Fernando RL, Dekkers JCM, Fernandez SA and Guldbrandtsen B 2003. A comparison of alternative methods to compute conditional genotype probabilities for genetic evaluation with finite locus models. Genetics Selection Evolution 35, 585-604.

Thompson E 1981. Pedigree analysis of Hodgkin's disease in a Newfoundland genealogy. Annals of Human Genetics 45, 279-292.

Van Arendonk JAM, Smith C and Kennedy BW 1989. Method to estimate genotype probabilities at individual loci in farm livestock. Theoretical and Applied Genetics 78, 735-740.

Wang T, Fernando RL, Stricker C and Elston RC 1996. An approximation to the likelihood for a pedigree with loops. Theoretical and Applied Genetics 93, 1299-1309.

Westell RA, Quaas LR and Van Vleck LD 1988. Genetic groups in an animal model. Journal of Dairy Science 71, 1310-1318.

https://doi.org/10.1017/S1751731107392628 Published online by Cambridge University Press