J. Dairy Sci. 84:944–950

Bias and Power in the Estimation of a Maternal Family Variance Component in the Presence of Incomplete and Incorrect Pedigree Information

T. Roughsedge,*,‡ S. Brotherstone,† and P. M. Visscher*,†
*Institute of Ecology and Resource Management,
University of Edinburgh, Edinburgh EH9 3JG
†Institute of Cell, Animal and Population Biology,
University of Edinburgh, Edinburgh EH9 3JT
‡Current Address: Animal Biology Division, Scottish Agricultural College,
Edinburgh EH26 0PH

ABSTRACT

Several studies over the last 15 yr have estimated the magnitude of cytoplasmic inheritance of production and type traits in dairy cattle. Pedigree information can be used to assign maternal lineages, and the between-maternal lineage variance is then assumed to be an estimate of cytoplasmic inheritance. Two potential sources of bias and reduction of the power of estimation of cytoplasmic inheritance using such a method are 1) incomplete and 2) incorrect pedigree information being used in the assignment of maternal lineages. The theoretical bias introduced by these two sources of error is investigated and the results of a simulation study varying the number of families, the percentage of pedigree errors, and the level of incomplete lineage assignment are presented. Pedigree errors were found to have the biggest impact. A pedigree error rate of 8% per generation would result in a 75% reduction in the estimable magnitude of a 5% true component of variance after nine generations. The effect that these mechanisms have on the power of estimation are discussed and investigated by simulation. It was concluded that using historical pedigree, with incomplete and incorrect maternal family information, to assign maternal lineage would cause a downward bias in the magnitude of the cytoplasmic effect estimated. In the future, it will be possible to overcome pedigree problems by using molecular information to directly assign cytoplasmic lineage groups.
(**Key words:** maternal lineage, cytoplasmic effect, pedigree errors, variance component)

INTRODUCTION

There has always been a belief among dairy cattle breeders that certain cow families produce better cows than bulls in terms of genetic merit for production and conformation. Also, heritability estimates from daughter dam regressions are consistently higher than those obtained from paternal half-sib estimation (Seykora and McDaniel, 1983; Visscher and Thompson, 1992). This evidence may point towards a mechanism of inheritance, in addition to nuclear genetic inheritance, being transmitted through the female line, which is not being accounted for by current evaluations. One possible explanation is the almost exclusive maternal transmission of mtDNA in mammals (Gyllensten et al., 1991; Hutchinson et al., 1974).

If maternal lineages are responsible for a component of variance that is not being accounted for in current breeding value estimations in dairy cattle, then it is important that the effect is identified and quantified. Southwood et al. (1989), using an animal model to adjust for the additive nuclear genetic variance component, were able to estimate a simulated component of maternal lineage variance. The animal model approach has been adopted by various studies over recent years (Boettcher et al., 1996a; Roughsedge et al., 2000a; Schnitzenlehner and Essl, 1999). However, this approach assumes that if variance exists between the true maternal lineages, i.e., families formed from the points of mtDNA divergence, then it is possible to estimate this variance component by assigning maternal lineage as a random component of variance in an animal model. Two main mechanisms are involved by which a reduction is likely in the estimate of the magnitude of the maternal lineage variance component. The first of these mechanisms is the incomplete assignment to true maternal families, i.e., the assignment of several maternal subfamilies within one true larger family, which is not detected because insufficient generations in the establishment of complete maternal

Received June 14, 2000.
Accepted November 21, 2000.
Corresponding author: T. Roughsedge; e-mail: t.roughsedge@ed.sac.ac.uk.

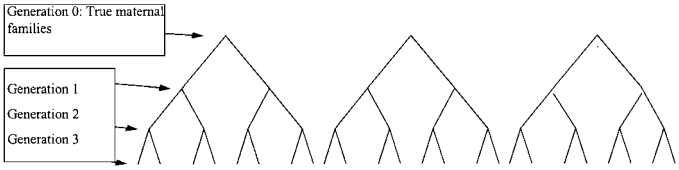

Figure 1. Illustration of section of family structure demonstrating subfamilies within true families.

families have been traced. The second of these mechanisms is the incorrect assignment of pedigree, leading to the accumulation of pedigree errors over generations from maternal lineage origin. With the quality of pedigree information often available from large field data sets, where it is often the case that only partial pedigree information is available, it is not possible to trace the true cytoplasmic origin or indeed to know how many generations the data are removed from that origin. The aim of this study was to establish the magnitude of these effects on the estimation of the maternal lineage variance component. The effect that the two mechanisms have on the power to detect maternal lineage variance is also investigated.

MATERIALS AND METHODS

Theory

Family structure. The effect of tracing maternal lineages in sufficient generations to establish the points of cytoplasmic origin can be illustrated by considering a simple balanced family design (Figure 1). In this design, we assume that all cows in the current generation are an equal number of generations from their cytoplasmic origin and that phenotypes have been adjusted for fixed and random effects other than a cytoplasmic effect.

Incomplete pedigree information. If full pedigree information is available, i.e., all records are assigned to the true maternal lineage, generation 0 in Figure 1, and the design is balanced, i.e., equal numbers per family, then the between family variance component ($\hat{\sigma}^2_f$) can be estimated with the ANOVA table shown in Table 1. If we then move to the situation in which incomplete family information is available due to tracing insufficient generations to the origin, i.e., to generation 1, 2, or 3 in Figure 1, we move to the situation shown in Table 2. Here r refers to the number of replicate families identified within the true family, i.e., if the pedigree is traced to generation 1 in Figure 1, then r = 2, because two families have been assigned within the one true family. In this table an extra level of sums of squares is now due to the between assigned subfamily replicate variance. However, when performing statistical analyses with incomplete pedigree information, a combination of the between family and between replicate sums of squares is used, since the true structure is not recognized. In the situation in which complete family assignment occurs (i.e., r = 1) Table 2 collapses to Table 1.

Table 1. Analysis of variance describing between-family variance.

Source	Degrees of freedom	Mean squares	E[Mean squares]
Between families	f-1	C	$(n)\sigma^2_f + \sigma^2_w$
Residual	f(n-1)	W	σ^2_2

f = Number of true cow families.

n = Assumed number of cows per family.

Table 2. Analysis of variance table describing between-family variance in the presence of incomplete family assignment.

Source	Degrees of freedom	Mean squares	E[Mean squares]
Between families	f-1	B	$(nr)\sigma^2_f + \sigma^2_w$
Between replicates	f(r-1)	R	σ^2_w
Residual	fr(n-1)	W	σ^2_w

f = Number of true cow families.

r = Number of replicates (i.e., number of subfamilies assigned to original family).

n = Assumed number of cows per family.

To obtain an estimate of the maternal family variance component we have an estimate of the sum of the sums of squares between replicates (f(r – 1)R) and the sums of squares between true maternal lineages ((f – 1)B). In practice we use the sum (SSC) of the between family (SSB) and between replicates (SSR) sums of squares (Equation 1).

$$SCC = SSB + SSR = (f - 1)B + f(r - 1)R \qquad [1]$$

The SSC has (f – 1) + f(r – 1) = fr – 1 degrees of freedom because the apparent number of families is fr. The mean squares of this sum of two sums of squares is obtained by equation 2.

$$C = \frac{SSC}{fr - 1} = \frac{[(f - 1)B + f(r - 1)R]}{fr - 1} \qquad [2]$$

The distribution of C is not proportional to a central χ^2 because B and R have different expectation. The expectation of the estimated maternal lineage variance ($\hat{\sigma}^2_f$) is then given by equation 3, and it can be seen that the between lineage variance component is reduced as the number of replicates within assigned families is increased.

$$E\,[\hat{\sigma}_f^2\,] = \frac{E[C - W]}{n}$$
$$= \sigma_f^2\left(\frac{(f-1)r}{fr-1}\right) + \sigma_w^2\left(\frac{(f-1)}{n(fr-1)} + \frac{f(r-1)}{n(fr-1)} - \frac{1}{n}\right) \quad [3]$$
$$= \sigma_f^2\left(\frac{(f-1)r}{fr-1}\right)$$

The downward bias introduced by incomplete maternal lineage assignment is the proportion of the maternal lineage variance component that is estimable after removing the bias demonstrated in equation 4.

$$Bias = \frac{\hat{\sigma}_f^2 - \sigma_f^2}{\sigma_f^2} = \frac{1-r}{fr-1} \quad [4]$$

Incorrect pedigree information. Pedigree errors cause a build-up of error in the estimation of maternal lineage variance, where the magnitude of underestimation is affected by the level of assignment error per generation and the number of generations from true cytoplasmic origin. In the estimation of maternal lineage variance, we assume that the tracing of maternal lineage provides correct families with identical mtDNA. If errors in the assignment of pedigree occur, information is irretrievably lost from the system and the magnitude of this loss can be approximated by equation 5, as suggested by Gibson et al. (1997).

$$E[\hat{\sigma}_f^2\,] = \sigma_f^2[(1-p)^g]^2 \quad [5]$$

where g = generations from origin to current generation; p = proportion of pedigree errors per generation. For example, an error rate of 5% per generation results in a reduction in the magnitude of the estimated component of $[(1 - 0.05)^8]^2 = 0.44$ after eight generations from the cytoplasmic origin.

Combining the effect of incomplete and incorrect pedigree information on the downward bias of maternal lineage variance estimation gives,

$$E[\hat{\sigma}_f^2] = \left[\frac{(f-1)r}{fr-1}\right]\cdot\left((1-p)^{2g}\right)\cdot\sigma_f^2$$

,

and the proportional bias in the estimation is,

$$Bias = \frac{(f-1)r}{fr-1}\cdot(1-p)^{2g} - 1 \quad [6]$$

Power of detecting a maternal lineage component. When the incorrect assignment of pedigree is considered, the power of detection of a variance component is the power of detection of the estimable component as obtained with equation 6. If complete pedigree information were available, the variance ratio test statistic obtained in the presence of incorrect pedigree follows approximately an F distribution with the degrees of freedom (fr – 1) for the numerator and fr(n – 1) for the denominator of the ratio. However, when the incomplete pedigree situation is considered, it is not easy to ascertain the distribution of the test statistic. The test statistic obtained is the ratio of the mean squares C from equation 2 over the residual mean squares. The mean squares C is derived from the sum of two sums of squares on different scales (equation 2) and is therefore not a standard χ^2.

Simulation Study

To demonstrate the previous two effects, simulations were run 10,000 times, each based on the structure shown in Table 2 and the family structure described in example 1, with the magnitude of the true maternal lineage variance and the error rate per generation being varied between simulations. Phenotypic records were simulated based on eight true maternal families with a random error component assigned to each of the 4096 individuals and the same cytoplasmic component to all individuals within the same maternal family. The pedigree error rate was then accounted for by calculating the proportion of the current generation that we would expect to be incorrectly assigned after the number of generations between the origin and the generation being simulated. The proportion of incorrect records expected were then randomly chosen and given a phenotype comprising a random error component and a random cytoplasmic component, assuming that the misidentified records are from random unknown maternal families. With this phenotypic information, a one-way ANOVA was used to estimate the between-family variance component, with the true cytoplasmic family assignment then with subdivision of the true simulated families to represent the incomplete tracing of pedigree information.

RESULTS

Bias with Incomplete Pedigree Information

The bias introduced by incomplete pedigree information is shown in Figure 2 for different family size and number of subfamilies assigned within true family. It can be seen that the bias is very small when the number of true families is large regardless of the size of r; however, when there are few true families the assignment of duplicate maternal families within the true family causes a downward bias in cytoplasmic variance component estimation. Following the structure

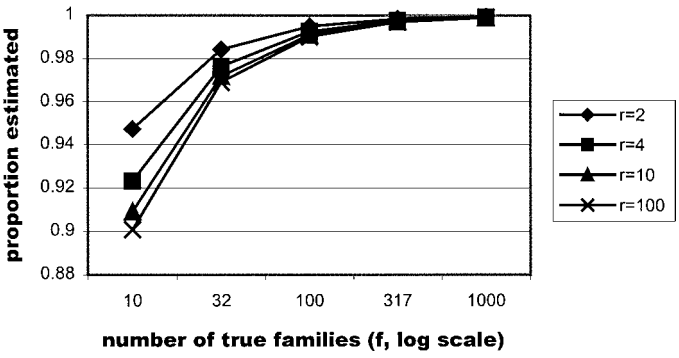

Figure 2. The proportion of the between lineage variance estimable with assignment to r subfamilies within true family (f).

Table 4. Simulated versus predicted magnitude of variance component where true variance component is 0.05 and pedigree error rate per generation is 0.08.

Number of assigned families	Number in assigned family	Predicted magnitude of variance component	Variance component estimated in simulation	Predicted as a proportion of simulated variance component
2048	2	0.0115	0.0118	0.97
1024	4	0.0115	0.0116	0.99
512	8	0.0115	0.0115	1.00
256	16	0.0116	0.0116	1.00
128	32	0.0116	0.0117	0.99
64	64	0.0117	0.0118	0.99
32	128	0.0119	0.0120	0.99
16	256	0.0123	0.0124	0.99
8	512	0.0132	0.0132	1.00

used in the simulation, if we consider that the true number of maternal families, f, is equal to 8, and that the total number of individuals, N, is equal to 4096 and the current generation is 9 generations from the true origin. Then, if at every generation the number of females is doubled, and the true between-maternal lineage variance component is 5% of the phenotypic variance, the magnitude of the variance component estimated from generation 9 data for different numbers of generations traced is shown in Table 3. The bias introduced ranges from 7 to 12% as we move from tracing to within one generation of the true family structure down to the situation in which only one generation is traced, resulting in a family size of two individuals with records per family. The increase in bias is nonlinear over the number of generations traced. In Table 3, the bias remains constant until more than four generations are traced. The bias is then seen to fall at an increasing rate until the true family structure is traced.

The downward bias introduced by not tracing families to their true cytoplasmic origin increases as the number of true families decreases, equation 4. For example, a field data set may have 30,000 records assigned to 10,000 families. If we assume that family size is equal in this case (though it may be far from equal) and we then assume that there are 1000, 100, or 10 true families then the downward bias is, respectively, 0.1, 1, and 10%. It can clearly be seen that the impact is noticeable only if there are a small number of true families. However, in a modern dairy breed, there could potentially be a low number of original mtDNA sources.

Bias with Incomplete and Incorrect Pedigree Information

Equation 6 was able to predict the outcome of the simulation consistently to within ~5% of the true component. Two extreme examples of the results of this simulation are shown in Tables 4 and 5. The pedigree error rate is 8% per generation, and the records are eight generations from the true cytoplasmic origin resulting in only $(1 - 0.08)^8 = 0.51$, i.e., 51% of records

Table 3. Expectation of estimation of variance component by tracing different numbers of generations when the true maternal lineage variance component is 0.05.

Number of generations traced	Number of cows in assigned family	Number of assigned families	Variance component estimated
1	2	2048	0.0438
2	4	1024	0.0438
3	8	512	0.0438
4	16	256	0.0438
5	32	128	0.0441
6	64	64	0.0444
7	128	32	0.0452
8	256	16	0.0467
9	512	8	0.0500

Table 5. Simulated versus predicted magnitude of variance component where true variance component is 0.05 and pedigree error rate per generation is 0.08.

Number of assigned families	Number in assigned family	Predicted magnitude of variance component	Variance component estimated in simulation	Predicted as a proportion of simulated variance component
2048	2	0.1153	0.1145	1.01
1024	4	0.1153	0.1147	1.01
512	8	0.1155	0.1148	1.01
256	16	0.1157	0.1150	1.01
128	32	0.1161	0.1155	1.01
64	64	0.1171	0.1164	1.01
32	128	0.1190	0.1182	1.01
16	256	0.1229	0.1222	1.01
8	512	0.1317	0.1309	1.01

Table 6. Power of detection between lineage variance with incomplete pedigree information and no incorrect pedigree assignment.

Number of assigned families	Number in assigned family	Power of detecting variance component simulated		Predicted power if number of assigned families is true number families	
		1%	5%	1%	5%
2048	2	0.107	0.580	0.117	0.732
1024	4	0.179	0.802	0.195	0.982
512	8	0.288	0.915	0.322	1.000
256	16	0.430	0.961	0.510	1.000
128	32	0.578	0.983	0.723	1.000
64	64	0.719	0.993	0.876	1.000
32	128	0.831	0.998	0.942	1.000
16	256	0.901	0.998	0.958	1.000
8	512	0.941	1.000	0.941	1.000

in the current generation being correctly assigned. It can be seen in Tables 4 and 5 that the simulation estimates of the variance component are very close to the prediction of the estimable variance component. In the most extreme case, in which only two of the 512 individuals per family are assigned and pedigree errors lead to 51% incorrect assignment of records to true families, the estimate and prediction are within 3% of each other. For all other situations, the estimation and prediction are within 1% of each other. Tables 4 and 5 represent true variance components of 0.05 and 0.5, respectively, illustrating that the prediction remains accurate in extreme cases. The actual impact of incorrect pedigree assignment can be seen to cause a dramatic reduction in the proportion of the true variance component that is estimable. In Tables 3 and 4 the same family structure was considered in the estimation of the same true variance component; however, in Table 4 incorrect pedigree assignment of 8% per generation was introduced. Even when the family structure was traced to the generation in which the true cytoplasmic origin was simulated, only 26% of the magnitude of the true variance component was estimated. This highlights the danger of incorrect pedigree assignment, indicating that it is better to exclude unknown pedigree information from genetic evaluations than to use incorrect pedigree information.

Power of Detection

With the simulation described, the power of detection of a 1 and 5% between-lineage variance component was obtained (Table 6). The power was obtained comparing the ratio C/W to an F ratio test statistic with the degrees of freedom as previously described in this section to either accept or reject the variance component estimated at a 5% type-1 error rate in each simulation. Also presented in Table 6 is the power of detection of a between lineage variance component given that the assigned families are the true families. For this it was assumed that assigned families were true families and an F ratio power test was used again at a 5% level of type-1 error rate (Lynch and Walsh, 1998). The actual power of detection was consistently lower than the situation in which the number of families assigned was the true number of families. The true magnitude of the variance component was 5% then, given the family structure shown in Table 6, the power of estimating this component with the full structure traced is 100%. For a 5% component, the power falls below 90 to 80% only when assigned family size is 4 in comparison to the true 512 records per true family. When a 1% component is considered, the power of estimation is seen to fall more rapidly and to be lower than 70% when the tracing is four generations from the true family structure. These are further illustrated in Figure 3, where the power is plotted for different magnitudes of true variance component for the situation where f × r true families exist and where each of

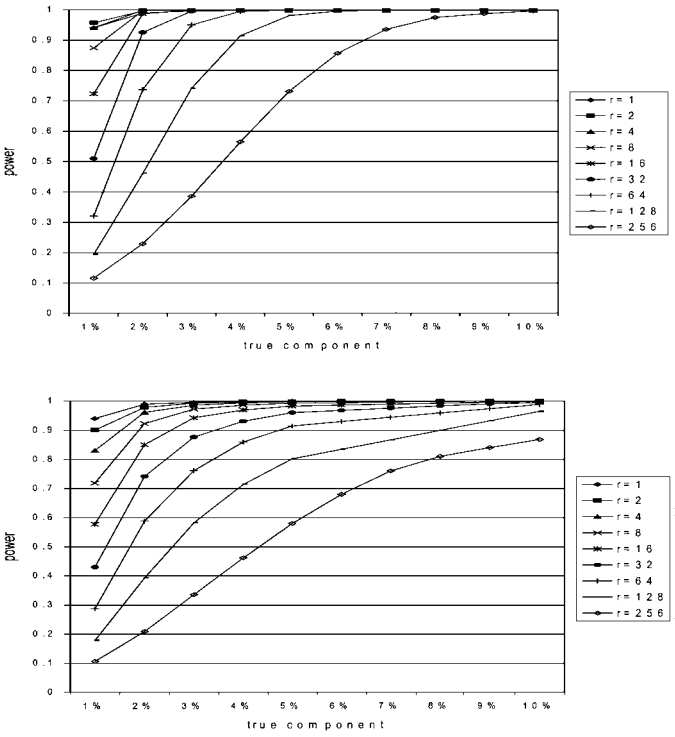

Figure 3. Power of detecting a true variance component if N = 4096 at a 5% type-1 error rate. Top graph illustrates the power where assigned number (8*r) of families is true number of families. Bottom graph shows the simulation results when eight true families (f) are incorrectly assigned to r subfamilies.

the f families are assigned to r subfamilies. No pedigree errors are included in the power results.

DISCUSSION

Two main mechanisms have been clearly demonstrated by which we can expect to underestimate maternal lineage variance. The first mechanism is the result of tracing insufficient generations of the maternal pedigree, which to some extent can be overcome by a more detailed tracing procedure. For large data sets the feasibility of tracing of further generations of maternal lineage is restricted by the size of the pedigree already available on a database. For some situations, tracing historic pedigrees further than the currently computerized pedigree may be possible. The increase in the accuracy of estimation was not seen to be linear, and it is dependent on the data structure of the true families. Again, our knowledge of true family structure is limited to the extent of the pedigree available.

The second mechanism is the accumulation of pedigree assignment errors. Ron et al. (1996) summarized estimates of paternity misidentification rates of between 1.3 and 23% across a number of European studies. Although these are paternity estimates, it would not be unreasonable to hypothesize that some degree of maternity misidentification occurs. This is a source of underestimation, which beyond ensuring the current recording is accurate, is historically out of our control. The only correction that can be made for the accumulation of pedigree errors is in the prediction of errors that have historically occurred and the use of DNA tests to make an appropriate adjustment to the component estimated. The problem that can be encountered with a large accumulation of error is that whatever upward adjustment we can make to the variance component that is estimated, the power of estimation is diminished in relation to the size of the estimable component.

Many of the studies that have estimated maternal lineage variance have involved experimental herds (e.g., Boettcher et al., 1996b, Roughsedge et al., 1999). It is unlikely that pedigree misidentification occurs in these herds on the scale that we might expect in field data. However, these studies suffered from other problems such as size of data set available and common maternal environment effects (Roughsedge et al., 2000a), and the problem of identifying maternal family origin still exists in these experimental herds.

With incomplete pedigree information the power of detection of a maternal lineage variance component is reduced in comparison to the power of detection where the assigned maternal lineage is the true maternal lineage.

Such a demonstration suggests that with available pedigree information and current methodology we may not be able to estimate maternal lineage variance but, cannot yet dismiss the possibility that such an effect exists. It is also possible that in studies in which significant maternal lineage variance components have been estimated, e.g., persistency in the study of Roughsedge et al. (2000b), the true magnitude of the effect has been underestimated. Roughsedge et al. (2000b) estimated a maternal lineage variance component of 4% of the overall phenotypic variance for persistency, with a heritability of 0.1. Given that this may be an underestimation, maternal lineage variance must be considered if persistency is used in a selection program.

Looking to the future, with molecular technology becoming more available and economically feasible, it may be possible to overcome these problems. If databases based of completed mtDNA sequence or, more realistically, marker information are developed then problems of maternal family assignment can be overcome and the hypothesis that within family mtDNA is identical can be tested. The opportunity will exist to look for direct associations between mtDNA polymorphisms and the expression of traits.

ACKNOWLEDGMENTS

T. Roughsedge thanks the UK Ministry of Agriculture Fisheries and Food for studentship funding.

REFERENCES

Boettcher, P. J., A. E. Freeman, S. D., Johnston, R. K. Smith, D. C. Beitz, and B. T. McDaniel. 1996a. Relationships between polymorphism for mitochondrial deoxyribonucleic acid and yield traits of Holstein cows. J. Dairy Sci. 79:647–654.

Boettcher, P. J., D.W.B. Steverink, D. C. Beitz, A. E. Freeman, and B. T. McDaniel. 1996b. Multiple herd evaluation of the effects of maternal lineage on yield traits of Holstein cattle. J. Dairy Sci. 79:655–662.

Gibson, J. P., A. E. Freeman, and P. J. Boettcher. 1997. Cytoplasmic and mitochondrial inheritance of economic traits in dairy cattle. Livest. Prod. Sci. 47:115–124.

Gyllensten, U., D. Wharton, A. Josefsson, and A. C. Wilson. 1991. Paternal inheritance of mitochondrial DNA in mice. Nature 352:255.

Hutchinson, C. A., J. E. Newbold, S. S. Potter, and M. H. Edgell. 1974. Maternal inheritance of mammalian mitochondrial DNA. Nature 251:536–538.

Lynch, M., and B. Walsh. 1998. Pages 885 *in* Genetics and Analysis of Quantitative Traits. Sinauer Associates, Inc., Sunderland, MA.

Ron, M., M. Blanc, E. Band, E. Ezra, and J. I. Weller. 1996. Misidentification rate in the Israelei dairy cattle poulation and its implications for genetic improvement. J. Dairy Sci. 79:676–681.

Roughsedge, T., S. Brotherstone, and P. M. Visscher. 1999. Estimation of variance of maternal lineage effects at the Langhill dairy herd. Anim. Sci. 68:79–86.

Roughsedge, T., S. Brotherstone, and P. M. Visscher. 2000a. Effects of cow families on type traits in dairy cattle. Anim. Sci. 70:373–381.

Roughsedge, T., P. M. Visscher, and S. Brotherstone. 2000b. Effects of cow families on production traits in dairy cattle. Anim. Sci. 71:49–57.

Schnitzenlehner, S., and A. Essl. 1999. Field data analysis of cytoplasmic inheritance of dairy and fitness-related traits in dairy cattle. Anim. Sci. 68:459–466.

Seykora, A. J., and B. T. McDaniel. 1983. Heritabilities and correlations of lactation yields and fertility for Holsteins. J. Dairy Sci. 66:1486–1493.

Southwood, O. I., B. W. Kennedy, K. Meyer, and J. P. Gibson. 1989. Estimation of additive maternal and cytoplasmic genetic variances in animal models. J. Dairy Sci. 72:3006–3012.

Visscher, P. M., and R. Thompson. 1992. Comparison between genetic variances estimated from different types of relatives in dairy cattle. Anim. Prod. 55:315–320.