Heredity **68** (1991) 299–311
©The Genetical Society of Great Britain

Received 5 April 1991

Methods of segregation analysis for animal breeding data: a comparison of power

S. A. KNOTT, C. S. HALEY* & R. THOMPSON*

*Institute of Cell, Animal and Population Biology, University of Edinburgh, Zoology Building, West Mains Road, Edinburgh EH9 3JT and *AFRC Institute of Animal Physiology and Genetics Research, Edinburgh Research Station, Roslin, Midlothian EH25 9PS, UK*

Maximum likelihood segregation analysis provides potentially the most powerful method for the detection of segregating major genes. Segregation analysis requires the comparison of the likelihood of the data under the combined model (allowing both polygenic and major gene genetic variation) with the likelihood of the data under the polygenic model (allowing only polygenic genetic variation). In this study three approximations to the combined model likelihood were compared using simulated data, both with and without a segregating major gene, containing observations on paternal half-sibs. The use of Hermite integration to replace the integration in the combined model likelihood provided the most powerful test for a major gene. Two approximations, based on extensions of linear-mixed-model theory and estimating transmitting abilities for sires, were also considered. These approximations were less powerful than the use of Hermite integration, although the approximation estimating a transmitting ability for each major genotype for the sires was an improvement over the approximation estimating a single transmitting ability. For each approximation the frequency of detection of a major gene depended on the proportion of the genetic variance explained by the simulated major gene and whether the major gene caused the distribution to be skewed.

Keywords: animal breeding, major genes, maximum likelihood, segregation analysis.

Introduction

Classical animal breeding theory is based on the assumption that traits are controlled by many genes each having a small effect. The action of individual genes cannot be observed directly and traits are generally described in terms of summary statistics such as the heritability. However, genes with a large effect on commercial traits have been identified in favourable circumstances. Notable examples are the dwarfing gene in poultry (Merat & Ricard, 1974), the Booroola gene affecting ovulation rate in sheep (Piper & Bindon, 1982; Piper *et al.*, 1985), the double muscling gene in cattle (Rollins *et al.*, 1972; Hanset & Michaux, 1985a,b), and the gene determining halothane sensitivity in pigs (Smith & Bampton, 1977). Where genes can be identified and individual animals genotyped, exploitation of the genetic variance can be optimized. Major genes can also provide raw material for genetic engineering programmes.

Despite large phenotypic effects, major genes are often not immediately apparent due to the obscuring effects of polygenic and environmental variation. It is likely, therefore, that genes of major phenotypic and potential economic importance have yet to be detected.

Segregation analysis has been proposed as a suitable method to detect a segregating major gene (Elston & Stewart, 1971). It involves maximizing and comparing the likelihood of the data under different genetic models to ascertain the most likely genetic structure. To identify a major gene the maximum likelihood (ML) of the data under a polygenic model is compared with that under the combined model (containing a major gene and polygenic component). A significant improvement in the likelihood obtained by incorporating a major gene in the model provides evidence for the segregation of a major gene in the population under study.

The size of pedigrees that can be considered under the exact combined model is effectively limited to between 10 and 20 by the fact that the number of

Correspondence: S. A. Knott.

calculations required to calculate the likelihood increases exponentially with the pedigree size (see below).

This paper explores the behaviour of approximations to the combined-model likelihood suitable for animal breeding data.

Likelihoods

Equations for the exact combined model and polygenic likelihoods are obtainable following Morton & MacLean (1974). A simple sire model with one offspring per dam and balanced structure will be used. All parents are assumed to be unrelated and randomly mated. A single trait is considered with one observation for each offspring. Fixed effects, such as herd or year, will be ignored in the development of the likelihoods, but an extension to include these or more complex pedigree structures is possible.

The model to describe the data for offspring *j* of sire *i*, when offspring *j* has major genotype *d*, can be represented as

$$y_{ij} = \mu + \mu_d + u_i + e_{ij}$$

where y_{ij} is the performance of the *j*th offspring of the *i*th sire, μ is the overall population mean of the polygenic and environmental components, *d* is the offspring major genotype, set to zero for the polygenic model, μ_d is the effect of major genotype *d* (for polygenic model μ_0 equals zero). u_i is the random effect for sire *i* (i.e. polygenic component) independent of μ_d; $u \sim N(0, \sigma_u^2)$, and e_{ij} is the residual random effect for each individual, independent of u_i and μ_d; $e \sim N(0, \sigma_w^2)$.

Following Morton & MacLean (1974), the polygenic likelihood can be written as follows

$$L(\text{poly}) = \prod_{i=1}^{s} \int_{-\infty}^{+\infty} h(u_i) \prod_{j=1}^{n} k_0(y_{ij} \mid \mu, u_i, \sigma_w^2) \cdot du_i \qquad (1)$$

and the combined model likelihood

$$L(\text{MM}) = \prod_{i=1}^{s} \int_{-\infty}^{+\infty} \sum_{c=1}^{m} p(c) h(u_i) \prod_{j=1}^{n} \sum_{d=1}^{m}$$

$$\times \operatorname{trans}(d \mid c) k_d(y_{ij} \mid \mu, \mu_d, u_i, \sigma_w^2) \cdot du_i \qquad (2)$$

where *s* is the number of sires, *n* is the number of half-sib offspring per sire, *m* is the number of major genotypes, *p*(*c*) is the frequency of major genotype *c* in the population of sires, and trans(d|c) is the probability of the offspring having major genotype *d* when the sire has genotype *c*, which is based on Mendelian transmission probabilities and the allele frequency in the population of dams, and $h(u_i)$ is the likelihood that sire *i* has polygenic transmitting ability u_i.

$$h(u_i) = \frac{1}{\sqrt{2\pi\sigma_u^2}} \exp\left[-\frac{1}{2\sigma_u^2}(u_i)^2\right]$$

$k_d(y_{ij} \mid \mu, \mu_d, u_i, \sigma_w^2)$ is the conditional likelihood that offspring *j* of sire *i* has phenotype $y_{ij}$when offspring *j* has major genotype *d* and the polygenic component contributed from sire *i* is u_i.

$$k_d(y_{ij} \mid \mu, \mu_d, u_i, \sigma_w^2) = \frac{1}{\sqrt{2\pi\sigma_w^2}}$$

$$\times \exp\left[-\frac{1}{2\sigma_w^2}(y_{ij} - \mu - \mu_d - u_i)^2\right].$$

The combined-model likelihood involves the integration of a complicated function, which gives a summation over each combination of major genotypes for the pedigree. With *n* offspring $2^{n+1} + 3^n$ summations for each sire are required. This calculation of the likelihood becomes impracticable even with only a small number of offspring per sire, for example, with five offspring ($n = 5$), 307 combinations have to be considered and with 10 offspring, 61, 097 combinations. Hence, several approximations to this likelihood will be considered.

Hermite integration

A standard numerical approximation to an integration is to replace it with a weighted summation, so that effectively a continuous density function [$c(x)$] is replaced by a discrete histogram (for example, see Hildebrand, 1974).

$$\int_a^b c(x) f(x) \cdot dx = \sum_{g=1}^{G} w_g f(x_g) \qquad (3)$$

where *G* is the number of points in the summation, x_g are the abscissae within the range *a* to *b*, and w_g are the weights.

Suitable weights and abscissae need to be supplied. As the number of points in the summation increases the approximation improves, integration being equivalent to an infinite number of points. By taking into account the distribution of the function to be integrated, the abscissae and weights can be optimized to reduce the number of points required in the summation to provide a reasonable approximation. In the case of the combined model likelihood, the variable over which integration takes place appears in the form $\exp[-x^2]$ [see equation (2)] and hence efficient abscissae and weights can be obtained from the Hermite polynomial (Hildebrand, 1974). Tables of the weights and abscissae exist for various numbers of points in the summation and are given for a standard curve, symmetrically placed about the origin (e.g. Abramowitz & Stegun, 1972).

The exact combined model likelihood contains an integration over the transmitting ability (u_i) for each sire. To allow the flexibility to take the summation around a value other than zero and to alter the variance of the transmitting ability distribution to have the same range as the abscissae of the standard curve, the transmitting abilities can be transformed thus

$$x_i = \frac{u_i - l_i}{V_i}$$

where l_i are the location parameters, and V_i are the scaling parameters.

The likelihood can be rewritten as follows

$$L(\mathrm{MM}) = \prod_{i=1}^{s} \int_{-\infty}^{+\infty} \frac{\sqrt{2\pi V_i^2}}{\sqrt{2\pi\sigma_u^2}} \exp\left[-\frac{u_i^2}{2\sigma_u^2} + \frac{x_i^2}{2}\right] \frac{1}{\sqrt{2\pi V_i^2}} \times \exp\left[-\frac{x_i^2}{2}\right] \sum_{c=1}^{m} p(c) \prod_{j=1}^{n} \sum_{d=1}^{m} \mathrm{trans}(d|c) \times \frac{1}{\sqrt{2\pi\sigma_w^2}} \exp\left[\frac{-(y_{ij} - \mu - \mu_{\mathrm{d}} - u_i)^2}{2\sigma_w^2}\right] \cdot du_i .$$

Rewriting the transform $u_i = V_i x_i + l_i$ and hence $du_i = V_i \cdot dx_i$ then changing the integration over u_i to an integration over x_i and finally replacing the integration over x_i with a summation gives the following combined model likelihood

$$L(\mathrm{MM}) = \prod_{i=1}^{s} \frac{\sqrt{2\pi V_i^2}}{\sqrt{2\pi\sigma_u^2}} \sum_{g=1}^{G} \left[\left[\sum_{c=1}^{m} p(c) \exp\left[-\frac{(V_i x_g + l_i)^2}{2\sigma_u^2} + \frac{x_g^2}{2}\right] \prod_{j=1}^{n} \sum_{d=1}^{m} \mathrm{trans}(d|c) \frac{1}{\sqrt{2\pi\sigma_w^2}} \times \exp\left\{\frac{-[y_{ij} - \mu - \mu_d - (V_i x_g + l_i)]^2}{2\sigma_w^2}\right\}\right]\right] w_g. \quad (4)$$

The Hermite polynomial is appropriate when $c(x)$ (3) is of the form $\exp[-x^2]$, whereas here the variable to be integrated is $1/\sqrt{2\pi}\exp[-(x_i^2/2)]$; hence the abscissae obtained from standard tables (e.g. Abramowitz & Stegun, 1972) should be multiplied by $\sqrt{2}$ and the weights divided by $\sqrt{\pi}$. This approximation will be denoted Herm.

On the basis of a study examining the number of points required in the summation to obtain a reasonable approximation (Knott, 1990), it was found that 20 points gave a value for the likelihood that was exact to more than five decimal places. Further refinements to the scaling of the abscissae and weights could reduce the number of points required for the same precision, for example, the use of the sire variance component to scale the abscissae reduced the required number of points to 10.

Estimating the mode of each sire's transmitting ability distribution

Following Le Roy *et al.* (1989) and incorporating the mode of the transmitting distribution for each sire for each combination of major genotypes for the half-sib family, the exact combined model likelihood can be rewritten as follows

$$L(\mathrm{MM}) = \prod_{i=1}^{s} \frac{1}{(2\pi)^{n/2}|\mathbf{V}_i|^{1/2}} \sum_{c=1}^{m} \sum_{D=1}^{m^n} p(c)\mathrm{trans}(D|c) \times \exp\left[-\frac{1}{2\sigma_u^2}\hat{u}_{icD}'\hat{u}_{icD}\right] \exp\left[-\frac{1}{2\sigma_w^2} \times (\mathbf{y}_i - \mathbf{1}\mu - \mathbf{W}_D\mu_d - \mathbf{Z}_i\hat{u}_{icD})'(\mathbf{y}_i - \mathbf{1}\mu - \mathbf{W}_D\mu_d - \mathbf{Z}_i\hat{u}_{icD})\right] \quad (5)$$

where $\mathbf{V}_i$ is the phenotypic variance–covariance matrix for the offspring of sire i, trans$(D|c)$ is the probability that the offspring have major genotype combination D given that the sire has genotype c, $\hat{u}_{icD}$ is the mode of the distribution of transmitting ability for sire i, when he has major genotype c and the combination of major genotypes for his offspring is D, $\mathbf{y}_i$ is a vector of phenotypes for the offspring of sire i, $\mathbf{W}_D$ is an nxm design matrix for combination D containing a one for the major genotype being considered for each offspring and a zero otherwise, μ_d is a vector of major genotype means, and $\mathbf{Z}_i$ is the column relating to the ith sire of the design matrix for random effects.

The mode of the distribution of transmitting ability for each sire has to be calculated for each possible combination of major genotypes for the sire and offspring. This calculation soon becomes infeasible as the number of offspring per sire increases, as described for the exact likelihood. Therefore an approximation to this likelihood is suggested, where a single estimate of the mode of each sire's transmitting ability distribution is calculated taking into account the possible major genotypes for the sire and offspring. The following expression is obtained for the combined model likelihood (Le Roy *et al.*, 1989), which will be denoted ME1

$$L(\mathrm{MM}) = \prod_{i=1}^{s} \left(\frac{2\pi\sigma_w^2}{n+\lambda}\right)^{1/2} h(\hat{u}_i) \sum_{c=1}^{m} p(c) \prod_{j=1}^{n} \sum_{d=1}^{m} \times \mathrm{trans}(d|c) k_d(y_{ij}|\mu, \mu_d, \hat{u}_i, \sigma_w^2) \quad (6)$$

where $\hat{u}_i$ is the mode of the transmitting ability distribution for sire i. $h(\hat{u}_i)$ and $k_d(y_{ij}|\mu,\mu_d,\hat{u}_i,\sigma_w^2)$ are as defined previously, now using the mode of the transmitting ability.

This is equivalent to Hermite integration with a single point in the summation for each sire, the point taken at an estimate for his polygenic breeding value, and would be equivalent to the exact combined-model likelihood if the major genotype of all individuals was known. A similar likelihood was considered by Hoeschele (1988).

Estimating the mode of each sire's transmitting distribution for each major genotype

Given the phenotypes of a group of half-sib offspring, the estimate of the polygenic transmitting ability of the sire should be dependent on his major genotype. Hence an approximation is proposed where the transmitting ability for each sire is estimated for each of his possible major genotypes.

Estimating the mode of each sire's transmitting ability distribution under the hypotheses of each major genotype for the sire, so that three estimates are obtained, gives the following approximation to the combined model likelihood (ME3) for the half-sib family structure defined previously

$$L(\mathrm{MM}) = \prod_{i=1}^{s} \left(\frac{2\pi\sigma_w^2}{n+\lambda}\right)^{1/2} \sum_{c=1}^{m} p(c)h(\hat{u}_{ic}) \prod_{j=1}^{n} \sum_{d=1}^{m} \times \mathrm{trans}(d|c)k_d(y_{ij}|\mu,\mu_d,\hat{u}_{ic},\sigma_w^2) \qquad (7)$$

where $\hat{u}_{ic}$ is the mode of sire i's transmitting ability distribution given that he has genotype c, $h(\hat{u}_{ic})$ and $k_d(y_{ij}|\mu,\mu_d,\hat{u}_{ic},\sigma_w^2)$ are as defined previously, but now the transmitting ability of the relevant genotype (c) is used.

This approximation would be equivalent to the exact combined model likelihood if the genotypes of all the offspring were known.

Analyses

A simulation study was carried out to investigate the performance of the three approximations. Phenotypes of 20 half-sib progeny from each of 50 sires were simulated, with all parents unrelated and randomly mated. Two polygenic models were simulated with phenotypes composed only of a polygenic component and an individual environmental component. In one model, the expected heritability was 0.2 and in the other 0.4. Four combined models, with a major gene, polygenes and an individual environmental component, were considered. In all simulations the genetic component comprised the effects of 25 unlinked loci, each with two alleles in Hardy–Weinberg equilibrium. For the polygenic components the alleles at each locus were at equal frequency with the same additive effect. The expected frequencies of the two alleles (denoted A and a) at the major locus and the relative effects of the three major genotypes are given in Table 1 for the four combined models. In each combined model there are two within major genotype standard deviations $(\sqrt{\sigma_u^2+\sigma_w^2})$ between the effects of the homozygotes. One hundred replicates were simulated for each model.

In the analysis the mean effect of the low-scoring homozygous genotype (μ) at the major locus, and the deviation from this mean of the other two major genotype means (μ_{AA} and μ_{Aa}), were estimated. The population was assumed to be in Hardy–Weinberg equilibrium and a single allele frequency [p(A)] was estimated. Two analyses of each dataset were carried out, the first assuming that the polygenic heritability was known, and estimating just the residual variance, and the second estimating the polygenic heritability as well as the residual variance.

Hermite integration was used with 20 points in the summation with the abscissae located around zero ($l_i=0$) and using the square root of the sire variance estimate (σ_u) from the previous iteration as the scaling parameter (V_i). ·

The Herm likelihood was maximized using a quasi-Newton algorithm (Numerical Algorithms Group, 1988) and the ME1 and ME3 likelihoods were maximized using an EM algorithm. This algorithm is described by Le Roy *et al.* (1989) and Hoeschele (1988) for the ME1 likelihood and can easily be extended for the ME3 likelihood. Parameter estimates are required from which the maximization process can start. If these are close to the global maximum, convergence to this maximum is more likely to be obtained. In these simulations the expected (i.e. simulated) parameters are known and hence these were used as initial estimates. In practice this would not be the case and several starting points would be used in order to confirm that the global maximum has been attained. The EM algorithm is sensitive to the initial parameter estimates and therefore the ME1 and ME3 likelihoods were maximized from an additional set of starting values. These alternative starting values explained the expected total mean and variance of the data but contained a major gene of different effect. They are given in Table 1. For each sire the initial estimate for his transmitting ability was zero. The polygenic heritability was fixed for the first few iterations otherwise the sire estimates remain at zero and convergence cannot be attained. For the polygenic data, the starting values for the combined model analyses

Table 1 Expected parameter values for the mixed models simulated and the alternative set of starting values used for the maximization process for each model

Model	$p(\mathrm{A})$	μ_{AA}	μ_{Aa}	μ_{aa}	σ_u^2	σ_w^2	h_{poly}^2	h_{mg}^2
Additive 0.2	0.5	20	10	0	5	95	0.2	0.33
Additive 0.4	0.5	20	10	0	10	90	0.4	0.33
Dominant	0.5	20	20	0	5	95	0.2	0.43
Rare	0.2	20	10	0	5	95	0.2	0.24
Alternative starting values for maximization								
Additive 0.2	0.5	25	12.5	0	4	68	0.2	0.52
Additive 0.4	0.5	25	12.5	0	7	65	0.4	0.52
Dominant	0.5	20	10	0	6	119	0.2	0.29
Rare	0.5	20	10	0	4	78	0.2	0.38

$p(\mathrm{A})$ = frequency of the high scoring allele in the parent population.

μ_d = the effect of the major genotype d, relative to the effect of the low scoring homozygote (μ_{aa}).

σ_u^2 = the additive polygenic sire variance component.

σ_w^2 = the residual variance.

$$h_{\mathrm{poly}}^2 = \frac{4\sigma_u^2}{\sigma_u^2 + \sigma_w^2}$$

$$h_{\mathrm{mg}}^2 = \frac{\sigma_{\mathrm{mg}}^2}{\sigma_u^2 + \sigma_w^2 + \sigma_{\mathrm{mg}}^2}$$

σ_{mg}^2 = the variance contributed by the major gene.

assumed that the major gene was additive with equal allele frequencies and explained 50 per cent of the total phenotypic variation and additionally for the ME1 and ME3 methods, assuming the major gene explained 13 per cent of the variation.

In exact ML analyses, a test statistic is provided by twice the difference between the natural logarithms of the MLs under the combined model and the polygenic model (2[In L(MM)—In L(poly)]). Under the null hypothesis of no major gene component, this test statistic is expected asymptotically to follow a chi-squared distribution with 3 d.f. (Wilks, 1938), as three parameters [$p(\mathrm{A})$, μ_{AA}, μ_{Aa}] are estimated in the combined model but fixed in the polygenic model.

The distribution of the test statistic for the three approximations for data simulated with and without a major gene can be used to explore their usefulness and power in the analysis of animal breeding data.

To enable a prediction of the number of individuals required to obtain a certain power, the analyses were repeated using different numbers of sires and half-sib offspring per sire. The combined models simulated with an additive major gene with equal allele frequencies and a polygenic heritability of 0.2 (additive 0.2 model) and with a dominant major gene (dominant model) were used. To consider the effect of the size of the major gene, two additional combined models were simulated. The data were simulated for 50 sires each with 20 half-sib progeny under a model with a polygenic heritability of 0.2 and with a major gene with additive effect and alleles at equal frequency. In one case there was one within-major gene phenotypic standard deviation ($\sqrt{\sigma_u^2 + \sigma_w^2}$) between the homozygotes and in the other three standard deviations. The results from 10 repeat simulations of each situation were considered.

Results

Analyses of polygenic data

The results from analyses of the simulated polygenic data are summarized in Table 2 as the mean and standard deviation of the test statistic distributions and the number of analyses giving significant results at the 5 and 1 per cent significance levels of a chi-squared distribution with 3 d.f. To compare the observed test statistic distribution with the expected distribution, the observed number of statistics falling within 10 equal probability classes of a chi-squared distribution were compared with the expected number (i.e. 10) using a chi-square test. The results are given in Table 3.

Using Herm to analyse data simulated under the null hypothesis produced mean values of the test statistic

Table 2 Mean and standard deviation of the test statistic (setting negative test statistics to zero) from the analysis of polygenic data, the number of analyses where the test statistic was zero and the number significant at the 5 and 1 per cent significance levels of a chi-square distribution with 3 d.f.

Model (heritability)	Analysis	Herm					ME1					ME3				
		Mean	s.d.	Zero	5%	1%	Mean	s.d.	Zero	5%	1%	Mean	s.d.	Zero	5%	1%
0.2	Fixed	2.89	2.33	0	5	0	0.16	0.76	95	0	0	0.41	1.14	78	0	0
0.2	Estimated	3.35	2.51	1	6	0	2.07	2.72	43	4	1	2.66	2.76	20	5	1
0.4	Fixed	3.06	2.99	1	5	3	0.01	0.00	99	0	0	0.30	1.54	89	1	1
0.4	Estimated	3.48	3.24	0	8	4	0.21	1.14	94	1	0	1.11	2.84	79	6	3
0.4	Fixed at 0.2	8.50	5.98	0	47	28	0.40	1.84	88	1	0	4.48	4.57	15	19	7

Based on 100 replicates of each simulation.

Table 3 Comparison of the observed test statistic distribution from analysis of polygenic data with the expected chi-square distribution with 3 d.f. The number of observed test statistics falling in 10 equal regions of the expected distribution were compared with the expected number (10) using a chi-square test

Model (heritability)	Analysis	χ^2 value		
		Herm	ME1	ME3
0.2	Fixed	4.4	803.6	642.8
0.2	Estimated	10.4	155.2	38.0
0.4	Fixed	6.6	880.2	747.6
0.4	Estimated	7.8	806.8	532.6
0.4	Fixed at 0.2	295.2	747.6	612.0

close to three with no significant difference between the observed and expected distributions. This suggests that the test statistic distribution does follow a chi-squared distribution with 3 d.f. This is the case both when the polygenic heritability was assumed to be known and hence fixed at its expected value and when the heritability was estimated.

The results presented from the ME1 and ME3 methods are based on the combined model analysis that gave the highest likelihood. For some of the analyses the test statistic obtained was negative for the ML obtained from both initial estimates, i.e. the combined model was less likely than the polygenic model, and these test statistics have been set to zero. All of the observed distributions using the ME1 and ME3 methods were significantly different from the expected distribution when the expected and observed number of analyses were compared for 10 equal regions of the chi-squared distribution. This was mainly because of the high number of analyses resulting in a zero test statistic. Removal of these analyses gave mean test statistics closer to the expected value of three and the test statistic distribution was not significantly different from the expected distribution (results not shown). Using the ME3 method the results were closer to the expected distribution than using the ME1 method.

When the polygenic heritability was fixed at a value less than that simulated, the mean and variance of the test statistic distribution increased along with the number of significant analyses, especially for the Herm and ME3 methods. The observed distribution of test statistics with the Herm method became significantly different from the expected distribution.

Analyses of combined-model data

The results from the analyses of combined-model data are given in Table 4. The mean test statistic over the 100 analyses is always highest using the Herm and lowest using the ME1 method. Using a comparison of the test statistic with the 5 and 1 per cent quantiles of the chi-squared distribution with 3 d.f. as a criterion for significance, major genes were detected most frequently using the Herm method. The test statistics obtained from the ME1 and ME3 methods, plotted against the test statistic from the Herm method for the same set of data, are shown in Fig. 1 for data simulated under the additive model with equal allele frequencies and an expected polygenic heritability of 0.2 (additive 0.2 model), fixed in the analyses, and in Fig. 2 for the simulated dominant major gene analysed (dominant model) with fixed polygenic heritability. There is a strong, positive linear relationship between the test statistics for each method if the zero test statistics are excluded, however, for each analysis, the Herm test statistic is highest, the ME1 test statistic is lowest and the ME3 intermediate.

For all three approximations, when the simulated major gene had an allele with dominant effect a major gene was detected in virtually all analyses. Considering the simulated additive major genes, for all the approxi-

Table 4 Mean and standard deviation of the test statistic (setting negative test statistics to zero) from the analysis of mixed-model data, the number of analyses where the test statistic was zero and the number significant at the 5 and 1 per cent significance levels of a chi-square distribution with 3 d.f.

		Herm					ME1					ME3				
Model	Analysis	Mean	s.d.	Zero	5%	1%	Mean	s.d.	Zero	5%	1%	Mean	s.d.	Zero	5%	1%
Additive 0.2	fix	12.80	6.84	0	75	59	0.58	2.38	90	4	1	6.70	4.96	5	33	14
Additive 0.2	est	5.09	3.71	0	20	7	3.31	5.31	33	13	5	3.34	4.43	30	13	5
Additive 0.4	fix	7.01	4.86	0	36	16	0.00	0.00	100	0	0	2.27	2.69	59	3	1
Additive 0.4	est	4.34	3.52	0	15	5	1.36	8.30	78	8	4	1.71	4.87	71	9	5
Dominant	fix	47.28	14.72	0	100	100	31.32	14.52	1	92	92	38.26	14.31	0	99	97
Dominant	est	41.13	12.89	0	100	100	37.22	13.53	0	99	98	37.34	13.44	0	99	99
Rare	fix	12.04	7.24	0	65	41	2.63	5.09	47	11	7	7.45	8.63	3	38	19
Rare	est	6.47	4.51	0	33	13	4.21	5.29	28	18	7	4.58	4.42	17	18	7

fix = analyses with the polygenic heritability fixed at the expected value in the analyses.
est = analyses with the polygenic heritability estimated.
Based on 100 replicates of each simulation.

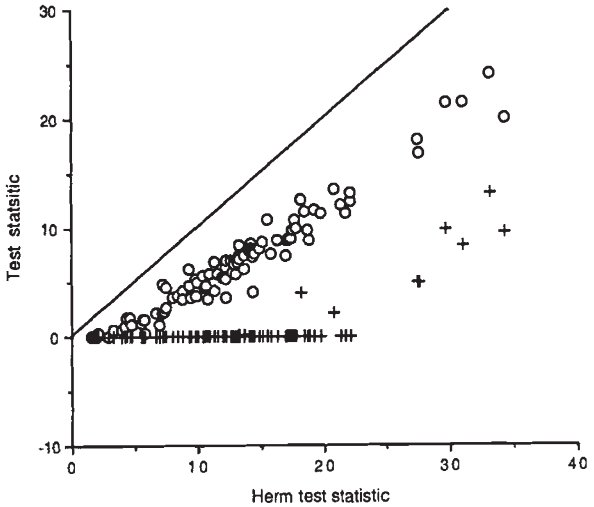

Fig. 1 ME3 (o) and ME1 (+) test statistic plotted against the Herm test statistic obtained from analysis of the same set of data for the simulated additive major gene with equal allele frequencies and an expected polygenic heritability of 0.2 fixed in the analyses. All negative test statistics have been set to zero. The solid line indicates a line of unity.

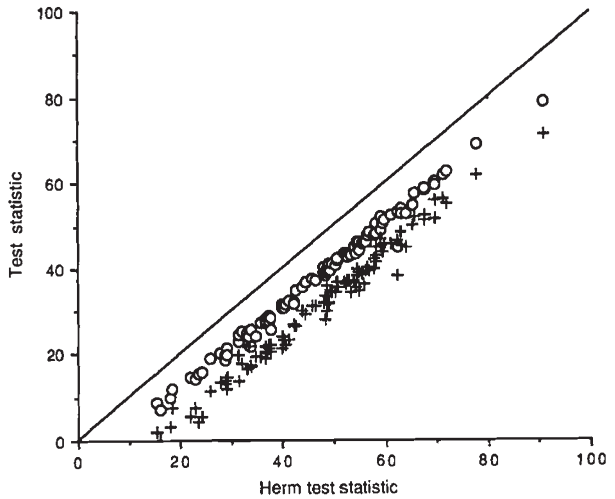

Fig. 2 ME3 (o) and ME1 (+) test statistic plotted against the Herm test statistic obtained from analysis of the same set of data for the simulated major gene with an allele with dominant effect and an expected polygenic heritability of 0.2 fixed in the analyses. The solid line indicates a line of unity.

mations except when using the Herm method with fixed polygenic heritability, evidence for a major gene was found in the highest number of analyses when one of the alleles was rare. For the Herm method with fixed heritability the number of analyses in which evidence for a major gene was found was dependent on the amount of genetic variance explained by the major gene (i.e. the additive 0.2 model is higher than the rare model which is higher than the additive 0.4 model). For all three approximations, both when the polygenic heritability was fixed and when it was estimated, the mean test statistic and number of significant results were higher when the simulated polygenic heritability was 0.2 than when it was 0.4, although the major gene explains the same proportion of the total variance.

Using the Herm and ME3 methods the mean test statistic decreased when the polygenic heritability was estimated and hence fewer analyses gave significant results. Using the ME1 method the mean test statistic distribution increased when the heritability was estimated, mainly because of the decreased number of analyses that resulted in a polygenic model with zero test statistic.

Sample size and gene magnitude effects

The mean test statistic from the analyses of 10 replicate datasets generated under each of the dominant (dominant) and additive (additive 0.2) models were considered for different numbers of sires and half-sib progeny. The relationship between the mean test statistic obtained using the Herm method and the total number of offspring in the data, regardless of the number of sires, is shown in Fig. 3 for the data containing the dominant major gene. A regression of the mean test statistic on the total number of progeny gives a linear relationship for all of the methods when analysing data containing the dominant major gene, both when estimating the polygenic heritability and when fixing it at the expected value. A quadratic relationship was also significant for most of the methods. However, fitting both the linear and quadratic components together explained less than an additional 2 per cent of the variance in the test statistic means compared with fitting just the linear component (which explains, on average 56 per cent). The constant was not significantly different from three, as expected, this being the degrees of freedom under this test. With the dominant major gene, considering only analyses with a total of 1000 offspring (with 5, 10, 20, 25, 40, 50, 100 or 200 sires), there was found to be a small negative relationship between the mean test statistic and the number of sires in the data and a small positive relationship with the number of progeny. Regression of the mean test statistic on the number of sires and the total number of progeny shows that the number of sires explain much less (only 2–3 per cent) of the variation in the mean test statistics than the total number of offspring (on average 56 per cent). However, fitting the number of sires together with the total number of offspring to the test statistic means removed or reduced the evidence for the quadratic component. Considering the combined model simulated with an additive major gene (additive 0.2 model), a linear relationship between the total number of offspring and the mean test statistic was observed. For this model, however, some of the analyses (on average 30 out of 80 analyses for the ME1 and 8 out of 80 for the ME3 method) resulted in a zero test statistic. Considering the ME1 method, where a large proportion of analyses resulted in zero test statistics, a contingency chi-square test indicated that the number of zero test statistics was not dependent on the sample size. With the ME3 method, the number resulting in zero was small, and hence a test cannot be carried out.

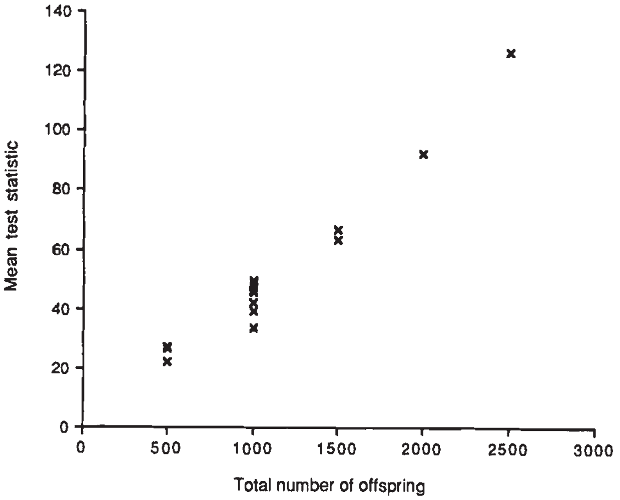

Fig. 3 Mean test statistic for different numbers of sires and half-sib progeny per sire plotted against the total number of progeny in the data from the analysis of data simulated under a mixed model containing a dominant major gene. Means are based on 10 replicates of each sample size, except for the mean from data containing 50 sires each with 20 half-sib progeny which is based on 100 replicates.

Using the data containing 1000 offspring, as the number of offspring per sire increased the mean test statistics for the ME1 and ME3 methods became closer to the value obtained from Herm. Figure 4 illustrates this, giving the mean ME1 and ME3 test statistics as a deviation from the Herm test statistic. With an increased number of offspring per sire the probability of being each major genotype for the sires is closer to one or zero for the three methods. When the sires are genotyped with absolute certainty the ME1 likelihood is equivalent to the ME3 likelihood. The increased probability with which sires are genotyped with increasing number of offspring also reduces the differ-

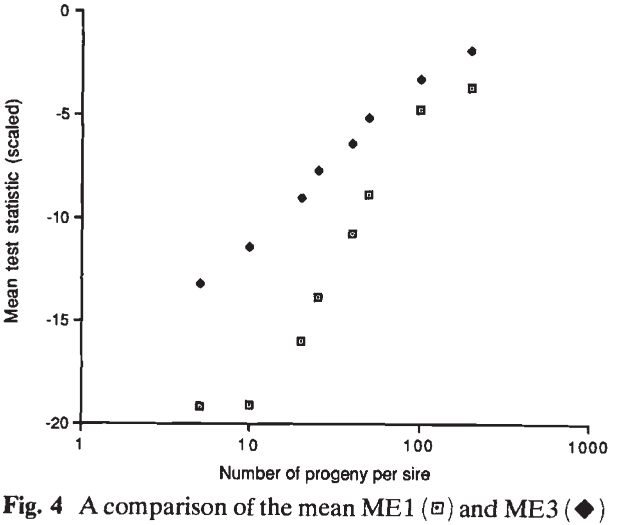

Fig. 4 A comparison of the mean ME1 (▣) and ME3 (◆) test statistics with the Herm test statistics, as a deviation from the Herm statistic, for different numbers of progeny per sire with the total number of progeny fixed at 1000.

Table 5 Mean and standard deviation of the test statistic (setting negative test statistics to zero) from the analyses of mixed-model data containing additive major genes of different effect with alleles at equal frequency. Also given are the proportion of analyses which resulted in a zero test statistic

		Herm			ME1			ME3		
Gene effect	Analysis	Mean	s.d.	Zero	Mean	s.d.	Zero	Mean	s.d.	Zero
3	fix	34.36	14.54	0.0	11.54	12.87	0.2	22.09	13.32	0.0
2	fix	12.80	6.84	0.0	0.58	2.38	0.9	6.70	4.96	0.1
1	fix	3.56	3.08	0.0	0.00	0.00	1.0	1.52	2.17	0.6
3	est	16.29	5.91	0.0	14.34	6.10	0.0	14.34	6.10	0.0
2	est	5.09	3.71	0.0	3.31	5.31	0.3	3.34	4.43	0.3
1	est	2.45	1.20	0.0	1.50	1.71	0.5	1.91	1.44	0.2

Gene effect = phenotypic standard deviations simulated between the major gene homozygotes.
fix = analyses with the polygenic heritability fixed at the expected value in the analyses.
est = analyses with the polygenic heritability estimated.
Models 1 and 3 are based on 10 repeat simulations and model 2 is based on 100.

ence between the ME3 and Herm likelihoods. However, although correct account is taken of the uncertainty in major genotype of the sire, the uncertainty in the offspring major genotypes, given the sire's is not correctly accounted for in the ME3 likelihood, and so the ME3 method is not expected to be asymptotically equivalent to the Herm likelihood.

Table 5 gives the mean test statistic from the analyses of combined model data containing additive major genes of different effect. Increasing the effect of the major gene increases the mean test statistic, the increase being in proportion to the variance of the major genes. In addition, increasing the effect of the major gene improves the ME1 and ME3 methods giving mean test statistics closer to that obtained with the Herm method. When the simulated gene has a large effect the sires are genotyped with high probabilities giving the same effect as increasing the number of offspring per sire. With the polygenic heritability estimated in the analyses, the ME1 and ME3 methods result in the same major gene models for the data with the largest gene effect.

Discussion

In this study we have developed methods of segregation analysis that can be applied to a half-sib population structure typical of many cattle and sheep populations and have shown that it is possible to detect segregating major genes in such populations. The three approximations to the combined model likelihood developed have been compared in terms of the test statistic obtained when analysing data containing phenotypes controlled by only polygenic and environmental components or these components plus a major gene.

From the analyses of polygenic data, the chi-square distribution with 3 d.f. provides a reasonable description of the observed distribution of test statistics using the Herm approximation and, hence, a suitable criterion against which to compare the test statistic for evidence of a major gene. With the ME1 and ME3 methods the observed test statistic distribution did not follow this chi-square distribution because a high proportion of analyses resulted in zero test statistics. However, the chi-square distribution does provide a conservative test for the detection of a major gene for the ME1 and ME3 methods.

The number of analyses in which evidence for a major gene was detected was dependent on the distribution of the phenotypes, as well as the proportion of the genetic variance explained by the simulated major gene. The dominant major gene both explained a high proportion of the genetic variance (79 per cent) and caused the phenotypic distribution to be skewed (mean skewness of the 100 simulations was -0.317). This skewed distribution cannot be explained by the polygenic model and hence the inclusion of a major gene can improve the likelihood significantly. Considering major genes simulated with the same effect but with different allele frequencies (rare versus additive 0.2 model), a major gene was detected more frequently when one of the alleles was rare despite the major gene explaining less of the variance. This could, again, be caused by the skewness of the distribution (mean skewness over the 100 analyses was 0.118 for the simulated major gene with a rare allele, and 0.003 when the alleles were at equal frequency). For the two simulations with an additive major gene with equal allele frequencies, a major gene was detected most frequently when the simulated gene explained a higher proportion of the genetic variance even though it explained the same proportion of the total variance.

Using the ME1 approximation to the combined model likelihood there is a problem in identifying the major gene because the test statistic is frequently zero. The relationship of non-zero ME1 test statistics with the test statistics from the same data obtained with the Herm method is strong, positive and linear, suggesting that it might be possible to reduce the threshold for significance for the test statistic or obtain a threshold by simulation. The latter approach was used by Le Roy *et al.* (1989) but involves intensive computer usage. Furthermore the large number of zero test statistics obtained analysing data containing a major gene suggests that the power will never be high.

It has been suggested that the polygenic heritability can be fixed in segregation analysis, thus reducing the number of parameters to be estimated (Le Roy *et al.*, 1989). Certainly, analysing data simulated with a major gene with the polygenic heritability fixed at the value at which it was simulated, the polygenic likelihood is much less than the combined model likelihood using the Herm and, to some extent, ME3 approximations. This occurs in part because the fixed polygenic heritability poorly explains the total genetic variation, both major gene and polygenic. When the polygenic heritability is estimated, the difference between the polygenic and combined model likelihood is reduced, because an increased heritability in the polygenic model can explain some of the major gene variance. Thus when the heritability is estimated the test statistics are smaller and this results in the major gene being detected less frequently for the additive models. A corollary to this, however, is that, if the polygenic heritability was underestimated and fixed in analyses at that value, a combined model is sometimes inferred, simply because the major gene can explain the additional polygenic variation in the data. Thus, analyses in which the polygenic heritability is fixed must be treated cautiously.

The results from the analyses with different numbers of individuals can be used to predict the number of individuals required to obtain, on average, a certain power with a given error for the simulated major genes considered here. In these analyses the non-central component of the mean test statistic approximately doubled with a doubling of the sample size, whether the increase in sample size was caused by an increase in the number of sires or the number of offspring per sire. Hence, the mean test statistic obtained from analysis of the 100 replicates, each with 50 sires and 20 half-sib progeny, given here, can be used to obtain an indication of the number of individuals required for a given power. To obtain 90 per cent power with an error of 5 per cent, assuming a non-central chi-square distribution, the non-central component of the mean test statistic needs to be 14.171 (Pearson & Hartley, 1976). Using Herm with fixed heritability, the mean test statistic for the 100 analyses was 47.28 (Table 4). Assuming the linear relationship described above only 320 individuals would be required to obtain this mean test statistic for this model. It would seem likely that for more complicated pedigree structures this simple relationship between the number of individuals and the mean test statistic will not hold. Even for the simple pedigree considered here this linear relationship is a simplification, but should give an indication of the number of individuals required in a balanced half-sib design.

For a given pedigree structure, increasing the effect of the simulated major gene increases the Herm mean test statistic approximately in proportion to the increase in the variance explained by the major gene. This relationship can be used to obtain an estimate of the number of individuals required to detect a major gene of given size and frequency.

What is wrong with the ME1 likelihood?

The ME1 approximation to the combined model likelihood is appealing because of its similarity to the classical mixed-model methods, breeding value estimation and ease of inclusion of fixed effects. Its performance in terms of the power, however, was poor. Further investigation of this approximation was carried out in order to explain this poor performance. The ME1 likelihood was investigated using a simple model where the sires can have the major genotype AA or aa and the dams have the major genotype AA. Hence, there is no segregation of the major genotype within half-sib families and the exact likelihood (5) can be written in the same form as the ME3 likelihood, estimating a transmitting ability for each major genotype of each sire. This likelihood can be simplified, because the difference between the two sire estimates (one for each major genotype) is constant over all sires and is a function of the difference between the major genotype means. This difference will be called Δ and its maximum likelihood estimate ($\hat{\Delta}$) is equal to

$$\frac{n(\mu_1-\mu_2)}{n+\lambda},$$

where μ_1 and μ_2 are the mean effects of the two major genotypes. The ME1 likelihood can be obtained by setting Δ to 0. Data were simulated with 50 unrelated sires each with 20 half-sib offspring. The polygenic heritability was 0.2 and the two major genotypes each had a frequency of 0.5 and with about two-thirds phenotypic standard deviations between the means. The major genotype means (μ_1,μ_2), population genotype frequency $[p(1)]$ and the residual variance (σ_w^2) were fixed at their ML estimates for the exact likelihood. For each method the sire effects (u_{i1} and Δ or u_i) were estimated using an EM algorithm based on first derivatives. Under these conditions the maximum value of the exact likelihood is greater than that of the ME1 likelihood. The difference between the likelihoods can be explained in terms of the sire effects, Δ, and the conditional sire genotype probabilities $[q_i(c)]$, which are different in the two methods as a consequence of the other changes. The solid line in Fig. 5 gives the difference between the exact and ME1 log likelihoods for each sire plotted against his conditional probability of being genotype 2 estimated under the ME1 likelihood $[q_{i0}(2)]$. When $q_{i0}(2)$ is equal to 1 or 0 the exact and approximate models give the same likelihood value, as then, effectively the genotype of the sire, and hence also of his offspring, is known. Otherwise the approximation always underestimates the likelihood, the largest difference being when the conditional probability under ME1 is equal to 0.5.

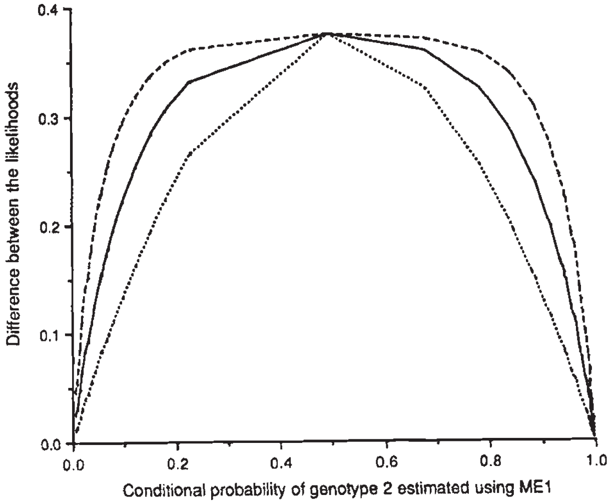

Fig. 5 Approximations to explain the difference between the exact and ME1 likelihoods for each sire plotted against the conditional probability of the sire being genotype 2 estimated using ME1. The major genotype means and frequencies and the residual variance are fixed at their ML estimates. A two genotype model with no segregation within half-sib families is used. See text for an explanation of the approximations. (——) ln L(exact)-ln L(ME1), (·····) approx. 1, (- - -) approx. 2.

A Taylor expansion can be used to explain the exact likelihood in terms of the ME1 likelihood. A first approximation would be to assume that the difference can be explained in terms of Δ. The partial second derivatives with respect to u_i and Δ can be approximated assuming that the conditional probabilities of each genotype for each sire $[q_{i0}(c)]$ are constants. The first derivatives can be arranged into a series of equations and minus the coefficient matrix from these equations used to approximate the second derivatives. The derivatives with respect to the transmitting abilities can be absorbed into the derivatives with respect to Δ to give the following equation for the exact likelihood, ignoring terms beyond a quadratic:

$$\ln L(\hat{\Delta}) \approx \ln L(0) + \frac{n+\lambda}{2\sigma_w^2}\hat{\Delta}^2 \sum_{i=1}^{s} q_{i0}(1)q_{i0}(2).$$

Alternatively the ME1 likelihood could be approximated in terms of the exact likelihood and second derivatives could be approximated using the first derivatives from the exact likelihood giving the following equation:

$$\ln L(0) \approx \ln L(\hat{\Delta}) - \frac{n+\lambda}{2\sigma_w^2}\hat{\Delta}^2 \sum_{i=1}^{s} q_{i\Delta}(1)q_{i\Delta}(2).$$

These approximations, rearranged as the difference between the likelihoods $[\ln L(\hat{\Delta}) - \ln L(0)]$, are given in Fig. 4 as approximations 1 and 2 respectively.

It can be seen that in both cases the difference between the exact and ME1 likelihoods can be correctly estimated when the conditional probability of being genotype 2 is equal to 0.5. Otherwise the conditional probabilities from ME1 underestimate the difference and those from the exact likelihood overestimate the difference. At 0.5 the conditional probability does not change with a change in Δ and the approximation can correctly estimate the difference between the approximate and exact likelihoods, because correct account is taken of the change in the transmitting ability which accompanies a change in Δ. At other values of the conditional probabilities they are not constant with a change in Δ, and as Δ approaches zero the conditional probabilities become more extreme (closer to 0 or 1).

Incorporating the observed second derivatives and making different assumptions about the change in the transmitting abilities that accompany a change in Δ were not improvements on the first approximations given above.

Fixing Δ at zero causes a term to be excluded from the log likelihood. A large component of this term is a function of the difference between the major genotype means, Δ, and the conditional sire probabilities. Using the two genotype model and assuming that there is prior knowledge that the difference between the means is zero, or more precisely, that the difference follows a normal distribution with expectation zero and with variance σ_m^2 would give the following likelihood:

$$\ln L(\text{prior}) = \ln L(\hat{\Delta}) - \frac{(\mu_1 - \mu_2)^2}{2\sigma_m^2}.$$

It can be shown that the ME1 likelihood can be written in a similar form. Δ is a function of the difference between the major genotype means and its variance can be obtained from the inverse of the matrix of partial second derivatives (which in the exact case is the same as the first derivative coefficient matrix).

It is suggested, therefore, that estimating a single sire effect is similar to using the exact likelihood with prior information that the difference between the major genotype means is zero. If there is considerable evidence in the data to suggest that this is not the case then a combined model will be more likely than a polygenic model. Otherwise, the prior will outweigh the data information and a polygenic model will be suggested with zero test statistic. From the simulation study it seems that evidence from the data has to be strong before a major gene is inferred.

Concluding remarks

The methods described here can be used to detect major genes in half-sib populations of farm animals. Using the Herm method, the power to detect a major gene was reasonable. However, with larger pedigrees and the inclusion of fixed effects the computation required may become prohibitive. The ME1 method, in comparison with the Herm, appears to require strong evidence in the data before a major gene is detected. The ME3 method detected a major gene more frequently than the ME1 but still provides a less powerful test than the Herm method.

Using paternal half-sibs the mean test statistic has an approximately linear relationship to the total number of offspring, with the constant term being equal to the degrees of freedom of the test. Hence from the mean test statistic obtained with a given sample size the number of individuals required to achieve a particular power can be predicted using the expected distribution of the non-central chi-square. The mean test statistic increases in proportion to the variance explained by the major gene for alleles of different effect but the same mode of action and frequency.

The results given here suggest that approximations to segregation analysis are capable of detecting a segregating major gene. In this study the effects of the major gene were fairly large (explaining 24 , 33 and 43 per cent of the total variance for the additive major gene with a rare allele, additive with equal allele frequencies and dominant models respectively) making the conditions for finding a major gene favourable. Nonetheless, the half-sib data structure used omits many potentially useful relationships and an improvement in the power of the methods might be obtained by using more complex pedigrees.

Acknowledgements

SAK acknowledges the financial support of the AFRC firstly through a studentship at IAPGR, Edinburgh, and subsequently through a grant at ICAPB.

References

ABRAMOWITZ, M. AND STEGUN, I. 1972. *Handbook of Mathematical Functions*. New York, Dover.

ELSTON, R. C. AND STEWART, J. 1971. A general model for the genetic analysis of pedigree data. *Hum. Hered.*, **21**, 523-542.

HANSET, R. AND MICHAUX, C. 1985a. On the genetic determinism of muscular hypertrophy in the Belgian White and Blue cattle breed. I. Experimental data. *Genet., Select., Evol.*, **17**, 359-368.

HANSET, R. AND MICHAUX, C. 1985b. On the genetic determinism of muscular hypertrophy in the Belgian White and Blue cattle breed. II. Population data. *Genet., Select., Evol.*, **17**, 369-386.

HILDEBRAND, F. B. 1974. *Introduction to Numerical Analysis. (International Series in Pure and Applied Mathematics).* McGraw-Hill Inc., New York.

HOESCHELE, I. 1988. Genetic evaluation with data presenting evidence of mixed major gene and polygenic inheritance. *Theoret. Appl. Genet.*, **76**, 81-92.

KNOTT, S. A. 1990. *Statistical methods for the detection of major genes in farm animal populations*. Ph.D. Thesis, University of Edinburgh.

LE ROY, P., ELSEN, J. M. AND KNOTT, S. 1989. Comparison of four statistical methods for detection of a major gene in a progeny test design. *Genet., Select., Evol.*, **21**, 341-357.

MERAT, P. AND RICARD, F. H. 1974. Etude d'un gene de nanisme lie aus sexe chez la poule: importance de l'etat d'engraissement et gain de poids chez l'adulte. *Ann. Génét. Sélect. Anim.*, **6**, 211-217.

MORTON, N. E. AND MACLEAN, C. J. 1974. Analysis of family resemblance. III. Complex segregation of quantitative traits. *Am. J. Hum. Genet.*, **26**, 489-503.

NUMERICAL ALGORITHMS GROUP. 1988. *The NAG Fortran Library Manual — Mark 13.* NAG Ltd. Oxford.

PEARSON, E. S. AND HARTLEY, H. O. (eds) 1976. *Biometrika tables for statisticians*. Vol. II. Biometrika Trust. Cambridge University Press, Cambridge.

PIPER, L. R. AND BINDON, B. M. 1982. Genetic segregation for fecundity in Booroola Merino sheep. In: Barton, R. A. and

Smith, W. C. (eds) *Proceedings of the World Conference on Sheep and Beef Cattle Breeding,* Vol. 1, Dunmore Press, Palmerston North, New Zealand, pp. 395–400.

PIPER, L. R., BINDON, B. M. AND DAVIS, G. H. 1985. The single gene inheritance of the high litter size of the Booroola Merino. In: Land, R. B. and Robinson, D. W. (eds) *Genetics of Reproduction in Sheep.* Butterworths, London, pp. 115–125.

ROLLINS, W. C., TANAKA, M., NOTT, C. F. G. AND THIESSEN, R. B. 1972. On the mode of inheritance of double-muscled conformation in bovines. *Hilgardia,* **41**, 433–456.

SMITH, C. AND BAMPTON, P. R. 1977. Inheritance of reaction to halothane anaesthesia in pigs. *Genet. Res.,* **29**, 287–292.

WILKS, S. S. 1938. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Stat.,* **9**, 60–62.