APPROXIMATIONS TO SEGREGATION ANALYSIS FOR THE DETECTION OF MAJOR GENES

S.A. Knott, C.S. Haley, and R. Thompson
A.F.R.C. Institute of Animal Physiology and Genetics Research, Edinburgh Research Station, Roslin,
Midlothian, EH25 9PS, UK.

SUMMARY

Segregation analysis can be used to detect an allele of large effect segregating against a polygenic background. Simulation was used to study the power of three approximations to the exact mixed model likelihood (major gene and polygenic component) for the detection of a major gene and the estimation of its effect. The approximations differed in their ability to detect a major gene, however the best method, which involves an approximation by Hermite integration, provided good estimates of its effect. Analysing the data assuming the polygenic heritability is known can provide a more powerful test for a major gene, but could potentially detect a spurious major gene if the polygenic heritability was fixed at too low a value.

INTRODUCTION

Several major genes affecting quantitative traits in farm animals have been detected in recent years. It is likely that genes of major phenotypic and potential economic importance remain to be detected. Several methods have been suggested for detecting major genes and estimating their effects (Hill and Knott,1989), but segregation analysis (Elston and Stewart,1971), which involves comparing the likelihoods of the data under different genetic models, is likely to be the most generally appropriate. To identify a major gene the likelihood of the data under the polygenic model is maximised and compared with the maximum likelihood of the data under the mixed model, containing both a polygenic component and major gene. A test statistic is provided by minus twice the difference between the natural logarithms of the maximum likelihoods under the mixed model and under the polygenic model. This deviance is expected asymptotically to follow a χ^2 distribution with degrees of freedom equal to the number of parameters fixed under the polygenic model but maximised under the mixed model. Using maximum likelihood methods the parameter estimates for the effect and frequency of the major gene in the population can be obtained.

MATERIAL AND METHODS

Likelihoods: For the purposes of this paper a sire model will be considered (i.e. paternal half-sibs) and the problem of fixed effects ignored. Under this model expressions for the polygenic and mixed model likelihoods are as follows:

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

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

Polygenic likelihood:

$$\prod_{i=1}^{s}\int_{-\infty}^{+\infty} h(u_i) \prod_{j=1}^{n} k_0(y_{ij}\,|\,\mu,u_i,\mu_0,\sigma_w^2,\sigma_u^2)\,.\,\delta u_i$$

Mixed model likelihood:

$$\prod_{i=1}^{s}\int_{-\infty}^{+\infty} \sum_{c=1}^{m} p(c)\,h(u_i) \prod_{j=1}^{n}\sum_{d=1}^{m} p(d\,|\,c)\,k_d(y_{ij}\,|\,\mu,u_i,\mu_d,\sigma_w^2,\sigma_u^2)\,.\,\delta u_i$$

Where: s - number of sires n - number of offspring per sire
m - number of major genotypes c - sire major genotype
$p(c)$ - probability of the sire's major genotype
$p(d|c)$ - probability of the offspring's major genotype given the sire's major genotype
$h(u_i)$ - probability of the sire's transmitting ability

$k_d(y_{ij}|\mu,u_i,\mu_d,\sigma_w^2,\sigma_u^2)$ - conditional probability of the offspring's phenotype given the sire's transmitting ability and the offspring's major genotype.

The mixed model likelihood involves the integration of a complex function for each sire, the function involving a summation over all possible combinations of genotypes for the half-sibs (m^n). As n increases it soon becomes impossible to calculate the likelihood. The descriptions follow of three approximations which overcome this problem.

1) Hermite integration (Herm): Approximates the integration with a weighted summation giving the mixed model likelihood:

$$\prod_{i=1}^{s}\sqrt{2\pi V^2}\sum_{g=1}^{G}\left(\sum_{c=1}^{m}p(c)\frac{1}{\sqrt{2\pi\sigma_u^2}}\exp\left[-\frac{(uc_i+Vx_g)^2}{2\sigma_u^2}+\frac{x_g^2}{2}\right]\prod_{j=1}^{n}\sum_{d=1}^{m}p(d|c)\frac{1}{\sqrt{2\pi\sigma_w^2}}\exp\left[-\frac{(y_{ij}-\mu-\mu_d-(uc_i+Vx_g))^2}{2\sigma_w^2}\right]\right)W_g$$

Where: G - number of summations

W_g - weights

uc_i - location parameter

X_g - abscissae

V - scaling parameter

2) Modal estimation (ME1): Replaces the integration with a single estimate of the mode of each sire's transmitting ability distribution, (Hoeschele, 1988; Le Roy *et al.* 1989), giving the mixed model likelihood:

$$\prod_{i=1}^{s}\sqrt{\left(\frac{2\pi\sigma_w^2}{n+\lambda}\right)}\sum_{c=1}^{m}p(c)\,h(\hat{u}_i)\prod_{j=1}^{n}\sum_{d=1}^{m}p(d|c)\,k_d(y_{ij}|\mu,\hat{u}_i,\mu_d,\sigma_w^2,\sigma_u^2)$$

Where: λ - ratio of residual to sire variance, $\hat{u}_i$ - mode of the distribution for sire i.

3) Modal estimation 3 (ME3): An extension of ME1, estimating 3 modes for each sire's transmitting ability distribution, one for each possible major genotype of the sire, giving the mixed model likelihood.

$$\prod_{i=1}^{s}\sqrt{\left(\frac{2\pi\sigma_w^2}{n+\lambda}\right)}\sum_{c=1}^{m}p(c)\,h(\hat{u}_{ic})\prod_{j=1}^{n}\sum_{d=1}^{m}p(d|c)\,k_d(y_{ij}|\mu,\hat{u}_{ic},\mu_d,\sigma_w^2,\sigma_u^2)$$

Where: $\hat{u}_{ic}$ - mode of the sire distribution given that the sire has genotype c.

Simulation: Data were simulated containing 20 half-sib progeny from each of 50 sires with all parents unrelated and randomly mated. Two mixed models were used, each with an additive polygenic variance equal to a quarter the environmental variance and a major locus with two alleles (A and a) at equal frequency and with 2 within major genotype standard deviations between the homozygotes. In the first model major gene action was additive and in the second model the allele A was completely dominant. Polygenic data were also simulated with heritability of 0.2. For each model 100 data sets were simulated.

The likelihoods were maximised using, for Herm, a quasi-Newton algorithm (Numerical Algorithms Group,1989) and for ME1 and ME3 an EM algorithm (Dempster *et al.*,1977). Initial parameter estimates for the mixed models were those used to simulate the data. For ME1 and ME3 additional initial estimates were used with the major gene explaining more of the variance for the additive model and with an additive major gene for the dominant model. The analyses of the polygenic data were started from mixed model estimates. Analyses were performed assuming that the polygenic heritability was known and fixing it at 0.2 and then repeated, relaxing this assumption and estimating the heritability. The deviance was calculated by comparing the mixed model likelihood with fixed or maximised heritability with the polygenic likelihood maximised under the same assumption. The analyses assumed that the population was in Hardy-Weinberg equilibrium and estimated the major gene allele frequency and the deviation of 2 major genotype means from the third. Thus 3 parameters were estimated in the mixed model in addition to those estimated in the polygenic model and the distribution of the deviances under the null hypothesis is expected to follow a χ^2 distribution with 3df. Herm was used with 20 summations making it virtually exact.

RESULTS

The number of analyses in which evidence of a major gene was found is summarised in table 1. The results from the polygenic data suggest that a χ^2 distribution with 3df gives an appropriate test for a major gene, at least for Herm. A dominant gene is more easily detected than an additive one. Herm detected the major gene more frequently than the other approximations.

Table 1. Number of analyses where the deviance was significant at the 5% level of a χ^2 distribution with 3df with the number of analyses giving non-zero deviances in parentheses.

Model	Herm		ME1		ME3	
	h^2 fixed	h^2 max	h^2 fixed	h^2 max	h^2 fixed	h^2 max
Polygenic	5 (100)	6 (99)	0 (5)	4 (57)	0 (22)	5 (80)
Additive	75 (100)	20 (100)	4 (20)	13 (95)	33 (95)	13 (90)
Dominant	100 (100)	100 (100)	92 (99)	99 (100)	99 (100)	99 (100)

If the estimates are unbiassed, the mean parameter estimates over the 100 simulations should give good estimates of the population parameters. Table 2 gives the average results for Herm. In all cases the results are in good agreement with the expected parameter values, although with a dominant major gene the residual variance was under estimated. When maximising the heritability, 20 of the analyses for the additive model went to a major gene model, losing the polygenic variance.

Table 2: Mean (and standard deviation) of parameter estimates from Herm.

Model	deviance	p(A)	μ(AA)	μ(Aa)	σ^2_u	σ^2_e
Additive						
Expected		0.5	20.0	10.0	5.0	95.0
h^2 fixed	12.80 (6.84)	0.50 (0.13)	18.75 (4.65)	9.11(5.16)	4.97 (0.63)	94.36(11.95)
h^2 max	5.09 (3.70)	0.50 (0.16)	19.24 (6.13)	9.13(6.62)	5.43 (5.21)	93.95(13.42)
Dominant						
Expected		0.5	20.0	20.0	5.0	95.0
h^2 fixed	47.28(14.72)	0.50 (0.05)	20.69 (3.79)	20.15(1.75)	4.78 (0.42)	90.81 (8.06)
h^2 max	41.13(12.89)	0.51 (0.05)	20.49 (3.88)	20.29(1.92)	5.03 (4.07)	90.05(12.89)

When the data were analysed with fixed heritability for the additive model, only 20 of the ME1 deviances are non-zero (table 1) with 10 greater than zero and 10 negative, the latter having gone to a local maximum. The ME3 method is a marked improvement over ME1 with only 5 zero deviances and none negative. For the dominant major gene there were no negative deviances and only one at zero (for ME1). Table 3 gives a summary of the results for the deviance and major gene parameters and their relationship with Herm results for the same data set, for those analyses which went to a non-polygenic model. Although for both methods and models the deviances (excluding the zero ones) show a good linear relationship and are highly correlated with the Herm deviances they are always less. For both models, when a non-zero deviance was obtained for ME1 and ME3 with fixed heritability, the parameter estimates were similar to the estimates from Herm. However ME1 and ME3 always over estimate the residual variance in comparison with Herm and consequently there is a consistent under estimation of the effects of the major gene compared to the Herm results for the same set of data. However, those simulations which resulted in a non-zero deviance for ME1 tended to be those which had largest deviances (mean 22.2, versus. 12.8 for all the simulations) and produced the largest major gene estimates in Herm. Thus the major gene estimates tend to be larger than those simulated.

When maximising the heritability for the additive model, 28 of the ME1 and 20 of the ME3 deviances were negative, having gone to local maxima. For the dominant model there were no negative deviances. In both models the correlations with the Herm deviances (table 3) were high although lower than those obtained with fixed heritability. For the additive model, all the non-polygenic models for ME1 (95) and all but 2 for ME3 (88) ended at a major gene model with σ^2_u equal to zero and the major gene effects were over estimated compared with Herm. For the dominant major gene, 99 ME1 and 86 ME3 analyses resulted in a major gene model, the remaining simulations giving mixed model results

DISCUSSION

Using Herm, the power to detect a major gene was reasonable and good parameter estimates were obtained. However with larger pedigrees and the inclusion of fixed effects the computation required may become prohibitive. ME1 and ME3 are faster and produce transmitting abilities immediately. However for ME1, by comparison with Herm, it appears that strong evidence is required in the data before a major gene is detected. ME3 detected a major gene more frequently than ME1.

Analysing mixed model data with the polygenic heritability fixed at the value simulated, the polygenic likelihood is much less than the mixed model likelihood. 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 mixed model likelihood is reduced, because an increased heritability in the polygenic model can explain some of the major gene variance. Thus the deviances when the heritability is maximised are smaller, and this results in the major gene being detected less frequently for the additive model. A corollary to this is that, if an underestimate of the polygenic heritability was used in analyses with fixed heritability, a mixed model might be inferred, simply because the major gene can explain the additional polygenic variation.

Table 3. Parameter estimates from ME1 and ME3 analyses with non-zero deviances, and the correlation with and regression on Herm estimates for the same set of data.

		h^2 fixed				h^2 maximised			
		dev	p(A)	μ(AA)	μ(Aa)	dev	p(A)	μ(AA)	μ(Aa)
Additive major gene									
ME1	mean	1.91	0.47	19.20	9.33	2.43	0.50	22.10	10.96
	sd	5.14	0.22	4.12	5.44	5.42	0.09	2.13	1.71
	slope	0.70	1.65	1.29	1.67	1.13	0.40	0.12	0.11
	r	0.85	0.88	0.87	0.90	0.79	0.71	0.33	0.40
ME3	mean	7.05	0.49	16.22	8.03	3.33	0.51	21.86	10.82
	sd	4.84	0.18	3.68	3.66	4.55	0.10	2.46	2.18
	slope	0.72	1.14	0.65	0.60	1.03	0.50	0.27	0.27
	r	0.97	0.82	0.70	0.78	0.86	0.70	0.55	0.67
Dominant major gene									
ME1	mean	31.64	0.52	18.75	20.74	37.23	0.52	23.44	19.05
	sd	14.24	0.05	2.46	1.53	13.53	0.06	4.04	1.78
	slope	0.98	0.58	0.57	0.83	0.99	0.77	0.66	0.63
	r	0.99	0.58	0.88	0.94	0.95	0.60	0.63	0.68
ME3	mean	38.26	0.51	20.24	19.80	37.34	0.51	22.72	19.26
	sd	14.31	0.04	2.50	1.63	13.44	0.05	3.75	1.65
	slope	0.97	0.66	0.61	0.90	1.00	0.93	0.70	0.67
	r	1.00	0.82	0.93	0.97	0.96	0.86	0.73	0.79

When maximising the heritability, both major gene and polygenic variation are no longer required to explain the proportion of genetic to environmental variation in the data. The methods, especially ME1 and ME3, seemed to have difficulty in distinguishing the two sources of genetic variation and suggest a model containing either all major gene or all polygenic variation. However, the major gene model was only obtained when a major gene existed in the data and gave a good indication as to whether it had additive or dominant effect. Hence, although more conservative than with the heritability fixed, there does not seem to be a potential problem of incorrectly inferring major genes.

The results given here suggest that approximations to segregation analysis are capable of finding and estimating the effects of a segregating major gene. In this study the effects of the major gene were fairly large (explaining 33% and 43% of the total variance for the additive and dominant models respectively) making the conditions for finding a major gene favourable. Work is required to investigate more realistic models including, for example, fixed effects. This half-sib data structure ignores many potentially useful relationships and an improvement in the power of the methods might be obtained by using a complete pedigree.

ACKNOWLEDGEMENTS

SAK acknowledges receipt of an AFRC studentship.

REFERENCES

DEMPSTER, A.P., LAIRD, N.M. and RUBIN D.B. 1977. *J. Royal Statist. Soc., Series B*, **39**: 1-38.
ELSTON, R.C. and STEWART, J. 1971. *Hum. Hered.* **21**: 523-542.
HILL, W.G. and KNOTT, S.A. 1989 In *Advances in Statistical Methods for Genetic Improvement of Livestock*. (Eds. D.Gianola and K.Hammond) pp 517-537. AGBU, University of New England, Australia.
HOESCHELE, I. 1988. *Theor. Appl. Genet.* **76**: 81-92.
LE ROY ,P., ELSEN, J.M. and KNOTT, S. 1989. *Genet. Sel. Evol.* **21**: 341-357.
NUMERICAL ALGORITHMS GROUP 1988. *The NAG Fortran Library Manual - Mark 13.* NAG Ltd.