Theor Appl Genet (1994) 89:318–322

S. A. Knott

Prediction of the power of detection of marker-quantitative trait locus linkages using analysis of variance

Received: 6 July 1993 / Accepted: 28 February 1994

Abstract Analysis of variance can be used to detect the linkage of segregating quantitative trait loci (QTLs) to molecular markers in outbred populations. Using independent full-sib families and assuming linkage equilibrium, equations to predict the power of detection of a QTL are described. These equations are based on an hierarchical analysis of variance assuming either a completely random model or a mixed model, in which the QTL effect is fixed. A simple prediction of power from the mean squares is used that assumes a random model so that in the mixed-model situation this is an approximation. Simulation is used to illustrate the failure of the random model to predict mean squares and, hence, the power. The mixed model is shown to provide accurate prediction of the mean squares and, using the approximation, of power.

Key words Quantitative trait loci · Analysis of variance · Genetic linkage

Introduction

Analysis of variance would seem to provide a simple method of combining information from many families in order to detect linkage of a quantitative trait locus (QTL) to a marker. In an outbreeding population the same marker allele will not be associated with the same QTL allele in all families. Therefore, evidence for a linked QTL cannot be obtained at a population level from overall mean differences between marker genotypes. A linked QTL is, however, expected to produce mean differences between marker genotypes within families. Using an hierarchical analysis of variance, the test for a linked QTL comes from the comparison of the between-marker within-families mean square with the residual mean square and can be tested as an F ratio (Hill 1975; Soller and Genizi 1978). Following standard procedures (for example see Searle 1971, p. 394 ff.), expected values of the mean squares and hence the F ratio can be written in terms of the variance components, functions of fixed effects, and coefficients that depend on the number of observations. The power is the probability that this F ratio is greater than the critical value from the central F distribution with the same degrees of freedom. Hence, given the pedigree structure, it is possible to predict the power of detection of a given QTL (Hill 1975; Soller and Genizi 1978). Hill (1975) described an hierarchical model with all effects random for the detection of a linked QTL using a population of full-sib families. A simulation study illustrates that the random model does not provide an adequate prediction of the mean squares and thus of the power. Soller and Genizi (1978) used an alternative mixed model and in the present paper a more detailed description of the analysis of variance for this model is presented. The full-sib family types described by Hill (1975) are used. This mixed model has a random effect between families and a fixed effect due to the QTL linked to the marker. An approximation to power is obtained using equations suitable for a simple random model. A simulation study illustrates the agreement between both the predicted and the observed mean squares and power.

Model

Consider a population of full-sib families (parents plus offspring) where the parents mated at random. Information on a segregating marker is available for parents and offspring, phenotypic information is recorded only on the offspring. A QTL is segregating in the population and is assumed to be in linkage equilibrium with the marker. The model for the kth full-sib with marker genotype j in family i can be written as follows:

$$y_{ijk} = \mu + u_i^* + \gamma_{ij}^* + e_{ijk}^* \quad (1)$$

where μ is the population mean, u_i^* is the effect of family i, γ_{ij}^* is the effect of the jth marker genotype in the ith family, and e_{ijk}^* is the

Communicated by C. Smith

S. A. Knott (✉)
Institute of Cell, Animal and Population Biology, University of Edinburgh, West Mains Road, Edinburgh, EH9 3JT, UK

residual effect for the kth full-sib of the jth marker genotype in the ith family.

The population will consist of different types of families with respect to the marker genotypes of the parents. If both parents are homozygous at the marker the family will be completely uninformative for linkage. Informative families consist of those where one parent is homozygous, and those where neither parent is homozygous, for the marker. In the latter case the parents can either have the same genotype or, in situations where there are more than two alleles at the marker locus, different genotypes. The population can either be divided on the basis of the marker genotypes of the parents or analysed as a whole. Although in practice it would be more efficient to analyse informative families together, to allow comparison with previous work (e.g., Hill 1975) the families from different marker classes will be considered separately. Two types of family will be investigated, those where one parent is homozygous at the marker locus and one is heterozygous (backcross-type or BC families) and those where both parents are heterozygous for the same genotype at the marker locus (intercross-type or IC families). A simple genetic model, with an additive QTL with two alleles at equal frequency, with half the difference between QTL homozygotes equal to δ, and with no recombination between the marker and QTL, will be considered to illustrate the problem.

Analysis of variance

Hierarchical analysis with main and nested effect random

As the population is assumed to be in linkage equilibrium, the data can be analysed as an hierarchical analysis of variance (Hill 1975), with variation between families, between marker genotypes within families, and within marker genotypes within families. Hill (1975) assumes all effects are random. With a constant number of full-sibs per family and the expected segregation ratios at the marker within families (i.e., with half the full-sibs in each marker class for the BC situation, and with one quarter of the full-sibs in each homozygous marker class and half in the heterozygous class for the IC), the analysis of variance can be written as follows (Hill 1975):

Source	df	E[MS]
Between families	$(N-1)$	$MS_f = \sigma_w^2 + k_2\sigma_m^2 + n\sigma_f^2$
Between marker genotypes within families	$N(M-1)$	$MS_m = \sigma_w^2 + k_1\sigma_m^2$
Within marker genotypes within families	$N(n-M)$	$MS_r = \sigma_w^2$

where N is the number of families, M is the number of marker genotypes in each family, i.e., $M = 2$ for the BC and $M = 3$ for the IC, n is the number of full-sibs per family, σ_f^2 is the between-family variance component, which, for the simple model described above, is equal to $\delta^2/4 + \sigma_u^2$, where σ_u^2 is the non-QTL variance between families, σ_m^2 is the between-marker-genotypes within-families variance component, which is equal to $\delta^2/8$, and σ_w^2 is the within-marker-genotypes within-families variance component, which is equal to $\sigma_e^2 + \delta^2/8$, where σ_e^2 is the individual environmental variance component plus the within-full-sib family genetic variance due to genes other than the QTL linked to the marker.

For all variance components the contribution due to the QTL is a function of $p(1-p)\delta^2$, where p is the QTL allele frequency ($p = 0.5$ in this example), i.e., a function of the contribution of the QTL to the variance.

With one parent heterozygous and one homozygous at the marker (BC families) $k_1 = k_2 = n/2$. With both parents heterozygous for the same genotype at the marker (IC families) $k_1 = 5n/16$ and $k_2 = 3n/8$.

In both cases the test for evidence of linkage comes from the ratio of the 'between-marker-genotypes within-families' mean square and the 'within-marker-genotypes within-families' mean square.

Hierarchical analysis with the QTL effect fixed

The analysis of Hill (1975) assumes all effects are random; however, it is generally assumed that the effects of the QTL genotypes are the same in all families and, hence, should be considered as fixed. This alternative model for the data can be written as follows:

$$y_{ijk} = \mu + u_i + q_i + \gamma_{ij} + \varepsilon_{ijk} + e_{ijk} \quad (2)$$

The family component u_i^* in (1) is now composed of a random component (u_i, distributed with variance σ_u^2) and a fixed component due to the QTL (q_i). Likewise the residual (e_{ijk}^*) in (1) is composed of both a random component (e_{ijk}, with variance σ_e^2) and a fixed component due to the QTL (ε_{ijk}). γ_{ij} is the fixed effect of the jth marker genotype in the ith family.

For a QTL with two alleles there are 16 different combinations of QTL genotypes for the parents taking account of the origin of the QTL. For each family type the expected value of q_i, γ_{ij} and ε_{ijk} can be written in terms of the effect of the QTL. The expected mean QTL effect of the family relative to the QTL mid-homozygote value is q_i. The expected deviation of each marker genotype from the expected family mean is γ_{ij}. The expected deviation of each QTL genotype from the expected marker genotype mean is ε_{ijk}. For example, consider a BC family with the additive QTL described previously. If both parents are heterozygous at the QTL (e.g., with parental genotypes aq/AQ and Aq/AQ) there are four possible offspring genotypes expected at equal frequency

Genotype:	AQ	AQ	aq	aq
	AQ	Aq	AQ	Aq
Effect:	δ	0	0	$-\delta$

$$q_i = 0$$

$$\gamma_{i1} = \frac{\delta}{2} \qquad \gamma_{i2} = -\frac{\delta}{2}$$

$$\varepsilon_{i11} = \frac{\delta}{2} \quad \varepsilon_{i12} = -\frac{\delta}{2} \quad \varepsilon_{i21} = \frac{\delta}{2} \quad \varepsilon_{i22} = -\frac{\delta}{2}$$

For given QTL genotypes of the parents, q_i and γ_{ij} are known precisely for each full-sib. The value of ε_{ijk}, however, is not always known as it depends on the QTL genotype of the individual which, in some circumstances with this model, can be one of two alternatives. Hence, the mean effect due to the QTL of the marker genotypes within families is not fixed but will depend on the random sampling of the possible QTL genotypes. That is, it is the effect of the QTL and not the marker that is fixed and, therefore, even in the absence of any residual random variance ($\sigma_e^2 = 0$), unlike the usual fixed-effect models, the residual mean square ('within-marker-genotypes within-families') will not be zero.

Derivation of expected mean squares

Following Searle (1971, 394 ff.) the equations for the mean squares can be written in terms of the components of model (2) rather than the phenotypes. The expectation of these equations can be simplified to obtain expressions for the mean squares composed of the variance components and functions of the fixed effects.

The usual assumptions were made about the random effects, that is:

$$E[u_i] = 0 \quad E[u_i^2] = \sigma_u^2 \quad E[u_i, u_{i' \neq i}] = 0$$

$$E[e_{ijk}] = 0 \quad E[e_{ijk}^2] = \sigma_e^2 \quad E[e_{ijk}, e_{ijk' \neq k}] = 0$$

Extensions to the standard models presented by Searle (1971) are required to account for the random segregation of the QTL within the marker class and random sampling of the parents with respect to their QTL genotype. First, QTL segregation within the offspring marker class will be considered. Depending on the QTL genotype of the parents, with the model presented here there may be either one or two possible QTL genotypes for full-sibs with the same marker genotype. When two QTL genotypes are possible an offspring can be either

genotype with equal probability. The number of full-sibs with each QTL genotype within the marker genotype is therefore binomially distributed with a probability of one half. The expectation of the square of the numbers of offspring in the two QTL classes within a marker genotype within a family (i.e., $E[n_1 n_1]$, $E[n_1 n_2]$ and $E[n_2 n_2]$ where n_1 is the number of full-sibs with one QTL genotype within a marker class and n_2 the number of the second QTL genotype) is required to derive the mean squares and this expectation involves the variance and covariance of the numbers in each class {i.e., $E[n_1 n_1] = E[n_1]E[n_1] + var(n_1)$ and $E[n_1 n_2] = E[n_1]E[n_2] + cov(n_1 n_2)$, as n_1 and n_2 are binomially distributed with $n_1 + n_2 = n_m$ (the number of full-sibs in a marker class) and $p_1 = p_2 = 0.5$ (the probability of each QTL genotype within a marker genotype), $E[n_1] = n_m p_1$, $var(n_1) = n_m p_1(1 - p_1)$ and $cov(n_1 n_2) = -n_m p_1 p_2$}. The covariance of the number of offspring with a given QTL genotype in different marker genotype classes is zero.

Secondly, the effect of random sampling of parents with respect to the QTL will be considered. Families contribute different amounts to the 'between-marker within-family' mean square and to the 'within-marker within-family' mean square depending on the QTL genotypes of the parents. Terms have to be accumulated over all family types in proportion to their expected frequency in the population. Assuming random mating, the expected frequency of each of the 16 types of family can be obtained simply from the QTL allele frequency. With two QTL alleles at equal frequency, all family types have the same expected frequency of 1/16. The 'between-family' mean square involves the expectation of the square of the numbers of each family type. The number of families of each type is multinomially distributed. The expected square of the numbers can be obtained in the same way as for the numbers of offspring in each QTL genotype within markers, where the binomial distribution was used [i.e., n_m would be the total number of families (N) and $p_1 \ldots p_{16}$, the frequency of each family type, would depend on the QTL allele frequency].

Incorporating the values described above into the equations for the expected mean squares and simplifying them gives the following:

One parent heterozygous and one homozygous at the marker (BC families):

Source	$E[MS]$
Between families	$\sigma_e^2 + \frac{\delta^2}{8} + n\frac{\delta^2}{4} + n\sigma_u^2$
Between marker genotypes within families	$\sigma_e^2 + \frac{\delta^2}{8} + n\frac{\delta^2}{8}$
Within marker genotypes within families	$\sigma_e^2 + \frac{\delta^2}{8}$

Both parents heterozygous for the same genotype at the marker (IC families):

Source	$E[MS]$
Between families	$\sigma_e^2 + \frac{\delta^2}{8} + n\frac{\delta^2}{4} + n\sigma_u^2$
Between marker genotypes within families	$\sigma_e^2 + \frac{1}{(M-1)}\frac{\delta^2}{8} + \frac{n}{(M-1)}\frac{\delta^2}{8}$
Within marker genotypes within families	$\sigma_e^2 + \frac{(n-2)}{(n-M)}\frac{\delta^2}{8}$

The ratio of the 'between-marker-genotypes within-families' and the 'within-marker-genotypes within-families' mean squares for the two different models can be written as follows:

	Random model (Hill 1975)	Mixed model
BC	$1 + \frac{n\delta^2}{2(8\sigma_e^2 + \delta^2)}$	$1 + \frac{n\delta^2}{8\sigma_e^2 + \delta^2}$
IC	$1 + \frac{5n\delta^2}{16(8\sigma_e^2 + \delta^2)}$	$1 + \frac{\frac{n^2 - 4n + 1}{n-3}\delta^2}{2\left(8\sigma_e^2 + \frac{n-2}{n-3}\delta^2\right)}$

Prediction of power

For a balanced random model assuming normally-distributed random effects, power can be predicted from a central F distribution using the following relationship (e.g., Scheffé 1959, p. 227):

$$\text{Prob}(T > F_0) = \text{Prob}\left(F > \frac{F_0}{G}\right)$$

Where T is the test variable (F ratio) appropriate for the test with n_1 and n_2 degrees of freedom, F_0 is the critical value (e.g., 5% level) from a central F distribution with the same degrees of freedom, F is a central F variable, again with n_1 and n_2 degrees of freedom, and G is a function of the parameters such that $T = FG$. For a simple balanced random model, G is equal to the value of the F ratio calculated using the true parameter values.

An approximation to power was obtained for the situations considered here using the equation given above replacing G with the expected value of the F ratio. This is no longer exact as the data are unbalanced and the model not random and, hence, $T \neq FG$ as the mean squares are no longer distributed as chi-square variables.

Simulation study

In order to check the predicted mean squares given above, data were simulated containing independent full-sib families. A marker locus and a single QTL with additive effect and two alleles at equal frequency were segregating with no recombination between the marker and the QTL. The parental generation was in linkage equilibrium. There was no environmental variance, no genetic variance other than that generated by the QTL or any additional common family component (i.e., $\sigma_e^2 = 0$ and $\sigma_u^2 = 0$). In one set of 1 000 simulations one parent of each family was homozygous at the marker and the other heterozygous (BC families) and in the other set both parents were heterozygous for the same marker genotype (IC families).

Data were also simulated in order to see whether the predicted power was in good agreement with that observed over multiple simulations. The data were similar to those described above except that a random environmental component was also included and in some sets a random component common to families. Power was predicted as for a balanced random design, using the equation given previously. One-thousand replicate data sets for each family type were simulated and analysed.

Results

The averages of the mean squares over the 1 000 replicates were calculated and are shown in Table 1 with the values that would be predicted with the two models described above and the parameter values used to simulate the data. The predicted mean squares from the random model do not agree with the values observed in the simulation study, with the 'between-family' mean square over estimated and the 'between-marker within-family' mean square underestimated. For the IC situation the residual mean square is also very different from expected. In both cases the predicted F ratio is lower than that observed, which would lead to the predicted power being an underestimate. The predicted mean squares from the mixed model, however, are in agreement with the simulated values and would give an unbiased prediction of the F ratio.

Table 2 gives the observed mean F ratios and power (defined as the percentage of significant analyses) and

Table 1 Observed mean squares (with empirical standard error in parentheses) compared with predicted values from an analysis of variance. One-thousand replicate simulations were carried out with N families each with n full-sibs and the expected number of full-sibs per marker class within families. Sets of families where either one parent was homozygous and one heterozygous at the marker locus (BC families), or where bothwere heterozygous for the same genotype (IC families), were simulated. An additive gene was simulated with two alleles at equal frequency and effect δ (half the QTL homozygote difference). There was no recombination between the marker and the QTL. Variation between individuals was due to the QTL only

Marker type	N	n	δ^2	Simulated[a]			Predicted:random			Predicted:mixed		
				MS_f^*	MS_m	MS_r	MS_f	MS_m	MS_r	MS_f	MS_m	MS_r
BC	64	40	200	2029 (9.84)	1022 (4.15)	25.02 (0.10)	2525	525.0	25.0	2025	1025	25.00
IC	64	40	200	2024 (9.78)	513.2 (2.42)	25.84 (0.13)	2400	337.5	25.0	2025	512.5	25.68
BC	32	20	100	509.9 (3.55)	263.9 (1.52)	12.44 (0.07)	637.5	137.5	12.5	512.5	262.5	12.50
IC	32	20	100	511.0 (3.66)	130.9 (0.85)	13.30 (0.09)	606.3	90.6	12.5	512.5	131.3	13.24

[a] MS_f is the between-family mean square; MS_m is the between-marker within-family mean square; MS_r is the within-marker within-family mean square

Table 2 Observed power and F-ratios and those predicted using the mixed model. One-thousand replicate simulations were carried out with N families each with n full-sibs and the expected number of full-sibs per marker class within families. Sets of families where either one parent was homozygous and one heterozygous at the marker locus (BC families), or where both were heterozygous for the same genotype (IC families), were simulated. An additive gene was simulated with two alleles at equal frequency. There was no recombination between the marker and the QTL

Marker type	N	n	λ[a]	d	ϕ	Observed		Predicted	
						F	Power	F	Power
BC	25	40	0.05	0.32	0.0	1.51	44.6	1.49	44.2
IC	25	40	0.05	0.32	0.0	1.25	30.0	1.24	29.7
BC	25	40	0.05	0.22	1.0	1.49	41.7	1.49	44.2
IC	25	40	0.05	0.22	1.0	1.23	29.0	1.24	29.7
BC	50	20	0.10	0.45	0.0	1.50	66.0	1.49	63.8
IC	50	20	0.10	0.45	0.0	1.23	40.6	1.23	41.5
BC	50	20	0.10	0.32	1.0	1.49	63.4	1.49	63.8
IC	50	20	0.10	0.32	1.0	1.23	39.8	1.23	41.5

[a] λ and ϕ are the expected population QTL variance (calculated from the simulated parameter values, i.e., $\delta^2/2$, where δ is half the QTL homozygote difference) and the between-family variance component (σ_u^2) used to simulate the data as a proportion of the individual environmental variance (σ_e^2), respectively (i.e., $\lambda = \delta^2/2\sigma_e^2$ and $\phi = \sigma_u^2/\sigma_e^2$). d is the additive effect of the QTL in residual standard deviations [i.e., $\delta = d(\sigma_e^2 + \sigma_u^2)^2$]

those predicted from the mixed model for a range of situations. It can be seen that the predicted power provides a good indication of the observed power, despite using a method which is only expected to provide an approximate result.

Extensions to more general prediction formulae

The most basic situation has been assumed in order to illustrate the failure of the random model presented by Hill (1975) to explain the data and to describe a more suitable mixed model. A more general model could account for recombination between the QTL and the marker or allow for random segregation of the marker alleles within families or unequal family size. In addition, the molecular markers now being used frequently have more than two alleles at a locus. This creates an additional family type with respect to the markers; that is, one where all four marker genotypes can be distinguished in the offspring. The expected mean squares obtained assuming a mixed model can be extended to incorporate these effects but this requires more laborious algebra.

It would also be useful to consider using information from the whole population rather than subsets of it (i.e., combining IC, BC and any other informative family types). As the F ratio of interest does not depend on the 'between-family' mean square, the F ratio for whole populations of mixed family types can be obtained simply by calculating the mean expected within- and between-marker within-families sums of squares using the expected frequencies of the different family marker types. These can be divided by the expected degrees of freedom to obtain an expression for the relevant mean square.

Discussion

The mixed model presented here provides a useful prediction of the power of a given situation using analysis of variance. This is despite the fact that families are heterogeneous with respect to the QTL, and therefore also for the within-family variance, so that the assumptions underlying the use of analysis of variance are not strictly correct. The power predicted would be higher than that suggested by the equations given by Hill (1975).

The equations for the mean squares, and hence power, depend on the QTL allele frequency which will be unknown in practice. In the example given in this paper, the allele frequency used for prediction was the same as that for simulation of the data. Likewise, there

was no recombination between the marker and the QTL in either the prediction or the simulation. This is unlikely to be the case and any recombination will reduce the power. In both of these cases predictions could be made under a range of values for the relevant parameters. Soller and Genizi (1978) discuss the effect of the QTL allele frequency and recombination on the power.

For ease and speed of computation, power was predicted from the F ratio using the simple formula for a balanced random model. In an unbalanced situation (as used here for the IC situation) this no longer holds, but an approximation using the central F distribution can be obtained (Scheffé 1959, p. 254). For a fixed-effects model the F ratio under the alternative hypothesis is distributed as a non-central F distribution. Estimates of power for given degrees of freedom, significance level, and a non-centrality parameter, have been calculated (for example Pearson and Hartley 1954). Alternatively Scheffé (1959, p. 414) gives a central approximation to the non-central F distribution. In the mixed model used here, however, the expected 'within-marker within-family' mean square is not a central chi-square variable and hence the F ratio obtained will not be distributed as a non-central F. Genizi and Soller (1979) derive a formula to approximate power for this type of mixed model using a Laguerre series expansion. The approach used here, following the equations for balanced random models, is much easier to calculate and, despite being formally inappropriate, gave predicted powers close to those observed for the same F ratio in the situations examined.

Soller and Genizi (1978) present the numbers of individuals required to obtain 90% power in given situations. The test statistics, although not derived in their paper, appear to be consistent with those derived here using a mixed rather than random model. This is supported by my simulations of the marker and the QTL situations which they explore (data not shown). Soller and Genizi suggest the omission from the IC families of the offspring that are heterozygous at the marker locus, and compare this situation with BC marker families. In both cases the F ratio is written in terms of the within-family QTL variance that is explained by the markers as a proportion of the residual variance. They suggest that this proportion in the IC is twice that in the BC, giving F ratios of $1 + n\delta^2/8\sigma_e^2$ and $1 + n\delta^2/4\sigma_e^2$ for the BC and IC (omitting heterozygotes), respectively. This is approximate for the BC families (compare with the ratio given in this paper), but will hold for most situations in practice, where the effect of the QTL is small compared with the residual variance.

Using the same families but omitting the heterozygous offspring in the IC situation increases the predicted power, despite the loss in degrees of freedom, as it increases the difference between the marker genotypes and reduces the residual variance. For the populations considered in Table 2, the predicted power for these families would be 44% and 63% which is similar to that obtained for the BC families. For a range of parameter values investigated, omitting the heterozygous offspring resulted in greater predicted power (data not shown).

Although prediction of power in more general models, with multiple marker alleles, recombination between the QTL and marker etc., requires tedious algebra, the implementation of analysis of variance with real or simulated data is relatively simple. Hence analysis of variance provides a useful tool to enable quick screening of a population preliminary to the use of computationally-demanding methods such as maximum likelihood. Maximum likelihood may, however, provide more power as well as a better framework for the estimation of QTL effects.

Acknowledgements I am grateful to Chris Haley and Bill Hill for some valuable discussions and to A. Genizi and the referees for their comments on the manuscript. The support of the Agricultural and Food Research Council (AFRC) is acknowledged.

References

Genizi A, Soller M (1979) Power derivation in an ANOVA model which is intermediate between the "fixed-effects" and the "random-effects" models. J Stat Planning and Inference 3:127–134

Hill AP (1975) Quantitative linkage: a statistical procedure for its detection and estimation. Ann Hum Genet 38:439–449

Pearson ES, Hartley HO (1954) Biometrika tables for statisticians. Biometrika Trust, Cambridge University Press, England

Scheffé H (1959) The analysis of variance. Wiley, New York

Searle SR (1971) Linear models. Wiley, New York

Soller M, Genizi A (1978) The efficiency of experimental designs for the detection of linkage between a marker locus and a locus affecting a quantitative trait in segregating populations. Biometrics 34:47–55