Theor Appl Genet (1996) 93:1193–1201

A. Charcosset · A. Gallais

Estimation of the contribution of quantitative trait loci (*QTL*) to the variance of a quantitative trait by means of genetic markers

Received: 30 April 1995 / Accepted: 2 February 1996

Abstract The estimation of the contribution of an individual quantitative trait locus (QTL) to the variance of a quantitative trait is considered in the framework of an analysis of variance (ANOVA). ANOVA mean squares expectations which are appropriate to the specific case of QTL mapping experiments are derived. These expectations allow the specificities associated with the limited number of genotypes at a given locus to be taken into account. Discrepancies with classical expectations are particularly important for two-class experiments (backcross, recombinant inbred lines, doubled haploid populations) and F_2 populations. The result allows us firstly to reconsider the power of experiments (i.e. the probability of detecting a QTL with a given contribution to the variance of the trait). It illustrates that the use of classical formulae for mean squares expectations leads to a strong underestimation of the power of the experiments. Secondly, from the observed mean squares it is possible to estimate directly the variance associated with a locus and the fraction of the total variance associated to this locus (r_l^2). When compared to other methods, the values estimated using this method are unbiased. Considering unbiased estimators increases in importance when (1) the experimental size is limited; (2) the number of genotypes at the locus of interest is large; and (3) the fraction of the variation associated with this locus is small. Finally, specific mean squares expectations allows us to propose a simple analytical method by which to estimate the confidence interval of r_l^2. This point is particularly important since results indicate that 95% confidence intervals for r_l^2 can be rather wide: 2–23% for a 10% estimate and 8–34% for a 20% estimate if 100 individuals are considered.

Communicated by J. Beckmann

A. Charcosset (✉) · A. Gallais
INRA-UPS-INAPG, Station de Génétique Végétale du Moulon, Ferme du Moulon, 91190 Gif sur Yvette, France

A. Gallais
INAPG, 16 rue Claude Bernard, 75005 Paris, France

Key words Quantitative trait locus (QTL) · Genetic markers · Analysis of variance (ANOVA)

Introduction

The use of markers in investigating the genetics of the quantitative traits has been undertaken in many studies since the pioneer work of Sax (1923). Interest for this approach was enhanced during the last decade by the development of molecular techniques, such as restriction fragment length polymorphism (RFLP), which provide high numbers of neutral markers (Beckmann and Soller 1983; Burr et al. 1983). Basically, the aim of these and present studies is to resolve the variation of a quantitative trait into individual factors, i.e. assign the loci involved in the variation of a quantitative trait (QTLs) at chromosomal locations and estimate their effects. The estimation of the contribution of each QTL to the variance of the trait of interest is then particularly interesting. For instance, it is classically reported that the QTL displaying the largest effect accounts for a given percentage of the variation of the trait of interest.

The aim of this paper is to investigate, from a quantitative genetics point of view, the estimation of the contribution of an individual QTL or marker locus to the variation of a quantitative trait. Basically, this problem can be partitioned into three parts: (1) will the QTL be detected or not; (2) if detected, what is the most appropriate estimator of the contribution of the QTL to the variance of the trait; (3) what is the precision of the estimation of this contribution. These points will be considered in the framework of an analysis of variance (ANOVA).

ANOVA mean squares expectation is the key factor to achieve previous objectives. In the specific case of QTL mapping experiments, several studies have considered the expectation of the mean square associated with locus effect (Hill 1975; Knapp and Bridges 1990; Knott 1994). In opposition to Hill (1975), Knapp and Bridges (1990) and Knott (1994) proposed that QTL

effects should be considered as being fixed. One drawback associated with this last approach is that the expected mean squares associated with a locus is not a direct function of the variance associated with this locus. In this study, we will reconsider the expectation of the mean square associated with locus effect. We will show that, in most cases, the expectation of the mean square associated with locus effect can be expressed as a function of the variance associated with this locus. The formula of mean squares expectation derived for the specific case of locus effect differs from classical formulae due to the restricted number of genotypes at a given locus.

We will show that this result allows the power of experiments to detect QTLs to be reconsidered. Furthermore, it allows an easy and unbiased estimation of the variance associated to a locus and of the fraction of the total variance that is associated with this locus. This estimator avoids the bias associated with the classical estimation (R^2) derived from the ratio of the sums of squares or other methods. We will present a simple analytical method to estimate its confidence interval.

Genetic model

We will consider a general model for any plant population structure: F_2, backcross (BC), recombinant inbred lines (RIL), doubled haploid (DH) populations and populations that involve higher numbers of genotypes at a given locus. We will define the value of an individual in a general way, that is to say that we will consider either (1) observations on the individual per se, or (2) the average value of its progeny in a given experimental design. At a given locus l, the population of interest (called the reference population in what follows, infinite in size) has C_l genotypes, which will be noted i_l ($i_l = 1$ to $i_l = C_l$). These genotypes have frequencies f_{i_l} in the population.

For the loci involved in the variation of the quantitative trait of interest (i.e. QTLs), further indicated as q, each genotype has an effect g_{i_q}. Definition of genotype effects follows the condition that $\Sigma_{i_q=1}^{C_q} f_{i_q} g_{i_q} = 0$. The genetic value of individual j at locus q can be defined as: $y_q^j = \Sigma_{i_q=1}^{C_q} g_{i_q} \theta_{i_q}^j$, where $\theta_{i_q}^j = 1$ if the genotype of individual j at locus q is i_q, $\theta_{i_q}^j = 0$ otherwise. Within the reference population, $\theta_{i_q}^j$ is a random variable that takes values 1 or 0 with probabilities f_{i_q} and $1 - f_{i_q}$, respectively. The variance of the value of individuals at locus q is $\sigma_{g_q}^2 = \Sigma_{i_q=1}^{C_q} f_{i_q} g_{i_q}^2$. When several QTLs are involved, the genetic value of an individual (j) of the population (assuming no epistasis) is: $Y^j = \mu + \Sigma_q y_q^j$, where μ is the mean of the population. Under the hypothesis that QTLs are independent (no linkage disequilibrium), the genetic variance of the population is: $\sigma_G^2 = \Sigma_q \sigma_{g_q}^2$ (see Gallais 1974 for a discussion of the effect of linkage disequilibrium and epistasis on variance expression). If the environmental variance associated with the trait of interest is σ_e^2, phenotypic variance is $\sigma_P^2 = \sigma_e^2 + \sigma_G^2$. Trait heritability (broadsense) is $h^2 = \sigma_G^2/\sigma_P^2$.

In general, markers cannot be considered as QTLs themselves. If l is a marker, the effects of the genotypes at locus l are functions of the effect(s) of the QTL(s) and linkage disequilibrium parameters between these QTLs and the marker. This was, for instance, illustrated by Edwards et al. (1987) in the case of F_2 populations. In this situation, the variance associated with l ($\sigma_{g_l}^2$) has to be considered in terms of prediction. It accounts for the fraction of the variance of the trait that can be predicted knowing the genotype of individuals at locus l. $\sigma_{g_l}^2$ will be a function of the effect(s) of the QTL(s) and linkage disequilibrium parameters between these QTLs and the marker. For instance, in the case of two-class experiments (RIL, BC or DH), and F_2 populations in the case of additive effects, $\sigma_{g_l}^2 = \lambda_{ql}^2 \sigma_{g_q}^2$, where $\sigma_{g_q}^2$ is the variance associated with QTL q, λ_{ql} is the parameter defined by Schnell (1961), which in this situation is the linkage disequilibrium between loci q and l ($\lambda_{ql} = 1 - 2c_{ql}$ for an F_2 population, where c_{ql} is the recombination fraction between loci q and l).

We will define the fraction of the genetic variance associated with a given locus (l) as $m_l^2 = \sigma_{g_l}^2/\sigma_G^2$. Similarly, we will define the fraction of the total phenotypic variance associated with locus l as $r_l^2 = \sigma_{g_l}^2/\sigma_P^2 = m_l^2 h^2$. Moreover, we will define the phenotypic variance that is not associated with locus l as: $\sigma_R^2 = \sigma_G^2 - \sigma_{g_l}^2 + \sigma_e^2$, which illustrates that this variance depends on the genetic variance that is not associated with locus l, and on environmental variance.

For a given QTL mapping experiment, individuals are randomly sampled from the population. For a given experiment that involves N individuals or progenies, the numbers of individuals for each genotype at locus l will be quoted n_{i_l} ($\Sigma_{i_l} n_{i_l} = N$). Thus, the observed frequency of genotype i_l is $f_{i_l}^\circ = n_{i_l}/N$. The number of individuals for each genotype (n_{i_l}) will differ from those expected for a balanced experiment (N/C_l) for three reasons: (1) it can be expected a priori in some cases (e.g. for F_2 populations where frequencies are 0.25 for homozygous genotypes, 0.50 for the heterozygous genotype); (2) it can be the result of random sampling of the individuals; (3) it can be due to systematic selection pressure. This last situation is refered to as segregation distortion. The deviations to the frequencies of the reference infinite size population (i.e. 0.5–0.5 for recombinant inbred lines, doubled haploids or backcrosses, 0.25–0.5–0.25 for an F_2) will be noted $d_{i_l} = f_{i_l}^\circ - f_{i_l}$.

Analysis of variance and expected mean squares, estimation of the variance associated with locus l ($\sigma_{g_l}^2$)

Several statistical methods have been proposed to investigate the relationship between genetic markers and the quantitative trait of interest within segregating populations and to infer results about the position and effect of the QTL(s). A first approach is to compare marker genotype means using a normal test or one-way analysis of variance (Soller et al. 1976; Ellis 1986; Edwards et al.

1987). It has been widely accepted that this approach gives no information about the recombination rate between a given marker and linked QTL and that distant linkage cannot be distinguished from a small phenotypic effect. Thus, several approaches have been proposed to estimate the recombination rate between markers and putative QTL and the effect of the QTL. Use of a single marker via maximum likelihood was considered by Weller (1986) and Simpson (1989). The simultaneous use of two genetic markers flanking a putative QTL (interval mapping) proposed by Lander and Botstein (1989) has been widely used by geneticists. Complementary approaches have been proposed recently (e.g. Rodolphe et Lefort 1993; Zeng 1994), which should lead to an increased accuracy in QTL mapping.

For reasons of simplicity, we will consider ANOVA at a given marker locus (l) in order to discuss the estimation of the variance associated with either (1) this marker locus or (2) a neighbouring QTL (q), the position of which has been determined by an interval mapping procedure. It has to be noted that (1) and (2) will tend to be equivalent when the QTL is located at the end of an interval or in very short intervals (very dense maps). The taking into account simultaneously of the information from two marker loci flanking the interval where the QTL is located to estimate the variance associated with this QTL should then deserve a specific investigation. If the statistical effect of a given locus l on the variation of the quantitative trait is analysed through classical ANOVA methods (see Scheffé 1959), the model can be written (e.g. Zeng 1994) as:

$$Y^j = \mu + \sum_{i_l=1}^{C_l} g_{i_l}\theta_{i_l}^j + R_l^j, \tag{1}$$

where Y^j is the phenotypic value of individual (or progeny) j, $\theta_{i_l}^j$ is the dummy variable which describes the observed genotypes of individual j at marker locus l, g_{i_l} is the effect of genotype i_l, defined for the reference population, μ is the mean of the population and R_l^j is the residual term.

In general, as was discussed previously, the locus involved in the analysis of variance is not a QTL itself. This may result in heterogeneous variances within genotypic classes, as noted by Asins and Carbonell (1988), and so affect F tests. However, the aim of this study is to investigate effects of relatively small magnitude, so heterogeneous variances should not be a major problem since environmental variance and unexplained genetic variance will buffer the genotypic variance differences within marker locus genotypes (Asins and Carbonell 1988).

Since $\theta_{i_l}^j$ is a random variable within a reference population, it is clear that the effect of locus l is a random effect. However, one specificity of ANOVA in the case of the association between a quantitative trait and a marker is that the expectation of the mean square associated with locus l effect can be considered in two different ways. This first method (model 1) is to consider the expectation over the experiments that have given numbers of individuals per genotype at the locus of interest (fixed class size). The second method (model 2) is to consider this expectation over all possible experiments, with a total of N individuals, (random class size).

Model 1

Expected mean square can be expressed as a function of the parameters that define the reference population (infinite size, no segregation distortion) at the locus (l) considered for the ANOVA model: g_{i_l}, f_{i_l} and σ_R^2, conditionally to the numbers of individuals per genotype (n_{i_l}) obtained for that experiment. The expectation of the residual mean square is: $E(MSr) = \sigma_R^2$. The expectation of the mean square associated with locus l effect is: $E(MSl|n_{i_l}) = 1/(C_l - 1)\,E(\Sigma_{i_l} n_{i_l}(Y_{i_l\cdot} - Y_{\cdot\cdot})^2)$. Using previous definitions and following a classical approach, this expectation becomes:

$$E(MSl|n_{i_l}) = \frac{1}{C_l - 1}\left(\sum_{i_l}^{C_l} n_{i_l} g_{i_l}^2 - \frac{1}{N}\left(\sum_{i_l}^{C_l} n_{i_l} g_{i_l}\right)^2\right) + \sigma_R^2. \tag{2}$$

Using the deviation parameters (d_{i_l}) defined in the previous section, this expression is equivalent to:

$$E(MSl|n_{i_l}) = \frac{N}{C_l - 1}\sigma_{gl}^2 + \frac{N}{C_l - 1} \times \left[\sum_{i_l} d_{i_l} g_{i_l}^2 - \left(\sum_{i_l}^{C_l} d_{i_l} g_{i_l}\right)^2\right] + \sigma_R^2. \tag{3}$$

When the observed genotype frequencies are equal to the frequencies within the reference population ($d_{i_l} = 0$), this expression reduces to $E(MSl|n_{i_l}) = \tilde{n}\sigma_{gl}^2 + \sigma_R^2$, with $\tilde{n} = N/(C_l - 1)$. The variance associated with locus l can then be estimated as:

$$\hat{\sigma}_{gl}^2 = \frac{MSl - MSR}{\tilde{n}} \tag{4}$$

If l cannot be assumed to be a QTL itself, this approach can be used to estimate the variance associated to a neighbour QTL ($\sigma_{g_q}^2$), provided it is possible to establish that $\sigma_{g_q}^2 = \lambda_{ql}^2 + \sigma_{gl}^2$ (see model section). In this situation:

$$\hat{\sigma}_{g_q}^2 = \frac{1}{\lambda_{ql}^2}\frac{MSl - MSR}{\tilde{n}} \tag{5}$$

In general, when $d_{i_l} \neq 0$, σ_{gl}^2 cannot be estimated directly from Eq. 3. However, solutions exist for several cases of special interest (see below).

Model 2

The second approach is to consider that class size varies at random across the experiments due to the random

sampling of the individuals. We will first assume no systematic distortion segregation. It is then possible to derive the expression of $E(MSl)$ over all the possible experiments (with size N).

Since $E(\Sigma_{i_l} d_{i_l} g_{i_l}^2) = 0$, from Eq. 3 $E(SSl) = N\sigma_{g_l}^2 - NE((\Sigma_{i_l} d_{i_l} g_{i_l})^2) + (C_l - 1)\sigma_R^2$. $(\Sigma_{i_l} d_{i_l} g_{i_l})$ is the deviation of the average value (at locus l) of a sample of N individuals to the mean of the population. Thus, $E((\Sigma_{i_l} d_{i_l} g_{i_l})^2)$ is the variance (over the experiments) of the average. Since individuals are sampled independently, $E((\Sigma_{i_l} d_{i_l} g_{i_l})^2) = 1/N\sigma_{g_l}^2$. This approach takes into account the covariances between class sizes observed across the experiments. For instance, for recombinant inbred lines, the condition $(n_1 + n_2 = N)$ generates a covariance between n_1 and n_2 (see Knott 1994). Thus, $E(SSl) = (N-1)\sigma_{g_l}^2 + (C_l - 1)\sigma_R^2$, and

$$E(MSl) = \frac{N-1}{C_l - 1}\sigma_{g_l}^2 + \sigma_R^2 \qquad (6)$$

Thus, for all populations, the general expression of the coefficient attributed to $\sigma_{g_l}^2$ is: $\tilde{n} = (N-1)/(C_l - 1)$. Estimation of $\sigma_{g_l}^2$ and $\sigma_{g_q}^2$ can then be performed in any situation using Eqs. 4 and 5, respectively.

If we now consider that deviations to the expected frequencies are generated by a systematic selection pressure in addition to random sampling of the individuals (i.e. segregation distortion), the expectation over the experiments is $E(MSl) = (N-1)/(C_l - 1)\,\sigma_{g_l(D)}^2 + \sigma_{R(D)}^2$, where $\sigma_{g_l(D)}^2$ and $\sigma_{R(D)}^2$ are the parameters defined for a reference population that has been submitted to the selection pressure.

Comparison of analyses 1, 2 and literature studies

We will consider first the case of two-class experiments (RIL, BC and DH), considering frequencies 0.5–0.5 in the reference population and genotype numbers n_1 and n_2 for the experiment that is considered. In this situation:

$$\tilde{n} = N - \frac{(n_1 - n_2)^2}{N} = \frac{4n_1 n_2}{N} \quad (model\ 1)$$

$$\tilde{n} = N - \frac{n_1^2 + n_2^2}{N} = \frac{2n_1 n_2}{N} \quad (ls)$$

$$\tilde{n} = N - 1 \quad (model\ 2) \qquad (7)$$

where the coefficient for model 1 is derived from Eq. 2 considering that for the reference population $g_{1_l} = -g_{2_l}$, $f_{1_l} = f_{2_l} = 0.5$, so that $g_{1_l}^2 = g_{2_l}^2 = \sigma_{g_l}^2$. The coefficient for model 2 is obtained from Eq. 6, and (ls) is the coefficient attributed to marker effect (ϕ_l^2) in the study of Knapp and Bridges (1990), which can also be inferred from the study of Hill (1975).

It first has to be noted that the coefficent obtained for model 1 is clearly superior to that of ls. The explanation for the difference between these two coefficients $(2n_1 n_2/N)$ is related to the fact that model 1 takes into account the fact that n_1 and n_2 are not independent $(n_1 + n_2 = N)$. Thus, classical coefficients, which are appropriate for independent class size experiments, are not adapted to the specific case of marker experiments unless the class number gets high. When two genotypes are segregating at a given locus (recombinant inbred lines, backcross ...), the variance which is estimated using appropriate coefficients is two times smaller than the variance estimated using classical coefficients.

With respect to the comparison of models 1 and 2, one has to reach fairly large deviations to the expected frequencies to get a large difference between the $\tilde{n}$ parameters for model 1 and 2. For instance, if one considers 100 RILs, $\tilde{n} = 99$ for model 2. For model 1, $\tilde{n}$ will depend on the observed genotype numbers. For genotype numbers 40 and 60, $\tilde{n} = 96$ for model 1. Under the hypothesis of no systematic segregation distortion, deviations of this magnitude or higher will occur for 5.6% of the experiments. If genotype numbers are 30 and 70, $\tilde{n} = 84$ for model 1. Deviations of this magnitude or higher will occur for only 0.008% of the experiments. Thus, important deviations are very unlikely. They should occur mostly in experiments of limited sizes or because of strong segregation distortion effects.

A second situation of interest is the case of F_2 populations. In this case, the expected frequencies of genotypes 11, 1L and LL are 0.25, 0.5 and 0.25, and the observed numbers of individuals will be defined as n_1, n_2, n_3, respectively. To compare the three possible approaches, we will first consider that the genetic effects are additive, for instance as for testcross progenies (each F_2 individual crossed to a tester). In this situation,

$$\tilde{n} = \frac{N}{2} + N\left[\left(\frac{1}{2} - \frac{n_2}{N}\right) - \left(\frac{n_1}{N} - \frac{n_3}{N}\right)^2\right] \quad (model\ 1)$$

$$\tilde{n} = \frac{N}{2} - \frac{n_1^2 + n_2^2 + n_3^2}{2N} \quad (ls)$$

$$\tilde{n} = \frac{N-1}{2} \quad (model\ 2) \qquad (8)$$

where the coefficient for model 1 is derived from Eq. 2 considering that for the reference population $g_{1_l} = -g_{3_l}$, $g_{2_l} = 0, f_{1_l} = f_{3_l} = 0.25, f_{2_l} = 0.5$ so that $g_{1_l}^2 = g_{3_l}^2 = 2\sigma_{g_l}^2$, $g_{2_l}^2 = 0$. Note that the hypothesis concerning the additivity of genetic effects only affects model 1. Coefficients for ls and model 2 do not depend on this hypothesis. The coefficient attributed to $\sigma_{g_l}^2(\tilde{n})$ in model 1 depends on (1) heterozygote deficiency (or excess) $(1/2 - n_2/N)$ and (2) the difference between homozygote class sizes $((n_1/N - n_3/N)^2)$. It has to be noted that a given heterozygote deficiency (or excess) has a larger effect on the $\tilde{n}$ value than a comparable difference between class sizes of the homozygote genotypes. Large heterozygote excess, which is sometimes observed for outcrossing plants, can be taken into account using model 1.

When no deviations to expected class sizes are observed, the coefficients for models 1 and 2 tend to be equivalent ($\tilde{n} = N/2$) and clearly differ from ls ($\tilde{n} = N/2 \times 5/8$). Thus, as for two-class experiments, the use of appropriate coefficients leads to a substantial modification when compared to classical coefficients.

In the general case, the genetic variance associated with a given QTL in the reference F_2 population is $\sigma^2_{g_l} = a^2/2 + d^2/4$ where a is half the difference between homozygote values and d is the difference between heterozygote value and mid-homozygote value (i.e. dominance effect). If deviations to the observed 0.25–0.50–0.25 ratio are respectively d_{ho1}, d_{he}, d_{ho2}, one can show using Eq. 3 that:

$$E(MSl|n_i) = \frac{N}{C_l - 1}\sigma^2_{g_l} + \frac{N}{C_l - 1}[(d_{ho1} + d_{ho2})a^2 + d_{he}d^2 - ((d_{ho1} - d_{ho2})a + d_{he}d)^2] + \sigma^2_R \quad (9)$$

This illustrates that in cases other than those described previously, expected mean squares cannot be simplified as simple functions of the genetic variance at the QTL ($\sigma^2_{g_l}$), since $\sigma^2_{g_l}$ depends on several genotypic effects. Thus, in these situations, the only possibility is to consider model 2 ($\tilde{n} = N/C_l - 1$). Unless class number gets high, this coefficient is clearly more appropriate than the classical formula for balanced experiments ($\tilde{n} = N/C_l$).

Estimation of the fraction of the phenotypic variance (r^2_l) and genetic variance (m^2_l), and the heritability (h^2_l) associated with locus *l*

The fraction of the phenotypic variation associated with locus l (r^2_l) is classically estimated as: $R^2_l = SSl/SStot$. If we follow the approach of Theil (1971, p 178; also described by Judge et al. 1985, p 862), which leads to the definition of the adjusted R^2 statistics, another possibility is to use the estimators of the variances of interest: $\hat{r}^2_{g_l} = \hat{\sigma}^2_{g_l}/\hat{\sigma}^2_P$. If l cannot be assumed to be a QTL itself, this approach can be used to estimate the fraction of the phenotypic variation associated to a neighbour QTL ($r^2_{g_q}$), provided it is possible to establish that $\sigma^2_{g_q} = \lambda^2_{ql}\sigma^2_{g_l}$ (see model section). In this situation, $\hat{r}^2_q = \hat{\sigma}^2_{g_q}/\hat{\sigma}^2_P$, where $\hat{\sigma}^2_{g_q}$ is derived from Eq. 5. Following Eq. 5 and considering that $\tilde{n} = N - 1/C_l - 1$:

$$\hat{r}^2_l = 1 - \frac{N-1}{N-C_l}(1 - R^2_l) \quad (10)$$

This expression yields $R^2_l = (N - C_l)/(N-1)\hat{r}^2_l + (C_l - 1)/(N-1)$. R^2_l is a biased estimator of r^2_l that leads to overestimated values, whereas $\hat{r}^2_l$ is an unbiased estimator. Differences between $\hat{r}^2_l$ and R^2_l will increase when (1) the experimental size (N) is limited, (2) the number of genotypes (C_l) at locus l is large, and (3) the fraction of the variation associated with locus l is small. If 100 individuals are considered for an F_2 population, this difference will be approximately 2% for small effects. The use of $\hat{r}^2_l$ rather than R^2_l should be particularly recommended to evaluate models which include several marker loci, since in this situation the number of estimated parameters is generally large.

Estimation of (m^2_l) can be derived from Eq. 10 provided an estimator of the heritability ($\hat{h}^2$) of the trait is available (which means that the environmental variance, σ^2_e can be estimated using replicates): $\hat{m}^2_l = \hat{r}^2_l/\hat{h}^2$.

Another parameter of interest is the heritability associated with locus l (h^2_l). h^2_l provides a means to evaluate the accuracy of the prediction of the value of an individual (j) by the estimated effect at locus l. It can be defined as the determination coefficient: $h^2_l = \rho^2(\hat{y}^j_l, y^j_l)$, where y^j_l and $\hat{y}^j_l$ are the actual (defined at the level of the reference population) and estimated effects at locus l, respectively. The variance of the estimated values is: $\sigma^2_{p_l} = \sigma^2_{g_l} + (1/\tilde{n})\sigma^2_R$. Thus,

$$h^2_l = \frac{\sigma^2_{g_l}}{\sigma^2_{p_l}} = \frac{\sigma^2_{g_l}}{\sigma^2_{g_l} + \frac{1}{\tilde{n}}\sigma^2_R} = \frac{r^2_l}{r^2_l + \frac{1}{\tilde{n}}(1 - r^2_l)} \quad (11)$$

h^2_l is less than 1 and will decrease when (1) the experimental size (N) is limited, (2) the number of genotypes (C_l) at locus l is large and (3) the fraction of the variation associated with locus l is small. For an F_2 experiment with 100 individuals, $h^2_l = 0.84$ for $r^2_l = 0.10$; $h^2_l = 0.72$ for $r^2_l = 0.05$.

Power of QTL detection

Theoretical aspects of the use of ANOVA to investigate the relationship between a genetic marker and a quantitative trait were developed by Soller et al. (1976) to investigate the power of F_2 populations to detect QTL effects. This study was later extended to more complex population structures (Soller and Genizi 1978). Soller and Beckmann (1990), Weller et al. (1990) and Knapp and Bridges (1990) considered the effect of progeny replication on the power of QTL detection.

Most of these early studies (e.g. Knapp and Bridges 1990) considered marker effect to be fixed, that is to say implicitly considered model 1 (genotype numbers considered as fixed). In this situation, the power of the test of locus l effect can be estimated following the approach described by Soller et al. (1976) and Knapp and Bridges (1990), $Pr[F_{df_q,df_e,\phi} > F_{\alpha,df_q,df_e,0}]$, where $F_{df_q,df_e,\phi}$ is a random variable from a noncentral F distribution ($\phi \neq 0$) and $F_{\alpha,df_q,df_e,0}$ is the critical value from a central F distribution used to test significance of QTL effect, for a α probability of type I error. Following O'Brien (1986) and Scheffé (1959), the noncentrality parameter is:

$$\phi = (C_l - 1)\tilde{n}\frac{r^2_l}{1 - r^2_l} \quad (12)$$

When investigating *a priori* the efficiency of a given design (size N), one should take into account that class sizes that will be obtained for a given experiment are unknown since deviations will occur due to the sampling of the individuals. In the case of recombinant inbreds, for instance, the exact power of an experiment integrates the power obtained for classes sizes $n1$ and $n2$ over the distribution of these two values (following a binomial law). The exact power can be obtained using model 2 expected mean square (Eq. 6), that is to say consider $\tilde{n} = (N-1)/(C_l - 1)$ in Eq. 12. However, unless very low numbers of individuals are considered (for instance fewer than 30 LR), the exact power of the test is very close to that obtained under the assumption that no deviation will be observed ($\tilde{n} = N/(C_l - 1)$).

In the case of recombinant inbred lines, Fig. 1a illustrates the relationship between the fraction of the phenotypic variation associated with locus l (r_l^2) and the power of the test for various experimental situations and α probability levels. Computations were made using appropriate procedures of SAS (1988). The case of F_2 populations is illustrated by Fig. 1b. It was observed that results obtained for recombinant inbred lines were in excellent agreement with those obtained by Simpson (1989 for comparison of marker genotype means; corrected results of 1992 for likelihood ratio test statistic) by means of simulations, where power was determined as the fraction of the experiments for which the effect of the marker was significant. Conversely, the results differed markedly from those of Knapp and Bridges (1990). For $r_l^2 = 0.20$, the power obtained by these authors with 100 individuals at a $\alpha = 0.01$ risk level is about 0.82 for recombinant inbred lines and 0.61 for an F_2; this can be compared to 0.99 and 0.97, respectively, in the present study. This large underestimation of the power of experiments when using classical expected mean squares formulae is associated to the differences in $\tilde{n}$ values that were reported previously in Eqs. 7 and 8.

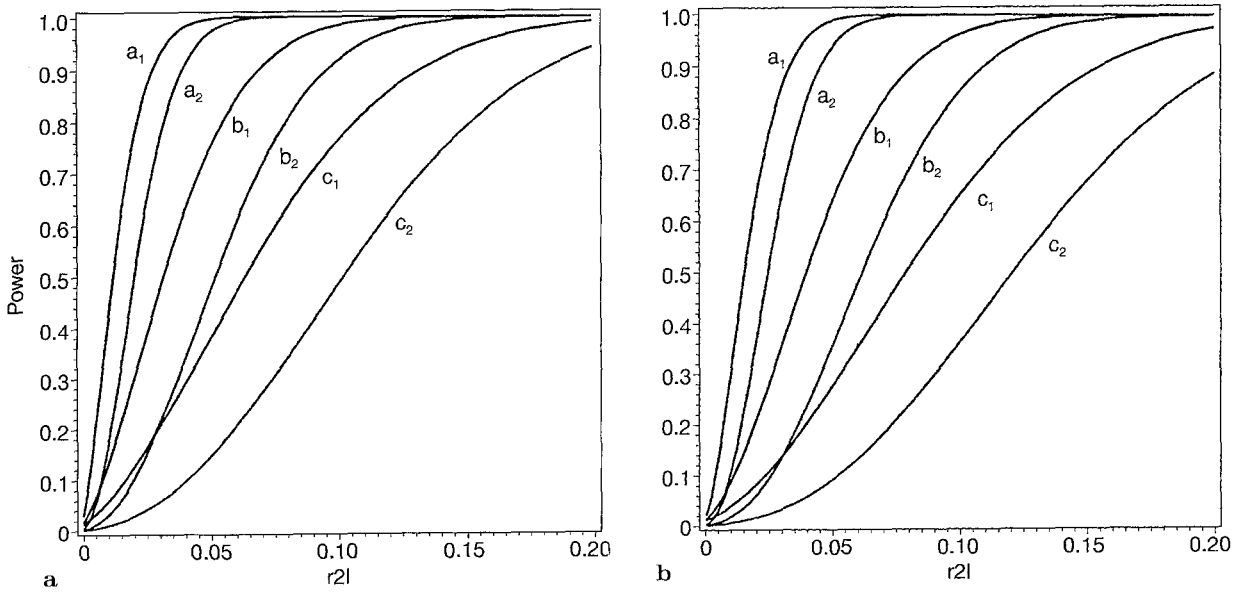
For a given value of r_l^2 and a given number of individuals, the power of QTL testing in F_2 populations is inferior to that in recombinant inbred lines. This is related to differences in degrees of freedom (two independent parameters are estimated for the F_2 against one for recombinant inbred lines), as discussed by Rebai and Goffinet (1993). Thus, the inclusion of the heterozygous class in the analysis or of the dominance effect in the model will lead to a decrease in power unless $d > \frac{a}{2}$ (Soller et al. 1976). More generally, the approach presented here can be undertaken to compare the power of different designs to detect QTLs under various hypotheses.

Confidence intervals for r_l^2 estimates

Once the contribution of a given locus to the variance of the trait has been estimated, it is important to evaluate the accuracy of this estimation. Similarly to power computation, the confidence intervals for r_l^2 estimate ($\hat{r}_l^2$) can be derived from F distributions. For given genotype numbers, the lower limit of the confidence interval ($r^2_{l_{inf}}$), at the α error level is determined by:

$$Pr[F_{df_q, df_e, \phi(r^2_{l_{inf}})} > F_o] = \frac{\alpha}{2} \qquad (13)$$

Fig. 1 Power of experimental designs to underline a marker effect associated with a fraction r_l^2 (r21) of the phenotypic variance of the trait of interest. *a*, *b* and *c* indicate the size of the experiment: 500, 200 and 100 individuals, respectively. *Subscripts 1* and *2* indicate the type-I risk level: $\alpha = 0.01$ and $\alpha = 0.001$, respectively. **a** describes the case of two-class experiments (RIL, BC and DH, see text for designs description); **b** describes the case of three-class experiments (F_2 populations)



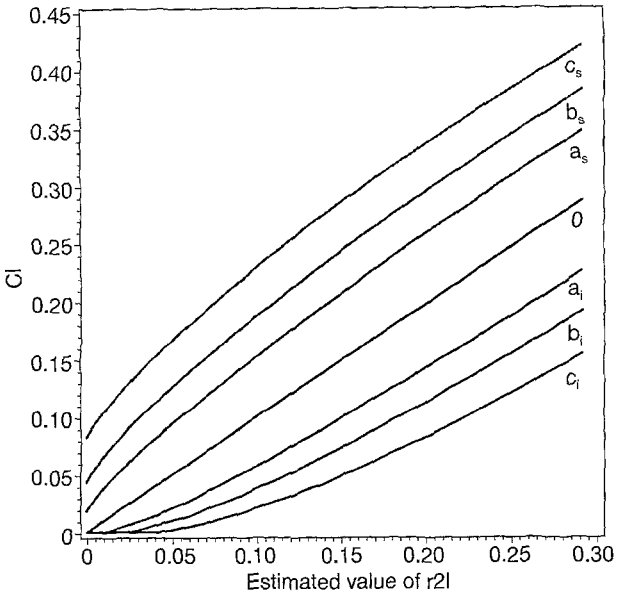


Fig. 2 Confidence intervals at the 95% level (CI) associated with the estimation of r_l^2 (r21). *a*, *b* and *c* indicate the size of the experiment: 500, 200 and 100 individuals, respectively; *i* and *s* indicate the lower and upper limits of the interval, respectively; *O* indicates the $x = y$ line

where F_o is the observed F value ($F_o = 1 + \tilde{n})\hat{r}^2/1 - \hat{r}^2)$), $\phi(r^2_{l_{inf}})$ is the noncentrality parameter corresponding to $r^2_{l_{inf}}$ (using Eq. 12). When no value for r_l^2 fitted condition 13 within the [0; 1] interval, which was the case of small $\hat{r}_l^2$ values, $r^2_{l_{inf}}$ was set at 0. In a similar way, the upper limit of the confidence interval ($r^2_{l_{sup}}$) is determined by:

$$Pr[F_{df_q, df_e, \phi(r^2_{l_{sup}})} < F_o] = \frac{\alpha}{2} \tag{14}$$

Using these formulae, we computed ($r^2_{l_{sup}}$) and ($r^2_{l_{inf}}$) values using appropriate SAS procedures, and these are represented in Fig. 2 in the case of recombinant inbred lines for $\alpha = 0.05$. It was checked that the confidence intervals obtained for F_2 populations were equivalent to those obtained for the RIL. This was consistent with the analyses of Cramer (1987) who reported similar results for the standard deviation of R^2 with two and three classes when the number of observations exceeded 30. This is related to the fact that the number of degrees of freedom which are used to estimate the residual variance are very close for both types of experiments (except if very low numbers of individuals are considered).

It was checked that in the case of two-class experiments with a very large size (so that the bias reported in Eq. 10 can be neglected) the confidence interval provided by this approach for r_l^2 was consistent with the results of Darvarsi et al. (1993) concerning the standard errors for the estimate of the difference between the values of the genotypes at a given locus. In the study of Darvarsi et al. (1993), the standard error for this difference is SE = 0.063 for the 500-individual experiment when considering the infinite number of markers model. A 0.5 difference between genotypes (with a variance within each class set to 1) corresponds to $r_l^2 = 0.058$. The 95% confidence interval for the difference is [0.374; 0.626], which corresponds to a [3.3%; 8.9%] confidence interval for r_l^2. These values are in close agreement with those observed in Fig. 2. In comparison to other methods, the approach which is developed here allows computation of the confidence interval for $\hat{r}_l^2$ for any number of genotypes at the locus of interest. Furthermore, in the case of experiments of limited size, it enables avoidence of the bias reported in Eq. 10.

If *l* cannot be assumed to be a QTL itself, this approach can be used to estimate the confidence interval of the fraction of the phenotypic variation associated to a neighbour QTL ($r^2_{g_q}$) provided it is possible to establish that $\sigma^2_{g_q} = \lambda^2_{ql}\sigma^2_{gl}$ (see model section). In this situation, ($r^2_{q_{sup}} = r^2_{l_{sup}}/\lambda^2_{ql}$) and $r^2_{q_{inf}} = r^2_{l_{inf}}/\lambda^2_{ql}$). It can be checked from Fig. 2 that, for a given estimated value of $r^2_{g_q}$, the size of the confidence interval increases when the distance between the QTL and the marker increases.

Fig. 2 illustrates that when the number of individuals is limited, confidence intervals for r_l^2 can be rather wide (2–23% for a 10% estimate, 8–34% for a 20% estimate if 100 individuals are considered). Precision increases dramatically with the number of individuals evaluated.

Conclusion

Analysis of variance (ANOVA) displays several specificities in the framework of QTL mapping experiments. As a consequence, mean squares expectations associated with locus effects show discrepancies when compared to classical formulae, as was illustrated in Eqs. 7 and 8. These discrepancies increase in importance when the number of genotypes at a given locus is small (typically RIL, BC, DH and F_2 populations). The specific mean squares expectations developed in this study has several consequences.

First, it has been emphasized that ANOVA is a simple and appropriate method by which to compare the power (i.e. the probability to detect a QTL with a given contribution to the variation of the trait) of various experiments, depending on the type of population considered, the size of the experiment etc... This study illustrates that classical formulae for mean squares expectations lead to large underestimations of the power of the experiments when only a few genotypes are segregating at the loci of interest (e.g. recombinant inbred line populations). The power of two-class experiments, RIL, backcross and doubled haploid populations, was reconsidered in this study, as was that of F_2 populations. The same approach can be applied to most populations classically used in plant experiments and used to determine the size of the experiments (*N*).

Secondly, appropriate mean squares expectations allows the variance associated with a given locus, or the fraction of the total variance associated with this locus (r_l^2), to be estimated. This approach has to be compared with two alternative strategies. The first strategy is to use classical mean squares expectations to estimate the variance. For RIL, the variance estimated using this approach is twice the correct estimation. The second strategy is to estimate the variance from the estimated effects (e.g. for a two-class experiment: $\hat{\sigma}^2_{g_l} = 1/4(\hat{g}_{1_l} - \hat{g}_{2_l})^2$), where $\hat{g}_{1_l}$ and $\hat{g}_{2_l}$ are the effects estimated for genotypes 1_l and 2_l, respectively. Even if this strategy does not lead to such large overestimations as the previous one, it leads to biased estimations of the variance, whereas the method proposed in this study (Eq. 10) leads to unbiased estimations. Unbiased estimators increase in importance when (1) the experimental size (N) is limited, (2) the number of genotypes (C_l) at locus l increases and (3) the fraction of the variation associated with locus l is small. Unbiased estimators should also be recommended for models that include several loci.

In addition to the two previous points, appropriate expected mean squares allow evaluation of the precision of the estimation of the fraction of the total variance associated with a given locus (r_l^2). Even if other approaches can be used for two-class experiments, based on the confidence interval of the estimated effects, this possibility is particularly interesting for higher numbers of classes (e.g. F_2 populations). This point is important, since it appears from Fig. 2 that, with the experimental size classically used in plant breeding experiments, confidence intervals for (r_l^2) can be fairly large. In addition to its consequences for the interpretation of the results of QTL mapping experiments, this precision should be taken into account to investigate the efficiency of marker-assisted breeding methods.

It has to be noted that, for sake of simplicity, this study considers analysis of variance at a single locus. If this locus cannot be assumed to be a QTL itself, the fraction of the total phenotypic variation associated to a neighbour QTL and its confidence interval can be estimated, provided it is possible to establish a simple relationship between the variances at both loci. However, the taking into account simultaneously of two markers flanking the QTL of interest should be a preferable protocol to estimate the fraction of the phenotypic variation associated to this QTL and its confidence interval. The development of an unbiased estimator and a confidence interval for (r_l^2) in the framework of interval mapping deserves additional investigation. This should be particularly useful when the QTL of interest lies in the middle of a large interval, since the actual amount of information at these positions is much less than at the end of intervals.

Acknowledgements We are grateful to B. Goffinet and A. Rebai for helpful discussions concerning the statistical aspects developed in this paper. We are grateful to A. Maurice and V. Geffroy for helpful discussions on QTL mapping. We gratefully acknowledge M. Causse's helpful discussions and remarks on the manuscript.

References

Asins MJ, Carbonell EA (1988) Detection of linkage between restriction fragment length polymorphisms and quantitative traits. Theor Appl Genet 76:623–626

Beckmann JS, Soller M (1983) Restriction fragment length polymorphism in varietal identification and genetic improvement: methodologies, mapping and costs. Theor Appl Genet 67:35–43

Burr B, Evola SV, Burr FA, Beckmann JS (1983) The application of restriction fragment length polymorphism to plant breeding. In: Setlow JK, Hollaender A (eds) Genetic engineering principle and methods, vol 5. Plenum Press, London pp 45–59

Cramer JS (1987) Mean and variance of R^2 in small and moderate samples. J Econometrics 35:253–266

Darvarsi A, Weinberg A, Minke V, Weller JI, Soller M (1993) Detecting marker-QTL linkage and estimating QTL gene effect and map location using a saturated genetic map. Genetics 134:943–951

Edwards MD, Stuber CW, Wendel JF (1987) Molecular marker-facilitated investigation of quantitative trait loci in maize. I. Numbers, distribution, and types of gene action. Genetics 116:113–125

Ellis THN (1986) Restriction fragment length polymorphism markers in relation to quantitative characters. Theor Appl Genet 72:1–2

Gallais A (1974) Covariance between arbitrary relatives with linkage and epistasis in the case of linkage disequilibrium. Biometrics 30:429–446

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

Judge GG, Griffiths WE, Hill RC, Lütkepohl H, Lee TS (1985) The theory and practice of econometrics. Wiley, New York

Knapp SJ, Bridges WC (1990) Using molecular markers to estimate quantitative trait locus parameters: power and genetic variances for unreplicated and replicated progeny. Genetics 126:769–777

Knott SA (1994) Prediction of the power of detection of marker-quantitative trait locus linkages using analysis of variance. Theor Appl Genet 89:318–322

Lande R, Thompson R (1990) Efficiency of marker assisted selection in the improvement of quantitative traits. Genetics 121:185–199

Lander ES, Botstein D (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185–199

O'Brien RG (1986) Power analysis for linear models. In: Proc 11th Annu SAS Users Group Conf. SAS, Cary, N.C. pp 915–922

Rebai A, Goffinet B (1993) Power of tests for QTL detection using replicated progenies derived from a diallel cross. Theor Appl Genet 86:1014–1022

Rodolphe F, Lefort M (1993) A multi-marker model for detecting chromosomal segments displaying QTL activity. Genetics 134:1277–1288

SAS (1988) SAS user's guide: statistics, version 6. SAS Institute, Cary, N.C.

Sax K (1923) The association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris*. Genetics 8:552–560

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

Schnell FW (1961) Some general formulations of linkage effects in inbreeding. Genetics 46:947–957

Simpson SP (1989) Detection of linkage between quantitative trait loci and restriction length polymorphisms using inbred lines. Theor Appl Genet 77:815–819

Simpson SP (1992) Correction: detection of linkage between quantitative trait loci and restriction length polymorphism using inbred lines. Theor Appl Genet 85:110–111

Soller M, Beckmann JS (1990) Marker-based mapping of quantitative trait loci using replicated progenies. Theor Appl Genet 80:205–208

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

Soller M, Genizi A, Brody T (1976) On the power of experimental designs for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines. Theor Appl Genet 47:35–59

Theil H (1971) Principles of econometrics. Wiley, N.Y.

Weller JI (1986) Maximum likelihood techniques for the mapping and analysis of quantitative trait loci with the aid of genetic markers. Biometrics 42:627–640

Weller JI, Kashi Y, Soller M (1990) Power of daughter and granddaughter designs for determining linkage between marker loci and quantitative trait loci in dairy cattle. J Dairy Sci 73:2525–2537