Journal of Experimental Botany, Vol. 49, No. 327, pp. 1619–1623, October 1998

Journal of Experimental Botany

The principles of QTL analysis (a minimal mathematics approach)

M.J. Kearsey[1]

Plant Genetics Group, School of Biological Sciences, The University of Birmingham, Birmingham B15 2TT, UK

Received 16 April 1998; Accepted 1 June 1998

Abstract

The combination of molecular marker and trait data to explore the individual genes concerned with quantitative traits, QTL analysis, has become an important tool to allow biologists to dissect the genetics of complex characters. However, the mathematical and statistical techniques involved have deterred many from understanding what the methods achieve and appreciating their strengths and weaknesses.

This paper is designed to give a non-mathematical explanation of the principles underlying these analyses, to discuss their potential and to provide an introduction to the techniques used in the subsequent papers in this series of articles based on the SEB symposium.

Key words: Gene mapping, QTL analysis, quantitative genetics.

Introduction

Although most of the advances in genetics over the last century have been concerned with structural variation in single, so-called 'major genes', much of the natural variation observed in our own species and in the crops, domestic animals and other populations that are studied, are due to much more minor genetic changes in many genes. QTL analysis is the phrase used currently to study this genetic variation, to locate the genes responsible and to explore their effects and interactions.

QTL is the acronym for Quantitative Trait Loci, genes which underlie quantitative traits (Gelderman, 1975). Before the discovery of molecular markers they were known as polygenes (Mather, 1949). Little was known about what these genes were or how they controlled the traits apart from the fact that for any given trait there were several such genes segregating in a Mendelian fashion in any given population and their effects were approximately additive (Kearsey and Pooni, 1996).

It is difficult to define a quantitative trait precisely. The best that can be said is that such a trait appears to show a continuous range of variation in a population, which is more or less normally distributed. There are no obvious discontinuities in the distribution as might be expected of a classical, single gene trait, such as the 1 : 2 : 1 distribution of genotypes and phenotypes in an F_2. Such qualitative genes have a large effect on the phenotype compared to the environment and, dominance apart, genotypes have recognizably different phenotypes. Very often one of the alleles is non-functional or very dysfunctional, which results in the clear phenotype.

However, allelic differences may occur in structural or regulatory genes which alter the genes' action slightly and so produce much smaller phenotypic effects. This type of allelic variation is what is assumed to underlie quantitative variation (Kearsey and Pooni, 1996). Thus, yield in cereals is the result of the combined effects of many genes, from those which control grain and tiller number, through those that affect photosynthesis and metabolism, to genes controlling root development and germination time. It is not difficult to imagine that minor allelic variants exist at many of these loci resulting in a wide range of yields when assembled in different combinations in a population. Although each gene is segregating in a standard Mendelian manner, the overall effect of all the genes is to produce a wide range of phenotypes and this variation is further blurred by differences in the environment giving the Normal range of variation. This does not mean that some of the genes involved might not also be genes whose mutant effects are well known, nor does it mean that some of the larger, qualitative allelic differences are not also present, but concealed in the other variation. It is also possible to obtain a continuous phenotypic distribu-

[1] Fax: +44 121 414 5463/5925. E-mail: m.j.kearsey@bham.ac.uk

tion with very few QTL, given a modest amount of environmental variation.

One of the aims of QTL analysis is to explore the individual QTL and discover more of their action, interaction and precise location. Because they are also very important in agriculture and medicine, there are major practical reasons for knowing more about them.

Background information about quantitative traits

The variation, V_P, among individuals in a population such as an F_2 for any trait can be easily measured. It is caused by genetic and environmental components, so $V_P = V_G + V_E$. It is relatively simple to devise experiments, involving comparisons between parents and offspring or with various types of family, to estimate V_G and V_E (Falconer and Mackay, 1996; Kearsey and Pooni, 1996). For example, if a family of inbred individuals were raised in the experiment, the variation between them would be V_E alone. Hence V_G could be estimated as $V_P – V_E$. The proportion of V_P arising from genetical causes is the heritability of the trait in that population, $h^2 = V_G/V_P$. For most traits of economic or medical interest, heritabilities are characteristically less than 50%, often much less, so most of the phenotypic variation among individuals is environmental.

It is possible to estimate the combined effects of all the QTL to the variation. The genetical variation, V_G, is a function of Σa^2, where a is half the difference in phenotype between alternative homozygotes of a QTL. For example, if there are two allelic forms of a QTL in a population, Q^+ and Q^-, with Q^+ causing a greater phenotype, then $a = (Q^+Q^+ – Q^-Q^-)/2$. If one was to perform divergent selection for the trait over several generations, one would produce high lines with Q^+ alleles at each QTL and low lines with all the Q^- alleles. The difference between these lines would be $2\Sigma a$. So the combined effects of all QTL (Σa^2 and Σa) can be estimated, but not their individual effects. The combination of trait data and molecular markers enables these individual effects to be identified and forms the next step of QTL analysis.

Principles

The basic problem with studying a quantitative trait has always been that the phenotype of a given genotype tells us little about the genotype itself; two plants could be 1 m high but have very different genotypes. The discovery of extensive and easily recognizable molecular variation has opened up the possibility of studying individual QTL (Lander and Botstein, 1989).

The principle is simple. Molecular markers give unambiguous, single site genetic differences that can easily be scored and mapped in most segregating populations. It is not difficult in populations of most species to identify and map 10 to 50 segregating markers per chromosome. Most will be in non-coding regions and will not affect any trait directly, but some at least will be linked to QTL which do affect the trait of interest. QTL analysis depends on the fact that where such linkage occurs, the marker locus and the QTL will not segregate independently and so differences in those marker genotypes will be associated with different trait phenotypes. Situations where genes fail to segregate independently are said to display ‘linkage disequilibrium’.

Almost any type of population is amenable to such analysis. The most useful are those derived from a cross between two inbred lines (F_2, backross, recombinant inbred lines, doubled haploid lines) because the marker-QTL linkages in the F_1 cause the derived populations to be in linkage disequilibrium. With natural populations, such as man, farm animals and tree species, the linkage associations have to be explored within families. This is because a consistent association between QTL and marker genotype will not exist across the species except in the unlikely situation that a given marker is completely linked to the QTL. This account will concentrate on the analysis of populations derived from an F_1; the other situations are analogous in principle but statistically more complex.

If the F_1 is heterozygous for a large number of molecular markers, M_1, M_2, ... M_i, etc. these and the trait can be scored in individuals of the F_2 or other derived population, and the markers mapped. For any particular marker locus, M_i, the average trait score of each of the marker genotypes can be calculated, i.e. of $M_{i1}M_{i1}$, $M_{i1}M_{i2}$ and $M_{i2}M_{i2}$, (Table 1). If this marker is on a different chromosome to any QTL, then the QTL alleles will be segregating completely independently of the marker, i.e. $M_{i1}M_{i1}$ will occur with all possible QTL genotypes. So will the other two marker genotypes and hence the average phenotype of all three will be the same and intermediate, i.e. about the population mean (Table 1a). If the marker locus is on the same chromosome and close to a QTL, then $M_{i1}M_{i1}$ homozygotes will mostly be Q^+Q^+ whilst the $M_{i2}M_{i2}$ will mostly be Q^-Q^-, and so they will differ in phenotype (Table 1b). The size of this difference will depend on the effect of that QTL, a, and how close the marker is to the QTL; the closer they are, the greater the difference (Fig. 1). The difference between the genotypes will be maximal and equal to $2a$ if the marker and QTL are so close that they do not recombine, i.e. M_i and the QTL cosegregate.

Figure 2 shows some actual data to illustrate how the difference between the marker means varies with proximity to a QTL. The actual marker differences are subject to error variation of course, but one can see by eye that they peak at 37 cM, so suggesting a single QTL, and that the effect of the QTL, a, is about 3 d. It is important to note that all the markers along the chromosome show some effect of the QTL, even those well separated from

Table 1. *Relationship between marker genotype and mean trait value (a) marker and QTL on different chromosomes (unlinked), (b) marker and QTL on same chromosome (linked)*

(a)

$$F_1 \qquad \frac{M_1}{M_2} \qquad \frac{Q^+}{Q^-}$$

Frequencies of genotypes in F_2

Marker genotype (x)	Q^+Q^+	Q^+Q^-	Q^-Q^-	Mean trait score (y)
M_1M_1	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	Intermediate
M_1M_2	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	Intermediate
M_2M_2	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	Intermediate

Difference between trait scores of M_1M_1 and M_2M_2 zero.

Conclusion: No relationship between trait score (y) and marker genotype (x).

(a)

$$F_1 \qquad \frac{M_1 \quad Q^+}{M_2 \quad Q^-}$$

Frequencies of genotypes in F_2

Marker genotype (x)	Q^+Q^+	Q^+Q^-	Q^-Q^-	Mean trait score (y)
M_1M_1	Most	Few	Rare	High
M_1M_2	Few	Most	Few	Intermediate
M_2M_2	Rare	Few	Most	Low

Difference between trait scores of M_1M_1 and M_2M_2 large.

Conclusion: Strong relationship between trait score (y) and marker genotype (x).

it. This is because there are few crossovers on a chromosome and a high proportion of chromosomes emerge from meiosis without being involved in any crossover. Hence a single QTL will always show some association with all linked markers.

In the example in Fig. 2, the estimated additive effect of $a = 3$ d can be compared with the estimate of Σa^2 obtained from this cross (43.3), $3^2/43.3$, which shows that 21% of the genetic variation has been explained by this one QTL. Similar analyses of other chromosomes will identify some of the remaining QTL.

Although the principle of QTL analysis has been illustrated here without resort to statistics, in practice, one needs statistical methods to identify the most likely QTL positions and effects, to test their significance and to indicate their reliability. This subject has exercised the minds of many statisticians and a variety of approaches designed to increase precision and cope with more complex data sets have been devised and statistical software developed.

The simplest approach is to use a '*t*-test' or 'ANOVA' to test if the differences between the marker means are significant for the trait. This does not locate the QTL, but simply confirms that the 'eyeballed' location indicates a real effect. An analogous approach is to regress the trait value (y) onto the marker genotype (x) (Table 1). If the marker and QTL are on different chromosomes there will be no regression, while if they are on the same chromosome, the regression will be maximal at those markers closest to the QTL, mirroring the shape of Fig. 1. Another approach is to attempt to use regression analysis to fit a line to the combined marker data for a given chromosome as has been done in Fig. 2. This will locate the most likely QTL position, estimate its effect, *a*, test that it is significant and test if the single QTL adequately explains the data for that chromosome (Kearsey and Hyne, 1994).

The most commonly used analytical approaches explore the interval between pairs of markers for the presence of QTL. Hence they are known as 'interval mapping' techniques (Lander and Botstein, 1989). They essentially look at the trait information from each adjacent pair of marker loci and use this to infer the likelihood of a QTL being at any given position between them. For example, in Fig. 2, the markers at positions 37 and 55 cM have the highest trait scores, so the QTL is likely to be between them. Had both marker values been the same, then the most likely QTL position would be midway between them, whilst if the QTL was closer to the left hand marker, then that marker would have the higher score, as indeed it does. By calculating the likelihood of a QTL across all intervals (compared to getting that result by chance alone) gives a likelihood ratio (or LOD) profile. A similar result can be achieved by regression (Haley and Knott, 1992), where the profile of probability associated with the *F* test for the regression (*pF*) is used instead of LOD. Where the LOD or *pF* exceed some significance threshold indicates the likely location of the QTL and provides information on its confidence interval (Churchill and Doerge, 1994; Mangin *et al.*, 1994).

These techniques can be elaborated in various ways to improve their precision and reliability (Jansen, 1993; Jansen and Stam, 1994). Thus every time a QTL is identified, its effect can be removed from the error, so increasing the precision of future tests. If all QTL are identified, only V_E should remain. Parameters can be built into the models to allow for environmental effects such as sites and years, or sex effects in animals.

QTL analysis of humans and other outbreeding populations involves the additional problem that each individual family has to be handled separately and the data combined. A particular pair of parents may represent the equivalent to the parents of an F_2 or backcross with respect to a particular marker locus and a QTL and hence the methods discussed above can be applied to that family. However, the family is very small and so marker QTL data from many families have to be combined. Moreover, the linkage phase in the parents, i.e. whether the marker gene and QTL are in coupling or repulsion, is not known and will vary from family to family and

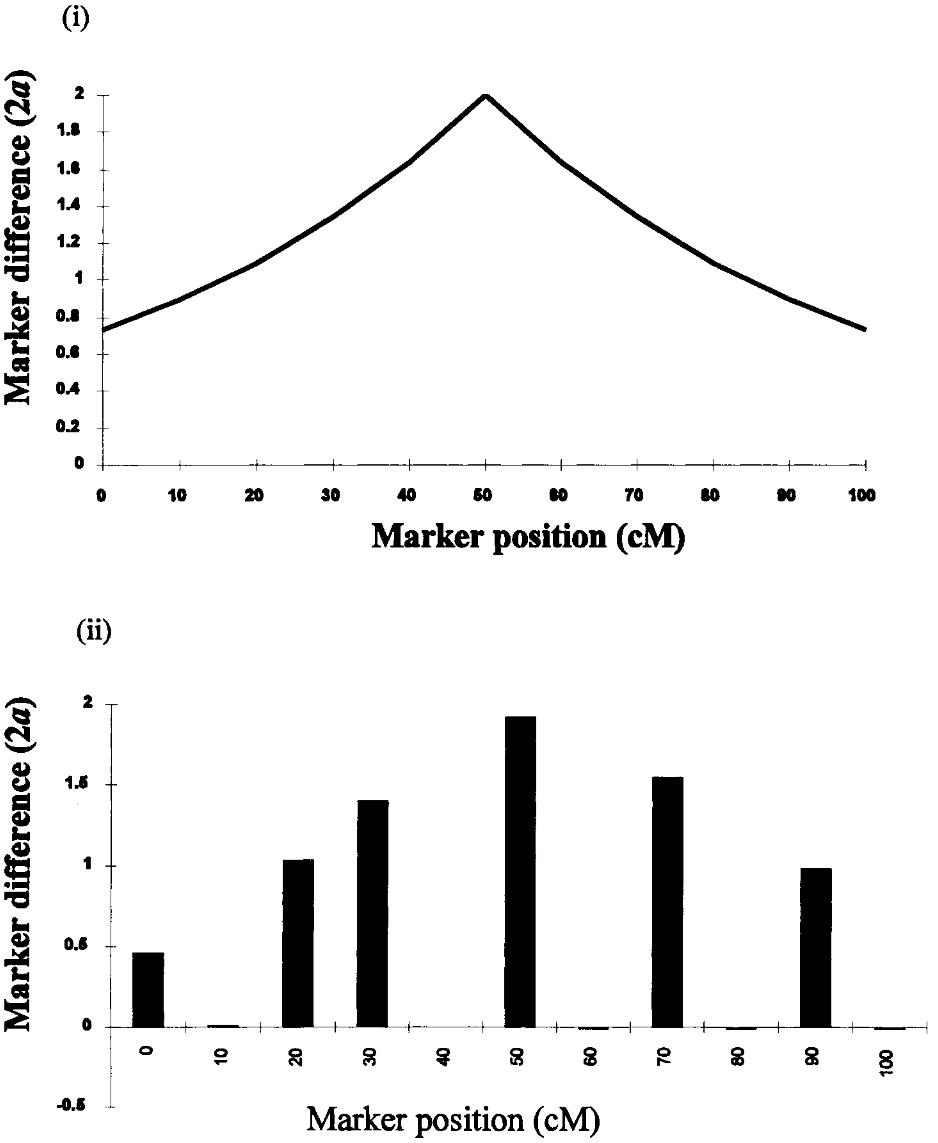

Fig. 1. Relationship between difference between means of marker homozygotes ($M_{i1}M_{i1}$–$M_{i2}M_{i2}$) and position on chromosome with a QTL of $a = 1$ unit at 50 cM. (i) Expected values. (ii) Possible observed values with six markers unevenly spaced along the chromosome.

will have to be deduced from the data. The approaches are therefore similar, but more elaborate and less precise.

The major problem associated with all QTL analyses is that the individual QTL effects are small. As stated above, heritabilities for most traits are generally less than 50%, so that the heritability associated with individual QTL is a small fraction of this. The more QTL there are in the population, the smaller their individual contribution and the more difficult they are to detect. It thus follows that only the larger QTL are ever detected and this leads to the biased impression that there are few QTL and they have large effects. These low individual QTL heritabilities also cause the estimates of QTL location to have large confidence intervals (Hyne *et al.*, 1995). Thus, although the analysis of a particular set of individuals in a population suggests that the QTL is at x cM on a given chromosome, its true position may well be anywhere within a range ±10–20 cM from this, i.e. over quite a large region of the chromosome. Unless the QTL effect is large and the environmental variation is greatly reduced by replication, it is difficult to reduce the confidence interval to less than about 10 cM. Such accuracy is not very helpful for positional cloning, but may be perfectly adequate for Marker Assisted Selection, MAS. It is a popular belief that ever denser marker maps help to resolve this problem, but beyond a density of about one marker per 10 cM, there is very little gain. By far the most important factor is the number of genotypes tested, but there are practical limits to this, particularly in an agricultural context.

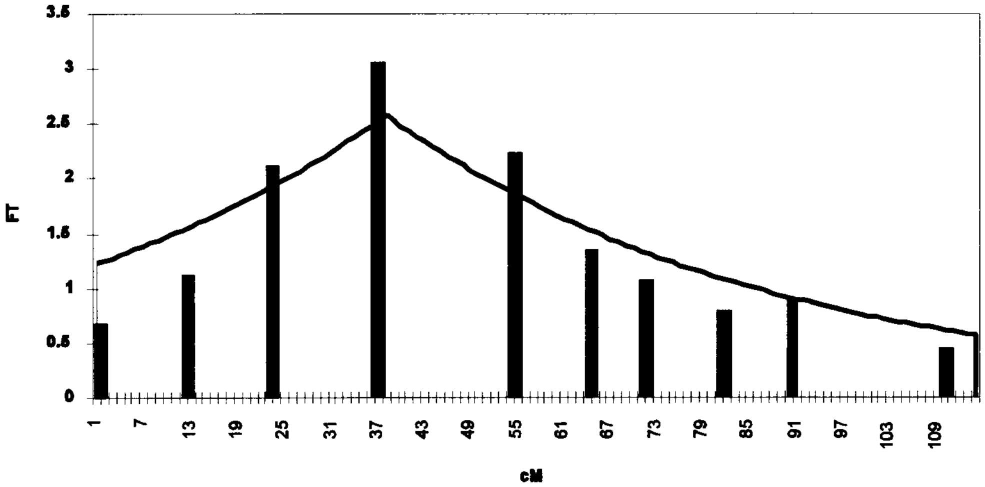

Fig. 2. Observed data (cf. Fig. 1) for flowering time for chromosome 9 of *Brassica oleracea*, suggesting the position of a single QTL, a = 3 d, at 37 cM. Best fitting line is also shown.

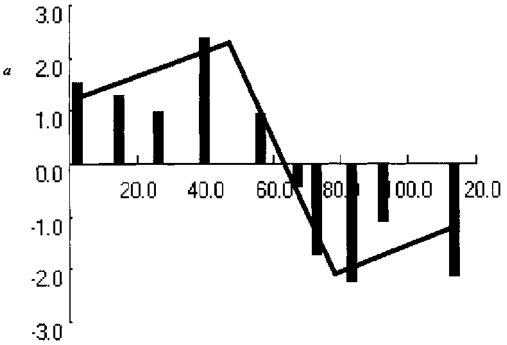

Fig. 3. Observed data (as Fig. 2) but illustrating two QTL in dispersion; one at 46 cM with a positive effect of 5 d and a second at 78 cM with a decreasing effect of 5 d. The solid line shows the expected marker means. The close proximity (32 cM) of the QTL in dispersion reduces the observed marker values because they are cancelling each others effect, i.e. the QTL of 5 d at 46 cM appears as a peak of just 2 d.

The relatively large size of the confidence intervals makes it more difficult to distinguish two QTL on a chromosome unless they are far apart. Figure 3 illustrates the effects on marker means of two QTL in dispersion. There is a characteristic pattern of changes in marker means from positive to negative along the chromosome, but one QTL tends to reduce the effects of the other so making the detection of either more difficult. Had the two QTL been in association, their individual effects could combine to give the appearance of a false, ghost QTL somewhere between them.

Despite these problems, QTL analyses of segregating populations have been very useful in identifying major QTL which may suggest candidate loci (Osborn *et al.*, 1997; Lagercrantz *et al.*, 1996) and permit accelerated MAS. For further reviews of QTL analysis, the reader is referred to the following: Tanksley (1993) and Kearsey and Farquhar (1998).

References

Churchill GA, Doerge RW. 1994. Empirical threshold values for quantitative trait mapping. *Genetics* **138,** 963–71.

Falconer DS, Mackay TFC. 1996. *Introduction to quantitative genetics,* 4th edn. London: Longman.

Gelderman H. 1975. Investigation on inheritance of quantitative characters in animals by gene markers. I. Methods. *Theoretical and Applied Genetics* **46,** 300–19.

Haley CS, Knott SA. 1992. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* **69,** 315–24.

Hyne V, Kearsey MJ, Pike DJ, Snape JW. 1995. QTL analysis: unreliability and bias in estimation procedures. *Molecular Breeding* **1,** 273–82.

Jansen RC. 1993. Interval mapping of multiple quantitative trait loci. *Genetics* **135,** 205–11.

Jansen RC, Stam P. 1994. High resolution of quantitative traits into multiple loci via interval mapping. *Genetics* **136,** 1447–55.

Kearsey MJ, Hyne V. 1994. QTL analysis: a simple 'marker regression' approach. *Theoretical and Applied Genetics* **89,** 698–702.

Kearsey MJ, Pooni HS. 1996. *The genetical analysis of quantitative traits.* London: Chapman and Hall.

Kearsey MJ, Farquhar AGL. 1998. QTL analysis in plants: where are we now? *Heredity* **80,** 137–42.

Lagercrantz U, Putterill J, Coupland G, Lydiate D. 1996. Comparative mapping in *Arabidopsis* and *Brassica*, fine scale collinearity and congruence of genes controlling flowering time. *The Plant Journal* **9,** 13–20.

Lander ES, Botstein D. 1989. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121,** 185–99.

Mangin B, Goffinet B, Rebai A. 1994. Constructing confidence intervals for QTL location. *Genetics* **138,** 1301–8.

Mather K. 1949. *Biometrical genetics,* 1st edn. London: Methuen.

Osborn TC, Kole C, Parkin IAP, Sharpe AG, Kuiper M, Lydiate DJ, Trick M. 1997. Comparison of flowering time genes in *Brassica rapa*, *B. napus* and *Arabidopsis thaliana. Genetics* **156,** 1123–9.

Tanksley SD. 1993. Mapping polygenes. *Annual Review of Genetics* **27,** 205–33.