# Mapping of quantitative trait loci by using genetic markers: an overview of biometrical models used

*Ritsert C. Jansen, Centre for Plant Breeding and Reproduction Research (CPRO-DLO), P.O. Box 16, 6700 AA Wageningen, The Netherlands*

## Introduction

In crop plants quantitative variation is a feature of many important traits, such as yield, quality or disease resistance. Means of analyzing quantitative variation and especially of uncovering its potential genetic basis are therefore of prime importance for breeding purposes. It has been demonstrated in the early $20^{th}$ century that such quantitative variation results from the combined action of multiple segregating genes and environmental factors (Johannsen 1909). An intrinsic feature of such traits is, however, that the individual genes contributing to quantitative variation can hardly be distinguished. The genetics of such complex traits is therefore studied in general terms (population means and variances, covariances between progenies, heritabilities and so on) of classical quantitative genetics (Mather & Jinks 1971), rather than in terms of individual gene effects. Only by the use of genetically marked chromosomes, is it possible to detect and locate the loci affecting quantitative traits ("quantitative trait loci" or "QTLs"). Linkage between QTLs and morphological markers (Sax 1923; Rasmusson 1933; Thoday 1961) has been reported, but accurate and systematic genetic mapping has been hampered by the lack of a sufficient number of genetic markers covering an entire genome. Recently, new tools have become available by the advent of molecular markers, such as restriction fragment length polymorphisms (RFLPs) (Botstein *et al.* 1980, Beckmann & Soller 1983). Now, dense genetic linkage maps exist for many plant and animal species, which heralds a new era for quantitative genetics (Tanksley *et al.* 1989).

Powerful and accurate biometrical methods are needed, so as to make possible the dissection of quantitative variation of complex characters into individual QTL effects. Mapped QTLs can be traced in breeding programmes, for instance, indirectly by selection for linked markers, or they can be cloned and introgressed via molecular or cell-biological techniques. The traditional methods for mapping of QTLs are, however, neither powerful nor accurate and the development of better methods is an area open to research. Not surprisingly, the detection and mapping of QTLs is gaining rapidly growing attention from biometrical geneticists.

### Biometrical models

Here, we give a short overview of the advancements in biometrical modelling of the QTL mapping problem. The models will be briefly described for backcross progenies, but the same ideas also apply to other types of progeny, in which linkage association between markers and QTLs is manifest.

*Studying single markers one by one*

The traditional approach to detecting and mapping QTLs involves studying single markers one by one (Sax 1923, Soller & Brody 1976). Allele substitution effects at a marker locus indicate the presence of one or more linked QTLs. In the case of a backcross progeny, the expected difference between the two marker classes, say Mm and mm, is:

$$\mu_{Mm} - \mu_{mm} = \sum a_i(1 - 2r_i) \ , \tag{1}$$

where the summation is over QTLs, $r_i$ is the recombination frequency between the marker and the $i^{th}$ QTL, and $a_i$ is the allele substitution effect of the $i^{th}$ QTL. The realized value of $1-2r_i$ is likely to be close to 0 for unlinked QTLs (unless the progeny size is small), and the effect of those QTLs is negligible. The $F$-test in analysis of variance is commonly used to test for the allele substitution effect at the marker locus. It is assumed that $Y=\mu_{Mm}+E$ for individuals in marker class Mm, and $Y=\mu_{mm}+E$ for individuals in marker class mm, where $Y$ is the value of the phenotypic trait and $E$ is a random normally distributed error. In short regression notation:

$$Y = \mu_{mm} + x(\mu_{Mm} - \mu_{mm}) + E \ , \tag{2}$$

where the indicator variable $x$ takes the value 0 and 1 for the genotypes mm and Mm, respectively, and $\mu_{Mm}-\mu_{mm}$ is the allele substitution effect.

This marker-one-by-one approach has a number of shortcomings. In the case of a single segregating QTL, (a) tight linkage to a single QTL with a small effect cannot be distinguished from loose linkage to a single QTL with a large effect; (b) the position of a single QTL relative to the marker is not defined accurately. In the case of multiple QTLs, (c) the method is not powerful since QTLs are mapped one a time, ignoring the effects of other mapped QTLs; (d) the method cannot separate linked QTLs; (e) effects of QTLs with opposite sign effects cancel so that the test for the allele substitution effect at a marker locus is not even a proper test for QTL activity; (f) the presence of QTLs with effects of equal sign can lead to the false detection of a single "ghost-QTL" at an intermediate marker; Finally, (g) the error distribution is actually a mixture of (normal) distributions (due to recombinations between the marker and QTLs; see below).

*Mixture models for a single QTL with one or two flanking markers*

Weller (1986) emphasized that the trait should be considered to follow a mixture of (normal) distributions and he developed mixture models for estimating the linkage between a single marker and a single QTL. Suppose that $F_1$ individuals with genotype MQ/mq are backcrossed to the parent with genotype mq/mq. For individuals in marker class Mm the model is $Y=\mu_{Qq}+E$ when no recombination between the marker and the QTL has occurred (chance $1-r$), and $Y=\mu_{qq}+E$ otherwise (chance $r$). Similarly, for individuals in marker class mm, the model is $Y=\mu_{qq}+E$ when no recombination between the marker and the QTL has occurred (chance $1-r$) and $Y=\mu_{Qq}+E$ otherwise (chance $r$). In short regression notation:

$$Y = \mu_{qq} + X(\mu_{Qq} - \mu_{qq}) + E \; , \qquad (3)$$

where $\mu_{Qq}-\mu_{qq}$ is the allele substitution effect at the QTL and $X$ is a random indicator variable which takes values 0 and 1 for the genotypes qq and Qq, respectively, with probabilities $r$ or $1-r$ depending on the marker genotype. If the phenotypic values are not affected by a QTL, then $Y=\mu+E$, *i.e.*, $\mu_{Qq}=\mu_{qq}=\mu$. The test for the presence of a putative QTL is commonly based on a comparison of the likelihood of the model with the QTL and that of the model without the QTL (the likelihood-ratio test).

Weller's approach has been generalized so as to make possible the analysis of single QTLs enclosed by a pair of flanking markers (Simpson 1989, Lander & Botstein 1989, Jensen 1989, Knapp *et al.* 1990). This flanking marker procedure has been termed "interval mapping". The regression model (3) is still used, but the distribution of $X$ now depends on the two flanking markers. Expressions for the (conditional) probabilities of the various genotypes can be derived straightforwardly.

The interval mapping method has several advantages over the traditional approach. In the case of a single segregating QTL, (a) the location and the effect of the QTL can be assessed more accurately; (b) the likelihood for the presence of a putative QTL can be plotted along the genetic map, so as to present the evidence for QTLs at the various positions of the genome; (c) the test for the presence of a QTL is more powerful. The principal shortcoming of interval mapping is that still only models for a single QTL are used, which is in clear contradiction with the commonly assumed oligogenic or polygenic nature of quantitative traits. Therefore, interval mapping has a number of shortcomings when two or more QTLs are segregating; see the points (c)−(f) listed in the previous section. This has motivated theoretical research for multiple QTL mapping methods.

*Standard multiple regression of the trait on the markers*

The simple method based on regression of phenotype on markers one by one has been generalized to multiple regression methods in which the trait can be regressed on a large number of markers (Cowen 1989, Stam 1991, Rodolphe & Lefort 1993, Jansen 1993, Zeng 1993, Jansen & Stam 1994). If the marker map sufficiently covers the whole genome, the major part of the QTL induced variation will be absorbed by marker cofactors. The regression model reads:

$$Y = \mu + \sum x_i a_i + E \,, \qquad (4)$$

where the summation is over marker loci, and $x_i$ and $a_i$ are the indicator variable and the allele substitution effect for the $i^{\text{th}}$ marker, respectively. Individuals with any missing marker observation might be eliminated from the regression, but in regression of the trait on many markers only a very limited set of data would then remain. Jansen & Stam (1994) developed the exact model, *i.e.*, a mixture model, in which the indicator variable $x_i$ is replaced by a random indicator variable $X_i$, the probability distribution of which is based on the observations at the linked marker loci (see below). Rodolphe & Lefort (1993) replaced the indicator variable $x_i$ by the expectation of $X_i$ given the observations at linked marker loci.

The multiple regression approach has several clear advantages: (a) the background "noise" is reduced (but not minimized) by taking into account the effects of QTLs by nearby markers; (b) by starting with a 'polygenic' model (regression on all markers) it gets around detection and mapping problems with interfering QTLs; (c) in regression on all markers, the test for QTL activity in a certain region is generally unaffected by QTLs that are located in other regions; (d) standard procedures for selection of important variables in regression can be used, so as to identify the "important" markers, hopefully those flanking the QTLs. Compared to interval mapping, the multiple regression approach has the disadvantage that (a) no precise information for the QTL location or the QTL effect is obtained and (b) no QTL likelihood plots are produced. Further, (c) in regression on all markers, the test for QTL activity is not powerful due to genetic correlation between the QTL and markers outside the region under study; (d) the overall significance level in QTL detection is unclear when standard selection methods are used.

*Multiple regression models based on the expected values of the marker class means*
Several authors (Knapp *et al.* 1990, Knapp 1991, Haley & Knott 1992, Martinez & Curnow 1992, Moreno-Gonzalez 1992) have developed similar approximate interval mapping methods, which could be generalized so as to map several QTLs simultaneously. These models are based on the expected phenotypic values of the marker classes, which are non-linear functions of QTL effects and recombination frequencies. The interval mapping model given by expression (3) is approximated by the model:

$$Y = \mu_{qq} + \mathscr{E}_M(X)(\mu_{Qq} - \mu_{qq}) + E \,, \tag{5}$$

*i.e.*, $X$ in expression (3) is replaced by its expectation $\mathscr{E}_M(X)$, given the observed genotype at the flanking marker loci. For multiple QTLs the regression model reads:

$$Y = \mu + \sum \mathscr{E}_M(X_i)a_i + E \,, \tag{6}$$

where the summation is over putative QTLs; the variables $X_i$ are the indicator variables for the QTLs, and the $a_i$ are the allele substitution effects of the QTLs. Knapp *et al.* (1990) and Knapp (1991) ignore double and multiple crossovers to simplify the model. They estimate the recombination parameters in the non-linear models by direct means. Like in the interval mapping method, Haley & Knott (1992) and Martinez & Curnow (1992) move the QTL along the chromosome, and at each map location the likelihood for the presence of a putative QTL is plotted. At a given map location the recombination frequencies are known (and with that $\mathscr{E}_M(X)$), so that expression (5) is a standard regression model with unknown parameters $\mu_{Qq}$ and $\mu_{qq}$. This approach can be generalized to a two-dimensional search for two QTLs (by moving independently two QTLs along the chromosomes) or to a multidimensional search for multiple QTLs (by moving independently multiple QTLs along the chromosomes). To simplify the models, Moreno-Gonzalez (1992) ignores double crossovers between flanking markers and locates putative QTLs at a fixed position, namely halfway between their flanking markers. This makes it possible to regress the trait on many QTLs in a way similar to standard multiple regression of the trait on markers (in which case putative QTLs are "located at marker positions"). The models of Moreno-Gonzalez are, however, much more complex.

The advantages of these methods compared to interval mapping are: (a) the effects of linked QTLs can be unravelled more efficiently and more accurately; (b) when two QTLs are simultaneously searched for, the simultaneous likelihood for the presence of these QTLs can still be plotted in a three-dimensional graph; (c) the computer programme is easy and fast. There are, however, several disadvantages: (a) the complexity of the models increases with the number of putative QTLs in the model; (b) the computation involved with all these models is almost unfeasible when the number of QTLs is larger than two or three; (c) two or three putative QTLs can be moved simultaneously along the chromosomes but other (mapped or not yet mapped) QTLs will be ignored; (d) the random variable $X$ for the QTL in the mixture model is replaced by its expected value, but this approximation is not efficient in the case of major QTLs or QTLs located in the middle of wide marker intervals.

*Mixture models and approximate mixture models for multiple QTLs*

Jansen (1992) developed exact models for multiple QTLs. We number the loci (markers and putative QTLs) according to their map order; $X_i$ is the indicator variable for the $i^{\text{th}}$ locus. The regression model reads:

$$Y = \mu + \sum X_i a_i + E , \qquad (7)$$

where the summation is over putative QTLs. Jansen (1992) demonstrated how the simultaneous likelihood of the trait ($Y$), the QTLs ($X_i$) and their flanking markers ($X_{i-1}$ and $X_{i+1}$) can be maximized; in fact it was demonstrated that the mixture model can easily be embedded in the framework of multiple linear regression models and even in that of generalized linear models. The problem can be considered as a multiple regression problem with missing genetic data. The core of the method is to augment and complete the data: in case of a single QTL all data are replicated twice; the first replication is completed with the QTL genotype qq, the other replication with Qq, and corresponding weights (conditional probabilities) can be calculated. Parameter estimation is carried out by iterative weighted regression of the augmented data on the QTLs, alternating updating of the weights and updating of the parameter estimates. If many QTLs are assumed, the number of possible genotypes becomes so large that computation is no longer feasible. Disregarding genotypes with negligible weights can be a solution, without substantial loss of information.

Jansen (1992) described a "hybrid" method, combining interval mapping with standard multiple regression methods (see also Jansen (1993) and Zeng (1994)). The regression model reads:

$$Y = \mu_{qq} + X(\mu_{Qq} - \mu_{qq}) + \sum X_i a_i + E \;, \tag{8}$$

where $X$ is the random indicator variable for the single QTL, and the summation is over markers used as cofactors. Jansen & Stam (1994) developed a very general method of multiple linear regression of a quantitative trait on genotype (QTLs and markers). This regression model is the same as that in expression (7), but now the summation is over loci in general, *i.e.*, over QTLs and over those markers used as cofactors. Here, the method will be termed "MQM mapping", where MQM is an acronym for "multiple-QTL models" as well as for "marker-QTL-marker", which reflects the insertion of QTLs between markers on the genetic map. The basic idea is the completion of any missing genotypic (QTL or marker) data by augmenting and weighting the data. Marker observations can be fortuitously missing, but also other types of missing marker data occur in a natural way. For instance in an $F_2$, when markers are dominant and the heterozygote cannot be distinguished from one of the homozygotes. Or in outbred progeny, when markers with different information are located in mixed order on the chromosomes (only one of the gametes gives information on recombination if a marker segregates according to backcross rules, whereas both gametes are informative if a marker segregates according to $F_2$ rules). Jansen (1994) studied the chance of type I or type II errors in MQM mapping.

Advantages of the models for MQM mapping are: (a) the full power of complete linkage maps is exploited as much as it is computationally feasible, to complete any missing genetic (QTL and marker) data; (b) the likelihood for the presence of a putative QTL can be plotted along the genome when marker cofactors are used; (c) Models, which are exact for major QTLs and approximate for minor QTLs, can be fitted.

## Concluding remarks

We have sketched the recent developments of QTL mapping methods from the traditional marker-one-by-one approach, via the "single QTL" interval mapping approach to more advanced methods based on exact or approximate models for multiple QTLs. Presently the traditional marker-one-by-one approach and the interval mapping method are still widely used (*cf.* Paterson *et al.* 1991, Stuber *et al.* 1992, De Vicente & Tanksley 1993). But it is now generally recognized that simultaneous mapping of multiple QTLs is more efficient and more accurate. Therefore, the methods based on simultaneous mapping of multiple QTLs should provide the method of choice for the analysis of QTL mapping data. These methods date, however, from the past two years and their properties are still being studied analytically or by simulation.

## References

Beckmann, J.S. & M. Soller, 1983. Restriction fragment length polymorphisms in genetic improvement methodologies, mapping and costs. Theor. Appl. Genet. 67: 35-43.

Botstein, D., R.L. White, M. Skolnick & R.W. Davis, 1980. Construction of a genetic map in man using restriction length polymorphisms. Am. J. Hum. Genet. 32: 314-331.

Cowen, N.M., 1989. Multiple linear regression analysis of RFLP data sets used in mapping QTLs. *In:* Helentjaris T, B. Burr (Eds.) Development and application of molecular markers to problems in plant genetics. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, pp. 113-116.

De Vicente, M.C. & S.D. Tanksley, 1993. QTL analysis of transgressive segregation in an interspecific tomato cross. Genetics 134: 585-596.

Haley, C.S. & S.A. Knott, 1992. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity 69: 315-324.

Jansen, R.C., 1992. A general mixture model for mapping quantitative trait loci by using molecular markers. Theor. Appl. Genet. 85: 252-260.

Jansen, R.C., 1993. Interval mapping of multiple quantitative trait loci. Genetics 135: 205-211.

Jansen, R.C. & P. Stam, 1994. High resolution of quantitative traits into multiple loci via interval mapping. Genetics 136: 1447-1455.

Jansen, R.C., 1994. Controlling the type I and type II errors in mapping quantitative trait loci. Genetics: in press.

Jensen, J., 1989. Estimation of recombination parameters between a quantitative trait locus (QTL) and two marker gene loci. Theor. Appl. Genet. 78: 613-618.

Johannsen, W., 1909. Elemente der exakten Erblichkeitslehre. Fisher, Jena.

Knapp, S.J., W.C. Bridges & D. Birkes, 1990. Mapping quantitative trait loci using molecular marker linkage maps. Theor. Appl. Genet. 79: 583-592.

Knapp, S.J., 1991. Using molecular markers to map multiple quantitative trait loci: models for backcross, recombinant inbred, and doubled haploid progeny. Theor. Appl. Genet. 81: 333-338.

Lander, E.S. & D. Botstein, 1989. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121: 185-199.

Martinez, O. & R.N. Curnow, 1992. Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. Theor. Appl. Genet. 85: 480-488.

Moreno-Gonzalez, J., 1992. Genetic models to estimate additive and non-additive effects of marker-associated QTL using multiple regression techniques. Theor. Appl. Genet. 85: 435-444.

Paterson, A.H., S. Damon, J.D. Hewitt, D. Zamir, H.D. Rabinowitch, S.E. Lincoln, E.S. Lander & S.D. Tanksley, 1990. Mendelian factors underlying quantitative traits in tomato: comparison across species, generations, and environments. Genetics 127: 181-197.

Rasmusson, J.M., 1933. A contribution to the theory of quantitative character inheritance. Heriditas 18: 245-261.

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

Sax, K., 1923. Association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris.* Genetics 8: 552-560.

Simpson, S.P., 1989. Detection of linkage between quantitative trait loci and restriction fragment length polymorphisms using inbred lines. Theor. Appl. Genet. 77: 815-819.

Soller, M., T. Brody & A. Genizi, 1976. On the power of experimental design for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines. Theor. Appl. Genet. 47: 35-39.

Stam, P., 1991. Some aspects of QTL analysis. In: Proceedings of the eighth meeting of the Eucarpia section

"Biometrics in plant breeding", BRNO.

Stuber, C.W. S.E. Lincoln, D.W. Wolff, T. Helentjaris & E.S. Lander, 1992. Identification of genetic factors contributing to heterosis in a hybrid from two elite maize inbred lines using molecular markers. Genetics 132: 823-839.

Tanksley, S.D., N.D. Young, A.H. Paterson & M.W. Bonierbale, 1989. RFLP mapping in plant breeding: new tools for an old science. Biotechnology 7: 257-264.

Thoday, J.M., 1961. Location of polygenes. Nature 191: 368-370.

Weller, J.I., 1986. Maximum likelihood techniques for the mapping and analysis of quantitative trait loci with the aid of genetic markers. Biometrics 42: 627-640.

Zeng, Z.-B., 1993. Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. Proc. Natl. Acad. Sci. USA. 90: 10972-10976.

Zeng, Z.-B., 1994. Precision mapping of quantitative trait loci. Genetics 136: 1457-1468.