Ritsert C. Jansen

Genetic mapping of quantitative trait loci in plants - a novel statistical approach

Proefschrift

ter verkrijging van de graad van doctor
in de landbouw- en milieuwetenschappen
op gezag van de rector magnificus,
dr. C.M. Karssen,
in het openbaar te verdedigen
op maandag 27 februari 1995,
des namiddags te vier uur in de aula
van de Landbouwuniversiteit te Wageningen

CIP-DATA KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Jansen, R.C.

Genetic mapping of quantitative trait loci - a novel statistical approach/
Ritsert C. Jansen - [S.l.: s.n.].
Thesis Wageningen - With summary in Dutch
ISBN 90-73771-14-5
Subject headings: genetic mapping, quantitative trait loci, statistics.

Aan mijn ouders

Aan Henny, Rianne, Ymke en Yde

VOORWOORD

In dit boekje zijn een aantal artikelen gebundeld, zoals die de afgelopen jaren op mijn beeldscherm (en daarna in wetenschappelijke tijdschriften) verschenen zijn. Het moment is dan ook voor mij daar om, naar goede wetenschappelijke traditie, dit werk als 'proeve van bekwaamheid tot het zelfstandig uitoefenen van de wetenschap' te verdedigen ten overstaan van de promotiecommissie.

Hoe heeft het toch zo ver kunnen komen? Dit proefschrift gaat in hoge mate over toeval en genetische en omgevings-factoren, zodat een verklaring in deze hoek voor de hand ligt. Allereerst: met een moleculair geneticus als vader en een zoöloge als moeder is de kans op een individu met interesse in biologie, en genetica in het bijzonder, natuurlijk aanzienlijk. Wanneer zo'n individu dan ook nog in de juiste voedingsbodem mag opgroeien, dan kan het haast niet meer mislopen. Hoe dan ook, Gerard en Netty, heel veel dank voor jullie genen en ondersteuning.

In 1981 ben ik wiskunde gaan studeren aan de Rijksuniversiteit Groningen met als uiteindelijk doel het toepassen van wiskundige gereedschappen in 'de groene hoek'. Vooral statistiek was (en is) een zeer leuk vak. Willem Schaafsma heeft mij op bijzondere wijze in de wereld van statistiek ingevoerd.

Begin 1988 kwam ik op het toenmalige Instituut voor de Veredeling van Tuinbouwgewassen (IVT) werken. De wijze waarop Hans Jansen mij begeleidde en toch ergens volledig vrij liet, heb ik zeer gewaardeerd. Ten gevolge van fusies kwamen wij in 1990 met twee kwantitatief genetici, Piet Stam en Johan van Ooijen, in de nieuwe afdeling 'Populatiebiologie' van het nieuwe 'Centrum voor Plantenveredelings- en Reproductieonderzoek' (CPRO-DLO). Al weer een (toevallig?) schot in de roos, want hierdoor passeerden nieuwe en uitdagende genetische problemen mijn pad en ik prijs mij dan ook zeer gelukkig dat ik zo met de neus in de 'QTL mapping' boter mocht vallen. Het is zonder meer duidelijk dat de kruisbevruchting binnen onze afdeling geleid heeft tot 'hybride' krachten. Ik wil na Hans ook met name Piet zeer nadrukkelijk bedanken: op vele momenten ben jij de aangever geweest van interessante problemen en hebben onze discussies bijgedragen tot de oplossing ervan. Onder jouw begeleiding en met jouw soms nadrukkelijke ondersteuning heb ik mij in het wetenschappelijke strijdgewoel rond 'QTL mapping' gestort. Het is en blijft dan ook jammer dat de eenheid binnen de afdeling Populatiebiologie door jouw vertrek verstoord is. Aan de andere kant doet het me juist veel plezier dat jij nu als mijn promotor kan optreden. Ik hoop van harte dat de plezierige samenwerking met jou en Johan nog verder mag groeien.

Ook vele andere collega's, zowel binnen als buiten het CPRO-DLO, hebben bijgedragen tot een goede werkomgeving. Ik heb van verschillende collega's (Paul Odinot en Pim Lindhout, afdeling Groente en Fruit, CPRO-DLO; Peter de Boer, vakgroep Erfelijkheidsleer, LUW; Clare Lister en Caroline Dean, Department of Molecular Genetics, John Innes Centre) fraaie gegevens uit echte experimenten gekregen en, na analyse, verwerkt in dit boekje. Maarten Koornneef (vakgroep Erfelijkheidsleer, LUW) heeft mij voorzien van achtergrond-informatie over *Arabidopsis*.

De leden van de promotiecommissie, Robert Curnow als tweede promotor in het bijzonder, wil ik bedanken voor hun bereidheid zich over mijn werk te buigen.

Tenslotte, Henny, Rianne, Ymke en Yde: genetica is fantastisch zowel in theorie als ook in praktijk!

CONTENTS

[1] *Biometrics in Plant Breeding: applications of molecular markers.* pp. 116-124, 1994

[2] *Theoretical and Applied Genetics* 85:252-260, 1992

[3] *Genetics* 135:205-211, 1993

[4] *Genetics* 136:1447-1455, 1994, with P. STAM

[5] *Genetics* 138:871-881, 1994

[6] *Theoretical and Applied Genetics*, 1995, with J.W. VAN OOIJEN, P. STAM, C. LISTER and C. DEAN

[7] *Biometrics* 49:227-231, 1993

ABSTRACT

Quantitative variation is a feature of many important traits such as yield, quality and disease resistance in crop plants and farm animals, and diseases in humans. The genetic mapping, understanding and manipulation of quantitative trait loci (QTLs) are therefore of prime importance. Only by using genetically marked chromosomes is it possible to detect and map these QTLs. The recent advent of complete genetic maps of molecular markers for many plant and animal species therefore heralds a new era for quantitative genetics. The "interval mapping" approach to QTL mapping is now widely used, though true resolution of quantitative variation into QTLs is hampered because only single-QTL models are used. Here we develop a novel analytical approach, "MQM mapping", where MQM is an acronym for multiple-QTL models as well as for marker-QTL-marker. Computer simulation work and practical experiments in tomato and in the model organism *Arabidopsis thaliana* demonstrate the superiority of the new approach over the conventional one in genetic mapping of multiple genes underlying quantitative variation.

OUTLINE OF THE THESIS

In chapter I the issue of genetic mapping of quantitative trait loci (QTLs) is introduced and an overview of biometrical methods used is presented.

In chapter II a novel, general and flexible biometrical framework for QTL mapping is developed. A very simple algorithm to obtain estimates of the parameters of the models is described. Our approach can be applied to various types of QTL background, and to many types of progeny, trait, experimental setup, etcetera. Two simulated backcross examples are worked out to demonstrate the models in "full action". The problem of simultaneously mapping multiple QTLs is also addressed. Exact models for multiple QTLs can be fitted to the data, at least in principle, but much computational work is necessary when the number of QTLs is large. An adaptive approach to the mapping of multiple QTLs is therefore suggested. In this approach models are exact for a single putative QTL at a given map location. They are however approximate for other putative QTLs, due to the fact that these QTLs are replaced by nearby markers (i.e. markers are used as "cofactors").

In chapter III a simple simulation study is presented with three QTLs, two of which are located on the same chromosome. A more realistic simulation study concerning the detection of eleven QTLs on a genome of ten chromosomes is also included. These studies illustrate the use of marker cofactors in the detection of multiple QTLs.

In chapter IV the problem of missing observations at marker loci is solved, a problem which so far hampered the use of markers as cofactors in practical experiments. The core of the very general method described is the completion of any missing genotypic (QTL and marker) observations. A practical example is described in which multiple QTLs for plant height in tomato are mapped in an F_2 progeny. It is demonstrated how additional parental data can be used in QTL mapping.

In chapter V the chance of a type I error (i.e. a QTL is indicated at a location where actually no QTL is present) and the chance of a type II error (i.e. a QTL is not detected) are studied by computer simulation. We address problems concerning the selection of "important" marker cofactors, and problems concerning the fitting of models with many marker cofactors relative to the number of plants. Our mapping approach is refined, so as to make it possible to exploit the full power of complete marker linkage maps. The approach is thereupon called MQM mapping, where MQM is an acronym for "multiple-QTL models" as well as for "marker-QTL-marker".

In chapter VI a practical example is presented, in which QTLs time and QTL by environment interactions are detected for flowering time in recombinant inbred lines of *Arabidopsis thaliana*.

Finally, in chapter VII it is demonstrated that our method of parameter estimation makes it easy to handle complex mixture models in other areas of research. A practical example using data on non-disjunction in the mouse is given.

I. MAPPING OF QUANTITATIVE TRAIT LOCI BY USING GENETIC MARKERS: AN OVERVIEW OF BIOMETRICAL MODELS USED

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 20th 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 and 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 and 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 and 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

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

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 $1\text{-}2r_i$ 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

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

where the indicator variable x takes the value 0 and 1 for the genotypes mm and Mm, respectively, and μ_{Mm}-μ_{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

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

where μ_{Qq}-μ_{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 and 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 and Lefort 1993; Jansen 1993; Zeng 1993; Jansen and 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

$$Y=\mu+\sum x_i a_i+E\ , \tag{4}$$

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-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 and 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 and 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 and Knott 1992; Martinez and 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 and Knott (1992) and Martinez and 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 μ_{Qq} and μ_{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-th locus. The regression model reads

$$Y = \mu + \sum X_i a_i + E \ , \tag{7}$$

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 and 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 and 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.

LITERATURE CITED

BECKMANN, J.S. and 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 and 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: Development and application of molecular markers to problems in plant genetics, edited by Helentjaris T, Burr B. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, pp. 113-116

DE VICENTE, M.C. and S.D. TANKSLEY, 1993 QTL analysis of transgressive segregation in an interspecific tomato cross. Genetics 134:585-596

HALEY, C.S. and S.A. KNOTT, 1992 A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity: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. and 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 and 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. and D. BOTSTEIN, 1989 Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185-199

MARTINEZ, O. and 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 and S.D. TANKSLEY, 1991 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. and 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 and 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 and 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 and 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

II. A GENERAL MIXTURE MODEL FOR MAPPING QUANTITATIVE TRAIT LOCI BY USING MOLECULAR MARKERS

ABSTRACT

In a segregating population a quantitative trait may be considered to follow a mixture of (normal) distributions, the mixing proportions being based on Mendelian segregation rules. A general and flexible mixture model is proposed for mapping quantitative trait loci (QTLs) by using molecular markers. A method is described to fit the model to data. The model makes it possible to (1) analyse non-normally distributed traits such as lifetimes, counts or percentages in addition to normally distributed traits, (2) reduce environmental variation by taking into account the effects of experimental design factors and interaction between genotype and environment, (3) reduce genotypic variation by taking into account the effects of two or more QTLs simultaneously, (4) carry out a (combined) analysis of different population types, (5) estimate recombination frequencies between markers or use known marker distances, (6) cope with missing marker observations, (7) use markers as covariables in detection and mapping of QTLs, and finally to (8) implement the mapping in standard statistical packages.

INTRODUCTION

The advent of complete linkage maps of molecular markers has recently stimulated interest in studying the genetics underlying quantitative traits (cf. PATERSON et al 1988; SOLLER and Beckmann 1983). Several methods have been proposed for mapping quantitative trait loci (QTLs). Methods proposed by WELLER (1986) and LUO and KEARSEY (1989) are based on estimation of linkage between a single putative QTL and a single marker. JENSEN (1989), LANDER AND BOTSTEIN (1989) and KNAPP et al. (1990) used a model involving flanking markers for detection and mapping of a single QTL. In this case linkage between a putative QTL and two markers is estimated. LANDER and BOTSTEIN (1989) developed a software package (MAPMAKER-QTL) for backcross (BC) populations and F_2 populations. KNAPP et al. (1990) mentioned that they were also developing a software package (GENEMAP) for BC and F_2 populations. WELLER (1986) emphasized that in a BC or F_2 population a quantitative trait may be considered to follow a mixture of (normal) distributions. The mapping algorithm in both MAPMAKER-QTL and GENEMAP uses maximum likelihood methods based on the EM-algorithm to estimate parameters of the mixture model of normal distributions. LANDER AND BOTSTEIN (1989), KNAPP et al. (1990) and KNAPP (1991) mentioned the need for accurate and efficient methods which can handle multiple QTLs. Methods are also required that can cope adequately with non-normally distributed traits, such as lifetimes, percentages or counts. Similarly, methods are required which can cope with designed experiments, in which populations are tested at a number of locations and in various years to study interactions between genotype and environment, or in which randomised blocks or other designs are used to control variation in experiments. LANDER and BOTSTEIN (1989) stated that standard computer programs for linear regression cannot be used. KNAPP et al. (1990) and KNAPP (1991) developed linear models for multiple unlinked QTLs and non-linear models for two and three linked QTLs, and for interactions between QTLs and environment. However, these models are no mixture models.

In the present paper a mixture model is developed to overcome some of the shortcomings of the methods mentioned previously. Extensions of mixture models and parameter estimation methods based on the EM-algorithm, as proposed by LANDER and BOTSTEIN (1989) and KNAPP et al. (1990), are described. In this paper the emphasis is on genetical and statistical modelling of the mapping problem, not on the detection problem. Two simulated examples are included. The first example illustrates modelling for a non-normally distributed trait and some problems concerning the robustness of the traditional approaches for deviations from normality. The second example illustrates modelling for multiple QTLs and some problems concerning detection of QTLs.

GENETICAL AND STATISTICAL MODELS

In the QTL-mapping problem the phenotype of the quantitative trait and the allelic constitution at the marker loci are observed, whereas the allelic constitution at the QTLs remains unobserved. However, for each individual weights may be specified, which quantify the (conditional) probability for each possible allelic constitution at the QTLs (KNAPP et al. 1990). In the present paper it is demonstrated that this enables one to reduce the QTL-mapping problem to two classical problems, one concerned with genetic linkage and the other with regression of phenotype on genotype. Genetic linkage models and models for regression of phenotype on genotype will be recapitulated in the next two sections. In these two sections it is assumed that the allelic constitution at the QTL is known. Then, in a third section it is supposed that the allelic constitution at the QTLs is unknown and the method for mapping QTLs will be developed. Consequences of a single QTL and two QTLs are considered in the cases of selfing F_1 individuals (M_1QM_2/m_1qm_2 and $M_1Q_1M_2Q_2M_3/m_1q_1m_2q_2m_3$, respectively) to obtain an F_2 population, and back crossing F_1 individuals to one of the parents (say m_1qm_2/m_1qm_2 and $m_1q_1m_2q_2m_3/m_1q_1m_2q_2m_3$, respectively) to obtain a BC-population. Extension to any other number of QTLs and to other population types is straightforward.

Genetic Linkage. A general model for estimation of genetic linkage between markers and QTLs will be described. The model makes it possible to (1) take into account a single QTL, or two or more QTLs simultaneously, (2) analyse BC populations, F_2 populations and many other populations, (3) estimate recombination parameters between markers, or use known marker distances, and (4) implement the parameter estimation in standard statistical packages.

The classical theory of genetic linkage has been described by BAILEY (1961). In this section the problem of estimation of genetic linkage parameters will be treated differently, namely by using log-linear models. Moreover, it will be assumed that the complete allelic constitution of chromosomes is observed, which implies that repulsion and coupling phases can be distinguished and that recombination events can be counted. The adaptive approach enables one to implement the mapping of QTLs readily in statistical packages, as will be made clear in one of the following sections.

First, the case of a single QTL with flanking markers is considered, which corresponds to the classical 'three point' linkage analysis. Let r_1 and r_2 denote the recombination frequency between the QTL and its flanking markers, respectively. Table 1 shows the gametes produced by F_1 individuals (M_1QM_2/m_1qm_2), classified by the recombination events in a 2 x 2 table. Table 1 also shows the expected frequencies of the four categories in the absence of interference. Let p_{00}, p_{01}, p_{10} and p_{11} denote the frequencies of the four categories of gametes in Table 1. The recombination events follow a multinomial distribution with parameters p_{00}, p_{01}, p_{10} and p_{11}, while the eight gamete types follow a multinomial distribution with parameters $\frac{1}{2}p_{ij}$ (i,j = 0,1). The usual log-linear model holds for the eight gamete types:

$\log(\frac{1}{2}p_{00}) = \lambda$, if the gamete is M_1QM_2 or m_1qm_2,
$\log(\frac{1}{2}p_{10}) = \lambda + \nu$, if the gamete is m_1QM_2 or M_1qm_2,
$\log(\frac{1}{2}p_{01}) = \lambda + \zeta$, if the gamete is M_1Qm_2 or m_1qM_2, and
$\log(\frac{1}{2}p_{11}) = \lambda + \nu + \zeta$, if the gamete is M_1qM_2 or m_1Qm_2,

Table 1. Gametes produced by F_1 individuals and expected frequencies of the four categories of gametes

Recombination between QTL and first marker[a]	Recombination between QTL and second marker[a]	
	0	1
0	M_1QM_2, m_1qm_2 $(1-r_1)(1-r_2)$	M_1Qm_2, m_1qM_2 $(1-r_1)r_2$
1	m_1QM_2, M_1qm_2 $r_1(1-r_2)$	M_1qM_2,m_1Qm_2 r_1r_2

[a] 0, No recombination; 1, recombination

Table 2. Coefficients of the genetic linkage parameters and the parameters for regression of phenotype on genotype in the progeny obtained by backcrossing F_1 individuals M_1 QM_2/m_1 qm_2 to the parent m_1 qm_2/m_1 qm_2

Observed incomplete allelic constitution	Unobserved complete allelic constitution	Genetic linkage[a]			Regression of phenotype on genotype[b]		
		λ	ν	ζ	*m*	*a*	*d*
M_1m_1 M_2m_2	M_1QM_2/m_1qm_2	1	0	0	1	0	1
	M_1qM_2/m_1qm_2	1	1	1	1	-1	0
M_1m_1 m_2m_2	M_1Qm_2/m_1qm_2	1	0	1	1	0	1
	M_1qm_2/m_1qm_2	1	1	0	1	-1	0
m_1m_1 M_2m_2	m_1QM_2/m_1qm_2	1	1	0	1	0	1
	m_1qM_2/m_1qm_2	1	0	1	1	-1	0
m_1m_1 m_2m_2	m_1Qm_2/m_1qm_2	1	1	1	1	0	1
	m_1qm_2/m_1qm_2	1	0	0	1	-1	0

Example: For M1 qm2/m1 qm2 individuals the coefficients of the genetic linkage parameters are $1\cdot\lambda$, $1\cdot v$ and $0\cdot\zeta$, since $\log[½r_1(1-r_2)]=\log[½(1-r_1)(1-r_2)]+\log[r_1/(1-r_1)]=1\cdot\lambda+1\cdot v+0\cdot\zeta$; for M_1qm_2/m_1qm_2 individuals the coefficients of the parameters for regression of phenotype on genotype are $1\cdot m$, $-1\cdot a$ and $0\cdot d$, since the genotypic value satisfies $G=m-a$

[a] λ, v and ζ denote the parameters for the linear genetic linkage model: $\lambda=\log(½(1-r_1)(1-r_2))$; $v=\log(r_1)-\log(1-r_1)$; $\zeta=\log(r_2)-\log(1-r_2)$, with r_1 and r_2 denoting the recombination frequencies between the QTL and its flanking markers

[b] *m,a,* and *d* denote the parameters for linear regression of phenotype on genotype: *m* is the mean of the expected phenotypes of individuals with QQ and qq at the QTL, respectively; *a* is the additive effect; *d* is the dominance effect

where $\nu=\log(r_1)-\log(1-r_1)$ and $\zeta=\log(r_2)-\log(1-r_2)$. The parameters are subject to the constraint $p_{00}+p_{01}+p_{10}+p_{11}=1$. In BC data only the chromosome originating from the F_1 parent provides information on the recombination parameters r_1 and r_2. Table 2 shows coefficients of the genetic linkage parameters for each of the eight possible allelic constitutions. For example M_1QM_2/m_1qm_2 has coefficients $1\cdot\lambda$, $0\cdot\nu$ and $0\cdot\zeta$, since $\log(½p_{00})=1\cdot\lambda+0\cdot\nu+0\cdot\zeta$.

In F_2 data both homologous chromosomes originate from F_1 parents and therefore both homologous chromosomes are informative. When calculating probabilities it is useful to distinguish chromosomes of maternal and paternal origin. Let M_1QM_2/M_1qM_2 denote the genotype of an individual with chromosome M_1QM_2 of maternal origin and chromosome M_1qM_2 of paternal origin. Other genotypes are defined similarly. Maternal and paternal chromosomes are independent, so that pairs of chromosomes occur in

Table 3. Coefficients of the genetic linkage parameters and the parameters for regression of phenotype on genotype in the F_2 progeny of selfed M_1QM_2/m_1qm_2 individuals

Observed incomplete allelic constitution	*Unobserved complete allelic constitution*	*Genetic linkage*[a]				*Regression of phenotype on genotype*[b]		
		λ	ν	ζ	Offset	*m*	*a*	*d*
M_1M_1 M_2M_2	M_1QM_2/M_1QM_2	2	0	0	0	1	1	0
	M_1QM_2/M_1qM_2, M_1qM_2/M_1QM_2	2	1	2	log(2)	1	0	1
	M_1qM_2/M_1qM_2	2	2	1	0	1	-1	0
M_1M_1 M_2m_2	M_1QM_2/M_1Qm_2, M_1Qm_2/M_1QM_2	2	0	1	log(2)	1	1	0
	M_1QM_2/M_1qm_2, M_1qm_2/M_1QM_2	2	1	0	log(2)	1	0	1
	M_1Qm_2/M_1qM_2, M_1qM_2/M_1Qm_2	2	1	2	log(2)	1	0	1
	M_1qM_2/M_1qm_2, M_1qm_2/M_1qM_2	2	2	1	log(2)	1	-1	0
M_1M_1 m_2m_2	M_1Qm_2/M_1Qm_2	2	0	2	0	1	1	0
	M_1Qm_2/M_1qm_2, M_1qm_2/M_1Qm_2	2	1	1	log(2)	1	0	1
	M_1qm_2/M_1qm_2	2	1	0	0	1	-1	0
M_1m_1 M_2M_2	$M_{1Q}M_2/m_1QM_2$, m_1QM_2/M_1QM_2	2	1	0	log(2)	1	1	0
	M_1QM_2/m_1qM_2, m_1qM_2/M_1QM_2	2	0	1	log(2)	1	0	1
	M_1qM_2/m_1QM_2, m_1QM_2/M_1qM_2	2	2	1	log(2)	1	0	1
	M_1qM_2/m_1qM_2, m_1qM_2/M_1qM_2	2	1	2	log(2)	1	-1	0
M_1m_1 M_2m_2	M_1QM_2/m_1Qm_2, M_1Qm_2/m_1QM_2, m_1QM_2/M_1Qm_2, m_1Qm_2/M_1QM_2	2	1	1	log(4)	1	1	0
	M_1QM_2/m_1qm_2, m_1qm_2/M_1QM_2	2	0	0	log(2)	1	0	1
	M_1Qm_2/m_1qM_2, m_1qM_2/M_1QM_2	2	0	2	log(2)	1	0	1
	m_1QM_2/M_1qm_2, M_1qm_2/m_1QM_2	2	2	0	log(2)	1	0	1
	m_1Qm_2/M_1qM_2, M_1qM_2/m_1Qm_2	2	2	2	log(2)	1	0	1
	M_1qM_2/m_1qm_2, M_1qm_2/m_1qM_2, m_1qM_2/M_1qm_2, m_1qm_2/M_1qM_2	2	1	1	log(4)	1	-1	0
M_1m_1 m_2m_2	M_1Qm_2/m_1Qm_2, m_1Qm_2/m_1Qm_2	2	1	2	log(2)	1	1	0
	M_1Qm_2/m_1qm_2, m_1qm_2/m_1Qm_2	2	0	1	log(2)	1	0	1
	M_1qm_2/m_1Qm_2, m_1Qm_2/m_1qm_2	2	2	1	log(2)	1	0	1
	M_1qm_2/m_1qm_2, m_1qm_2/m_1qm_2	2	1	0	log(2)	1	-1	0
m_1m_1 M_2M_2	m_1QM_2/m_1QM_2	2	2	0	0	1	1	0
	m_1QM_2/m_1qM_2, m_1qM_2/m_1QM_2	2	1	1	log(2)	1	0	1
	m_1qM_2/m_1qM_2	2	0	2	0	1	-1	0
m_1m_1 M_2m_2	m_1QM_2/m_1Qm_2, m_1Qm_2/m_1QM_2	2	2	1	log(2)	1	1	0

	m_1QM_2/m_1qm_2, m_1qm_2/m_1QM_2	2	1	0	log(2)	1	0	1
	m_1qM_2/m_1Qm_2, m_1Qm_2/m_1qM_2	2	1	2	log(2)	1	0	1
	m_1qM_2/m_1qm_2, m_1qm_2/m_1qM_2	2	0	1	log(2)	1	-1	0
m_1m_1 m_2m_2	m_1Qm_2/m_1Qm_2	2	2	2	0	1	1	0
	m_1Qm_2/m_1qm_2, m_1qm_2/m_1Qm_2	2	1	1	log(2)	1	0	1
	m_1qm_2/m_1qm_2	2	0	0	0	1	-1	0

Example: For M_1QM_2/M_1qM_2 individuals the coefficients of the genetic linkage parameters are $2\cdot\lambda$, $1\cdot v$ and $1\cdot\zeta$, since $\log[\tfrac{1}{2}(1-r_1)(1-r_2)\cdot\tfrac{1}{2}r_1r_2]=2\log[\tfrac{1}{2}(1-r_1)(1-r_2)]+\log[r_1/(1-r_2)]=2\cdot\lambda+1\cdot v+1\cdot\zeta$; For M_1QM_2/M_1qM_2 individuals the coefficients of the parameters for regression of phenotype on genotype are $1\cdot m$, $0\cdot a$ and $1\cdot d$, since its genotypic value satisfies $G=m+d$. M_1QM_2/M_1qM_2 and M_1qM_2/M_1QM_2 have the same coefficients and are grouped together. An extra offset of log (2) appears, since $\log[2\tfrac{1}{2}(1-r_1)(1-r_2)\tfrac{1}{2}r_1r_2]=\log[\tfrac{1}{2}(1-r_1)(1-r_2)\tfrac{1}{2}r_1r_2]+\log(2)$

[a] λ, v and ζ denote the parameters for the genetic linkage model: $\lambda=\log(\tfrac{1}{2}(1-r)(1-r_2))$; $v=\log(r_1)-\log(1-r_1)$; $\zeta=\log(r_2)-\log(1-r_2)$, where r_1 and r_2 are the recombination frequencies between the QTL and its flanking markers

[b] m, a and d are the parameters for the linear regression of phenotype on genotype: m is the mean of the expected phenotypes of individuals with QQ and qq at the QTL, respectively; a is the additive effect; d is the dominance effect

expected frequencies $\tfrac{1}{2}p_{hi}\cdot\tfrac{1}{2}p_{jk}$ (h,i,j,k = 0,1). Since $\log(\tfrac{1}{2}p_{hi}\cdot\tfrac{1}{2}p_{jk}) = \log(\tfrac{1}{2}p_{hi}) + \log(\tfrac{1}{2}p_{jk})$, it follows that the linear model is the sum of the linear models for the separate chromosomes. Table 3 shows coefficients of the genetic linkage parameters for each of the 64 allelic constitutions. For example M_1QM_2/M_1qM_2 has coefficients $2\cdot\lambda$, $1\cdot\nu$ and $1\cdot\varsigma$, since $\log(\tfrac{1}{2}p_{00}\cdot\tfrac{1}{2}p_{11}) = \log(\tfrac{1}{2}p_{00})+\log(\tfrac{1}{2}p_{11}) = 1\cdot\lambda+(1\cdot\lambda+1\cdot\nu+1\cdot\varsigma) = 2\cdot\lambda+1\cdot\nu+1\cdot\varsigma$. Genotypes M_1QM_2/M_1qM_2 and M_1qM_2/M_1QM_2 have the same coefficients (and the same phenotype, see the next section) and may be grouped together. The probability that a genotype is either M_1QM_2/M_1qM_2 or M_1qM_2/M_1QM_2 equals $2\cdot\tfrac{1}{2}p_{00}\cdot\tfrac{1}{2}p_{11}$. Therefore, in the log-linear model an extra offset of log(2) appears, since $\log(2\cdot\tfrac{1}{2}p_{00}\cdot\tfrac{1}{2}p_{11}) = 2\cdot\lambda + l\cdot\nu + 1\cdot\varsigma + \log(2)$.

Next, the case of an F_1 ($M_1Q_1M_2Q_2M_3/m_1q_1m_2q_2m_3$) with two QTLs in adjacent intervals is considered. Let r_{11} and r_{12} denote the recombination frequencies between the first QTL and its flanking markers. Similarly, let r_{21} and r_{22} denote the recombination frequencies between the second QTL and its flanking markers. In the absence of interference the recombination events in the first interval are independent of those in the second interval. The expected proportions of gametes of F_1 individuals are the products of the expected proportions in the first (p_{1hi}) and second interval (p_{2jk}). Since
$\log(\tfrac{1}{2}p_{1hi}\cdot\tfrac{1}{2}p_{2jk}) = \log(\tfrac{1}{2}p_{1hi}) + \log(\tfrac{1}{2}p_{2jk})$, it follows that the linear model is the sum of the linear models for the separate QTLs (h,i,j,k=0,1).

Maximum likelihood estimates for the parameters of the log-linear model (a so-called generalized linear model for multinomial data) may be obtained easily (McCULLAGH and NELDER 1989): computations may be carried out by statistical packages having facilities for generalized linear models. It will be shown in one of the following sections that in solving the QTL mapping problem the genetic linkage analysis is carried out by fitting the log-linear model to weights which quantify the probability for each possible allelic constitution at the QTLs and marker loci. For example, suppose that the actual allelic constitution of a BC individual is M_1QM_2/m_1qm_2. In the QTL mapping problem only the allelic constitution M_1M_2/m_1m_2 can be observed. Two probabilities may be calculated, namely the probability that the complete allelic constitution is M_1QM_2/m_1qm_2, and the probability that it is M_1qM_2/m_1qm_2. Probabilities which are calculated, are conditional probabilities given the observed phenotypic value and given the observed marker genotype.

For each observation coefficients of the genetic linkage parameters are specified and stored into a design matrix or into explanatory variables to be analysed. If applicable, offsets are stored into an offset variable. Estimation is often carried out by the Newton-Raphson method or by the method of scoring. Note that since the log-linear models for BC-populations, F_2 populations and many other populations are specified in the same parameters, the corresponding data can be easily analysed by the same computer program. It is also possible to carry out a combined analysis of data of different population types.

If the map distance between markers is unknown or based on insufficient information, the multinomial proportions are free of further constraints. However, once a proper map of the markers is available, one may add additional constraints. For example in the case of a single QTL, the extra constraint becomes $p_{10} + p_{01} = t$, where t is the known recombination frequency between the two markers. Finally, it is remarked that models may also be extended to include interference constraints, e.g. $t=r_1+r_2-2Cr_1 \cdot r_2$ with C=1. Estimation may be carried out again by applying the Newton-Raphson method or by the method of scoring.

Regression of phenotype on genotype. A general model for regression of phenotype on genotype will be described. The model makes it possible to (a) analyse non-normally distributed traits such as lifetimes, counts or percentages in addition to normally distributed traits, (b) reduce environmental variation by taking into account the effects of experimental design factors and interaction between genotype and environment, (c) reduce genotypic variation by taking into account the effects of two or more QTLs simultaneously, and (d) implement the parameter estimation in standard statistical packages.

LANDER and BOTSTEIN (1989), KNAPP et al. (1990) and KNAPP (1991) discuss the traditional approach of regression of phenotype on genotype. We use the notation of BULMER (1985) and denote the phenotypic value by Y, the genotypic value by G, and the environmental variation by E. In this section it is assumed again that the allelic constitution at the QTLs is known. The simplest model is $Y = G + E$. The genotypic contribution is often decomposed into additive (A) and dominance (D) components. The following linear model for the genotypic values at a single diallelic locus is formulated by Bulmer (1985) in short notation as $G=m+A+D$, or written out

$G=m+a$, if an individual's genotype is QQ,
$G=m+d$, if its genotype is Qq, and
$G=m-a$, if its genotype is qq,

where m is the mean of the expected values of the genotypes QQ and qq, the additive component A takes values $+a$, 0 or $-a$ and the dominance component D takes values 0 or d.

In Table 2 coefficients of the regression parameters are presented for each of the 8 genotypes of a BC population. For example M_1QM_2/m_1qm_2 has coefficients $1 \cdot m$, $0 \cdot a$ and $1 \cdot d$. In a BC population additive and dominance components are aliased (Table 2). Therefore, the parameters μ_{Qq} and μ_{qq} will be used below to denote the expected values of individuals with allelic constitution Qq and qq at the QTL, respectively. In Table 3 coefficients of the regression parameters are presented for each of the 64 genotypes of an F_2 population. For example, M_1QM_2/m_1Qm_2 has coefficients $1 \cdot m$, $1 \cdot a$ and $0 \cdot d$, M_1QM_2/m_1qm_2 has coefficients $1 \cdot m$, $0 \cdot a$ and $1 \cdot d$, and M_1qM_2/m_1qm_2 has coefficients $1 \cdot m$, $-1 \cdot a$ and $0 \cdot d$.

The model is readily extended to take into account two or more QTLs simultaneously. For example, the two loci linear model is $G=m+A_1+A_2+D_1+D_2+AA_{12}+AD_{12}$ $+AD_{21}+DD_{12}$ (BULMER, 1985).

Experimental design factors, such as blocks, have to be incorporated into the model to provide a certain degree of control over environmental variation. But also interactions between genotype and environment, such as year x genotype or location x genotype interactions, are of particular interest. The model is also readily extended to take such explanatory variables into account. For example, the single QTL model may be extended to $G=m+A+D+X'ß$, where $X'ß$ relates the genotypic value to the explanatory variables ($ß$ is a vector of regression parameters and X is a vector of coefficients of regression parameters).

Usually, the environmental variation E is assumed to be normally distributed with mean 0 and variance σ^2. However, it may actually have some other continuous distribution, such as the log-normal or exponential distribution. It may even be discrete rather than continuous, such as is the case when percentages, counts or ordinal data are recorded. Generalized linear models provide an extension of classical linear models for normally distributed data to binomial data (percentages), Poisson data (counts), ordinal data (severity scores) and other types of data. Maximum likelihood methods for normally distributed data can be found in many statistical text books. Generalized linear models, and how to fit them to data, are extensively discussed by McCullagh and Nelder (1989).

It will be shown in the next section that in solving the QTL mapping problem a weighted regression analysis is carried out, in which the weights quantify the conditional probability for each possible allelic constitution at the QTLs and marker loci.

Mapping quantitative trait loci. A general mixture model for mapping QTLs will be described now. The model makes it possible to (a) transfer to the QTL-mapping problem all facilities developed above for the two classical problems, one concerned with genetic linkage and the other with regression of phenotype on genotype (facilities such as analysis of non-normally distributed traits, or analysis of designed experiments), (b) cope with missing QTL data and missing marker data, and (c) implement the mapping in standard statistical packages.

In the QTL-mapping problem the phenotype of the quantitative trait and the allelic constitution at the marker loci are observed, whereas the allelic constitution at the QTLs remains unobserved. However, for each individual weights may be specified, which quantify the conditional probability for each possible allelic constitution at the QTLs (Knapp et al. 1990). Note that the information on the allelic constitution at the marker loci is in general also incomplete, since the phases (coupling or repulsion) remain unobserved. The information on the marker genotype may also be incomplete due to dominance, or to problems in classification. A special case of missing marker data is so-called selective genotyping, in which case marker data are collected only for the extreme phenotypic values (Lander and Botstein 1989). An adaptive approach is to specify weights, which quantify the conditional probability for each possible allelic constitution at the QTLs and the marker loci simultaneously. It will be shown below that this enables one to implement the mapping of QTLs readily in statistical packages.

The EM-algorithm, proposed by Demster et al. (1977), may be used to specify and update weights iteratively. It will be demonstrated here that application of the EM-algorithm enables one to reduce the QTL-mapping problem to two classical problems, one concerned with genetic linkage and the other with regression of phenotype on genotype.

Each iteration of the EM-algorithm consists of two steps:

Step 1: specify or update weights, and

Step 2: update the parameter estimates by

(1) a genetic linkage analysis based on the weights, and

(2) a weighted regression of phenotype on genotype.

In step 1 the weights are updated by calculating the conditional probabilities given the current parameter estimates according to the Bayes theorem (KNAPP et al. 1990; MCLACHLAN and BASFORD 1988; TITTERINGTON et al. 1985). In step 2 the classical problems are solved by using the weights. In the preceding sections the solutions of the corresponding classical problems have been discussed.

Let us suppose again that in the BC population the phenotype *y* and the marker genotype M_1m_1 M_2m_2 were observed. Coefficients of the parameters for the two possible complete genotypes $M_1QM_2/m1qm_2$ and M_1qM_2/m_1qm_2 are stored into a design matrix or into explanatory variables to be analysed (Table 2). The corresponding weights are stored into an extra variable.

Let us suppose next that in the F_2 population the phenotype *y* and the marker genotype M_1M_1 M_2m_2 were observed. Since the complete genotype has one of the following eight allelic constitutions M_1QM_2/M_1Qm_2, M_1Qm_2/M_1QM_2, M_1QM_2/M_1qm_2, M_1qm_2/M_1QM_2, M_1qM_2/M_1Qm_2, M_1Qm_2/M_1qM_2, M_1qM_2/M_1qm_2 or M_1qm_2/M_1qM_2, the phenotype may be assumed to follow a mixture of eight distributions (Table 3). However, genotypes having the same coefficients of the regression parameters can be grouped together, so that the complete genotype is in one of the following four groups: $\{M_1QM_2/M_1Qm_2$ or $M_1Qm_2/M_1QM_2\}$, $\{M_1QM_2/M_1qm_2$ or $M_1qm_2/M_1QM_2\}$, $\{M_1qM_2/M_1Qm_2$ or $M_1Qm_2/M_1qM_2\}$ and $\{M_1qM_2/M_1qm_2$ or $M_1qm_2/M_1qM_2\}$. Therefore, the number of components in the mixture can be reduced, so that the phenotype y can be assumed to follow a mixture of four distributions. As a consequence, an offset of log(2) appears in the log-linear model for genetic linkage. It can be derived analogously that the phenotype y follows a mixture of three distributions when an individual is homozygous at both marker loci, of four distributions when it is homozygous at only one of the marker loci, and finally of six distributions when it is heterozygous at both loci (Table 3). Therefore, individuals are replicated three, four or six times in the design matrix or explanatory variables, depending on their observed marker genotype. The weights of the corresponding allelic constitutions are stored again into an extra variable.

The two steps of the algorithm are alternated until convergence. The algorithm is conveniently started by (arbitrary) thresholding of the data, giving initial weights equal to 0 or 1. Alternatively, the algorithm can be started by setting the parameters to (well choosen) initial values. The analyses can be carried out by statistical packages which have facilities for generalized linear models.

Formal justification. Continuous phenotypic data, such as observed when the trait is

Notation:

y	phenotype
h	genotype (incomplete information)
g	genotype (complete information)
$p(h)$	expected proportion of h
$p(g)$	expected proportion of g
$p(g\|h)$	expected proportion of g given h
$p(g\|y,h)$	expected proportion of g given h and y
$f(y\|h)$	probability density function given h
$f(y\|g)$	probability density function given g

normally distributed, will be considered here. Expressions for discrete phenotypic data, such as counts or percentages, can be obtained by substituting probabilities for densities. The likelihood $\mathcal{L}$ of observations $(y_1,h_1),(y_2,h_2),\dots,(y_I,h_I)$ is

$$\mathcal{L}((y_1,h_1),(y_2,h_2)\dots(y_I,h_I))=\prod_{i=1}^{I} f(y_i,h_i)=\prod_{i=1}^{I} p(h_i)\cdot\prod_{i=1}^{I} f(y_i\mid h_i).$$

Parameter estimation will be carried out by maximum likelihood. The likelihood equations are

$$0=\frac{\partial}{\partial\theta}\log\mathcal{L}=\sum_{i=1}^{I}\frac{\partial}{\partial\theta}\log p(h_i)+\sum_{i=1}^{I}\frac{\partial}{\partial\theta}\log f(y_i\mid h_i)$$

$$=\sum_{i=1}^{I}\frac{\partial}{\partial\theta}\log p(h_i)+\sum_{i=1}^{I}\frac{1}{f(y_i\mid h_i)}\frac{\partial}{\partial\theta}\sum_{g}p(g\mid h_i)f(y_i\mid g)$$

$$=\sum_{i=1}^{I}\frac{\partial}{\partial\theta}\log p(h_i)+\sum_{i=1}^{I}\sum_{g}\left(\frac{p(g\mid h_i)\cdot f(y_i\mid g)}{f(y_i\mid h_i)}\frac{\partial}{\partial\theta}\log(p(g\mid h_i)f(y_i\mid g))\right)$$

$$=\sum_{i=1}^{I}\frac{\partial}{\partial\theta}\log p(h_i)+\sum_{i=1}^{I}\sum_{g}p(g\mid y_i,h_i)\frac{\partial}{\partial\theta}\log(p(g\mid h_i)f(y_i\mid g))$$

$$=\sum_{i=1}^{I}\frac{\partial}{\partial\theta}\log p(h_i)+\sum_{i=1}^{I}\sum_{g}p(g\mid y_i,h_i)\frac{\partial}{\partial\theta}\log p(g\mid h_i)+\sum_{i=1}^{I}\sum_{g}p(g\mid y_i,h_i)\frac{\partial}{\partial\theta}\log f(y_i\mid g)$$

$$=\sum_{i=1}^{I}\sum_{g}p(g\mid y_i,h_i)\frac{\partial}{\partial\theta}\log p(g)+\sum_{i=1}^{I}\sum_{g}p(g\mid y_i,h_i)\frac{\partial}{\partial\theta}\log f(y_i\mid g).$$

The problem can be considered as a missing data problem. The likelihood equation can be solved by applying the EM algorithm, proposed by DEMPSTER et al. (1977). Each iteration consists of two steps. First, in the so-called E-step, the conditional probability

$$p(g\,|\,y_i,h_i)=\frac{p(g\,|\,h_i)\cdot f(y_i\,|\,g)}{f(y_i\,|\,h_i)}$$

is evaluated for all possible allelic constitutions g, given the current parameter estimates and given the observed incomplete information h_i on the genotype. Next, in the so-called M-step, the likelihood equation is solved by fixing the weights $p(g\,|\,y_i,h_i)$ whereby updated parameter estimates are obtained. Note that p(g) is a function of recombination parameters only, whereas $f(y|g)$ is a function of parameters for the regression of phenotype on genotype. Therefore, the likelihood equation can be split into two terms: the first term refers to the genetic linkage problem, the second term to the problem of regression of phenotype on genotype. Thus, the one M-step for the mixture problem is split into two M-steps for the two classical non-mixture problems.

EXAMPLES

Two simulated backcross examples will be worked out here: (1) the case of mapping a single QTL affecting lifetime (assumed to be exponentially distributed), and (2) the case of two QTLs in adjacent intervals with genes in repulsion phase and the QTLs affecting a normally distributed trait. These cases show the general mixture model in 'full action'. The first example serves to illustrate the modelling for non-normally distributed traits and to discuss the robustness of the traditional approach in which normality is assumed. The second example serves to illustrate modelling for multiple QTLs and to discuss some problems concerning detection of QTLs. In both examples data were simulated for 200 individuals. Genotypes were generated assuming absence of interference. The markers were set at a distance of 20 cM apart, which gives a recombination frequency of approximately 0.16 according to Haldane's mapping function (HALDANE 1919). The QTLs were located halfway between their flanking markers, which gives recombination frequencies of approximately 0.09.

Example 1. A simulated backcross example will be elaborated for the case of a single QTL with an exponentially distributed trait and F_1 individuals $M_1Q_1M_2/m_1q_2m_2$. The exponential distribution is of considerable importance and has a widespread use in the analysis of data in which the response variable is a lifetime (MCCULLAGH and NELDER 1989). The probability density function of the exponential distribution is $f(y)=\mu^{-1}\exp(-y/\mu)$, where $y\geq 0$. The mean of the exponential distribution is μ; its variance is μ^2. The mean values of the genotypes qq and Qq were set to $\mu_{qq}=10$ and $\mu_{Qq}=15$, respectively.

Table 4 shows log-likelihoods and parameter estimates for various models. A comparison of the log-likelihoods shows that the models under the correct distributional

Table 4. Example 1: a simulated backcross of F_1 individuals M_1QM_2/m_1qm_2 to the parent m_1qm_2/m_1qm_2 with an exponentially distributed trait; log-likelihood and parameter estimates for various models are presented

QTL fitted (yes/no)	Exponential or normal distribution assumed (e/n)	Log-likelihood	Genetic linkage[a]		Regression of phenotype on genotype[b]	
			$\hat{r}_1$	$\hat{r}_2$	$\hat{\mu}_{qq}$	$\hat{\mu}_{Qq}$
n	n	-794.8	-	-	-	-
y	n	-792.1	0.02	0.17	11.2	16.7
n	e	-685.3	-	-	-	-
y	e	-681.7	0.04	0.16	11.0	16.9

The parameter values used to simulate the data were r_1=r_2=0.09, μ_{qq}=10, μ_{Qq}=15 and n=200 individuals.

[a] r_1 and r_2 denotethe recombination frequencies between the QTL and its flanking markers
[b] μ_{Qq} and μ_{qq} denote the mean value of individuals with Qq and qq at the QTL, respectively

assumption fit much better than the models under the false distributional assumption do. Parameter estimates under both the correct and the false assumption are still much the same. Detection of a single QTL is usually based on the LOD-score $^{10}\log\mathcal{L}_1$-$^{10}\log\mathcal{L}_0$, or on the deviance 2($\log\mathcal{L}_1$ -$\log\mathcal{L}_0$), where $\mathcal{L}_1$ and $\mathcal{L}_0$ are the likelihoods of the models with and without a QTL, respectively (KNAPP et al. 1990). However, distributional properties of the test statistic are not completely known due to failure of the regularity conditions (cf. MCLACHLAN and BASFORD 1988; TITTERINGTON et al. 1985). In our example the values of the test statistic 2($\log\mathcal{L}_1$-$\log\mathcal{L}_0$) are 5.4 and 7.2 under the assumptions that the distribution is normal and exponential, respectively. Using the threshold $\chi^2_{2,0.95}$=5.99 as a rule of thumb (KNAPP et al. 1990), the QTL will be detected only under the correct distributional assumption.

Example 2. A simulated backcross example will be elaborated now for the case of two QTLs in adjacent intervals with three markers and F_1 genotypes $M_1Q_1M_2q_2M_3/m_1q_1m_2Q_2m_3$. Note that the genes at the QTLs are in repulsion phase. Let λ_1, ν_1 and ζ_1 denote the genetic linkage parameters for the first QTL, and similarly λ_2, ν_2 and ζ_2 those for the second QTL. The environmental contribution was normally distributed with unit variance (σ^2=1). The effects of the genes at the QTLs were additive (G=m+A_1+A_2) and set to one unit (a_1=a_2=1). As an example, coefficients of the parameters are presented in Table 5 for an individual with observed marker genotype M_1m_1 M_2m_2 m_3m_3. Coefficients for individuals with other allelic constitutions at the

Table 5. Example 2: a simulated backcross of F_1 individuals $M_1Q_1M_2q_2M_3/m_1q_1m_2Q_2m_3$ to the parent $m_1q_1m_2q_2m_3/m_1q_1m_2q_2m_3$ with a normally distributed trait.

Coefficients of the genetic linkage parameters and the parameters for regression of phenotype on genotype for an individual with observed marker genotype $M_1m_1M_2m_2M_3m_3$

Unobserved complete allelic constitution at the QTL	Genetic linkage[a]						Regression of phenotope on genotype[b]		
	λ_1	ν_1	ζ_1	λ_2	ν_2	ζ_2	m	a_1	a_2
Q_1Q_2/q_1q_2	1	0	0	1	0	0	1	0	0
Q_1q_2/q_1q_2	1	0	0	1	1	1	1	0	-1
q_1Q_2/q_1q_2	1	1	1	0	0	0	1	-1	0
q_1q_2/q_1q_2	1	1	1	1	1	1	1	-1	-1

Log-likelihood and parameters estimates for various models

QTL fitted (yes/no)		Log-likelihood	Genetic linkage[a]				Regression of phenotype on genotype[b]		
1	2		$\hat{r}_{11}$	$\hat{r}_{12}$	$\hat{r}_{21}$	$\hat{r}_{22}$	$\hat{a}_1$	$\hat{a}_2$	$\hat{\sigma}^2$
n	n	-477.2	-	-	-	-	-	-	1.2
n	y	-476.1	-	-	0.16	0.00	-	0.2	1.2
y	n	-473.9	0.00	0.16	-	-	0.4	-	1.2
y	y	-467.0	0.06	0.12	0.13	0.04	1.0	0.9	1.0

Log-likelihood and parameters estimates for various models with markers as covariables

QTL fitted (yes/no)		Marker fitted (yes/no)			Log-likelihood	Genetic linkage[a]				Regression of phenotype on genotype[b]		
1	2	1	2	3		$\hat{r}_{11}$	$\hat{r}_{12}$	$\hat{r}_{21}$	$\hat{r}_{22}$	$\hat{a}_1$	$\hat{a}_2$	$\hat{\sigma}^2$
n	n	y	n	n	-476.1	-	-	-	-	-	-	1.2
n	n	n	n	y	-473.8	-	-	-	-	-	-	1.2
n	y	y	n	n	-467.7	-	-	0.16	0.00	-	0.6	1.1
y	n	n	n	y	-467.0	0.05	0.13	-	-	0.9	-	1.1

The effects of the QTL's were additive ($G=m+A_1+A_2$). The parameters values used to simulate the data were $r_{11}=r_{12}=r_{21}=r_{22}=0.09$, $a_1=a_2=1$, $\sigma^2=1$ and $n=200$ individuals.

[a] r_{12} denote the recombination frequencies between the first QTL and its flanking markers; r_{21} and r_{22} denote the recombinaton frequencies between the second QTL and its flanking markers; λ_1, v_1, and ζ_1 denote the genetic linkage parameters for the first QTL: $\lambda_1=\log(½(1-r_{11})(1-r_{12}))$; $v_1=\log(r_{11})-\log(1-r_{11})$;$\zeta_1=\log(r_{12})-\log(1-r_{12})$; λ_2, v_2 and ζ_2 denote the genetic linkage parameters for the second QTL: $\lambda_2=\log(½(1-r_{21})(1-r_{22}))$; $v_2=\log(1-r_{21})$; $\zeta_2=\log(r_{22})-\log(1-r_{22})$

[b] m, a_1 and a_2 denote the parameters for regression of phenotype on genotype: m is the mean of the expected phenotypes of individuals with $Q_1Q_1Q_2Q_2$ and $q_1q_1q_2q_2$ at the QTL, respectively; a_1 and a_2 are the additive effects of the first and second QTL, respectively; σ^2 denotes the variance of the fitted normal distribution

marker loci may be derived easily by using Table 2. Since the allelic constitution at the QTL can be Q_1Q_2/q_1q_2, Q_1q_2/q_1q_2, q_1Q_2/q_1q_2 or q_1q_2/q_1q_2, the phenotype follows a mixture of four distributions.

Models were fitted with the markers at the known distance of 20 cM. Table 5 shows log-likelihoods and parameter estimates for various models. In this example the values of the test statistic $2(\log\mathcal{L}_1\text{-}\log\mathcal{L}_0)$ are 6.6 and 2.2 for the first and second QTL, respectively. Using again the threshold $\chi^2_{2,0.95}$=5.99 as a rule of thumb, only the first QTL will be detected. However, estimates of the location of the QTLs on the linkage map and estimates of the QTL effects are highly biased. Deviances between the 'true' model (in which the two QTLs are fitted simultaneously) and the two single QTL models (in which a single QTL is fitted at a time) are large (18.2 and 13.8). This suggests that the detection procedure may be improved by testing models versus the true model instead of versus a 'no-QTL' model. However, in real applications the true model is unkown.

An adaptive procedure is to fit a single QTL at a time by using its flanking markers, and to incorporate the remaining marker as covariable into the linear model for the response variable. Table 5 shows log-likelihoods for the two single QTL models with marker covariables. It demonstrates that the likelihoods (-467.7 and -467.0) are now very close to the likelihood of the true model in which the two QTLs are fitted simultaneously (-467.0). The parameter estimates are much better than in the two single QTL models without using marker covariables. Note that the likelihoods of the 'no-QTL' models with marker covariables (-476.1 and -473.8) are also very close to the likelihoods of the single QTL models without using marker covariables (-476.1 and -473.9).

DISCUSSION

In this paper a general and flexible mixture model is developed for mapping QTLs by using molecular markers. The computational idea is that, by adopting the EM algorithm for parameter estimation, the mixture problem can be split into two solvable non-mixture problems, one concerning genetic linkage analysis, the other concerning regression of phenotype on genotype. Moreover, by using generalized linear models a framework is provided covering regression techniques for many types of data. More accurate and efficient mapping of QTLs can be achieved by these procedures, which are extensions of methods proposed by LANDER and BOTSTEIN (1989) and KNAPP et al. (1990). The computational work can be done by statistical packages having facilities for generalized linear models, such as GENSTAT (GENSTAT 5 COMMITTEE 1987).

The included examples illustrate the generality and flexibility of the described mixture model. For the sake of brevity other examples, such as modelling experimental design factors or modelling of epistatic QTLs, have not been included. It will be obvious that these are easily dealt with.

Testing for the number of components in a mixture is an important and difficult problem which has not been resolved completely (cf. McLachlan and Basford 1988; Titterington et al. 1985). As suggested by our second example, the procedure for detection of QTLs may be improved by testing versus a polygenic model instead of testing versus a 'no-QTL' model. One strategy could be to use a hypothetical polygenic model, e.g. a dense map of QTLs at distances of 20 cM. However, there will be problems of model selection as in multiple regression, and computational problems to cope with. Important work still has to be done to develop adaptive detection procedures and to study their behaviour for various situations in the QTL-mapping case. An adaptive detection procedure might be to fit a single QTL at a time (or two or more QTLs simultaneously) by using flanking markers, and to incorporate the remaining markers as covariables into the regression model of phenotype on genotype. This procedure shows promise, as was suggested in the second example.

The robustness of the method against deviations from the model assumptions also needs further consideration. In the first example it was shown that (at least) complications in testing may arise when the underlying phenotypic component distributions are non-normal, whereas normality is assumed. In such cases a transformation analysis should be carried out to find a suitable transformation such that the normality assumption holds. Alternatively, mixtures of other types of distribution should be used (McCullagh and Nelder 1989).

LITERATURE CITED

Bailey, N.T.J., 1961 Introduction to the mathematical theory of genetic linkage. Oxford Univ Press, London

Bulmer, M.G., 1985 The mathematical theory of quantitative genetics. Clarendon Press, London

Dempster, A.P., N.M. Laird and D.B. Rubin, 1977 Maximum likelihood from incomplete data via the EM-algorithm. JR Statist Soc B, 39:1-38

Genstat 5 Committee, 1987 Genstat 5 reference manual. Clarendon Press, Oxford

Haldane, J.B.S., 1919 The combination of linkage values, and the calculation of distance between the loci of linked factors. J Genet 8:299-309

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

Knapp, S.J., W.C. Bridges and 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 imbred, and doubled haploid progeny. Theor Appl Genet 81:333-338

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

Luo Z.W. and M.J. Kearsey, 1989 Maximum likelihood estimation of linkage between a marker gene and a quantitative trait locus. II Application to backcross and doubled haploid populations. Heredity 66:117-124

McCullagh P., and J.A. Nelder, 1989 Generalized linear models. Monographs on statistics and applied probability 37, Chapman and Hall, London

McLachlan G.J., and K.E. Basford, 1988 Mixture models: inference and applications to clustering. Marcel

Dekker, New York

PATERSON, A.H., E.S. LANDER, J.D. HEWITT, S. PETERSON, S.E. LINCOLN, and S.D. TANKSLEY, 1988 Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. Nature 335:721-726

SOLLER, M., and J.S. BECKMANN, 1983 Genetic polymorphism in varietal identification and genetic improvement. Theor Appl Genet 47:179-190

TITTERINGTON, D.M., A.F.M. SMITH, and U.E. MAKOV, 1985 Statistical analysis of finite mixture distributions. Wiley, New York

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

III. INTERVAL MAPPING OF MULTIPLE QUANTITATIVE TRAIT LOCI

ABSTRACT

The interval mapping method is widely used for the mapping of quantitative trait loci (QTLs) in segregating generations derived from crosses between inbred lines. The efficiency of detecting and the accuracy of mapping multiple QTLs by using genetic markers are much increased by employing multiple QTL models instead of the single QTL models (and no QTL models) used in interval mapping. However, the computational work involved with multiple QTL models is considerable when the number of QTLs is large. In this paper it is proposed to combine multiple linear regression methods with conventional interval mapping. This is achieved by fitting one QTL at a time in a given interval and simultaneously using (part of) the markers as cofactors to eliminate the effects of additional QTLs. It is shown that the proposed method combines the easy computation of the single QTL interval mapping method with much of the efficiency and accuracy of multiple QTL models.

INTRODUCTION

Conventional methods for the detection of quantitative trait loci (QTLs) are based on a comparison of single QTL models with a model assuming no QTL. For instance in the 'interval mapping' method (LANDER and BOTSTEIN 1989) the likelihood for a single putative QTL is assessed at each location on the genome. However, QTLs located elsewhere on the genome can have an interfering effect. As a consequence, the power of detection may be compromised, and the estimates of locations and effects of QTLs may be biased (LANDER and BOTSTEIN 1989; KNAPP 1991). Even non-existing so-called 'ghost' QTLs may appear (HALEY and KNOTT 1992; MARTINEZ and CURNOW 1992). Therefore, it is obvious that multiple QTLs could be mapped more efficiently and more accurately by using multiple QTL models. KNAPP (1991), HALEY and KNOTT (1992) and MARTINEZ and CURNOW (1992) developed approximate methods for mapping QTLs using the information in the expected values of marker genotype means. JANSEN (1992) described a general mixture model for the case of multiple QTLs. Unfortunately, the computation involved with all these methods is almost infeasible when the number of QTLs is large. Also, standard multiple linear regression procedures are used in mapping QTLs (COWEN 1989, STAM 1991). The regression method is available in many statistical packages, but suffers from the relative lack of interpretability in terms of genetical models. In these standard multiple linear regression procedures the quantitative trait is regressed on the markers, so that all markers are treated as if they are QTLs themselves. The effects of QTLs will be absorbed (partially) by linked markers. STAM (1991) showed that in a backcross population of infinite size QTL effects are fully absorbed by their flanking markers when these are used as regressors. Although this will rarely be the case in finite populations (due to random deviations from the theoretical cosegregation ratio of markers), flanking markers will tend to absorb the effects of nearby QTLs. JANSEN (1992) suggested a detection and mapping approach that is basically a hybrid between the interval method and the multiple regression method. It was proposed to fit single QTL models (one per marker interval) and use (selected) markers to eliminate the effects of possible QTLs in other intervals. This can be achieved by using markers as cofactors in the regression of phenotype on genotype. Again, single QTL models may be compared with the model assuming no QTL, but now markers are used as cofactors. In the present paper this hybrid approach is worked out and illustrated for backcross populations, but the same ideas apply to other types of population; emphasis will be on detection aspects. A simple simulation study with three QTLs, two of them located on the same chromosome, is presented to illustrate the potential use of marker cofactors in the detection of multiple QTLs. A simulated example concerning detection of 11 QTLs on a genome of 10 chromosomes is also included.

SOME PRELIMINARY INVESTIGATIONS

A genome of two chromosomes was simulated 100 times in a backcross of F_1-individuals to one of the parental lines with two markers (M) and a single QTL (Q_1) on the first chromosome (MQ_1M/mq_1m), and with three markers and two other QTLs (Q_2 and Q_3) on the second chromosome (MQ_2MQ_3M/mq_2mq_3m). The markers were set at a distance of 20 cM apart. The QTLs were located halfway between between their flanking markers. The environmental contribution was normally distributed. The effects of the genes at the QTLs were additive; the additive deviations (half the differences between the homozygotes) were set to one standard deviation. In all simulations data were generated for 200 individuals assuming absence of interference.

Tables 1 and 2 show the specification of the various models which were fitted to the simulated data. Expressions for the simultaneous likelihood of the observed phenotypic and genotypic (marker) data are given by JANSEN (1992). Let $\mathcal{L}_A$, $\mathcal{L}_B$, $\mathcal{L}_C$, $\mathcal{L}_D$, $\mathcal{L}_{A'}$ and $\mathcal{L}_{B'}$ denote the maximum log-likelihoods of the corresponding models. For instance, $\mathcal{L}_A$ can be written as follows

$$\mathcal{L}_A=\sum_{i=1}^{N}\log P(h)+\sum_{i=1}^{N}\log\left[\frac{1}{\sqrt{2\pi\sigma^2}}\sum_{g}P(g\,|\,h)\exp(-\frac{(y-m(g))^2}{2\sigma^2})\right],$$

where y is the phenotype, h is the observed marker genotype with probability $P(h)$, g is

Table 1. Outline of the models fitted to compare different strategies for detection of QTL 1 (see also Figure 1)

Model	QTL fitted			Marker co-factors fitted
	1	2	3	
A	Yes	Yes	Yes	No
B	No	Yes	Yes	No
C	Yes	No	No	No
D	No	No	No	No
A'	Yes	No	No	Flanking markers of QTLs 2 and 3
B'	No	No	No	

Table 2. Outline of the models fitted to compare different strategies for detection of QTL 2 (see also Figure 2)

Model	QTL fitted			Marker co-factors fitted
	1	2	3	
A	Yes	Yes	Yes	No
B	Yes	No	Yes	No
C	No	Yes	No	No
D	No	No	No	No
A'	No	Yes	No	Flanking markers of QTLs 1 and 3
B'	No	No	No	

the complete genotype (markers and QTLs) with conditional probability P(g|h), $m(g)$ is the normal mean and σ^2 the normal variance of individuals with genotype g. In general $m(g)$=m+A+M, where A is now the additive component of the QTLs and M is the component for the marker cofactors with two levels per marker.

In conventional interval mapping, the detection of a QTL is based on $\mathcal{L}_C$-$\mathcal{L}_D$. $\mathcal{L}_A$-$\mathcal{L}_B$ is a similar expression, but now the QTLs on the other chromosome are also accounted for. Figure 1a shows that $\mathcal{L}_C$-$\mathcal{L}_D$ is less than $\mathcal{L}_A$-$\mathcal{L}_B$ in almost all simulations. Thus higher power for detection of QTL 1 is achieved when taking the QTLs on chromosome 2 into account. Figure 1b shows that taking the QTLs on chromosome 2 into account by using multiple QTL models results in about the same power as using the markers of chromosome 2 as cofactors. Contrary to Figure 1a, Figure 2a shows now that $\mathcal{L}_C$-$\mathcal{L}_D$ exceeds $\mathcal{L}_A$-$\mathcal{L}_B$ in all simulations. The single QTL model (model C) now absorbs the simultaneous effect of QTL 2 and 3, so that $\mathcal{L}_C$-$\mathcal{L}_D$ represents approximately the

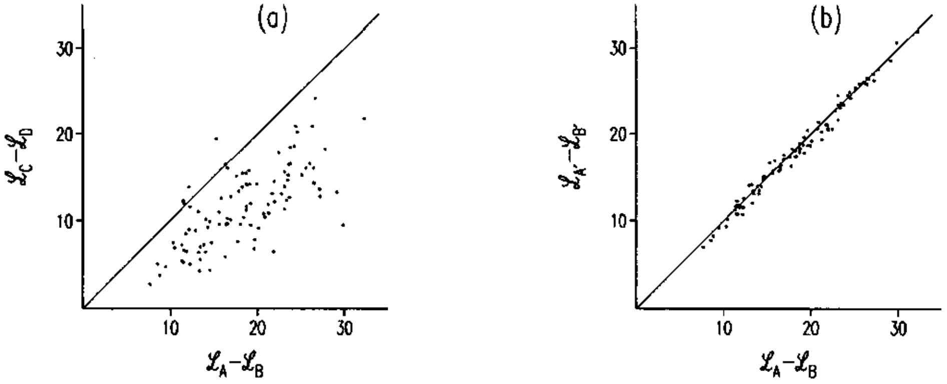


Figure 1. A comparison of strategies for detection of QTL 1 (see also Table 1). A simulated backcross of F_1 individuals with MQ_1M/mq_1m on the first chromosome and MQ_2MQ_3M/mq_2mq_3m on the second chromosome to the parent with a normally distributed trait. The effects of the QTLs were additive and set to one standard deviation. All distances between QTLs and flanking markers were set to 10 cM. $\mathcal{L}_A$, $\mathcal{L}_B$, $\mathcal{L}_C$, $\mathcal{L}_D$, $\mathcal{L}_{A'}$ and $\mathcal{L}_{B'}$ denote the maximum log-likelihoods of the corresponding models (Table 1). Differences $\mathcal{L}_A$-$\mathcal{L}_B$, $\mathcal{L}_C$-$\mathcal{L}_D$ and $\mathcal{L}_{A'}$-$\mathcal{L}_{B'}$ represent the contribution of QTL 1 to the log-likelihood.
(a) Interval mapping approach versus a multiple QTL approach. (b) Interval mapping approach using marker cofactors versus a multiple QTL approach.

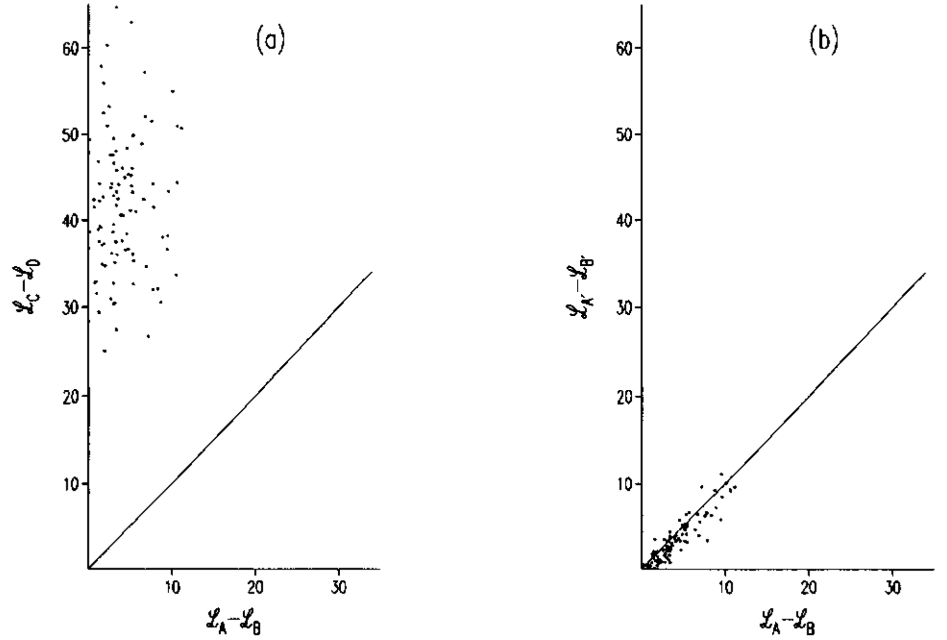


Figure 2. A comparison of strategies for detection of QTL 2 (see also Table 2). A simulated backcross of F_1 individuals with MQ_1M/mq_1m on the first chromosome and MQ_2MQ_3M/mq_2mq_3m on the second chromosome to the parent. Differences $\mathcal{L}_A$-$\mathcal{L}_B$, $\mathcal{L}_C$-$\mathcal{L}_D$ and $\mathcal{L}_{A'}$-$\mathcal{L}_{B'}$ represent the contribution of QTL 2 to the log-likelihood, otherwise as Figure 1.

simultaneous contribution of QTLs 2 and 3, while $\mathscr{L}_A$-$\mathscr{L}_B$ represents the contribution of an additional QTL in the model. The fact that $\mathscr{L}_C$-$\mathscr{L}_D$ exceeds $\mathscr{L}_A$-$\mathscr{L}_B$ therefore indicates the possible presence of multiple QTLs on chromosome 2. Figure 2b shows that again $\mathscr{L}_A$-$\mathscr{L}_B$ may be well approximated by $\mathscr{L}_{A'}$-$\mathscr{L}_{B'}$ using the flanking markers of QTLs 1 and 3 as cofactors.

In this example the 'saturated' multiple QTL model still involves only three QTLs, which can be dealt with satisfactorily in terms of computational efforts. However, when the number of QTLs to be fitted simultaneously increases, the computational complexity quickly becomes prohibitive. Though representing a simple situation, the example clearly demonstrates the following points. First, searching for one QTL at a time by using markers as cofactors to absorb the effects of additional QTLs is (approximately) as powerful as searching for QTLs by dropping a single QTL from the full multiple QTL model. Second, the comparison of (a) the difference between the full multiple QTL model and one from which a single QTL is dropped, and (b) the difference between the conventional single and no-QTL model, is indicative of the presence of multiple QTLs on the same chromosome. In the next sections these ideas are extended to a general strategy for the detection of multiple QTLs.

A GENERAL STRATEGY FOR THE DETECTION OF MULTIPLE QTLS

The log-likelihoods of various models when maximized over unknown parameters provide a basis for choosing the genetic model that best fits the data: the genetic model which gives rise to the largest likelihood is the best fitting one. However, it is clear that, for instance, by adding an extra QTL or an extra marker cofactor to the model, the likelihood will increase. To allow for the fact that different genetic models depend on different numbers of parameters, we choose the genetic model that leads to the largest value of the log-likelihood ($\mathscr{L}$) minus a penalty for the number of free parameters (k) in the model. Equivalently, Akaike's Information Criterion (AIC)

$$\text{AIC} = -2(\mathscr{L} - k)$$

may be minimized (SAKAMOTO, ISHIGURO and KITAGAWA 1986). If the difference between AICs for two models is larger than 2, then the difference is considered to be significant (SAKAMOTO, ISHIGURO and KITAGAWA 1986). A single QTL model with the QTL located at a marker position is equivalent to the model with that specific marker as cofactor, i.e. that marker is also considered to represent a QTL. We impose no penalty on the AIC for the additional recombination parameter in a single QTL model. Then the above single QTL model and marker cofactor model have the same AIC. Our detection procedure consists of two stages: (1) selection of markers located closely to QTLs, and (2) interval mapping using (subsets of) the selected markers to absorb effects of other QTLs. An example in the next section serves to illustrate the procedure. Figure 3 shows a flow diagram for the detection procedure; the details will now be described below.

The first stage starts with multiple regression of the quantitative trait on all markers. By a subset selection method for multiple regression (the method of backward elimination) markers are dropped from the model until no further reduction in AIC can be achieved. The final model is denoted by B_1. The final subset of markers will be used in the second stage of the procedure. Models A_1, A_2, A_3, B_1, B_2, B_3, C and D as used below, now refer to the models specified in Figure 3 and in the example given in the next section. Models A_1, A_2 and A_3 are single QTL models, models B_1, B_2 and B_3 are 'no QTL' models. Models C and D are the commonly used models with and without a single

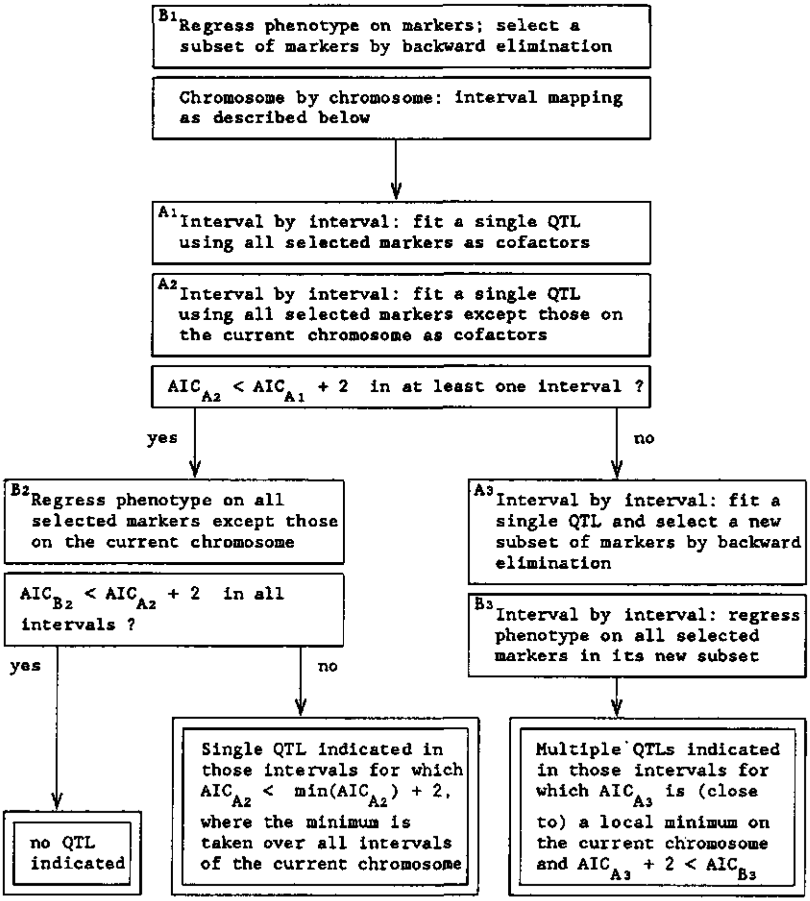

B1 Regress phenotype on markers; select a subset of markers by backward elimination
Chromosome by chromosome: interval mapping as described below
A1 Interval by interval: fit a single QTL using all selected markers as cofactors
A2 Interval by interval: fit a single QTL using all selected markers except those on the current chromosome as cofactors
$AIC_{A2} < AIC_{A1} + 2$ in at least one interval ?
yes
no
B2 Regress phenotype on all selected markers except those on the current chromosome
$AIC_{B2} < AIC_{A2} + 2$ in all intervals ?
A3 Interval by interval: fit a single QTL and select a new subset of markers by backward elimination
B3 Interval by interval: regress phenotype on all selected markers in its new subset
yes
no
no QTL indicated
Single QTL indicated in those intervals for which $AIC_{A2} < \min(AIC_{A2}) + 2$, where the minimum is taken over all intervals of the current chromosome
Multiple QTLs indicated in those intervals for which AIC_{A3} is (close to) a local minimum on the current chromosome and $AIC_{A3} + 2 < AIC_{B3}$

Figure 3. Flow diagram for interval mapping of multiple QTLs.

QTL (no marker cofactors used). It is the difference among AICs that matters and not the actual values themselves. Therefore, we present AICs relative to the AIC of the multiple regression model using all selected markers (model B_1).

Selected markers (hopefully) indicate locations of QTLs or at least regions where QTLs are located. Important QTLs are located on those chromosomes for which the dropping of markers from the final multiple regression model results in a large increase of AIC (model B_2 is compared with model B_1).

In the second stage (the interval mapping stage) selected markers are used as cofactors in the regression of phenotype on genotype (JANSEN 1992). Interval by interval, the AICs of several models are calculated. Firstly, a single QTL model is fitted using all selected markers as cofactors (model A_1). However, by fitting the putative QTL, some (or all) of the selected markers on the current chromosome may now be redundant. This may be studied by dropping some or even all selected markers on the current chromosome (model A_1 is compared with models A_3 and A_2, respectively). Suppose that for some interval all selected markers on the current chromosome may be dropped without a loss in AIC (model A_2 fits better than model A_1). In that case a single QTL suffices to take over the role of these markers. However, if the putative QTL may also be dropped (model B_2 fits better than model A_2) without a loss in AIC, then no QTL is detected on the current chromosome. Alternatively, suppose that the AIC of model A_2 exceeds the AIC of model A_1 in all intervals. This indicates that a single QTL cannot take over the role of the markers, and the presence of multiple QTLs on the current chromosome is indicated. Then a second selection procedure is carried out interval by interval. Starting from the single QTL model using all selected markers, it is studied which markers still may be dropped (those markers previously explained the effect of the putative QTL), and which markers cannot be dropped (these markers possibly absorb the effects of other QTLs on the current chromosome). Dropped markers will often be the markers flanking the interval. Interval by interval, detection of the putative QTLs is now carried out by dropping the QTL (model B_3 is compared with model A_3) using for each interval its own subset of selected markers.

EXAMPLE

A simulated backcross example will be worked out in the case of a genome of 10 chromosomes and a quantitative trait which is affected by 11 QTLs spread over the chromosomes (Figure 4). The example serves to illustrate the behaviour of our new interval mapping approach and to compare this approach with the traditional interval mapping method. Data were simulated for 500 individuals. Genotypes were generated assuming absence of interference. The markers were set at a distance of 20 cM apart,

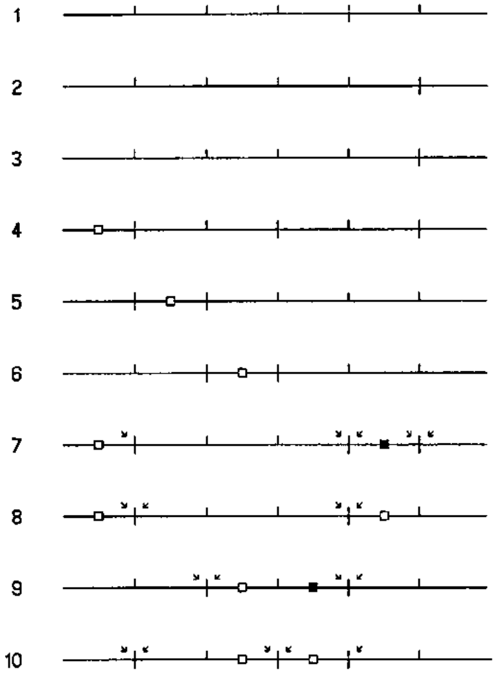


Figure 4. A simulated backcross of 500 individuals with a genome of 10 chromosomes and 11 QTLs spread over the chromosomes. The markers were set at a distance of 20 cM apart. The QTLs were located halfway between their flanking markers. Lines indicate chromosomes. Blocks indicate QTL positions, the effect of a QTL is either 1 (□) or -1 (■). Marker locations are indicated by a ⊥ and +. A subset of all markers was selected by backward elimination in multiple linear regression of the trait on the markers; selected markers are indicated by +, the other markers by ⊥. Arrows indicate per marker interval when the left (↙) or the right (↘) flanking marker is dropped in the second marker subset selection (see also text, Table 3 and Figure 3). Markers are numbered 1 to 5 from the left to the right on each chromosome.

Table 3. Outline of the models fitted in the example

		Selected marker cofactors used on this/other chromosome(s)	
Model	QTL fitted	This	Other
A_1	Yes	All	All
B_1	No	All	All
A_2	Yes	None	All
B_2	No	None	All
A_3	Yes	See Figure 4	All
B_3	No	See Figure 4	All
C	Yes	None	None
D	No	None	None

the QTLs were located halfway between their flanking markers. The environmental contribution was normally distributed. The effects of the genes were additive and the additive deviations were set to either 1 or -1 standard deviation. Expressions for the simultaneous likelihood of the observed phenotypic and genotypic (marker) data are given by JANSEN (1992).

Table 3 shows the specification of the various models fitted. Some results are presented in Table 4 for chromosomes 1, 6, 7, 8, 9 and 10; emphasis is on detection aspects. Results for other chromosomes were similar. A complete overview of the results would contain not only AICs as given in Table 4, but also a monitoring of AICs and parameter estimates during the whole detection process and for the complete genome.

The total phenotypic variance in the simulated data was equal to 3.76, which consisted of environmental variance (1.02) and genotypic variance (2.74). The explained (genotypic) variance was in the range of 0.00 to 0.74 when using single QTL models (conventional interval mapping), and in the range of 2.18 to 2.36 when using single QTL models with marker cofactors. This clearly demonstrates that a considerable part of the genotypic (QTL) variance was absorbed by marker cofactors. Table 4 shows that the AICs of single QTL models (conventional interval mapping) were large relative to the single QTL models with marker cofactors. Thus, the better fit of the model to the data is achieved when using marker cofactors.

The procedure indicates correctly the presence of no QTL on chromosomes 1, 2 and 3, the presence of one QTL on chromosomes 4, 5 and 6, and the presence of multiple QTLs on chromosomes 7, 8, 9 and 10. The multiple QTLs could be well separated on chromosomes 7 and 8. The estimates of the QTL effects on chromosome 9 take values -0.06, 0.36, 0.45, -0.67, -0.67 and 0.01 in the first up to the sixth interval (model A_3), which shows a clear changeover at the third marker. No clear separation of the two QTLs on chromosome 10 could be obtained. The estimates of the QTL effects on chromosome 10 are 1.27, 1.20, 1.63, 1.60, 1.11 and 0.94 in the first up to the sixth interval when using model A_2, and they are 0.44, 0.38, 1.24, 1.22, 0.30 and 0.05 when using model A_3. This change clearly represents the effect of the marker cofactors. However, the selected marker 1 was not replaced by the single QTL in the third interval. Nevertheless, the AIC for the model without marker 1 was close to the AIC for the given optimum model. Similar results hold for the selected marker 4 and the single QTL in the fourth interval. Thus discrimination between the various models was poor.

The choice of a genetic model may also be based on additional considerations. For instance, a QTL is indicated in the fourth, fifth or sixth interval on chromosome 7, but the markers 4 and 5 are simultaneously redundant as cofactors only when fitting a QTL in the fifth interval (model A_3). Therefore, only a QTL in the fifth interval can take over the role of these markers. It may also be worthwhile to force a marker cofactor to be

Table 4. Interval mapping multiple QTLs: AIC values for various models in a simulated backcross (see Figure 4 for a description of the backcross and see Table 3 and Figure 3 for an outline of the models A_1, A_2, A_3, B_1, B_2, B_3, C and D)

Chromosome	Model	Marker interval					
		-1	1-2	2-3	3-4	4-5	5-
1	A_1	1.9	2.0	1.9	1.9	1.5	1.5
(0.6)	A_2	2.5	2.2	1.1	0.0	0.0	2.2
	C	391.5	390.9	385.7	385.6	391.8	392.3
6	A_1	2.0	2.0	2.0	2.0	1.8	1.6
(43.4)	A_2	28.3	2.5	-1.4	11.1	30.4	35.4
	C	385.0	377.4	374.0	377.6	388.0	389.0
7	A_1	2.0	2.0	1.0	0.0	2.0	2.0
(82.8)	A_2	60.2	60.3	79.4	61.5	54.7	57.4
	A_3	-2.7	-3.5	0.5	-2.2	0.0	0.0
	B_3	55.9	0.0	0.0	8.9	60.2	10.9
	C	371.1	371.3	392.3	381.0	374.7	375.8
8	A_1	1.5	1.9	1.4	1.1	-0.7	0.6
(108.0)	A_2	43.6	42.6	53.8	41.0	41.3	69.4
	A_3	-0.3	0.0	1.4	0.0	-2.5	0.6
	B_3	42.4	42.4	0.0	43.5	43.5	0.0
	C	346.5	341.7	346.9	348.4	349.3	369.3
9	A_1	1.9	1.9	1.4	1.5	2.0	2.0
(21.9)	A_2	23.9	23.9	22.1	5.3	5.7	16.9
	A_3	1.9	0.0	-0.5	0.0	0.0	2.0
	B_3	0.0	5.3	5.3	23.9	23.9	0.0
	C	392.2	391.7	392.0	389.3	389.0	391.2
10	A_1	2.0	2.0	2.0	2.0	1.9	1.9
(157.3)	A_2	106.9	79.4	5.2	7.0	87.7	125.0
	A_3	0.0	0.0	0.0	0.0	0.0	1.9
	B_3	6.1	6.1	55.0	55.0	2.1	0.0
	C	349.5	333.9	297.0	297.6	340.3	363.2

A subset of markers was selected by backward elimination in multiple linear regression of the trait on the markers. Only selected markers (or subsets of selected markers) were used as cofactors. All AICs are relative to the AIC for the multiple regression of the trait on the selected markers (model B_1). AICs for model B_2 are printed between brackets below the chromosome number, the AIC for model D equals 390.3.

included in the selected subset. For instance, when observing the changeover in the estimated QTL effects at marker 3 of chromosome 9, a marker cofactor for marker 3 can be reentered into the model. The estimates of the QTL effects on chromosome 9 now take the slightly better values 0.52 and -0.77 in the third and the fourth interval, respectively (model A_3).

In the conventional interval mapping approach the detection and mapping of QTLs is based on models with a single QTL (model C) and without a single QTL (model D). The AIC of the latter model is equal to 390.3. A LOD threshold of about 2.4 (LANDER and BOTSTEIN 1989; their Figure 4), or equivalently, an AIC threshold of about $2(2.4/\log_{10}e - 1) \approx 9.1$ is commonly used as a threshold for the detection of QTLs. A QTL would be indicated then in those intervals for which the AIC of model C is less than 381.2. Following this approach, QTLs would be indicated on all chromosomes but chromosome 9. The putative QTL on chromosome 8 is most likely (but incorrectly) located in the second interval.

The example clearly demonstrates the following points. Firstly, the AIC profile is much steeper around QTLs when using model A_2 instead of when using model C (see for instance chromosome 6 in Table 4). Therefore, the locations of the QTLs can be assessed more accurately when using marker cofactors. Secondly, the difference for AIC between model A_2 and model B_2 is often (much) larger than the difference for AIC between model C and model D in case a single QTL is indicated on a specific chromosome (see chromosome 6, results were similar for chromosomes 4 and 5). This difference is indicative for the effect of dropping the QTL, so that detection is more powerful when using marker cofactors. Finally, contrary to our method, conventional interval mapping does not indicate the presence of multiple QTLs on chromosomes 8, 9 and 10. In conclusion, the example demonstrates that more efficient detection and more accurate mapping can be achieved by the interval mapping approach proposed here than by conventional interval mapping.

DISCUSSION

Detection of multiple QTLs is hampered by two main problems. First, though exact models for mapping multiple QTLs can be formulated (JANSEN 1992), the computational work involved is almost infeasible for large numbers of QTLs. Second, many genetic models have to be compared, so that problems of model selection arise. In the present paper an approach is developed to get around these problems. In this approach only single QTL models are used, while effects of other QTLs are (hopefully well) eliminated by their flanking markers. A small simulation study demonstrated the usefulness of this approach for the detection of multiple QTLs. The Akaike Information Criterion (AIC) is used to evaluate the goodness of the assumed models (SAKAMOTO, ISHIGURO and KITAGAWA 1986). A model which minimizes the AIC, or models for which the AIC is close to the minimum, are considered to be the most appropriate. This procedure shows promise, as is suggested by the example: the results indicate that more efficient detection and more accurate mapping can be achieved by using our approach than by using the conventional single QTL interval mapping approach. However it should be noted that, even when it is detected that a specific chromosome contains multiple QTLs, large data sets may still be required to unravel the separate effects of closely linked QTLs.

Conventional interval mapping starts with a 'no QTL' model and compares this model with a single QTL model. The test statistic shows the improvement in fitting a single QTL over fitting no QTL. If the improvement is significant, a second test may be carried out and the test statistic shows the improvement in fitting two QTLs over a single QTL, and so on. However, the first test may not be significant due for instance to linked genes with opposite effects or to unaccounted segregation on other QTLs. In conventional interval mapping the error of 'missing an existing QTL' is uncontrolled and may therefore be high. It has also been reported that non-existing 'ghost' QTLs can appear, due to interference between undetected multiple QTLs (HALEY and KNOTT 1992; MARTINEZ and CURNOW 1992). The interval mapping method proposed in this paper starts with a hypothetical 'polygenic' model to get around such detection and mapping problems concerned with interferring QTLs. This method has like multiple regression methods the advantage of controlling the error of 'missing an existing QTL'. In conventional interval mapping the probability of 'detection by error of a QTL somewhere on the genome, whereas no QTL is actually present' is controlled. However, the assumption that 'no QTL is actually present' makes no sense whenever a QTL is detected. In that case the significance level of the test is no longer known. Probabilities (and costs) of both error-types ('missing an existing QTL' or 'detecting by error a QTL') may be balanced by the researcher; he may prefer to choose an AIC-threshold with a value other than the one used here (=2) for the comparison of models. Further research has to be done to study the probabilities of both error-types under various circumstances (e.g. for different levels of heritability, different numbers of multiple QTLs, linked or unlinked QTLs, linked QTLs in repulsion or coupling phase, different population sizes and so on).

The general and flexible facilities of the mixture model approach described by JANSEN (1992) also apply to the interval mapping method proposed in this paper. For instance, it is possible to analyse non-normally distributed traits in addition to normally distributed traits, to take experimental design factors into account, or to carry out a (combined) analysis of different population types. Furthermore, the interval mapping method can readily be programmed in statistical computer packages that have facilities for generalised linear models. The observed quantitative trait and the observed marker genotypes should be specified by the user and standard output may then be produced. But a general procedure would make it also possible to specify the type of distribution for the trait or to include the experimental design. More advanced users may also be able to leave the beaten track and may try to fit alternative models. For instance, specific markers which were dropped during the process may be added again to the model. Alternatively, specific markers may now be excluded. For instance, selected markers which are located on chromosomes for which no QTL is detected may be dropped. The advanced user may also want to fit multiple QTL models, for instance two or three QTLs simultaneously on one chromosome, while taking into account additional QTLs on other chromosomes by marker cofactors. This is possibly the most accurate, efficient and still feasible way to unravel the separate effects of closely linked QTLs.

LITERATURE CITED

COWEN, N.M., 1989 *Multiple linear regression analysis of RFLP data sets used in mapping QTLs in Development and application of molecular markers to problems in plant genetics,* edited by T. HELENTJARIS and B.BURR. Cold Spring Harbor Laboratory

HALEY, C.S. and 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

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 79:583-592

LANDER, E.S., and D.BOTSTEIN, 1989 Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185-199

MARTINEZ, O., and 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

SAKAMOTO, Y., M. ISHIGURO and G. Kitagawa, 1986 *Akaike Information Criterion Statistics.* KTK Scientific Publishers, Tokyo

STAM, P., 1991 Some aspects of QTL analysis in *Proceedings of the eighth meeting of the Eucarpia section Biometrics in Plant Breeding.* BRNO, july 1991

IV. HIGH RESOLUTION OF QUANTITATIVE TRAITS INTO MULTIPLE LOCI VIA INTERVAL MAPPING

ABSTRACT

A very general method is described for multiple linear regression of a quantitative phenotype on genotype (putative QTLs and markers) in segregating generations obtained from line crosses. The method exploits two features, (*a*) the use of additional parental and F_1 data, which fixes the joint QTL effects and the environmental error, and (*b*) the use of markers as cofactors, which reduces the genetic background noise. As a result, a significant increase of QTL detection power is achieved in comparison with conventional QTL mapping. The core of the method is the completion of any missing genotypic (QTL and marker) observations, which is embedded in a general and simple EM algorithm to obtain maximum likelihood estimates of the model parameters. The method is described in detail for the analysis of an F_2 generation. Because of the generality of the approach, it is easily applicable to other generations, such as backcross progenies and recombinant inbred lines. An example is presented in which multiple QTLs for plant height in tomato are mapped in an F_2 progeny, using additional data from the parents and their F_1 progeny.

INTRODUCTION

Since the pioneering papers of WELLER (1986), LANDER and BOTSTEIN (1989) and PATERSON et al. (1988), the detection and genetic mapping of quantitative trait loci (QTLs) by using molecular markers is gaining growing attention from biometrical geneticists. A variety of genetic models and estimation procedures for QTL mapping has been proposed, some focusing on specific breeding designs. A widely applied QTL mapping method is "conventional" interval mapping, first described by LANDER and BOTSTEIN (1989) and successfully applied in a number of case studies (e.g. PATERSON et al. 1988,1991; STUBER et al. 1992). Addressing the issues of the power of detecting QTLs and the precision of QTL mapping in F_2's and backcross progenies obtained from line crosses, VAN OOIJEN (1992) showed that, generally speaking, efficient "conventional" interval mapping requires population sizes which are beyond the sizes commonly used in this type of experiment.

In interval mapping, QTLs are usually mapped one at a time, ignoring the effects of other (mapped or not yet mapped) QTLs. It is now generally recognized that simultaneous mapping of multiple QTLs is more efficient and more accurate (cf. KNAPP 1991; HALEY and KNOTT 1992). In the ideal case all genotypic variation in for example an

F_2 is explained by putative QTLs, i.e. the residual variation after fitting QTLs should be approximately equal to the phenotypic variation observed in the isogenic parents and F_1. Also the observed difference between the parents and that between each of the parents and the F_1 should ideally be explained by the joint QTL effects.

In this paper we present an approach to QTL detection and mapping which combines two important features for power improvement: (*a*) the use of markers as cofactors (as a working substitute for simultaneous mapping of multiple QTLs) and (*b*) the use of parental and F_1 data (which fixes the joint QTL effects and the environmental error). Both features tend to decompose more powerfully the phenotypic variation into genetic and environmental variation and thus improve the accuracy of QTL mapping. We present an example on plant height in tomato which demonstrates that with this method the ideal situation sketched above can even be reached with a data set of moderate size.

WELLER (1986), LANDER and BOTSTEIN (1989) and other authors have shown that a quantitative trait derives from a mixture of (normal) distributions, so that statistical methods for maximum likelihood estimation in finite (normal) mixture models can be applied. Recently it has been demonstrated that the finite mixture model can be embedded easily in the framework of multiple linear regression models, and even in that of generalized linear models (JANSEN 1992,1993a).

Estimating the effects of QTLs and also mapping of QTLs by using molecular markers can be considered as a multiple regression problem with missing genotypic data. The basic idea of our unified approach to this problem is the completion of any missing genotypic data. The formulation of multiple linear regression models or generalized linear models (GLMs) for the completed data is straightforward. Parameter estimation is carried out by iterative weighted regression. The details will be worked out in this paper for an F_2 progeny.

The phenotype can be regressed on a single QTL, on two or more QTLs simultaneously, on markers and so on. Here we follow the method described by JANSEN (1993b), which is essentially a computationally feasible alternative to simultaneous mapping of multiple QTLs. In this method the phenotype is regressed on a single putative QTL in a given marker interval and at the same time on a number of markers that serve as cofactors. The rationale behind using markers as cofactors is that these will eliminate the major part of the variation induced by QTLs located elsewhere on the genome, thus reducing the genetic background variation.

MULTIPLE LINEAR REGRESSION OF PHENOTYPE ON GENOTYPE IN AN F_2

Segregation analysis for quantitative traits and QTL mapping can be viewed as problems in which the data are incomplete: the observations of the genotypes at the quantitative trait loci are missing. Complete data models and incomplete data models for an F_2 progeny are described in the next two sections.

Genotype known. We will adopt the following notation for the genotypes at a diallelic locus: A and B denote homozygous (parental) genotypes and H denotes the heterozygote. Let us assume that the genotype at all loci affecting a quantitative trait is known. Then, assuming absence of epistatic effects, the regression model reads

$$Y = m + \sum_{i} x_{ai}\, a_i + \sum_{i} x_{di}\, d_i + E$$

where Y is the phenotypic trait, m is the mean, a_i and d_i are the additive and dominance effects of individual loci and E is the environmental error; the summation is over loci affecting the trait. The x_{ai} and x_{di} are indicator variables for the genotype; x_{ai} takes the value −1, 0 and +1 for the genotypes A, H and B, respectively; x_{di} takes values 0, 1 and 0 for A, H and B, respectively. E is generally assumed to be normally distributed.

The genotypes at QTLs are, of course, not known. However, marker loci may take over the role of QTLs. In fact, the loci in the regression model may be either a set of markers, a single QTL, multiple QTLs or any combination of markers and QTLs. In order to be able to regress on the unknown QTL genotypes, one can complete the missing QTL genotypic data. This is elaborated in the next section.

Missing genotypic observations. All genotypic data at QTLs can be viewed as missing. In practice it also occurs frequently that the observation of a molecular marker genotype fails for a number of plants, for instance due to faint bands on the autoradiogram. It is quite common that (up to) 5 percent of the marker data are missing. Apart from these fortuitously missing data, another type of missing marker data may occur in a natural way, namely when markers are dominant and the heterozygote cannot be distinguished from one of the homozygotes. Plants with any missing marker data might be eliminated from the regression, but in multiple linear regression of the trait on many markers only a very limited set of data would then remain. A general solution to the problem of missing genotypic data is to complete them in the way described below.

The basis of completing missing genotypic observations is to assign weights to the possible genotypic states at a locus for which the observation fails. These weights are conditional probabilities of the genotypic states given the observed phenotype and the observed genotypes at other (linked) loci. In this way both phenotypic and genetic linkage information is used to complete the missing genotypic observation. Having completed the data, estimates of the regression parameters are obtained by weighted regression of phenotype on the completed genotype. Repeated updating of weights, based on the current parameter estimates, followed by parameter estimation are the basic steps of an iterative EM algorithm to obtain maximum likelihood estimates.

The completion of missing genotypic observations not only applies to a putative QTL, but also to any missing marker genotype. Since both putative QTLs and markers are factors (in statistical sense), they are dealt with in exactly the same way. We will now describe in detail how phenotypic information is used; next the use of genetic linkage information is dealt with, and finally the simultaneous use of phenotypic and linkage information are discussed.

The phenotype can be used to complete missing genotypic data in the following way. Suppose, for the moment, that it is known that genotypes A, B and H at a specific locus have different mean phenotypic values, genotype A having the largest mean phenotype. An observed large phenotypic value *y* then indicates that the missing observation is most likely to be A. This could be expressed by assigning weights of, for instance, 0.6 to A, 0.3 to H and 0.1 to B. The basic idea of an iterative EM algorithm described by JANSEN (1992, 1993a) consists of the replacement of the single incomplete observation *y* by its three complete observations (*y*,A), (*y*,B) and (*y*,H), and weighting the three complete observations by specified or updated (conditional) probabilities. The conditional probability $P(A|y)$ that the missing observation has constitution A equals $P(A|y)=P(A)\cdot f(y|A)/f(y)$, where $f(y) = P(A)\cdot f(y|A)+P(B)\cdot f(y|B)+P(H)\cdot f(y|H)$, $P(A)=P(B)=¼$, $P(H)=½$ and $f(y|A)$, $f(y|B)$ and $f(y|H)$ are the probability density functions of observations with genotypes A, B and H, respectively. Similar expressions hold for $P(B|y)$ and $P(H|y)$. Generally, parameter values are unknown and their maximum likelihood estimates can be obtained iteratively by the following alternating steps:

Step 1: specify or update weights,

Step 2: update the estimates of the regression parameters by a weighted regression of phenotypes on the completed genotype.

The weights in step 1 are calculated by using the current parameter estimates. When the environmental error is assumed to be normally distributed, the updates in step 2 are

$$\hat{\beta}=(X^TWX)^{-1}X^TWY,$$

$$\hat{\sigma}^2=\frac{1}{N}(Y-X\hat{\beta})^TW(Y-X\hat{\beta}),$$

where *Y* is the complete data vector, X is the design matrix for the complete data, W is the diagonal matrix of weights, ß is the vector of regression parameters for the normal mean, σ^2 is the normal variance and *N* is the number of individuals. The algorithm is conveniently started by setting the parameters to (well-chosen) initial values. The same procedure can be used to estimate the parameters of a multiple linear regression of the trait on two or more loci. The data of a single plant are replicated three times for any missing genotypic observation (–) and completed with the three possible outcomes A, B and H, the three possibilities being properly weighted. Similarly, all data of a plant are replicated twice for incomplete observations 'non-A' or 'non-B' which occur in the case of dominance, and completed with B and H, and A and H, respectively.

Flanking loci can also be informative to complete missing genotypic data. For instance, suppose that for two adjacent loci the score is A–, which means that the observation on the second locus is missing. The observation on the neighbour locus indicates that the missing observation most likely will also be A. The single incomplete observation is replaced by its three complete observations AA, AB and AH. The conditional probability P(AA | A–) that the missing observation has constitution A equals P(AA | A–)=$(1-r)^2$, where r is the recombination frequency between the two loci. The other two conditional probabilities are P(AB | A–)=r^2 and P(AH | A–)=$2r(1-r)$. Similarly, conditional probabilities are calculated for the genotypes B and H when the missing observation is scored as non-A, or for the genotypes A and H when it is scored as non-B. These conditional probabilities can be calculated directly when the value of r is known. In practice the genetic linkage map of the markers is often fixed and a putative QTL is moved along the genetic map, so that for a given map position of the QTL all recombination frequencies are fixed. If r must be estimated from the same data an iterative procedure may be followed with the above step 1 and a new step 2:

Step 2: update the estimate of the recombination frequency based on the weights.

The APPENDIX describes how to update the estimates of recombination frequencies for an F_2. The same procedure also applies to scores for multiple loci such as HHH, A–H, H– –H or A– –B.

The information contained in the phenotypic values and in the marker map can also be used simultaneously to calculate conditional probabilities given the observed marker data and given the phenotypic values: the above procedures can be combined and this leads to our QTL mapping method. Given the current parameter estimates the conditional probability in step 1 is updated as follows:

$$P(g\,|\,y,h)=\frac{P(g\,|\,h)\cdot f(y\,|\,g)}{f(y\,|\,h)}$$

where $P(g\,|\,h)$ is the conditional probability for the complete genotype g given the incomplete genotype h, $f(y\,|\,g)$ is the probability density function of the trait y given the complete genotype g, and $f(y\,|\,h)=\Sigma_g\, P(g\,|\,h)\cdot f(y\,|\,g)$ is the mixture of probability density functions of the trait y given the incomplete genotype h. In step 2 the regression parameters are updated and so are the recombination frequencies if the map is not fixed.

This method is a modification of the approach proposed by Jansen (1992). The method described here allows more efficient computer programming. A computer program has been written in Genstat (Genstat 5 Committee 1987), exploiting weighting options for (generalized) linear models.

The completed data are used for the weighted regression of phenotype on genotype and residuals may be calculated in the usual way. A measure for the discrepancy between the data and their fitted values can be obtained by calculating the weighted sum of the squared residuals

$$\Delta^2 = \sum_g P(g\,|\,y,h) \cdot (y - m_g)^2,$$

where m_g is the mean of genotype g. For observations obtained from one of the parents or from the F_1 progeny, the weighted sum of squared residuals is in fact a squared residual. For non-mixture data the squared residual follows approximately a chi-squared distribution with one degree of freedom, multiplied by the residual variance. No standard theory is currently available on the distributional properties of the weighted sum of squared residuals in the case of mixture models; as an ad hoc approximation we used the chi-squared distribution with one degree of freedom, multiplied by the residual variance.

Generalizations. In our approach outlined above, phenotypic data of the parental lines and their F_1 progeny can be included without any further modification. The genotypes at the marker loci are completely known; no data completion is required. By definition then, all markers and putative QTLs have genotype A for one parent, B for the other parent, and H for the F_1.

Other generations, such as doubled haploids, backcross progenies and F_3's, can be dealt with in a similar way to the F_2. In a backcross progeny, for example, an incomplete observation (y) is replaced by two weighted complete observations y(A) and y(H) (or y(H) and y(B), depending on the direction of the backcross). When using information from linked markers in a backcross, the weighting rules must be adapted accordingly. Recombinant inbred lines (RILs) can also be dealt with easily, the modification being that only homozygotes can occur; and again the weighting rules must be adapted accordingly when using linkage information.

When the experimental set-up involves fixed effects, like block effects or replicates, these are accommodated for straightforwardly by adding corresponding terms in the regression model.

The above procedure applies not only to multiple linear regression models, assuming a normal error distribution, but also to generalized linear models (GLM). Generalized linear models can be used to describe the dependence of phenotype on genotype for grouped normal, gamma, binomial, multinomial, Poisson, ordinal data, and so on (MCCULLAGH and NELDER 1989). This is of particular importance since the distribution of many agronomic traits in crop species, for which QTL mapping is relevant, is of one of the above listed types. The same procedure also applies to variance component models that are often used for QTL mapping in animals.

Model selection. We choose the genetic models that maximize the value of the log-likelihood ($\mathfrak{L}$) minus a penalty for the number of free parameters (k) in the model. Equivalently, Akaike's Information Criterion AIC=-2($\mathfrak{L}$-k) may be minimized. The number of parameters should not be too large, preferably less than 2√(number of observations) (SAKAMOTO, ISHIGURO AND KITAGAWA 1986).

In many experiments designed to detect associations between marker genotypes and quantitative characters, the number of segregating molecular markers may be fairly large. Since in an F_2 each marker that is used as a cofactor corresponds to two parameters, the number of parameters may readily exceed 2√(number of observations). In order to avoid this situation we have used the following procedure to select only the most influential markers as cofactors. Linkage group by linkage group, the AICs for several models are calculated and subsets of markers are selected. First, the phenotype of the F_2 progeny is regressed on the markers of only the first linkage group, and the corresponding AIC is calculated. Some of these markers may be dropped from the model to reduce the AIC; the subset of markers with the smallest AIC is retained. Next, the phenotype of the F_2 progeny is regressed on the markers of only the second linkage group, and the corresponding AIC is calculated. Some of these markers may be dropped to reduce the AIC of the second linkage group, and so on. In the end the selected markers of all linkage groups are amalgamated and a new, overall AIC value is calculated for the regression of the phenotype of the F_2 progeny on all selected markers.

In the process of interval mapping, a single putative QTL is moved along the genetic marker map and at each position the deviance (twice the log likelihood ratio) or the LOD score (deviance divided by 2ln(10) ≈ 4.6) between the model with and that without the assumed QTL is calculated and plotted along the marker map. Table 1 lists the models for which it makes sense to calculate (maximum) likelihoods (same notation as JANSEN 1993b). For the example data we have calculated the deviances between models A_2 (with QTL) and B_2 (without QTL) of Table 1; in both cases the selected markers on the other chromosomes were used as cofactors. We also calculated the deviances between models A_1 (with QTL and all selected markers) and B_2 (without QTL, with selected markers on other chromosomes only), which expresses the joint effect of a putative QTL and the selected markers on the same chromosome; the resulting deviance curve will be (approximately) a level line if there is a single QTL the effect of which is absorbed by selected flanking markers. If there is an additional QTL on the same chromosome, the deviance curve may show a peak at the position of that second QTL, and so on (see JANSEN 1993b for more details). For the sake of comparison we also calculated and plotted the deviance between models C and D, which corresponds to "conventional" interval mapping.

The use of AIC provides a decision strategy for model selection and enables us to compare nested and unnested hypotheses. One should consider all models which have approximately equal AICs (i.e. models with an AIC difference less than 2 or some other chosen threshold). Regular methods can be used for testing of nested hypotheses. Tests for the presence of a QTL (model C versus model D, or model A_2 versus model B_2) can be based on the deviance, but its (asymptotic) distribution is not exactly known. As a rule of thumb, we use the chi-squared distribution with 3 degrees of freedom (one degree of freedom for the recombination parameter, one for the additivity parameter of the QTL and one for the dominance parameter of the QTL). Each additional marker in the model takes two extra degrees of freedom. It takes 4 degrees of freedom to test for the simultaneous effect of two markers in multiple regression on markers; it takes 5 degrees of freedom to test model A_1 versus model B_2 for the simultaneous effect of a single QTL and one marker, and so on. Many tests are performed when moving along the genetic map. An overall significance level cannot be guaranteed due to the current lack of knowledge about the statistical behaviour of the (interdependent) tests. Using a significance level of 0.001 per test, the overall significance level in conventional interval mapping would be between 1% and 5% for a genome of 12 chromosomes covered with 50 markers (KNOTT and HALEY 1992). We use the same significance level per test (0.001) in the practical example on tomato plants described in the next section, but an overall significance level for our mapping approach cannot be guaranteed. The chi-

Table 1. Outline of the models fitted

QTL fitted	Selected markers used on no/other/all chromosomes		
	no	other	all
yes	C	A_2	A_1
no	D	B_2	B_1

Models C and D are compared in "conventional" interval mapping. Models A_1, A_2, B_1 and B_2 make use of additional marker cofactors to reduce genotypic variation induced by QTLs located elsewhere on the genome squared threshold at a significance level of 0.001 per test equals 13.8 for 2 degrees of freedom; it is 16.3, 18.5, 20.5, 22.5, 24.3 and 26.1 for 3, 4, 5, 6, 7 and 8 degrees of freedom, respectively. By using a high significance level per test the probability of missing any existing QTL may become undesirably large. QTLs the presence of which cannot be demonstrated significantly may still partly explain the differences for phenotypic values between the parents, F_1 and F_2. Therefore, selected markers may be retained in the regression even though no QTLs are indicated significantly in the nearby region.

APPLICATION

A practical example on plant height in an F_2 progeny of tomato will be used to illustrate the methods described in the previous section; additional parental and F_1 data and marker cofactors are used in the interval mapping. The data are part of a larger experiment, the details and results of which will be reported elsewhere.

The parents were a commercial tomato cultivar (*Lycopersicon esculentum*) and a wild species (*Lycopersicon pennellii*). In the F_2 52 restriction fragment length polymorphism (RFLP) markers were scored. Plant height was measured six weeks after sowing. Mean phenotypic values and variances for the parents, the F_1 and the F_2 progeny are presented in Table 2. A log-scale was used as is commonly done for young plants when growth is nearly exponential. Four percent of the marker data were missing. Two of the 84 F_2 plants had broken tops so that their observations of plant height were missing. Nevertheless, their marker data could still be used for mapping markers.

The markers were assigned to linkage groups and mapped (and the recombination frequencies between adjacent markers were estimated) by using the computer package JOINMAP (STAM 1993). The total number of markers is 52, so that the total number of

Table 2. Mapping QTLs for plant height: some population parameters for *L. esculentum*, *L. pennellii*, the hybrid F_1 and the F_2

Population	Number of plants	Mean phenotype	Phenotypic variance
L.Esculentum	18	4.009	0.0199
L.Pennellii	20	3.885	0.0219
F_1	11	4.049	0.0877
F_2	82[a]	4.022	0.1483

Plant height (cm) has been log-transformed.
[a] RFLP data for 84 plants, plant height data for 82 plants.

parameters in the regression of the phenotype on all markers is equal to 104. This number exceeds the number of F_2 plants (82), and is still too large for reliable model selection even when parental and F_1 data are added (49 plants). Therefore, we applied the procedure of marker selection described above, using the F_2 data. These selected markers were subsequently used as cofactors in interval mapping (also some non-selected marker cofactors were added again during the interval mapping stage; see below). Next, the phenotypes of the F_2 progeny, the parents and the F_1 progeny were simultaneously regressed on a single QTL and on selected markers. This putative QTL was moved along the genetic maps of the various chromosomes. The results are shown in Figure 1. The impact of a single putative QTL on a given chromosome is indicated by the deviance between models A_2 (with QTL) and B_2 (without QTL); in both cases the selected markers on the other chromosomes were used as cofactors (finely dashed lines). The joint effect of the putative QTL and selected markers on the same chromosome is expressed by the deviance between models A_1 (with QTL and all selected markers) and B_2 (without QTL, with selected markers on other chromosomes only) (coarsely dashed lines).

At least six QTLs were indicated, one on each of the chromosomes *6*, *7*, *8* and *9* (in the regions were the finely dashed lines in Figure 1 exceed the critical level of 16.3) and two QTLs on chromosome *2* (in the regions close to the marker cofactors; see below). Selected markers on chromosomes *3*, *5* and *10* were retained in the regression to absorb effects of possible QTLs whose presence could not be demonstrated significantly, but which still explain a part of the phenotypic variation. On chromosome *8* the smallest AIC value of model A_1 is much less than the smallest AIC value of model A_2 (the AIC difference is $41.93\text{-}27.96\text{-}2k=5.97>2$, where k is the number of free parameters for the additional two cofactors; see Figure 1). This indicates multiple QTLs on chromosome *8*. However, the deviance difference of 13.97 is still not significant: it is less than the critical value of 18.5. We did present only the most apparent result (estimates for a single QTL on chromosome *8*), but we should bear in mind that the true genetic background can be more complex (multiple QTLs on chromosome *8*). On chromosome *2* the joint contribution of the two marker cofactors to the deviance is significant: the coarsely dashed line in Figure 1 exceeds the critical value of 18.5. The effect of the cofactors are opposite, which indicates an extremely difficult case to unravel: linked QTLs with opposite effects. The finely and coarsely dashed lines in Figure 1 result from using either none or both of the two cofactors, respectively. We also fitted model A_1 with either the first or the second cofactor; the estimates of the two QTLs are based on these models. The effect of one QTL is estimated on the assumption that the effect of the other QTL is eliminated by the marker cofactor.

Table 3 presents estimates of the QTL effects. Three out of the six QTLs have large positive additive effects, the other three have large negative additive effects. Note that the parents, the F_1 and the F_2 have approximately the same mean height (Table 2), so that the effects of the QTLs should approximately cancel. The discrepancy between the summed QTL effects and the observed differences between the parents could be due to undetected QTLs; their effects are hopefully eliminated by the marker cofactors. The pooled environmental variance for the original parents and the F_1 equals 0.0273 (after removing one F_1 plant; see below). Table 3 shows that this value is approximated very

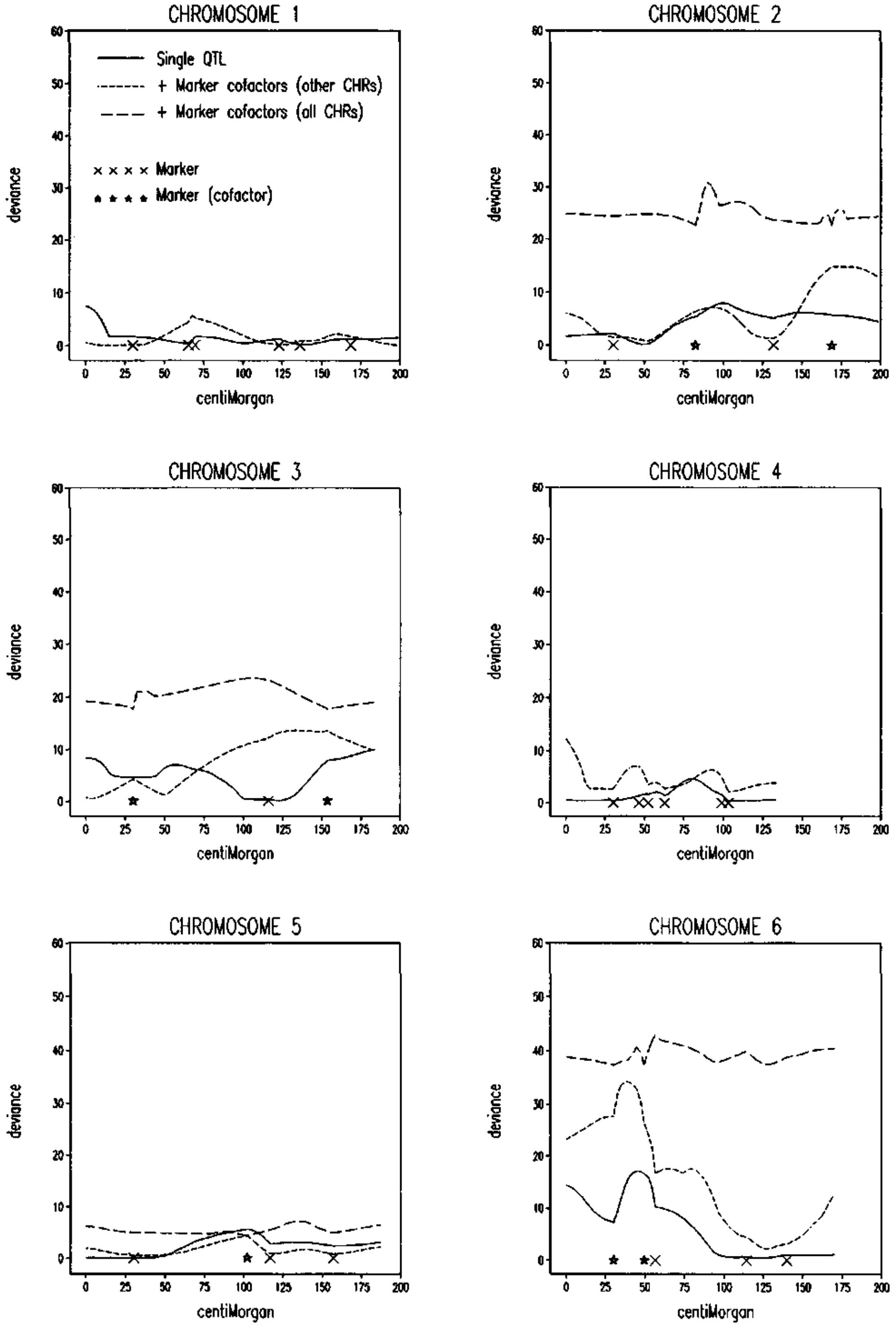
Table 3. Estimates of QTL effects, residual variance and recombination frequency between QTL and left flanking marker

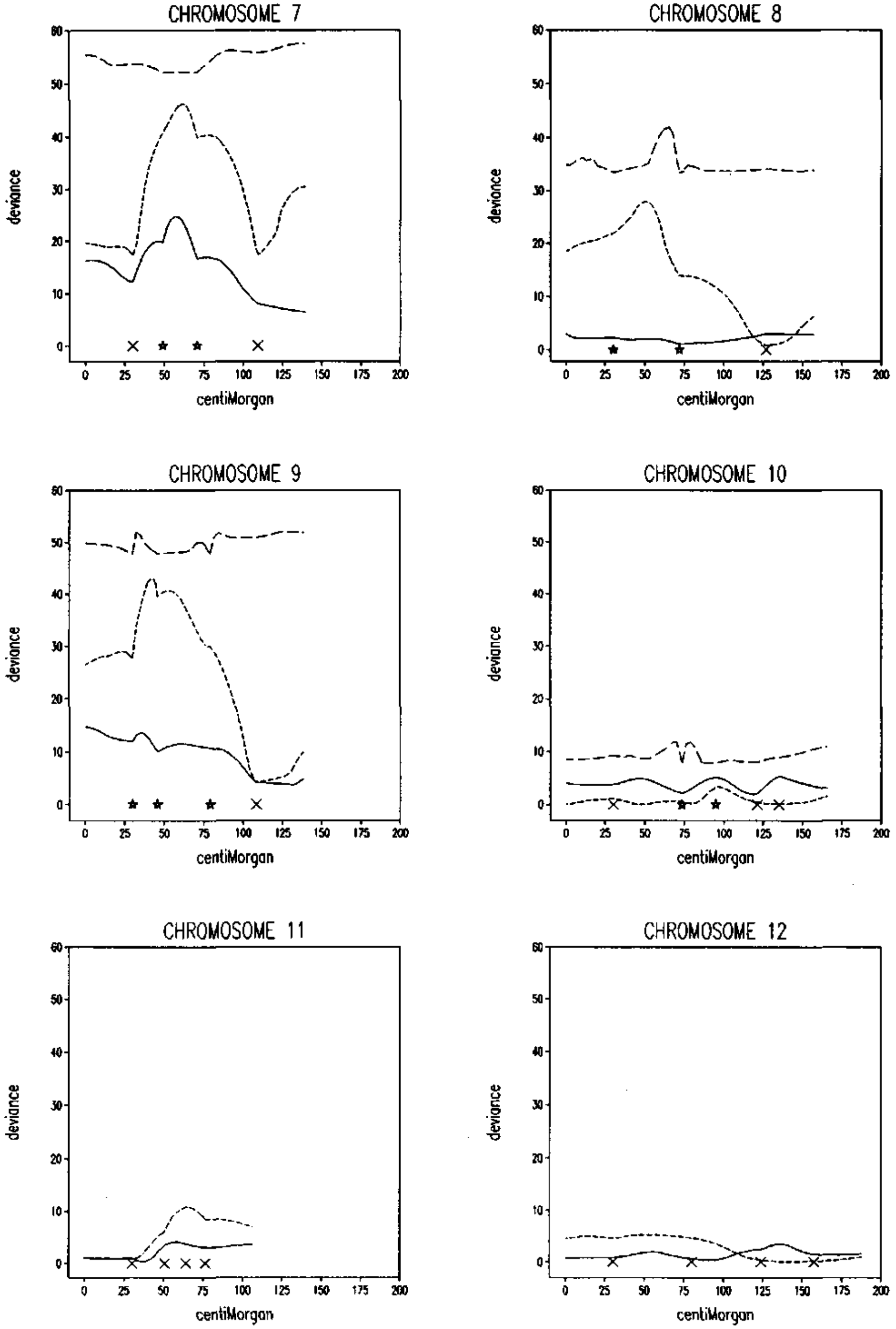
Chromosome (and marker interval)	QTL effects		Variance	Recombination frequency[a]
	Additive	Dominance		
2 (2-3)	0.255 (0.050)	0.026 (0.065)	0.0197 (0.0053)	0.130 (0.041)
2 (4-)	-0.247 (0.043)	-0.071 (0.057)	0.0208 (0.0050)	0.091 (0.040)
6 (1-2)	-0.204 (0.044)	0.205 (0.063)	0.0244 (0.0043)	0.070 (0.039)
7 (2-3)	-0.248 (0.047)	-0.114 (0.067)	0.0236 (0.0043)	0.111 (0.039)
8 (1-2)	0.272 (0.058)	0.118 (0.064)	0.0181 (0.0028)	0.165 (0.038)
9 (1-2)	0.249 (0.037)	-0.087 (0.048)	0.0249 (0.0042)	0.111 (0.036)

Standard errors of the estimates are presented between brackets.

[a] The QTL was moved along the genetic map with steps of 2.5 cM; the recombination between the QTL and its left flanking marker is reported.

Figure 1. Deviance plots for plant height in an F_2 progeny of tomato. The phenotypes of the F_2 progeny were regressed on a putative QTL, which was moved along the genetic map of each chromosome ("conventional" interval mapping). The deviance between the model with the QTL (model C) and the model without the QTL (model D) was plotted (solid line). The phenotypes of the F_2 progeny, the parents and the F_1 progeny were simultaneously regressed on a putative QTL and a number of selected markers; again the QTL was moved along the genetic map. The finely dashed line shows the plot of the deviance between model A_2 (with the QTL) and model B_2 (without the QTL); in both cases all selected markers from other chromosomes were used. The coarsely dashed line represents the plot of the deviance between model A_1 (with the QTL and all selected marker cofactors) and model B_2 (without the QTL but with selected markers only on other chromosomes).


CHROMOSOME 1
CHROMOSOME 2
CHROMOSOME 3
CHROMOSOME 4
CHROMOSOME 5
CHROMOSOME 6
Single QTL
+ Marker cofactors (other CHRs)
+ Marker cofactors (all CHRs)
× × × × Marker
★ ★ ★ ★ Marker (cofactor)
deviance
0
10
20
30
40
50
60
centiMorgan
0
25
50
75
100
125
150
175
200


CHROMOSOME 7
CHROMOSOME 8
CHROMOSOME 9
CHROMOSOME 10
CHROMOSOME 11
CHROMOSOME 12
deviance
0
10
20
30
40
50
60
centiMorgan
0
25
50
75
100
125
150
175
200

well by using single QTL models with marker cofactors on other chromosomes, indicating that these models explain the total genetic variation satisfactorily.

It should be mentioned that the interval mapping stage was passed through several times. The first time all preselected markers were used as cofactors (so far chromosome *8* contained no selected markers). Then the deviance plot for chromosome *8* showed a clear peak, indicating a QTL between marker 1 and 2. Therefore, the second time two cofactors were added on chromosome *8* to eliminate the putative QTL effect. Next the weighted sums of squared residuals were checked for outliers. Figure 2 presents a histogram of the weighted sum of squared residuals obtained from the multiple linear regression of the phenotype of the F_2 progeny, the parents and the F_1 progeny on all selected markers. At a significance level of 0.01 the critical value equals approximately 0.24, so that one observation from the F_1 may be considered to be an extreme outlier. One plant of the F_2 progeny has a weighted sum of squared residuals just exceeding the critical value. The F_2 outlier also caused narrow sharp peaks in the coarsely dashed lines close to marker cofactors (not shown): the factor for a putative QTL absorbed the effect of the outlier rather than an effect of a true QTL. The plant heights of these two outliers were removed, which reduced the variance among F_1 plants from 0.0877 to 0.0512, and changed the variance among F_2 plants from 0.1483 to 0.1499 (see Table 2). For the third and final time the interval mapping was then passed through. After each successive passing of interval mapping the peaks shown in Figure 1 for chromosomes *6*, *7*, *8* and *9* became more pronounced.

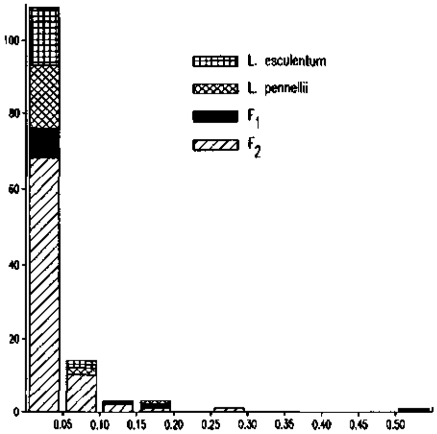


Figure 2. Histogram of the weighted sum of squared residuals, used for the detection of outliers for plant height in an F_2 progeny of tomato. The residuals were obtained from the multiple linear regression of the phenotypes of the F_2 progeny, the parents and the F_1 progeny on all selected markers. Two outliers are indicated, namely the plants with the weighted sum of squared residuals > 0.24.

To compare the above results with conventional interval mapping, the phenotypes of the 82 F_2 plants were regressed on a single putative QTL, which was moved along the genetic map. The deviance between the model with the single QTL (model C) and that without the single QTL (model D) was plotted at each map position (solid line in Figure 1). A comparison of deviance curves for chromosomes *6*, *7*, *8* and *9* demonstrates that our approach is much more powerful than conventional interval mapping. Only two QTLs are detected by conventional interval mapping (one QTL on chromosome *6* and one on chromosome *7*).

DISCUSSION

Powerful and accurate QTL mapping can serve several important goals. First, dissecting quantitative characters into Mendelian factors yields a position from where the genetics of complex characters can be studied in terms of individual gene effects rather than in the statistical terms (variances, covariances, etc.) of classic quantitative genetics. Secondly, the application of indirect selection via markers and other forms of tracing individual genes in breeding programs, such as guided introgression, gains substantially from powerful QTL mapping methods. In this paper none of these ultimate goals was aimed at directly; nevertheless, the example given illustrates the potential contribution of our new analytical method to progress in these areas. The phenotypic variation of the quantitative trait was resolved into at least six putative QTLs and an environmental error component. These results should still be regarded as preliminary; they have to be confirmed by further experiments. F_3 lines, isogenic for regions of putative QTLs, may be produced and tested (PATERSON et al. 1991); also backcross inbred lines may be used for this purpose (BECKMANN and SOLLER 1989).

Our approach to QTL mapping uses the unified concept of completing missing genotypic data for both a putative QTL and markers. If many data are missing, this may give rise to computational problems: in an F_2 one missing marker observation may actually have one of three allelic constitutions, two missing marker observations (for the same plant) result in nine possible constitutions, and so on. If in a data set with many markers a certain proportion of the marker genotypes is missing, the number of weighted completed data may become so large that computation is no longer feasible. Molecular geneticists, who are generally collecting the marker data, should be aware of the consequences of missing marker data, so that they hopefully will strive for completeness of their data. However, to complete data it is not necessary to use all available information; the amount of computation can be reduced considerably by a limited completion of missing data: genotypes with negligible weights may be disregarded, without substantial loss of information.

In conventional interval mapping data from the parents and the F_1 progeny cannot be used; if the parental and F_1 data were included, the results would be seriously biased because the single QTL would be called upon to explain all the mean differences between the parents and the F_1 progeny. It is only because markers are used as cofactors in our approach that data from parents and F_1 can be included; QTL mapping may become much more powerful when marker cofactors explain a large proportion of the genetic variation (or at least the mean difference between the parents and the F_1 progeny). In other cases, for instance when there are numerous QTLs of small effect distributed throughout the genome, the power of QTL mapping may be reduced by using parental and F_1 data, because the additional constraints on the parameters are too exacting.

In our example data set, an interaction between marker cofactors and a putative QTL is indicated (Figure 1, chromosome *8*): if the inclusion of marker cofactors simply reduced the residual variance, the solid and finely dashed lines should be approximately similar in shape, although the finely dashed lines might be higher. We speculate that in the small F_2 progeny of 84 plants in our example, deviant segregation ratios for two or more unlinked QTLs have masked the effect of the QTL on chromosome *8* when we applied the conventional interval mapping method. In our approach, the effects of the QTLs involved could be unravelled by the use of marker cofactors. This problem for small populations should be explored in more detail by simulation.

Little is known about the influence of outliers on QTL mapping; we proposed a weighted sum of squared residuals to indicate outliers. Two particular observations in the example data set were detected as potential outliers. It was observed that such outliers can incorrectly indicate multiple linked QTLs. Also they may hamper efficient and accurate resolvability of QTLs.

In the example we have come across a situation which represents a "worst case" configuration: linked QTLs with opposite effects. As indicated by STAM (1991), and confirmed by the present study, in such a case multiple regression will be more powerful than "conventional" interval mapping. Our single data set cannot answer the general question as to what resolution power is attainable with our method. To answer this question a number of known configurations of QTLs and QTL effects, as well as heritability and population size, need to be studied by simulation.

The regression models that are used in our approach assume additivity of effects over loci. Though epistatic effects can in principle be modelled straightforwardly as well, we have chosen not to do so because of the rapid increase of the number of parameters, relative to the amount of data. In our view, however, the detection of epistatic effects requires a different type of experimental approach, such as raising the F_3 offspring of deliberately chosen F_2 multilocus marker genotypes.

LITERATURE CITED

BECKMANN, J. S., and M. SOLLER, 1989 Backcross inbred lines for mapping and cloning of loci of interest, pp. 117-122 in *Development and application of molecular markers to problems in plant genetics*, edited by B. BURR and T. HELENTJARIS. Brookhaven National Laboratory

GENSTAT 5 COMMITTEE, 1987 *Genstat 5 reference manual*. Clarendon Press, Oxford

HALEY, C. S., and S. A. KNOTT, 1992 A simple 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. Theoretical and Applied Genetics 85: 252-260

JANSEN, R. C., 1993a Maximum likelihood in a generalized linear finite mixture model by using the EM algorithm. Biometrics 49: 227-231

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

KNAPP, S. J., 1991 Using molecular markers to map multiple quantitative trait loci: models for backcross, recombinant inbred, and doubled haploid progeny. Theoretical and Applied Genetics 81: 333-338

KNOTT, S. A., and C. S. HALEY, 1992 Aspects of maximum likelihood methods for the mapping of quantitative trait loci in line crosses. Genetical Research 60: 139-151

LANDER, E. S., and D. BOTSTEIN, 1989 Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121: 185-199

MCCULLAGH, P., and J. A. NELDER, 1989 *Generalized linear models. Monographs on statistics and applied probability 37.* Chapman and Hall, London

PATERSON, A. H., E. S. LANDER, J. D. HEWITT, S. PETERSON, S. E. LINCOLN and S. D. TANKSLEY, 1988 Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment polymorphisms. Nature 335: 721-726

PATERSON, A. H., S. D. DAMON, J. D. HEWITT, D. ZAMIR, H. D. RABINOWITCH, S. E. LINCOLN, E. S. LANDER and S. D. TANKSLEY, 1991 Mendelian factors underlying quantitative traits in tomato: comparison across species, generations and environments. Genetics 127: 181-197

SAKAMOTO, Y., M. ISHIGURO and G. KITAGAWA, 1986 *Akaike information criterion statistics.* KTK Scientific Publishers, Tokyo

STAM, P., 1991 Some aspects of QTL mapping, in Proceedings of the Eighth Meeting of the Eucarpia section Biometrics in Plant Breeding. Brno, July 1991

STAM, P., 1993 Constructing integrated genetic linkage maps by means of a new computer package: JOINMAP. The Plant Journal 3: 739-744

STUBER, C. W., S. E. LINCOLN, D. W. WOLFF, T. HELENTJARIS and 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

VAN OOIJEN, J. W., 1992 Accuracy of mapping quantitative trait loci in autogamous species. Theoretical and Applied Genetics 84: 803-811

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

APPENDIX

Updating the estimates of the recombination frequencies in the EM algorithm runs parallell to the "normal" EM procedure for estimation of *r* from F_2 data, as outlined below. In an F_2 recombinant the F_1 gametes could be counted directly from the frequencies of the genotypes AA, AH, AB, HA, HH, HB, BA, BH and BB if the contribution of repulsion and coupling phase to HH were known. Given the current estimate, r, the ratio of repulsion and coupling phase within the double heterozygotes equals $r^2 : (1-r)^2$. Denoting the observed genotypic frequencies by n(AA), n(AH), etc., the EM procedure runs as follows:

Step 1: update the unknown number of repulsion heterozygotes,

Step 2: obtain the new estimate by counting recombinant gametes.

This leads to the following update

$$\hat{r}=\frac{n(AH)+n(HA)+n(BH)+n(HB)+2\left[n(AB)+n(BA)+\frac{r^2}{r^2+(1-r)^2}\cdot n(HH)\right]}{2\sum n(\cdot)}.$$

When updating the estimate of *r* in our QTL mapping method, the above equation is used; the numbers n(·) are replaced by the updated summed weights w(·), where w(·) and n(·) are defined analogously.

V. CONTROLLING THE TYPE I AND TYPE II ERRORS IN MAPPING QUANTITATIVE TRAIT LOCI

ABSTRACT

Although the interval mapping method is widely used for mapping quantitative trait loci (QTLs), it is not very well suited for mapping multiple QTLs. Here, we present the results of a computer simulation to study the application of exact and approximate models for multiple QTLs. In particular, we focus on an automatic two-stage procedure in which in the first stage "important" markers are selected in multiple regression on markers. In the second stage a QTL is moved along the chromosomes by using the preselected markers as cofactors, except for the markers flanking the interval under study. A refined procedure for cases with large numbers of marker cofactors is described. Our approach will be called MQM mapping, where MQM is an acronym for "multiple-QTL models" as well as for "marker-QTL-marker". Our simulation work demonstrates the great advantage of MQM mapping compared to interval mapping in reducing the chance of a type I error (i.e. a QTL is indicated at a location where actually no QTL is present) and in reducing the chance of a type II error (i.e. a QTL is not detected).

INTRODUCTION

The advent of maps of molecular markers enables geneticists to detect and map individual loci affecting quantitative traits (cf. PATERSON et al. 1988). In the ideal case all genetic variance of the trait is explained by detected QTLs. In practice a number of QTLs may be missed (a type II error) and at the same time a number of false positives may occur, indicating QTLs at map positions (or regions) where actually no QTLs are present (a type I *error). The actual balance between the cost of false positives and the benefit* of detected QTLs depends on the aim of the experiment (e.g. map-based cloning or introgression breeding). Nevertheless, one often strives for keeping at least the chance of a type I error below 5%. Therefore, the QTL mapping method used should keep the chance of a type I error below 5%, but at the same time it should minimize the chance of a type II error. The interval mapping method (LANDER and BOTSTEIN 1989) is widely used, but it is now generally recognized that the chance of a type I or a type II error is higher in interval mapping than it is in simultaneous mapping of multiple QTLs (cf. HALEY and KNOTT 1992; MARTINEZ and CURNOW 1992; JANSEN 1993b). This has motivated theoretical research for multiple QTL mapping methods. Recently, JANSEN (1992,1993b) and JANSEN and STAM (1994) developed a unifying framework of exact and approximate models for multiple QTLs, from now on called MQM mapping. MQM is an acronym for "multiple-QTL models" but also for "marker-QTL-marker" (which reflects the insertion of QTLs between markers on the genetic linkage map). The framework includes interval mapping and regression on markers (COWEN 1989; STAM 1991; RODOLPHE and LEFORT 1993; ZENG 1993) and also includes their "hybrid" in which the phenotype is regressed on a single putative QTL in a given marker interval, and at the same time on a number of markers located elsewhere on the genome. The rationale behind using markers as cofactors is that these markers will eliminate the major part of the variation induced by nearby QTLs. Some simulation work (JANSEN 1993b) and a practical application (JANSEN and STAM 1994) indicated that the MQM mapping method is computationally feasible and substantially more powerful than interval mapping. For the present paper a computer simulation study was set up to study more thoroughly the chances of a type I or type II error in MQM mapping, and to compare MQM mapping with interval mapping. A number of QTL configurations were studied by simulation, covering the most relevant multiple-QTL configurations; the results are presented and discussed.

STATISTICAL MODELS FOR MQM MAPPING

In this section statistical aspects of MQM mapping are summarized. For more details see JANSEN (1992,1993b) and JANSEN and STAM (1994). Further refinements to MQM mapping are proposed, concerning the testing for the presence of a putative QTL, and concerning the parameter estimation for the case that many marker cofactors are used.

The framework. We restrict ourselves to backcross progenies, but the same method applies to other inbred or outbred progenies. Furthermore, we assume a normally distributed environmental error. The general model in MQM mapping is $Y = m + \Sigma x_i a_i + E$, where Y is the phenotypic trait, m is the mean, a_i are the allele substitution effects of individual loci and E is the (environmental) error; the summation is over all loci affecting the trait. The x_i are indicator variables specifying the genotype. In a backcross progeny they can take two values: 0 or 1. The loci in the above expression can be one or more QTLs, but -as an approximation- markers can be used as well. Therefore, the model includes interval mapping (LANDER and BOTSTEIN 1989), but also exact models for multiple QTLs (JANSEN 1992,1993b; JANSEN and STAM 1994), multiple regression on markers (COWEN 1989; STAM 1991; RODOLPHE and LEFORT 1993; ZENG 1993), and the hybrid between interval mapping and multiple regression on markers (JANSEN 1993b) in which marker cofactors are selected prior to the analyses considering a QTL in each interval in turn. Parameter estimation is based on the simultaneous distribution of the genotype and phenotype; the core of our method is completion of any missing genotypic (QTL and marker) information, which is embedded in a general and simple EM algorithm to obtain maximum likelihood estimates of the model parameters (JANSEN and STAM 1994). In the case of exact models for multiple QTLs, this procedure makes simultaneous estimation of QTL positions possible.

Preselection of marker cofactors. Markers can be used in the regression to take over the role of nearby QTLs. STAM (1991) demonstrated that in multiple regression the effect of a QTL is absorbed only by its flanking markers, at least if the progeny size is large; other markers are then redundant. Since the locations of the QTLs are generally unknown, the question is which markers have to be used as cofactors in MQM mapping. A standard regression selection procedure can be used to select the "important" markers. One such procedure is backward elimination of marker cofactors in multiple regression of the phenotype on the markers. Jansen (1993b) minimized Akaike's Information Criterion, AIC=$-2(\mathcal{L}-k)$, where $\mathcal{L}$ is the log-likelihood and k is the number of free parameters in the model. Here, we minimize $-2(\mathcal{L}-3k)$, i.e. a more stringent penalty for the number of free parameters is used. In "ordinary" regression with adequate degrees of freedom to estimate σ^2, a penalty of k is equivalent to the use of (about) the 16% point of the F-test for the comparison of two nested models, which differ only by the inclusion of one free parameter; a penalty of $3k$ is equivalent to the use of (about) the 2% point (MCCULLAGH and NELDER 1989). At each step of the backward elimination process a marker is dropped, namely the marker which gives the largest decrease of the criterion; the process is stopped when no further reduction of the criterion can be achieved. In the next stage (the actual mapping stage), the selected markers will be used as cofactors. For proper marker selection a reasonable number of recombinants between flanking markers is required (the larger the QTL effect, the fewer recombinants are required). Because of the near collinearity of closely linked marker cofactors, it makes little sense to use a very dense map in a progeny of, say, 100 individuals.

Very recently, ZENG (1994) presented a simulation study in which all markers were used as cofactors, except for the markers flanking the interval under study. He, however, also suggested preselecting the markers which explain most of the genetic variation in the genome.

Testing for the presence of a putative QTL. In MQM mapping at each map location the log-likelihood $\mathcal{L}_1$ for a single QTL in a given interval can be calculated and compared with the log-likelihood $\mathcal{L}_0$ of no QTL in the given interval, using in both models the same set of marker cofactors (or the same set of QTLs in other intervals when exact models for multiple QTLs are used). The likelihood-ratio test statistic for the presence of a putative QTL in a given interval is then expressed as the maximum of $2(\mathcal{L}_1-\mathcal{L}_0)$ over the interval. The distribution of the test statistic for the presence of a QTL in a specific interval is not exactly known. However, when no QTLs are segregating, the asymptotic distribution is expected to be between the χ_1^2 and χ_2^2 distribution (TITTERINGTON, SMITH and MAKOV 1985). The latter distribution is justified by the difference in the number of parameters (one for the allele substitution effect *a* of the putative QTL, and one for the location of the QTL in the marker interval). The former is justified by the fact that the null hypothesis is defined by the single constraint *a*=0. LANDER and BOTSTEIN (1989) and VAN OOIJEN (1992) simulated the distribution of the test statistic. Based on extensive simulations these authors published appropriate thresholds for the test statistic so that the chance of a false positive occurring anywhere on the genome is at most 5% (still under the assumption that no QTLs are segregating). We here suggest that these thresholds are also suitable for MQM mapping: they can be used when no QTLs are segregating, since in that case it is expected that no or only a very few markers will be selected in MQM mapping. Moreover, these thresholds can also be used when QTLs are segregating, the effects of which are eliminated by marker cofactors in MQM mapping. One condition is, however, that the number of degrees of freedom for estimating σ^2 is large enough (see below).

Marker cofactors should not replace the putative QTL in the interval of current interest. It was decided to study a simple approach to prevent this: for a given interval all selected markers are used as cofactors, except the ones flanking the interval of current interest. We expect that this approach applies well if marker selection is properly based on reasonable numbers of recombinants between flanking markers (see above). Otherwise, a general (but more computer intensive) selection approach can be used (JANSEN 1993b): starting from the single-QTL model using all selected markers, it is assessed which nearby markers still may be dropped (those markers previously explained the effect of the QTL), and which markers cannot be dropped (these markers possibly absorb the effects of other QTLs on the current chromosome).

When the number of marker cofactors is large. In ordinary regression the number of parameters estimated from the data should not be too large when maximum likelihood is used. Asymptotic relations such as the χ^2-approximations do not necessarily hold in the case of large numbers of parameters. The main reason for this is the bias of the maximum likelihood estimate of the residual variance. The usual bias adjustment of the estimate of the variance is to multiply the estimate by *N*/(*N*-*p*), where *N* is the number of individuals and *p* is the number of free parameters used for modelling the relation between the mean and explanatory variables. When comparing a sequence of (nested) models we have the option of using a common estimate of variance for all models in the sequence, or using separate estimates derived from the fit of each model in turn (MCCULLAGH and NELDER 1989). In "ordinary" regression analysis a single estimate of the variance obtained from the most complex model is usually considered. This estimate of the variance is used for all models in the sequence, which at the same time guarantees that the test statistic takes only positive values (cf. HALEY and KNOTT 1992). This property does not hold if for each model a separate bias-adjusted estimate of the variance is used. Here, we deal with mixture models instead of "ordinary" regression models, because of missing QTL and marker observations (it is quite common that a small proportion of the marker data are missing). Variable selection and bias adjustment of the maximum-likelihood estimate of the residual variance in mixture models is an area open to research, probably because mixture models with many parameters did not occur before. Mixture analysis can be viewed as "ordinary" regression with missing values for one or more factors (JANSEN 1992,1993a). Therefore, it is natural to adapt the approach for variable selection and bias adjustment in regression models to the case of mixture models. In MQM mapping with complete linkage maps we propose the use of the following heuristic three-step procedure:

(1) Obtain maximum-likelihood estimates for the most complex model (usually the model for regression of phenotype on all markers);
(2) Adjust the estimate of the residual variance for bias;
(3) Obtain maximum-likelihood estimates in the sequence of models (in the models for regression of phenotype on subsets of the markers during the selection process, or in single-QTL and no-QTL models with selected marker cofactors), keeping the variance fixed at the value obtained from step 2.

Following this approach, the distribution of the test statistic for the presence of a QTL in a specific interval is expected to be between the $F_{1,df}$ and $2F_{2,df}$ distribution rather than between the χ_1^2 and χ_2^2 distribution, where df are the degrees of freedom for estimating σ^2 (HALEY and KNOTT 1992). Therefore, appropriate thresholds for an entire genome should also be functions of the number of residual degrees of freedom. Of course, F and χ^2 distributions are closely related if the number of residual degrees of freedom is large.

SIMULATIONS

For a number of specified configurations of QTLs and QTL effects, we studied the distribution of the test statistic for the presence of a putative QTL. These configurations include no QTL, a single QTL or two QTLs, the two QTLs being unlinked, linked in repulsion (i.e with opposite sign effects) or linked in coupling phase (i.e. with equal sign effects). Furthermore, we considered small and large numbers of markers. In MQM mapping with many cofactors, a common and bias-adjusted estimate of the variance was used for all models according to the procedure described above. See Figures 1-9 for the description of the various settings. Putative QTLs are detected via the following procedures: (a) by MQM mapping using (selected) markers as cofactors, (b) by MQM mapping with exact models for multiple QTLs, and (c) by interval mapping. In all cases we simulated by computer and according to the Mendelian segregation rules the genotypes and phenotypes of 100 individuals as if they had been produced by backcrossing F_1 individuals to one of the parents. For each genetic setting 500 simulations were run. Marker distances were assumed to be known and to be equal to the values used for simulation. For each genetic setting we plotted the simulated distribution of the test statistic in a given interval (the maximum of $2(\mathcal{L}_1-\mathcal{L}_0)$ over all map locations in the given interval). The distributions turned out to be markedly skewed. To have a better presentation we plotted the square root of the test statistic. Two types of simulation were run: (a) simulations concerning configurations with no QTL in the interval of interest, aiming at a study of the type I error, and (b) simulations with a QTL in the interval of interest, aiming at a study of the type II error. They are dealt with in the next two sections, respectively.

Type I error

The distribution of the test statistic for the presence of a putative QTL was simulated for a given marker interval, in which actually no QTL is located. When no QTLs are segregating or when the effects of QTLs are sufficiently eliminated, this distribution is expected to be between the χ^2_1 and χ^2_2 distribution. If the number of degrees of freedom for estimating σ^2 is small, this distribution is expected to be between the $F_{1,df}$ and $2F_{2,df}$ distribution.

We successively considered the following situations. A single QTL is located on the same chromosome as the interval under study, or on another chromosome; or two QTLs in coupling phase are located on the same chromosome, one on either side of the marker interval of interest. We also considered how the distribution of the test statistic is affected by the number of free parameters to be estimated from the data. Finally, we studied the maximum value of the test statistic in an entire genome in absence of segregating QTLs. See Figures 1-5 for the description of the QTL configurations.

A single QTL present on another chromosome. First, we studied the case that no QTL is present in the interval of interest, while a single QTL is present on another chromosome (Figure 1). In interval mapping the distribution of the test statistic may be affected by an unlinked major QTL when the marker interval 1-2 is wide (curve I): the curve deviates from the χ^2 distributions. In MQM mapping, the distribution of the test statistic is unaffected by the QTL when the markers 3 and 4 are used as cofactors (curve C(3,4)). It is of course generally unknown where the QTLs are and therefore one does not know which markers should be used as cofactors to absorb them. The QTL has,

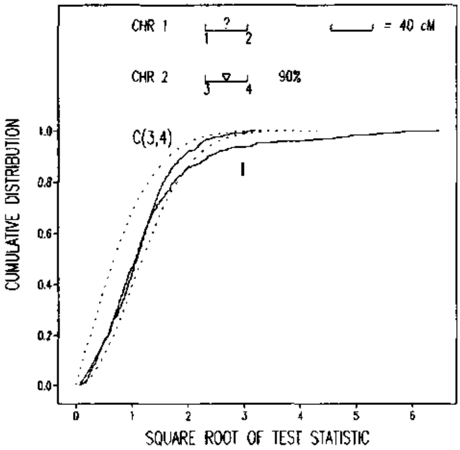


Figure 1. A study of the type I error in case a single QTL is present on another chromosome

Backcross progenies of 100 individuals were simulated. Markers are numbered from the left to the right on each chromosome. The question mark (?) indicates the marker interval under study in which the likelihood for the presence of a putative QTL is assessed. The symbols ▿ and ▵ indicate the position of a QTL with an effect of positive and negative sign, respectively. The percentage (beside the chromosome) indicates what percentage of the expected total phenotypic variance is attributable to the expected (simultaneous) genetic variance of the QTLs on the given chromosome. The finely dashed curves indicate the χ_1^2 distribution (left) and the χ_2^2 distribution (right). The coarsely dashed curves indicate the $F_{1,17}$ distribution (left) and the $2F_{2,17}$ distribution (right). The solid curves indicate:

I: Interval (I) mapping;
C(3,4): MQM mapping with markers 3 and 4 as cofactors (C). Analogous definition for other sets of marker cofactors;
S: MQM mapping with selected (S) markers as cofactors (but markers flanking the interval under study are not used as cofactors, even if they were selected);
E: MQM mapping with exact (E) models for multiple QTLs;
A: MQM mapping with two adjustments (A): (1) the estimate of the variance in the most complex model is adjusted for bias and, (2) the variance in any other model is fixed at the value obtained from the most complex model;
IA: Combination of I and A;
CA: Combination of C and A;
SA: Combination of S and A.

however, a major effect and when marker selection was applied, markers 3 and 4 were selected in almost all simulations (not shown). Therefore, the curve for MQM mapping with selected markers as cofactors almost coincides with the curve C(3,4). As the expected genetic variance of the QTL represents 90% of the expected phenotypic variance, these simulations show the maximal influence of a single QTL on the type I error in an interval on another chromosome.

A single QTL present on the same chromosome. Next, we studied the case that no QTL is present in the interval of interest, while a single QTL is present on the same chromosome (Figure 2a-d). Both in interval mapping and in MQM mapping the presence of a major QTL in marker interval 2-3 has a very strong influence on the test statistic in marker interval 1-2, even when the markers 2 and 3 are used as cofactors (Curves I and C(2,3), respectively, in Figure 2a). We also considered the case that a major QTL is not in marker interval 2-3, but in marker interval 3-4 (Figure 2b). In interval mapping, the test statistic in marker interval 1-2 is still highly affected by the QTL (curve I in Figure 2b); in MQM mapping however, it is unaffected when the markers 3 and 4 are used as cofactors (curve C(3,4) in Figure 2b). In practice it is not apriori known that, for instance, a QTL is located in marker interval 3-4 and that therefore marker 3 and 4 should be used as cofactors to absorb the effect of the QTL. When marker selection was applied, in almost all simulation runs the two markers flanking the QTL were selected: marker 2 and 3 in Figure 2a, and marker 3 and 4 in Figure 2b. In the first case the corresponding curve S deviates even more from the χ^2 distributions than that curve C(2,3) deviates from them, and in the second case curve S coincides with curve C(3,4) (S curves are not plotted). As the expected genetic variance of the QTL forms the major part of the expected phenotypic variance (90%), these simulations show the maximal influence of a single QTL on the type I error in another interval on the same chromosome.

We also simulated the same configurations with a QTL with a much smaller effect (Figures 2c and 2d). In interval mapping, the test statistic is still highly affected by the presence of a QTL in marker interval 2-3 (curve I in Figure 2c), or by the presence of a QTL in marker interval 3-4 (curve I in Figure 2d). In neither case is the test statistic influenced in MQM mapping when the markers 2-3 or 3-4 are used as cofactors (curves C(2,3) and C(3,4) in Figures 2c and 2d, respectively). When marker selection was applied, in many simulation runs only one of the two markers flanking the QTL was selected (marker 2 or 3 in Figure 2c, marker 3 or 4 in Figure 2d). Since markers flanking the interval of interest are not used as cofactors when applying marker selection, this seriously affects the test statistic in the case of a QTL in the interval adjacent to the interval of interest (curve S in Figure 2c). However, when an additional marker between the interval of interest and the QTL is available, the test statistic corresponding to marker selection is hardly affected (curve S in Figure 2d).

Two linked QTLs in coupling phase. Then, we studied the case of two linked QTLs in coupling phase (i.e. with effects of equal size and equal sign; see Figure 3). It is well known that in interval mapping the test statistic in this case will often be at its maximum in one of the intermediate intervals (MARTINEZ and CURNOW 1992). This can lead to the detection of a single QTL in the wrong interval (a type I error). Therefore, we studied the effect of both QTLs on the test statistic in the intermediate marker interval 4-5. The effect of the second QTL in marker interval 6-7 on the test statistic for the first QTL in

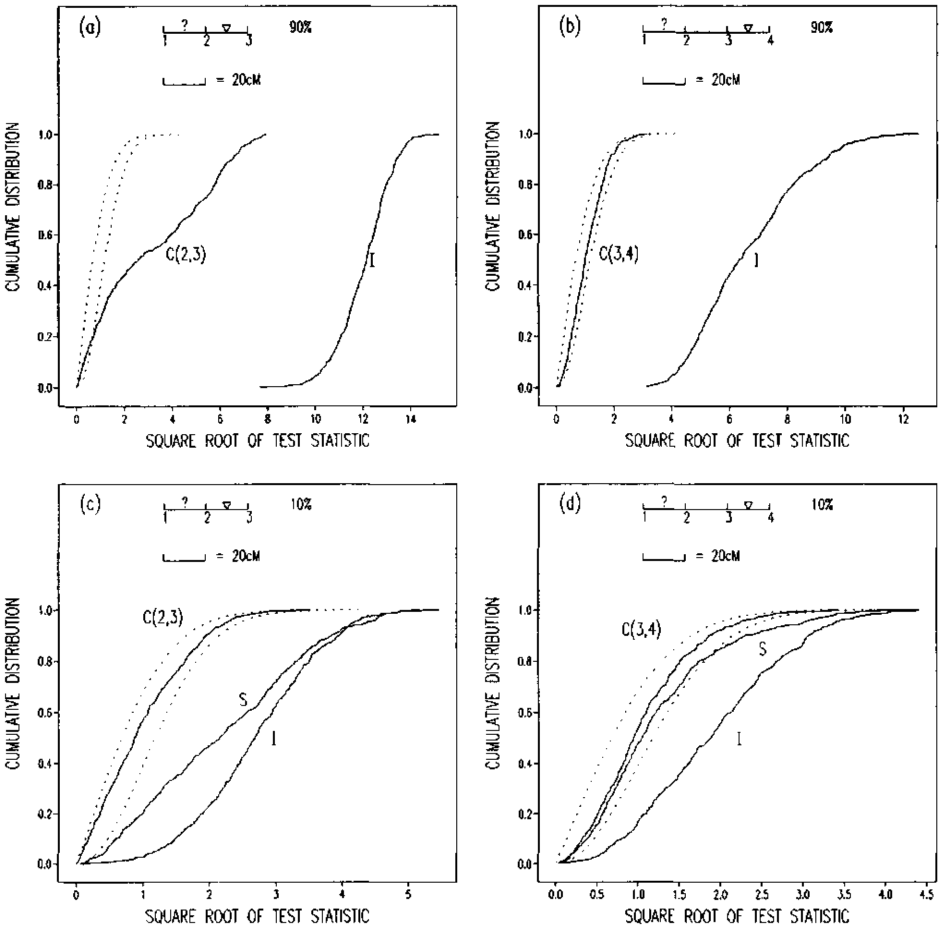


Figure 2. A study of the type I error in case a single QTL is present on the same chromosome (see Figure 1 legend)

marker interval 2-3 is dealt with in the next section (aiming at a study of the type II error).

In MQM mapping, the distribution of the test statistic for the presence of a putative QTL in marker interval 4-5 is unaffected by the two QTLs when the markers 2, 3, 6 and 7 are used as cofactors (curve C(2,3,6,7)), and only slightly affected when selected markers are used (curve S). In interval mapping, the test statistic takes very large values (curve I).

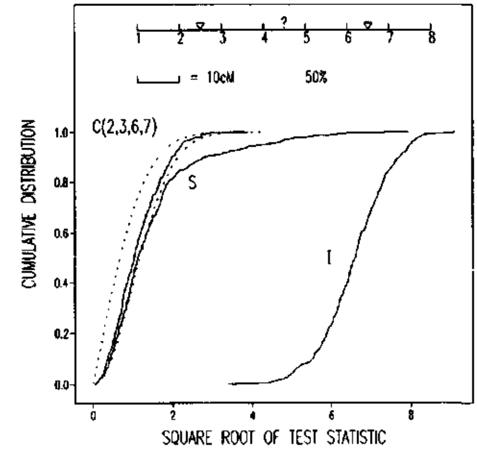


Figure 3. A study of the type I error at an interval between two QTLs in coupling phase (see Figure 1 legend)

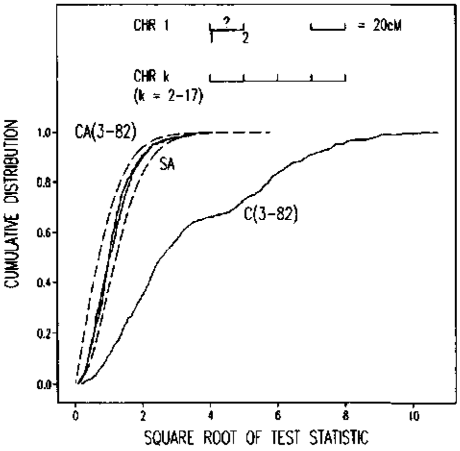


Figure 4. A study of the type I error in case of many marker cofactors (see Figure 1 legend)

The effect of the number of cofactors. Furthermore, we studied the effect of the number of cofactors on the type I error in MQM mapping with 80 marker cofactors distributed over 16 other chromosomes (Figure 4). This figure shows that the test statistic is seriously affected (curve C(3-82)). Therefore, it is clear that the number of redundant cofactors should not be too large in maximum likelihood estimation. Bias adjustment of the estimate of the variance could be a solution to this problem and therefore we reanalysed the case, using the bias adjustment procedure described above. The distribution of the test statistic for the case of 80 cofactors with the bias adjustment is between the $F_{1,17}$ and $2F_{2,17}$ distribution (curve CA(3-82)), and so is the distribution of the test statistic when marker selection is combined with bias adjustment (curve SA). This confirms that the bias adjustment works.

The maximum value of the test statistic in an entire genome. Finally, we studied the type I error in a genome with 40 markers distributed over 8 chromosomes (Figure 5). In MQM mapping, we applied the variable selection and bias adjustment procedure, developed above for the case of many marker cofactors (curve SA). No or only a very few markers were selected (394 times no markers, 72 times one marker, 16 times two markers, 9 times three markers, 8 times four markers and only once five markers; together 500 simulation runs). The same case was reanalysed using the interval mapping method (curve I). The two distributions of the maximum value of the test statistic in the entire genome are very close to each other. This demonstrates that the results by LANDER and BOTSTEIN (1989) and VAN OOIJEN (1992) can be generally used to choose a threshold for the test statistic such that the probability of a type I error is about 5%. Of course, the methods differ with respect to the estimation of σ^2, so that at least small differences can be expected. Moreover, the thresholds from interval mapping are less appropriate for MQM mapping if there are only a few degrees of freedom for estimating σ^2. Further simulation should reveal the thresholds for these situations (see also discussion below).

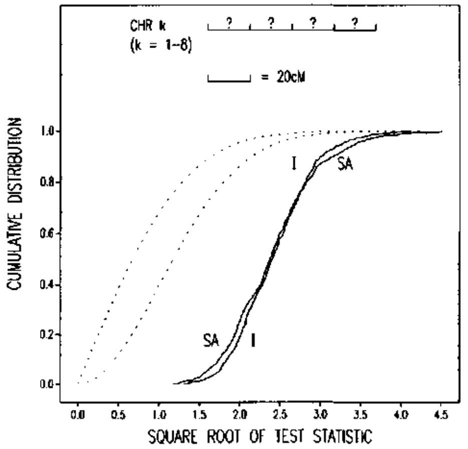


Figure 5. A study of the type I error in an entire genome (see Figure 1 legend)

Type II error

The distribution of the maximum value of the test statistic for the presence of a putative QTL was simulated in an interval in which a QTL is actually segregating. We successively considered the following situations: another QTL is also segregating and the two QTLs are either unlinked, linked in repulsion, or linked in coupling phase. See Figures 6-9 for the description of the QTL configurations. We studied the effect which the second QTL has on the test statistic for the presence of the first QTL. We say that the QTL is detected if the test statistic exceeds the threshold at a significance level of 5% for a 1000 cM genome. This means that we assume the simulated intervals to be part of a large genome. The value of this threshold is 2.4·2·ln(10)≈11.05 (LANDER and BOTSTEIN 1989; LOD threshold=2.4, see their Figure 4). The square root of the threshold is equal to 3.32.

Finally, we also considered the effect of bias adjustment of the estimate of the variance and the effect of a common estimate of the variance in sequences of models (Figure 9).

Two unlinked QTLs. First, we studied the case of two unlinked QTLs (Figure 6). In interval mapping, the first QTL in marker interval 2-3 is detected with a chance of 0.14 (curve I). In MQM mapping with markers 6 and 7 as cofactors the first QTL is detected with a chance of 0.74 (curve C(6,7)). In general the locations of the QTLs are unknown, so that markers to be used as cofactors should be selected. In some cases marker 1 or marker 4 may be selected and used as cofactor. The markers 1 and 4 are linked to the first QTL and can also (partially) absorb the effect of this QTL. As a consequence, the test statistic takes the smaller values in these cases (lower tail of curve S in Figure 6). Nevertheless, the chance of detecting the first QTL is still 0.70 when selected markers are used. This demonstrates that QTLs can be detected more powerfully by MQM mapping than by interval mapping (the chance of detection of the first QTL is 0.70 versus 0.14, respectively). As the expected genetic variance of the QTL forms the major part of the expected phenotypic variance (90%), these simulations show the maximal increase in power.

MQM mapping with marker cofactors was also compared to MQM mapping with exact models for two QTLs. To that order, a putative QTL, or no QTL, was fitted in marker interval 2-3, while in either case a second QTL was fitted in marker interval 6-7 (curve E in Figure 6). It is clear from Figure 6 that the curves C(6,7) (or curve S) and E are still rather different. This means that a proportion of the genetic variation of the second QTL could not be eliminated by the marker cofactors, due to recombinants between marker 6 (or 7) and the second QTL. Thus, MQM mapping with exact models for multiple QTLs is sometimes much more powerful than MQM mapping with marker cofactors. It is clearly beneficial to use exact models for those putative QTLs that have a major effect on the trait (and the corresponding marker cofactors can be dropped).

Two linked QTLs in repulsion phase. Next, we studied the case of two linked QTLs in repulsion phase (with effects of equal size but opposite sign; see Figure 7). In interval mapping, the first QTL in marker interval 2-3 is detected with a chance of only 0.42 (curve I). In MQM mapping with markers 4 and 5 as cofactors the first QTL is detected with a chance of 0.93 (curve C(4,5)). When only selected markers are used in MQM mapping (excluding the markers 2 and 3, which flank the interval under study), the chance of detecting the first QTL increases even to 0.97 (curve S). The increase is due to the fact that marker 4 sometimes partially absorbs the large effect of the first QTL, while marker 5 absorbs the large effect of the second QTL in marker interval 4-5; the value of the test statistic for the presence of the first QTL then increases by dropping marker 4 or 5. Our simulations demonstrate that linked QTLs in repulsion phase can be detected and separated much more powerfully by MQM mapping with marker cofactors than by interval mapping.

MQM mapping with marker cofactors was also compared to MQM mapping with

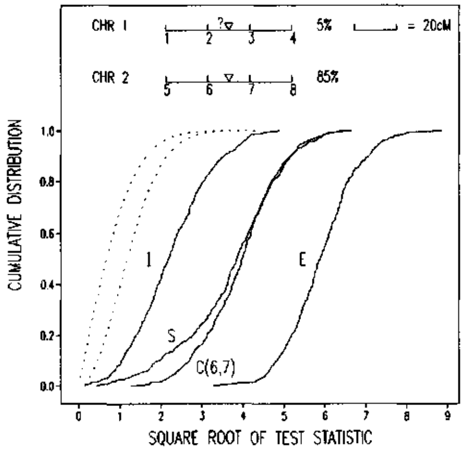


Figure 6. A study of the type II error in case of two unlinked QTLs (see Figure 1 legend)

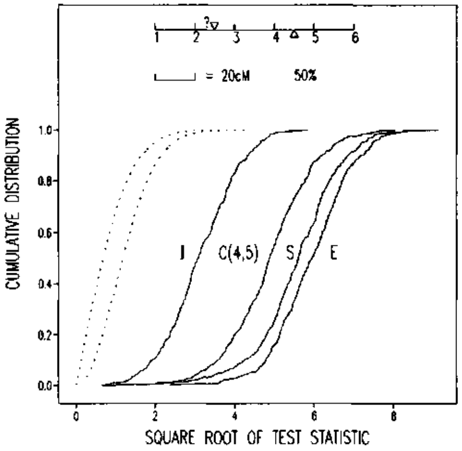


Figure 7. A study of the type II error in case of two linked QTLs in repulsion phase (see Figure 1 legend)

exact models for two QTLs. To that order, a putative QTL, or no QTL, was fitted in marker interval 2-3 while in either case a second QTL is fitted in marker interval 4-5 (curve E). It is clear from Figure 7 that curve S and curve E are only slightly different, i.e. MQM mapping with marker cofactors is almost as powerful as MQM mapping with exact models for multiple QTLs, when QTLs are in repulsion phase.

Two QTLs in coupling phase. Then, we studied the case of two linked QTLs in coupling phase (with effects of equal size and equal sign; Figure 8). In interval mapping, the test statistic for the presence of the first QTL in marker interval 2-3 exceeds the threshold with a chance of 1.00 (curve I). In MQM mapping with markers 6 and 7 as cofactors the test statistic for the first QTL exceeds the threshold with a chance of 0.92 (curve C(6,7)). Thus, in contrast to the results for the previous configurations, MQM mapping now leads to smaller values for the test statistic than interval mapping does. The reason for this is that the effect of the first QTL, but also the major part of the effect of the second QTL in marker interval 6-7, are absorbed when interval mapping of a single putative QTL is carried out in marker interval 2-3. In MQM mapping, the effect of the second QTL and that of the major part of the first QTL are absorbed by the marker cofactors 6 and 7; the test statistic for marker interval 2-3 gives the likelihood for the presence of multiple linked QTLs, one being located in marker interval 2-3, the other being located nearby markers 6 and 7. When only selected markers are used in MQM mapping (excluding markers 2 and 3, which flank the interval under study), the test

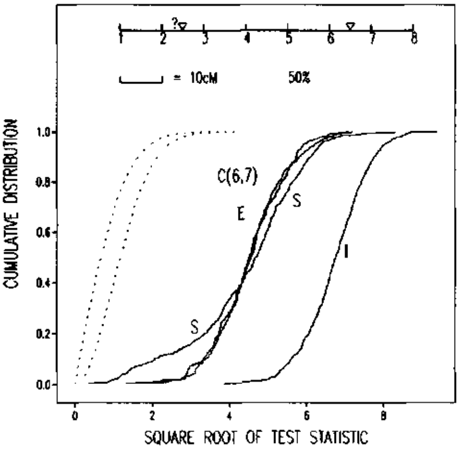


Figure 8. A study of the type II error in case of two linked QTLs in coupling phase (see Figure 1 legend)

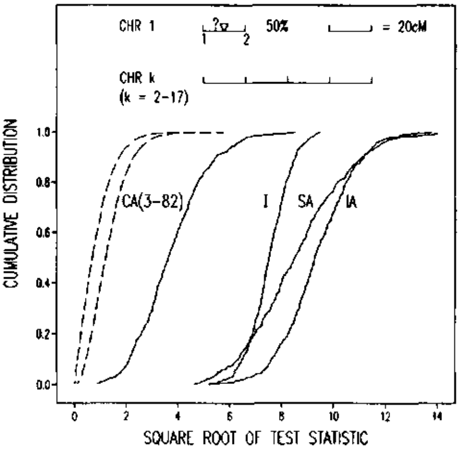


Figure 9. A study of the type II error in case of many marker cofactors (see Figure 1 legend)

statistic for the presence of a QTL in marker interval 2-3 decreases slightly (the chance of detection of the QTL is 0.81; curve S). The reason for this is that markers 1 or 4 were selected (as well) in some simulation runs. The upper tail of curve S exceeds the upper tail of curves E and C(6,7) slightly. The reason for this is that either marker 6 or marker 7 was selected in a number of simulations (rather than selecting markers 6 and 7 simultaneously), thereby absorbing slightly less variation induced by the QTLs.

MQM mapping with marker cofactors was also compared to MQM mapping with exact models for two QTLs. To that order, a putative QTL, or no QTL, was fitted in marker interval 2-3 while in either case a second QTL was fitted in marker interval 6-7 (curve E). It is clear from Figure 8 that curve C(6,7) and curve E are very close; however, the lower tail of curve S deviates from them. Thus, not unexpectedly, MQM mapping with selected marker cofactors is not always as powerful as MQM mapping with exact models for multiple QTLs, when QTLs are in coupling phase.

The effect of the number of cofactors. Finally, we will discuss the effect on the type II error of two changes to MQM mapping, namely the use of a single estimate of the variance in a sequence of models and the bias adjustment of the estimate of the variance (Figure 9). Firstly, the test statistic for the presence of a putative QTL in marker interval 1-2 on chromosome 1 is considered without using marker cofactors for other chromosomes (curves I and IA). Figure 9 shows the distribution for the test statistic when the usual maximum likelihood method is used (interval mapping with a separate estimate of the variance in the single and the no-QTL model; curve I) and also when the variance in the no-QTL model is fixed at the estimate from the single-QTL model (which was adjusted for bias; curve IA). In this case the bias adjustment of the estimate of the variance will be almost negligible. The simulations clearly demonstrate that the use of a common estimate of the variance can lead to a more powerful QTL detection.

Secondly, the test statistic for the presence of a putative QTL in marker interval 1-2 is considered when 80 markers (or a subset) on 16 other chromosomes are used as cofactors (curves CA(3-82) and SA). In such a case the estimate of the variance will be highly biased and bias adjustment is needed. The estimate of the variance in the most complex model (the model with all marker cofactors) was adjusted for bias as described above and we used this estimate as a common estimate in the subset selection procedure and in the single-QTL and the no-QTL models. The test statistic takes much smaller values when all 80 markers are used as cofactors (curve CA(3-82)) than it does when no marker cofactors are used (curve IA). This demonstrates that the 80 cofactors partially absorb the effect of the QTL in marker interval 1-2, even though these markers are located on other chromosomes. However, when preselected marker cofactors are used (curve SA), the distribution of the test statistic is much closer to the one found when no marker cofactors are used (curve IA).

DISCUSSION

The simulations presented in this paper demonstrate the great advantage of MQM mapping over interval mapping in controlling the chances of type I and type II errors. The nice feature of MQM mapping is that marker cofactors are generally selected only in regions were QTLs are segregating. Because of this feature, thresholds for the test statistic, which were obtained for the case that no QTLs are segregating (LANDER and BOTSTEIN 1989; VAN OOIJEN 1992), are also suitable for MQM mapping. These thresholds can be used when no QTLs are segregating, since in that case no or only a few markers will be selected; moreover, these thresholds can still be used when there are QTLs segregating, the effects of which are eliminated by marker cofactors. One condition is, however, that the residual degrees of freedom for estimating the variance (or the dispersion parameter in generalized linear models) are adequate. In such cases, the choice of the appropriate threshold for the test statistic (so that the chance of a type I error is small, say 5%) can be made satisfactorily in MQM mapping. Further simulation work is required to reveal the appropriate thresholds for the cases in which the number of residual degrees of freedom is small. In interval mapping, the threshold for the test statistic should be used with caution. It is known that a single QTL affects the test statistic in all intervals on the same chromosome; the test statistic often exceeds the threshold in a number of intervals on either side of the QTL, although one should not "detect" multiple QTLs in this region. In MQM mapping on the other hand, the effect of a QTL diminishes rapidly when the distance between the QTL and the interval of interest increases; a QTL often affects the test statistic only in the two intervals adjacent to the one of the QTL.

The use of marker cofactors reduces the unexplained variance, so that the chance of a type II error in the case of unlinked QTLs is generally smaller in MQM mapping than in interval mapping. Our simulations also demonstrate that the detection and unravelling of the separate QTL effects in the case of linked loci is much easier in MQM mapping than in interval mapping. Linked QTLs with opposite (and mutually neutralizing) effects are worst case configurations for interval mapping: often no QTLs will be detected. Also, linked QTLs with equal sign effects are a difficult configuration for interval mapping: often a single "ghost" QTL will be detected somewhere between the two QTLs (MARTINEZ and CURNOW 1992). Again, our simulations make it clear that separation of such QTLs is much easier in MQM mapping than in interval mapping.

In our simulations the progeny size is fixed at 100 individuals, because we are involved in real experiments of that size. For such cases, VAN OOIJEN (1992) demonstrated that the chance of detecting a specific QTL is small, unless the QTL explains a large proportion of the phenotypic variance. Therefore, we considered QTL configurations for relatively high levels of heritability. Our simulations make it clear that QTLs can be mapped more powerfully by MQM mapping than by interval mapping. In some of our simulations, the gene was even of qualitative rather than quantitative nature (Figure 1, 2a, 2b and 6). We expect that a similar power improvement can be achieved when several QTLs instead of one major gene contribute to the genetic variation. Furthermore, we expect that similar results can also be obtained for smaller levels of heritability if the progeny size is larger. On the other hand, in some types of progeny such as recombinant inbred lines, the heritability can be increased at will by using more plants per line, leading to similar configurations.

In QTL-mapping experiments, large numbers of markers are commonly scored. In this paper we addressed problems concerning fitting models with many marker cofactors, and problems concerning selection of "important" marker cofactors. The maximum-likelihood estimate of the residual variance will be biased when many markers are used as cofactors; the number of parameters should not be too large, preferably less than $2\sqrt{}$(number of observations) (JANSEN and STAM 1994). We propose a heuristic three-step procedure to adjust for the bias. Our simulations demonstrate that the bias adjustment works. This makes it possible to use many markers as cofactors in MQM mapping. However, the use of redundant marker cofactors can lead to a loss of detection power.

There are two causes for this loss of detection power: (a) any redundant marker which is used as a cofactor and which is located nearby the QTL can also (partially) absorb the effect of the QTL; and (b) the marker data are generally unbalanced so that the effect of a QTL can even be absorbed by redundant markers on other chromosomes, especially in small progenies. Therefore, selection of the "important" markers is beneficial. In order to exclude redundant markers, the selection criterion should be stringent, but not so stringent that important markers (those flanking the QTLs) are thereby excluded. JANSEN (1993b) proposed to maximize the log-likelihood minus the number of free parameters (*k*) in the model; this is equivalent to minimizing Akaike's Information Criterion AIC=-2($\mathcal{L}$-*k*). In general, a penalty in the range of *k* to 3*k* may provide plausible initial models (MCCULLAGH and NELDER 1989). In the present paper, we use the more stringent penalty of 3*k*. Our simulations demonstrate that (a) this penalty is stringent, since no or only a few markers are generally selected in the case of no QTLs segregating and (b) this penalty is still not too stringent, since markers are selected for those QTLs that considerably affect the test statistic in their nearby region; the effects of such QTLs are satisfactorily eliminated by selected markers. Nevertheless, we feel that it is still worthwhile to study the properties of the method for other levels of the penalty in the range from *k* to 3*k*. In particular, we believe that the penalty should depend on the aim of the experiment. For instance, consider an experiment in which the aim is prediction of phenotypic value followed by indirect selection via markers. In the case of prediction, the penalty should be probably *k* rather than 3*k* (MCCULLAGH and NELDER 1989). KNOTT and HALEY (1992) discuss another situation which should be investigated in more detail: a trait with a reasonable level of heritability, which is affected by very many genes of small effect distributed throughout the genome. In general, the benefit of using a small penalty is that more variation induced by QTLs is eliminated. The cost is loss of power, since also (many) redundant markers are selected. Also, the threshold for the test statistic should become more stringent, when the penalty decreases. In order to obtain thresholds as a function of the penalty and also as a function of the residual degrees of freedom, further simulation work should be done. LANDER and BOTSTEIN (1989) and VAN OOIJEN (1992) studied the mapping of a single QTL with no markers as cofactors (equivalent to a penalty of ∞) and ZENG (1994) studied the mapping of a single QTL with all markers as cofactors, except the ones flanking the interval under study (nearly equivalent to a penalty of 0). Simulation work by Zeng (1994; his Figure 1) demonstrated that $\chi^2_{2;\alpha/M}$ can be used as an upper bound for the 100α% threshold for the overall test with *M* intervals, unless the number of parameters is too large. It should now be obvious that the $\chi^2_{2;\alpha/M}$ relation does not hold if the number of parameters exceeds 2$\sqrt{}$(number of observations) (JANSEN and STAM 1994). Our work, however, makes it possible to fit properly models with many parameters. It also indicates that $2F_{2,df;\alpha/M}$ can be used as an upper bound, where df are the degrees of freedom for estimating σ^2. Finally, we note that our selection criterion applies not only to "ordinary" regression models, assuming a normal error distribution, but also to generalized linear models (GLMs; MCCULLAGH and NELDER 1989). In comparing a sequence of GLMs, a single estimate of the dispersion parameter (σ^2 in "ordinary" regression) based on the most complex model is usually considered.

In the present paper we study an automatic MQM mapping procedure. In practice the user may wish to step in interactively. Some marker cofactors could be dropped and others could be added by hand. Also, exact models for multiple QTLs could be fitted for those putative QTLs that have a major effect on the trait (and the corresponding marker cofactors may be dropped). One can still take into account the effects of less important putative QTLs by using marker cofactors. Also, exact models for two (or more) QTLs could be fitted to separate the effects of QTLs located in adjacent intervals. Such an interactive approach is possibly the most accurate and efficient way to map multiple QTLs, which is still feasible.

LITERATURE CITED

COWEN, N. M., 1989 Multiple linear regression analysis of RFLP data sets used in mapping QTLs, pp. 113-116 in *Development and application of molecular markers to problems in plant genetics*, edited by T. HELENTJARIS and B. BURR. Cold Spring Harbor Laboratory, N.Y.

HALEY, C. S., and 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., 1993a Maximum likelihood in a generalized linear finite mixture model by using the EM algorithm. Biometrics 49: 227-231

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

JANSEN, R. C., and P. STAM, 1994 High resolution of quantitative traits into multiple loci via interval mapping. Genetics 136: 1447-1455

KNOTT, S. A., and C. S. HALEY, 1992 Aspects of maximum likelihood methods for the mapping of quantitative trait loci in line crosses. Genet Res Camb 60: 139-151

LANDER, E. S., and D. BOTSTEIN, 1989 Mapping Mendelian factors underlying quantitative traits by using RFLP linkage maps. Genetics 121: 185-199

MARTINEZ, O, and 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

MCCULLAGH, P., and J. A. NELDER, 1989 Generalized linear models, in *Monographs on Statistics and Applied Probability* 37. Chapman & Hall, London

PATERSON, A. H., E. S. LANDER, J. D. HEWITT, S. PETERSON, S. E. LINCOLN, and S. D. TANKSLEY, 1988 Resolution of quantitative traits into Mendelian factors, using a complete linkage map of restriction fragment length polymorphisms. Nature 335: 721-726

RODOLPHE, F., and M. LEFORT, 1993 A multi-marker model for detecting chromosomal segments displaying QTL activity. Genetics 134: 1277-1288

STAM, P., 1991 Some aspects of QTL analysis, in *Proceedings of the Eighth Meeting of the Eucarpia Section Biometrics in Plant Breeding*. BRNO, July 1991

TITTERINGTON, D. M., A. F. M. SMITH, and U. E. MAKOV, 1985 Statistical analysis of finite mixture distributions. Wiley, N.Y.

VAN OOIJEN, J. W., 1992 Accuracy of mapping quantitative trait loci in autogamous species. Theor. Appl. Genet. 84: 803-811

ZENG, Z.-B., 1993 Theoretical basis of separation of multiple linked gene effects on 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

VI. GENOTYPE BY ENVIRONMENT INTERACTION IN GENETIC MAPPING OF MULTIPLE QUANTITATIVE TRAIT LOCI

ABSTRACT

The interval mapping method is widely used for the genetic mapping of quantitative trait loci (QTLs), though true resolution of quantitative variation into QTLs is hampered with this method. Separation of QTLs is troublesome, because single-QTL models are fitted. Further, genotype by environment interaction, which is of great importance in many quantitative traits, can only be approached by separately analyzing the data collected in multiple environments. Here, we demonstrate for the first time a novel analytic approach (MQM mapping) that accommodates both the mapping of multiple QTLs and genotype by environment interaction. MQM mapping is compared to interval mapping in the mapping of QTLs for flowering time in *Arabidopsis thaliana* under various photoperiod and vernalization conditions.

FLOWERING TIME IN *ARABIDOPSIS*

Arabidopsis thaliana is a model organism for genetic analysis because of its small genome size, short generation time and ease of propagation (MEYEROWITZ and PRUITT 1985). Transition to flowering is one of the current issues in *Arabidopsis* research (MARTINEZ-ZAPATER et al. 1994). At least twelve loci for flowering time were identified by mutational analysis (KOORNNEEF et al. 1991). Also, large differences between ecotypes exist for flowering time; the FRI locus was found to be responsible for some of these differences (CLARKE and DEAN 1994). The group of early genotypes, which includes the widely used ecotypes Columbia (Col) and Landsberg *erecta* (L*er*), was not analyzed extensively. Small differences in flowering time within this group have been reported (KOORNNEEF et al. 1991), and it has been suggested that the FLC locus is involved (KOORNNEEF, personal communication). Flowering time strongly depends on many environmental factors, amongst which photoperiod and temperature (vernalization treatment) are most important. Distinct norms of reaction have been reported for several mutants and ecotypes (MARTINEZ-ZAPATER et al. 1994). Here, we report the genetic mapping of quantitative trait loci (QTLs) underlying the differences in flowering time between Col and L*er*. Flowering time was recorded under various photoperiod and vernalization conditions in a set of recombinant inbred lines (RILs, Table 1) derived from a cross between Col and L*er*; details of the experimental conditions will be presented elsewhere (Lister and Dean, manuscript in preparation). We used 37 of the previously mapped RFLP markers (LISTER and DEAN 1994).

We successfully applied a novel method of analysis based on multiple-QTL models

Table 1 Genetic mapping of QTLs for flowering time (expressed by leaf number[a]) in *Arabidopsis thaliana*: some population parameters

Environmental conditions[b]	Phenotypic mean			Phenotypic variance between RILs	Multiple regression of RIL phenotypes on all 37 markers	
	Col[c]	Ler[c]	RILs[d]		Residual variance	Variance explained
LD	9.9	7.1	8.60	1.63	0.79	52%
LDV	9.0	7.4	8.58	0.38	0.23	39%
SD	32.9	28.3	29.41	29.69	10.12	66%
SDV	22.2	19.5	21.23	5.51	3.37	39%
CL	18.1	11.5	12.84	9.76	5.55	43%
CLV	11.3	8.3	10.29	0.78	0.37	53%

[a] Leaf number is often taken as a measure of flowering time; leaf numbers in this table represent the total number of rosette and cauline leaves per plant.
[b] SD=short day (10 hours of light); LD=long day (16 hours of light); CL=continuous light; LDV, SDV and CLV=LD, SD and CL + vernalization, respectively; Col=Columbia and Ler=Landsberg *erecta*
[c] Two sets of five plants per environment were tested.
[d] In total 99 recombinant inbred lines (RILs) were tested, each RIL with five plants per

Fig. 1 Genetic mapping of QTLs for flowering time (expressed by leaf number) in *Arabidopsis thaliana*: QTL likelihood maps and QTL effect maps produced by interval mapping (IM). Chromosome number is indicated at the right-hand top corner of each graph, markers are plotted along the abscissa. The solid, dashed and dotted curves represent the test statistic (twice the log of the likelihood ratio) for the hypothesis of a QTL (with no QTL by environment interaction) in the environment indicated. The overall 5% significance threshold for the test is 10. Solid, dashed and dotted bars represent two lod (^{10}log of likelihood ratio) support intervals for the map locations of detected QTLs (pattern of curves and bars are corresponding). SD=short day (10 hours of light); LD=long day (16 hours of light); CL=continuous light; LDV, SDV and CLV=LD, SD and CL + vernalization, respectively.

Fig. 2 Genetic mapping of QTLs for flowering time (expressed by leaf number) in *Arabidopsis thaliana*: QTL likelihood maps and QTL effect maps produced by MQM mapping. Chromosome number is indicated at the right-hand top corner of each graph, markers are plotted along the abscissa. Selected markers are indicated by '+' when interaction with environment is still assumed, otherwise by '^'. Solid curves indicate the test statistic (twice the log of the likelihood ratio) for the hypothesis of a QTL with no QTL by environment interaction assumed (upper part) and the estimated QTL effect (lower part). The overall 5% significance threshold for this test is 11. Dashed curves represent the test statistic for the hypothesis of a QTL with QTL by environment interaction (upper part) and the estimated QTL effects (lower part). The overall 5% significance threshold for the interaction test (the difference between solid and dashed curve) is 19. Bars along the abscissa indicate two lod (^{10}log of likelihood ratio) support intervals for the map locations of detected QTLs. The QTL effect is expressed proportionally, i.e. the replacement of the putative QTL allele of Col by that of *Ler* (a) has no effect if the QTL effect is equal to 1, (b) proportionally increases the number of leaves, if the QTL effect is larger than 1 and (c) proportionally decreases the number of leaves, if the QTL effect is smaller than 1. SD=short day (10 hours of light); LD=long day (16 hours of light); CL=continuous light; LDV, SDV and CLV=LD, SD and CL + vernalization, respectively.

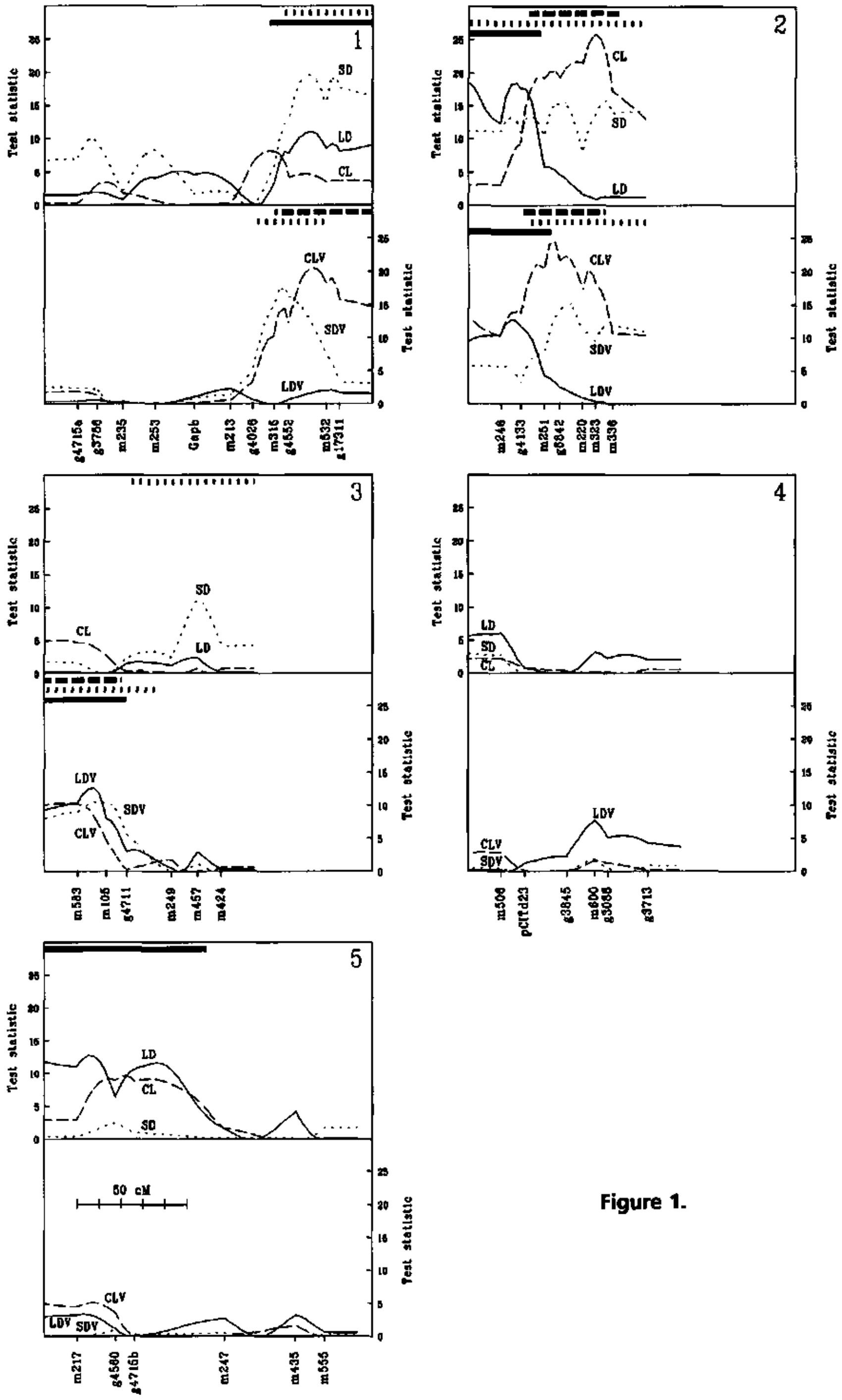

1
2
3
4
5
Test statistic
0
5
10
15
20
25
SD
LD
CL
CLV
SDV
LDV
g4715a
g3766
m235
m253
Gapb
m213
g4026
m316
g4552
m532
g17311
m246
g4133
m251
g6842
m220
m323
m336
m583
m105
g4711
m249
m457
m424
m506
pCITd23
g3845
m600
g3088
g3713
m217
g4560
g4716b
m247
m435
m555
50 cM

Figure 1.

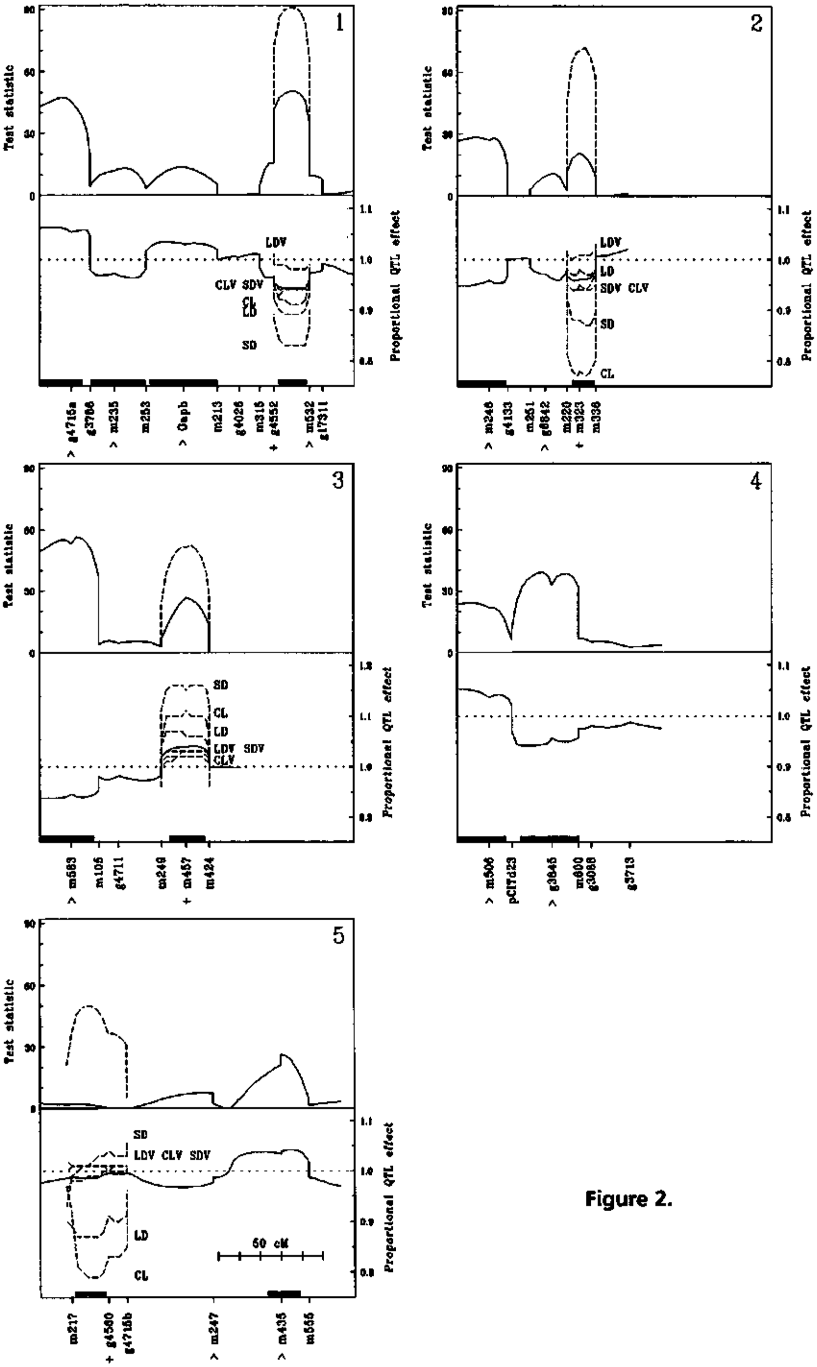

1
Test statistic
0
30
60
90
Proportional QTL effect
0.8
0.9
1.0
1.1
LDV
CLV SDV
CL
LD
SD
> g4715a
g3786
> m235
m253
> Oapb
m213
g4026
m315
+ g4552
> m532
g17311
2
LDV
LD
SDV CLV
SD
CL
> m246
g4133
m251
> g6842
m220
+ m323
m336
3
SD
CL
LD
LDV SDV
CLV
1.2
> m583
m105
g4711
m249
+ m457
m424
4
> m506
pCI7d23
> g3645
m600
g3086
g3713
5
SD
LDV CLV SDV
LD
CL
50 cM
m217
+ g4580
g4715b
> m247
> m435
m555

Figure 2.

environment.(MQM mapping; JANSEN 1994); QTL by environment interaction is part of the models fitted. For comparison we also applied interval mapping (LANDER and BOTSTEIN 1989), analyzing the data for each environment separately.

COMPARISON OF INTERVAL MAPPING AND MQM MAPPING

In interval mapping (IM) the likelihood for a single putative QTL is assessed at each map location on the genome and, in case of multiple environments, a QTL likelihood map is produced for each environment separately (Fig. 1). MQM mapping is an automatic two-stage procedure in the first stage of which "important" markers and marker by environment interactions are identified by the backward elimination method in multiple regression on all markers and environments (including interactions). In the second stage the likelihood for a single putative QTL is assessed at each map location (like in IM), but the preselected markers are used as cofactors, except for selected markers flanking the interval under study (Fig. 2). Marker cofactors will hopefully eliminate the major part of the variation induced by QTLs located elsewhere on the genome. A likelihood curve for a single putative QTL with QTL by environment interaction is plotted in regions where interaction with environment is still assumed (Fig. 2). The overall 5% significance thresholds for the tests (for the hypothesis of a QTL with no QTL by environment interaction and for the hypothesis of QTL with QTL by environment interaction) were obtained by computer simulation, using the actual marker data and analyzing 1000 replicates. Observed flowering times were log-transformed prior to analysis and distinct variance parameters for each of the environments were included.

With MQM mapping, we found evidence for twelve QTLs; four of these display QTL by environment interaction (Fig. 2). The QTL by environment effects indicate a QTL by vernalization interaction, where vernalization decreases the effects of these QTLs (Fig. 2). This result is not unexpected, because vernalization considerably decreases both environmental and genetic variance (Table 1). Further, QTL by photoperiod interactions are indicated. For instance, on chromosome 2 the QTL near marker m323 has little effect at LD but a large effect at CL. A full analysis of the identified QTLs and their relationship to previously mapped flowering time loci will be presented elsewhere (Lister and Dean, manuscript in preparation).

We now compare the results of MQM mapping with those of IM and discuss the features of MQM mapping and IM that contribute to the differences. In IM, single-QTL models are used and independence of residual errors is a basic assumption. In our experiment, however, each RIL is tested in six environments and the six observations are correlated via genetic identity of the underlying genes. Therefore, the usual assumption of independent residual errors may be seriously violated and a joint analysis accommodating the information from all environments is not possible with IM; QTL likelihood maps can only be produced for each environment separately (Fig. 1). Although environment-specific QTLs may be detected this way, the approach is intrinsically weak, because the interaction is not part of the genetic model that is being fitted with IM. In MQM mapping with a complete linkage map however, the major part of this correlation is removed by markers which are used as cofactors in the model. This makes it possible to produce a joint map, including QTL by environment interaction, in the univariate regression frame of MQM mapping (Fig. 2).

The IM analysis indicates the presence of at least four QTLs in several environments (Fig. 1). The fact that IM detects a QTL at a specific map region in one environment but not in another environment, may indicate QTL by environment interaction (for instance, a QTL is detected near marker g4552 on chromosome 1 in environments SD and LD but not in CL). In the absence of true QTL by environment interaction, however, a QTL can also be detected in one environment and not in another environment, because the chance of simultaneous detection in both environments is small. Therefore, the IM analysis may be indicative but can not be conclusive on the presence of QTL by environment interaction. However, if a pattern of environment-specific QTLs really results from QTL by environment interaction, this is readily, and more powerfully detected by MQM mapping.

In MQM mapping, genetic background "noise" is removed by using marker cofactors (JANSEN 1994). Therefore, the chance of detecting QTLs is generally higher in MQM mapping than in IM. Further, separation of linked QTLs is much easier in MQM mapping than in interval mapping (JANSEN 1994). In IM, linked QTLs of unidirectional effect tend to be mapped as a single "ghost-QTL" at some intermediate position on the marker map (MARTINEZ and CURNOW 1992; JANSEN 1994); also, linked QTLs of opposite effect may go unnoticed because of their mutually neutralizing effects (JANSEN 1994). Both situations and even the more complex configuration of multiple linked QTLs with effects of alternating sign, have been encountered in our *Arabidopsis* experiment (Figs. 1 and 2). For instance on chromosome 2, MQM analysis indicates the presence of two QTLs. In IM, the QTL near m246 is found in LD and LDV. In the other environments a QTL is mapped at various positions (in the middle of the chromosome near g6842 in CLV, SDV and SD, and near m323 in CL), but support intervals are very large. This illustrates the problems in separating linked QTLs with unidirectional effects by IM. Here, the situation is even more complex due to QTL by environment interaction for the QTL near m323. Another example of linked QTLs is indicated on chromosome 3, where MQM mapping detects two QTLs. In IM, the QTL near m583 is detected in LDV, SDV and CLV but not in LD, SD and CL. The presence of the second QTL, with opposite effect only in LD, SD and CL, near m457 is one of the reasons for this. The other chromosomes exemplify similar problems in mapping of linked QTLs by IM.

The present study clearly illustrates the advantages of the MQM approach over IM in detection and mapping of multiple genes underlying quantitative traits, especially when data have been collected in multiple environments. Therefore, we feel that our approach is another step forward towards understanding the genetics of quantitative characters. Our results also suggest that re-analysis of several QTL experiments reported in literature (cf. PATERSON et al. 1988,1991; STUBER et al. 1992; DE VICENTE and TANKSLEY 1993; HAYES et al. 1993; SCHÖN et al. 1993; LAURIE et al. 1994) may further lift the veil that covers the link between phenotype and genotype.

LITERATURE CITED

CLARKE, J.H., and C. DEAN, 1994 Mapping FRI, a locus controlling flowering time and vernalization response. Mol Gen Genet 242:81-89

DE VICENTE, M.C. and S.D. TANKSLEY, 1993 QTL analysis of transgressive segregation in an interspecific tomato cross. Genetics 134:585-596

HAYES, P.M., B.H. LIU, S.J. KNAPP, F. CHEN, B. JONES, T. BLAKE, J. FRANCKOWIAK, D. RASMUSSON, M. SORRELLS, S.E. ULLRICH, D. WESENBERG and KLEINHOFS A, 1993 Quantitative trait locus effects and environmental interaction in a sample of North american barley germ plasm. Theor Appl Genet 87:392-401

JANSEN, R.C., 1994 Controlling the type I and type II errors in mapping quantitative trait loci. Genetics (in press)

KOORNNEEF, M., C.J. HANHART AND J.H. VAN DER VEEN, 1991 A genetic and physiological analysis of late flowering mutants in *Arabidopsis thaliana*. Mol Gen Genet 229:57-66

LANDER, E.S., and D. BOTSTEIN, 1989 Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185-199

LAURIE, D.A., N. PRATCHETT, J.H. BEZANT and J.W. SNAPE, 1994 Genetic analysis of a photoperiod response gene on the short arm of chromosome 2(2H) of *Hordeum vulgare* (barley). Heredity 72:619-627

LISTER, C., and C. DEAN, 1993 Recombinant inbred lines for mapping RFLP and phenotypic markers in *Arabidopsis thaliana*. The Plant Journal 4:745-750

MARTINEZ, O, and 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

MARTINEZ-ZAPATER, J.M., G. COUPLAND, C. DEAN and M. KOORNNEEF, 1994 The transition to flowering in *Arabidopsis*. In: Sommerville SR, Meyerowitz EM (eds) *Arabidopsis*. Cold Spring Harbor (in press)

MEYEROWITZ, E.M., and R.E. PRUITT, 1985 *Arabidopsis thaliana* and plant molecular genetics. Science 229:1214-1218

PATERSON, A.H., E.S. LANDER, J.D. HEWITT, S. PETERSON, S.E. LINCOLN and S.D. TANKSLEY, 1988 Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. Nature 335:721-726

PATERSON, A.H., S. DAMON, J.D. HEWITT, D. ZAMIR, H.D. RABINOWITCH, S.E. LINCOLN, E.S. LANDER and S.D. TANKSLEY, 1991 Mendelian factors underlying quantitative traits in tomato: comparison across species, generations and environments. Genetics 127:181-197

SCHÖN, C.C., M. LEE, A. MELCHINGER, W.D. GUTHRIE and W.L. WOODMAN, 1993 Mapping and characterization of quantitative trait loci affecting resistance against second-generation European corn borer in maize with the aid of RFLPs. Heredity 70:648-659

STUBER, C.W., S.E. LINCOLN, D.W. WOLFF, T. HELENTJARIS and 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

VII. MAXIMUM LIKELIHOOD IN A GENERALIZED LINEAR FINITE MIXTURE MODEL BY USING THE EM ALGORITHM

ABSTRACT

A generalized linear finite mixture model and an EM algorithm to fit the model to data are described. By this approach the finite mixture model is embedded within the general framework of generalided linear models (GLMs). Implementation of the proposed EM algorithm can be readily done in statistical packages with facilities for GLMs. A practical example is presented were a generalized linear finite mixture model of ten Weibull distributions is adopted. The example is concerned with the flow cytometric measurement of the DNA content of spermatids in a mutant mouse, which shows non-disjunction of specific chromosomes during meiosis.

INTRODUCTION

Generalized linear models (GLMs) have been proved very useful in many agricultural and biological applications (MCCULLAGH and NELDER 1989). Surprisingly, little attention has been paid to the use of GLMs in finite mixture models. In the past decades much literature on finite mixture models appeared, including important monographs by EVERITT and HAND (1981), TITTERINGTON, SMITH and MAKOV (1985), and MCLACHLAN and BASFORD (1988). The more straightforward situation is commonly dealt with, where the components have separate parameters for mixing proportions and separate parameters for mixing distributions. In this paper it is shown that, by adopting a simple EM algorithm (DEMPSTER, LAIRD and RUBIN 1977), the mixture problem can be split into two solvable non-mixture problems. This makes it possible to transfer all GLM facilities to the corresponding finite mixture equivalent. Moreover, standard statistical packages can be readily used to do the computational work. A general procedure, which requires specification of the GLM for the mixing proportions and specification of the GLM for the mixing distributions, can be easily written in for instance GENSTAT (GENSTAT 5 COMMITTEE 1987). The distribution of the component counts may be either multinomial or Poisson. The mixing distribution can be for example univariate normal, Weibull, binomial or Poisson, but also for example multivariate normal or grouped normal. An illustration using data on non-disjunction in the mouse will also be given.

A GENERALIZED LINEAR FINITE MIXTURE MODEL

DEMPSTER, LAIRD and RUBIN (1977) considered the mixture problem as one of many examples in which the data can be viewed as incomplete. They interpreted the mixture data as incomplete data by regarding an observation on the mixture as missing its component (or category) of origin. Complete data models and incomplete data models are discussed in the next two sections.

Complete data. Suppose that each individual in a sample of size *N* is classified to one of *M* categories. Let the random variable N_j denote the number of individuals in category j (j=1,2...*M*), and let the random variable Y_j denote the response of individual i with respect to an observable quantity (i=1...*N*).

In some cases it may be assumed that the numbers $N_1,N_2,\ldots,N_M$ are independently distributed according to the Poisson distribution. More commonly there are constraints on the $N_1,N_2,\ldots,N_M$. For example, the total number of observations *N* is fixed. In that case the joint distribution of the $N_1,N_2,\ldots,N_M$ is the multinomial distribution. The usual hypotheses can all be formulated as multiplicative models (GLMs for count data).

The response variable is often assumed to be normally distributed, in which case the usual hypotheses can be formulated as regression models. The distribution of the response variable is assumed to depend on the category, which is therefore one of the explanatory variables. The response variable may have some other continuous distribution, such as the log-normal or the Weibull distribution. It may even be discrete rather than continuous, such as is the case when percentages, counts, grouped or ordinal data are recorded. GLMs provide an extension of classical linear models for normally distributed data to these and many other types of data.

Incomplete data. Suppose now that it cannot be observed to which category an individual belongs. The observed response variable does now have a finite mixture distribution. The categories are usually referred to as components of the mixture. The previously described GLMs for the counts $N_1,N_2,\ldots,N_M$ and the responses $Y_1,Y_2,\ldots,Y_M$ may still hold, and the model may now be called a 'generalized linear finite mixture model'. It will be shown below that the parameters can be estimated by fitting GLMs to updated complete data in an iterative way.

The case of a continuous response variable will be considered here. Expressions for a discrete response variable may be obtained by substituting probabilities for densities. Suppose that the likelihood of the i-th observation is

$$f(y_i)=\sum_{i=1}^{M} p_j f_j(y_i) \ ,$$

where $f_j(\cdot)$ is the probability density function of the j-th component. The likelihood equations are

$$0=\sum_{i=1}^{N}\sum_{j=1}^{M}p_{j|i}\frac{\partial}{\partial\theta}\log p_j+\sum_{i=1}^{N}\sum_{j=1}^{M}p_{j|i}\frac{\partial}{\partial\theta}\log f_j(y_i)\ ,$$

where

$$p_{j|i}=\frac{p_j f_j(y_i)}{f(y_i)}$$

(EVERITT and HAND 1981). The likelihood equations can be solved by applying the EM algorithm (DEMPSTER, LAIRD and RUBIN 1977). Each iteration consists of two steps. First, in the so-called E-step, $p_{j|i}$ is evaluated given the current parameter estimates. Next, in the so-called M-step, the likelihood equations are solved by fixing the $p_{j|i}$ whereby new parameter estimates are obtained. Note that in this case the likelihood equations are split into two terms; the first term is a function of the mixing proportions only, the second term is a function of the parameters of the component distributions only. Therefore the estimation of mixing proportions and the estimation of parameters of the component distributions are separated in the iterative scheme (EVERITT and HAND 1981). Until now it has not been recognized that each term can be treated as a likelihood equation for non-mixture problems of $N\times M$ observations. Each of the N observations is replicated and the number of replicas is equal to the number of components M. As a result, an M-step for the mixture problem can be split into two M-steps for standard non-mixture problems. This makes it possible to embed the finite mixture model into the general framework of GLMs. The first problem is solved by fitting a GLM for multinomial or Poisson data to the 'data' $p_{j|i}$. The second problem is solved by fitting a GLM to the response variable by using weights $p_{j|i}$.

APPLICATION

An example is now discussed where the fitting of a generalized linear finite mixture model provides an informative interpretation of the data.

The DNA contents of 6817 spermatids of a mutant mouse, which shows meiotic non-disjunction of specific chromosomes, and the DNA contents of 5488 spermatids of a control mouse, were measured by flow cytometry (FCM). Fig. 1 shows the observed FCM histograms. The control mouse produces spermatids of two different DNA content levels, namely spermatids carrying an X-chromosome and spermatids carrying an Y-chromosome. In addition to these two spermatid types, the mutant mouse was known to produce eight other spermatid types resulting from meiotic non-disjunction of chromosomes 11 or 13^1 (a so-called reciprocal translocation was used to enlarge chromosome 13 to 13^1). As a consequence of non-disjunction some spermatids have one

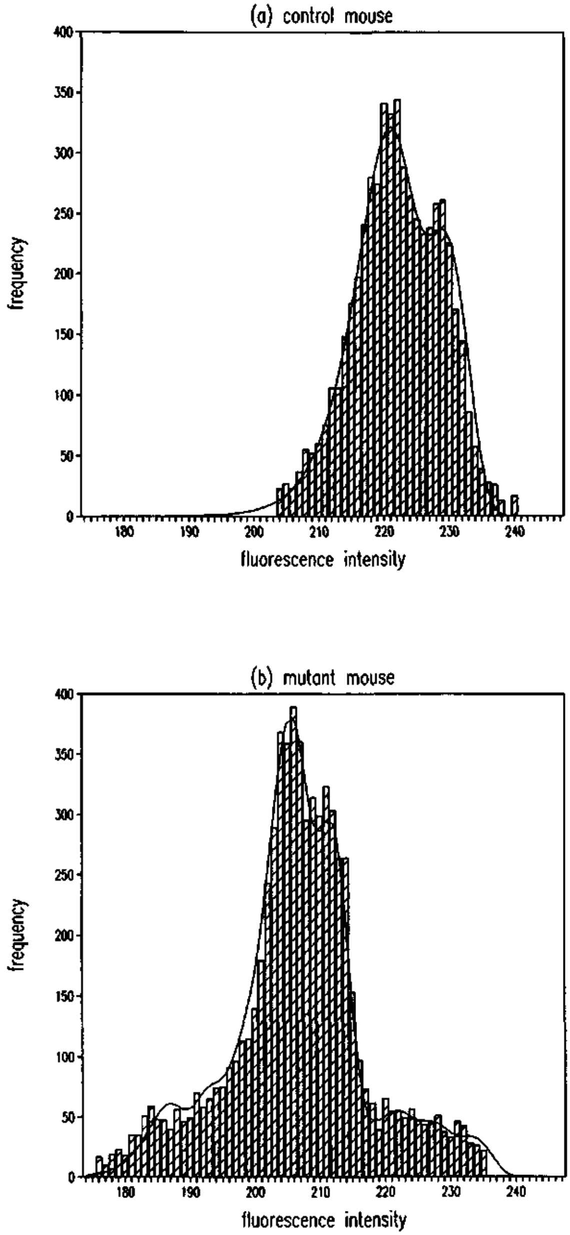


Figure 1. Flow cytometry (FCM) histograms of 5488 spermatids from a control mouse and of 6817 spermatids from a mutant mouse, which shows meiotic non-disjunction of specific chromosomes. Fitted mixtures of Weibull distributions are superimposed on top of the histograms.

extra chromosome, either chromosome 11 or 13[1], while other spermatids lack one of these. In accompanying cytological experiments no indication was found for simultaneous non-disjunction of chromosomes 11 and 13[1]. Conventional methods to estimate non-disjunction frequencies are based on chromosome counting. Statistical evaluation of the above data should reveal the usefulness of flowcytometry as a new and fast method for estimating non-disjunction frequencies in the mouse.

Let y_x and y_y denote the DNA content of normal haploid spermatids carrying an X-chromosome and spermatids carrying an Y-chromosome, respectively. Next, let y_{11} and y_{13} denote the DNA content of chromosome 11 and 13[1], respectively. Obviously, the DNA contents of the four hyperploid and the four hypoploid spermatid types can be expressed in terms of y_x, y_y, y_{11} and y_{13} in a linear way. For example, the DNA content of an X-chromosome carrying spermatid which lacks chromosome 11, denoted by $y_{x,-11}$, is given by $y_{x,-11}= y_x-y_{11}$. Similarly, the DNA content of an X-chromosome carrying spermatid with an extra chromosome 11, denoted by $y_{x,+11}$, is given by $y_{x,+11}= y_x+y_{11}$. The DNA contents $y_{x,-13}$, $y_{x,+13}$, $y_{y,-11}$, $y_{y,+11}$, $y_{y,-13}$ and $y_{y,+13}$ are defined and expressed in terms of y_x, y_y, y_{11} and y_{13} in the analogous way.

The expected frequencies of spermatids carrying an X-chromosome and spermatids carrying an Y-chromosome (denoted by p_x and p_y, respectively) are equal, so that in control mice $p_x=p_y=½$. The expected frequencies of the ten spermatid types in mutant mice are derived from the frequencies of non-disjunction. Let P_{11} and P_{13} denote the probability of non-disjunction for chromosome 11 and 13[1], respectively. The frequencies of the ten spermatid types, using analogous definitions for their frequencies as for their DNA contents, satisfy the equations $p_x=p_y=½(1-P_{11}-P_{13})$, $p_{x,-11}=p_{x,+11}=p_{y,-11}=p_{y,+11}=¼P_{11}$ and $p_{x,-13}=p_{x,+13}=p_{y,-13}=p_{y,+13}=¼P_{13}$. In the log-linear formulation offsets appear, i.e.,

$$\log(p_x)=\log(p_y)=\log(½)+\log(1-P_{11}-P_{13}),$$

$$\log(p_{x,-11})=\log(p_{x,+11})=\log(p_{y,-11})=\log(p_{y,+11})=\log(¼)+\log(P_{11}),$$

$$\log(p_{x,-13})=\log(p_{x,-13})=\log(p_{y,-13})=\log(p_{y,+13})=\log(¼)+\log(P_{13}).$$

We supposed that the FCM measurement x arose from a mixture of Weibull distributions, i.e. that, using the notation of McCULLAGH and NELDER (1989), the probability density function $f(x|y)$ for spermatids with DNA content y is given by $f(x|y)=\alpha x^{\alpha-1}\exp(y-x^{\alpha}\exp(y))$. Parameter estimation was carried out by the method described above. The computational work could be done easily in GENSTAT by exploiting its offset and weighting options for generalized linear models (GENSTAT 5 COMMITTEE 1987). The estimated frequency distributions are plotted on top of the histograms (Figure 1), the fit being satisfactory to the main body of the data. It should be noted that the FCM histograms were (unfortunately) automatically thresholded at the lower and upper tail of the distribution to eliminate (a low level of) background noise. An adaptive procedure is to add an extra component to the mixture distribution to take background effects into account. The exponential distribution (the Weibull distribution with α=1) is often used in flow cytometry to model background noise (cf. BALDETORP, DALBERG and LINDGREN 1989).

Table 1. Results of fitting a generalized linear finite mixture model of ten and two Weibull distributions to a mutant and a control mouse, respectively

	y_x	y_y	y_{11}	y_{13}	α	P_{11}	P_{13}
Mutant mouse	-366.25	-368.33	2.91	6.74	68.80	.45	.46
Control mouse	-288.28	-290.43	-	-	53.50	-	-

Parameter estimates are presented in Table 1. All standard deviations were close to 0, due to the huge numbers of observations. Estimates of y_x and y_y differ between control and mutant mice, so that accurate quantification of DNA contents of spermatids is unfeasible. However, the estimate of the sex-chromosome effect y_x–y_y is fairly constant over the two mice (≈2.1). The estimated DNA contents of chromosome 11 and chromosome 13[1] are 1.4 and 3.2 times the estimated sex-chromosome effect, respectively. These values are close to 1.4 and 3.1 respectively, which values can be derived from estimated chromosome lenghts presented in genetical literature (EVANS 1989).

LITERATURE CITED

Baldetorp, B., M. Dalberg and G. Lindgren, 1989 Statistical evaluation of cell kinetic data from DNA flow cytometry (FCM) by the EM algorithm. Cytometry 10: 695-705

Dempster, A.P., N.M. Laird and D.B. Rubin, 1977 Maximum likelihood from incomplete data via the EM-algorithm. J R Statist Soc B, 39: 1-38

Evans, E.P., 1989 Standard normal chromosomes, standard idiogram, in *Genetic variants and strains of the laboratory mouse*, edited by Lyon, M.F., and A.G. Searle. Oxford University Press, Oxford: 576-578

Everit, B.S., and D.J. Hand, 1981 *Finite mixture distributions*. Chapman and Hall, London

Genstat 5 Committee, 1987 *Genstat 5 reference manual*. Clarendon Press, Oxford

McCullagh, P., and J.A. Nelder, 1989 *Generalized linear models*. Chapman and Hall, London

McLachlan, G.J., and K.E. Basford, 1988 *Mixture models. Inference and applications to clustering.* Marcel Dekker, New York

Titterington, D.M., A.F.M. Smith and U.E. Makov, 1985 *Statistical analysis of finite mixture distributions.* New York, Wiley

SUMMARIZING DISCUSSION

In this thesis a new, general and powerful method is developed for the detection and mapping of QTLs in plants. The method is termed "MQM mapping", where MQM is an acronym for "multiple-QTL models" as well as for "marker-QTL-marker". The first term indicates that the models take into account the individual effects of the QTLs on the trait as well as their joint effect. The second term reflects the fitting of putative QTLs between markers on the genetic linkage map. The contributions of our method to progress in QTL mapping methodology (and some related topics) are summarized and discussed below.

Related methods. The interval mapping method (LANDER and BOTSTEIN 1989) has become the most widely used method for QTL analysis. In this method the likelihood for the presence of a single segregating QTL is assessed for each location on the genetic map. In statistical sense the trait is regressed on a single putative QTL and the unknown QTL-genotype is recovered via marker and phenotypic data as best it may. The use of single-QTL models in interval mapping is, however, in clear contradiction with the commonly assumed oligogenic or polygenic nature of quantitative traits. It is now generally recognized that interval mapping has several serious deficiencies. In interval mapping "ghost-QTLs" may be falsely detected between linked QTLs with unidirectional effects (coupling phase), or no QTL may be detected in the case of linked QTLs with opposite effects (repulsion phase). Moreover, the method is not powerful since QTLs are mapped one at a time, ignoring the effects of mapped or not yet mapped other QTLs. These deficiencies were already recognized to some extent by LANDER and BOTSTEIN (1989) and they made the first move towards fitting two QTLs simultaneously. Several authors (cf. KNAPP 1991; HALEY and KNOTT 1992; MARTINEZ and CURNOW 1992) subsequently developed methods for dissecting the effects of two or three linked QTLs.

Another method of QTL analysis is based on standard multiple regression of the trait on markers (cf. COWEN 1989; STAM 1991). In this approach the effects of QTLs will be absorbed by nearby markers; or in statistical sense, markers are treated as if they are QTLs themselves. Of course the regression models only approximate genuine multiple-QTL configurations. Genome regions displaying QTL activity can be found by testing for and selection of the influential markers (hopefully markers flanking QTLs will be traced). Unfortunately, application of the multiple regression method is seriously hampered if part of the marker observations is missing. In practice this is nearly always the case. In this thesis, we develop a solution to this problem (see below). [*chapters II-VI*]

MQM mapping. The statistical framework of MQM mapping consists of multiple linear regression of the trait on any set of loci (QTLs and markers). The MQM mapping framework therefore includes interval mapping and multiple regression on markers. Moreover, in MQM mapping it is possible to map multiple QTLs simultaneously. In theory the latter approach leads to the most efficient and most accurate mapping of multiple QTLs. However, when many QTLs are included in a multiple-QTL model, computation may become unfeasible (see also below). In this thesis an approximation of the genuine multiple-QTL configuration is described and extensively studied: a single putative QTL is moved along the chromosomes and the exact position of this QTL is assessed. Other putative QTLs are replaced by nearby markers (these markers are used as 'cofactors' in the model). In other words, these QTLs are 'located' at nearby marker positions instead of at their true (but unknown) positions between markers. This method is a combination of interval mapping and standard multiple regression on the markers. Recently, ZENG (1994) studied a similar approach. We focus on an automatic two-stage procedure, in the first stage of which "important" markers (hopefully those flanking the QTLs) are selected for multiple regression on markers. In the second stage a putative QTL is moved along the chromosomes by using the preselected markers as cofactors, except for selected markers flanking the interval under study.

In MQM mapping also models can be fitted in which two or three QTLs are combined with marker cofactors that eliminate effects of other QTLs. Such an approach is currently the most efficient and most accurate way to map multiple QTLs, which is still computationally feasible. [*chapters II-VI*]

Exploiting the full power of complete linkage maps of markers. Complete linkage maps of molecular markers are now available for many species. In MQM mapping the trait is regressed on these marker loci simultaneously and thereby the full power of complete linkage maps is exploited as much as it is computationally feasible, to detect and map QTLs. We address problems concerning the selection of "important" marker cofactors; a backward-elimination procedure for marker selection is described. Problems concerning the fitting of models with many marker cofactors, the over-fitting of the data and the estimation of the error variance are solved. Related approaches for marker selection have recently been discussed by HACKETT (1994).

For proper selection of markers close to QTLs, a reasonable number of recombinants between flanking markers is required. Because of the near collinearity of closely linked marker cofactors, it makes little sense to use a very dense map in a progeny of, say, 100 individuals. In order to increase the resolution, one should increase the progeny size (to produce more recombinants) rather than the number of markers. [*chapters II-VI*]

QTLs: how often are they missed, or mapped at the wrong location? In practice often a number of QTLs will be missed (a type II error) and at the same time a number of false positives may occur, indicating QTLs at map positions where actually no QTLs are present (a type I error). One often strives for keeping at least the chance of a type I error below 5%, while at the same time the chance of a type II error should be minimized. Our simulation work demonstrates that the chances of type I and type II errors are generally much smaller in MQM mapping than in interval mapping. The reasons are that the unexplained variance is much smaller in MQM mapping and that linked QTLs can be detected and separated much better by MQM mapping. The chance of detecting a "ghost-QTL" is also much smaller in MQM mapping than in interval mapping.

In interval mapping, the threshold for the test statistic should be used with caution. It is known that a single QTL affects the test statistic in all intervals on the corresponding chromosome; the test statistic often exceeds the threshold in a number of intervals on either side of the interval under study, although no multiple QTLs are present. In MQM mapping on the other hand, the influence of a QTL diminishes rapidly when the distance between the QTL and the interval under study increases; a QTL often affects the test statistic only in the two intervals adjacent to the one that contains the QTL.

ZENG (1994) studied the use of all markers as cofactors, except the markers flanking the interval under study, and developed an upperbound for the corresponding type I error. The benefit of using many (or all) markers as cofactors is the elimination of as much QTL-induced variation as possible. The cost is a (sometimes tremendous) loss of power for QTL detection, due to over-fitting the data. [*chapter V*]

QTL mapping: a problem of incomplete data. QTL mapping can be viewed as a problem in which the data are incomplete: the observations of the genotypes at the QTLs are missing. Since marker genotypes are (generally) known, markers can be informative to reveal QTL genotypes. In this thesis we develop a general and flexible EM algorithm to recover information about QTL genotype. When there are many QTLs or when there are marker cofactors with many missing observations, the computations in MQM mapping may become time consuming: it is unfeasible to take into account all possible genotypes (combinations of alleles) for the QTLs and marker cofactors. Disregarding genotypes with negligible probability of occurrence can be a solution to this computational problem. We adopt this approach in the analysis of *Arabidopsis* data (see below). For each plant we disregard all candidate genotypes which are a certain factor less likely than the most likely genotype. The value of the factor may be decreased to 10 without substantial influence on the results.

In practice it occurs frequently that the observations of marker genotypes fail. In addition to fortuitously missing data, another type of missing marker data may occur in a natural way, namely when markers are dominant, or when unequally informative markers are used in experiments with cross breeders. In the first case, the heterozygote cannot be distinguished from one of the homozygotes; in the second case some markers may segregate according to, for instance, backcross rules, so the gametes from only one parent are informative, while other markers may segregate according to F_2 rules, so that the gametes from both parents are informative (HALEY and KNOTT 1994; MALIEPAARD and VAN OOIJEN 1994). This happens in cross-breeding species where, for instance, the parents are homozygous for some markers and heterozygous for others. In the MQM mapping approach missing (QTL and marker) observations are recovered by using all information on phenotype and genotype (i.e. for the putative QTLs as well as for *all* markers) simultaneously.

Recovering genetic information may be difficult if the marker map is sparse: too many candidate genotypes have a probability of occurrence that is relatively small but not negligible. Then, disregarding the relatively unlikely genotypes does not solve the computational problem in MQM mapping. In particular, problems arise when the genetic data are highly incomplete, for instance when many markers are dominant, when many QTLs are assumed, or when unequally informative markers are used. In such situations Monte Carlo solutions rather than analytic solutions can be used for updating parameter estimates in the M-step of the EM algorithm (GUO and THOMPSON 1992). The Monte Carlo EM algorithm is straightforwardly implemented, but computation requires very much computer time. [*chapters II and IV*]

Type of population. In this thesis we mainly concentrate on QTL mapping in self-fertilizing crops. Computer simulations for backcross progenies are given. In addition, practical applications in tomato and *Arabidopsis thaliana* for F_2 and recombinant inbred lines, respectively, are presented. MQM mapping can also be used for other types of segregating progeny in self-fertilizing crops, such as doubled haploids or F_3 lines. Furthermore, MQM mapping can be used in outcrossing species, such as apple and farm animals. In outcrossing species, chromosomes may contain markers of distinct type of segregation (see above). The interval under study may, for instance, be flanked by markers of the backcross-type. However, nearby markers of the F_2-type may provide additional information on genotype and it is therefore important to use the information of multiple markers simultaneously (HALEY and KNOTT 1994; MALIEPAARD and VAN OOIJEN 1994). Furthermore, QTLs and markers may segregate with two or more alleles per locus. MQM mapping easily deals with this multiple allelism.

In MQM mapping we can easily analyse data obtained from different types of progeny simultaneously; this increases the power of QTL detection. In the case of complete linkage maps, phenotypic data from the parents can also be incorporated in the analysis. MQM mapping then exploits the information from parents and their progeny simultaneously. The observed differences between the parents reflect the joint QTL effects. At the same time the environmental variation can be assessed from the parental data. The inclusion of data from individuals with known genotype (such as parental data), can lead to more efficient and more accurate QTL mapping.

[*chapters II-VI*]

Genotype by environment interaction. Many traits exhibit genotype by environment interaction when a set of cultivars (genotypes) is tested in diverse environments. In terms of gene effects, the expression of QTLs may change from one year to another, from one location to another, etcetera. In interval mapping, all information on this type of interaction would be lost if observations are averaged over years and locations; QTL effects may even become more masked by this averaging. Alternatively, the data for the multiple environments can be analysed separately in interval mapping. In the absence of true QTL by environment interaction, however, a QTL may be detected in one environment and not in another, because the chance of simultaneous detection in both environments is small. HAYES et al. (1993) use a model for interaction between a single QTL and environment and they neglect that observations are correlated due to the genetic identity of other QTLs. In MQM mapping with a complete linkage map, however, the major part of this correlation is removed by markers which are used as cofactors in the model. The models in MQM mapping can accommodate QTL by environment interactions, so that clear information about these interactions can be obtained. When the experimental setup involves other factors, such as blocks, these can also be accommodated straightforwardly. In this thesis a practical experiment is presented in which recombinant inbred lines of *Arabidopsis thaliana* are tested under diverse light conditions and with or without vernalization (see below). [*chapters II and VI*]

QTL by QTL interaction. The expression of a given QTL may depend on the expression of one or more other QTLs (epistasis). The models developed in this thesis can accommodate such interactions between QTLs. Though epistatic effects can in principle be modelled straightforwardly, this will cause a rapid increase in the number of parameters in the model relative to the amount of data. Therefore, the detection of epistatic effects probably requires a different type of experimental approach, such as raising plants of deliberately chosen multilocus marker genotypes. [*chapter II*]

Disease scores and such. In research on resistance against diseases one often uses disease scores, such as 0=no symptoms, 1=moderate infection, 2=severe infection, and 3=dead. Obviously, there is no question of normally distributed and continuous variation. In MQM mapping other types of distribution for the trait can be assumed in addition to the commonly assumed normal distribution. This is not only important for disease scores measured on an ordinal scale, but also for percentages, counts, life times, etcetera. In this thesis it is demonstrated that mixture models can be embedded in generalized linear models. Hereby, many types of distribution can be used in QTL mapping. We present an example in which the exponential distribution is used. HACKETT and WELLER (in press) deal with traits measured on an ordinal scale. Also, variance component models, which are often used in animal breeding research, can be used in the MQM mapping framework. [*chapter II*]

Computer software. In MQM mapping we use a new and simple iterative EM algorithm to estimate the parameters of the genetic model. This algorithm can be readily implemented using standard statistical software packages (such as Genstat; JANSEN 1994). There are several advantages: the computer programme is general, flexible, short, easy to read (for Genstat writers) and reliable. The disadvantages are that computation may take much time and that the software is not (yet) generally accessible to non-statisticians. [*chapters II and IV*]

Application in two practical experiments. We detect QTLs for plant height in an F_2 progeny of tomato. We have evidence for two QTLs in interval mapping and evidence for six QTLs (four additional QTLs) in MQM mapping. In addition to F_2 and marker data plant heights of parental and F_1 plants are used in the MQM mapping analysis. In a second experiment we detect QTLs for flowering time in recombinant inbred lines of *Arabidopsis thaliana*. These lines were tested under diverse light conditions and with or without vernalization. In MQM mapping twelve QTLs are detected, four of which display QTL by environment interaction. Unlike this, in interval mapping only four QTLs are indicated with much less precision for map location, and for these QTLs there is no conclusive information on QTL by environment interaction. These examples clearly illustrate the superiority of MQM mapping over interval mapping. [*chapters IV and VI*]

Future work. Currently, QTL mapping in plants is a very active area of theoretical research. Many important issues are being investigated, such as thresholds for tests for QTL detection (REBAÏ, GOFFINET and MANGIN 1994), construction of confidence intervals for QTL location (MANGIN, GOFFINET and REBAÏ; in press) and analysis of outbreeding progenies (HALEY and KNOTT 1994; MALIEPAARD and VAN OOIJEN 1994). However, these issues are studied within the frame of interval mapping and they should now be re-investigated for the more complex situation of mapping multiple QTLs. Other challenges are: the selection of markers to be fitted in the models as cofactors, the detection of epistasis, the use of very dense marker maps, the fine mapping of QTLs, etcetera. These issues may bear upon statistical problems of multiple linear regression with many correlated regressors (and statistical solutions may or may not yet be available). The mapping of QTLs with sparse marker maps and very incomplete genetic data also needs further consideration. The use of the Monte Carlo EM algorithm as a computational tool needs further exploration. A last but certainly not least area of work is the development of user-friendly software, a prerequisite for making QTL mapping a successful tool in plant breeding.

Impact on plant breeding. Molecular markers have become available around 1980 and expectations of their usefulness in plant breeding ran high. Molecular markers are now basic tools in scientific research. Powerful biometrical methods, such as developed in this thesis, make possible the detection and genetic mapping of multiple QTLs affecting complex traits. This leads to improved understanding and more efficient manipulation of many important processes in plants. Mapped QTLs can be traced in breeding programmes, for instance, indirectly via linked markers. New strategies aiming at accumulation of favourable alleles of QTLs are now within reach. Fine-mapped QTLs can also be cloned and transferred via molecular and cell-biological techniques. Such molecular marker techniques are now breaking through in applied plant breeding. Success stories stimulate breeding companies to change their strategy from classical breeding to marker assisted breeding. We believe that solutions to important public matters such as food production with less requirements for chemical pesticides can be suitably realised with the aid of the new techniques.

Detection of major genes and other genetic applications. In this thesis we develop methods which can be applied to a much larger range of quantitative genetic problems than to "just" the current problem of mapping QTLs. In all these quantitative genetic problems the phenotype is observed in a segregating population, whereas the genotype is completely or partially masked. For instance, the detection of genes with major effects on a quantitative trait (major genes) without the use of genetic markers is closely related to the mapping of QTLs (JANSEN 1994; see COLÓN, JANSEN and BUDDING 1995 for an example). Here, we present a rather different application: the analysis of flow cytometric measurements of the DNA content in a segregating population of spermatids from a mutant mouse, which shows non-disjunction for specific chromosomes during meiosis. Another practical example is described by JANSEN and DEN NIJS (1993): the estimation of the proportion of unreduced pollen grains in perennial ryegrass via the size of pollen grains. [*chapter VII*]

Mixture models in other areas of research. Our method of parameter estimation makes it relatively easy to handle complex mixture models. A large number of statistical tools has become available, because we embedded mixture models within the frame of generalized linear models. The method can be readily implemented in statistical packages. This offers new and important possibilities for research areas where (complex) mixture models are appropriate, such as chemistry, pharmacology, medicine, psychology and technology. [*chapter VII*]

LITERATURE CITED

COLON, L.T., R.C. JANSEN and D.J.BUDDING, 1995 Partial resistance to late blight (*Phytophthora infestans*) in hybrid progenies of four South American *Solanum* species crossed with diploid *S. Tuberosum*. Theor Appl Genet, in press

COWEN, N.M., 1989 Multiple linear regression analysis of RFLP data sets used in mapping QTLs, in *Development and application of molecular markers to problems in plant genetics*, edited by T. HELENTJARIS and B. BURR. Cold Spring Harbor Laboratory, New York, pp. 113-116

GUO, S.W. and E.A. THOMPSON, 1992 A Monte Carlo Method for combined segregation and linkage analysis. Am J Hum Genet 51:1111-1126

HACKETT, C.A., 1994 Selection of markers linked to quantitaitve trait loci by regression techniques, in *Biometrics in plant breeding: applications of molecular markers*, edited by J.W. VAN OOIJEN and J.JANSEN. CPRO-DLO, The Netherlands

HACKETT, C.A. and J.I. WELLER, 1995 Genetic mapping of quantitative trait loci for traits with ordinal distributions. Biometrics, in press

HALEY, C.S. and S.A. KNOTT, 1992 A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity:315-324

HALEY, C.S. and S.A. KNOTT, 1994 Mapping quantitative trait loci between outbred lines using least squares. Genetics 136:1195-1207

HAYES, P.M., B.H. LIU, S.J. KNAPP, F. CHENN, B JONES et al., 1993 Quantitative trait locus effects and environmental interaction in a sample of North American barley germ plasm. Theor appl Genet 87:392-401

JANSEN, R.C., and A.P.M. DEN NIJS, 1993 A statistical mixture model for estimating the proportion of unreduced pollen grains in perennial ryegrass (*Lolium perenne* L.) via the size of pollen grains. Euphytica 70:205-215

JANSEN, R.C., 1994 Maximum likelihood in a finite mixture model by exploiting the GLM facilities of Genstat. Genstat Newsletter 30:25-27

MANGIN, B., B. GOFFINET and A. REBAÏ, 1994 Constructing confidence intervals for QTL location. Genetics, in press

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. and D. BOTSTEIN, 1989 Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185-199

MALIEPAARD and J.W. VAN OOIJEN, 1994 QTL mapping in a full-sib family of an outcrossing species, in *Biometrics in plant breeding: applications of molecular markers*, edited by J.W. VAN OOIJEN and J. JANSEN. CPRO-DLO, The Netherlands

MARTINEZ, O. and 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

REBAÏ, A., B. GOFFINET and B. MANGIN, 1994 Approximate thresholds of interval mapping tests for QTL detection. Genetics 138:235-240

STAM, P., 1991. Some aspects of QTL analysis, in *Proceedings of the Eighth Meeting of the Eucarpia Section Biometrics in Plant Breeding*, BRNO

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

GENETISCHE KARTERING VAN GENEN VOOR KWANTITATIEVE EIGENSCHAPPEN (SAMENVATTING)

Veel voor cultuurgewassen belangrijke eigenschappen, zoals opbrengst, kwaliteit en ziekte-resistentie, vertonen een continue variatie. Methoden om deze variatie te analyseren en vooral om de mogelijke genetische basis ervan te ontrafelen zijn derhalve van het grootste belang voor veredelingsdoeleinden; dit is het werkterrein van de kwantitatieve genetica. Reeds aan het begin van deze eeuw is aangetoond dat continue variatie het gecombineerde effect is van omgevingsfactoren en segregatie van verschillende, helaas niet direct traceerbare genen. Deze genen ("quantitative trait loci" of "QTLs") kunnen alleen goed opgespoord worden als men de beschikking heeft over genetisch gemarkeerde chromosomen. Voor zo'n gemarkeerde positie (een "merker") kan het genotype bepaald worden, bijvoorbeeld met moleculaire technieken. De gemiddelde waarde van een kwantitatieve eigenschap kan dan berekend worden voor elk van de genotypen die mogelijk zijn voor zo'n merker. Deze gemiddelden zullen verschillen als een merker in de buurt van een QTL ligt (tenzij er sprake is van zogenaamd tussen-locus-evenwicht). In de loop van de jaren tachtig zijn de eerste typen moleculaire merkers ontwikkeld en nu al zijn voor diverse plante- en diersoorten genetische kaarten gemaakt met grote aantallen van dergelijke merkers, gelijkmatig verspreid over de chromosomen. Daarmee is een nieuw tijdperk voor de kwantitatieve genetica aangebroken.

Bij kartering van QTLs bestaan de gegevens uit metingen aan een eigenschap (het fenotype) en waarnemingen aan merkers (het genotype, met moleculair-biologische middelen vastgesteld). Er is behoefte aan efficiënte en nauwkeurige biometrische methoden om dergelijke gegevens te analyseren. Daarmee kunnen dan QTLs voor (complexe) eigenschappen worden gelocaliseerd. De overerving van de QTLs kan vervolgens worden gevolgd in veredelingsprogramma's, bijvoorbeeld via indirecte selectie op merkers. Of men kan deze QTLs kloneren en ze daarna met moleculair- of celbiologische technieken weer overbrengen naar planten. De traditionele karteringsmethoden zijn echter nog verre van optimaal en het onderzoek naar betere methoden is het werkveld van een toenemend aantal biometrici en kwantitatief genetici. Het onderhavige proefschrift vormt het verslag van een dergelijk onderzoek.

In hoofdstuk I wordt de kartering van QTLs in een historisch perspectief geplaatst en wordt een overzicht van de in de loop der tijd gebruikte biometrische modellen gegeven.

In hoofdstuk II wordt een nieuw, algemeen en flexibel biometrisch raamwerk ontwikkeld voor de kartering van QTLs. Een eenvoudig algoritme voor het schatten van de parameters van de modellen wordt beschreven. Deze methode kan in veel verschillende omstandigheden toegepast worden (een of meer QTLs, diverse typen kruisingspopulaties, eigenschappen en proefopzetten, etcetera). In twee met de computer gesimuleerde voorbeelden worden enkele van de nieuwe mogelijkheden geïllustreerd. In dit hoofdstuk wordt ook het probleem van de gelijktijdige kartering van meervoudige QTLs aangepakt. Exacte modellen voor meervoudige QTLs kunnen nu gebruikt worden, tenminste in principe; maar er is veel rekenwerk als er veel QTLs zijn. Een benaderende methode voor het karteren van meervoudige QTLs wordt voorgesteld. In deze aanpak zijn de modellen exact voor een QTL op een veronderstelde kaartpositie. Ze zijn echter benaderend voor andere mogelijke QTLs; dit komt doordat in de analyse deze QTLs vervangen worden door nabijgelegen merkers (d.w.z. merkers worden als "cofactoren" in de analyse gebruikt).

In hoofdstuk III wordt de methode voor het opsporen en karteren van meervoudige QTLs verder uitgewerkt. Enkele simulatie-studies illustreren de potentiële kracht van merker cofactoren hierbij.

In hoofdstuk IV wordt het probleem van ontbrekende waarnemingen voor merkers opgelost. Er ontbreken altijd waarnemingen en dit bemoeilijkt het gebruik van merkers als cofactoren in praktische experimenten. Een zeer algemene methode wordt beschreven, die ons in staat stelt toch zo goed mogelijk alle ontbrekende genetische (QTL en merker) waarnemingen boven tafel te krijgen. Als eerste praktisch voorbeeld worden verschillende QTLs voor plant-hoogte gekarteerd in een F_2-kruisingspopulatie bij de tomaat. Tevens wordt gedemonstreerd hoe gegevens van ouders en F_1-populatie gebruikt kunnen worden bij de kartering van QTLs.

In hoofdstuk V worden de kans op een fout van de eerste soort (d.w.z. een QTL is gekarteerd op een plaats waar helemaal geen QTL ligt) en de kans op een fout van de tweede soort (d.w.z. een QTL wordt niet opgespoord) bestudeerd door middel van simulatie met de computer. Problemen met betrekking tot de selectie van "belangrijke" merker cofactoren alsmede met betrekking tot het schatten van parameters in modellen met veel merker cofactoren worden opgelost. De karteringsmethode wordt verder verfijnd zodat het nu mogelijk is om de volledige kracht van complete koppelingskaarten van merkers optimaal te benutten. Deze methode heeft de naam "MQM mapping" gekregen. MQM staat voor "meervoudige-QTL modellen" alsmede voor "merker-QTL-merker". Het eerste geeft aan dat de modellen niet alleen rekening houden met de individuele effecten van QTLs op de eigenschap, maar ook met hun gezamenlijke effect; het tweede geeft aan dat QTLs geplaatst worden tussen merkers op de al bestaande merkerkaart.

In hoofdstuk VI worden verschillende QTLs en interacties tussen QTLs en milieu opgespoord voor het bloeitijdstip in ingeteelde lijnen van het modelgewas *Arabidopsis* *thaliana*.

Tenslotte wordt in hoofdstuk VII getoond dat het biometrische raamwerk ook zijn toepassing kent bij complexe mengselmodellen op andere gebieden van onderzoek. Een praktisch voorbeeld betreffende non-disjunctie bij de muis is uitgewerkt.

CURRICULUM VITAE

The author was born on 21 August 1963 in Utrecht. In 1981 he finished secondary education at the municipal grammar-school (Stedelijk Gymnasium) in Utrecht and began his studies of mathematics at the University of Groningen. In 1987 he graduated cum laude with *mathematical statistics* as main subject and *advanced calculus* and *numerical methods and computer programming* as minor subjects. In 1988 the author became affiliated with the Institute of Horticultural Plant Breeding (IVT), which is now merged in the Centre for Plant Breeding and Reproduction research (CPRO-DLO). His current position is scientist for statistics in the Department of Population Biology. The author served as a member and chairman of the examining-board for Statistical Analyst A-VVS of the Netherlands Society for Statistics and Operation Research. He now serves as a committee member of the Biometrics Society region The Netherlands. The author will be invited speaker on the Fifth Gordon Conference on Quantitative Genetics and Biotechnology (USA 1995) and on the National Meeting of the Biometric Society region Italy (Italy 1995).