High Resolution of Quantitative Traits Into Multiple Loci via Interval Mapping

Ritsert C. Jansen* and Piet Stam*,†

**Centre for Plant Breeding and Reproduction Research (CPRO-DLO), Department of Population Biology, P.O. Box 16, 6700 AA Wageningen, The Netherlands, and †Department of Genetics, Wageningen Agricultural University, Dreÿenlaan 2, 6703 AH Wageningen, The Netherlands*

Manuscript received July 2, 1993
Accepted for publication December 7, 1993

ABSTRACT

A very general method is described for multiple linear regression of a quantitative phenotype on genotype [putative quantitative trait loci (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 expectation maximization (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.

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 study 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 report 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

$$Y = m + \sum_i x_{ai} a_i + \sum_i x_{di} d_i + E \qquad (1)$$

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. 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% 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 expectation maximization (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 $f(y) = P(A) \cdot f(y|A) + P(B) \cdot f(y|B) + P(H) \cdot f(y|H)$, $P(A) = P(B) = \frac{1}{4}$, $P(H) = \frac{1}{2}$ 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

$$\hat{\beta} = (X^T W X)^{-1} X^T W Y,$$

$$\hat{\sigma}^2 = (1/N)(Y - X\hat{\beta})^T W (Y - X\hat{\beta}),$$

where Y is the complete data vector, X is the design matrix for the complete data, W is the diagonal matrix of weights, β 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 neighbor 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 $P(AA \mid A–)$ that the missing observation has constitution A equals $P(AA \mid A–) = (1 - r)^2$, where r is the recombination frequency between the two loci. The other two conditional probabilities are $P(AB \mid A–) = r^2$ and $P(AH \mid A–) = 2r(1 - r)$. 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

$$P(g \mid y, h) = \frac{P(g \mid h) \cdot f(y \mid g)}{f(y \mid h)} \qquad (2)$$

where $P(g \mid h)$ is the conditional probability for the complete genotype g given the incomplete genotype h, $f(y \mid g)$ is the probability density function of the trait y given the complete genotype g, and $f(y \mid h) = \sum_g P(g \mid h) \cdot f(y \mid g)$ 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

$$\Delta^2 = \sum_g P(g \mid y, h) \cdot (y - m_g)^2, \qquad (3)$$

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 setup 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, γ, 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 ($\mathscr{L}$) minus a penalty for the number of free parameters (k) in the model. Equivalently, Akaike's information criterion, AIC = $-2(\mathscr{L} - k)$ may be minimized. The number of parameters should not be too large, preferably less than $2\sqrt{\text{number of observations}}$ (SAKAMOTO *et al.* 1986).

In many experiments designed to detect associations between marker genotypes and quantitative characters, the number of segregating molecular markers maybe 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 $2\sqrt{\text{number of observations}}$. 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.

TABLE 1

Outline of the models fitted

	Selected markers used on no/other/all chromosomes		
QTL fitted	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.

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 2 ln(10) $\approx$ 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 *vs.* model D, or model A_2 *vs.* 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 (d.f.) (1 d.f. 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 d.f. It takes 4 d.f. to test for the simultaneous effect of two markers in multiple regression on markers; it takes 5 d.f. to test model A_1 *vs.*

TABLE 2

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

Population	No. 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.

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 behavior 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-squared threshold at a significance level of 0.001 per test equals 13.8 for 2 d.f.; it is 16.3, 18.5, 20.5, 22.5, 24.3 and 26.1 for 3, 4, 5, 6, 7 and 8 d.f., 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 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 $41.93 - 27.96 - 2k = 5.97 > 2$, 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

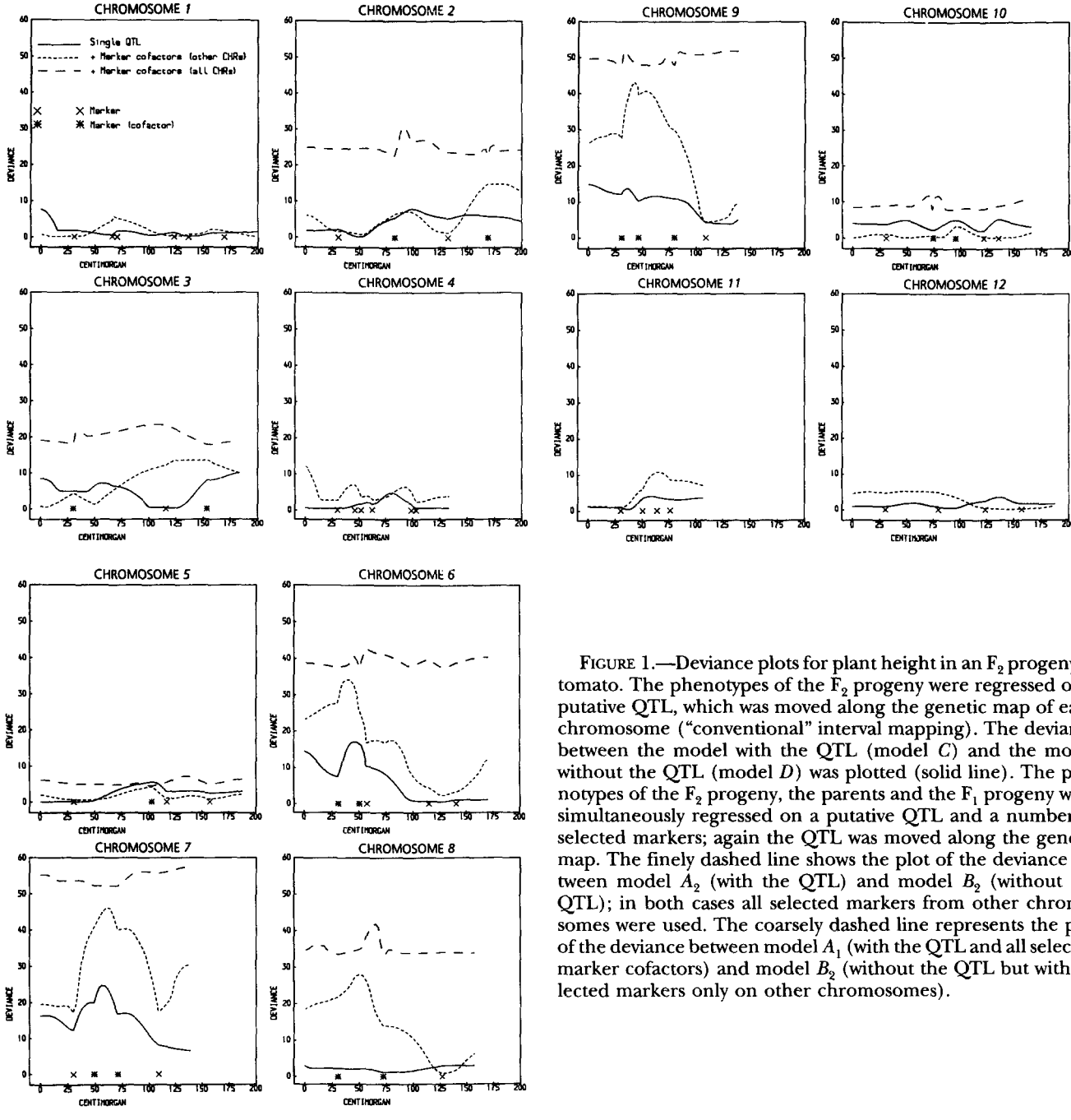

FIGURE 1.—Deviance plots for plant height in an F_2 progeny of tomato. The phenotypes of the F_2 progeny were regressed on a putative QTL, which was moved along the genetic map of each chromosome ("conventional" interval mapping). The deviance between the model with the QTL (model C) and the model without 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).

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 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

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	0.255	0.026	0.0197	0.130
(2-3)	(0.050)	(0.065)	(0.0053)	(0.041)
2	−0.247	−0.071	0.0208	0.091
(4-)	(0.043)	(0.057)	(0.0050)	(0.040)
6	−0.204	0.205	0.0244	0.070
(1-2)	(0.044)	(0.063)	(0.0043)	(0.039)
7	−0.248	−0.114	0.0236	0.111
(2-3)	(0.047)	(0.067)	(0.0043)	(0.039)
8	0.272	0.118	0.0181	0.165
(1-2)	(0.058)	(0.064)	(0.0028)	(0.038)
9	0.249	−0.087	0.0249	0.111
(1-2)	(0.037)	(0.048)	(0.0042)	(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.

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.

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).

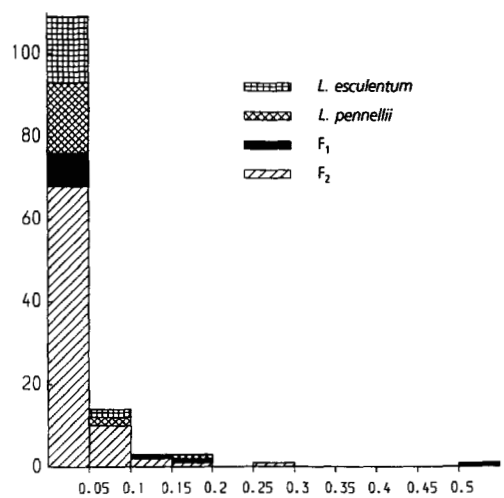

FIGURE 2.—Histogram of the weighted sum of squared residuals, used for the detection of outliers for plant height in an F_2 progeny of tomato. The residuals were obtained from the multiple linear regression of the phenotypes of the F_2 progeny, the parents and the F_1 progeny on all selected markers.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

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 (*i.e.*, variances, covariances) of classic quantitative genetics. Second, 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 study 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 unraveled 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 modeled 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.

The authors are greatly indebted to P. ODINOT and W. H. LINDHOUT, Department of Vegetables and Fruit Crops of CPRO-DLO, for supplying the data of the example.

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. 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.

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.

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. **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, 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 *et al.*, 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 *et al.*, 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. Plant J. **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. Theor. Appl. Genet. **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.

Communicating editor: W. G. HILL

APPENDIX

Updating the estimates of the recombination frequencies in the EM algorithm runs parallel 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:

E step: Update the unknown number of repulsion heterozygotes.

M step: Obtain the new estimate by counting recombinant gametes.

This leads to the following update

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

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