Theor Appl Genet (1992) 85:252–260

THEORETICAL AND APPLIED GENETICS

A general mixture model for mapping quantitative trait loci by using molecular markers

R. C. Jansen

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

Received November 20, 1991; Accepted April 23, 1992
Communicated by J. W. Snape

Summary. 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 discribed 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.

Key words: EM-algorithm – Generalised linear model – Genetic linkage map – Mixture of distributions – Molecular marker – Quantitative trait locus

Introduction

The advent of complete linkage maps of molecular markers has recently stimulated interest in studying the genetics underlying quantitative traits (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 the 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 the detection and mapping of a single QTL in which linkage between a putative QTL and two markers is estimated. Lander and Botstein (1989) developed a software package (MAPMAKER-QTL) for the backcross (BC) populations and F_2 populations, while 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) all mentioned the need for accurate and efficient methods that 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 programmes 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 not mixture models.

In the present paper a mixture model is developed to overcome some of the shortcomings of the methods mentioned above. Extensions of mixture models and parame-

ter 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 of the problems concerning the robustness of the traditional approaches for deviations from normality. The second example illustrates modelling for multiple QTLs and some of the problems concerning the 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 that 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 the regression of phenotype on genotype. Genetic linkage models and models for the 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 backcrossing 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 the 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 2×2 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,

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 $\log(\frac{1}{2}p_{00})=1\cdot\lambda+0\cdot\nu+0\cdot\zeta$.

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 expected frequencies $\frac{1}{2}p_{hi}\cdot\frac{1}{2}p_{jk}$ $(h,i,j,k=0,1)$. 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

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	
	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_1QM_2/m_1qm_2 to the parent m_1qm_2/m_1qm_2. *Example*: For M_1qm_2/m_1qm_2 individuals the coefficients of the genetic linkage parameters are $1\cdot\lambda$, $1\cdot\nu$ and $0\cdot\zeta$, since $\log[\frac{1}{2}r_1(1-r_2)]=\log[\frac{1}{2}(1-r_1)(1-r_2)]+\log[r_1/(1-r_1)]=1\cdot\lambda+1\cdot\nu+0\cdot\zeta$; 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 $G=m-a$

Observed incomplete allelic constitution	Unobserved complete allelic constitution	Genetic linkage			Regression of phenotype on genotype		
		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

[a] λ, ν and ζ denote the parameters for the linear genetic linkage model: $\lambda=\log(\frac{1}{2}(1-r_1)(1-r_2))$; $\nu=\log(r_1)-\log(1-r_1)$; $\zeta=\log(r_2)-\log(1-r_2)$, 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

has coefficients $2\cdot\lambda$, $1\cdot\nu$ and $1\cdot\zeta$, since $\log(\frac{1}{2}p_{00}\cdot\frac{1}{2}p_{11})=\log(\frac{1}{2}p_{00})+\log(\frac{1}{2}p_{11})=1\cdot\lambda+(1\cdot\lambda+1\cdot\nu+1\cdot\zeta)=2\cdot\lambda+1\cdot\nu+1\cdot\zeta$. 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 $2\cdot\frac{1}{2}p_{00}\frac{1}{2}p_{11}$. Therefore, in the log-linear model an extra offset of $\log(2)$ appears, since $\log(2\cdot\frac{1}{2}p_{00}\cdot\frac{1}{2}p_{11})=2\cdot\lambda+1\cdot\nu+1\cdot\zeta+\log(2)$.

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_{1hj}) 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 generalised linear model for multinomial data) may be obtained easily (McCullagh and Nelder 1989): computations may be carried out by statistical packages having facilities for generalised 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 that 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 programme. 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 noted that models may also be extended to include interference constraints, e.g.

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. *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 $\log[\frac{1}{2}(1-r_1)(1-r_2)\,\frac{1}{2}r_1r_2]=2\log[\frac{1}{2}(1-r_1)(1-r_2)]+\log[r_1/(1-r_1)]+\log[r_2/(1-r_2)]=2\cdot\lambda+1\cdot v+1\cdot\zeta$; 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 $\log[2\frac{1}{2}(1-r_1)(1-r_2)\,\frac{1}{2}r_1r_2]=\log[\frac{1}{2}(1-r_1)(1-r_2)\,\frac{1}{2}r_1r_2]+\log(2)$

Observed incomplete allelic constitution	Unobserved complete allelic constitution	Genetic linkage[a]				Regression of phenotype on genotype[b]		
		λ	v	ζ	Offset	m	a	d
$M_1M_1M_2M_2$	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	1	log(2)	1	0	1
	M_1qM_2/M_1qM_2	2	2	2	0	1	−1	0
$M_1M_1M_2m_2$	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
$M_1M_1m_2m_2$	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
$M_1m_1M_2M_2$	M_1QM_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
$M_1m_1M_2m_2$	[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
$M_1m_1m_2m_2$	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
$m_1m_1M_2M_2$	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
$M_1M_1M_2m_2$	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
$m_1m_1m_2m_2$	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

[a] λ, v and ζ denote the parameters for the genetic linkage model: $\lambda=\log(\frac{1}{2}(1-r_1)(1-r_2))$; $v=\log(r_1)-\log(1-r_1)$; $\zeta=\log(r_2)-\log(1-r_2)$, 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

$t=r_1+r_2-2Cr_1r_2$ with $C\neq1$. 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 (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 and (4) imple-

ment 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 was 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 eight 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. However, interactions between genotype and environment, such as year × genotype or location × genotype interactions, are also 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 $G=m+A+D+X'\beta$, where $X'\beta$ relates the genotypic value to the explanatory variables (β is a vector of regression parameters and X is a vector of the 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. Generalised 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. Generalised 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 (1) 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), (2) cope with missing QTL data and missing marker data and (3) 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 that 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 generally 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 that 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 Dempster 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/m_1qm_2 and M_1qM_2/m_1qm_2 are stored into a design matrix or into explanatory variables to be analysed (Tables 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. Consequently, 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, which gives initial weights equal to 0 or 1. Alternatively, the algorithm can be started by setting the parameters to (well-chosen) initial values. The analyses can be carried out by statistical packages that have facilities for generalised linear models.

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

Formal justification

Continuous phenotypic data, such as that observed when the trait is 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 $\mathscr{L}$ of observations $(y_1,h_1), (y_2,h_2), \ldots, (y_I,h_I)$ is

$$\mathscr{L}((y_1,h_1),(y_2,h_2)\ldots(y_I,h_I)) = \prod_{i=1}^{I} f(y_i,h_i) = \prod_{i=1}^{I} p(h_i) \prod_{i=1}^{I} f(y_i|h_i).$$

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

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

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

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

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

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

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

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 the 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 the 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 $f(y) = \frac{1}{\mu}\exp(-y/\mu)$, 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. *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. The parameter values used to simulate the data were $r_1 = r_2 = 0.09$, $\mu_{qq} = 10$, $\mu_{Qq} = 15$ and $n = 200$[a]. 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[b]		Regression of phenotype on genotype[c]	
			$\hat{r}_1$	$\hat{r}_2$	$\hat{\mu}_{qq}$	$\hat{\mu}_{Qq}$
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

[a] n denotes the number of individuals in the BC progeny
[b] r_1 and r_2 denote the recombination frequencies between the QTL and its flanking markers
[c] μ_{Qq} and μ_{qq} denote the mean value of individuals with Qq and qq at the QTL, 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 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 (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 $M_1Q_1M_2q_2M_3/m_1q_2m_2Q_2m_3$. Note that the genes at the QTLs are in the 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

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. The effects of the QTLs were additive ($G=m+A_1+A_2$). The parameter 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 $n=200$[a]

Coefficients of the genetic linkage parameters and the parameters for regression of phenotype on genotype for an individual with observed marker genotype $M_1m_1\,M_2m_2\,M_3m_3$

Unobserved complete allelic constitution at the QTL	Genetic linkage[b]						Regression of phenotype on genotype[c]		
	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	1	1	0	0
Q_1q_2/q_1q_2	1	0	0	1	1	0	1	0	−1
q_1Q_2/q_1q_2	1	1	1	1	0	1	1	−1	0
q_1q_2/q_1q_2	1	1	1	1	1	0	1	−1	−1

Log-likelihood and parameter estimates for various models

QTL fitted (yes/no)		log-likelihood	Genetic linkage[b]				Regression of phenotype on genotype		
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 parameter estimates for various models with markers as covariables

QTL fitted (yes/no)		Marker fitted (yes/no)			log-likelihood	Genetic linkage[b]				Regression of phenotype on genotype		
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

[a] n denotes the number of individuals in the BC progeny

[b] r_{11} and r_{12} denote the recombination frequencies between the first QTL and its flanking markers; r_{21} and r_{22} denote the recombination frequencies between the second QTL and its flanking markers; λ_1, ν_1 and ζ_1 denote the genetic linkage parameters for the first QTL: $\lambda_1=\log(\frac{1}{2}(1-r_{11})(1-r_{12}))$; $\nu_1=\log(r_{11})-\log(1-r_{11})$; $\zeta_1=\log(r_{12})-\log(1-r_{12})$; λ_2, ν_2 and ζ_2 denote the genetic linkage parameters for the second QTL: $\lambda_2=\log(\frac{1}{2}(1-r_{21})(1-r_{22}))$; $\nu_2=\log(r_{21})-\log(1-r_{21})$; $\zeta_2=\log(r_{22})-\log(1-r_{22})$

[c] 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

($\sigma^2=1$). 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 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 also shows log-likelihoods and parameter estimates for various models. In this example the values of the test statistic $2(\log\mathscr{L}_1-\log\mathscr{L}_0)$ 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 unknown.

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 generalised 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 generalised linear models, such as GENSTAT (Genstat 5 Comittee 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 should be obvious that these are easily dealt with.

Testing for the number of components in a mixture is an important and difficult problem that has not been resolved completely (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).

Acknowledgements. Thanks are due to G. J. O. Jansen, J. Jansen, P. Stam, R. E. Voorrips and W. E. van de Weg for their helpful suggestions and critical reading of the manuscript.

References

Bailey NTJ (1961) Introduction to the mathematical theory of genetic linkage. Oxford University Press, London

Bulmer MG (1985) The mathematical theory of quantitative genetics. Clarendon Press, London

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM-algorithm. J R Stat Soc Ser B 39:1–38

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

Haldane JBS (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 SJ (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

Knapp SJ, Bridges WC, Birkes D (1990) Mapping quantitative trait loci using molecular marker linkage maps. Theor Appl Genet 79:583–592

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

Luo ZW, Kearsey MJ (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, Nelder JA (1989) Generalized linear models. Monographs on statistics and applied probability 37. Chapman and Hall, London

McLachlan GJ, Basford KE (1988) Mixture models: inference and applications to clustering. Marcel Dekker, New York

Paterson AH, Lander ES, Hewitt JD, Peterson S, Lincoln SE, Tanksley SD (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, Beckmann JS (1963) Genetic polymorphism in varietal identification and genetic improvement. Theor Appl Genet 47:179–190

Titterington DM, Smith AFM, Makov UE (1985) Statistical analysis of finite mixture distributions. Wiley, New York

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