Statistica Applicata Vol. 8, n. 1, 1996

BIOMETRICAL MODELLING IN MAPPING QUANTITATIVE TRAIT GENES BY USING GENETIC MARKERS

Ritsert C. Jansen

Centre for Plant Breeding and Reproduction Research (CPRO–DLO), Wageningen, the Netherlands.

SUMMARY

An introduction is made to statistical problems presented by the recent use of new marker technology in the mapping of genes affecting quantitative traits. The problems include estimation and inference in complex genetic models with missing or incomplete genotypic data. The key to the problems is to consider all candidate complete genotypes, assign weights to them and do a weighted linear regression of the trait on the complete genotype using a (Monte Carlo) EM algorithm for parameter estimation.

Keywords: genetic Map, QTL, multiple regression, E.M. algorithm.

1. INTRODUCTION

Many, if not most, traits of interest to plant, animal and human geneticists are controlled by genes of which the inheritance can hardly be assessed (quantitative trait loci or QTLs). Recently, new biotechnological tools have become available by the advent of molecular markers, which heralds a new era for studying the genetics of complex traits. In only a few years time it has had a major impact on fundamental plant and animal genetics and on human medical genetics (Tanksley, 1993; Lander and Schork, 1994). Powerful and accurate biometrical methods are needed, so as to make possible the efficient dissection of complex traits into individual gene effects. Not surprisingly, this area is gaining fast growing attention of biometricians (Jansen, 1994). In this paper we make a survey of some of the challenging problems presented to biometricians. At least a basic knowledge of genetics is essential for understanding of our paper. For readers unfamiliar with this area we first describe the main features of the genetic mechanisms involved.

2. SOME BASICS OF MOLECULAR AND QUANTITATIVE GENETICS

Genes are distributed along several linear chromosomes. Diploid organisms like humans, animals and many plant species have two sets of chromosomes, one set from each parent. The members of a pair of chromosomes are called homologous chromosomes. Gametes (egg and sperm cells) receive only one set of chromosomes, which conserves the number of chromosomes from generation to generation. In sexual reproduction genetic material is recombined in two ways. Firstly, maternally and paternally derived homologous chromosomes physically exchange chromosome parts by symmetrical breakage and crosswise rejoining (crossovers; see Fig. 1). Secondly, gametes randomly receive one chromosome from each pair of chromosomes. The maternal (white) and paternal (black) forms of a gene are termed alleles. The recombination frequency (r) between two genes is not linearly related to the distance between those genes. As distance between genes increases, the incidence of multiple crossovers causes the observed recombination frequency to be an underestimate of the crossover frequency and hence of the true genetic "map distance" m (in Morgan units). It is often assumed that the number of crossovers between two genes follows a Poisson distribution with expectation m. Therefore, the recombination frequency can be calculated as the sum of probabilities of odd crossovers, which is $r = 0.5(1-e^{-2m})$; note that 1 centiMorgan $\approx$1% recombination.

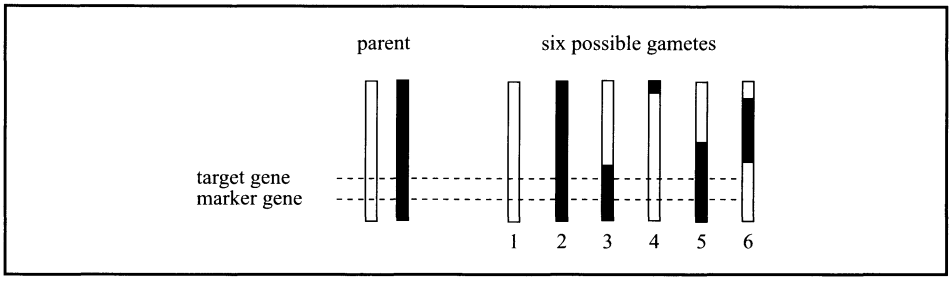


Fig. 1: Example of how the genetic material of two homologous parental chromosomes may be reshuffled by the crossover mechanism when gametes (egg and sperm cells) are formed. Gametes 1 and 2 are non-recombinant; gametes 3–6 show single or double recombination; recombination between the target and marker gene did not occur.

For quantitative traits we know nothing about how many genes are involved, where the genes are located and what effects the genes have. The recent development of molecular markers provides geneticists with new technology that can be used for detecting and mapping genes. A molecular marker may be considered as a gene of which the allelic constitution (=genotype) can be observed with biochemical methods. In Fig. 1 the genotype of the marker gene (white or black per gamete) can be observed, but not that of the target gene. Since the two genes are at nearby map locations, most gametes will not be recombinant for the two genes (proportion

1–*r*): white is associated with white and black is associated with black. Therefore, in a progeny an indirect observation of the genotype at the target gene can be obtained from marker information. This works well only if any target gene is located close to a marker, i.e. ideally the set of markers should cover all chromosomes. Now, genetic marker maps exist for many plant and animal species. For example Fig. 2 shows a marker map of *Arabidopsis*. A subset of these markers was used in the application presented in Box 1. A marker map of lily is not yet available; an application with a small number of widely spaced markers is shown in Box 2.

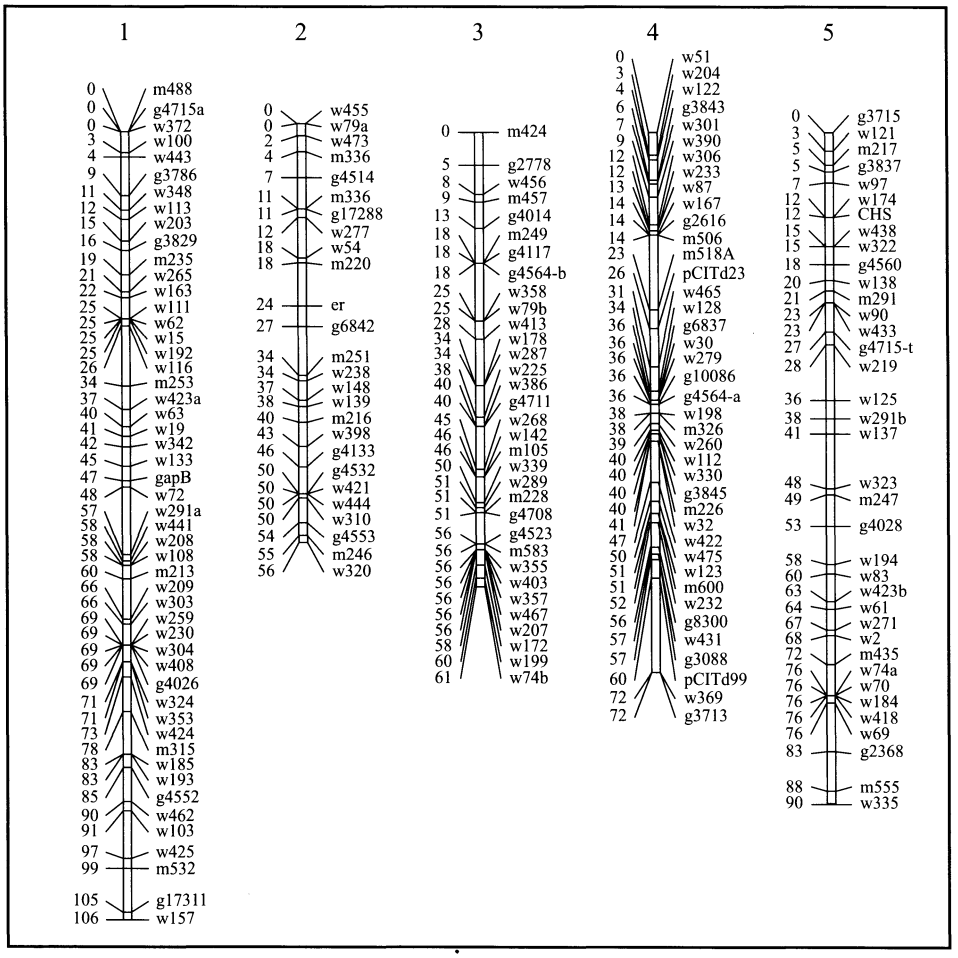


Fig. 2: A marker map of the five chromosomes of Arabidopsis. Map distances (in centiMorgans) are indicated on the left hand side of each chromosome, marker names on the right hand side.

If a trait is encoded by many genes, the distribution of trait values may appear continuous because numerous genotypes exist in the population. If the trait is affected by a few genes, it may still show continuous variation when environmental factors influence the trait. In most situations both genetic and environmental factors are active. Boxes 1 and 2 present histograms of the trait values in experiments concerning germination in *Arabidopsis* and resistance to *Fusarium* in lily, respectively. In the first application many genes affect the trait. In the latter probably only three genes affect the trait, but the presence of multiple candidate alleles at each locus increases the number of genotypes. In both experiments the variation is largely attributable to genetic factors.

3. MAPPING QUANTITATIVE TRAIT GENES

The easiest approach to mapping quantitative trait genes is to consider markers one by one. Differences between the genotypes of a marker with regard to a trait may indicate the presence of one or more linked genes (see Box 2 for an example). Simple linear regression (or a non–parametric method) can be used for testing. This one–by–one approach is simple and has been used in many applications. However, using more markers simultaneously in a multiple regression is more efficient.

Unfortunately, the use of more markers simultaneously is usually hampered by missing marker data. In practice frequently about 5% of the observations on markers fail. In addition to fortuitously missing data, other types of 'missing information' may occur, e.g. when the marker technique provides only partial information about the allelic constitution of markers (see Box 2 for an example). One way to proceed is to eliminate individuals with missing data, but this could mean that only a very limited set of data remains.

Another way to overcome the problem of missing information is obtained by noting that the missing genotype of a marker belongs to a limited set of candidate allelic states, e.g. double white, white plus black or double black. An observation at a flanking marker, e.g. double white, may help to reveal information: it is likely that both markers are double white because white is associated with white and black with black. If the distance between the markers is not very small, we should also take into account the occurrence of recombination, e.g. the observed genotype is double white at one marker whereas the (unobserved) actual genotype at the other marker is white plus black. For this purpose we can assign probabilities to the three candidate complete genotypes (double white at one marker and double white, white plus black or double black at the other marker). In a full (maximum likelihood)

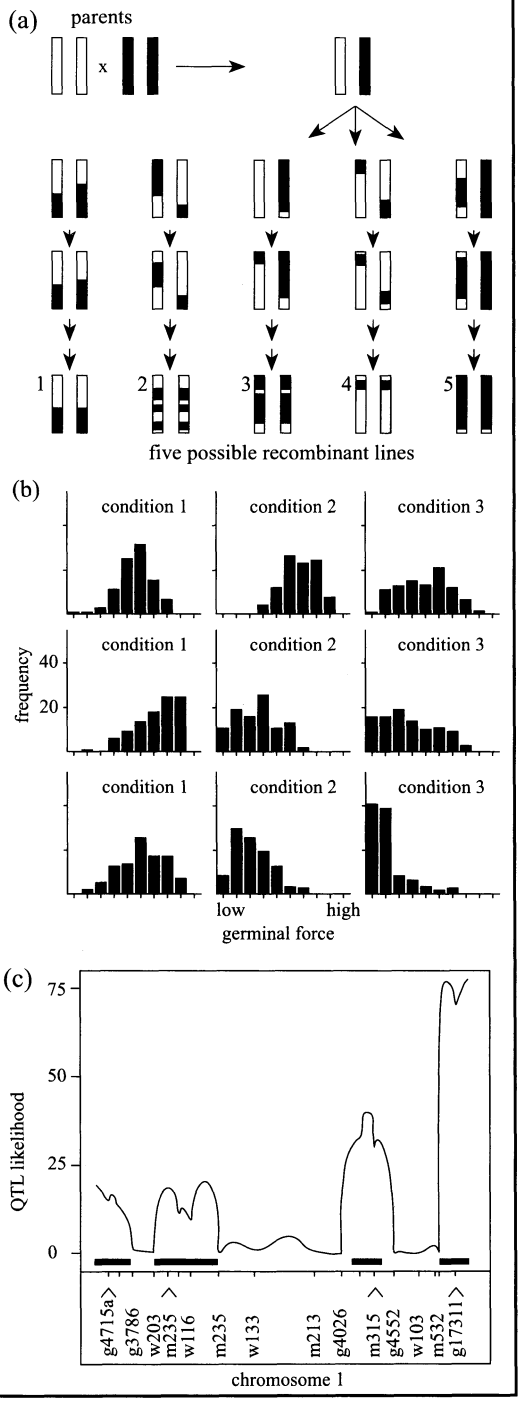
approach the observations on both markers and the trait considered are used for calculating these probabilities. Then probabilities depend on unknown trait parameters, but estimation can be done in an iterative manner (see below). It should be noted that in practice the complete genotype may involve many loci rather than two.

The key to the estimation problem is to consider an artificial complete set of data that per individual consists of all pairs of candidate complete genotype and observed trait value. Calculation of the probabilities associated with candidate complete genotypes given observed marker and trait data constitutes the E(xpectation)–step of an EM algorithm. Bayes' theorem is used to obtain a simple expression for these probabilities. The M(aximization)–step of the EM algorithm consists of a weighted linear regression using the artificial complete set of data. The replicated trait values are regressed on candidate complete genotypes and the weights involved are the probabilities assigned to the candidate complete genotypes (Jansen and Stam, 1994). Usually it is assumed that trait values follow a normal distribution but it is easy to extend the model to accommodate for non–normal distributions. To keep things simple it is often assumed that gene effects are additive.

It should be noted that the genotype may not only involve marker loci but also 'true' quantitative trait loci (QTLs at hypothetical map positions). The genotypes of the QTLs are always unknown but again complete data can be constructed. 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. We often use an approach in which the trait is regressed on selected markers and a single putative QTL. By moving the putative QTL along the chromosome, we can produce a profile of QTL likelihood at any map position (see Box 1 and 2).

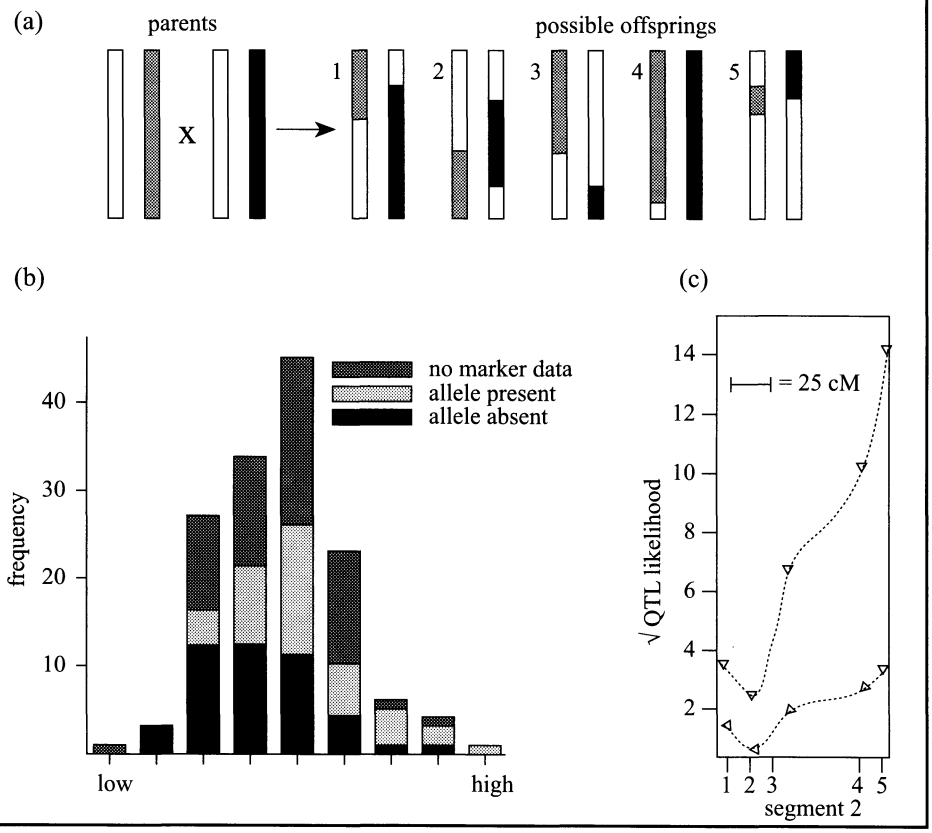
For large progenies with very incomplete marker information exact computations are not feasible due to the extremely large number of candidate complete genotypes. One solution to this problem is to disregard unlikely genotypes in the calculations. This approach works well in our first example (Box 1). However, in our second example (Box 2), the set of candidate genotypes is still too large. Therefore a Monte Carlo solution rather than an analytic solution for updating parameter estimates in the M–step of the EM algorithm is used. In each E–step of the EM algorithm candidate complete genotypes are sampled from the conditional probabilities calculated in the E–step. Again, an artificial data set can be constructed involving multiple copies of the data. Now the number of copies is equal to the number of Monte Carlo runs. Sums of squares and products (SSP) can be

Box 1:
QTL mapping for germination in *Arabidopsis*. (a) Two ecotypes were crossed. Recombinant inbred lines were derived by self–pollinating plants for a number of generations. Each generation genetic material may be reshuffled by recombination in regions where the homologous chromosomes have different alleles (white and black). When inbreeding proceeds, homologous chromosomes tend to become identical and the chance of recombination between a target gene (e.g. double white) and a marker gene (double black) is $R=r/(1+2r)$. (b) Recombinant inbred lines (99) were tested under three different conditions in three replicates (nine environments). (c) The trait was regressed on a set of markers (covering all five chromosomes) and environment including marker by environment interactions. By backward elimination important markers were selected (indicated by ^); markers at chromosome 1 showed no interaction with environment. Next, the trait was regressed on all selected markers and a putative QTL. By moving the QTL along the chromosome, we produced a profile of QTL likelihood; a selected marker was (interactively) dropped from the model if the putative QTL was nearby (<15 cM). Bars indicate 95% support intervals for the QTL detected.



Box 2:

QTL mapping for *Fusarium* resistance in lily. (a) Two cultivars with one common set of chromosomes were crossed. In the offspring three candidate alleles are present at each locus (white, black or grey). Unfortunately, the marker technique used provides only partial information about the actual allelic configuration. Markers at chromosome segment 2 can only assess the presence or absence of the grey alleles. The type of the other allele (white or black) cannot be assessed. The white and black alleles may have different effects on the trait. Two complete allelic configurations are possible for each marker observation. For several markers, the number of candidate genotypes soon becomes extremely large (also because some plants have not yet been genotyped in which case we have four candidates per marker). We used Monte Carlo EM for parameter estimation. (b) The offspring was tested for Fusarium resistance. To illustrate the traditional analysis of markers–one–by–one, we show the distribution of individuals with the grey allele present and the distribution of individuals with this allele absent for marker 5 at segment 2. The difference is significant using Kruskal–Wallis test. Markers in three segments displayed QTL activity. (c) Results of the traditional approach are denoted by Δ. We also implemented the full maximum likelihood approach (denoted by ∇). We fitted three QTLs simultaneously, one in each segment. One QTL was moved along the chromosome, while keeping the two other QTLs at their nearest marker position obtained in the preliminary analysis.



accumulated sequentially for each Monte Carlo run in turn. Regression calculations are based on the final SSP matrix.

Unfortunately, there is no feasible way to generate the Monte Carlo samples, because it is difficult to update the genotype for all loci simultaneously. In our second application (Box 2) we use the Gibbs sampler, a simple iterative approach treating the problem locus by locus. If an individual has incomplete genotypic information at a certain locus, then a complete genotype at that locus is sampled from the conditional distribution given the incomplete genotype at that locus, the current complete genotype at other loci, the trait values and the current parameter estimates. Expressions for the conditional probabilities can be derived in a straightforward manner using Bayes' theorem. These calculations are now much easier because of the small number of candidate genotypes per step. One cycle of this Gibbs approach is terminated if genotypes have been updated once for all loci. It should be noted that subsequent cycles produce dependent Monte Carlo realizations and usually only a subset of the realizations is used.

We finally mention that the use of Monte Carlo techniques in combination with Gibbs sampling opens up ways for tackling complicated gene mapping problems, such as arise when data originate from human pedigrees rather than from controlled experiments in plants or animals (Guo and Thompson, 1992).

4. CONCLUDING REMARKS

Currently, mapping of quantitative trait genes is a very active area of theoretical research. Many important issues are being investigated, such as optimal approaches to selection of markers, thresholds for tests of QTL detection, construction of confidence intervals for QTL location, problems of (over)parameterization and bias, robustness of the mapping approach and development of diagnostics for diverse purposes. In this paper we have chosen to describe a general frame for QTL mapping and to emphasize the relation with standard multiple linear regression models. We believe that many of the genetic problems may bear upon statistical problems of multiple linear regression and solutions may or may not yet be available. Anyhow, challenging problems arise at the interface of statistics and genetics and statisticians can have a key role in solving them.

REFERENCES

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

JANSEN R.C., (1994), Mapping of quantitative trait loci by using genetic markers: an overview of biometrical models used, in: Van Ooijen J.W. and Jansen J. (eds) *Biometrics in plant breeding: applications of molecular markers,* CPRO–DLO, the Netherlands.

JANSEN R.C., STAM P., (1994), High resolution of quantitative traits into multiple loci via interval mapping, *Genetics,* 136, 1447–1455.

LANDER E.S., SCHORK N.J. (1994), Genetic dissection of complex traits, *Science,* 265, 2037–2048.

TANKSLEY S.D., (1993), Mapping Polygenes, *Annu. Rev. Genet.*, 27, 205–233.

MODELLI BIOMETRICI PER IL MAPPAGGIO DI GENI PER CARATTERI QUANTITATIVI TRAMITE MARCATORI MOLECOLARI

RIASSUNTO

Viene proposta una introduzione alle problematiche statistiche connesse all'impiego, introdotto di recente, dei marcatori molecolari per il mappaggio di geni responsabili della variabilità di caratteri quantitativi. Si tratta essenzialmente di stima e di inferenza nell'ambito di modelli genetici complessi che possono introdurre dati genotipici mancanti od incompleti. Il problema può essere affrontato considerando tutti i dati genotipici completi disponibili ed assegnando ad essi dei pesi opportuni. Il carattere di interesse può essere quindi analizzato tramite regressione multipla pesata su tutti i dati genotipici, tramite algoritmo EM (Monte Carlo) per la stima dei parametri.

Parole chiave: mappe genetiche, QTL, regressione multipla, algoritmo E.M.