Statistical Issues in the Search for Genes Affecting Quantitative Traits in Experimental Populations

R.W. Doerge, [1] Z-B. Zeng and B.S. Weir
Program in Statistical Genetics, Department of Statistics
North Carolina State University, Raleigh, NC 27695-8203

BU-1292-M June 1995

Keywords: genetic map, interval mapping, QTL, molecular marker, single point analysis.

[1]present address: Department of Statistics, Purdue University, West Lafayette, IN 47907-1399

Running Head: A Review of Statistical Issues Involved in Searching for QTL

Corresponding author: R.W. Doerge
Department of Statistics
1399 Mathematical Sciences Building
Purdue University
West Lafayette, IN 47907-1399

Telephone number: (317) 494-6030
Fax number: (317) 494-0558
E-mail: doerge@stat.purdue.edu

Key words: interval mapping, interval testing, multiple markers,
mixture distribution, QTL
single markers

Abstract

The goal of this article is to review key contributions in the area of statistics as applied to the use of molecular marker technology and quantitative genetics in the search for genes affecting quantitative traits responsible for specific (human) diseases, and often times economically important agronomic traits. Since an exhaustive literature review is not possible, the limited scope of this work is to encourage further statistical work in this vast field by first reviewing human and domestic species literature, and then concentrating on the statistical developments for experimental breeding populations. Most traits pertaining to production in domestic species are quantitative, and substantial gains, due to the availability of genetic markers, have been made over the years by both plant and animal breeders toward a long–term goal of locating genes affecting quantitative traits (quantitative trait loci, QTL) for the eventual characterization and manipulation of these genes in order to develop improved agronomically important crops. Our main concern is that the care and expense that is required in generating both genetic marker data and quantitative trait data is accompanied by equal care in the statistical analysis of the data, thus continuing the long tradition of statistics in quantitative genetics. Through an example using an F_2 male genetic map of mouse chromosome 10, and quantitative trait values measured on weight gain, we implement much of the reviewed methodology for the purpose of detecting/locating a QTL having its effect on weight gain.

1 Introduction

One of the early benefits of the human genome project has been the establishment of genetic maps for human and many domestic species. For example, in crop plants, maps have been established for barley (Graner *et al.* 1991), brassica (Slocum *et al.* 1990), corn (Coe *et al.* 1990), soybean (Keim *et al.* 1990), and tomato (Bernatzky and Tanksley 1986). For animals, maps have been developed for the cow (Barendse *et al.* 1994) and the mouse (Copeland *et al.* 1993). An account of the human map in late 1992 was given by the NIH/CEPH Collaborative Mapping Group (1992). A compendium of genetic maps for many species is provided by O'Brien (1993). These maps, consisting of identifiable features or *markers* on the genome at known locations, can be used in the search for genes affecting traits of interest. Notable successes have been in human diseases; cystic fibrosis (Kerem *et al.* 1989), Huntington's disease (Huntington's Collaborative Group 1993), and familial dysautomia (Blumfield *et al.* 1993). Although methodologies are still being developed, the accomplishments represented by these successes are substantial. They were also the easiest in the sense that the traits being studied were discrete. By and large, there was little ambiguity over which individuals had the disease. Discrete traits are also being mapped in domestic animals (Georges *et al.* 1993).

In this discussion, we consider the much more difficult task of searching for genes affecting quantitative, or continuous, traits. Many of the issues we cover were treated by Doerge (1993). It is often the case that these traits are controlled by more than one gene, as well as by non-genetic causes, which further complicates the searches. Most traits pertaining to production in domestic species are quantitative, and substantial gains have been made over the years by plant and animal breeders. The immediate hope is that the possibility of identifying specific portions of the genome will enhance breeding programs. The long–term hope is that finding the location of genes affecting quantitative traits (the so–called QTLs, or quantitative trait loci) will lead to characterization and possible manipulation of these genes. It will not even be necessary to perform the initial localization in the species of concern. The possibility of using genes mapped in animals to aid in the study of human disease was illustrated by the location of genes for elevated blood pressure in rats (Hilbert *et al.* 1991, Jacob *et al.* 1991). Because of a great deal of similarity, or *synteny*, between the rat and human genomes, reflecting evolutionary relatedness, a gene found in rat is likely to be found at

the corresponding position in humans. Even though success did not follow in this particular case, (Jeunemaitre *et al.* 1992), the basic strategy is sound. The mapping of genes for fat deposition in pigs (Andersson *et al.* 1994), for example, may have implications for understanding human obesity.

At this point it is necessary to distinguish between physical and genetic mapping. The set of hereditary material transmitted from parent to offspring is known as the *genome*, and it consists of molecules of deoxyribonucleic acid (DNA) arranged in chromosomes. The DNA itself is characterized by its nucleotide sequence – the sequence of bases A, C, G or T that bind in complementary pairs $A-C$, $G-T$ between the two strands of the DNA helical molecule. DNA sequences therefore have lengths measured in base pairs, *bp*. A physical map is an ordering of features of interest along the chromosomes in which the metric is the number of *bp* between features. This is the level of detail needed for molecular studies, and there are several techniques available for physical mapping of discrete genetic markers or traits. In the present discussion, however, we are concerned with genetic mapping where the metric is itself a variable under genetic control.

Genetic map distances depend on the level of recombination expected between two points. An individual receives one copy of each heritable unit (allele) from each parent, but the combination of units (genotype) at different locations that the individual transmits to the next generation need not be one of the parental sets. Recombination may have taken place during the process of meiosis producing eggs or sperm. That is, through *crossing over* events, alleles may come from either of the two parental chromosomes (diploid) to form the egg or sperm. Recombination between two elements on the same chromosome is more likely the further apart the elements are, with a limiting value of 50%. Although there is generally a monotonic relation between physical and recombinational distances, allowing genes to be ordered on the basis of recombination distances between them, the relation is not a simple one. The distance over which one recombinational event is expected to occur depends on the region of the genome, as well as on genes at other places in the genome. The most striking evidence of variability in the genetic map metric is provided by the human genetic maps for males and females being of different lengths.

Genetic mapping of QTLs rests on the simple idea that genetic markers which tend to be transmitted together with specific values of the trait are likely to be close to a gene affecting that trait. In other words, an association is sought between marker variants (genotypes) and

trait values (phenotypes), with higher levels of association suggesting closer genetic map distance. Locating QTLs has a long history, initially with visible markers. Recent progress rests on the availability of an almost inexhaustible supply of molecular markers that has overcome "The main practical limitation of the technique seems to be the availability of suitable markers" (Thoday 1961). Associations with molecular markers have already been reported for yield, quality traits and insect resistance in tomato (Nienhuis *et al.* 1987, Paterson *et al.* 1991), and for yield, abiotic stress and morphological characters in maize (Edwards *et al.* 1987, Stuber *et al.* 1987; Abler *et al.* 1991; Reiter *et al.* 1991). Milk protein genes have been used as markers for dairy cattle traits (Bovenhuis and Weller 1994). Work is even proceeding in the search for genes affecting behavioral traits in mice (Plomin *et al.* 1991). Evidently, these searches for associations will be statistical, continuing the long tradition of the use of statistics in quantitative genetics.

2 Notation

Genetic markers (often referred to as markers) are neutral markers having no affect on an individuals phenotype. Through molecular techniques, these markers may be identified and arranged so that each chromosome is represented by a linear arrangement of neutral markers. The markers are then used as a *genetic map* of the organism's genome (genetic structure) for the purpose of detecting regions of the genome associated with a specific trait of interest. Genetic markers will be represented by letters **M, N, L,** .... Generally markers will be used that have two variants (alleles), denoted by subscripts, e.g. M_1, M_2. Traditional experimental designs for locating QTLs start with two parental lines differing both in trait values and in the marker variants they carry. Quantitative trait alleles are denoted by Q_1 and Q_2, with Q_1/Q_2 denoting the unknown quantitative trait loci (QTL) genotype. Our goal is to detect the QTL by relying on the association between the measured trait values recorded for each individual and the genetic map information. In practice, markers are sought that have different alleles in the parents. Without loss of generality, suppose two pure-breeding (inbred) lines of parents have marker genotypes M_1N_1/M_1N_1 and M_2N_2/M_2N_2 (homozygous). Crossing these lines produces an offspring, or F_1, generation that is heterozygous at both loci: M_1N_1/M_2N_2, where the slash separates the contributions from the two parents (chromosomes). Each F_1 individual produces four possible gametes, or marker allele combinations,

for transmission to the next generation. The proportions of these four gametes can be expressed in terms of the recombination fraction r_{MN} between the two markers, and is referred to as the genotypic array:

$$\frac{1-r_{MN}}{2}M_1N_1 + \frac{r_{MN}}{2}M_1N_2 + \frac{r_{MN}}{2}M_2N_1 + \frac{1-r_{MN}}{2}M_2N_2$$

and this serves to define r_{MN}. Unlinked markers, those on different chromosomes for example, recombine freely so that all four gametes will be equally frequent, illustrating that $0 \leq r_{MN} \leq 0.5$.

2.1 Recombination and Map Functions

For more than two markers, a simplifying assumption is that recombination between any two of them is independent of recombination between any other two. With this assumption called *no interference*, and a Poisson–process assumption for the phenomenon of crossing over between DNA strands, recombinational fractions r are related to genetic distances x by means of Haldane's mapping function (Haldane 1919):

$$r = \frac{1}{2}(1 - e^{-2x})$$

Genetic distances are expressed in terms of centiMorgans, cM, with one Morgan being the distance over which one recombinational event is expected to occur, and are sometimes preferred to the recombination probability because cM distances are additive, whereas recombination fractions are not. When recombination is not independent, interference is assumed and the Kosambi map function (Kosambi 1941) is appropriate. Further details on modeling interference in genetic recombination are discussed in Speed *et al.* (1992), McPeek and Speed (1995), Zhao *et al.* (1995a,b).

2.2 Variation

Values for the measurable quantitative trait of interest will be denoted by Y and, for genetically homogeneous populations, will be taken to be normally distributed, possibly after transformation. Trait values contain genetic and environmental components G and E, with the simplest model being

$$Y = G + E$$

For uncorrelated genetic and environmental effects, the total (*phenotypic*) variance of the trait can be partitioned into genetic and environmental components

$$V_Y = V_G + V_E$$

For a trait affected by a single gene **Q**, individuals with genotype Q_iQ_j have genotypic value expressed in terms of a mean, additive and dominance effects:

$$G_{ij} = \mu + a_i + a_j + d_{ij}$$

Multilocus traits may include epistatic interactions between the loci.

It is often not made explicit that the magnitude of the various genetic components depends on the genetic constitution of the population. Suppose a population has genotypic array

$$P_{11}(Q_1Q_1) + P_{12}(Q_1Q_2) + P_{22}(Q_2Q_2)$$

where P_{ij} is the frequency of the Q_iQ_j genotype. Fitting the mean, additive and dominance effects by least squares, under the constraints

$$\begin{aligned}(2P_{11} + P_{12})a_1 + (P_{12} + 2P_{22})a_2 &= 0\\ (2P_{11} + P_{12})d_{11} + (P_{12} + 2P_{22})d_{12} &= 0\\ (2P_{11} + P_{12})d_{12} + (P_{12} + 2P_{22})d_{22} &= 0\end{aligned}$$

provides

$$\begin{aligned}\mu &= P_{11}G_{11} + P_{12}G_{12} + P_{22}G_{22}\\ a_1 &= (P_{11} + P_{12}/2)G_{11} + (P_{12}/2 + P_{22})G_{12} - \mu\\ a_2 &= (P_{11} + P_{12}/2)G_{12} + (P_{12}/2 + P_{22})G_{22} - \mu\\ d_{ij} &= G_{ij} - a_i - a_j - \mu\end{aligned}$$

Although the genotypic values G depend only on the genotype, the additive, dominance and epistatic components depend on genotypic frequencies and so are population-dependent. Partitioning the genotypic values leads to a partitioning of the genetic variance into additive and dominance

components:

$$\begin{aligned} V_G &= P_{11}G_{11}^2 + P_{12}G_{12}^2 + P_{11}G_{11}^2 - \mu^2 \\ &= V_A + V_D \end{aligned}$$

Finally, the ratio of additive genetic variance to total variance is termed the heritability h^2, and quantifies the degree to which a trait is resolved genetically.

3 Numbers of Loci Affecting a Trait

A preliminary investigation of how many loci affect a quantitative trait may give some indication of the chances of success in locating QTL. It will be easier to locate genes (QTL) when only a few affect the trait than when many genes are involved. A simple approach was given by Wright (in Castle 1921). If M loci affect a character, then Wright gave

$$M = \frac{(\mu_1 - \mu_2)^2}{8\sigma^2} \tag{1}$$

where μ_1, μ_2 are the means of two parental populations and σ^2 is the additive genetic variance stemming from differences in allele frequencies of the parental populations. Theoretically, the estimate of the number of genes is possible if they are of large effect. Equation (1) assumes additivity and equality of the effects of the M loci. Cockerham (1986) modified Wright's approach to accommodate bias in the estimated values of $(\mu_1 - \mu_2)^2$. Zeng *et al.* (1990) allowed for unequal gene effects and for linkage between the loci. Lande (1981) and Comstock and Enfield (1981) have also suggested derivations of the number of genes (loci) affecting a trait.

4 Single-Marker, Single QTL Analyses

4.1 Comparison of Marker Means

The use of genetic markers to locate QTL is well established (Sax 1923, Thoday 1961, Elston and Stewart 1973, Soller *et al.* 1976, Edwards *et al.* 1977, Darvasi and Weller 1992). Investigations by Sax (1923) were initiated through the association of seed coat pattern and pigmentation with the seed size differences in *Phaseolus vulgaris* (common name, the bean). This study was one of the

initial demonstrations of linkage between major gene differences and determinants of quantitative variation. The findings of Sax showed color difference to be characterized by a single gene difference.

Considerable attention has been paid to the case of associations between a single marker and a quantitative trait (Weller 1986, Beckman and Soller 1988, Luo and Kearsey 1989, Luo and Woolliams 1993) and we now review the statistical issues. Observations on marker genotype and trait value are taken in order to test the hypothesis that the two loci are unlinked, i.e. the recombination fraction between them is 0.5. Rejection of this hypothesis has a dual implication. Not only does it confirm a genetic basis for the trait, but also it suggests that the trait is affected by a gene close to the marker.

Classical work is conducted within the two experimental designs shown in Figure 1. Two inbred lines P_1, P_2 are chosen as parents. Often these will have been selected in opposite directions for the trait, to ensure that they differ in trait values because they carry different variants, or alleles, at the trait locus. Similarly, markers are chosen with different alleles in the two parents. Inbreeding of P_1, P_2 means that these lines are homozygous at trait and marker loci. The F_1 generation can be either backcrossed to P_1 or P_2, or mated among itself (selfing or crossing) to produce the second filial, or F_2, generation. Observations on marker and trait values for the backcross, B_1, B_2, or F_2 individuals are used in tests of association. For the purpose of notational development, we continue the statistical derivation in terms of a backcross model. An F_2 experimental design will serve as an example of methodology later in the paper.

Under a completely additive model, the trait mean for the F_1 individuals is the average of the two parental means. Since all three groups, P_1, P_2, F_1, are genetically uniform, they are assigned the same trait variance σ^2. Individuals within the backcross and F_2 generations, however, have mixtures of trait and marker genotypes with the mixing proportions depending on the recombination fraction between the two loci.

For the B_1 design (see Appendix 1 for analogous derivation of F_2 design), the genotypic array is

$$\frac{1-r_{MQ}}{2}M_1Q_1/M_1Q_1 + \frac{r_{MQ}}{2}M_1Q_1/M_1Q_2 + \frac{r_{MQ}}{2}M_1Q_1/M_2Q_1 + \frac{1-r_{MQ}}{2}M_1Q_1/M_2Q_2$$

with a similar expression for B_2. Only the marker genotype can be directly observed, so the B_1 individuals can be separated into two observable classes: marker types M_1/M_1 and M_1/M_2. The

expected trait distributions within these two classes are

$$\begin{aligned} M_1/M_1: &\quad (1-r_{MQ})N(\mu_1,\sigma^2)+r_{MQ}N(\mu_{12},\sigma^2) \\ M_1/M_2: &\quad r_{MQ}N(\mu_1,\sigma^2)+(1-r_{MQ})N(\mu_{12},\sigma^2), \end{aligned}$$

where $N(\mu,\sigma^2)$ denotes a normal distribution with mean μ and variance σ^2. The means and variances of these two mixture distributions are

$$\begin{aligned} \mu_{M_1/M_1} &= (1-r_{MQ})\mu_1+r_{MQ}\mu_{12} \\ \mu_{M_1/M_2} &= r_{MQ}\mu_1+(1-r_{MQ})\mu_{12} \\ \sigma^2_{M_1/M_1}=\sigma^2_{M_1/M_2} &= \sigma^2+r_{MQ}(1-r_{MQ})(\mu_1-\mu_{12})^2 \\ &= \sigma^2+r_{MQ}(1-r_{MQ})\delta^2 \end{aligned}$$

This defines δ as half the difference between the P_1 and F_1 means. The expected difference in average trait values between the two classes is

$$\mu_{M_1/M_1}-\mu_{M_1/M_2} = (1-2r_{MQ})\delta \tag{2}$$

Providing lines P_1 and F_1 have different mean trait values, the hypothesis that trait and marker loci are unlinked, $r_{MQ}=0.5$, is therefore equivalent to the hypothesis that the two marker classes in a backcross generation have equal means. Since the original lines P_1 and P_2 were chosen because they differed for the trait, the condition $\delta \neq 0$ will be satisfied unless allele Q_1 is completely dominant to Q_2. The classic test appeals to the robustness of the t-test and uses the test statistic

$$t = \frac{\tilde{\mu}_{M_1/M_1}-\tilde{\mu}_{M_1/M_2}}{\sqrt{s^2(\frac{1}{n_{M_1/M_2}}+\frac{1}{n_{M_1/M_2}})}}$$

where tildes denote sample means, the sample sizes of the two marker classes are $n_{M_1/M_1}, n_{M_1/M_2}$, and the pooled estimate of the variance within the two classes is s^2.

The issue could be raised as to the validity of either t-tests or analyses of variance since the trait distributions within marker classes are mixtures of normals rather than normals themselves. In the backcross B_1 population, the coefficients of skewness S and kurtosis K in the two marker classes are

$$S_{M_1/M_1}=-S_{M_1/M_2} = \frac{2r_{MQ}(1-r_{MQ})(1-2r_{MQ})\Delta^3}{[1+r_{MQ}(1-r_{MQ})\Delta^2]^{3/2}}$$

$$K_{M_1/M_1} = K_{M_1/M_2} = \frac{r_{MQ}(1-r_{MQ})(1-6r_{MQ}+6r_{MQ}^2)\Delta^4}{[1+r_{MQ}(1-r_{MQ})\Delta^2]^2}$$

where $\Delta = (\mu_1 - \mu_{12})/\sigma = \delta/\sigma$ is the standardized difference between the P_1 and F_1 means. The mixtures are therefore symmetric when the trait locus is either completely linked ($r_{MQ} = 0$) or completely unlinked ($r_{MQ} = 0.5$) to the marker locus. Otherwise there is skewness that has maximum numerical value at a point depending on Δ. The mixtures have zero kurtosis for $r_{MQ} = 0, 0.21$ (Doerge 1993). Both skewness and kurtosis, and hence non-normality, increase with Δ. From work of Eisenberger (1964), a sufficient condition that the mixtures will be unimodal for all values of r_{MQ} is $\Delta < 1.84$, whereas a sufficient condition that there exists an r_{MQ} value between zero and one giving bimodality is that $\Delta > 2$. Departures from the nominal distributions of the test statistic for the t-test and analysis of variance are therefore anticipated only for parental populations with large differences between means, but this is the condition for which it is most likely there will be departures from the null hypothesis. The generally satisfactory nature of the t-test for detecting linkage between a single QTL and a single marker has been demonstrated by simulation (Doerge 1993).

4.2 Regression

In work that anticipates later multi-marker approaches, we now consider regressing the trait value on marker genotype. For the jth individual in backcross population B_1, the model is

$$Y_j = \beta_0 + \beta_{YX} X_j + \epsilon_j \tag{3}$$

where the indicator variable X_j takes the values 1 or 0 according to whether the individual has marker genotype M_1/M_1 or M_1/M_2, and ϵ_j is a random error term (not normally distributed). The regression coefficient for Y on X

$$\beta_{YX} = (1-2r_{MQ})\delta$$

is the expected difference between the recurrent parent and the F_1 (2). The hypothesis of the marker and trait loci being unlinked can be tested by testing for a non-zero slope to the regression line of trait value on marker indicator. This approach is valid for all non-trivial partitions of the

sample into two marker classes, but it still assumes that the trait values are distributed normally within each marker class. Care should be taken in applying the test: if δ is known to be positive (or negative) from observations on the parents, then the alternative to $H_0 : \beta_{YX} = 0$ is $H_1 : \beta_{YX} > 0$ ($H_1 : \beta_{YX} < 0$) since there is a biological constraint that $(1 - 2r_{MQ})$ is not negative.

4.3 Likelihood

The fact that trait values have mixtures of normal distributions within marker classes can be taken into account properly with likelihood analyses. Estimates of the recombination fraction can also be derived in the likelihood framework (in the other approaches, moment estimators can be constructed for the recombination fraction). If Y_{1i}, Y_{2i} are the trait values for the ith individuals in B_1 marker classes $M_1/M_1, M_1/M_2$, then the likelihood L for the parameters $\mu_1, \mu_{12}, \sigma^2, r_{MQ}$ is

$$L = \prod_{i=1}^{n_{M_1/M_1}} \left[\frac{1 - r_{MQ}}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(Y_{1i} - \mu_1)^2}{2\sigma^2} \right) + \frac{r_{MQ}}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(Y_{1i} - \mu_{12})^2}{2\sigma^2} \right) \right]$$

$$\times \prod_{i=1}^{n_{M_1/M_2}} \left[\frac{r_{MQ}}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(Y_{2i} - \mu_1)^2}{2\sigma^2} \right) + \frac{1 - r_{MQ}}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(Y_{2i} - \mu_{12})^2}{2\sigma^2} \right) \right]$$

The hypothesis of interest can be tested for with the likelihood ratio statistic

$$\lambda = -2 \ln \left[\frac{L(\hat{\mu}_1, \hat{\mu}_{12}, \hat{\sigma^2}, r_{MQ} = 0.5)}{L(\hat{\mu}_1, \hat{\mu}_{12}, \hat{\sigma^2}, \hat{r}_{MQ})} \right]$$

with carets denoting maximum likelihood estimates. The estimates for $\mu_1, \mu_{12}, \sigma^2$ will be different in the numerator and denominator in this and subsequent likelihood ratios. The ratio is often assumed to be distributed as chi-square with one d.f. under the null hypothesis $r_{MQ} = 0.5$, although there is the problem that this hypothesis puts the parameter r_{MQ} at a boundary value (Self and Liang 1987).

Even at the simple level of a single marker and single trait locus, the likelihood calculations are not trivial. One possibility is to use prior estimates of the trait means and variance, μ_1, μ_2, σ^2, possibly from the parental lines. Care would be needed to check for consistency of non-genetic effects for the three generations, P, F_1, B, and a check that the F_1 had the postulated distribution of trait values should be performed. Use of such prior estimates reduces the likelihood to a function of a single parameter, although iterative methods for solution will still be necessary.

Another procedure is to evaluate the test statistic over a grid of r_{MQ} values, as is done in human pedigree linkage studies (Ott 1991, Morton 1995). Following the convention for those analyses, results are expressed in terms of the LOD score:

$$\text{LOD} \quad = \quad -\log_{10}\left[\frac{L(\hat{\mu}_1, \hat{\mu}_{12}, \hat{\sigma^2}, r_{MQ} = 0.5)}{L(\hat{\mu}_1, \hat{\mu}_{12}, \hat{\sigma^2}, r_{MQ})}\right]$$

The maximum LOD score indicates the grid value r_{MQ} closest to the maximum likelihood estimate $\hat{r}_{QM}$. If a smooth curve is fitted to the set of LOD values, an indication of precision is provided by the 2-LOD interval which is the range of values between those r_{MQ}'s at which the LOD is two less than its maximum value. Under the assumption that the likelihood ratio follows a function of a chi-square distribution with one d.f. this interval corresponds approximately to a 95% confidence interval. Jansen (1992) uses the EM–algorithm (Dempster *et al.* 1977) to estimate the model parameters, the same algorithm may be used for single marker regression situations.

5 Genetic Map

There exists an underlying complexity to the search for QTL which begins with the ordering of genetic markers into chromosomes, for the eventual representation of the entire genome. As mentioned in the introduction, it is generally the case that many markers are available to use in the search for loci affecting quantitative traits. Genetic markers may be arranged in linear order across chromosomes with the measure of association between them being either recombination or map distance (cM). The closer together two markers are, the smaller their distance/recombination will be. When recombination between pairs of markers is used to order markers, this is called *two–point* analysis (Ott, 1991; p.54). When all possible recombinant classes are calculated, *multipoint analysis* (Lathrop, 1985) may be used to estimate a more accurate genetic map. The genetic marker ordering problem is analogous to the historic *traveling salesman problem* in which a salesman is asked to travel between cities in the shortest possible route. Several useful methods have been described for the purpose of estimating genetic maps, including branch and bound methods (Thompson, 1984), simulated annealing (Corona *et al.*, 1987, Weeks and Lange, 1987; Falk, 1992), and seriation (Buetow and Chakravarti, 1987a,b). The associations among genetic markers may be exploited for the purpose of having more information available in the search for QTL. One continuing controversy

between statisticians and plant breeders is the issue of sample size versus number of genetic markers. A reasonably large number of individuals must be measured and genotyped in order to assess the quantitative variation and phenotype–genotype association. However, an acceptable number of genetic markers must be used in order to cover the entire genome. Due to the costly laboratory techniques, greenhouse space, field plots, marker scoring and data entry, the question of sample size versus genome coverage arises. Is it better to grow more individuals and score fewer markers, or score more markers on fewer individuals? Clearly, from the parameter estimation standpoint large sample size on a uniformly distributed genetic map is sensible. Realistically, since the goal is to locate QTL, a *dense map* (many markers) is preferred over a *sparse map* (fewer markers) since it allows a greater precision of location.

From this point forward we will assume that a known genetic map has already been estimated. Although it is certainly possible to apply single-marker tests for each marker in turn, a more efficient procedure is one in which the markers are used all together. This is the rationale behind current multiple regression approaches, but we first review the use of pairs of markers.

6 Interval Mapping

Any indication that the recombination fraction r_{MQ} is less than the value 0.5 from single-marker analyses is confounded by the size of effects of locus **Q**, since it is actually the product $(1-2r_{MQ})\delta$ that is being tested for departures from zero. A marker close to a QTL of small effect will give the same signal as a marker some distance from a QTL of large effect. Also, it will not be known whether the two loci are in order **QM** or **MQ** on a genetic map. If two markers **M, N** are used, however, it should be possible to separate the recombination and size of effect as well as to infer the position of **Q** relative to both. It is also expected that more precision and power will follow simply from the use of the extra information from a second marker.

6.1 Likelihood method

Two markers Continuing the treatment of the backcross mating scheme, suppose the two parental lines have marker genotypes M_1N_1/M_1N_1 and M_2N_2/M_2N_2. Backcrossing M_1N_1/M_2N_2

F_1 individuals to P_1 results in four distinguishable marker classes (Tables 1 and 2), in expected proportions depending on the recombination fraction r_{MN} between the two markers, which contribute to the genotypic array:

$$\frac{1-r_{MN}}{2}M_1N_1/M_1N_1 + \frac{r_{MN}}{2}M_1N_1/M_1N_2 + \frac{r_{MN}}{2}M_1N_1/M_2N_1 + \frac{1-r_{MN}}{2}M_1N_1/M_2N_2$$

The trait distributions within each marker class depend in the first place on whether the trait locus is inside or outside the interval **MN**. For each of the three possible orders of trait and marker loci, the frequencies of the eight possible genotypes is shown in Table 1.

Primary interest is in the order that places the trait locus between the two markers. Under the assumption of order being true, calculations are performed by stepping along the marker interval and assigning appropriate recombination values r_{MQ}, r_{NQ}. Specifically, the likelihood for **Q** being unlinked to both markers is compared to the likelihoods that it is at specific interior points in the interval. An hypothesis testing approach would instead use mutually exhaustive alternatives:

$$\begin{aligned} H_0: &\quad r_{MQ} = r_{NQ} = 0.5 \quad \text{QTL unlinked to markers} \\ H_1: &\quad \min(r_{MQ}, r_{NQ}) < 0.5 \quad \text{QTL linked to markers} \end{aligned}$$

or

$$\begin{aligned} H_0: &\quad \min(r_{MQ}, r_{NQ}) > r_{MN} \quad \text{QTL exterior to interval} \\ H_1: &\quad \min(r_{MQ}, r_{NQ}) < r_{MN} \quad \text{QTL interior to interval} \end{aligned}$$

Under the assumption of no interference mentioned earlier, the three recombination fractions r_{MQ}, r_{NQ}, r_{MN} are related. When **Q** is interior to **MN**, the event of no recombination between **M, N** is equivalent to no recombination in both intervals **MQ** and **QN**, or recombination in both intervals:

$$\begin{aligned} (1-r_{MN}) &= (1-r_{MQ})(1-r_{NQ}) + r_{MQ}r_{NQ} \\ r_{MN} &= r_{MQ} + r_{NQ} - 2r_{MQ}r_{NQ} \\ (1-2r_{MN}) &= (1-2r_{MQ})(1-2r_{NQ}) \end{aligned}$$

For the order **QMN**, the relationship becomes

$$(1-2r_{NQ}) = (1-2r_{MQ})(1-2r_{MN})$$

It is taken that r_{MN} is known, so that there is only one independent unknown recombination fraction. Note that neither r_{MQ} nor r_{NQ} can equal 0.5 when $r_{MN} < 0.5$. A very useful discussion of the hypotheses of relevance was given by Knott and Haley (1992).

The mixture distributions for the four marker classes can be written as

$$\begin{aligned}
M_1N_1/M_1N_1: &\quad c_{11}N(\mu_1,\sigma^2)+(1-c_{11})N(\mu_{12},\sigma^2)\\
M_1N_1/M_1N_2: &\quad c_{12}N(\mu_1,\sigma^2)+(1-c_{12})N(\mu_{12},\sigma^2)\\
M_1N_1/M_2N_1: &\quad c_{21}N(\mu_1,\sigma^2)+(1-c_{21})N(\mu_{12},\sigma^2)\\
M_1N_1/M_2N_2: &\quad c_{22}N(\mu_1,\sigma^2)+(1-c_{22})N(\mu_{12},\sigma^2)
\end{aligned}$$

For the F_2 design, there are nine distinguishable marker classes, each having a mixture of three normals for the trait distribution.

From Table 1, the backcross mixing proportions for the **MQN** order are

$$c_{11} = 1 - c_{22} = \frac{(1-r_{MQ})(1-r_{NQ})}{(1-r_{MN})}$$

$$c_{21} = 1 - c_{12} = \frac{r_{MQ}(1-r_{NQ})}{r_{MN}}$$

The four marker-class trait means cannot be equal.

For the order **QMN**, the mixing proportions are

$$c_{11} = c_{12} = 1 - c_{21} = 1 - c_{22} = 1 - r_{MQ}$$

so that the distributions are the same for the M_1N_1/M_1N_1 and M_1N_1/M_1N_2 classes and there is no need to record the N type. No additional information is provided from outside the working interval. Similarly, for order **MNQ**

$$c_{11} = c_{12} = 1 - c_{21} = 1 - c_{22} = 1 - r_{NQ}$$

and there is no need to record the M type. Outside the marker interval, calculations reduce to those for one marker (the nearest) and are based on only two marker classes. In either of these two cases of a QTL outside the marker interval, the two marker class means are equal if and only if $r_{MQ} = r_{NQ} = 0.5$, suggesting that the first of the pairs of hypotheses above be addressed by

a t-test on marker class observations. Certainly rejection of the hypothesis of equal marker class means would imply that **Q** was linked to either or both of **M, N**, although it would not necessarily place **Q** between **M** and **N**.

It is straightforward to evaluate the likelihood L from observations on the two or four marker classes, although computationally demanding if the parameters $\mu_1, \mu_{12}, \sigma^2$ have to be estimated. Matters are simplified by assigning values to r_{MQ}, r_{NQ}. This means specifying a map position for the QTL, relative to the marker interval, and invoking a mapping function to provide the two recombination fractions. For positions to the left (or right) of the interval **MN**, the one-marker LOD scores can be evaluated using marker **M** (or **N**). For positions inside the interval, it is usual to use the two-marker LOD score evaluated for the four marker classes

$$\text{LOD} \quad = \quad -\log_{10}\left[\frac{L(\hat{\mu}_1, \hat{\mu}_{12}, \hat{\sigma^2}, r_{MQ} = r_{NQ} = 0.5)}{L(\hat{\mu}_1, \hat{\mu}_{12}, \hat{\sigma^2}, r_{MQ}, r_{NQ})}\right]$$

even though the denominator is not the unconstrained likelihood over all possible recombination values. It is important to recognize that the LOD score does not provide a test for the presence of a QTL between the two markers, and so is not leading to a true interval test. Instead the LOD compares the likelihood of the QTL being at the position characterized by recombination fractions r_{MQ}, r_{NQ} against the likelihood that it is at some position unlinked to the interval. Of course, the map position at which the LOD score is greatest is likely to be close to the location of the QTL.

The LOD scores at the interval boundaries are the same, whether they result from setting $r_{MQ} = 0$ in the analysis using only marker **M**, or from setting $r_{MQ} = 0, r_{NQ} = r_{MN}$ in the analysis using both markers.

Lander and Botstein One of the most influential papers of the late 1980's pertaining to the locating a single QTL can be credited to Lander and Botstein (1989). Working from a known genetic map, the Lander–Botstein interval mapping method employs a simple linear regression model similar to the one defined in (2). Since the distance between each pair of genetic markers is known, the method steps through the interval in specified increments, using a map function, and then estimates the model parameters at the analysis point. The likelihood equation is calculated under the estimated parameters, and then again under the null hypothesis of $\beta_{XY} = 0$ (no QTL present). The ratio of the two likelihood evaluations is calculated in the form of a LOD score for

each analysis point in the genome. The maximum LOD score over all analysis points is indication of a single QTL if the maximum LOD score is larger than some specified threshold value. We will discuss the implications of multiple tests and the distribution of the trait values on the distribution of the test statistic in a later section of this paper. The essence of the Lander-Botstein approach is that trait loci are postulated to occur at a series of positions within a set of adjacent marker intervals, and the trait observed value (*the phenotype*) is regressed on the number of F_1 trait alleles (*the genotype*). The regression approach was expanded upon by Martinez and Curnow (1992), as well as many others.

Many Markers Martinez and Curnow (1992) considered the four marker classes for the case of two markers in a backcross. Within each marker class they regressed the trait values on the probability that the individual had the F_1 trait genotype. As this probability depends on the unknown recombination fractions between trait and marker loci, they performed the regressions at a series of specified recombination values. They then formed a residual sum of squares of differences between trait observations and fitted values, summing over all four marker classes, and took the minimum to indicate the best estimate of the position of the trait locus. This approach allows an analytical treatment whereas likelihood methods do not.

The usual procedure for interval mapping is to calculate LOD scores at interior points of a series of adjacent marker intervals. For markers **L, M, N**, for example, there will be two intervals **LM** and **MN**. The maximum value of the curve fitted to the LOD scores indicates the probable position of the QTL, and 2-LOD intervals can be constructed. As in the single-interval case, the LOD scores at each marker are the same whether the marker is at the left or the right of an interval. If there is a QTL in one interval, adjacent intervals may also show peaks with "significant" likelihood ratios, often called *ghosting effects* (Knapp *et al.* 1990, Martinez and Curnow 1992, Jansen 1994).

Ghost Effects Ghosting effects occurs when a QTL is linked to one genetic marker and by the definition of genetic mapping, the additional adjacent genetic markers associated through linkage may also exhibit significant test statistics. A problem with traditional interval mapping is that it does not take account of all markers at once, but uses them only two at a time so that it is difficult to discriminate between actual QTL effects and ghost QTL effects that exist simply because of the

relative density of the genetic map being used. Martinez and Curnow (1992) illustrate numerically that "ghosting" can occur – if there are trait loci $\mathbf{Q}_1$, $\mathbf{Q}_2$ in non-adjacent intervals $\mathbf{M}_1$, $\mathbf{M}_2$ and $\mathbf{M}_3$, $\mathbf{M}_4$, there will be spurious indications of a trait locus in the intervening interval $\mathbf{M}_2$, $\mathbf{M}_3$. Haley and Knott (1992) also drew attention to the biases resulting from linked trait loci. The same phenomenon is expected for the traditional LOD-score approach of Lander and Botstein (1989). The "ghosting" shown numerically by Martinez and Curnow is a similar phenomenon to that anticipated by Paterson *et al.* (1991): "If a QTL is actually present in one interval, the hypothesis of a QTL in an adjacent interval will still fit the data better than the hypothesis of [*sic*] no QTL at all, and the more likely position of a QTL in this adjacent interval will often be near the middle of the interval (since this position is furthest from any potentially conflicting data at the observed markers.) Accordingly, multiple peaks correctly reflect the shape of the likelihood surface but need not indicate multiple QTLs." The authors meant to contrast the cases of linked or unlinked QTLs, rather than the presence of absence of QTLs. The fact that P_1, P_2, F_1 have different trait values means that there are QTLs. The detection of "ghosts" was also a concern of McMillan and Robertson (1974) in their important discussion of methods for detecting loci affecting quantitative traits in *Drosophila.* They referred to two errors "(i) The detection of loci which do not exist. (ii) The magnification of the estimated effect of those major loci which do exist by accumulating to their effect those of undetected loci close to them on the chromosome." Zeng (1993, 1994) has demonstrated ghosting effects by showing that interval mapping gives results that can be confounded by the presence of additional QTLs outside the interval being considered. Zeng's method (which will be discussed later) shows evidence of a QTL in the two intervals $\mathbf{M}_1$, $\mathbf{M}_3$ and $\mathbf{M}_2$, $\mathbf{M}_4$, but would avoid the problem if there were three markers between the QTL.

6.2 Regression Methods

There has been a growing realization that the appropriate way to relate quantitative traits to information on many markers is by multiple regression (Wright and Mowers 1994, Kearsey and Hyne 1994, Wu and Li 1994). Moreno-Gonzalez (1992a,b) set up a regression model containing additive, dominance and epistasis terms for putative QTLs associated with several marked chromosome segments. A more extensive discussion of the theoretical issues for regression on additive and

dominance effects was given by Jansen (1992, 1993). Jansen and Stam (1994) have included parental and F_1 information in their multiple regression analyses of F_2 and other crosses.

Regression on Marker Genotypes For a pair of linked markers, **M** and **N**, the trait value Y_j for individual j can be regressed on indicator variables X_{ij} that take the value 1 if, for the first marker, $\mathbf{M}_i$, individual j has the P_1 genotype and value 0 if it has the F_1 genotype. The model is

$$Y_j = \beta_0 + \beta_{YX_1.X_2}X_{1j} + \beta_{YX_2.X_1}X_{2j} + \epsilon_j$$

where $\beta_{YX_1.X_2}$ is the coefficient of regression of Y on X_1 conditional on the value of X_2.

The partial regression coefficients for the trait on one marker, holding the other marker constant, do depend on the marker ordering. Regressing on the indicator for **M**, holding constant the other indicator variable and invoking the relationships among the recombination values r_{MN}, r_{MQ}, r_{NQ} when there is no interference gives:

$$\beta_{YX_1.X_2} = \begin{cases} (1-2r_{MQ})\delta & \text{order } \mathbf{QMN} \\ \dfrac{r_{NQ}(1-r_{NQ})(1-2r_{MQ})}{r_{MN}(1-r_{MN})}\delta & \text{order } \mathbf{MQN} \\ 0 & \text{order } \mathbf{MNQ} \end{cases}$$

If a test of the hypothesis that this coefficient is zero is not rejected, there is support for **Q** either being unlinked to **M**, or coincident with **N**, or to the side of **N** away from **M**. If the tests for both $\beta_{YX_1.X_2}$ and $\beta_{YX_2.X_1}$ indicate non-zero values, then the QTL is placed within the marker interval. Testing procedures are given, for example, by Stuart and Ord (1991). A flow chart for interval mapping of many QTL is given by Jansen (1993).

When a series of markers are available there is a straightforward expansion of the previous regression equation. In a further extension, Zeng (1993, 1994) explicitly allows for several QTLs affecting the trait. If dominance and epistasis are ignored, the genetic model for the trait is

$$G = \mu + \sum_k (a_{ku} + a_{kv})$$

for individuals with genotype $Q_{ku}Q_{kv}$ at the kth QTL $\mathbf{Q}_k$, where u and v denote allele number. With several QTLs, B_1 individuals have a range of trait genotypes with frequencies depending

on the recombination between trait loci. If m QTLs are named according to their order, the B_1 genotypic array is

$$2^{m-1}\left(\frac{1}{2}Q_{11}+\frac{1}{2}Q_{12}\right)\prod_{k=2}^{m}\left(\frac{1-r_{Q_{k-1},Q_k}}{2}Q_{k1}+\frac{r_{Q_{k-1},Q_k}}{2}Q_{k2}\right)$$

and the genetic variance of this array is

$$\sigma_G^2 = \frac{1}{4}\sum_{k=1}^{m}\delta_k^2+\frac{1}{4}\sum_{k,k'=1;k\neq k'}^{m}(1-2r_{Q_k,Q_{k'}})\delta_k\delta_{k'}$$

where the recombination fractions between non-adjacent loci follow from the no-interference arguments shown above. The products of effects at different loci affect the variance only for linked loci in this additive model.

If we denote m ordered markers as $\mathbf{M}_1, \mathbf{M}_2, \cdots, \mathbf{M}_{i-}, \mathbf{M}_i, \mathbf{M}_{i+}, \cdots, \mathbf{M}_m$, the partial regression coefficient $\beta_{YX_i.S_i}$ of the trait on the indicator variable for the ith marker $\mathbf{M}_i$, conditional on the set S_i of all other markers, depends only on those QTLs in the two marker intervals ($\mathbf{M}_{i-}$, $\mathbf{M}_i$), ($\mathbf{M}_i$, $\mathbf{M}_{i+}$) that have marker $\mathbf{M}_i$ as a common boundary

$$\begin{aligned}\beta_{YM_i.S_i} = &\sum_{i-<k\leq i}\frac{r_{M_{i-}Q_k}(1-r_{M_{i-}Q_k})}{r_{M_{i-}M_i}(1-r_{M_{i-}M_i})}(1-2r_{Q_kM_i})\delta_k\\ &+\sum_{i<k<i+}\frac{r_{Q_kM_{i+}}(1-r_{Q_kM_{i+}})}{r_{M_iM_{i+}}(1-r_{M_iM_{i+}})}(1-2r_{M_iQ_k})\delta_k\end{aligned}$$

In other words, the partial regression coefficient of trait value on the indicator variable for marker $\mathbf{M}_i$ is non-zero only when there are QTLs in either or both of the two marker intervals with $\mathbf{M}_i$ as a common boundary. The logic of this, as well as the algebraic details, reduce correctly to those given above for two markers and one QTL. Partial regression therefore leads to a test for the presence of QTLs in the marker interval ($\mathbf{M}_{i-}$, $\mathbf{M}_{i+}$) only, regardless of the presence of other QTLs in the genome.

Regression on Trait Genotypes Trait values can also be regressed on the (unknown) trait genotypes. If trait locus **Q** is inside the marker interval **MN**, the frequencies of the four marker classes among backcross B_1 genotypes are given in Table 2. The trait value Y is regressed on the indicator variable X^* defined as

$$X^* = \begin{cases}1, & \text{if trait genotype is } Q_1Q_1\\ 0, & \text{if trait genotype is } Q_1Q_2\end{cases}$$

with regression equation for the jth individual:

$$Y_j = \beta_0 + \beta^* X_j^* + \epsilon_j$$

The sample regression coefficient $\tilde{\beta}^*$ depends on the numbers of individuals in each of the four marker classes, and on the recombination values r_{MN}, r_{MQ}, r_{NQ}. The last two of these recombination values depend, in turn, on the assumed location of the QTL. In expectation, however, recombination does not affect the regression coefficient. From the entries in Table 2, the expected values are

$$\begin{aligned} \mathcal{E}(X^*) &= \frac{1}{2} \\ &= \mathcal{E}(X^*)^2 \\ \mathcal{E}(Y) &= \frac{\mu_1 + \mu_{12}}{2} \\ \mathcal{E}(X^*Y) &= \frac{\mu_1}{2} \end{aligned}$$

leading to a regression coefficient of

$$\beta^* = (\mu_1 - \mu_{12})$$

This shows that the regression coefficient has an expected value that does not depend on the location of the QTL, and it will be non-zero whenever the P_1 and F_1 have different mean trait values. Regression on the trait locus, since it does not involve a marker locus, is not attenuated by the recombination value between marker and trait loci.

6.3 Composite Interval Mapping

Zeng (1993, 1994) set up a model involving regression both on QTL within an interval and on marker loci outside that interval. Inference is made by maximum likelihood. This method is essentially a combination of interval mapping (Lander and Botstein, 1989) and multiple regression, and a similar strategy was adopted by Jansen (1993). We present Zeng's regression equation

$$Y_j = \beta_0 + \beta^* X_j^* + \sum_k \beta_k X_{kj} + \epsilon_j \qquad (4)$$

where X^* refers to a QTL in the interval between adjacent markers $\mathbf{M}_i$ and $\mathbf{M}_{i+}$ (recall previous notation), and X_{kj} refers to all markers $\mathbf{M}_k$ except these two. If there is no QTL in the interval, $\beta^* = 0$, since the effects of all other QTLs are removed by the β_k terms. The model is designed to detect QTLs only within the interval $\mathbf{M}_i$, $\mathbf{M}_{i+}$, and a test for the presence of such QTLs is a test of the hypothesis $H_0 : \beta^* = 0$.

Other QTLs affecting the trait may be scattered throughout the genome. The effects of these other QTLs are removed through the regressions on markers outside the interval. The regression coefficients $\beta_0, \beta^*, \{\beta_k\}$ reflect the effects of all the QTLs, and replace the previous μ_1, μ_{12} parameters. When $X_j^* = 1$, the trait is normally distributed with mean $\beta_0 + \beta^* + \sum_k \beta_k X_{kj}$ and variance σ^2 and when $X_j^* = 0$, the mean is $\beta_0 + \sum_k \beta_k X_{kj}$. For convenience, X_{0j} is defined as 1 and the sum $\beta_0 + \sum_k X_{kj}\beta_k$ written as $X_j\beta$. When a total of m markers are used in the analysis, and two markers flank the interval of interest, the quantity β is a column vector with $m-1$ components and X_j a row vector with $m-1$ components. We write the density functions of these two normal distributions as $\phi_1(Y)$ and $\phi_0(Y)$, respectively.

If the sample sizes in each of four marker classes are written as $n_l, l = 1,2,3,4$, the likelihood function for the composite interval model is

$$\begin{aligned} L(\beta_0, \beta^*, \{\beta_k\}, \sigma^2) &= \prod_{j=1}^{n_1}[\phi_1(Y_{1j})] \prod_{j=1}^{n_2}[(1-p)\phi_1(Y_{2j}) + p\phi_0(Y_{2j})] \\ &\quad \times \prod_{j=1}^{n_3}[p\phi_1(Y_{3j}) + (1-p)\phi_0(Y_{3j})] \prod_{j=1}^{n_4}[\phi_0(Y_{4j})] \end{aligned} \tag{5}$$

The quantity $p = r_{MQ}/r_{MN}$ is assumed known.

It is relatively straightforward to find the maximum likelihood estimates of the various parameters (see Appendix 2).

The ratio of maximum likelihoods, and a test that $\beta^* = 0$, requires the parameters to be re-estimated under this hypothesis. Using a zero subscript for these estimates evaluated under the null hypothesis:

$$\begin{aligned} \hat{\beta}_0 &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\ \hat{\sigma^2}_0 &= (\mathbf{Y} - \mathbf{X}\hat{\beta}_0)'(\mathbf{Y} - \mathbf{X}\hat{\beta}_0)/n \end{aligned}$$

The only potential for false indications of QTLs with the composite interval approach arises if

there are QTLs in the intervals immediately adjacent to the interval being studied.

7 Threshold Values

Each methodology discussed in this review is based on the assumption of normality either on the quantitative trait distribution, or on the error term of the model. Since the actual genotype of the QTL is/are unknown, within each known genotypic marker class one must consider each possibility for the QTL genotypes, which gives rise to the mixture distributions described previously. It is well known that deviations from the normal distribution assumption will greatly affect the distribution of the test statistic used (in this case to detect or locate the QTL), and in fact it is further known that mixture distributions fail to follow a function of a standard distribution (Ghosh and Sen 1985, Hartigan 1985, Feng 1990). Some researchers (Lander and Botstein 1989, Darvasi *et al.* 1993, Jansen 1994, Rebaï 1994a) have relied on simulations to derive the distribution of the test statistic (often a LOD score) for the purpose of gaining a threshold value which represents a desired level of significance. Analytical work has also been provided by (Lander and Botstein 1989, 1994, Feingold 1993, Rebaï 1994b, Dupuis 1994) in order to lend asymptotic support to this issue. Nonparametric (permutation) (Fisher 1935, Good 1994) based methods have also been applied to the problem of estimating empirical threshold values (Churchill and Doerge 1994), as well as Wilcoxon rank-sum (Kruglyak and Lander 1995). An alternative is to permute the trait values of the sample and then simply compare marker class means (N.L. Kaplan, personal communication). Repeated permutations lead to a distribution of the difference of means under the hypothesis of no association of trait and marker loci.

There are many benefits to each of the aboved mentioned approaches, and while no one threshold value is the true value, each if used in an informed manner may provide an appropriate threshold value against which to compare test statistics for significant QTL location. The simulation based threshold values are model dependent, and if the correct model is used, the threshold values will be accurate. Unfortunately, simulation based threshold values do little to include the effect of missing data patterns and ghost QTL. Analytical threshold values accurately reflect the sample size and map density of the experiment, but sometimes the accuracy of these threshold values is limited by small sample size and sparse marker maps. Since environmental variation plays a

large role in any experimental system, one would expect permutation based methods to provide an accurate reflection of sample size, missing data patterns, environment, as well as multiple testing issues. The computational intensity of the permutation based methods is a limiting factor in its application. For a desired significance level of 5% upwards of 1,000 permutation of the trait data must be performed, more if a smaller significance level is desired.

8 Software

Using marker loci to locate genes affecting quantitative traits has been a matter under consideration for the past 70 years, ever since Sax (1923) associated seed coat characters with seed size in the bean *Phaseolus vulgaris*. Much attention has been paid to statistical issues, but the work with the greatest practical impact on the manner in which QTL data is statistically analysed by computer has been that of Lander and Botstein (1989). The LOD-score methodology of Lander and Botestein has been incorporated into a computer package MAPMAKER (EXP and QTL) and has been widely distributed and employed (e.g. Paterson *et al.* 1988, 1991; Stuber *et al.* 1992). Wide-scale application of a computer package is always accompanied by the possibility that the underlying methodology, and especially its assumptions, are not understood by the user. In this case, Luo and Kearsey (1992) stated "the approaches and the relevant program have been widely considered by plant/animal breeders as being difficult to understand and this has hindered the efficient use of the method." These authors elaborated on the discussion given in Lander and Botstein (1989) and gave details for the F_2 design.

One of the major issues in the proper location of quantitative trait loci is the availability of software to do the analysis. While many of the procedures covered in this review are available from standard statistical packages (*e.g.* SAS, MINITAB, etc.), many of the more complicated procedures require statistical expertise. Therefore, appropriate software must be developed and distributed so that the correct analyses may be performed, and so that a service is provided from the statisticians to the mapping community.

In the following section, we analyze a real data set using publically available computer packages: MAPMAKER/EXP (Lander and Green 1987, Lincoln *et al.* 1992a,b), MAPMAKER/QTL (Paterson *et al.* 1988, Lincoln *et al.* 1992a,b), and CARTOGRAPHER (Basten *et al.* 1994.

9 Example

As a working example to this review, we use an F_2 mouse data set (Horvat and Medrano 1995) containing 190 male individuals, scored at 9 genetic markers (microsatellites) with average density 3.85 cM. The goal of this published research was to locate the *high growth (hg)* locus (QTL), a region in the mouse genome that increases both weight gain and body size of mature mice. Energy metabolism is affected by the *hg* locus, with no apparent physical malformation to the body composition. The long range goal of such work is to rely on the syntenic relationship between mouse, humans and domestic species to advance analogous research in human studies, as well as economically important livestock traits. As a result of previous work in this area, the search for the *hg* locus (Medrano *et al.* 1992) is restricted to chromosome 10. Localization of QTL in specific regions of a genome is referred to *fine scale* mapping. The measurable trait of interest in this application is weight gain from 14 to 63 days of age.

We first review the quality of the data set, and then present the known estimated genetic map. The analyses are presented in the order that the topics were discussed. Finally, the results of each analysis are compared with the published findings.

Data We summarize the quality of the data by assessing the amount of missing marker and trait information. One individual trait measurement is missing, while complete genetic marker data is available on each of the nine markers. A histogram of weight gain is shown in Figure 2, with the average trait value being 16.2333 (variance 12.0737). There is a slight right hand skew in the trait distribution, having a skewness coefficient of 0.5732 and kurtosis of 3.1694. The quality of this data is exceptionally high. Traditionally (Lincoln *et al.* 1992a), data showing this level of skew would be transformed to normality. However, since the distribution of the trait values within the genotypic marker classes follows a mixture distribution, and the expectation that there is a single QTL, the skewing is anticipated. For the purpose of illustration we will work with untransformed data.

An abbreviated version of the data set is shown in Figure 3. Marker names are listed as rows, and each individual's score for that marker is recorded in the appropriate column. An individual in this F_2 data set may have one of three possible genotypes per marker. An 'A' is homozygous (parent 1), 'B' is homozygous (parent 2), and 'H' is heterozygous. All marker information is recorded and

the measured trait information on each individual follows (Figure 3). It is important to make sure that the order of individuals remains the same across marker and trait data.

Figure 4 displays the genetic map of chromosome 10. Map order and recombination estimates (Haldane mapping function) were estimated using MAPMAKER/EXP.

Single Marker Analysis For each genetic marker in this F_2 experimental population there are three possible genotypes. Using a single factor analysis of variance (ANOVA) on each marker tests the hypothesis of equal trait means in each of the three genotypic classes. Significant results will indicate a difference in the trait means, an indication of QTL action. If normality is assumed, a 5% significance level has a critical value of $F_{2,186} \approx 3.00$. Since multiple tests (one for each marker) will be made, a correction (Lander and Botstein 1989) to the significance level may be appropriate, or one can estimate a critical value by permuting the trait data for the purpose of representing the data under the null hypothesis. Empirical threshold values (Churchill and Doerge 1994) based on 1,000 permutations were estimated for each marker, and for an overall critical value of 5%. Table 3 shows the results of a single factor ANOVA, as well as the F test statistic as calculated by CARTOGRAPHER. CARTOGRAPHER tests that the marker is unlinked to the QTL through a one degree of freedom F–test. Based upon an estimated 5% threshold value of 4.5453 and a maximum test statistic of 24.9495, marker *D10MIT12* displays the highest test statistic. Since no information from the genetic map (*i.e.* marker order) is used, and recombination and QTL effect are confounded in the difference between the genotypic class means, location of the QTL relative to *D10MIT12* can not be determined. *D10MIT12* is simply the one marker that displays the highest level of genotype–phenotype association.

Single Marker Regression We continue with our single factor analysis by using a simple linear regression as specified in (2). Since there is a direct relationship between t–test, F–tests and regression, it is not surprising that the final results are the same. Within the computer program CARTOGRAPHER, the LRmapqtl (Linear Regression) option was employed. For each marker the slope of the regression equation was tested for equality to zero under the null hypothesis. Table 3 gives the results of this analysis. Marker *D10MIT12* displays the highest level of association to a QTL.

Interval Mapping Using the computer program MAPMAKER/QTL, interval mapping as described by Lander and Botstein was employed for locating a single QTL using the known fixed map (Figure 4). Figure 5 shows a typical QTL analysis from MAPMAKER/QTL. CARTOGRAPHER also has a module capable of reproducing MAPMAKER/QTL's effort. For the sake of illustration MAPMAKER/QTL is used for (2 cM increment) interval mapping. The original analysis by Horvat and Medrano (1995) uses incremental values of 0.5 cM. The interval *D10MIT41–D10MIT12* (Figure 5) displays the highest LOD score (10.679) 2 cM to the right of *D10MIT41*. Note that the interval separating these two markers is of length 3.3 cM. Analysis at the marker is equivalent to single factor analysis since no additional information is used from the map. The estimated 5% empirical threshold value to be used across the entire chromosome is 2.0590.

Composite Interval Mapping Composite interval mapping (3) was employed by implementing the model 1 (Zeng 1993) option of the Zmapqtl module of the CARTOGRAPHER computer program. Model 1 tests the current analysis point (increments of 2 cM) in an interval while conditioning on the remaining markers in the genome in order to control for genetic background (Table 4). Both additive and dominance effects were tested using a likelihood ratio test statistic. Since we are performing multiple tests across the entire chromosome, the 5% empirical threshold value (Churchill and Doerge 1994) was estimated (CARTOGRAPHER) based on 1,000 permutations of the original data. The most significant region is within the *D10MIT41–D10MIT12* interval. This result is consistent with previous findings, and is well above the 9.6975 empirical threshold value.

Results This data set illustrates a major single QTL effect. Each method of analysis confirms the published results of Horvat and Medrano (1995), namely that the *hg* locus is approximately in the middle of the *D10MIT41–D10MIT12*. Physical mapping of *hg* is the next step in the long-term goal of cloning *hg* (*i.e.* genetically engineering a replicate of the DNA sequence responsible for the *hg* locus). Cloning will allow functional definition of the *hg* locus, for the purpose of identifying similar loci in human and domestic animal species.

Many experimental situations are not as neat and straightforward as the one used here. Often multiple QTL are detected across the entire genome, in which case the analysis becomes more complicated since the model must reflect the correct genetic situation. Multiple QTL effects are

sometimes independent and their effects may be additive, but often times QTL interact (*epistasis*), and this too must be added to the model. In addition, the sample size is sometimes small, and the proportion of missing data is large (genotypic and phenotypic) making the accuracy of the parameter estimation questionable.

10 Discussion

We have attempted to review a vast amount of literature in a limited space. As a result of this limitation relevant statistical issues have not been discussed fully, yet are worthy of further discussion. The topics not sufficiently covered are, genotype by environment interaction, effects of missing data and sample size, nonlinear model methods of QTL analysis, as well as additional means by which parameter estimation may be accomplished, and issues of statistical power.

When experiments to locate QTLs are conducted in different environments, there is no guarantee that the same results will be found (Paterson *et al.* 1991; Stuber *et al.* 1992). This could be taken as evidence for, or even explanation of, genotype by environment interaction, and so is of biological interest. Caution is needed, however, to ensure that differences in LOD curves, for example, do not simply reflect sampling variation in these curves (Doerge 1993). Genotype by environment (G $\times$ E) interaction has been studied using ANOVA (Paterson *et al.* 1988, Guffy *et al.* 1989, and Zehr 1990), by recording the number of times a marker–QTL association occurs in varying environments (Patterson *et al.* 1991, Stuber 1992, Bubeck *et al.* 1993), as well as by indirect selection where the phenotypic correlation between multiple environments is exploited to study indirect response to selection given no correlation of error effects among environments. G $\times$ E interaction as studied by repeated association produce varying results which may be an artifact of the traits studied or simply because the number of replicates within each environment is too small. There are a number of exhaustive reviews that address G $\times$ E interaction (Freeman 1973, 1990, Fox and Rosielle 1982, Zobel 1990, Bull 1992, Cooper and DeLacy 1994), and even so a large amount of work remains in order for complete understanding. Cooper and DeLacy (1994) put forth two important questions. "The first is, are the aspects of G $\times$ E interaction observed in the multienvironment experiment repeatable? The second is what is the nature of the interaction and how relevant is it to the target population of environments for which the breeding program is responsible?"

Knapp *et al.* (1990) address issues of multiple QTL (unlinked) using linear models similar to those presented in this review, they also consider linked QTL. Using non–linear theory, multiple linked QTL models were developed for backcross, F_2 and F_3 experimental populations.

Several authors have presented heuristic algorithms for determining estimates of QTL distribution parameters and recombination fractions between QTL and trait loci. In the case of one QTL and one marker, Weller (1986) gave the likelihood function for the F_2 design. For specified values of r_{MQ}, he used first and second moments of trait values in each of the three marker classes to provide moment estimators for the means and variances of the P_1, P_2, F_1 trait distributions. Estimates for three parameters, the mean and variance of the F_1 type and the recombination fraction, were then varied over grids in an attempt to maximize the likelihood. Weller (1987) applied this method in a study of some traits in tomato, but did not use information on the marker heterozygotes in the F_2 population. In an even further departure from true maximum likelihood methods, Luo and Kearsey (1989) used the same six moment equations to assign values to the trait distribution parameters as functions of the single unknown r_{MQ}. They substituted these expressions into the likelihood function and then chose r_{MQ} to maximize this expression. Luo and Kearsey (1991) applied the same strategy to other mating designs, including the backcross. Darvasi and Weller (1992) then pointed out that Luo and Kearsey were producing "pseudo" maximum likelihood estimates, and showed numerical differences between such values and values found from a grid search of the full seven-parameter space. Darvasi and Weller (1992) also claimed that the EM-algorithmic approach of Lander and Botstein (1989) did not give true maximum likelihood estimates as it was based on likelihoods calculated at a series of specified r_{MQ} values. The debate has not been characterized by rigorous statistical theory, and now seems to be moot in light of the current regression approaches.

Finally, after methods have been established to detect linkage between trait and marker loci, it is of interest to determine sample size requirements. One of the early discussions was that of Soller *et al.* (1976). For the backcross design with a single trait locus and a single marker, they approximated the t-statistic with a standard normal, and determined the (equal) sample sizes needed in each marker class to have 90% power at a 5% significance level. Their treatment of the F_2 situation was more approximate since they compared only the two homozygous marker class means. They suggest that sensitivity of the F_2 design will be increased by including the

heterozygous markers when $d_{11} > a_1$.

Soller *et al.* (1979) considered how likely it is to find QTLs linked to arbitrary markers under a range of values assigned to marker spacing and genotypic effects at the loci contributing to a trait with specified heritability. Another extension from Soller and his colleagues was a treatment of the case when the two parental lines are segregating at the marker locus (Beckmann and Soller, 1988) rather than being fixed for alternative alleles. Larger sample sizes are needed to attain the same power as in the fixed populations case.

All the work described in this review has been based on crosses of inbred lines. For many species this is not practicable but crosses can be made between outbred lines. Haley *et al.* (1994) use least-squares methods to regress trait phenotypes onto additive and dominance effects of putative QTLs in marker intervals. The work is for the situation of crosses between outbred lines in which the trait loci are segregating but in which the markers used are fixed for alternative alleles.

Although a substantial amount of work has been done somewhat less attention has been paid to issues of statistical power. Carbonell *et al.* (1992) looked at power in analyses involving single marker intervals in F_2 populations and found higher power for F tests than for LOD-score tests, although this came at the expense of higher type I errors. Rebaï *et al.* (1995) compared likelihood methods and analysis of variance for interval mapping in a backcross population. They were able to provide approximate analytical expressions for both critical values and powers, and demonstrated the superiority of likelihood methods. Haley and Knott (1992) compare regression and maximum likelihood and made the point that regression provides a simple alternative to maximum likelihood for single intervals without the computational complexity.

As technology advances and the collective scientific community is able to generate even more molecular based data for the investigation of genetically formed phenomenon, methods of proper QTL analysis must be available. The fields of quantitative genetics and statistics have a long history of excellence, and in this forum (QTL mapping) have the potential to continue as "vital to the welfare of the nation and world" (Bailar 1995). We close with two dynamic examples of QTL research, the first in plant breeding, the second dealing with the synteny between mouse and human. Mutschler *et al.* (1995) present a QTL analysis of the production of acylsugar responsible for pest resistance in wild tomato. The aim of this work is to identify regions in the wild type tomato genome

associated with acylsugar production as related to pest control, and to incorporate these regions into crop species for the purpose of reducing reliance on synthetic pesticides. Horvat and Medrano (1995) demonstrate similar advances in the use of molecular technology and analysis for the location of the *high growth (hg)* locus in mouse (previous example). Molecular characterization of the *hg* locus has potential to direct similar studies in both human and domestic species. The impact of mouse work may be seen in future human diabetes, obesity and heart disease studies. Under growing concern about health and environmental issues associated with the use of environmental/chemical stimuli, quantitative genetics and "molecular" plant (animal) breeding (Rafalski 1993) coupled with proper statistical development has a huge potential for the general improvement of human health issues, as well as economically important food sources.

As a final word, the purpose of this review is to summarize the vast amount of work that has been done in statistical development of methodologies which facilitate the exciting advances in molecular and quantitative genetics as applied to inheritable functions. It is our hope that this review will peak interest in the interdisciplinary field of statistics and genetics by pointing out the statistical nuances of the field, review past and current work, as well as encourage further involvement from the statistical community.

11 Appendix 1:

We develop the specifics of the F_2 generation for single marker analysis considerations. Th F_2 generation is similar to the backcross, except that there are now ten trait-marker genotypes contributing to the genotypic array

$$\frac{(1-r_{MQ})^2}{4}[M_1Q_1/M_1Q_1 + M_2Q_2/M_2Q_2] + \frac{(1-r_{MQ})^2}{2}M_1Q_1/M_2Q_2$$
$$+ \frac{r_{MQ}(1-r_{MQ})}{2}[M_1Q_1/M_1Q_2 + M_2Q_1/M_2Q_2 + M_1Q_1/M_2Q_1 + M_1Q_2/M_2Q_2]$$
$$+ \frac{r_{MQ}^2}{4}[M_1Q_2/M_1Q_2 + M_2Q_1/M_2Q_1] + \frac{r_{MQ}^2}{2}M_2Q_1/M_1Q_2$$

The mixture distributions for the three distinguishable marker classes are

$$M_1M_1: \quad (1-r_{MQ})^2N(\mu_1,\sigma^2) + 2r_{MQ}(1-r_{MQ})N(\mu_{12},\sigma^2) + r_{MQ}^2N(\mu_2,\sigma^2)$$
$$M_1M_2: \quad r_{MQ}(1-r_{MQ})N(\mu_1,\sigma^2) + [r_{MQ}^2 + (1-r_{MQ})^2]N(\mu_{12},\sigma^2) + r_{MQ}(1-r_{MQ})N(\mu_2,\sigma^2)$$
$$M_2M_2: \quad r_{MQ}^2N(\mu_1,\sigma^2) + 2r_{MQ}(1-r_{MQ})N(\mu_{12},\sigma^2) + (1-r_{MQ})^2N(\mu_2,\sigma^2)$$

with means

$$\mu_{M_1M_1} = (1-r_{MQ})^2\mu_1 + 2r_{MQ}(1-r_{MQ})\mu_{12} + r_{MQ}^2\mu_2$$
$$\mu_{M_1M_2} = r_{MQ}(1-r_{MQ})\mu_1 + [r_{MQ}^2 + (1-r_{MQ})^2]\mu_{12} + r_{MQ}(1-r_{MQ})\mu_2$$
$$\mu_{M_2M_2} = r_{MQ}^2\mu_1 + 2r_{MQ}(1-r_{MQ})\mu_{12} + (1-r_{MQ})^2\mu_2$$

and variances

$$\sigma^2_{M_1M_1} = \sigma^2 + 2r_{MQ}(1-r_{MQ})[(\mu_1-\mu_{12}) - r_{MQ}(\mu_1+\mu_2-2\mu_{12})]^2$$
$$+ r_{MQ}^2(1-r_{MQ})^2(\mu_1+\mu_2-2\mu_{12})^2$$
$$\sigma^2_{M_1M_2} = \sigma^2 + r_{MQ}(1-r_{MQ})[(\mu_1-\mu_{12})^2 + (\mu_2-\mu_{12})^2]$$
$$- r_{MQ}^2(1-r_{MQ})^2(\mu_1+\mu_2-2\mu_{12})^2$$
$$\sigma^2_{M_2M_2} = \sigma^2 + 2r_{MQ}(1-r_{MQ})[(\mu_2-\mu_{12}) - r_{MQ}(\mu_1+\mu_2-2\mu_{12})]^2$$
$$+ r_{MQ}^2(1-r_{MQ})^2(\mu_1+\mu_2-2\mu_{12})^2$$

The variances are equal, in general, only for an additive trait and in that case reduce to

$$\sigma^2_{M_1M_1} = \sigma^2_{M_1M_2} = \sigma^2_{M_2M_2} = \sigma^2 + 2r_{MQ}(1-r_{MQ})\delta^2$$

with $\delta^2 = (\mu_1 - \mu_{12})^2 = (\mu_2 - \mu_{12})^2$. Once again, the hypothesis of no linkage between marker and trait loci can be tested by comparing the three marker class means, this time by an analysis of variance. Under this hypothesis, the three marker means and variances will be equal regardless of the degree of dominance. Conversely, equality of all three means implies that the hypothesis is true for all degrees of dominance, providing only that the two parental lines have unequal means. Edwards *et al.* (1987) pointed out that comparisons of the three marker class means allow statements to be made about the relative magnitudes of additive and dominance effects in F_2 populations.

12 Appendix 2:

We derive the maximum likelihood estimates of the various parameters involved in composite interval mapping (Zeng 1993, Jansen 1993). The likelihood equation is defined in (5), β^* is estimated in the following manner

$$\begin{aligned}\frac{\partial \ln L}{\partial \beta^*} &= \sum_{j=1}^{n_1}(Y_{1j} - \beta^* - X_j\beta)/\sigma^2 \\ &+ \sum_{j=1}^{n_2}\frac{(1-p)\phi_1(Y_{2j})(Y_{2j} - \beta^* - X_j\beta)/\sigma^2}{(1-p)\phi_1(Y_{2j}) + p\phi_0(Y_{2j})} \\ &+ \sum_{j=1}^{n_3}\frac{p\phi_1(Y_{3j})(Y_{3j} - \beta^* - X_j\beta)/\sigma^2}{p\phi_1(Y_{3j}) + (1-p)\phi_0(Y_{3j})}\end{aligned}$$

Setting this derivative to zero provides

$$\sum_{l=1}^{4}\sum_{j=1}^{n_l} P_{lj}(Y_{lj} - \beta^* - X_j\beta) = 0$$

where

$$\begin{aligned}P_{1j} &= 1 \\ P_{2j} &= (1-p)\phi_1(Y_{2j})/[(1-p)\phi_1(Y_{2j}) + p\phi_0(Y_{2j})] \\ P_{3j} &= p\phi_1(Y_{3j})/[p\phi_1(Y_{3j}) + (1-p)\phi_0(Y_{3j})] \\ P_{4j} &= 0\end{aligned}$$

This leads to the solution given by Zeng (1994) as

$$\beta^* = \sum_{l=1}^{4}\sum_{j=1}^{n_l} P_{lj}(Y_{lj} - X_j\beta)/\sum_{l=1}^{4}\sum_{j=1}^{n_l} P_{lj}$$

Differentiating the log-likelihood with respect to β:

$$\begin{aligned}\frac{\partial \ln L}{\partial \beta} &= \sum_{j=1}^{n_1} X_j(Y_{1j} - \beta^* - X_j\beta)/\sigma^2 \\ &+ \sum_{j=1}^{n_2}[P_{2j}X_j(Y_{2j} - \beta^* - X_j\beta) + (1-P_{2j})X_j(Y_{2j} - X_j\beta)]/\sigma^2 \\ &+ \sum_{j=1}^{n_3}[P_{3j}X_j(Y_{3j} - \beta^* - X_j\beta) + (1-P_{3j})X_j(Y_{3j} - X_j\beta)]/\sigma^2 \\ &+ \sum_{j=1}^{n_4} X_j(Y_{4j} - X_j\beta)/\sigma^2\end{aligned}$$

The equation $\partial \ln L/\partial\beta = 0$ is most easily expressed in matrix notation as

$$\begin{aligned}\mathbf{X}'(\mathbf{Y} - \mathbf{X}\hat{\beta}) &= \mathbf{X}'\mathbf{P}\beta^* \\ \hat{\beta} &= \mathbf{X}'\mathbf{X}^{-1}\mathbf{X}'(\mathbf{Y} - \mathbf{P}\beta^*)\end{aligned}$$

where $\mathbf{Y}$ is the $n \times 1$ vector of all $n = n_1 + n_2 + n_3 + n_4$ observations, $\mathbf{X}$ is the $n \times (m-1)$ matrix with elements X_{kj}, $\mathbf{P}$ is the $n \times 1$ vector with elements P_{lj} (from P_{11} to P_{4n_4}), and β is the $(m-1) \times 1$ vector with elements $\beta_0, \{\beta_k\}$. The same notation allows the expression

$$\beta^* = (\mathbf{Y} - \mathbf{X}\beta)'\mathbf{P}/c$$

if c represents the sum of all the elements of vector $\mathbf{P}$.

Differentiating the log-likelihood with respect to σ^2:

$$\begin{aligned}\frac{\partial \ln L}{\partial \sigma^2} &= \sum_{j=1}^{n_1}(Y_{1j} - \beta^* - X_j\beta)^2/2\sigma^4 \\ &\quad + \sum_{j=1}^{n_2}[P_{2j}\frac{(Y_{2j} - \beta^* - X_j\beta)^2}{2\sigma^4} + (1 - P_{2j})\frac{(Y_{2j} - X_j\beta)^2}{2\sigma^4}] \\ &\quad + \sum_{j=1}^{n_3}[P_{3j}\frac{(Y_{3j} - \beta^* - X_j\beta)^2}{2\sigma^4} + (1 - P_{3j})\frac{(Y_{3j} - X_j\beta)^2}{2\sigma^4}] \\ &\quad + \sum_{j=1}^{n_4}\frac{(Y_{4j} - X_j\beta)^2}{2\sigma^4} - \frac{n}{2\sigma^2}\end{aligned}$$

Setting this derivative to zero leads to the solution

$$n\hat{\sigma^2} = (\mathbf{Y} - \mathbf{X}\hat{\beta})'(\mathbf{Y} - \mathbf{X}\hat{\beta}) - c(\hat{\beta}^*)^2$$

So far, these solutions have been derived under the assumption that p was known. If it is regarded as being unknown, then the maximum likelihood estimate follows from

$$\begin{aligned}\frac{\partial \ln L}{\partial p} &= \sum_{j=1}^{n_2}\frac{-\phi_1(Y_{2j}) + \phi_0(Y_{2j})}{(1-p)\phi_1(Y_{2j}) + p\phi_0(Y_{2j})} \\ &\quad + \sum_{j=1}^{n_3}\frac{\phi_1(Y_{3j}) - \phi_0(Y_{3j})}{p\phi_1(Y_{3j}) + (1-p)\phi_0(Y_{3j})} \\ &= \sum_{j=1}^{n_2}[-\frac{P_{2j}}{1-p} + \frac{1-P_{2j}}{p}] + \sum_{j=1}^{n_3}[\frac{P_{3j}}{p} - \frac{1-P_{3j}}{1-p}]\end{aligned}$$

so that

$$\hat{p} = \frac{n_2 - \sum_{j=1}^{n_2} \hat{P}_{2j} - \sum_{j=1}^{n_3} \hat{P}_{3j}}{n_2 + n_3}$$

with carets on the P_{lj} values indicating that they are evaluated at the estimated regression and variance values. An iterative procedure is required: estimates of the regression coefficients and σ^2 are found for a specified p value and then this value updated by the last equation and the process is repeated.

13 Acknowledgements

The authors acknowledge S. Horvat and J.F. Medrano for the use of their F_2 data. J.F. Medrano may be contacted at the Department of Animal Sciences, University of California, Davis, CA 95616–8521. E-mail: jfmedrano@ucdavis.edu.

14 References

Abler, B.S.B., M.D. Edwards and C.W. Stuber. 1991. Isoenzymatic identification of quantitative trait loci in crosses of elite maize inbreds. Crop Science. 31:267–274.

Andersson, L., C.S. Haley, H. Ellegen, S.A. Knott, M. Johansson, K. Andersson, L. Andersson-Ekliund, I. Edfors-Lilja, M. Fredholm, I. Hansson, J. Hakansson and K. Lundstrom. 1994. Genetic mapping of quantitative trait loci for growth and fatness in pigs. Science 263:1771–1774.

Bailar, J.C. 1995. A larger perspective. The Amer Stat. Vol 49, No.1, p.10–11.

Barendse, W., S.M. Armitage. L.M. Kossarek, A. Shalom, B.W. Kirkpatrick, A.M. Ryan, D. Clayton, L. Li, H.L. Neibergs, N. Zhang, W.M. Grosse, J. Weiss, P. Creighton, F. McCarthy, M. Ron, A.J. Teale, R. Fries, R.A. McGraw, S.S. Moore, M. Georges, M. Soller, J.E. Womack and D.J.S. Hetzel. 1994. A genetic linkage map of the bovine genome. Nature Genetics 6:227–235.

Basten, C.J., B. S. Weir and Z.-B. Zeng. 1994. Zmap-A QTL Cartographer IN Proceedings of the 5th World Congress on Genetics Applied to Livestock Production. Volume 22: Computing Strategies and Software. Editors: C. Smith, J. S. Gavora, B. Benkel, J. Chesnais, W. Fairfull, J. P. Givson, B. W. Kennedy and E. B. Burnside. University of Guelph, Guelph, Canada. pp. 65-66.

Beckmann, J.S. and M. Soller. 1988. Detection of linkage between marker loci and loci affecting quantitative traits in crosses between segregating populations. Theoretical and Applied Genetics 76:228–236.

Bernatzky, R. and S.D. Tanksley. 1986. Toward a saturated linkage map in tomato based on isozymes and random cDNA sequences. Genetics 112:87–898.

Blumfield, A., S.A. Slaugenhaupt, F.B. Axelrod *et al.*. 1993. Localization of the gene for familial dysautomia on chromosome 9 and definition of DNA markers for genetic diagnosis. Nature

Genetics 4:160–163.

Bovenhuis, H. and J.I. Weller. 1994. Mapping and analysis of dairy cattle quantitative trait loci by maximum likelihood methodology using milk protein genes as genetic markers. Genetics 137:267–280.

Bubeck, D.M., M.M. Goodman, W.D. Beavis, D. Grant. 1993. Quantitative trait loci controlling resistance to grapy leaf spot in maize. Crop Sci. 33:838–847.

Buetow, K.H., A. Chakravarti. 1987. Multipoint gene mapping using seriation. II. Analysis of simulated and empirical data. Am J Hum Genet 41:189-201.

Bull, J.K., M. Cooper, I.H. DeLacy, K.E. Basford, D.R. Woodruff. 1992. Utility of repeated checks for hierarchial classification of data from plant breeding trials. Field Crops Res 30:70–95.

Carbonell, E.A., T.M. Gerig, E. Balansard and M.J. Asins. 1992. Interval mapping in the analysis of nonadditive quantitative trait loci. Biometrics 48:305–315.

Castle, W.E. 1921. An improved method of estimating the number of genetic factors concerned in cases of blending inheritance. Science 54:541–553.

Churchill, G.A. and R.W. Doerge. 1994. Empirical threshold values for quantitative trait mapping. Genetics 138:963–971.

Cockerham, C.C. 1986. Modifications in estimating the number of genes for a quantitative character. Genetics 114:659–664.

Coe, E.H., D.A. Hoisington and M.G. Nuffer. 1990. Linkage map of corn (maize) (*Zea mays* L.) (2n=20). In: O'Brien, S.J. (ed.) *Genetic maps*, 5th edn. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, p. 6.39–6.67.

Comstock, R.E. and F.D. Enfield. 1981. Gene number estimation when multiplicative genetic effects are assumed –growth in flour beetles and mice. Theor Appl Genet 59:373–379.

Cooper, M. and I.H. DeLacy. 1994. Relationships among analytical methods used to study genotypic variation and genotype–by–environmental interaction in plant breeding multi–environment

experiments. Theor Appl Genet 88:561–572.

Copeland, N.G., N.A. Jenkins, D.J. Gilbert, J.T. Eppig, L.J. Maltais, J.C. Miller, W.F. Dietrich, S.E. Lincoln, R.G. Steen, L.D. Stein, J.H. Nadeau and E.S. Lander. 1993. A genetic linkage map of the mouse: current applications and future prospects. Science 262:57–66.

Corana, A., M. Marchesi, C. Martini, and S. Ridella. 1987. Minimizing multimodal functions of continuous variables with the "simulated annealing" algorithm. ACM Trans on Math Software 13:2:262-80.

Darvasi, A. and J.I. Weller. 1992. On the use of the moments method of estimation to obtain approximate maximum likelihood estimates of linkage between a genetic marker and a quantitative locus. Heredity 68:43–46.

Darvasi, A., A. Weinreb, V. Minke, J.I. Weller, M. Soller. 1993. Detecting marker–qtl linkage and estimating QTL gene effect and map location using a saturated map. Genetics 134:943–951.

Dempster, A.P., N.M Laird, D.B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc. 39:1–38.

Doerge, R.W. 1993. *Statistical Methods for Locating Quantitative Trait Loci with Molecular markers.* Ph.D. Dissertation, Dept. Statistics, N.C. State University, Raleigh, NC.

Dupuis, J. 1994. *Statistical problems associated with mapping complex and quantitative traits from genomic mismatch scanning data.* Ph.D. Thesis of Stanford university, Department of Statistics, USA.

Edwards, M.D., C.W. Stuber and J.F. Wendel. 1987. Molecular-marker-facilitated investigations of quantitative-trait loci in maize. I. Numbers, genomic distribution and types of gene action. Genetics 116:113–125.

Eisenberger, I. 1964. Genesis of bimodal distributions. Technometrics 6:357–363.

Elston. R.C. and J. Stewart. 1973. The analysis of quantitative traits for simple genetic models from parental, F_1 and backcross data. Genetics 73:695-711.

Falk, C.T. 1992. Preliminary ordering of multiple linked loci using pairwise linkage data. Genet Epidemiol 9:367-375.

Feingold E., Brown P.O., Siegmund D. 1993. Gaussian models for genetic linkage analysis using complete high-resolution maps of identity by descent. *Am. J. Hum. Genet.*, **53,** 234-251.

Feng, Z. 1990. Statistical Inference Using Maximum Likelihood Estimation and the Generalized Likelihood Ratio Under Nonstandard Conditions. Cornell University. Ph.D. Dissertation.

Fisher, R.A. 1935. *The Design of Experiments*, Ed. 3, Oliver and Boyd Ltd., London.

Fox, P.N. and A.A. Rosielle. 1982. Reducing the influence of environmental main–effects on pattern analysis of plant breeding environments. Euphytica 31:645–656.

Freeman, G.H. 1973. Statistical methods for the analysis of genotype–environment interactions. Heredity 31:339–354.

Freeman, G.H. 1990. Modern statistical methods for analyzing genotype–by–environment interactions. In: Kang MS (ed) Genotype–by–environment interaction and plant breeding. Louisiana State University. Baton Rouge, Louisiana, pp. 118–125.

Georges, M., R. Drinkwater, T. King, *et al.*. 1993. Microsatellite mapping of a gene affecting horn development in *Bos taurus*. Nature Genetics 4:206–210.

Ghosh, J.K. and P.K. Sen. 1985. On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results. Proc. of the Berkeley Conf., Vol.II. Eds. Lucien M. LeCam and Richard A. Olshen. pp.789–807.

Good, P. 1994. Permutation Tests: A Practical Guide to Resampling for Testing Hypotheses. Springer-Verlag, New York.

Graner, A., A. Jahoor, J. Schondelmaier, H. Seidler, K. Pillen, G. Fishbeck, G. Wenzel and R.G. Herrmann. 1991. Construction of an RFLP map of barley. Theoretical and Applied Genetics 83:250–256.

Guffy, R.D., C.W. Stuber, and M.D. Edwards. 1989. Dissecting and enhancing heterosis in corn using molecular markers. pp. 99-120. In: Proc. 25th Illinois Corn Breeders School, Champaign, IL. 6–7 March 1989. Champaign, IL.

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

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

Haley, C.S., S.A. Knott and J-M. Elsen. 1994. Mapping quantitative trait loci in crosses between outbred lines using least squares. Genetics 136:1195–1207.

Hartigan, J.A. 1985. A failure of likelihood asymptotics for normal distributions. Proc. of the Berkeley Conf., Vol. II. Eds. Lucien M. LeCam and Richard A. Olshen. pp.807–810.

Hilbert, P., K. Lindpaintner, J.S. Beckmann, *et al.*. 1991. Chromosomal mapping of two genetic loci associated with blood-pressure regulation in hereditary hypertensive rats. Nature 353:521–529.

Horvat, S. and J.F. Medrano. 1995. Interval mapping of *high growth (hg)*, a major locus that increases weight gain in mice. Genetics 139:1737–1748.

Jacob, H.J., K. Lindpaintner, S.E. Lincoln, *et al.*. 1991. Genetic mapping of a gene causing hypertension in the stroke-prone spontaneously hypertensive rat. Cell 67:213–224.

Jansen, R.C. 1992. A general mixture model for mapping quantitative trait loci by using molecular markers. Theoretical and Applied Genetics 85:252–260.

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

Jansen, R.C. 1994 Controlling the type I and type II errors in mapping quantitative trait loci. Genetics 138:871-881.

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

Jeunemaitre, X., R.P. Lifton, S.C. Hunt, R.R. Williams and J-M. Lalouel. 1992. Absence of linkage between the angiotensin converting enzyme and human essential hypertension. Nature Genetics 1:72–75.

Keim, P., B.W. Diers, T.C. Olson and R.C. Shoemaker. 1990. RFLP mapping in soybean: association between marker loci and variation in quantitative traits. Genetics 126:735–742.

Kearsey, M.J. and V. Hyne. 1994. QTL analysis: a simple 'marker–regression' approach. Theor Appl Genet 89:698–702.

Kerem B-S, J.M. Rommens, J.A. Buchanan, D. Markiewicz, T.K. Cox, A. Chakravarti, M. Buchwald, L-C. Tsui. 1989. Identification of the cystic fibrosis gene: genetic analysis. Science 245:1073–1080

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

Knott, S.A. and C.S. Haley. 1992. Aspects of maximum likelihood methods for the mapping of quantitative trait loci in line crosses. Genetical Research 60:139–1521.

Kosambi, D.D. 1944. The estimation of map values from recombination values. Ann Eugen 12:172–175.

Kruglyak, L. and E.S. Lander. 1995. A nonparametric approach for mapping quantitative trait loci. Genetics 139:1421-1428.

Lande, R. 1981. The minimum number of genes contributing to quantitative variation between and within populations. Genetics 99:541–553.

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

Lander, E.S. and D. Botstein. 1994. Corrigendum. Genetics 36: 705.

Lander, E.S. and P. Green. 1987. Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci 84:2363-2367.

Lathrop, G., J. Lalouel, C. Julier, J. Ott. 1985. Multilocus linkage analysis in humans: detection of linkage and estimation of recombination. Am J Hum Genet 37:482-498.

Lincoln, S., M. Daly, E. Lander. 1992. Constructing Genetic Maps with MAPMAKER/EXP 3.0. Whitehead Institute Technical Report. 3rd edition.

Lincoln, S. M. Daly, E. Lander. 1992. Mapping Genes Controlling Quantitative Traits with MAPMAKER/QTL 1.1. Whitehead Institute Technical Report. 2nd edition.

Luo, Z.W. and M.J. Kearsey. 1989. Maximum likelihood estimation of linkage between a marker gene and a quantitative trait locus. Heredity 63:401–408.

Luo, Z.W. and M.J. Kearsey. 1991. 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.

Luo, Z.W. and M.J. Kearsey. 1992. Interval mapping of quantitative trait loci in an F_2 population. Heredity 69:236–242.

Luo, Z.W. and J.A. Woolliams. 1993. Estimation of genetic parameters using linkage between a marker gene and a locus underlying a quantitative character in F_2 populations. Heredity 70:245–253.

Mangin, B., B. Goffinet and A. Rebaï. 1994. Constructing confidence intervals for QTL location. Genetics 138:1301–1308.

Martinez, O. and R.N. Curnow. 1992. Estimating the locations and the size of the effects of quantitative trait loci using flanking markers. Theoretical and Applied Genetics 85:480–488.

McMillan, I. and A. Robertson. 1974. The power of methods for the detection of major genes affecting quantitative characters. Heredity 32:349–356.

McPeek, M.S. and T.P. Speed. 1995. Modeling interference in genetic recombination. Genetics 139:1031–1044.

Medrano, J.F., D. Pomp, B.A. Taylor and G.E. Bradford. 1992. The *high growth* gene *(hg)* in mice is located on chromosome 10 linked to *Igf1*. Advances in gene technology: Feeding the world in teh 21st century, edited by W. J. Whelan *et al.* The 1992 Miami Bio/Technology Winter Symposium. 1:12.

Moreno-Gonzalez, J. 1992a. Estimates of marker-associated QTL effects in Monte Carlo backcross generations using multiple regression. Theoretical and Applied Genetics 85:423–434.

Moreno-Gonzalez, J. 1992b. Genetic models to estimate additive and non-additive effects of marker-associated QTL using multiple regression techniques. Theoretical and Applied Genetics 85:435–444.

Morton, N.E. 1995. LODs past and present. Genetics 140:7–12.

Mutschler, M.A., R.W. Doerge, S-C Liu, J.P. Kuai, B.E. Liedl, J.A. Shapiro. 1995. QTL analysis of the production of acylsugars responsible for pest resistance in the wild tomato *Lycopersicon pennellii.* Poster. The Third International Conference on the Plant Genome. San Diego, CA.

Nienhuis, J., T. Helentjaris, M. Slocum, B. Ruggero and A. Schaefer. 1987. Restriction fragment length polymorphism analysis of loci associated with insect resistance in tomato. Crop Science 27:797–803.

NIH/CEPH Collaborative Mapping Group. 1992. A comprehensive genetic linkage map of the human genome. Science 258:148–162.

O'Brien, S.J.(Editor) 1993. Genetic Maps (6th Edition), Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Ott, J. 1991. Analysis of Human Genetic Linkage. The Johns Hopkins University Press.

Paterson, A.H., E.S. Lander, J.D. Hewitt, S. Peterson, S.E. Lincoln and S.D. Tanksley. 1988. Resolution of quantitative traits into Mendelian factors by using a complete map of restriction fragment length polymorphisms. Nature 335:721–726.

Paterson, A.H., S. Damon, J.D. Hewitt, D. Zamir, H.D. Rabinowitch, S.E. Lincoln, E.S. Lander and S.D. Tanksley. 1991. Mendelian factors underlying quantitative traits in tomato: comparison across species, generations, and environments. Genetics 127:181–197.

Paterson, A.H., J.W. Deverna, B. Lanini and S.D. Tanksley. 1990. Fine mapping of quantitative trait loci using selected overlapping recombinant chromosomes, in an interspecies cross of tomato. Genetics 124:735–742.

Plomin, R., G.E. McClearn and G. Gora-Maslak. 1991. Quantitative trait loci and psychopharmacology. J. Psychopharmacology 5:1–9.

Rafalski, J.A. and S.V. Tingey. 1993. Genetic diagnostics in plant breeding: RAPDs, microsatellites and machines. Trends in Genetics Vol. 9, No. 8, pp. 275–280.

Rebaï, A., B. Goffinet and B. Mangin. 1995. Comparing power of different methods of QTL detection. Biometrics 51:87-99.

Rebaï, A., B. Goffinet, B. Mangin. 1994. Approximate thresholds of interval mapping tests for QTL detection. Genetics 138:235-240.

Reiter, R.S., J.G. Cors, M.R. Sussman and W.H. Gabelman. 1991. Genetic analysis of tolerance to low-phosphorus stress in maize using restriction fragment length polymorphisms. Theoretical and Applied Genetics 82:561–568.

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

Self, S.G. and K-Y. Liang. 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. Journal of the American Statistical Association 82:605–610.

Slocum, M.K., S.S. Figdore, W.C. Kennard, J.Y. Suzuki and T.C. Osborne. 1990. Linkage arrangement of restriction fragment length polymorphism loci in *Brassica oleracea.* Theoretical and Applied Genetics 80:57–64.

Speed, T.J., M.S. McPeek and S.N. Evans. 1992. Robustness of the no–interference model for ordering genetic models. Proc. Natl. Acad. Sci. 89:3103-3106.

Soller, M, T. Brody and A. Genizi. 1976. On the power of experimental designs for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines. Theoretical Applied Genetics 47:35–39.

Soller, M, T. Brody and A. Genizi. 1979. The expected distribution of marker-linked quantitative effects in crosses between inbred lines. Heredity 43:179–190.

Stuart, A., J.K. Ord. 1991. *Kendall's Advanced Theory of Statistics.* Oxford Univ. Press, New York.

Stuber, C.W., M.D. Edwards, M.D. and J.F. Wendel. 1987. Molecular-marker-facilitated investigations of quantitative-trait loci in maize. II. Factors influencing yield and its component traits. Crop Science. 27:639–648.

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 inbred lines using molecular markers. Genetics 132:823–839.

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

Thompson, E.A. 1984. Information gain in joint linkage analysis. IMA J Math Appl Med Biol 1:31-49.

Weeks, D. and K. Lange. 1987. Preliminary ranking procedures for multilocus ordering. Genomics 1:236-242.

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.

Weller, J.I. 1987. Mapping and analysis of quantitative trait loci in *Lycopersicon* (tomato) with the aid of genetic markers using approximate maximum likelihood methods. Heredity 59:413–421.

Wright, S. 1952. The genetics of quantitative variability. In: Reeve ECR, Waddington CH (eds) Quantitative inheritance. Her Majesty's Stationary Office, London, pp.5–41.

Wright, A.J. and R.P. Mowers. 1994. Multiple regression for molecular–marker, quantitative trait data from large F_2 populations. Theor Appl Genet 89:305–312.

Wu, W.R. and W.M. Li. 1994. A new approach for mapping quantitative trait loci using complete genetic marker linkage maps. Theor Appl Genet 89:535–539.

Zehr, B.E. 1990. Use of RFLP markers in maize as an aid in selection during inbreeding. Ph.D. diss. Univ. of Illinois.

Zeng, Z-B. 1993. Theoretical basis of precision mapping of quantitative trait loci. Proceedings of the National Academy of Science USA 90:10972–10976.

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

Zeng, Z-B., D. Houle and C.C. Cockerham. 1990. How informative is Wright's estimator of the number of genes affecting a quantitative character? Genetics 126:235–247.

Zhao, H., M.S. McPeek, T.P. Speed. 1995. Statistical analysis of chromatid interference. Genetics 139:1057–1065.

Zhao, H., T.P. Speed, M.S. McPeek. 1995. Statistical analysis of crossover interference using the chi–square model. Genetics 139:1045–1056.

Zobel, R.W. 1990. A powerful statistical model for understanding genotype–by–environment interactions. In: Kang MS (ed) Genotype–by–environment interaction and plant breeding. Louisiana State University. Baton Rouge, Louisiana, pp. 126–140.

Table 1: Genotypic frequencies of trait and two marker loci in backcross population. Markers are denoted **M** and **N**, each with two alleles. The QTL is denoted **Q** with alleles $\mathbf{Q}_1$ and $\mathbf{Q}_2$. Recombination between loci i and j denoted r_{ij}, whether QTL or marker.

Marker Class	Genotype[a]	**QMN**	**MQN**[b]	**MNQ**
1	$M_1Q_1N_1$ / $M_1Q_1N_1$	$(1-r_{MQ})(1-r_{MN})$	$(1-r_{MQ})(1-r_{NQ})$	$(1-r_{MN})(1-r_{NQ})$
	$M_1Q_1N_1$ / $M_1Q_2N_1$	$r_{MQ}(1-r_{MN})$	$r_{MQ}r_{NQ}$	$(1-r_{MN})r_{NQ}$
2	$M_1Q_1N_1$ / $M_1Q_1N_2$	$(1-r_{MQ})r_{MN}$	$(1-r_{MQ})r_{NQ}$	$r_{MN}r_{NQ}$
	$M_1Q_1N_1$ / $M_1Q_2N_2$	$r_{MQ}r_{MN}$	$r_{MQ}(1-r_{NQ})$	$r_{MN}(1-r_{NQ})$
3	$M_1Q_1N_1$ / $M_2Q_1N_1$	$r_{MQ}r_{MN}$	$r_{MQ}(1-r_{NQ})$	$r_{MN}(1-r_{NQ})$
	$M_1Q_1N_1$ / $M_2Q_2N_1$	$(1-r_{MQ})r_{MN}$	$(1-r_{MQ})r_{NQ}$	$r_{MN}r_{NQ}$
4	$M_1Q_1N_1$ / $M_2Q_1N_2$	$r_{MQ}(1-r_{MN})$	$r_{MQ}r_{NQ}$	$(1-r_{MN})r_{NQ}$
	$M_1Q_1N_1$ / $M_2Q_2N_2$	$(1-r_{MQ})(1-r_{MN})$	$(1-r_{MQ})(1-r_{NQ})$	$(1-r_{MN})(1-r_{NQ})$

[a]chromosome 1 is the top genotype, chromosome 2 is the bottom genotype

[b]Twice the frequency if QTL is in the interval.

Table 2: Marker classes and trait probabilities in backcross B_1 population (ignoring double crossovers between markers) for marker **M** with alleles $\mathbf{M}_1$ and $\mathbf{M}_2$, and marker **N** with alleles $\mathbf{N}_1$ and $\mathbf{N}_2$. The chromosomes are separated by '/', and r_{ij} denotes recombination fractions between loci i and j (marker or QTL).

Marker Class	Frequency	$\Pr(Q_1Q_1) = \Pr(X^* = 1)$	$\mathcal{E}(Y)$
M_1N_1/M_1N_1	$\frac{1}{2}(1 - r_{MN})$	$\frac{(1-r_{MQ})(1-r_{NQ})}{(1-r_{MN})} \approx 1$	μ_1
M_1N_1/M_1N_2	$\frac{1}{2}r_{MN}$	$\frac{(1-r_{MQ})r_{NQ}}{r_{MN}} \approx 1 - \frac{r_{MQ}}{r_{MN}} = 1 - p$	$(1-p)\mu_1 + p\mu_{12}$
M_1N_1/M_2N_1	$\frac{1}{2}r_{MN}$	$\frac{r_{MQ}(1-r_{NQ})}{r_{MN}} \approx \frac{r_{MQ}}{r_{MN}} = p$	$p\mu_1 + (1-p)\mu_{12}$
M_1N_1/M_2N_2	$\frac{1}{2}(1 - r_{MN})$	$\frac{r_{MQ}r_{NQ}}{(1-r_{MN})} \approx 0$	μ_{12}

Table 3: Single marker analysis of Horvat and Medrano (1995) data set. 190 F_2 individuals scored for 9 genetic markers on chromosome 10 of the male mouse genome. Regression and F^* calculations are from CARTOGRAPHER. F° and critical values calculated using Fortran program (R.W. Doerge). 5% empirical threshold values calculated using 1,000 permutations of the original data. The 5% experimental empirical threshold value (for entire chromosome) using the F° test statistic is 4.522.

Marker	β_0 [a]	β_1 [b]	LR [c]	F^* [d]	F° [e]	Critical Value [f]
D10MIT31	14.820	1.291	12.198	12.466	7.390	3.060
D10MIT42	13.855	2.112	31.315	33.685	18.110	3.265
IGF1	13.827	2.166	32.993	35.651	18.058	3.235
D10MIT9	13.912	2.120	31.330	33.703	17.153	3.244
D10MIT10	13.870	2.169	33.473	36.218	18.683	3.201
D10MIT41	13.730	2.320	41.259	45.496	24.348	3.242
D10MIT12	13.674	2.349	42.207	46.765	24.950	3.077
D10NDS2	13.935	2.110	32.396	34.950	19.177	3.055
D10MIT14	14.654	1.422	15.691	16.185	9.563	2.976

[a] intercept of simple linear regression

[b] slope of simple linear regression

[c] Likelihood Ratio $-2\log(L_0/L_1)$

[d] F statistic for testing that the marker is unlinked to the QTL

[e] F-statistic for testing that there is no difference between the three genotypic class means

[f] Empirical threshold values (5%) for F°

Table 4: Composite mapping results for Horvat and Medrano (1995) data using CARTOGRAPHER. 190 F_2 individuals scored for 9 genetic markers on chromosome 10 of the male mouse genome. See composite interval section of paper for model specification, all markers are used to control for genetic background. Interval mapping is performed in approximate increments of 2 cM using a likelihood ratio test statistic and the hypotheses: $H_0 : a = 0$, $d = 0$, $H_1 : a \neq 0$, $d = 0$, $H_3 : a = 0$, $d \neq 0$. 5% empirical threshold values calculated using 1,000 permutations of the original data. The 5% experimental empirical threshold value (for entire chromosome) is 9.680 as calculated by CARTOGRAPHER.

Marker	Test Position [a]	$H_0 : H_3$ [b]	$H_1 : H_3$	$H_2 : H_3$
D10MIT31	0.0001	14.431	3.377	10.437
	0.0201	18.984	4.220	14.508
	0.0401	23.443	4.482	18.266
	0.0601	27.612	4.398	21.644
	0.0801	31.514	3.985	24.768
D10MIT42	0.0906	33.291	3.368	26.209
	0.1106	33.151	2.788	28.413
	0.1306	33.063	0.682	29.621
IGF1	0.1327	33.044	0.467	29.661
D10MIT9	0.1462	31.634	0.801	28.599
D10MIT10	0.1542	34.364	1.645	30.656
D10MIT41	0.1703	43.767	4.632	38.273
	0.1903	47.568	9.749	42.045
D10MIT12	0.2036	44.557	4.459	38.498
	0.2236	35.728	4.338	29.797
D10NDS2	0.2254	35.202	4.930	28.926
	0.2454	31.449	5.120	25.147
	0.2654	27.534	4.939	21.303
	0.2854	23.506	4.754	17.676
	0.3054	19.175	4.285	13.942
D10MIT14				

[a] over total length of chromosome

[b] likelihood ratio

Figure 5: MAPMAKER/QTL interval mapping computer output of Horvat and Medrano (1995) data. 190 F_2 individuals scored for 9 genetic markers on chromosome 10 of the male mouse genome. Haldane map function used to convert from recombination fraction to map distance (cM).

POS [a]	WEIGHT [b]	DOM[c]	%VAR[d]	LOG-LIKE [e]	Significance [f]
					D10MIT31-D10MIT42 [g] 9.1 cM [h]
0.0	-1.202	-0.743	7.4%	3.140	**********
2.0	-1.475	-0.907	11.0%	4.243	**********
4.0	-1.683	-0.912	13.4%	5.260	**********
6.0	-1.838	-0.845	15.0%	6.159	**********
8.0	-1.943	-0.769	16.0%	6.940	**********
					D10MIT42-IGF1 4.2 cM
0.0	-1.977	-0.733	16.3%	7.303	**********
2.0	-2.106	-0.571	17.3%	7.527	**********
4.0	-2.116	-0.383	16.5%	7.332	**********
					IGF1-D10MIT9 1.3 cM
0.0	-2.108	-0.358	16.3%	7.282	**********
					D10MIT9-D10MIT10 0.8 cM
0.0	-2.065	-0.395	15.6%	6.953	**********
					D10MIT10-D10MIT41 1.6 cM
0.0	-2.104	-0.509	16.8%	7.524	**********
					D10MIT41-D10MIT12 3.3 cM
0.0	-2.241	-0.757	20.8%	9.552	**********
2.0	-2.392	-0.998	24.6%	10.679	**********
					D10MIT12-D10NDS2 2.2 cM
0.0	-2.257	-0.757	21.2%	9.752	**********
2.0	-2.063	-0.815	18.0%	7.966	**********
					D10NDS2-D10MIT14 8.3 cM
0.0	-2.005	-0.824	17.1%	7.693	**********
2.0	-1.971	-0.894	17.2%	7.163	**********
4.0	-1.851	-0.943	16.0%	6.364	**********
6.0	-1.649	-0.956	13.6%	5.359	**********
8.0	-1.372	-0.854	10.0%	4.220	**********

[a] test position

[b] estimated additive effect

[c] estimated dominance effect

[d] estimated percent total variance explained by QTL

[e] LOD score

[f] one star is printed at a LOD score over 2.0, 0.25 increments are denoted with additional stars

[g] map interval

[h] map distance between markers which define interval

Figure 1: Standard backcross and F_2 mating designs for marker **M** with alleles $\mathbf{M}_1$ and $\mathbf{M}_2$ and QTL **Q** with alleles $\mathbf{Q}_1$ and $\mathbf{Q}_2$. The chromosomes are separated by '/', and the assumption of normality on the traits values, given the known genotype of the QTL, is imposed and denoted by $N(\mu, \sigma^2)$.

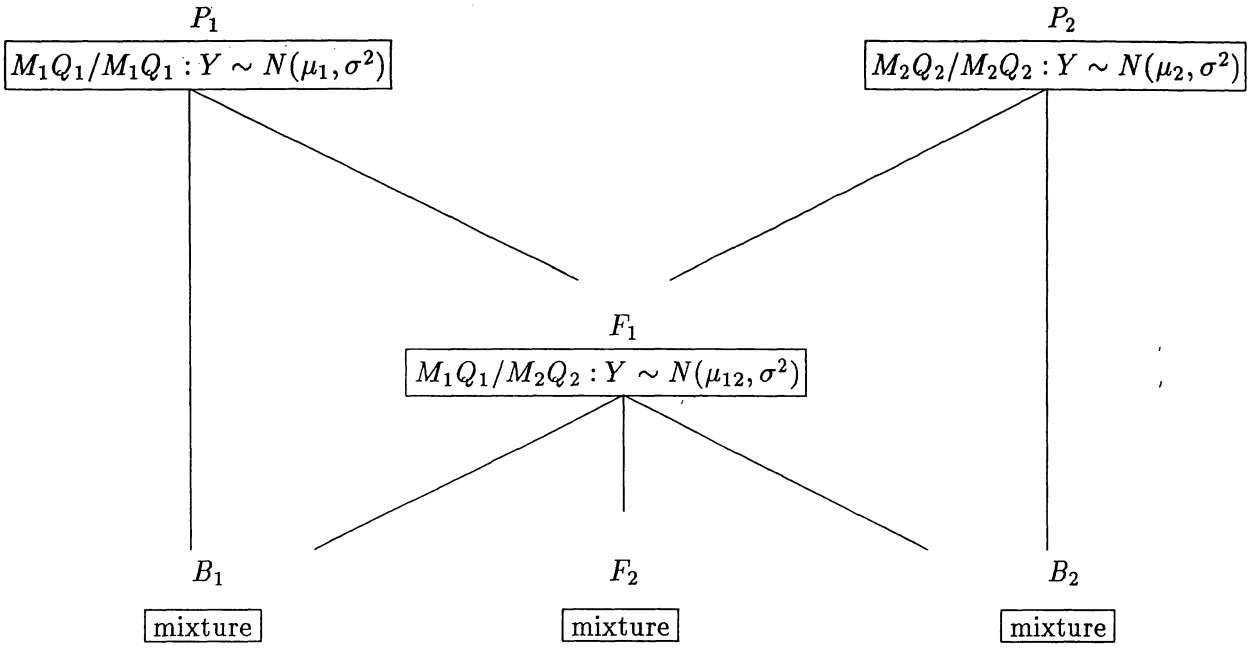

Figure 2: Histogram of weight gain from 14 to 63 days of age for the F_2 mouse data set (Horvat and Medrano 1995) containing 190 male individuals, 9 genetic markers with average density 3.85 cM on chromosome 10 of the male mouse genome.

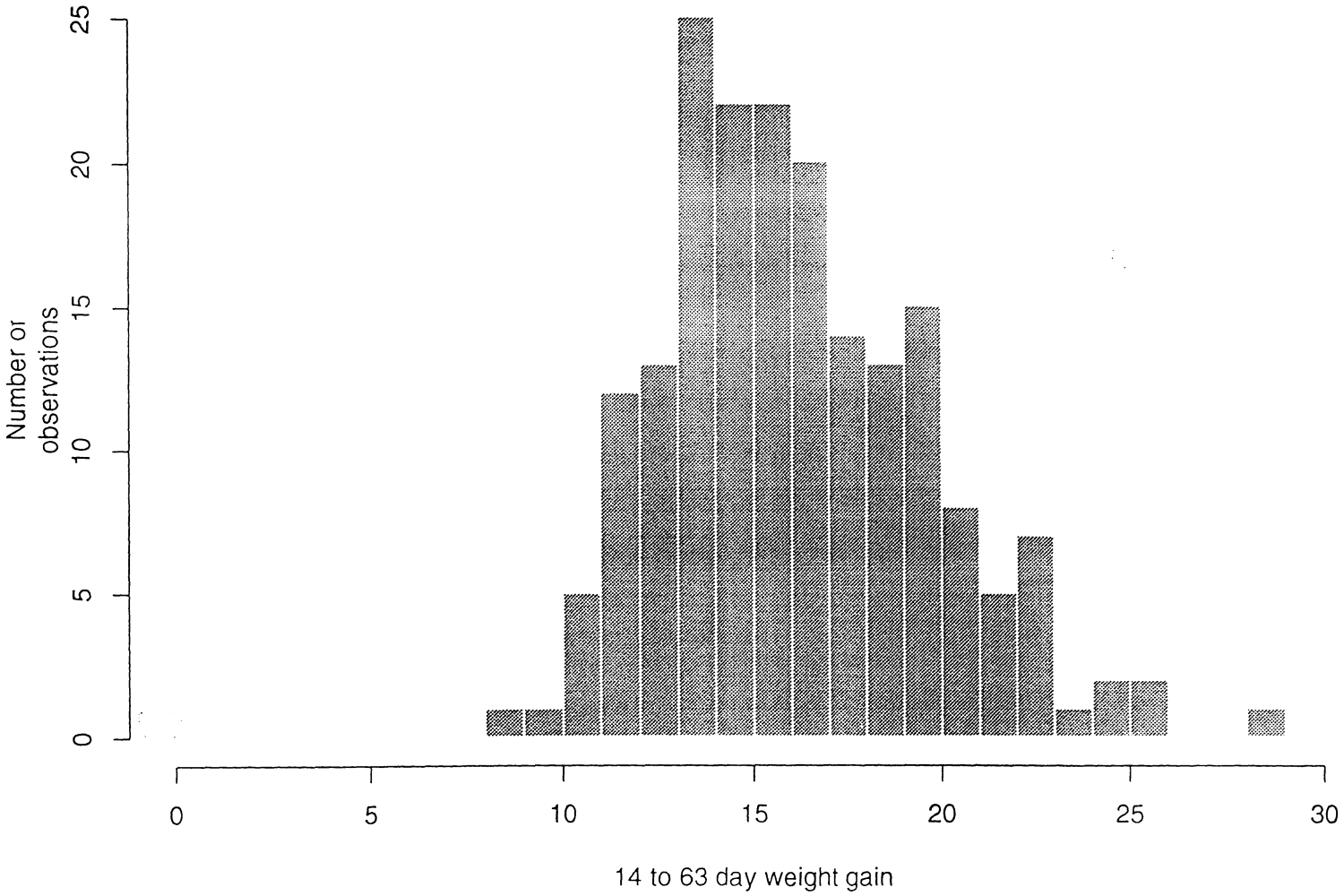

Figure 3: An example of the abbreviated genotypic and phenotypic data from Horvat and Medrano (1995). 190 F_2 individuals scored for 9 genetic markers on chromosome 10 of the male mouse genome. Markers names are in map order at the beginning of each row. Genetic markers are scored for each individual (columns). Homozygous genotypes of the first parental type are denoted A, homozygous genotypes of the second type are denoted B, and heterozygotes are H. The measured trait data is weight gain from 14 to 63 days of age, and the order of the individuals is the same for both genotypic and phenotypic data.

D10MIT31	H H H H B A H H B H $\cdots$ H H H H A H A A B B
D10MIT42	H B H H B H H H B H $\cdots$ H H B H A H A A H B
IGF1	H B H H B H H H B H $\cdots$ H H B B A H A A H B
D10MIT9	H B H H B H H H B H $\cdots$ A H B B A H A A H B
D10MIT10	H B H H B H H H B H $\cdots$ A H B B A H A A H B
D10MIT41	H B H H B H H H B H $\cdots$ A H B B A H A A H B
D10MIT12	H B H H B H H H B H $\cdots$ A H B B A H A A H B
D10NDS2	H B H H B H H H B H $\cdots$ A H B B A H A A H B
D10MIT14	H B A A B H H H B H $\cdots$ A H B B A H A A H B
weight	12.1 15.6 14.0 14.6 13.5 13.2 17.3 13.0 16.0 11.6 18.4
	$\cdots$ 17.8 14.6 12.0 10.3 11.2 16.0 19.2 20.8 13.3 11.8

Figure 4: MAPMAKER/EXP estimated genetic map of Horvat and Medrano (1995) data. 190 F_2 individuals scored for 9 genetic markers on chromosome 10 of the male mouse genome. Haldane map function used to convert from recombination fraction to map distance (cM).