Université catholique de Louvain
Faculté d'ingénierie biologique, agronomique et environnementale

Methodological aspects of the mapping of disease resistance loci in livestock

Pierre Tilquin

September 2003

Thèse présentée en vue de l'obtention
du grade de docteur en sciences agronomiques
et ingénierie biologique

Promoteurs : P. Baret et E. Le Boulengé

Part I

Context

Chapter 2

QTL mapping methods

In this chapter, the basic principle used for the mapping of genes underlying traits showing continuous variation is given [2.2]. A thorough explanation is first given in the single-marker context [2.2.3]; then, the interval mapping approach using multiple markers is outlined [2.2.4]. The different methods for the determination of the significance thresholds and confidence intervals are presented [2.2.5].

In the second part of this chapter, the possible extensions of the method are given regarding for example the type of animal populations, the use of multiple traits, and the analysis of non-normally distributed traits [2.3].

2.1. Introduction

Most of the variation observed within populations, or between lines or breeds is quantitative in nature. Examples of quantitative traits include all classical production traits such as liveweight, growth rate or milk production, but also disease resistance variables such as antibody levels or number of somatic cells in the milk.

Quantitative traits show a continuous distribution of phenotypic values rather than the discrete distribution observed for a qualitative trait. Therefore, under the term quantitative trait, traits such as morphological traits (e.g. coat colour) or inherited disorders (e.g. bovine leukocyte adhesion deficiency, BLAD) will not be included.

The classical quantitative genetics theory assumes that quantitative traits are controlled by an infinite number of genes, each with an infinitesimal effect, and are influenced by environmental factors. However, beyond this model, it is well known that there are loci that can have a major effect on quantitative traits (see e.g. Falconer & Mackay, 1996). Such genes are called quantitative trait loci (QTL). A QUANTITATIVE TRAIT LOCUS is defined as a region of the genome (i.e. segment of chromosome) that harbours one or more genes affecting a quantitative trait (Geldermann, 1975). The QTL concept is also used for traits with discrete distribution (such as sick *versus* healthy) for which a continuous underlying process can be assumed. These traits are often referred to as THRESHOLDS TRAITS and require specific methods.

The presence of a QTL is inferred from significant differences in terms of phenotypic value between individuals that have inherited different QTL ALLELES from their parents. Since QTL alleles are not identifiable, because their position and determinism are unknown, the information coming from molecular markers is used instead.

A MOLECULAR MARKER is a locus with a known position on the genome and that can take various forms (alleles) identifiable at the molecular level.

The QTL MAPPING strategy is defined as the methodology aimed at testing for – and positioning on chromosomes – putative QTL involved in the

determination of quantitative traits, owing to their linkage with molecular markers.

The hypothesis underlying this methodology is that, if several markers are scored (i.e. genotyped) throughout the genome, some at least will be linked to a QTL which has an influence on the trait of interest. If such linkage occurs, differences in marker genotypes will be associated with differences in phenotypic values for the trait controlled by the QTL.

There are two main strategies for finding QTL:

(a) association tests using single markers;

(b) genome scans based on linkage mapping with multiple markers that are used simultaneously.

Both types of strategies make use of two types of information (Figure 2.1): (1) phenotypic measurements for the quantitative trait(s) of interest; (2) genotypes at marker loci.

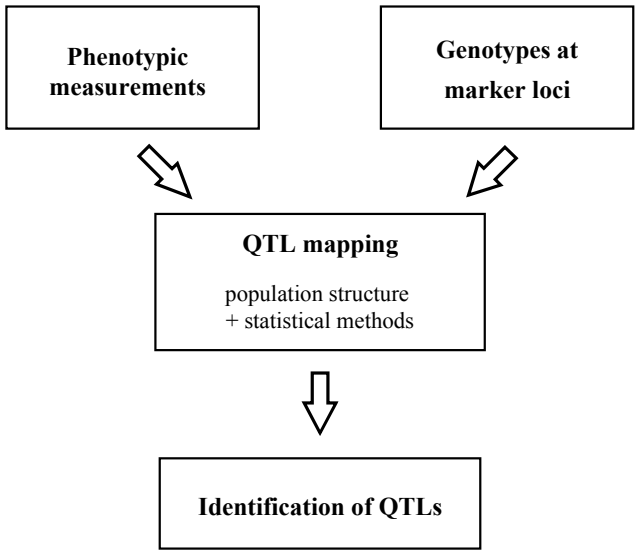

Figure 2.1. Schematic representation of the QTL mapping methodology

By taking into account the structure of the population, and by combining both types of information into a statistical analysis, the QTL mapping strategy enables an eventual identification of chromosome regions underlying the phenotypic variation observed between animals (Figure 2.1). Furthermore, the quantification of the QTL effect in terms of proportion of phenotypic variance explained by the QTL is also possible.

During the last 15 years, a large number of QTL mapping methods were developed. These techniques are the focus of this chapter. For the sake of simplicity, Bayesian QTL mapping methods will not be tackled.

2.2. Principle of QTL mapping

2.2.1. Assumptions

As an archetype of the methods, we will develop the model on one of the most common situations: a F_2 DESIGN, the scoring of MICROSATELLITES markers and the measurement of a normally distributed phenotype. Extensions to other designs, to other types of markers and to non-normal phenotypes will be discussed later in this chapter.

One major requirement of the QTL mapping strategy is that LINKAGE DISEQUILIBRIUM between a marker and a QTL is present. Linkage disequilibrium is defined as an association between alleles at adjacent loci that would not happen if the population was in Hardy-Weinberg equilibrium, i.e. in the case of random mating.

One way of introducing linkage disequilibrium in populations is by crossing lines that differ with respect to their allele frequencies at marker loci and QTL. Indeed, one can assumes that if the chosen lines are divergent for a trait of interest (e.g. one low-producing and one high-producing dairy cattle breed), differences between lines could be due to differences in terms of allelic frequencies at the QTL.

A special case of such a design is the use of inbred lines. It is assumed that, if inbred lines are used, they are fixed for alternative marker and QTL

alleles. Starting with two inbred parental lines, a F_1 is produced and is assumed to be heterozygous for the markers as well as for the QTL. An *inter se* cross of the F_1 individuals produces the F_2 generation. As all F_1 individuals are heterozygous, a variation on the phenotypic scale will appear in the F_2 between their offspring. This will generate linkage disequilibrium between marker and QTL. This design is referred to as the F_2 DESIGN.

MICROSATELLITES are short arrays of simple repeated DNA sequences (of one to six base pairs) and have the advantages of being abundant, multiallelic, highly polymorphic and co-dominant (as heterozygotes can be distinguished from homozygotes).

2.2.2. Conditional probabilities of QTL genotypes

The basic element upon which the formal theory of QTL mapping is built is the conditional probability that the QTL genotype is Q_k (where k is a suffix for the different QTL genotypes, i.e. QQ, Qq and qq), given the observed marker genotype M_j at position j:

$$P\left(Q_k \mid M_j\right) = P\left(Q_k M_j\right) \Big/ P\left(M_j\right) \tag{2.1}$$

The joint and marginal probabilities, $P(Q_kM_j)$ and $P(M_j)$, are functions of the experimental design and of the position of the putative QTL on the linkage map (i.e. relative to marker loci) (Lynch & Walsh, 1998).

Depending on the mapping strategy used (single marker or interval mapping), conditional probabilities will be used either for an *a posteriori* estimation of the effect of the QTL (single marker tests), or as an explanatory variable in the model (interval mapping).

2.2.3. Single marker analysis

2.2.3.1. Principle

To illustrate the method for detecting QTL by association with single markers, consider a marker locus (M) and a putative QTL (Q), with r the

RECOMBINATION FREQUENCY between them. Assume that both are co-dominant. Let the genotype of one parental line be *MQ/MQ* and of the other parental line be *mq/mq*. All F_1 individuals are *MQ/mq*, i.e. heterozygotes at the marker and the QTL.

If the hypothetical QTL has an effect on the phenotypic value, the quantitative genetic theory tells us that the genotypic value of *Q/Q* and *q/q* parents are *a* and *-a*, respectively, where *a* is the additive effect (or allele substitution effect) of the QTL (i.e. half the difference between both homozygotes: $(\mu_{qq}-\mu_{QQ})/2$). Note that μ_{qq} or μ_{QQ} are in fact means of the phenotype conditional on the genotype at the QTL (i.e. if genotype at the QTL is *q/q* or *Q/Q* respectively).

The genotypic value of the F_1 individuals that are *Q/q* is *d*, which is the dominance effect of the QTL. The value of *d* can be computed as $(2\mu_{Qq}-\mu_{qq}-\mu_{QQ})/2$ (Figure 2.2).

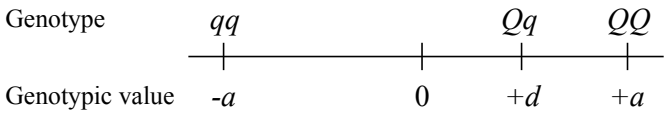

Figure 2.2: Genotypic values of QTL genotypes when allele *Q* is partially dominant over allele *q*

As a function of the recombination frequency between the marker and the QTL, *r*, the frequency of the parental F_1 gametes (*MQ* and *mq*) is (1-*r*)/2 and the frequency of the recombinant F_1 gametes (*Mq* and *mQ*) is *r*/2:

$$P(MQ) = P(mq) = (1-r)/2, \qquad P(Mq) = P(mQ) = r/2 \qquad (2.2)$$

According to these probabilities, the frequency of each F_2 genotypes at the marker and the QTL can be computed under the assumption of random mating of F_1 individuals (Table 2.1).

For example, the (joint) probability that an F_2 individual is *MQ/MQ* is

$$P(MQ/MQ) = P(MQ).P(MQ) = \left[(1-r)/2\right]^2 = (1-r)^2/4 \qquad (2.3)$$

Table 2.1: Probabilities of F_2 genotypes at one marker (*M*) and one linked QTL (*Q*)[a]

Genotype	Joint probabilities of genotypes (A)	Probabilities of marker genotypes (B)	Conditional probabilities of QTL genotypes (A/B)	Genotypic value (C)	Expected marker class mean Σ [(A/B)*C]
MQ/MQ	$(1\text{-}r)^2/4$	} 1/4	$(1\text{-}r)^2$	$+a$	$\mu_{M/M} =$
MQ/Mq	$r(1\text{-}r)/2$		$2r(1\text{-}r)$	d	$a(1\text{-}2r) +$
Mq/Mq	$r^2/4$		r^2	$-a$	$2dr(1\text{-}r)$
MQ/mQ	$r(1\text{-}r)/2$	} 1/2	$r(1\text{-}r)$	$+a$	$\mu_{M/m} = d[(1\text{-}r)^2+r^2]$
MQ/mq	$(1\text{-}r)^2/2$		$(1\text{-}r)^2$	d	
Mq/mQ	$r^2/2$		r^2	d	
Mq/mq	$r(1\text{-}r)/2$		$r(1\text{-}r)$	$-a$	
mQ/mQ	$r^2/4$	} 1/4	$r^2/4$	$+a$	$\mu_{m/m} =$
mQ/mq	$r(1\text{-}r)/2$		$r(1\text{-}r)/2$	d	$-a(1\text{-}2r) +$
mq/mq	$(1\text{-}r)^2/4$		$(1\text{-}r)^2/4$	$-a$	$2dr(1\text{-}r)$

[a] adapted from Table 21.2 of Falconer & Mackay (1996).

The probability of each F_2 genotypes is then expressed conditional on the marker class. For example, the probability of a QTL genotype Q/Q if marker class is M/M, is equal to:

$$P(Q/Q \mid M/M) = \frac{P(MQ/MQ)}{P(M/M)} = \frac{(1-r)^2/4}{1/4} = (1-r)^2 \qquad (2.4)$$

The expected mean of each marker class, expressed in terms of QTL effect and RECOMBINATION FRACTION, is obtained by multiplying the probability of each genotype by its genotypic value and then summing within marker genotype classes. For example, for marker class M/M: $\mu_{M/M} = a(1\text{-}r)^2 + 2dr(1\text{-}r) - ar^2$, which simplifies to $\mu_{M/M} = a(1\text{-}2r) + 2dr(1\text{-}r)$.

From Table 2.1, contrasts of expected marker class means can be built to evidence the presence of a QTL linked to the marker locus

$$\left(\mu_{M/M} - \mu_{m/m}\right)/2 = a\left(1-2r\right) \tag{2.5}$$

$$\mu_{M/m} - \left[\left(\mu_{M/M} + \mu_{m/m}\right)/2\right] = d\left(1-2r\right)^2 \tag{2.6}$$

These contrasts are functions of *a*, the additive effect of the QTL, *d*, the dominance effect of the QTL, and *r*, the recombination frequency between marker and QTL.

A significant difference in the mean value of the quantitative trait between both homozygous marker classes can be taken as an evidence of linkage between a QTL and the marker locus.

2.2.3.2. Analysis of variance

The simplest method to perform single-marker tests to detect QTL is the one-way analysis of variance (ANOVA) at marker loci (Neimann-Sørensen & Robertson, 1961; Hill, 1975). At each marker, the F_2 progeny is separated into three groups, according to their genotypes at the marker (*M*/*M*, *M*/*m* or *m*/*m*), and the phenotypic distributions of the three groups are compared. The null hypothesis which is tested is that the mean of the phenotypic value is independent of the genotype at the marker. Assumptions of this method are that variances are homogenous between marker classes and that the distribution of phenotypes is normal.

A significant between-marker variance indicates the presence of QTL linked to a marker (Figure 2.3). When only two genotypic classes are compared (such as with backcross and half-sib designs), marker genotype classes can be compared with a simple *t*-test.

When lines are partially inbred lines or when analysing outbred populations with large full-sib of half-sib families, the one-way ANOVA is extended to a two-way nested ANOVA to take into account allelic heterogeneity at the marker or at the QTL between families. In this case, the marker effects are nested within family (Hill, 1975; Soller & Genizi, 1978; Knott, 1994).

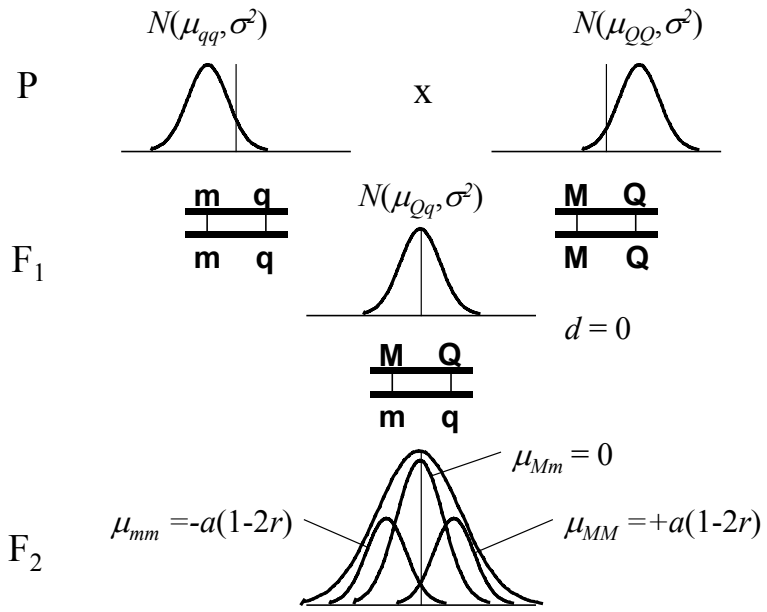

Figure 2.3: Single-marker test for QTL mapping in a F_2 design

2.2.3.3. Limitations

The single-marker approach for QTL mapping has two major limitations. Firstly, the position and the effect of the QTL cannot be estimated separately. A significant marker effect may be equally explained by a QTL with moderate effect tightly linked to a marker and a QTL with larger effect more loosely linked. As shown in equations (2.5) and (2.6), contrasts of marker class means are not only functions of the additive or dominance effects of the QTL but are also function of the RECOMBINATION FRACTION, *r*, between marker and QTL, which is an unknown parameter. The QTL effect is therefore underestimated by an amount that is proportional to the map distance between marker and QTL. Secondly, individuals for which the genotype is missing at the marker must be discarded from the analysis.

2.2.4. Interval mapping

2.2.4.1. Introduction

When markers are ordered on a genetic map, the problem of separately estimating the effect of the QTL and its location can be resolved by the use of interval mapping analysis, initially proposed by Lander & Botstein (1989). For the construction of a genetic map, the recombination between each pair of markers is first estimated, and then any marker that is linked to any other marker is placed in the same linkage group, or chromosome, and ordered on this linkage group. The linear arrangement of markers into linkage groups provides the framework to locate QTL relative to intervals of markers instead of to single markers.

2.2.4.2. Principle

To define conditional probabilities of QTL genotypes at every location along a chromosome, interval mapping uses the genotypes at both markers flanking the position of interest and takes into account the recombination frequencies between the QTL and both marker loci. Indeed, since the RECOMBINATION FRACTION (r) between pairs of flanking markers can be estimated, the recombination fraction between the left marker and the QTL (r_L) as well as the recombination fraction between the right marker and the QTL (r_R) can be inferred if no interference is assumed ($r_L + r_R = r$, for small distances between loci) (Figure 2.4).

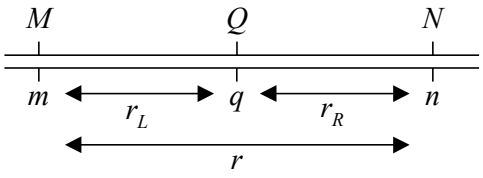

Figure 2.4: Genotype of the F_1 for a putative QTL flanked by two markers and recombination frequencies between each of them assuming no interference

Assume a QTL (*Q*) lying between two co-dominant flanking markers (*M* and *N*) and assume that a cross is performed between inbred lines carrying different alleles for all three loci. The genotypes of the two parental lines are thus *MQN*/*MQN* (with genotypic value +*a*) and *mqn*/*mqn* (with genotypic value -*a*) and of the F_1 is *MQN*/*mqn* (with genotypic value *d*) (Figure 2.4).

As in the single marker approach, the expected mean in terms of the putative QTL for each F_2 marker genotype can readily be derived, by successively computing: (1) joint probabilities of each F_1 marker-QTL gamete; (2) marginal probabilities of each F_2 marker genotype; (3) joint probabilities of each diploid marker-QTL F_2 genotype; (4) conditional probabilities of QTL genotypes given marker genotypes; and (5) expected marker class means from the product of conditional probabilities by genotypic values.

For example, the gamete *MQN* has en expected frequency of $(1-r_L)(1-r_R)/2$ and the gamete *MqN* has an expected frequency of $r_L r_R/2$. The homozygous marker genotype *MN*/*MN* has a (marginal) probability of $(1-r)^2/4$ in the F_2:

$$P(MN/MN) = P(MN).P(MN) = \left[(1-r)/2\right]^2 = (1-r)^2/4 \qquad (2.7)$$

The (joint) probability that an F_2 individual is *MQN*/*MQN* is

$$P(MQN/MQN) = P(MQN).P(MQN) = (1-r_L)^2(1-r_R)^2/4 \qquad (2.8)$$

For genotypes *MQN/MqN* and *MqN/MqN*, this probability is equal to $(1-r_L)(1-r_R)r_L r_R/2$ and $r^2_L r^2_R/4$ respectively. The conditional probability of an individual to be *Q*/*Q* at the QTL given a marker genotype *MN*/*MN* is then:

$$P(QQ \mid MN/MN) = \frac{P(MQN/MQN)}{P(MN/MN)} = \frac{(1-r_L)^2(1-r_R)^2}{(1-r)^2} \qquad (2.9)$$

The corresponding conditional probabilities of an individual to be *Q*/*q* or *q*/*q* at the QTL given a marker genotype *MN*/*MN* are then:

$$P(Qq \mid MN/MN) = \frac{P(MQN/MqN)}{P(MN/MN)} = \frac{2(1-r_L)(1-r_R)r_L r_R}{(1-r)^2} \qquad (2.10)$$

$$P(qq \mid MN/MN) = \frac{P(MqN/MqN)}{P(MN/MN)} = \frac{r_L^2 r_R^2}{(1-r)^2} \tag{2.11}$$

The expected mean of the *MN/MN* marker class, expressed in terms of QTL effect, is then obtained by multiplying these conditional probabilities by the genotypic values of QTL genotypes:

$$\begin{aligned} \mu_{MN/MN} = &+a\left[(1-r_L)^2(1-r_R)^2\right]/(1-r)^2 \\ &+d\left[2(1-r_L)(1-r_R)r_L r_R\right]/(1-r)^2 \\ &-a\left[r_L^2 r_R^2\right]/(1-r)^2 \end{aligned} \tag{2.12}$$

The expected mean values of the other eight marker classes produced in an F_2 are given in Haley & Knott (1992). Similarly to the single marker approach, contrasts of expected marker class means can be built to evidence the presence of a QTL between both marker loci, under the assumption of no interference, e.g.:

$$\begin{aligned} (\mu_{MN/MN} - \mu_{mn/mn})/2 &= a\left(\frac{1-r_L-r_R}{1-r_L-r_R+2r_L r_R}\right) \\ &\simeq a(1-2r_L r_R) \end{aligned} \tag{2.13}$$

As stated by Lynch & Walsh (1998), equation (2.13) is close to *a*, when the distance between flanking markers is inferior to 20 cM, which gives a minimum value for $(1\text{-}2r_L r_R)$ of 0.98. This shows that the interval mapping analysis in F_2 design has the advantage of giving an estimate of the QTL effect nearly independent of recombination frequencies between markers and QTL, as opposed to the single marker analysis. Furthermore, missing markers genotypes can be handled by interval mapping by computing conditional probabilities relative to the nearest informative marker.

An important feature of the interval mapping analysis is that, instead of using the marker class as explanatory variable of the phenotype, conditional probabilities are used. This allows to test for the presence of a QTL not only at marker location but also at every position increment (say 1 cM) between

every pairs of markers. In this manner, if a QTL is located somewhere in an interval, by performing a statistical test at every position in that interval, one can identify the position maximising the test statistic. The position giving the maximum test statistic value is the most likely position of a QTL. If this value is larger than some specified threshold value, a QTL is considered as significantly associated with that position.

2.2.4.3. Least-squares interval mapping

Haley & Knott (1992) and Martinez & Curnow (1992) independently proposed to use a least-squares-based (or regression-based) procedure to perform interval mapping. In this approach, a regression of phenotypes onto conditional probabities is performed in which regression coefficients are a function of the unknown QTL parameters: *a*, the additive effect and *d*, the dominance effect of the QTL. Recalling that genotypic values of QTL genotypes *QQ*, *Qq* and *qq* are respectively equal to $+a$, $+d$ and $-a$, the following regression is performed at each position *j*:

$$z_i = \mu + a.x_{ij} + d.y_{ij} + \varepsilon_{ij} \tag{2.14}$$

where z_i is the phenotypic value of individual *i*, μ is the overall mean equivalent to the polygenic effect (or mid-homozygote value), *a* and *d* are the additive and dominance effects of the QTL, variables *x* and *y* are functions of the conditional probabilities of QTL genotypes given the genotypes (M_j) at the two markers flanking the position *j* being tested (taking into account recombination), and e_{ij} is the residual effect. As detailed in (Lynch & Walsh, 1998), variables *x* and *y* are the following functions of the conditional probabilities:

$$x_j = P\left(QQ \mid M_j\right) - P\left(qq \mid M_j\right) \tag{2.15}$$

$$y_j = P\left(Qq \mid M_j\right) \tag{2.16}$$

The conditional probabilities of QTL genotypes are obtained as described in the preceding paragraph (pt 2.2.4.2). For example, from equations (2.7) to

(2.11), one can show that if the genotype at both flanking markers is *MN/MN*, *x* and *y* are computed as follows:

$$x_j = \frac{(1-r_L)^2(1-r_R)^2 - r_L^2 r_R^2}{(1-r)^2} \tag{2.17}$$

$$y_j = \frac{2(1-r_L)(1-r_R)r_L r_R}{(1-r)^2} \tag{2.18}$$

Owing to this method, interval mapping analysis in F_2 design can be performed by use of standard statistical packages. The only difficulty resides in the computation of conditional probabilities for each individual and at every position increments (cM) between markers.

Using least-squares interval mapping, the evidence in favour of a segregating QTL, at a given chromosome position, is assessed by an *F* test: ratio of the regression mean square to the residual mean square. The position on a linkage group maximising the *F* ratio, and therefore minimising the residual mean square, gives the most likely position of the QTL. A QTL is said to be present at that position if the *F* ratio value exceeds some specified threshold.

2.2.4.4. Maximum likelihood

The use of maximum likelihood to map QTL was originally proposed by Lander & Botstein (1989), in a backcross context. Knott & Haley (1992) extended the work of Lander & Botstein (1989) to the analysis of F_2 populations. The maximum likelihood interval mapping approach exploits the fact that if a QTL is segregating in a population the phenotypic distribution is the mixture different phenotypic distributions depending on the QTL genotype (see Figure 2.3). Indeed, different density functions are fitted to the data according to the genotype at the QTL. As in the regression approach, it is based upon the use of conditional probabilities of QTL genotypes given the genotype at the two flanking markers. These conditional probabilities are used as weighing factors of the different density functions. For an individual (*i*) with phenotypic value y_i, assumed to be normally

distributed, and genotype M_j at markers flanking position j, the general formula of the likelihood is:

$$L\left(y_i \mid M_j\right) = \sum_{k=1}^{3} P\left(Q_k \mid M_j\right) \frac{1}{\sqrt{2\pi\sigma_w^2}} \exp\left[-\frac{\left(y_i - \mu_{Qk}\right)^2}{2\sigma_w^2}\right] \qquad (2.18)$$

or more generally

$$L\left(y_i \mid M_j\right) = \sum_{k=1}^{3} P\left(Q_k \mid M_j\right) \varphi\left(y, \mu_{Qk}, \sigma_w^2\right) \qquad (2.19)$$

where k is a suffix for the different QTL genotypes (i.e. QQ, Qq and qq), $P(Q_k|M_j)$ is the conditional probability of genotype Q_k given the genotype M_j at both flanking markers, and $\varphi(y,\mu_{Qk}, \sigma^2_w)$ denotes the normal density function with mean μ_{Qk} and variance σ^2_w, the variance within a QTL genotype. This variance is assumed to be the same across all genotypes. For an F_2 design with N independent individuals, the total likelihood is the product of the N individual likelihoods:

$$L = \prod_{i=1}^{N} L\left(y_i \mid M_j\right) \qquad (2.20)$$

This total likelihood is a function of four parameters: the overall mean (μ), the additive (a) and dominance (d) effect of the QTL, since $\mu_{QQ} = \mu + a$, $\mu_{Qq} = \mu + d$, $\mu_{qq} = \mu - a$, and the common variance σ^2_w. For each position on the chromosome, the likelihood equation is calculated under the estimated parameters (μ, a, d, and σ^2_w), and then again under the null hypothesis of no QTL ($\mu=\mu_{QQ}=\mu_{Qq}=\mu_{qq}$). The ratio of the two likelihood is calculated either in the form of a LOD score ($LOD = -\log_{10}\left[L\left(H_0\right)/L\left(H_1\right)\right]$), where LOD stands for the logarithm of odds ratio, or in the form of a likelihood ratio test statistic ($LRT = -2\log_e\left[L\left(H_0\right)/L\left(H_1\right)\right]$). The LOD (or LRT) is then plotted against chromosome position. A QTL is said to be present at the position maximising the LOD (or LRT) value when this value exceeds some threshold value.

2.2.5. Statistical aspects

Two aspects of statistical significance are faced by statistical geneticists: (1) the threshold to apply to a QTL mapping test, and (2) the calculation of a confidence interval to QTL locations.

2.2.5.1. Significance thresholds

Results of QTL mapping are often reported as a plot of the test statistic against the chromosomal position (Figure 2.5). A peak in the test statistic curve may indicate that a QTL is located at that chromosomal position. Deciding whether or not this peak (i.e. the maximum value of the test statistic) is significant is a tricky problem (Lander & Botstein, 1989; Churchill & Doerge, 1994; Rebaï *et al.*, 1994; Lander & Kruglyak, 1995; Weller *et al.*, 1998; Piepho, 2001; Cheverud, 2001).

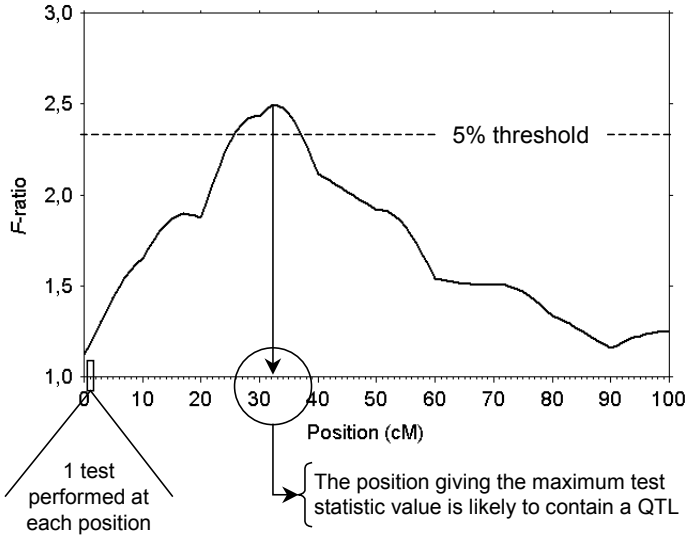

Figure 2.5: Schematic representation of a QTL mapping test on a 100 cM chromosome with 11 markers evenly spaced (10 cM)

Chromosome-wide

As for every statistical test, a null hypothesis is rejected when the probability of the test statistic under the null hypothesis (i.e. the *P*-value) is below a predetermined level, generally either 5 % or 1 % (risk of type I error rate, α).

As opposed to a single statistical test, the problem is to find a threshold value for the unknown distribution of the maximum values of the test statistic along the chromosome (*chromosome-wide threshold*). Furthermore, in QTL mapping studies, another difficulty resides in the fact that non-independent tests are performed. At every position on a chromosome those multiple tests involve markers that are linked. In consequence, the overall false positive rate is expected to be higher than the nominal 5 % level, and the significance level has to be corrected to avoid the detection of spurious QTL.

As stated by Lander & Botstein (1989), when mapping QTL, the dependence between statistical tests is a matter of marker density. If markers are sparse (*sparse-map* case), and if distance between marker is high, dependence is low. If the density is high (*dense-map* case), the distance between consecutive markers tends to zero, and dependence between tests performed at consecutive markers is high.

In the *sparse-map* case, Lander & Botstein (1989) suggest a sort of Bonferroni correction, i.e. to divide the nominal significance level by *M*, the number of intervals tested. In the *dense-map* case, maximum LOD scores were shown to vary according to an Ornstein-Uhlenbeck diffusion process (Lander & Botstein, 1989; Cierco, 1998) and an analytical formula was given to obtain LOD threshold values (function of the number of chromosomes and of the genetic length of the genome).

Out of this LOD score context, and more generally, we have identified four options to compute *chromosome-wide significance thresholds* in QTL mapping studies.

(1) The first option is to use a **theoretical threshold value**, based upon the theoretical distribution of the test statistic (F or χ^2); in this case an arbitrary low type I error rate (e.g. α=0.001) is chosen to correct for the

multiple and dependent tests problem. If the threshold value is correctly chosen, the theoretical threshold approach can have good results, but it doesn't take into account the experimental design (number of individuals, marker density and informativity) and furthermore, if a too stringent threshold value is chosen, the power of detection may be reduced (true QTL may be missed, type II errors).

(2) Another option is to perform **computer simulations** under the null hypothesis of no QTL, with simulation parameters as close as possible to the context of the study (number of individuals, number of markers and alleles per markers, marker density,…). By performing a sufficiently high number of simulations, an empirical null-distribution of the maximum values of the test statistic along the chromosome is obtained. Based on this distribution, a context-based threshold value is obtained by taking the 95 % (or 99 %) quantile of the ordered maximum values as 5 % (or 1 %) threshold; this approach is somewhat tedious and is model dependant; if the correct model is used, the threshold value will however be appropriate (e.g. Sewalem *et al.*, 2002).

(3) A third option is to use **empirical formulae** developed by statisticians to account for the multiple test problem; formulae are either derived from the Bonferroni correction for multiple tests (e.g. Lander & Botstein, 1989; Cheverud, 2001), or from the theory of the nuisance parameter (Davies, 1977; Davies, 1987). The latter considers that the position of the QTL is a nuisance parameter which is present only under the alternative hypothesis (Rebaï *et al.*, 1994; Rebaï *et al.*, 1995; Piepho, 2001); due to its complexity, this latter approach is not frequently used.

(4) The fourth option relies on a **permutation-based approach** (Churchill & Doerge, 1994; Doerge & Churchill, 1996); an empirical distribution of test statistic is obtained by randomly shuffling the observed trait value over individuals (marker genotypes), generating a sample with the original marker information but with trait values randomly assigned over genotypes. The analysis is repeated on this new sample and the maximum value of the test statistic isolated. By repeating this procedure

many times, an empirical distribution of the test statistic under the hypothesis of no marker-trait associations can be generated. For a desired significance level of 5 %, Churchill & Doerge (1994) suggest that at least 1000 permutations of the trait data must be performed, more if a smaller significance level is desired.

Among those four options, the permutation-based approach is the most commonly used.

Genome-wide

An additional dimension of the multiple-test problem is the fact that QTL are sought not only on one chromosome but on several chromosomes, which increases the chance of having a false positive due to the high number of tests performed. To guarantee a 5 % false positive rate at the genome level (*genome-wide threshold*), the 5 % threshold value obtained at the chromosome level cannot be used.

In a genome-wide scan of n independent linkage groups (chromosomes), a common practice to obtain an overall 5 % genome-wide significance level ($\alpha_{gw} = 5\ \%$) is to apply the chromosome-wide level (α_{cw}) solving the following Bonferroni equation (Knott *et al.*, 1998; de Koning *et al.*, 1998):

$$\alpha_{gw} = 1 - (1 - \alpha_{cw})^n \tag{2.21}$$

A very good and simple approximation for the solution of this equation is,

$$\alpha_{cw} = \alpha_{gw} / n \tag{2.22}$$

Another approach based upon permutations is to use the distribution of the maximum values of the test statistic across all chromosomes rather than over a single chromosome. The 95 % (or 99 %) quantile of the ordered genome-wide maximum values is taken as 5 % genome-wide significance threshold (Knott *et al.*, 1998).

Guidelines

In this genome-wide context, Lander & Kruglyak (1995) proposed a classification to report QTL mapping results. It is based on the number of

times that one would expect to see a significant result by chance in a dense, complete genome scan (i.e. with a high marker density and over all chromosomes). This classification is reproduced hereafter.

- *Suggestive linkage*: statistical evidence expected to occur 1/*n* times in a genome scan, where *n* is the number of independent linkage groups.
- *Significant linkage*: statistical evidence expected to occur 0.05 times in a genome scan (that is, with probability 0.05).
- *Highly significant linkage*: statistical evidence expected to occur 0.001 times in a genome scan.
- *Confirmed linkage*: significant linkage from one or a combination of initial studies that has subsequently been confirmed in a further sample, preferably by an independent group of investigators. For confirmation, a nominal *P*-value of 0.01 should be required.

According to Lander & Kruglyak (1995) suggestive linkage results are worth reporting although they will often be wrong. However, they should be accompanied by an appropriate warning label about their tenuous nature.

Instead of controlling the experiment-wide type I error (either chromosome-wide or genome-wide) as it lowers severely the power of QTL detection, Weller *et al.* (1998) proposed controlling the *false discovery rate* (FDR) for preliminary genome scans. The false discovery rate is defined as "*the expected proportion of true null hypothesis within the class of rejected null hypothesis*". The control of the FDR is implemented as follows: the set of *m* null hypotheses (e.g. each test performed on a chromosome) are ordered by their *p*-values $P_{(i)}$ so that: $P_{(1)} \leq \ldots \leq P_{(i)} \leq \ldots \leq P_{(m)}$. For a declared FDR level of q^*, one has to find *k*, the largest *i* for which $P_{(i)} \leq k.q^*/m$. Then, all hypotheses $H_{(i)}$ that correspond to $P_{(i)}$, $i=1,\ldots,k$ are rejected. It is expected

that the proportion of erroneously rejected hypotheses out of k rejected hypotheses is not greater that q^*.

However, the number of null hypotheses to include in the FDR computation is not straightforward. Weller *et al.* (1998) based the presentation of their approach on tests performed at marker position rather than tests performed at each 1-cM position on a chromosome. Lee *et al.* (2002) proposed to determine an FDR level based on the number of marker interval instead of considering each position. Beside these computational aspects, the FDR approach is attractive as it represents a compromise between the restrictive (low power) control methods of the experiment-wide error rate and the permissive results that occur when control for multiple testing is ignored. Although the FDR method to determine significance of QTL mapping tests has been already used in practice (Mosig *et al.*, 2001; Beh *et al.*, 2001; Lee *et al.*, 2002; MacLeod *et al.*, 2002), it is still controversial (Zaykin *et al.*, 2000; Weller, 2000) and its practical implementation needs still to be clarified (Lee *et al.*, 2002).

2.2.5.2. Confidence interval for QTL location

Four methods were proposed so far to estimate the confidence interval (CI) of an estimated QTL map location for a single QTL mapping experiment.

(1) In the **LOD drop-off method** (Lander & Botstein, 1989), a confidence interval (or support interval as they termed it) is calculated by finding the location at either side of the estimated QTL location that corresponds to a decrease in the LOD score of 1 or 2 units. The total width corresponding to a 1 or 2 LOD drop-off is then taken as the confidence interval, and, asymptotically, these should be approximately equivalent to 96.8 % and 99.8 % confidence intervals, respectively (Mangin *et al.*, 1994). This method is however biased since the number of LOD units to use to determine this support interval is influenced by the size of the effect of the QTL and by the type of population (Van Ooijen, 1992; Mangin *et al.*, 1994).

(2) A method was developed by Mangin *et al.* (1994) and Rebaï *et al.* (1994) for constructing the CI for QTL location in backcross populations. It is based upon a **maximum likelihood ratio test** for which the significance level depends on the nuisance parameters in the model (global mean, variance and QTL additive effect). Approximate analytical thresholds were proposed by Cierco & Mangin (1996) for F_2 and Mangin & Goffinet (1997) for backcross using test statistics whose asymptotic distribution under the null hypothesis does not depend on the nuisance parameters. However, those analytical solutions are not practical to use since a maximisation step is required. Therefore, this method is barely used in practice.

(3) Visscher *et al.* (1996b) proposed an approach based upon **non-parametric bootstrapping**. Using this method, bootstrap samples are created by sampling with replacement *n* individual observations out of the pool of *n* original observations. An observation consists of a marker genotype and a phenotype. Some records can appear more than once in a bootstrap sample, while others are not included at all. The QTL mapping analysis is then performed on the bootstrap sample. After *N* bootstrap samples followed by analysis (say 1000 samples), the empirical 95 % confidence interval of the QTL position is determined by ordering the *N* estimates of QTL position and taking the bottom and top 2.5^{th} percentiles. If positions are normally distributed, a Student-based CI can be calculated. This bootstrap procedure is a good alternative to the LOD drop-off method, but tends to give too large confidence intervals for small QTL effects (h^2_{QTL}=0.01 or 0.05) (Visscher *et al.*, 1996b).

To reduce the width of the estimated bootstrap confidence interval, a selective non-parametric bootstrap method was proposed by Lebreton *et al.* (1998). Their selection is based on criteria related to the assumed genetic model. Walling *et al*. (1998) provided a parametric bootstrap re-simulation method as an alternative to the non-parametric bootstrap approach and tested this method in a simulation study. As the authors themselves pointed out, the parametric bootstrap method produced poor results. Bennewitz *et al.* (2002) proposed a correction for confidence

intervals obtained by non-parametric bootstrapping taking into account the bias induced by marker positions (Walling *et al.*, 1998; Walling *et al.*, 2002) while maintaining approximately a 5 % error rate.

The approach of getting CI by bootstrapping has to be differentiated from the permutation approach (Churchill & Doerge, 1994) which is used for the estimation of significant thresholds. In this method, phenotypic observations are permuted among haplotypes (i.e. ordered marker genotypes), and within family if a half-sib design is used instead of inbred line crosses. The sampling is therefore made without replacement as opposed to the bootstrapping approach and furthermore the sampling unit is the phenotype rather than phenotype and genotype taken together (see point 2.2.5.1).

(4) Darvasi & Soller (1997) proposed the use of empirical formulas based on the effect of the QTL identified and on the sample size to compute the confidence interval for the QTL location (**resolving power** as they termed it) for a single experiment. Resolving power is defined as "*the 95 % confidence interval for QTL map location that would be obtained when scoring an infinite number of markers in a given constellation of a marker-QTL experiment*". Their formulas were built for F_2 and backcross designs, but not for half-sib designs.

Among the methods proposed for computing individual CI, the LOD drop-off method is no longer recommended in practice, as shown in several studies. If one wants to get CI representative of the specific conditions of every replicate (informativeness, design, …), bootstrap CI should be preferred. In F_2 and backcross designs, the empirical formulae proposed by Darvasi & Soller (1997) seem to give a good approximation.

2.3. Extensions

2.3.1. Designs

Starting with two inbred lines, a number of line crosses derived from the F_1 can be used for QTL mapping. The F_2 DESIGN examines marker-trait

associations in the progeny from an *inter se* cross of the F_1 individuals, while the BACKCROSS examines marker-trait associations in the progeny formed by the backcrossing of the F_1 to one of the parental lines. The backcross design is less powerful than the F_2 design because it only detects differences between heterozygotes and one of the homozygotes. As three genotypes are generated at each marker locus, the estimation of the degree of dominance associated with the detected QTL is possible in the F_2 design. The contrast between marker class means in a backcross give an estimate of the additive QTL effect (a) which is unbiased only in absence of dominance (d=0). Furthermore, recessive or partly recessive QTL may not be detected. This problem can be overcome by backcrossing to both parental lines (Falconer & Mackay, 1996). While F_2 and backcross designs are the most widely used, other line crosses can offer further advantages. For example, the F_1 can be used to create recombinant inbred lines (RIL) and doubled-haploid lines (DHL), which allow marker-trait associations to be scored in a completely homozygous background and across multiple environments (see Lynch & Walsh, 1998).

Designs using inbred lines are frequently used in laboratory animals (e.g. mice) and in plants. In most livestock species, inbred lines are seldom available. In some experiments, crosses between lines with divergent phenotypes have been used instead. In addition, rearing large number of F_1 and F_2 individuals is only possible for some farm animals (e.g. chicken) but not for others (e.g. cattle) because of the long generation interval and the costs of the experiment. An alternative to the cross between inbred lines, is to use existing outbred populations (i.e. non-inbred).

Linkage disequilibrium in an outbred population between markers and QTL can be found within families. This is due to the co-segregation of marker and QTL. In cattle, analysis can be performed within paternal half-sib families using either the daughter design or the granddaughter design (see Weller *et al.*, 1990). The basic idea of the daughter design, also referred to as HALF-SIB DESIGN) is to trace marker alleles from the sire to his offspring and to determine whether progenies that inherited alternative sire alleles differ with respect to the quantitative trait. In the daughter design, many progenies of a

sire are scored for markers and evaluated for the quantitative trait. In the GRANDDAUGHTER DESIGN, a number of sons of a proven sire are scored for genetic markers and granddaughters are evaluated for the quantitative trait. Then, the trait value for each son is taken to be the mean value of the trait in their daughters (rather than direct measures of the son itself).

There are some important differences between analysing data from a cross between inbred lines and an outbred population (see Bovenhuis *et al.*, 1997). If we consider outbred lines: (1) only a fraction of the sires will be heterozygous for the marker and the QTL, (2) sires might have different linkage phases (associations between marker and QTL alleles) and therefore, observations cannot simply be pooled across families, and (3) QTL might have more than two alleles and allele frequencies are unknown.

These differences between a design based on inbred line crosses and a design within an existing outbred population have important consequences for the statistical analysis of the data, both in terms of power and appropriate methodology (see Haley *et al.*, 1994; Knott *et al.*, 1996; Hoeschele *et al.*, 1997).

In the methodological part of this thesis, aspects such as the power of QTL detection and the accuracy of QTL position will be studied in the context of a half-sib designs (Chapters 5, 6 and 7).

2.3.2. Markers and linkage maps

Beside microsatellites, there are different types of molecular markers that can be used for QTL mapping (Table 2.2). All those marker types differ in terms of number of alleles (level of polymorphism), abundance and degree of dominance. For a review on molecular markers, see Vignal *et al.* (2002).

Genetic or linkage maps show the ordering of loci along chromosomes and the relative distance between them. Such maps are essential for the localisation of QTL. Every livestock species has its own linkage map with an always increasing number of markers. However, within a species, maps are varying according to the population under study and are re-estimated for

most of the data sets. So-called consensus maps are published to provide reference maps to the mapping community (Table 2.3).

Table 2.2: Main categories and properties of molecular markers

Marker acronym	Marker name	Number of alleles / locus	Potential abundance	Marker dominance
RFLP	Restriction fragment length polymorphisms	≥ 2	High	(Co)-dominant
AFLP	Amplified fragment length polymorphisms	2	Very high	Dominant
RAPD	Random amplified polymorphic DNA	2	Very high	Co-dominant
Microsatellite (SSR)	Simple sequence repeats	Many	Medium	Co-dominant
SNP	Single nucleotide polymorphisms	≤ 4	Very high	Co-dominant

Table 2.3: Status of genetic maps of livestock species

Species	Number of chromosomes	Number of markers	Length of linkage map	Reference
Cattle	30	1250	2990 cM	Kappes *et al.* (1997)
Pig	19	1042	2286 cM	Rohrer *et al.* (1996)
Chicken	39	1965	3800 cM	Schmid *et al.* (2000)
Sheep	27	1015	3500 cM	Maddox *et al.* (2001)
Goat	27	223	2300 cM	Vaiman *et al.* (1996)

2.3.3. Multiple QTL

The interval mapping procedure tests for one QTL at a time, as it assumes a single QTL at the position being tested in an interval. This procedure has a few limitations. Firstly, if there is more than one QTL on a chromosome, the test statistic at the position being tested will be affected by all those QTL, if they are affecting the same trait, as they contribute to the residual variation,

and the estimated positions and effects of the QTL identified by this method are likely to be biased (Haley & Knott, 1992; Martinez & Curnow, 1992; Zeng, 1994). Secondly, only two flanking markers are used at a time and the information from other markers is not used (Zeng, 1994). And thirdly, unlinked QTL can also have an effect as they contribute to the phenotypic variance.

To deal with multiple QTL problems, Jansen (1992; 1993) and Zeng (1993; 1994) independently proposed the idea of combining interval mapping with multiple regression analysis. Their MQM (multiple QTL mapping) and CIM (composite interval mapping) methods were implemented using maximum likelihood techniques. Their approach is to evaluate the probability that a marker or an interval between two markers is associated with a QTL affecting the trait, while simultaneously controlling for the effects of other selected markers on the trait, known as cofactors. The purpose of using these cofactors is to minimise the effects of other QTL in the remainder of the genome when attempting to identify a QTL in a particular region. This method considerably increases the power to detect a QTL, by decreasing the within-marker-class phenotypic variation, and improves the precision of estimates of QTL position (Jansen & Stam, 1994; Zeng, 1994).

Kao *et al.* (1999) and Zeng *et al.* (1999) extended the CIM method for mapping multiple QTL using multiple marker intervals simultaneously. Their statistical method (MIM, multiple interval mapping) can give an estimation of EPISTASIS between QTL. However, as the number of QTL increases in the model, the MIM method becomes rapidly computationally intractable, and e.g. Carlborg *et al.* (2000) proposed a genetic algorithm improving computational efficiency.

2.3.4. Multiple-trait

2.3.4.1. Introduction

Several different traits may be measured on the same population. The QTL search is usually conducted on a trait-by-trait basis. The traits, however, are

often correlated and, one QTL can affect more than one trait (PLEIOTROPY). If all traits were used simultaneously, the power to detect the QTL and the precision of the location estimate should be improved. Furthermore, models regarding the correlation between traits could be tested.

2.3.4.2. Methods

Several approaches for the multiple-trait analyses have been proposed. Ronin *et al.* (1995) used bivariate normal distributions to look at the joint likelihood of two correlated traits, affected by a QTL linked to a single marker; Korol *et al.* (1995) extended this method to full interval mapping for two correlated traits. They found, by simulation, that the power of QTL detection could be increased by taking into account correlation between traits, even if the QTL under consideration affected only one of the traits. Jiang & Zeng (1995) extended the CIM method (composite interval mapping, see here above) to include the effects of a locus on several traits simultaneously. Their approach gives a statistical test for PLEIOTROPY versus close linkage. By simulation they demonstrated an increase in power of QTL detection and in the accuracy of parameter estimation.

Weller *et al.* (1996) and Mangin *et al.* (1998) have both proposed the use of canonical transformations (i.e. principal components) of the original data followed by single-trait analyses on the resulting independent combination of traits. However, as stated by Knott & Haley (2000), a transformation that produces traits that are either phenotypically or genetically uncorrelated does not ensure that QTL only influence a single canonical trait: some QTL may be pleiotropic. Korol *et al.* (2001) also used a transformation of the trait space followed by single-trait analysis and subsequently back transformation. The specific feature of their method is that the multivariate transformations are interval specific.

Henshall & Goddard (1999) proposed to use logistic regression for multiple trait QTL mapping, by regressing the QTL genotype on one or multiple phenotypes. Rather than comparing phenotypic means for different marker genotype classes, they compared marker genotype

classes for different phenotypes. The main advantages of this method are that it is much simpler than maximum likelihood, that no assumption is made on phenotype's normality, and that standard software can be used.

Knott & Haley (2000) developed a multiple-trait least-squares method for a three generation pedigree with fixation of the QTL in the grand-parental generation, and provided tests to differentiate between linked and/or interacting QTL. This method can be extended to other population structures. Hackett *et al.* (2001) extended the regression approach for crosses between inbred lines by Haley & Knott (1992) to a multiple-trait analysis via multivariate regression and applied this to a doubled haploid population of barley.

The application of multiple-trait methods, however, is not yet common practice. They can present computational difficulties, especially when applied to an outbred population or if the number of traits is large. Schrooten & Bovenhuis (2002) proposed a simple and fast method for detecting pleiotropic QTL or closely linked QTL in outbred population. It is based upon the covariance between marker contrasts from single-trait regression analysis for different traits.

2.3.4.3. Significance thresholds

As correlated traits are analysed using the same genotype information, the dependency between individual tests is increased when multiple traits are analysed. An option is to reduce multiple correlated traits to a lower number of independent traits explaining all the variance by a principal component analysis on the phenotypic values (Spelman *et al.*, 1996; de Koning *et al.*, 1998). Using this option, the *multiple-trait genome-wide significance level* ($\alpha_{mt\text{-}gw}$) is obtained by solving equation (2.21) for genome-wide thresholds, but with n equal to the number of independent linkage groups (c) multiplied by the number of independent traits (t).

$$\alpha_{mt-gw} = 1-(1-\alpha_{cw})^{c\times t} \qquad (2.23)$$

Another approach based on permutations is to permute the multiple-trait values against the vector of genotypes (keeping together the trait values from an individual). Therefore, the correlations among traits and any linkage among markers are maintained, while relationships among markers and the trait values are randomised (Weller *et al.*, 1998). The 95 % (or 99 %) quantile of the ordered maximum values across all chromosomes and all traits is taken as 5 % (or 1 %) *multiple-trait genome-wide significance threshold*.

Unfortunately, the drawback of these two approaches (formula-based or permutation-based) is to dramatically reduce the power of QTL detection. For large QTL mapping populations that are powerful in nature, this approach is often applied (e.g. Spelman *et al.*, 1996), without taking the risk of having no significant results. However, for less powerful small pedigrees, or for comparison purposes between studies (de Koning *et al.*, 1999), this correction is sometimes omitted. Instead, the highest genome-wide threshold obtained among all traits can be used as in Knott *et al.* (2002).

2.3.5. The candidate gene approach

There are situations in which one can choose candidate loci in place of genotyping a number of anonymous markers spread over the genome. In this case, candidate genes are sequenced genes of known biological action that are involved in the development or physiology of the trait. For example, the genes of the major histocompatibility complex (MHC) are important candidate genes for resistance to infectious diseases. In the candidate gene approach, one tests for the association between the quantitative trait and a genetic polymorphism in the candidate gene.

The first step in performing an association test of a candidate gene with trait variation, is to detect one or more genetic polymorphisms in the gene. This requires that the sequence information is available for the gene. An association test (equivalent to a single-marker test) is then performed between trait values and alleles at individual polymorphic sites or a combination of marker sites (haplotypes) at the candidate gene locus. An

association test can be carried out either within a randomly mating population or in the F_2 progeny of a cross between divergent lines. A significant example of this approach is the estrogen receptor locus for which a specific allele was found to be associated with increased litter size in pigs (Rothschild *et al.*, 1996).

The choice of the candidate gene is sometimes based on a comparative mapping approach, i.e. loci identified to have a phenotypic effect in one species are taken as candidates in other species. It is also motivated by the lower cost involved in the genotyping of a small set of marker sites, instead of a large set of markers covering the entire genome. For a thorough discussion of this approach, see reviews by Rothschild & Soller (1997), Soller & Andersson (1998) and Haley (1999).

2.3.6. Non-normally distributed traits

2.3.6.1. Introduction

Most QTL mapping methods (least-squares-based [regression, ANOVA], or maximum likelihood-based) share a common assumption: that the phenotype follows a normal distribution. Many phenotypes of interest, however, do not satisfy this assumption. Two types of non-normally distributed traits can be distinguished: (1) continuous but non-normally distributed traits, and (2) categorical or discrete traits. The first group comprises e.g. exponentially distributed traits (such as lifetimes), log-normal data, or censored data (such as survival times in an experiment of limited duration), … . The second group includes e.g. binary traits (such as sick or healthy), ordinal traits (such as severity grades of disease susceptibility), count data either with a few categories (numbers of lesions or tumours, …), or with many categories (numbers of bacteria or parasites).

2.3.6.2. Continuous non-normal traits

Mathematical transformation – For continuous non-normally distributed traits, a classical approach to obtain data respecting the assumption of

normality is to apply a mathematical transformation that will convert the trait into an approximately normal variable (see Sokal & Rohlf, 1995, p409). The most frequently used transformations are logarithmic, square-root, Box-Cox and arcsine transformations. If correctly chosen, mathematical transformations can efficiently normalise the data.

Robustness – Another approach is to rely on the legendary robustness of parametric methods (e.g. least-squares-based methods) and analyse non-normal data as such. Compared to least-squares-based methods, maximum likelihood-based methods are likely to be less robust against non-normality when the normal density function is used for the structure of the likelihood.

Simulation studies – Several simulations studies focussed on the robustness of parametric QTL mapping methods to non-normal data. Different types of theoretical distributions were used to simulate non-normally distributed phenotypes (Table 2.4).

Based on position and effect estimates when simulating a skewed phenotypic distribution, i.e. χ^2 with 4 d.f., Knott & Haley (1992) concluded that interval mapping was not sensitive to the distribution of phenotypes. Jansen (1992) evidenced a decreased fit (log-likelihood) when an exponentially distributed phenotype was analysed, although parameter estimates were much the same. Rebaï (1997) showed that estimates of the QTL effect were close to the simulated value when an exponentially distributed phenotype was analysed with regression interval mapping. Hackett (1997) showed that, if data from a *log*-normal distribution are analysed as though they were normal, the maximum value of the test statistic is reduced and the support intervals are wider than if a transformation was used before QTL analysis.

Only two out of the ten simulation studies (Table 2.4) provided power estimates of a parametric method when analysing normal and non-normal data (Coppieters *et al.*, 1998; Zou, 2001). The eight other studies compared the values of different kind of parameters as threshold value, recombination fraction between marker and QTL, position and effect of the QTL. In a half-sib design context, Coppieters *et al.* (1998) evidenced that the power of least-squares-based interval mapping in half-sib designs was slightly reduced

(2-3 % of power) when the phenotypic distribution was close to a Student or to a χ^2 distribution. Those two distributions were chosen respectively as examples of homoscedastic skewed and homoscedastic with positive kurtosis distributions.

Table 2.4: Simulations of non-normally distributed phenotypes

		Distribution[b]						
Reference	Method[a]	Exp.	Ga.	*log*-N	χ^2	St.	Cau.	HC
Knott & Haley (1992)	ML				x			
Jansen (1992)	GLM	x						
Doerge & Rebaï (1996)	--		x					
Rebaï (1997)	LS	x	x					
Hackett (1997)	LS			x				
Muranty & Goffinet (1997)	ML				x			
Coppieters *et al.* (1998)	LS				x	x		x
Henshall & Goddard (1999)	LR				x			
Zou (2001)	LS	x				x	x	
von Rohr & Hoeschele (2002)	Bayes				x	x		

[a] ML=maximum-likelihood, GLM=generalised linear model, LS=least-squares, LR=logistic regression, Bayes=bayesian method.

[b] Exp.=exponential, Ga=Gamma, *log*-N=*log*-normal, St.=Student, Cau=Cauchy, HC=hemicircular.

In a backcross context, Zou (2001) showed that the power of the least-squares-based approach was drastically reduced when the distribution of the trait was non-normal: from 100% with a normally distributed trait, to 74 %, 55 % and 2 % respectively for exponential, Student and Cauchy distributions. However, in these two simulation studies, no mathematical transformation was applied to the non-normal trait before least-squares analysis, although it is done in practice in most of the cases.

Based on these simulation studies, robustness of parametric QTL mapping methods to non-normally distributed phenotypes is difficult to establish.

Jansen (1992) presented a general mixture model for mapping QTL which uses the distributional properties of the data by fitting a generalised linear model (GLM). The proposed approach is to use the knowledge about the appropriate distribution of the data to define the likelihood function rather than assuming normality. Jansen (1992) gave an example for a trait in which the residuals were from an exponential distribution and showed a better fit of the model when the right distributional assumption was done. Such an approach could be used for other types of continuous non-normal distributions.

Non-parametric – An alternative approach is to use non-parametric (or distribution-free) methods. Kruglyak & Lander (1995a) described a non-parametric interval mapping approach based upon the Wilcoxon rank-sum test applicable to backcross designs. Those authors stated that the non-parametric interval mapping test was expected to be more powerful than parametric methods in certain cases. Based on the book of Kendall & Stuart (1979), Kruglyak & Lander (1995a) gave the example of an exponential distribution in which the non-parametric test would outperform parametric ones. However, no power estimates of the non-parametric test regarding such a distribution of phenotypes was given in this first study.

Coppieters *et al.* (1998) adapted this method to outbred half-sib pedigrees (more common in livestock species). In a simulation context, the non-parametric test was shown to give a gain of power compared to least-squares based method in some specific conditions: when the distribution of phenotypes is either heteroscedastic normal, or homoscedastic skewed, or homoscedastic positively kurtosed. In this study, heteroscedasticity and homoscedasticity were referring to the condition of homogeneity of variances within QTL genotype classes. Coppieters *et al.* (1998) explained the relative lower efficiency of the least-squares-based method by the presence of outliers in the distributions which contribute excessively to the residual variance. However, in this study no mathematical transformations

was sought to normalise the data prior to analysis with the least-squares method.

In a backcross context, Zou (2001) evidenced a gain of power for the non-parametric test compared to the least-squares approach when analysing non-normal traits. This gain was respectively equal to 18 %, 45 % and 56 % of power when analysing exponential, Student and Cauchy distributed traits. As for Coppieters *et al.* (1998), no mathematical transformation was applied to the non-normal trait prior to analysis.

Although already implemented in some QTL mapping software (e.g. Mapmaker/QTL, Lander *et al.*, 1987; MapQTL, Van Ooijen & Maliepaard, 1996) and used in mapping QTL in mice or plants, Broman (2003) extended this method to F_2 designs for which three categories of QTL genotypes must be compared. A non-parametric analogue of the analysis of variance (Kruskal-Wallis test) was proposed.

Survival data – Finally, among all non-normally distributed data, one type of data requires a specific method of analysis. Indeed, survival times are often censored as experiments are of a limited duration. An adequate model for survival times is the Cox's proportional hazards model. In a single-marker context, Hu *et al.* (1997) and Sebastiani *et al.* (1998) used this model to compare marker genotype groups. Recently, an interval mapping method based on the Cox's proportional hazards model was briefly described by Symons *et al.* (2002) and in more details by Moreno (2003). It uses a maximum likelihood approach and provides a Cox likelihood ratio test for interval mapping. By means of simulation, Moreno (2003) evidenced a higher power of QTL detection using this approach compared to a maximum likelihood approach under the assumption of normality.

2.3.6.3. Discrete traits

Discrete traits are the result of a phenotypic assessment either on a binary scale, or an ordinal scale, or are based on a counting process.

Binary traits

Complex binary traits have a dichotomous phenotypic expression but do not show a simple Mendelian segregation ratio. These traits are considered to be jointly controlled by the actions of several genes and environmental factors. Because of this feature, these traits belong to quantitative traits. Complex binary traits are usually analysed using a THRESHOLD MODEL, in which it is assumed that the observed category is determined by the value of an underlying unobservable continuous variable (Falconer & Mackay, 1996). The underlying continuous variable, called the LIABILITY, can be considered as a regular quantitative trait which can be partitioned into genetic and environmental components. The binary phenotype and the continuous liability are linked through a fixed but unknown threshold.

For example, for a binary trait such as resistant *versus* susceptible to an animal disease, the probability of being resistant can be defined as:

$$P(R) = P(Y > y) \tag{2.24}$$

where Y is the underlying continuous variable, and y is the threshold. When mapping QTL for such traits, it is assumed that the value of this threshold is a function of the genotype at the putative QTL.

Methods for mapping QTL involved in the determination of binary traits have been developed in line-crossing experiments. These methods are less obvious than for normal traits due to the non-linear relationship between the observed phenotype and the unobservable genetic effects.

Note that methods for binary (and ordinal traits) are given hereafter for information but will not be used in the following sections of this thesis.

Using the threshold model, Hackett & Weller (1995) and Xu & Atchley (1996) developed methods to map QTL controlling the liability. These methods were based on a logit probability model using a backcross as an example. Hackett & Weller (1995) used a generalised linear model (GLM) for the estimation of the parameters. Xu & Atchley (1996) used an expectation-maximisation (EM) algorithm. In a GLM framework, Visscher

et al. (1996a) compared the classical parametric linear regression to the logistic regression using either a probit or a logit link function, and both link functions gave similar results. When compared to the classical linear regression (RIM) applied on binary data, the logistic regression gave similar power of QTL detection. Instead of using an EM algorithm, Rebaï (1997) proposed an iteratively weighted least-squares algorithm to estimate the parameters of the logistic model and Xu *et al.* (1998) used an heterogeneous residual variance model.

All these methods were primarily conducted in a single cross or family. However, plant and animals show typically large family sizes. Yi & Xu (1999b) developed a fixed-model approach to map QTL for complex binary traits from multiple full-sib families in outbred populations, and showed that their method is efficient when there are a small number of large families. As the number of families increases, the fixed model approach becomes inefficient because of the large number of parameters to be estimated. Yi & Xu (1999a) proposed therefore a random model approach in which the effect of each allele is treated as a random variable so that a single variance rather than individual allelic effects is estimated and tested. Kadarmideen *et al.* (2000) set up a (likelihood-based) generalised interval mapping method to map QTL for complex binary traits in multi-family half-sib designs based on the threshold theory and implemented using a Newton-Raphson algorithm for fitting the likelihood. This study showed that the least-squares-based method and their generalised interval mapping method had similar power of QTL detection.

Ordinal traits

Methods for binary traits can easily be extended for ordinal traits, as the threshold model is also valid for this type of traits (Hackett & Weller, 1995). Rao & Xu (1998) adapted the logit probability model used for binary trait to the analysis of ordinal traits in four-way crosses. In a GLM framework, Spyrides-Cunha *et al.* (2000) developed a proportional odds model for testing the presence of QTL at single-markers. Van Ooijen *et al.* (2001)

proposed the use of the non-parametric Kruskal-Wallis test for QTL mapping for traits assessed on an ordinal scale.

Count data

Two types of count data are usually encountered in practice:

(1) counts between zero and a small maximum value (say 20-25), such as litter size in pigs or mice, teat number in pigs, or tumour counts in mice (hereafter referred to as *small counts*);

(2) counts with large range (between zero and larger than 25) such as bacteria counts, faecal egg counts or somatic cell counts that are indicator traits of the disease resistance status of animals (hereafter referred to as *large counts*).

Small counts – Classically, when searching for QTL involved in the determination of *small counts* phenotypes, such as litter size or teat number in pigs, classical statistical methods based on the normal assumption are used (e.g. Rothschild *et al.*, 1996; Hirooka *et al.*, 2001; de Koning *et al.*, 2001). However, due to counting nature of the data, a Poisson model would be intuitively preferred. In genetic analysis, Poisson models were firstly used by Pérez-Enciso *et al.* (1993) in modelling litter size in pigs. Up to now, a few QTL mapping method were developed for the analysis of count data. Kayis *et al.* (1998a) proposed a single-marker single-QTL method for the QTL analysis of backcross mice litter size data. Litter size was assumed to have a Poisson distribution, and parameters were estimated using the expectation-maximization algorithm. Respectively for backcross and F_2 data, Kayis *et al.* (1998b) and Kayis *et al.* (1999b) extended this method in the GLM framework to allow for interval mapping and for the incorporation of fixed effects. Shepel *et al.* (1998) used a Poisson regression model for the estimation of the interaction between loci and of the gene effects on mammary tumour numbers in the a combined analysis of backcross and intercross data in the rat. This model was not aimed at locating the QTL along chromosome. Recently, Thomson (2003) presented a Poisson regression model for mapping QTL for litter size in mice. Both fixed and

random effects can be incorporated in this model and it is fitted using a generalised estimating equations approach, which allows for the specification of a reduced probability model.

As the Poisson distribution constrains the variance of the distribution to be equal to the mean, a negative binomial model to allow for overdispersion could be more appropriate for the analysis of count data (Tempelman & Gianola, 1999) when the Poisson conditions are not met (e.g. heterogeneous distribution).

In an analysis of the genetic determinism of tumour multiplicity in mice, Lee *et al.* (1995) used an equation (square-root transformation) proposed by Anscombe (1948) to improve the fit of negative binomial data to the normal distribution (Drinkwater & Klotz, 1981). Poole & Drinkwater (1996) used a *log*-linear model for the estimation of additive and dominance effects for liver tumour multiplicity in a combined analysis of backcross and intercross data. More recently, a negative binomial multiple regression model was used to perform genome scans in F_2 crosses for tumour multiplicity in mice (Nagase *et al.*, 1999; Lan *et al.*, 2001; Nagase *et al.*, 2001).

Large counts – Among large counts phenotypes, the somatic cell count used to assess the resistance of dairy cattle or sheep to mastitis is the most frequently used. For large counts, a *log*-normal distribution can frequently be assumed. Mapping QTL for somatic cell counts is usually performed on *log*-transformed data (Heyen *et al.*, 1999; Schrooten *et al.*, 2000) and standard methods based on the normality assumption can be applied. As somatic cell counts are measured repeatedly, the QTL mapping is preferably performed on lactational means of somatic cell counts, daughter-yield deviations or estimated breeding values. In such a context, the non-normality of the data can easily be overcome by a *log*-transformation or can even disappear by the use of aggregate measurements as the effect of outliers is attenuated.

Bacteria counts (or CFU for colony-forming units) and faecal egg counts (FEC) are indicator traits that are often used to assess the disease resistance status of animals to bacterial and parasitic diseases respectively. The distribution of this type of trait is right-skewed with a long tail, and a high

frequency of zero values is often observed (zero-inflated distribution). Such distributions are neither really continuous nor strictly categorical. Indeed, because of the dichotomous aspect of the trait (affected or unaffected), one can be tempted to dichotomise the data and analyse it as a binary trait. Furthermore, because of the counting aspect of the trait, one can think that a Poisson process could fit. Finally, since counts can take on high values resembling a continuous phenomenon and have sometimes non-integer values due to the sampling method, a *log*-normal distribution could be imagined.

In a study on QTL mapping in the case of a spike in the phenotype distribution, Broman (2003) proposed a two step approach. This approach is to first analyse the data, using only the individuals having a non-zero count, by standard interval mapping (Lander & Botstein, 1989), and then separately analyse the binary trait (0 or count > 0) according to the method of Xu & Atchley (1996) based on maximum likelihood. Furthermore, Broman (2003) proposed also a maximum likelihood approach called *two-part model*, in which the likelihood is composed of a mixture between a Bernoulli distribution and normal distributions, depending on the QTL genotype. This model assumes a sort of zero-inflated normal distribution.

Limitations

Among all the studies about mapping QTL for discrete traits, only two studies compared the proposed method to standard regression interval mapping in terms of power of QTL detection (Visscher *et al.*, 1996a; Kadarmideen *et al.*, 2000), respectively in F_2 and backcross and in half-sib designs. Both studies evidenced that the least-squares based method gave similar power than other models.

Regarding the analysis of overdispersed count data under the assumption of a negative binomial distribution, the model developed by Nagase *et al.* (1999) is restricted to the analysis of F_2 or backcross between inbred lines, in which only one negative binomial distribution can be assumed for all animals. As opposed to inbred line crosses, in half-sib (or outbred) designs, different distributions (and therefore parameters) should be assumed for each

half-sib families as overdispersion can be assumed to differ between sire-families. Up to now, no such model was proposed.

2.4. Conclusions

The QTL mapping methodology described in this chapter is complex, but fascinating. During the last decade, various methods have been developed to detect quantitative trait locus, ranging from the simple regression method, to maximum likelihood methods and even Bayesian methods (omitted in this introductory chapter for simplicity).

Although many publications treated this topic since the innovative paper of Lander & Botstein (1989) the problem of declaring a QTL mapping test significant is still under development. Methods were provided to correct significance levels for the multiple test problem, but people are still afraid of missing true QTL when applying a too stringent significance level.

Most of the QTL mapping methods available are based under the assumption that the trait follows a continuous (normal) distribution. However, many traits of scientific and economic interest have a non-normal distribution. For example, non-normal distribution is frequently encountered when assessing the disease resistance status of animals. Indicator traits of the disease resistance of animals which vary from binary traits (such as diseased versus healthy) to complex zero-inflated traits such as bacteria counts are the focus of the following chapter.