Chapter 9

Methods for QTL analysis

Julius van der Werf

In this Chapter we will discuss in more detail regression analysis and Maximum likelihood methods for QTL mapping. Regression methods are generally much easier to use (standard software like SAS or ASREML can easily be used), and the method is much faster computationally. Maximum likelihood is computationally more demanding, and specific software is needed. For many designs, results are very similar to regression. This makes regression analysis attractive as it can be used in resampling methods. Resampling methods are use to determine test statistics for hypothesis testing. In this Chapter we will discuss bootstrapping and permutation tests.

We will also discuss QTL mapping with multiple markers (more than 2) and methods to account for more than one QTL. Accounting for other QTL has been proposed by including cofactors, or by using composite interval mapping.

There are two classes of methods that are not discussed in the chapter. Those are the mixed model methods and Monte Carlo Markov Chain methods. In both methods, QTLs are modeled either as fixed or as random effects, and additional random effects can account for polygenic variation. Combined segregation and linkage analysis is needed to infer QTL genotype probabilities from marker data.

Both methods are useful in ‘complex pedigrees’, typical in animal breeding data from outbred populations. When line crosses are analysed, or half sib families ignoring relationships across families, such methods are less relevant, and they have not been extensively used in QTL detection studies. In most animal breeding applications, however, such methods are typically needed in genetic evaluations including QTLs. We will discuss mixed model methods including QTL effects in chapters 17 and 18.

Regression Methods

ANOVA analysis using single marker genotypes.

A marker genotype (or marker-haplotype) represents a fixed effect class.

$$y = \mu + MG_1 + e$$

The number of marker genotypes is 2 in backcrosses of inbred lines and 3 in F2 populations. However, most animal populations are not inbred and could have more genotypes, which will have less power.
The analysis gives an F statistic, and provides a quick and simple method to detect which markers are associated with a QTL.

ANOVA analysis using multiple marker genotypes.

Each marker genotype (or marker-haplotype) represents a fixed effects class.

$$y = \mu + MG_1 + MG_2 + \ldots\ldots\ldots + MG_n$$

This is a multiple regression model, and markers can drop out of the model if they are not significant. The set of markers that is significant in the final analysis point to the existence of a significant QTL effect (or more, depending how far the markers are apart). The analysis does not take into account any recombination rates between markers, or between QTL and markers. In that sense it is comparable with regression on single marker genotype. The multiple marker method is more powerful than single marker analysis, and when the markers are well spread over the genome, it is better able to distinguish the position of the QTL. Normally, after detection of such a location, analysis with interval mapping would be recommended.

Regression on QTL probability, conditional on marker haplotypes.

For a given marker genotype, or marker haplotype that was inherited from the sire, we can calculate the probability for having inherited the Q or the q allele. It seems therefore natural to regress phenotype on Q-probability. The model is

$$y = \mu + \alpha.x + e$$

where y is the observed phenotype
x is the probability of having inherited a paternal Q, given the observed marker genotypes, and marker/QTL positions: P(Q|mg1, mg2, r1,r12)

The coefficient for x are obtained as in Chapter 7 (Table 4). For a each QTL position, the residual sums of squares can be determined, and the estimate of the QTL position is there where SSE is minimum. This is interval mapping (see Chapter 7)

Haley-Knott regression

Haley and Knott (1992) have proposed a slight reparameterization from the previous model, but the principle is similar. Rather than dealing with marker haplotypes, they present a more general model where QTL genotypes are dependent on marker genotypes. The probability of carrying a certain QTL genotype depends on the marker genotypes and the design

$$y = \mu + \alpha.x_1 + \beta x_2 + e$$

where y is the observed phenotype
$x_1 = P(QQ|M_i) - P(qq|M_i)$
$x_2 = P(Qq|M_i)$

x_1 and x_2 are probabilities for QTL genotypes conditional the flanking marker genotypes. The regression coefficients α and β represent the difference between the homozygote QTL genotypes, and the QTL dominance effect, respectively.
Haley and Knott are well known for their proposed regression model, but an important result from their paper was the similarity that was shown with maximum likelihood. They proposed to use the following test statistic, indicated as 'approximate Likelihood ratio test':

$$LR = n\ln\left(\frac{SSE_{reduced}}{SSE_{full}}\right) = -n.\ln(1-r^2)$$

Which is ration of the residual sums of squares in a model with the QTL ("full') and a model without it ('reduced'). The term r^2 is the usual R-squared, used for the percentage of variance explained by the model (only applicable if there are no other fixed effects).

Regression of phenotype on marker type

The previous two regression models proposed regressing phenotype on Q-probability, conditional on marker type. As this probability depends on QTL position, relative to markers, interval mapping can be used. A regression analysis is needed for all possible positions (usually in 1 cM steps) within the marker bracket.

Whittaker et al. (1996) have shown that direct regression of phenotype on marker types, provides the same information about location and QTL-effect without having to step to all positions on the interval.

For interval mapping we used: $y = \mu + \alpha.x + e$ [1]

where x = P(Q|mg1, mg2, r1,r12)

Whittaker et al. (1996) proposed their model for a backcross or F2 population:

$$y = \mu + \alpha\lambda.x_L + \alpha\rho.x_R + e$$

Now $\lambda = P(Q|X_L = M1M1, X_R = m2m2)$ and $\rho = P(Q|X_L = m1m1, X_R = M2M2)$.
The term α is the effect of Q. The terms x_L and x_R refer to left and right marker, and have values –1, 0 and 1 for m_im_i, M_im_i and M_iM_i, respectively. From the regression coefficients: $\beta_1 = \alpha\lambda$, and $\beta_2 = \alpha\rho$, it was shown (Whittaker et al., 1996) that location and QTL effect can be estimated:

location (recombination between M1 and QTL)

$$r_1 = 0.5\left[1 - \sqrt{1 - \frac{4b_2q(1-q)}{b_2 + b_1(1-2q)}}\right]$$

and the estimate of the QTL effect:

$$a = \sqrt{\frac{[b_1 + (1-2q)b_2][[b_2 + (1-2q)b_1}{1-2q}}$$

where θ = r1+r2(1-2r1). Hence a single analysis can give the same result as a complete interval mapping. Note that the assumption is here that there are no QTL's in the neighboring marker-brackets.

Maximum Likelihood estimation:

In these notes, we will not discuss the detail of a maximum likelihood analysis (for interested readers are referred to Lynch and Walsh (1998). Only the principle is given here.

We have a probability of observing certain data (y) for a given set of parameters (θ):

$$F(y_i) = P(y|\theta)$$

This function F is indicated as probability density function (pdf). For example, if we take normally distributed observations, and the simplest model, with a mean (μ) and standard deviation (σ) the pdf looks like:

$$f(y_i|\ \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{\frac{1}{2}(y-\mu)^2}{\sigma^2}} \qquad [2]$$

The likelihood is the probability of certain parameters, given the observed data: $L(\theta|\ y)$. We can use the same function for this, e.g.

$$L(\mu, \sigma|y_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{\frac{1}{2}(y-\mu)^2}{\sigma^2}}$$

The total likelihood of data set **y** is calculated as the product of all likelihoods for each observation.

$$L(\ \mu, \sigma|\ \mathbf{y}) = \Pi_i\ L(\mu, \sigma|y_i)$$

As these likelihoods can become very small numbers, is better to work with the LogLikelihood

$$LogL(\ \mu, \sigma|\ \mathbf{y}) = \Sigma_i\ LogL(\mu, \sigma|y_i)$$

Also for an alternative model, e.g. with a QTL effect, we may have different means. A new set of parameters is then (μ_1, μ_2, α, and σ) and we can write the likelihood.

$$L(\mu_1, \mu_{2,}, \sigma|y_i) = P(\mu_1).\frac{1}{s\sqrt{2p}} e^{\frac{\frac{1}{2}(y-m_1)^2}{s^2}} + P(\mu_2).\frac{1}{s\sqrt{2p}} e^{\frac{\frac{1}{2}(y-m_2)^2}{s^2}} \qquad [3]$$

Typically, in QTL analysis, we are not sure about QTL genotype, i.e. whether an observation belongs to the Q-mean or to the q-mean. The likelihood is calculated as the sum of the two possibilities, each weighted with its probability (=$P(\mu_I)$).

The estimates of the model parameters are obtained for those values where the likelihood is at its maximum. The maximum can be found using maximization routines (EM; Newton Raphson; NAG-libraries).
A test of significance is obtained by comparing the maximum likelihood with the likelihood of a model with the tested parameter omitted (reduced model).

$$LR = -2\ln\frac{Max\_Likelihood(reduced\ \mathrm{mod}\ el)}{Max\_Likelihood(\ full\ \mathrm{mod}\ el)}$$

The reduced model refers to the null-hypothesis, e.g. "there is no QTL effect"

Using the log-likelihood: $LR = -2.(ln\_L_r - ln\_L)$ where L stands for LogLikelihood.

Example of simple QTL mapping with maximum likelihood

In QTL analysis the data consists not only of phenotypic observations of performance, but also of marker genotypes.
Using the example as in chapter 7, where we looked at a half sib family with known paternal marker haplotypes, we could calculate the probability of having inherited the paternal QTRL alleles for each of the four marker haplotypes (and given the recombination fractions, i.e. for a given QTL position)

If the dam alleles are fixed there are only two possible QTL genotypes, hence we can calculate the likelihood for each observation as in [3]. If the dam alleles are not fixed, we would have to sum over all three possibilities.
In a simple fixed effects model, the ML estimate of the fixed effect parameters is equal to the LS estimate of the fixed effects. Hence for a given QTL positions we can calculate μ and α from a regression as in [1] and subsequently calculate the likelihood as in [3].

The following Table shows a likelihood calculation of the example as in Chapter 7, for the QTL position M1-Q = 0.1

Phenotype	Marker haplotye	Prob(Q|markers)	Expected phenotype (H1-model)	LogL0	LogL
50.98	M1M2	0.9718	50.43	-1.18884	-0.81727
49.98	M1M2	0.9718	50.43	-0.4575	-0.65658
50.75	M1m2	0.7451	50.34	-0.73859	-0.59655
49.75	M1m2	0.7451	50.34	-0.73859	-0.91164
50.75	m1M2	0.2549	50.16	-0.73859	-0.91152
49.75	m1M2	0.2549	50.16	-0.73859	-0.59663
50.52	m1m2	0.0282	50.07	-0.4575	-0.65648
49.52	m1m2	0.0282	50.07	-1.18884	-0.81739
			sum	-6.24705	-5.96407

Model with no QTL:
The general mean = μ_0 = 50.25.
SST = (sum of deviations from general mean) = 2.21 giving a variance σ_0^2 = 0.316

The likelihood is calculated according to [2] using μ_0 and σ_0^2
The sum of the Log Likelihood over the whole data for the H0-model = -6.247

Model with a QTL
Regression analysis gave solutions μ = 50.057 and α = 0.386.
SSE = (sum of deviations from expected phenotype) = 2.05 giving a variance σ^2 = 0.292

The likelihood is calculated according to [3] using σ^2, and the two means are

$\mu_Q = \mu + \alpha = 50.443$ and $\mu_q = \mu = 50.057$ and the weights are P(Q) and 1-P(Q), where P(Q) is given for each individual in the third column of the Table.

The sum of the Log Likelihood over the whole data for the H0-model = -5.964

The LR-value = -2(L0 – L) = -2(-6.247 + 5.964) = 0.57.
(Note: this is NOT the Maximum Likelihood, as we have used the residual variance as (over) estimated by regression).

The approximate LR value from regression was

$$\text{appr.LR} = n\ln\left(\frac{SSE_{reduced}}{SSE_{full}}\right) = 8.\ln(2.21/2.05) = 0.63.$$

Comparison of likelihood and regression procedures

The difference between maximum likelihood and regression is that the last method assumes normality within a marker group, i.e. there is a homogeneous variance within a marker group (errors only due to e). Maximum likelihood accounts for the fact that within a marker group, some animals have obtained a q and some have obtained a Q, hence there are actually two distributions. The fact that the test statistics are practically very similar shows that accounting for this bimodality within marker genotypes is not very important. Most of the variation is explained from the differences between the marker genotypes. Xu(1995) shows that the regression method is somewhat biased: it overestimates the residual variance, and therefore tends to give lower values for the approximate LR test. This bias is larger if the difference between Q and q is larger, and when there is less certainty about QTL-allele inherited. The largest differences between the two methods will be found in the middle of a marker bracket, when there is most uncertainty about which QTL allele was inherited.
Xu's suggest correction is

$$\sigma^2_{e\_corrected} = \sigma^2_e - a^2 \sum_{i=1}^{4} p_i(1-p_i)$$

where p_i is the probability of having inherited Q in marker genotype class i and *a* is the regression coefficient on Q-probability in the regression model. Generally, this adjustment has only a small effect, unless the QTL effect is very large and markers are far from the QTL position

It should be noted that ML procedures depend on the distribution of the phenotypes. Regression analysis is much more robust against deviation from normal distributions. On the other hand, in outbred populations, ML is better able to use all possible relationships to infer upon marker- and QTL probabilities. With no markers, ML analysis would still boil down to a segregation analysis, whereas regression methods would not be able to make any inferences at all. However, regression methods combined with a genotype-probability-type algorithm could be very competitive to a ML analysis (see

Chapters 17 and 18).

Accounting for additional QTLs

In the examples discussed, we looked at detecting a single QTL in a marker bracket. Now, if there other QTL linked to the markers used in the analysis, we would tend to estimate the joint effect of two QTL's, and we would not be able to distinguish between one or multiple QTL. Moreover, the inference we would made from analysis regarding size of QTL effect and QTL position would both be biased. We may observe two peaks in a likelihood map, which would be an indication of the existence of two QTL, but both positions would be biased. Besides avoiding bias, another reason for accounting for additional QTL effects is to reduce residual variance, giving more power to an analysis. This would also hold for additional QTL on other chromosomes (unlinked).

A few approached have been proposed to avoid effects of additional linked QTL.

Multiple regression on marker genotypes,
The effect of a QTL on one marker is corrected for possible effects of linked QTL-effects. The effects of the linked QTL are taken away by effect by fitting markers close to these QTL. A simple regression method that considers all markers has been proposed by Kearsey and Hyne (1994).They propose to plot the difference between marker types, i.e. one difference for each marker locus. This is described in more detail by Lynch and Walsh (1998, p. 461), who refer to this method as marker-difference regression.

Interval mapping with marker co-factors (composite interval mapping)

Jansen (1993) proposed an interval mapping approach where additional markers were included in the model as *cofactors*. Such an additional QTL (say QTL2) can be accounted for if there is information about additional markers (outside the bracket) that are linked to QTL2. This analysis is also referred to as composite interval mapping (CIM) (Zeng, 1994). Regression is on the additional marker genotypes are, hence, additional QTL are accounted for as if they were at the marker locus.

$$y = \mu + \text{p(QTL1 given marker bracket M1M2)} + \text{markers near QTL2} \quad [5]$$

Several authors have shown that composite interval mapping gives a large increase in power, and much more precision in estimating QTL position.

As we discussed earlier in this chapter, Whittaker et al (1996) found that the regression coefficient for two adjacent markers contain all information about position and effect of a QTL between those markers. If the QTL is isolated, i.e. there are no QTL's in the adjacent brackets, than these regression coefficients can not be biased by other QTLs outside the bracket. However, no distinction can be made between on or more QTL within the bracket. hence, the position estimate within a marker bracket is only unbiased if there is only one QTL. If there are more QTL within the bracket, we can not estimate their positions.

rather than accounting for more QTL as in [5] we can also account for them with the following model:

$$y = \text{p(QTL1| M1M2)} + \text{p(QTL2| other markers near QTL2)} \quad [5]$$

hence this refers to a multiple interval mapping procedure (Kao et al., 1999).
Some problems here can be that 1) not all markers are informative, especially not in outbred populations 2) it is hard to search for the best fitting model (set of positions) as there are many combinations possible with multiple QTL.
The problem of multiple QTL will be further dealt with in chapter 10.

Precision of mapping and hypothesis testing

Maximum likelihood estimates are approximately normally distributed for large sample sizes and confidence intervals can be based don the sampling variances. However, these are often not so easy to obtain.
Approximate 95% confidence intervals for QTL position can be constructed using the 'one-LOD rule' (Lander and Botstein, 1989). All QTL with a LOD score value less than 1 from the maximum fall within this confidence interval. Note that 1 LOD score corresponds to a LR value of 4.61, which has a significance value of 4% for the χ^2_1-distribution.

LR tests have a χ^2_{df}-distribution, where df refers to the degrees of freedom of the tested parameter (i.e. the difference in df between the full model and the restricted model).

In QTL analysis, this statistic provides only an approximate test, as the null-hypothesis involves a non-mixture distribution whereas the QTL model involves a mixture distribution.
Also regression analysis provide only approximate test statistics, as they assume normal distributed errors within marker type, whereas the distribution is really a mixture of two (or 3).
Simulation studies have been used to examine distributions of test statistics, or to determine threshold values. However, such studies rely on the true data have the same distribution as the simulated data.

Permutation testing

Churchill and Doerge (1994) proposed permutation testing to obtain empirical distributions for test statistics. In a permutation test, the data is randomly shuffled over the marker data. Analysis of the permutated data provides a test statistic, as it is the result of the null-hypothesis (marker not associated with QTL).
The number of permutations required is about 10,000 for a reasonable approximation of threshold levels of 1% (Churchill and Doerge, 1994). The important property of this method is that it does not depend on the distribution of the data. A permutation test is typically used to determine a threshold value for significance testing of the existence of a QTL effect.

Bootstrapping

Bootstrapping, described by Visscher et al., (1996) is a resampling procedure. From the original dataset, N individual observation are drawn *with replacement*. An observation is a phenotype and its marker type, hence unlike in permutation testing, the observed combinations remain together. Note that some observation may appear twice in the bootstrap sample, whereas other may not appear at al. Visscher et al 91996) show that confidence are approximated very well with this method, with only 200 bootstrap samples used. A bootstrap method is typically used to determine an empirical confidence interval for the QTL location, assuming that the QTL effect exists.

Accounting for multiple testing
In QTL analysis, usually many markers are tested, often for multiple traits and in multiple families. The risk of false positives is very high with so many tests. If a 5% significance level would be used, we would expect 5% false positives! Therefore, a more stringent significance level is usually applied for gnome wide QTL detection, e.g. 0.1%.

In general (quoted from Lynch and Walsh, 1998):
If n independent tests with significance level α are conducted, the probability that at least one test is false positive is $\gamma = 1 - (1 - \alpha)^n$.

25 tests with a significance level of 1% would give a probability of 22% to find false positives. It is nearly one for a few hundred tests.

A more stringent level is required (known as the Bonferroni correction):

$$\alpha = 1 - (1 - \gamma)^{1/n} \approx \gamma/n.$$

Hence, for 200 tests we would need a significance level of 0.05/200 = 0.00025 to have a chance of false positives of about 5%. Usually, a significance level of around 0.1% is applied.
However, test statistics from common analysis are usually not valid. Empirical threshold values obtained by permutation testing are more reliable. Permutation testing can also be used to obtain genome-wide significance levels, by simply repeating the procedure across all markers.

References

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

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

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

Kao, C.H. , Z.B. Zheng, and R.D. Teasdale. 1999. Multiple interval mapping for quantitative trait loci. Genetics 152: 1203-1216.

Kearsey, M.J. and V. Hyne. 1994. QTL analysis: a simple ‘marker regression’ approach. Theor. Appl. genet. 698-702.

Lynch, M. and B. Walsh. 1998. Genetics and analysis of quantitative traits. Sinauer Associates Inc. ISBN 0-87893-481-2.

Visscher, P.M., R. Thompson and C.S. Haley. Confidence intervals in QTL mapping by bootstrapping. Genetics 143:1013-1020

Whittaker, J.C., Thompson, R., and P. Visscher. 1996. On the mapping of QTL by regression of phenotype on marker type. Heredity 77:23-32.

Xu, S. 1995. A comment on the simple regression method for interval mapping. Genetics 141:1657-1659.

Zeng, Z-B. 1994. Precision mapping of quantitative trait loci. genetics 136:1457-14.

.