Methods for QTL analysis

Julius van der Werf

Methods for QTL analysis

Julius van der Werf

In this Chapter we will discuss first the principles of mapping based on two markers (interval mapping) and then 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 further classes of methods that are not extensively discussed in this 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. Mixed model methods are based on the Gametic Relationship Matrix, which will be briefly discussed.

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 a next chapter.

Single versus multiple markers

Association between a quantitative trait and genetic markers can be evaluated using single markers or multiple markers. When using one single marker, it is possible to make inference about the segregation of a QTL linked tot that marker. However, in *the case of single markers it is not possible to distinguish between size of a QTL effect and its position (relative to the marker).* Also, single marker analyses have less power if the markers are far apart.

If two (or more) markers are jointly used in an analysis, there is a lot less confounding between the position and size of QTL effect, and there is more power in detecting a QTL, even if the markers are far apart. Inference about the QTL effect as well as the recombination rate between QTL and markers (i.e. position of QTL) is possible. The recombination rate between markers is usually assumed known.
Therefore mapping of a QTL therefore requires the use of multiple marker genotypes in the analysis.

Determining associations between genetic markers and QTL with two markers

For two markers, the QTL probability given the marker genotype depends on more recombinations: those are the recombination rates between M1 and QTL (=r1), between M2 and QTL (=r2) and between M1 and M2 (=r12).
We consider again a half sib design where we know the sires marker genotype for two markers, the sire is heterozygous for the QTL and we know the marker-QTL phase.

TABLE 1

Parental genotype			M1 m1	Q q	M2 m2
Possible gametes			**recombination?**		**Gamete probability**
M1	Q	M2	no		(1-r1)(1-r2)/2
M1	q	M2	double: M1-q, q-M2		r1.r2/2
M1	Q	m2	yes: Q-m2		(1-r1)r2/2
M1	q	m2	yes: M1-q		r1(1-r2)/2
m1	Q	M2	yes: m1-Q		r1(1-r2)/2
m1	q	M2	yes: q-M2		(1-r1)r2/2
m1	Q	m2	double: m1-Q, Q-m2		r1.r2/2
m1	q	m2	no		(1-r1)(1-r2)/2

Assume now also (for simplicity) that we know which marker alleles came from the sire. We can now work out the expected difference between the paternal marker genotype-groups in the sire's progeny:

TABLE 2

Marker alleles obtained from sire group	QTL allele obtained from sire	frequency	Expected mean of progeny
M1M2	Q	(1-r1)(1-r2)/2	$\mu + \alpha$
M1M2	q	r1.r2/2	μ
M1m2	Q	(1-r1)r2/2	$\mu + \alpha$
M1m2	q	r1(1-r2)/2	μ
m1M2	Q	r1(1-r2)/2	$\mu + \alpha$
m1M2	q	(1-r1)r2	μ
m1m2	Q	r1.r2/2	$\mu + \alpha$
m1m2	q	(1-r1)(1-r2)/2	μ

α = average effect of allele substitution of Q (over q).

Some tedious algebra would give the following means for the possible paternal marker-haplotypes in progeny (sum of frequency * mean of group and divide by frequency of marker haplotype group)

TABLE 3. Expected means of different marker haplotypes.

Mean of M1M2-group:	$\frac{\frac{1}{2}(1-r1)(1-r2)(\mu+\alpha)+\frac{1}{2}r1.r2.\mu}{\frac{1}{2}(1-r12)}$	=	$\mu + (1-\frac{r1r2}{1-r12})\alpha$
Mean of M1m2-group:	$\frac{\frac{1}{2}(1-r1).r2.(\mu+\alpha)+\frac{1}{2}r1(1-r2)\mu}{\frac{1}{2}r12}$	=	$\mu + \frac{r2-r1r2}{r12}\alpha$
Mean of m1M2-group:	$\frac{\frac{1}{2}r1(1-r2)(\mu+\alpha)+\frac{1}{2}(1-r1).r2.\mu}{\frac{1}{2}r12}$	=	$\mu + \frac{r1-r1r2}{r12}\alpha$
Mean of m1m2-group:	$\frac{\frac{1}{2}r1.r2(\mu+\alpha)+\frac{1}{2}(1-r1)(1-r2)\mu}{\frac{1}{2}(1-r12)}$	=	$\mu + \frac{r1r2}{1-r12}\alpha$

The difference between the M1M2 and m1m2 haplotypes is now equal to .

$$[\ \mu + (1-\frac{r1r2}{1-r12})\alpha\ ]\ -[\ \mu + \frac{r1r2}{1-r12}\alpha\ ] = (1-\frac{2r1r2}{1-r12})\alpha$$

and as r1r2 is usually a small number, this difference is quite close to the actual QTL allelic effect (α). The coefficient for α in Table 3 in the last column is the probability of having inherited Q from the sire, conditional on (given the) the paternal marker haplotype. This is shown more explicit in Table 4.

TABLE 4. Probabilities for having inherited the paternal Q-allele of different marker haplotypes.

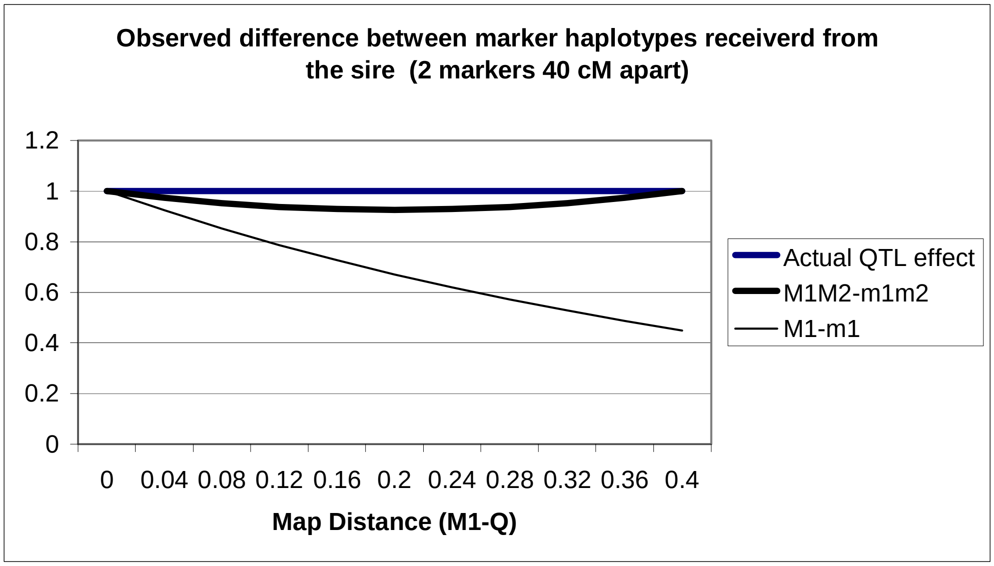
$\mathrm{Prob}(Q \mid M1M2)$	$\frac{\frac{1}{2}(1-r1)(1-r2)}{\frac{1}{2}(1-r12}$	=	$(1-\frac{r1r2}{1-r12})$
$\mathrm{Prob}(Q \mid M1m2)$	$\frac{\frac{1}{2}(1-r1)r2}{\frac{1}{2}r12}$	=	$\frac{r2-r1r2}{r12}$
$\mathrm{Prob}(Q \mid m1M2)$	$\frac{\frac{1}{2}r1(1-r2)}{\frac{1}{2}r12}$	=	$\frac{r1-r1r2}{r12}$
$\mathrm{Prob}(Q \mid m1m2)$	$\frac{\frac{1}{2}r1r2}{\frac{1}{2}(1-r12)}$	=	$\frac{r1r2}{(1-r12)}$

The following Table 5 gives an example of the probabilities of having inherited the Q-allele in a half-sib family, given the marker haplotypes (PQ|MiMj). The distance between the markers is 40 cM. The QTL location investigated is at 10 cM from M1. Haldane's mapping function is used to determine recombination rates based on these distances. Tables 1, 3 and 4 are used to determine probabilities. Table 3 is used to determine expected means of each marker type, assuming QTL genotypic means of 10 and 11 for qq and QQ, respectively. Note that in first instance only the probability of inheriting Q form the sire (given the observed marker types, i.e. P(Q|M1M2)) is relevant. However, in order to predict progeny means, we need to know the alleles contributed by dams, as well as the genetic model, e.g. existence of dominance. The dam population is assumed to have a q-frequency of 1 (comparable with a backcross design)

TABLE 5

Paternal	Probability of marker haplotypes			Qq mean =10	qq mean =9.5	Mean
Markertype	P(M1M2)	P(M1QM2)	P(Q|M1M2)	prob(Qq)	Prob(qq)	Expected
M1M2	0.362	0.352	0.972	0.972	0.028	9.986
M1m2	0.138	0.103	0.745	0.745	0.255	9.873
m1M2	0.138	0.035	0.255	0.255	0.745	9.627
m1m2	0.362	0.010	0.028	0.028	0.972	9.514

The following figure shows the difference between marker haplotype groups in progeny for a single marker (M-m) and for two markers (M1M2-m1m2), for different positions of the QTL relative to the M1.



The figure shows that the difference between the non-recombinant marker haplotypes is much less affected by the marker-QTL distance than the M1-m1 difference for the single markers. Moreover, the map position is now not confounded with the QTL effect. In a way, map position and QTL effect have become estimable with two markers.

The example shown here is based on half sib analysis. The interpretation of the genetic effect estimated depends on the constitution of the dam population, as shown in the previous chapter. If we want to estimate both a and d, we need a dam population that contributes both q and Q alleles, and where we can trace the inheritance from the dam. In other words, we need to identify also segregation from the dam. Choosing the dam population from a F1-cross of two extreme lines (extreme with respect to the putative QTL) would be the best choice.

Inbred lines have been used in QTL mapping to avoid uncertainty about the genetic effects estimated. However, in animal population, complete inbred lines (with marker- and QTL alleles fixated) are hardly feasible, and possibly less relevant for QTL's to be used in practical applications.

In outbred populations, there is less certainty about the animals' QTL genotypes. Lack of design usually means that the marker genotypes are frequently not informative about paternal or maternal origin. In the next chapter, the advantages and disadvantages of different design will be discussed.

At this stage we can continue that for 'any' design, the QTL estimation is based on two steps

1) What is the probability that an individual has a certain QTL genotype (give the observed marker genotypes)
2) What is the estimated effect of this particular genotype on the individuals' phenotypes

The first step is much easier in well-defined experiments. The second step can be quantified either by using the likelihood principle, or by using regression (where the match is measured in terms of residual sums of squares).

We present the principle briefly here, and thereafter we will discuss in more detail these different methods.

Interval mapping

<u>Maximum Likelihood</u>

The term 'interval mapping' is used for estimating the position of a QTL within two markers (often indicated as 'marker-bracket'). Interval mapping is originally based on the maximum likelihood but there are also very good approximations possible with simple regression.

The principle is:

1) The Likelihood can be calculated for a given set of parameters (particularly QTL-effect and QTL position) given the observed data on phenotypes and marker genotypes.

2) The estimates for the parameters are those were the likelihood are highest.

3) The significance can be tested with a likelihood ratio test:

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

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

Using the log-likelihood: $\text{LR} = -2.(\ln\_L_{\text{reduced}} - \ln\_L_{\text{full}})$

where ln_L is the $\log_e$ of the maximum likelihood.

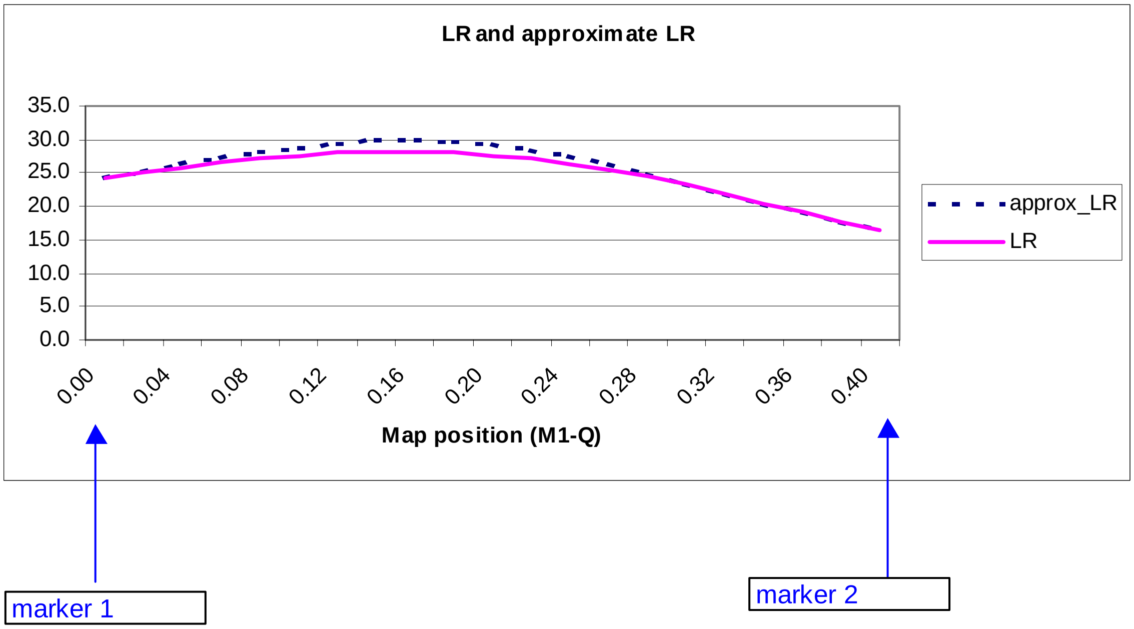
The evidence for a particular QTL at a particular chromosomal position can be displayed as a *likelihood map*, The LR-statistic is plotted against the map position of the QTL.

Lander and Botstein (1989) introduced first the concept of likelihood maps. The proposed to use the LOD-score as a test statistic. However, the LOD score is equal to a constant (1/4.61) time the LR test statistic, as shown:

The LOD score for a QTL at position c is:

$$\text{LOD(c)} = -\ \log_{10} \frac{Max\_Likelihood(reduced\,\text{mod}\,el)}{Max\_Likelihood(\,full\,\text{mod}\,el, c)} = \frac{LR(c)}{2\ln 10} \approx \frac{LR(c)}{4.61}$$

The following figure shows a likelihood map for a marker bracket based on simulated data from one half sib family (backcross) with 300 progeny. The simulated QTL effect was 0.5 within-family standard deviations. The figure shows the true LR value based on ML, and the approximate LR (upper line) based on regression analysis.



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. We illustrate the method for two marker, which is therefore like interval mapping.
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 for a given QTL position as in Table 4. Note that different positions give 4 different x values for the 4 haplotypes. 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.

For each recorded animal, we can then give a predicted phenotype with this “QTL-model” which is equal to

$$\hat{y}_i = \hat{\mu} + \hat{a}.x_i$$

where the “hats” refer to estimated (predicted) values.

A model ignoring a QTL would predict each observation as

$$\hat{y} = \hat{\mu}_0$$

where $\hat{\mu}_0$ is typically the general progeny mean

Now let the total sum of squares (SST) be the sum (over animals) of $(\hat{y} - \hat{\mu}_0)^2$

and let the residual sum of squares (SSE) be the sum (over animals) of $(\hat{y}_i - \hat{\mu} - \hat{a}.x_i)^2$

Each map position will yield an SSE and the position with the lowest SSE is the most likely position.

A test statistic for this method is for an experiment with n observations is

$$\text{LR} = n \ln(\frac{\text{SST}}{\text{SSE}})$$

where n is equal to the number of observations. The LR stands for “Likelihood Ratio”, as this test statistic is approximately similar to the LR from maximum likelihood.
Haley and Knott (1992) have shown that this similarity. If there are more fixed effects in the model, the test statistic is calculated as

$$\text{LR} = n\ln(\frac{SSE_{reduced}}{SSE_{full}})$$

Which is ratio of the residual sums of squares in a model with the QTL (”full’) and a model without it (‘reduced’).

The information about a QTL is only dependent on the flanking markers. If the QTL lays outside the bracket, it will only depend the nearest marker. Likelihood maps can be constructed for neighbouring marker brackets and they should exactly match up at each marker, and a map of multiple intervals M1-M2-M3....-Mk is smooth.

Example of QTL mapping by regression:

Data on 8 individuals with paternal marker haplotypes given. The probabilities are derived for different positions (dM1-Q is distance between marker 1 and QTL), with further the same assumptions as in this chapter (see Table 5).

```
[     X         ]     yhat       y           markers

1.0000    1.0000   50.3656   50.9813         M1M2
1.0000    1.0000   50.3656   49.9813         M1M2
1.0000    1.0000   50.3656   50.7500         M1m2
1.0000    1.0000   50.3656   49.7500         M1m2
1.0000         0   50.1344   50.7500         m1M2
1.0000         0   50.1344   49.7500         m1M2
1.0000         0   50.1344   50.5187         m1m1
1.0000         0   50.1344   49.5187         m1m1

 dM1-Q    SST       SSE       LR

    0    2.2139    2.1070    0.3961
```

```
[     X         ]     yhat       y
1.0000    0.9718   50.4321   50.9813
1.0000    0.9718   50.4321   49.9813
1.0000    0.7451   50.3446   50.7500
1.0000    0.7451   50.3446   49.7500
1.0000    0.2549   50.1554   50.7500
1.0000    0.2549   50.1554   49.7500
1.0000    0.0282   50.0679   50.5187
1.0000    0.0282   50.0679   49.5187

 dM1-Q    SST       SSE       LR

  0.1   2.2139     2.0455    0.6331
```

```
   [   X      ]     yhat       y

1.0000    0.9718   50.4321   50.9813
1.0000    0.9718   50.4321   49.9813
1.0000    0.2549   50.1554   50.7500
1.0000    0.2549   50.1554   49.7500
1.0000    0.7451   50.3446   50.7500
1.0000    0.7451   50.3446   49.7500
1.0000    0.0282   50.0679   50.5187
1.0000    0.0282   50.0679   49.5187


 dM1-Q    SST       SSE       LR

 0.3       2.2139    2.0455    0.6331
```

```
 [     X         ]     yhat       y
1.0000    0.9625   50.4813   50.9813
1.0000    0.9625   50.4813   49.9813
1.0000    0.5000   50.2500   50.7500
1.0000    0.5000   50.2500   49.7500
1.0000    0.5000   50.2500   50.7500
1.0000    0.5000   50.2500   49.7500
1.0000    0.0375   50.0187   50.5187
1.0000    0.0375   50.0187   49.5187

  dM1-Q    SST       SSE       LR


  0.2    2.2139    2.0000    0.8129
```

```
1.0000    1.0000   50.3656   50.9813
1.0000    1.0000   50.3656   49.9813
1.0000         0   50.1344   50.7500
1.0000         0   50.1344   49.7500
1.0000    1.0000   50.3656   50.7500
1.0000    1.0000   50.3656   49.7500
1.0000         0   50.1344   50.5187
1.0000         0   50.1344   49.5187


    dM1-Q    SST        SSE        LR


   0.4   2.2139    2.1070    0.3961
```

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’:

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

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)

$$r_1 = 0.5\left[1 - \sqrt{1 - \frac{4\beta_2\theta(1-\theta)}{\beta_2 + \beta_1(1-2\theta)}}\right]$$

and the estimate of the QTL effect:

$$\alpha = \sqrt{\frac{[\beta_1 + (1-2\theta)\beta_2][[\beta_2 + (1-2\theta)\beta_1}{1-2\theta}}$$

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(θ| 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

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

Also for an alternative model, e.g. with a QTL effect, we may have different means. A new set of parameters is then ($\mu_{1,}$ μ_2, α, and σ) and we can write the likelihood.

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

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(μ_I)).

The distribution of the data under the interval mapping model is generally assumed to follow a normal mixture distribution. For example, there are two normal distributions, each of which is for a QTL paternal haplotype. The iterative expectation-maximization (EM) algorithm is broadly used to calculate the maximum-likelihood estimates (MLE) of the normal mixture model. The ML mapping model used here for locating a QTL (Q) on an interval flanked by markers i and i + 1 (M_i and M_{i+1}) (assuming the order M_iQM_{i+1}) can be written as:

$$y_j = u + b\, x_j + e_j \qquad j = 1, 2, \ldots, n$$

where

y_j = the trait value of the jth individual

μ = mean

b = the effect of the putative QTL

$$x_j = \begin{cases} 1 & \textit{if the Q is inherited} \\ 0 & \textit{if the q is inherited} \end{cases}$$

$$e_j \sim N(0, \sigma^2)$$

x_j, which is unobserved, can take different values with probabilities depending on the genotype of the flanking markers (M_i, M_{i+1}) of the jth individual and the testing position. Here, the QTL genotype for an individual is usually missing value, but its distribution can be inferred from its flanking marker genotypes. Since the x_j could be 1 or 0, the likelihood for every position is then a normal mixture distribution with mixing proportions equivalent to the probabilities of having inherited paternal alleles Q and q from a heterozygous (Qq) sire, p_j and 1 - p_j. As in [3] but now for n individuals in a population, the likelihood function of the model is:

$$L(\mu, b, \sigma^2) = \prod_{j=1}^{n} \left[p_j\ \phi\left(\frac{y_j - \mu - b}{\sigma}\right) + (1 - p_j)\ \phi\left(\frac{y_j - \mu}{\sigma}\right) \right]$$

where $\phi(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-z^2/2}$ is the standard normal density function. It can be seen that the likelihood depends on p_j, which is determined by the QTL position and the marker genotypes, as well as phenotypic data (y_j). The maximum likelihood estimates of the parameters b, $\mu$and σ^2 are derived based on the EM algorithm, that is by iterating the following equations and beginning with the random starting values of each parameter. The EM approach proceeds as follows:

(1) **E-step**: we write down expected values for genotype probabilities (the ‘missing values’), given current estimates of the parameters:

$$P_j^{(k)} = \frac{p_j \phi\,(\frac{y_j - \mu^{(k)} - b^{(k)}}{\sigma^{(k)}})}{p_j\ \phi\,(\frac{y_j - \mu^{(k)} - b^{(k)}}{\sigma^{(k)}}) + (1 - p_j)\ \phi(\frac{y_j - \mu^{(k)}}{\sigma^{(k)}})} \qquad (5.1)$$

where k refers to iteration round and $P_j^{(k)}$ can be called the posterior probability of $x_j = 1$.

(2) **M-step**: estimate parameters given these probabilities

$$b^{(k+1)} = \sum_{j=1}^{n} (y_j - \mu^{(k)}) P_j^{(k)} / \sum_{j=1}^{n} P_j^{(k)} \qquad (5.2)$$

$$\mu^{(k+1)} = \sum_{j=1}^{n} (y_j - P_j^{(k)} b^{(k+1)}) / n \qquad (5.3)$$

$$\sigma^{(k+1)^2} = \frac{1}{n} \sum_{j=1}^{n} [\ (y_j - \mu^{(k)})^2 - P_j^{(k)} b^{(k)^2}\ ] \qquad (5.4)$$

From the above four equations, it can be seen that each parameter depends on estimates of other parameters. Therefore, in each iteration, the algorithm consists of one E-step, equation (5.1), and three M-steps, equations (5.2), (5.3) and (5.4). This process is iterated until convergence of estimates.

A *test of significance* is obtained by comparing the maximum likelihood with the likelihood of a model with the tested parameter omitted (reduced model).

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

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

<u>Model with no QTL:</u>

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

<u>Model with a QTL</u>

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

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

Multi-family analysis

With more families in an outbred population, the phase maybe different in different families, or the sires may be homozygote for the QTL. The QTL analysis in multi-family was performed according the following model [Kerr, 2000]:

$$y_{ij} = \mu_i + (Z_{111,ij} + Z_{122,ij})\, b – (Z_{112,ij} + Z_{121,ij})b + e_{ij}$$

where y_{ij} is the corrected phenotype of the jth progeny of the ith sire, μ_i is the mean of the ith sire family, b is the magnitude of the effect of QTL allele inherited from the sire, and e_{ij} is the random error term. Compared with the b in previous section, b in this model is only one half of the previous b because it is considered that the QTL effect is +b if the progeny inherited the Q allele and the QTL effect is -b if the progeny inherited the q allele in this section. (IN the previous section the QTL effect was either 0 if the progeny inherited the q allele or b if it inherited the b allele). The variables $Z_{111,ij}$, $Z_{122,ij}$ etc. are indicator variables taking the value 1 or 0 with the probability depending on the unknown probability that sire is heterozygous (h), the probability that the sire has one of two equally likely possible linkage phases, the genotypes of the flanking markers and the position being tested, where

$Z_{1,i}$ = 1, if the ith sire is heterozygous

$Z_{2,i}$ = 1, if the ith sire is homozygous

$Z_{111,ij}$ = 1, if the ith sire is heterozygous, has phase 1 and its jth progeny has inherited the Q allele

$Z_{112,ij}$ = 1, if the ith sire is heterozygous, has phase 1 and its jth progeny has inherited q allele

$Z_{121,ij}$ = 1, if the ith sire is heterozygous, has phase 2 and its jth progeny has inherited the q allele

$Z_{122,ij}$ = 1, if the ith sire is heterozygous, has phase 2 and its jth progeny has inherited the Q allele

When one progeny has been assumed to be one type of $Z_{111,ij}$, $Z_{112,ij}$, $Z_{121,ij}$ and $Z_{122,ij}$, this type equals 1 and the other three types are all 0.

Denoting all μ_i by the vector μ, the number of sires by n_s and the number of progeny in each sire family by n_i, the likelihood function for this model is [Knott, 1996 #43]:

$$L(h, \mu, b, \sigma^2) = \prod_{i=1}^{n_s} \left\{ \begin{array}{l} .5h\prod_{j=1}^{n_i}\left[p_{ij}\phi\left(\frac{y_{ij}-\mu_i-b}{\sigma}\right)+\left(1-p_{ij}\right)\phi\left(\frac{y_{ij}-\mu_i+b}{\sigma}\right)\right] \\ +.5h\prod_{j=1}^{n_i}\left[p_{ij}\phi\left(\frac{y_{ij}-\mu_i+b}{\sigma}\right)+\left(1-p_{ij}\right)\phi\left(\frac{y_{ij}-\mu_i-b}{\sigma}\right)\right] \\ +(1-h)\prod_{j=1}^{n_i}\phi\left(\frac{y_{ij}-\mu_i}{\sigma}\right) \end{array} \right\}$$

where p_{ij} is the specific prior probabilities that the progeny has inherited the Q allele, conditional on the genotypes of flanking markers i and i + 1 and the position being tested, 1 - p_{ij} is the prior probability that the progeny has inherited the q allele, $\phi(z)$ is the probability density for the normal distribution, and h is the probability if sire is heterozygous. Solutions again obtain using an EM algorithm (see Kerr 2000 or Song 2003).

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

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

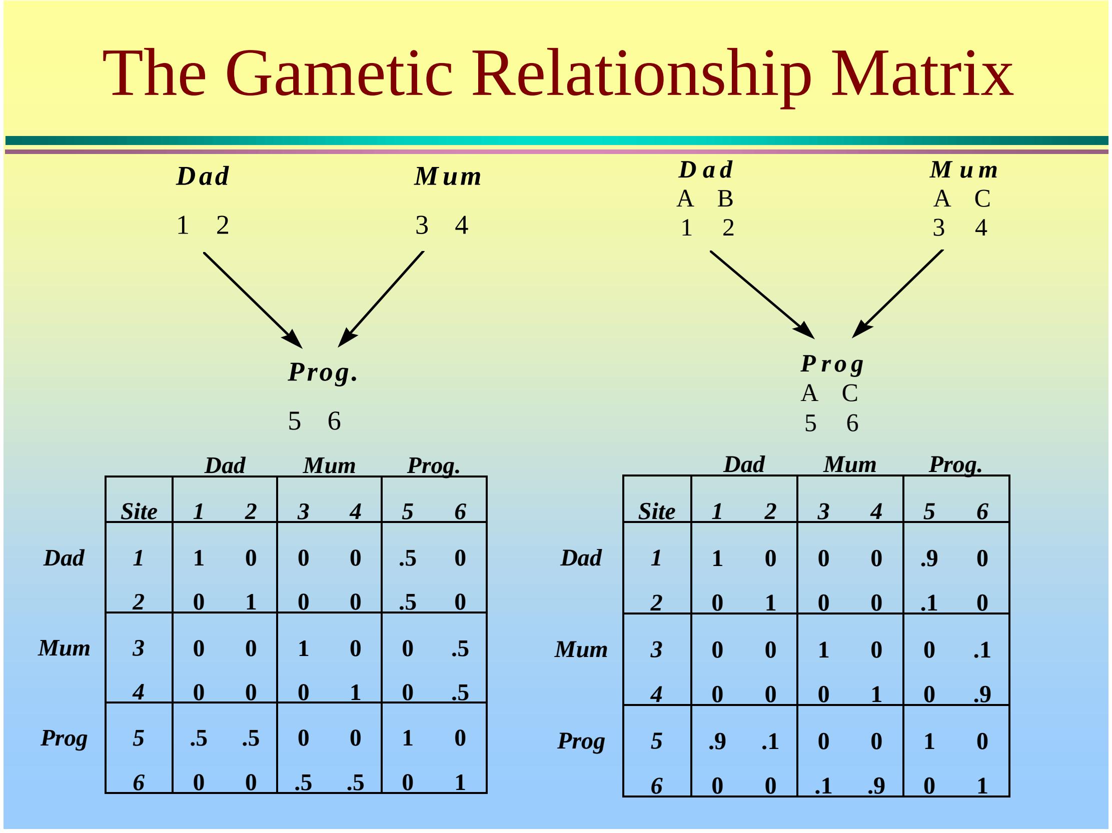
where p_i is the probability of having inherited Q in marker genotype class i and *a* is the regression coefficient on Q-probability in the regression model. Generally, this adjustment has only a small effect, unless the QTL effect is very large and markers are far from the QTL position

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 next).

The Gametic Relationship Matrix approach

This can be the most complete approach containing all information on pedigree in an outbred population. In this approach, we first set up a GRM; a symmetrical matrix that contains a row and column for each gametic haplotype (2 per animal, one from each parent) in the population of animals that we have. Such a matrix is specific to the chromosomal region of current interest. Each element in this matrix is then the probability of identity-by-descent for the representations of this region (one representation per gamete). Here are simple examples of this “Gametic Relationship Matrix” (GRM). Notice that without marker information we must resort to simple segregation probabilities – however, marker information allows us to be more ‘surgical’ in allocating identity-by-descent probabilities:



The Gametic Relationship Matrix

Dad

1 2

Mum

3 4

Prog.

5 6

Dad

A B

1 2

Mum

A C

3 4

Prog

A C

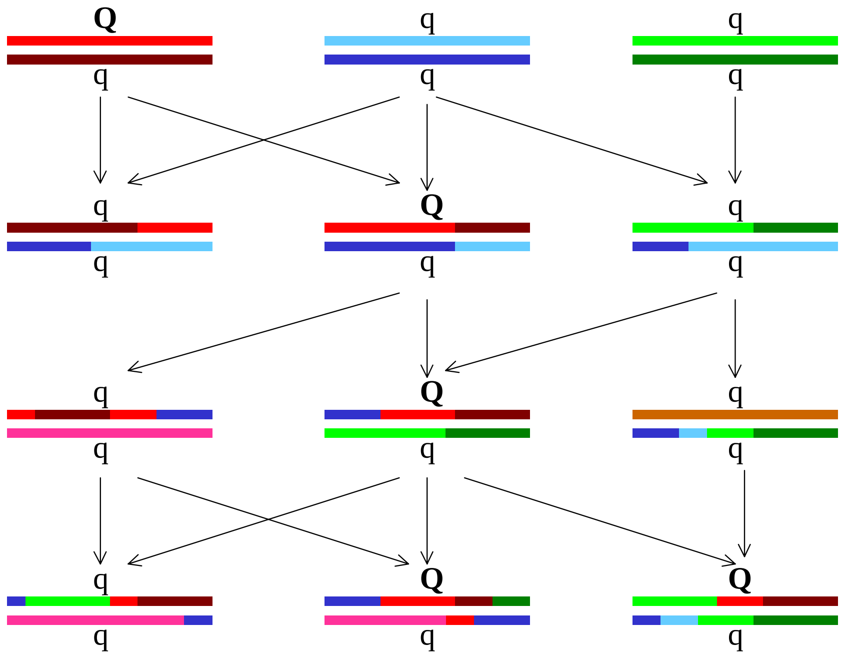
5 6

		Dad		Mum		Prog.	
	Site	*1*	*2*	*3*	*4*	*5*	*6*
Dad	*1*	1	0	0	0	.5	0
	2	0	1	0	0	.5	0
Mum	*3*	0	0	1	0	0	.5
	4	0	0	0	1	0	.5
Prog	*5*	.5	.5	0	0	1	0
	6	0	0	.5	.5	0	1

		Dad		Mum		Prog.	
	Site	*1*	*2*	*3*	*4*	*5*	*6*
Dad	*1*	1	0	0	0	.9	0
	2	0	1	0	0	.1	0
Mum	*3*	0	0	1	0	0	.1
	4	0	0	0	1	0	.9
Prog	*5*	.9	.1	0	0	1	0
	6	0	0	.1	.9	0	1

Figure. Gametic relationship matrices (GRM) for a QTL are of dimension 6 sites x 6 sites for the simple 3-animal pedigree shown. Elements of the GRM are probability of identity by descent of the alleles at the prevailing pair of sites. In the GRM to the left, no marker information is available, and, for example, probability of identity by descent between sites 4 and 6 is 0.5, as site 6 (maternal) could have inherited from sites 3 or 4 with equal probability. In the GRM to the right, a marker with alleles A, B and C is available, and for example, probability of identity by descent between sites 4 and 6 is 1, for the marker locus. If the QTL is linked with a recombination fraction of 0.1, then the probability of identity by descent between sites 4 and 6 is 0.9, for the QTL, with a 0.1 probability (in the event of recombination) for sites 3 and 6. Special attention is required where there is ambiguity of marker allele inheritance (Wang et al., 1995).

With a good data set, the GRM gives us a lot of information for mapping. You can visualise regions of identity-by-descent in the following diagram. In the diagram, the top-left founder animal has QTL allele Q in its paternally inherited region of haplotype (coloured red). For all its descendants, the GRM gives us probabilities that the have inherited the same bit of DNA, holding that Q allele. We can then simple ***regress*** their phenotypes on these probabilities to get an estimate of the effect of Q on phenotype.



The strategy is to construct a GRM (or a subset of it) for each location in the genome, and test the goodness of fit of the resulting regression. We end up with something like this for each chromosome:

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_{df}^2-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 (1996) 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.

Methods for detecting multiple interacting QTL

Julius van der Werf and Brian Kinghorn

Introduction

In the last lecture you found out about regression and maximum likelihood methods for detecting QTL. The extension to cater for multiple interacting QTL is best illustrated on the basis of analysis by regression.

The strategy is to construct a GRM (or a subset of it) for each location in the genome, and test the goodness of fit of the resulting regression. We end up with something like this for each chromosome:

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.

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

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 = p(QTL1| M1M2) + p(QTL2| other markers near QTL2)		[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.

Detecting multiple interacting QTL

In Composite Interval Mapping, once we are happy about the most likely position and effect of a QTL, we fix that in the analysis – we correct all the animal phenotypes for the most likely impact of that QTL on their performance – and then repeat the process to look for another QTL.

This has two problems:

- The estimated position of the first QTL can be influenced by the second QTL, and vice-versa. This is especially dangerous for linked QTL. A method to simultaneously locate the two QTL is preferable.

- Life is complex – and that means that genes (or gene products) interact with each other to produce the organisms that we all are. The value of a particular gene variant will differ between genetic backgrounds. In some cases it will be the weak link to achieving high merit, and in others it will not. This means that we should ideally look for interacting sets of genes. Otherwise we could miss some important genes – and opportunities to exploit them.

A more general approach is proposed by Carlborg et al. (2000).

We can nominate two separate positions in the genome as candidate locations for two QTL. We can then construct a GRM for each position, and carry out a 2-locus regression, as outlined below, fitting interaction effects between the two loci, as well as additive and dominance effects within each locus.

How can we find the best fitting two positions? Carlborg et al. (2000).demonstrate an approach that works efficiently, using a genetic algorithm The genetic algorithm (GA) works by "breeding" the best solution to the prevailing mathematical problem. In this case, the "DNA" that the GA uses is simply the candidate positions for the two (or more) QTL. Each of these is a candidate solution to the problem of QTL locations. Each candidate solution competes to become a "parent" in the next generation. They compete on a criterion that is simply the goodness of fit of these positions to the phenotypic data and pedigree on hand.

The successful "parent" solutions then combine in some way – exchanging information, and mutate to some extent, to generate a new generation of candidate solutions.

Model for fitting interacting QTL

Here is a simple one-locus model of genetic effects, similar to that found in all texts in this area. *II*, *Ii* and *ii* are the genotype values for combinations of the two alleles *I* and *i*, μ is a general mean, A_i is the additive affect and D_i the dominance effect at locus *i*.

$$\begin{pmatrix} II \\ Ii \\ ii \end{pmatrix} = \begin{pmatrix} \mu + A_i \\ \mu + D_i \\ \mu - A_i \end{pmatrix}$$

We can now expand this to cater for effects at two loci. The classical statistical approach (eg. Jana 1971) is typified as follows:

$$\begin{pmatrix} II\,JJ & II\,Jj & II\,jj \\ Ii\,JJ & Ii\,Jj & Ii\,jj \\ ii\,JJ & ii\,Jj & ii\,jj \end{pmatrix} = \begin{pmatrix} \mu + A_i + A_j + AA_{ij} & \mu + A_i + D_j + AD_{ij} & \mu + A_i - A_j - AA_{ij} \\ \mu + D_i + A_j + AD_{ji} & \mu + D_i + D_j + DD_{ij} & \mu + D_i - A_j - AD_{ji} \\ \mu - A_i + A_j - AA_{ij} & \mu - A_i + D_j - AD_{ij} & \mu - A_i - A_j + AA_{ij} \end{pmatrix}$$

The number of parameters to handle has increased from three (μ, A_i and D_i) to nine (μ, A_i, A_j, D_i, D_j, plus interaction terms AA_{ij}, AD_{ij}, AD_{ji}, and DD_{ij}). Notice that each locus here has two alleles.

More detail is here extracted from Carlborg et al. (2000):

“

The objective function used was the residual sum of squared errors from a weighted least squares approach to QTL mapping. The method is the extension of the method of Jansen (1992) to the two-loci linear model G = m+A1 +A2+D1 +D2 +AA12 +AD12 +AD21 +DD22 as indicated by the author. The parameters of the model will be explained below. Markers have not been used as cofactors and successive iterations in the EM algorithm have been removed to increase the computational efficiency during the evaluation procedure. The modifications needed to the single QTL mapping procedure described by Jansen and Stam (1994) when implementing the two QTL model included duplication of each individual nine times (instead of three times i.e. once for every possible two-QTL genotype) and the use of an expanded design matrix (X). The design matrix for the two-locus linear model has been described by Jana (1971). The weight for each observation was taken to be the product of the conditional probabilities of the single QTL-genotypes given the markers (Haley and Knott 1992) at each of the two fitted QTL. The estimates of the model parameters can be found as:

$\beta = (X^{T} WX)^{-1} X^{T} WY$

$\sigma^2 = (1/N)(Y - X\,\beta)^T\, W(Y - X\,\beta)$

where Y is the complete data vector, X is the design matrix for the complete data, W is the diagonal matrix of weights, β is the vector of the regression parameters, σ^2 is the normal variance and N is the number of individuals (Jansen and Stam 1994).

The residual sums of squared errors can then be calculated as:

$SSE = (Y - X\,\beta)^T\, W(Y - X\,\beta)$

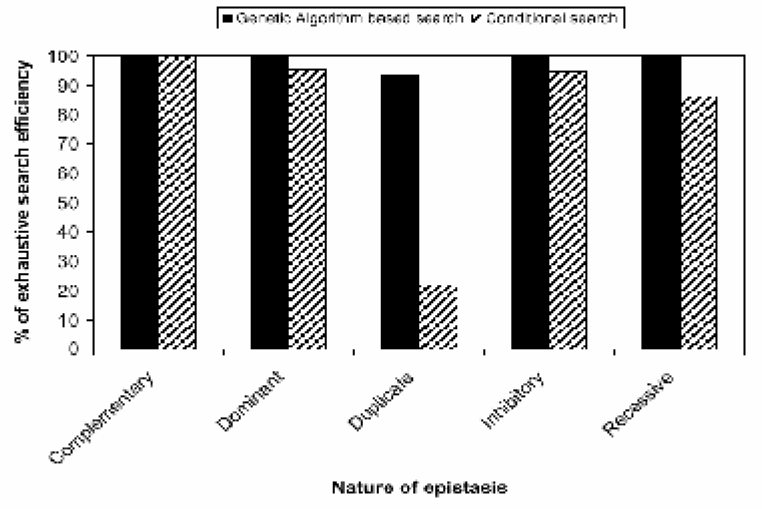
The method described above can easily be extended to take account of background QTL in the analysis. Two extra ga-genes are added to the genetic algorithm and two extra columns are added to the X matrix for each background QTL. The extra ga-genes represent the chromosomal location for the QTL and the columns in the design matrix are to contain the QTL indicator variables a and d (Haley and Knott 1992), for a QTL at the location given by the ga-genes. The rest of the evaluation procedure is the same as before. We have evaluated the increase in computational demand for a simultaneous search for more than two QTL using this method, but have not investigated any other properties. "

Some results

There are two advantages in this approach:

- The genetic algorithm gives a fast search, saving much computer time. It increases the computational demand by a factor of 3 to 5 when compared to the conditional search (Carlborg et al. 2000). The improvement in computational efficiency of the GA as compared to an exhaustive enumerative search (looking at all pairs of locations in a genome size of 2,000 cM using a resolution of 1 cM) was by a factor 133 for two QTL. An expansion of the search to additional dimensions by also searching for background QTL simultaneously leads to further computational advantages for the GA based search. Improvements are in the order of 65,000 for three QTL and 1.7 x 10 7 for four simultaneously fitted QTL.

- As Carlborg et al. (2000) report, the results from the simulation study with 18 QTL (Figure below) showed that the genetic algorithm based search had higher relative efficiency to detect the simulated pair of epistatically interacting QTL than the conditional search (ie. composite mapping approach, as described above) for all epistatic models tested. The genetic algorithm had a relative efficiency of 100% for all epistatic models except for the duplicate. The conditional search had between 86 and 96% relative efficiency for the dominant, recessive and inhibitory epistatic models and 100% relative efficiency for the complementary model. The difference in relative efficiency for the search methods was very large for the duplicate epistatic model, where the conditional search only had a relative efficiency of 21%, while the genetic algorithm based search had a relative efficiency of 93% (this could grow to 100% with better tuning of the GA parameters). In the simulation where two interacting QTL explained all genetic variation, both methods had a relative efficiency of 100%.



As Carlborg et al. note: “The genetic algorithm is a general tool to search large parameter spaces and could be of use in many other areas in QTL mapping. In this study we have used a genetic algorithm in the search for two interacting QTL in a cross between inbred lines, but the method can also be used for analyses of crosses between outbred lines and in searches for more than two QTL. For analyses of outbred lines, the genetic algorithm could also be used when testing for QTL segregation within the founder lines. This would be implemented by using a genetic algorithm to group the haplotypes from the founders in allelic groups and in this way obtain the most likely allelic constitution for the founders and other individuals in the pedigree. This results in greater detection power because of more extreme probabilities of identity-by-descent of chromosomal regions between phenotyped individuals and each founder.“

Multiple trait mapping of QTL

Jiang and Zeng (1995) have proposed a multiple trait version of the composite interval mapping. Their method is based on maximum likelihood, and requires special programs for analysis. The authors should considerable increase in power when using information from two correlated traits.

Most QTL detection studies comprise phenotypic data on multiple traits. Joint use of data from multiple traits in QTL analysis has two advantages: increased power and testing of models regarding the genetic correlation between two traits.

Increased power of QTL detection

Multiple traits that are correlated can add information to each other. To some extent, two measurements on correlated traits are somewhat like repeated measurements. Therefore, information from correlated traits can reduce the effect of error variance, therefore making it easier (more powerful) to detect QTL. Not only the power of QTL detection is increased, also the precision of the QTL map position is better.

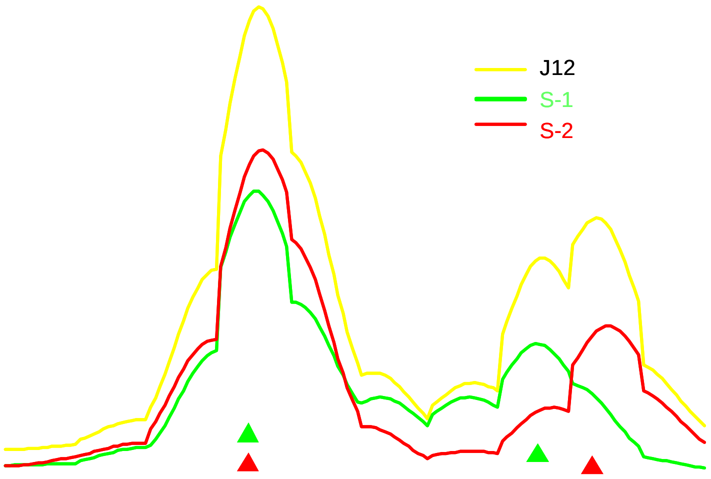


Illustration of increased power from using joint analysis of two traits (J12) over single trait analysis (S1 and S2)

Jiang and Zheng, 1995

Jiang and Zeng (1995) also discussed the increased power from multiple trait analyses in relation to the correlation structure.

In summary:

1. If the correlation between the traits (here: correlation between residual effects, this could be the sum of residual and polygenic effects) is zero, the joint test statistic is approximately the sum of the test statistics for the single traits

$$LR_j \approx LR_{S1} + LR_{S2} \qquad \text{if correlation} = 0$$

2. If the QTL is only affecting one of the two traits, say $\alpha_2 = 0$, then a joint analysis can increase the test statistic of detecting that trait, depending on the correlation (r) between the two traits.

$$LR_j \approx LR_{S1}/(1\text{-}r)^2 \geq LR_{S2}$$

3. The joint test statistic is equal or greater than the maximum of the single trait statistics.

$$LR_j \geq \text{maximum}[LR_{S1},LR_{S2}]$$

4. $r\ \alpha_1\ \alpha_2 < 0$ (i.e r and $\alpha_1\ \alpha_2$ have different signs)

$$LR_j > LR_{S1} + LR_{S2}$$

This is the most favourable situation for using multiple traits analysis.

Testing for linked QTL vs pleiotropic QTL

When two QTL are found in the same region, when using single trait analysis, the question arises whether these are actually the same genes affecting both traits, or whether these are two separate QTL.

Unravelling this difference allows to better understand the nature of a genetic correlation between two traits. This would provide information concerning the possibility to break a unfavourable genetic correlation between two characters (in the case of linkage) or whether this is impossible (in the case of pleiotropism).

The test can be carried out with H_0: position 1 = position 2

H_1: position 1 $\neq$ position 2

Also other genetic models could be compared and tested (depending on design)

- Existence of epistasis (see Chapter 10)
- QTLs effecting one trait only vs effect on both traits

Maximum likelihood might be a bit laborious for multiple trait analyses, especially when comparing a range of genetic models.

Multiple trait analysis using regression

Moser (2000) has proposed a multiple trait regression approach and showed that again regression is very similar to maximum likelihood methods (at least in designed experiments).

As in single trait analysis, the approximate LR $\approx$ n $ln(SSE_{reduced} / SSE_{full})$

Moser proposes to use for a multiple trait analysis

$$LR \approx n\, \mathit{ln}(\ |VE_{reduced}|\ /\ |VE_{full}|\ )$$

i.e. rather than the sum of squares of errors of a single trait analysis, he used the determinant of the matrix with residual sum of squares and sum of cross products of errors for two traits.

The advantage of the simple multiple trait regression method is that

1) permutation tests are feasible
2) a number of genetic models can bet tested and compared

Moser (2000) used a genetic algorithm to efficiently find the most likely genetic model (as described in the previous chapter).

Multiple trait analysis using logistic regression

Henshall and Goddard (1999) proposed to use logistic regression for multiple trait QTL mapping. In fact, this method is also very useful for single trait analysis.

Logistic regression is used for traits where the response variable has a binomial distribution. Henshall and Goddard (1999) regressed, within half sib families, QTL genotype on phenotype. The QTL genotype refers to which QTL allele was received from the heterozygous sire (either Q or q). This is a 0/1 response with a probability, hence binomially distributed. Hence, rather than comparing phenotypic means for different marker genotype classes, they compared marker genotype classes for different phenotypes.

The main advantages of this method:

1) It is much simpler than maximum likelihood and standard software (like SAS) can be used, even for multi trait analyses. Maximum likelihood methods would be much more complex, as all data that was used in selection would have to be included in the analysis. Logistic regression however, is nearly equivalent than ML.

Example: analysis of the traits Y and Z would require in SAS

```
proc logistic;
     model Q/n = Y Z
run;
```

The variable Q is the marker genotype (0 or 1) and n is the number of trials for each observation (=1)

2) The phenotypic observations can be subject to selection (as regression is not affected by regression on the 'x-variable'. Hence, logistic regression is a simple method that is applicable to data obtained from selective genotyping.

The principle of the method is as follows:

Let p = P(Q), i.e. probability of having inherited the Q-allele from the sire and assume genotype means of $\mu + \alpha$ and $\mu - \alpha$ for genotypes Q- and q- resp.

.

In single trait analysis, the logistic regression model is written as:

$$\log(\frac{p}{1-p}) = a + by$$

The QTL effect can be calculated as $\alpha = \frac{-1+\sqrt{1+b^2\sigma^2}}{b}$

where σ^2 is the sum of the residual variance σ_e^2 and the QRTL variance = $\alpha^{2.}$

In multiple trait analysis, the model is: $\log(\frac{p}{1-p}) = Y'\beta$

where Y and β are vectors. The vector of QTL effects is

$A = \dfrac{\sum \beta}{1 + \sqrt{\beta' \sum \beta + 1}}$ where Σ = V * AA' is the sum of the residual covariance matrix and the QTL covariance matrix.

If there is no recombination between marker and QTL, we can observe p. However, in case of recombination (r), p depends on r.

We can observe p if the marker is at the QTL (no recombination). Henshall and Goddard (1999) suggest that in case of recombination, the vector β can be estimated at each marker (as if it was the QTL), and the estimate for β at any position between two markers is obtained by linear interpolation. They also show how the log-likelihood can be calculated for any position of a QTL between two marked loci.

Acknowledgement: Thanks to Örjan Carlborg*, Leif Andersson* and Brian Kinghorn for permission to use direct quotation. Jie Song os acknowledged for summarizing the ML estimation procedure.

References

Carlborg, O., Andersson, L. and Kinghorn, B.P. 2000. The use of a genetic algorithm for simultaneous mapping of multiple interacting quantitative trait loci. *Genetics 2000 155: 2003-2010*

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

Henshall, J.M. and M.E. Goddard. 1999. Multiple trait mapping of quantitative trait loci after selective genotyping using logistic regression. Genetics 151:885-894.

Jana, S., 1971 Simulation of quantitative characters from qualitatively acting genes. I. Nonallelic gene interactions involving two or three loci. Theor. Appl. Genet. 41: 216-226.

Jansen, R. C., 1992 A general mixture model for mapping quantitative trait loci by using molecular markers. Theor. Appl. Genet. 85: 252-260.

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

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

Jiang, C., and Z-B. Zeng. 1995. Multiple trait analysis of genetic mapping for quantitative trait loci. Genetics 140:1111-1117.

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.

Kerr, 2001. AAABG p. 409

Kinghorn, B.P. and Clarke B. E. 1997. Genetic evaluation at individual QTL. Animal Biotechnology, 8:63-68.

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

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

Moser, G. and van der Werf 2001 AAABG 405. .

Song, J. 2003. MSc thesis, UNE.

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

Wang, T., Fernando, R.L., van der Beek, S., Grossman, M. and van Arendonk, J.A.M. 1995. Covariance between relatives for a marked quantitative trait locus. Gen. Sel. Evol. **27**:251-274.

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.

Wang, T., Fernando, R.L., van der Beek, S., Grossman, M. and van Arendonk, J.A.M. 1995. Covariance between relatives for a marked quantitative trait locus. Gen. Sel. Evol. **27**:251-274.

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

Exercise 1 Inheritance probabilities with two markers

Consider two markers that are 40 cM apart. The alleles are M1/m1 at locus 1 and M2/m2 at locus 2.

- Calculate the recombination frequency between the markers, assuming Haldane's mapping function
- Calculate the recombination frequency, assuming Kosambi's mapping function

From now on we will use Haldane's mapping function.

Now assume there is a QTL effect at 10 cM from the first marker locus. The QTL has two alleles (Q/q). Consider a bull that has received a M1QM2 gamete from the sire and a m1qm1 gamete from the mother.

- What are the expected paternal marker haplotypes in the offspring from this bull?
- What are the recombination frequencies between the marker loci and the QTL.
- How many paternal haplotypes for the three loci (M1-Q-M2) can be found in the offspring from this bull? What are their expected frequencies?
- Calculate conditional probabilities for carrying the Q-allele for each paternal marker haplotype.
- Calculate expected phenotypic means for each group of progeny of a particular paternal marker haplotype, given the genotypic means of QQ, Qq and qq genotypes are 9, 10 and 11, respectively. Assume that the dams of the progeny contribute q alleles only.

Exercise 2: Interval Mapping of QTL

We continue with the case as in Exercise 7.2. We have now 8 half sib progeny from the sire with the following data:

Paternal marker haplotype	phenotype
M1M2	9.7
M1M2	10.3
M1m2	10.2
M1m2	9.5
m1M2	9.8
m1M2	9.2
m1m2	9.3
m1m2	8.8

- Estimate relevant effects (which?) for a QTL that is positioned at 10 cM from marker locus 1 and 30 cM from Marker locus 2.

- Test whether there is a significant QTL effect at this location.

Now use the excel spreadsheet QTLDET.XLS.
The spreadsheet allows you to enter data in the blue cells about position of marker and QTL, and to give QTL parameters.

- Verify the answers you obtained from the previous exercise concerning maker haplotype probabilities, and Q-probabilities conditional on marker haplotypes.

The spreadsheet also allows you to simulate data for a half sib group.

- Simulate data for 200 progeny, using the parameters as given in Exercise 2.2 and 1.3. Use a variance of 1.
- What can you conclude concerning the QTL effect from your simulation? What evidence/criteria have you used to draw such conclusions.
- Now simulate 10 such data sets. Determine the average value for the relevant QTL parameters, and their standard deviation.
- Try to work out what would be a reasonable (minimal) progeny group size to detect this QTL
- You can also work out the minimal progeny group size needed to detect a QTL of 0.5 and 2, respectively.
- Does the size of the marker bracket have a large effect on the minimal progeny group size?

Exercise 3: Models for multiple QTL

Data is analysed from an experiment where we have hypothesised two QTL, each with two alleles (A and for the first locus and B and b for the second locus, respectively. You can use the spreadsheet LINMOD.XLS to analyse the data.

We have estimated the means for 16 QTL genotypes using phenotypic and marker genotype data from a QTL experiment. The QTL genotypes are inferred based on hypothetical QTL positions that were derived from the marker haplotypes.

- Investigate for the following data the genetic model.
- Estimate additive and dominance effects at each QTL.
- Test whether effects are significant
- Test whether there is epistasis between the QTLs

Genotype	Mean	Number observed
AABB	13.08	25
AABb	9.79	21
AabB	9.83	16
Aabb	8.72	23
AaBB	14.18	28
AaBb	9.62	32
AabB	10.15	17
Aabb	10.58	31
aABB	12.16	27
aABb	10.23	28
aAbB	9.79	19
aAbb	7.65	17
aaBB	7.97	34
aaBb	5.62	18
aabB	7.65	24
aabb	6.89	27