Establishing Marker-QTL Linkage: Principles, Requirements and Methodologies

M. Maheswaran
Tamil Nadu Agriculture University, Coimbatore

Introduction

The idea of using genetic markers to locate the individual quantitative trait locus (QTL) responsible for variation in quantitative traits goes back nearly to the beginning of modern genetics (Sax, 1923). With the availability of dense highly informative marker maps, it has recently become feasible to map genes or QTL accounting for part of the heritability of continuously distributed traits in experimental crosses as well as outbred populations. The most extensive comparative data set available at this point probably comes from QTL mapping efforts in plants. Interestingly, an unexpectedly high proportion of QTL affecting seed size, height, flowering and other complex traits do correspond among different taxa (Paterson, 1998). The process of QTL analysis requires 1) a suitable mapping population of phenotypically contrasting parents, 2) a linkage map of molecular markers, 3) mapping methods and software and 4) reliable phenotypic screening methods and generation of phenotypic data.

Suitable mapping population

It would be always advantageous using populations of early generations such as F2, F3, BC population etc, since these populations are amenable to make accurate predictions. However, the predictions made involving early generations would be misleading because of camouflaging effect in early generation of the major gene on many other minor genes. Continuous inbreeding to evolve recombinant inbred lines (RILs) can eliminate this camouflaging effect (Allard and Harding, 1963). Thus, RILs can remain as the best choice of population for QTL analysis. As an alternative doubled haploid (DH) lines can also be used. The inherent homozygosity prevailing in the individuals of these two populations make the RILs and DHLs as immortals and help to have as many replications as required by the experiment.

![Diagram](image-url)

Fig 1. Mapping populations for QTL mapping
A linkage map of molecular markers

Thoday (1961) emphasized that the main practical limitation in localizing QTL, seems to be the non-availability of suitable markers. This limitation was remedied by the construction of complete Restriction Fragment Length Polymorphism (RFLP) linkage maps, permitting systematic searches of an entire genome for QTL influencing a trait (Paterson et al 1988). The Amplified Fragment Length Polymorphism (AFLP) markers, (Vos et al., 1995), the markers of choice, remain the best alternative to construct the linkage maps in a very short period based on the existing RFLP maps (Maheswaran et al., 1997). Several linkage maps of molecular markers have been constructed exclusively for QTL analysis of various agronomic traits in crop species such as tomato, maize, rice and soybean.

Mapping methods and software

The basis of all QTL detection, regardless of the crop to which it is applied, is the identification of association between genetically determined phenotypes and specific genetic markers. The possible methods of analysis to detect QTL include: 1) single marker analysis (otherwise called as Marker-Trait (MT) Method) and 2) interval analysis.

QTL mapping methods

Conceptually, QTL mapping amounts to a three-step recipe: scan the entire genome with a dense collection of genetic markers; calculate an appropriate linkage statistic $S(x)$ at each position $x$ along the genome; and identify the regions in which the statistic $S$ shows a significant deviation from what would be expected under independent assortment. The underlying assumptions of QTL mapping involving molecular markers are: 1) genes controlling quantitative traits are located on the genome, just like simple genetic markers, 2) if the markers cover a large portion of the genome then there is a large chance that some of the genes controlling the quantitative traits are linked with some of the genetic markers and 3) if the genes and markers are segregating in a genetically defined population, then the linkage relationship among them may be resolved by studying the association between trait variation and marker segregation pattern. The association between quantitative trait variation and marker segregation pattern can be carried out by the following methods.

Single marker analysis

The single marker analysis (SMA) is a good start not only for learning QTL mapping, but also for practical data analysis. Single marker analysis is the method used in earliest studies on QTL mapping (Edwards et al., 1987; Weller et al., 1988). In this, one marker is involved at a time to find the QTL-marker association. The single marker analysis can be implemented as a simple t-test, ANOVA, linear regression, and likelihood ratio test and maximum likelihood estimation (Haley and Knott, 1992; Nienhuis et al., 1987; Wang et al., 1994). SMA is simple in terms of data analysis and implementation. It can be
performed using common statistical software. Gene orders and complete linkage map are not required.

![Diagram of association between a marker and putative QTL](image)

**Fig 2.** Association of a marker with a putative QTL

The disadvantages of the single marker analysis are: 1) the putative QTL genotypic means and QTL positions are confounded. These confounding cause the estimated QTL effects to be biased and the statistical power to be low particularly when linkage map density is low and 2) QTL positions cannot be precisely determined, due to the non-dependence among the hypothesis tests for linked markers that confound QTL effect and position. Worked out examples to do single marker analysis to establish marker-QTL association are given by Liu, (1998) and some of the key references for single marker analysis are given below.

**Table 1. Methods to carry out Single Marker Analysis**

<table>
<thead>
<tr>
<th>Method</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANOVA</td>
<td>Edwards <em>et al</em> (1987)</td>
</tr>
<tr>
<td>Simple t-test</td>
<td>Tanksley and Hewitt (1988)</td>
</tr>
<tr>
<td>Likelihood ratio</td>
<td>Weller (1986)</td>
</tr>
</tbody>
</table>

**Interval analysis or Interval mapping**

Interval mapping (IM) is considered as a second level of QTL mapping. QTL mapping by this method requires prior construction of a marker genetic map. The interval mapping approach is based on the joint frequencies of a pair of adjacent markers and a putative QTL flanked by the two markers.
Fig 3. Association of a putative QTL to two flanking markers

Interval mapping can be done by the following methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likelihood approach</td>
<td>Lander and Botstein (1989)</td>
</tr>
<tr>
<td>Regression approach</td>
<td>Knapp et al. (1990)</td>
</tr>
<tr>
<td>Combination of likelihood and regression approach</td>
<td>Zeng (1994)</td>
</tr>
</tbody>
</table>

The approach of interval mapping (IM), otherwise called as simple interval mapping (SIM) considers one QTL at a time in the model for QTL mapping. Therefore, SIM can bias identification and estimation of QTL when multiple QTL are located in the same linkage group (Zeng, 1994). SIM evaluates the association between the trait values and the expected contribution of hypothetical QTL (target QTL) at multiple analysis points between each pair of adjacent marker loci (the target interval). The expected QTL genotype is estimated from the genotypes of flanking marker loci and their distance from the QTL. Since there is usually uncertainty in the QTL genotype, the likelihood is sum of terms, one for each possible QTL genotype, weighted by the probability of that genotype given the genotypes of the flanking markers. The analysis point that yields the most significant association may be taken as the location of a putative QTL. Although IM represented a significant contribution to QTL analyses, it is based on the null hypothesis of no QTL: an incorrect assumption for quantitative traits.

Multiple QTL Mapping

Both SMA and IM are biased when multiple QTL are linked to the marker/interval being considered. To deal with multiple QTL problems, Jansen (1993) Rodolphe and Lefort (1993) and Zeng (1993) independently proposed the idea of combining SIM with multiple regression analysis in mapping. Multiple regression methods were integrated with IM to increase the probability of including all significant QTL in the model. This method was named as composite interval mapping (CIM). Though CIM produced more accurate and precise estimates than IM, the inclusion of too many cofactors reduced the power to identify QTL relative to IM (Zeng, 1994; Utz and Melchinger, 1996). Kao et al. (1999) proposed new method viz, multiple interval mapping (MIM) to deal with the mapping of multiple QTL. When compared to SIM and CIM, MIM tends to be more powerful and precise in detecting QTL.

Composite Interval Mapping (CIM): CIM evaluates the possibility of a target QTL at multiple analysis points across each intermarker interval. However, at each point it also includes the effect of one or more background markers, as defined in SIM. The inclusion of a background marker in the analysis helps in one of two ways, depending on whether the background marker and the target interval are linked. If they are not linked, inclusion
of the background marker makes the analysis more sensitive to the presence of a QTL in the target interval. If they are linked, inclusion of the background marker may help to separate target QTL from other linked QTL on the far side of the background marker (Zeng, 1993, 1994).

**Multiple Interval Mapping (MIM):** MIM method uses multiple marker intervals simultaneously to fit multiple putative QTL directly in the model for mapping QTL. The MIM method is based on Cockerham’s model for interpreting genetic parameters and the method of maximum likelihood for estimating genetic parameters. With the MIM approach, the precision and power of QTL mapping could be improved. Also, epistasis between QTL, genotypic values of individuals and heritabilities of quantitative traits can be readily estimated and analyzed.

**Power, precision and accuracy of QTL mapping**

QTL analysis includes three stages: detection, mapping and fine mapping. Detection and mapping (estimating a chromosomal location) are often accomplished simultaneously, but they are logically and statistically distinct (Beavis, 1998).

**Power of detection:** Power is the probability of identifying a QTL of known magnitude, given the predetermined frequency of false positive association (α). Each QTL detection experiment provides an estimate of the strength of a QTL. In some experiments, the QTL will be over estimated, in others, underestimated. This variability may determine whether the QTL appears to be statistically significant, that is, whether the QTL is detected in that experiment. The power of a QTL detection experiment, at a given level of statistical significance depends upon the strength of the QTL and the number of progeny in the population.

The strength of the QTL can be determined based on the fraction of the total trait variance that it explains. Those, which explain over 20 percent of the variance, are strong QTL; traits controlled by such QTL can be considered almost Mendelian. At the other extreme, weak QTL, which explain 1 percent or less of the trait variance, require at least a thousand progeny to detect them with high power. Detection of such QTL is routinely feasible. Between those extremes are moderate QTL, which can be detected with crosses of reasonable size but not necessarily at high power.
Precision of mapping: Precision is a measure of the dispersion of repeated independent estimates of genomic positions or genetic effects of the alleles at QTL and reported by inverse measures such as standard errors or confidence intervals. The size of a confidence interval is expected to be inversely proportional to the number of progeny in the mapping population and to the square of the strength of the QTL. Weak QTL, as defined above, can be assigned to a chromosome, but not located with more precision. Strong QTL can be located by a large backcross or intercross in confidence intervals as small as 11 cM. For strong QTL, precision is limited by the number of recombinants in backcross or intercrosses and QTL can be located more precisely in advanced intercrosses (Darvasi, 1998).

The power and precision of QTL mapping depends on the test statistic derived based on the asymptotic distribution. When single marker analysis is involved in QTL mapping, ‘t’ and ‘F’ statistics are used to assess the power and precision of contrasting marker genotypic classes (Soller et al., 1976; McMillan and Robertson, 1974).

Accuracy of mapping: Accuracy is a measure of how close the estimates are to the true values. In practice, accuracy is very difficult to estimate for experimental results because the true values are unknown.

Test statistic for claiming QTL detection

The QTL, by definition, are merely significant statistical associations. These significant associations are detected by having suitable test statistic, otherwise called as ‘critical value’ or ‘threshold statistic’. The reliability and efficiency of the QTL mapping depend considerably on the validity and relevance of the statistical tests used to detect the presence of QTL. Clear statistical guidelines for the interpretation of linkage results are needed to avoid a flood of false positive (presence of a QTL when actually it is not present) claims. At the same time, an overly cautious approach runs the risk of causing true hints of linkage to be missed (false negative).

‘Critical value’ or ‘threshold statistic’ is a limit fixed to eliminate the detection of spurious QTL and QTL with smaller effects. Fixing a suitable threshold statistic for each population size will help in improving the power of QTL mapping. In using single marker analysis, test of significance is used as threshold statistic. When analysis of variance is used as method to detect QTL, ‘F’ value is used as threshold statistic. In the same manner, for interval mapping LOD score is used as a threshold statistic. The LOD score summarizes the strength of evidence in favour of the existence of QTL with an effect at a position; if the LOD score exceeds a predetermined threshold (usually LOD score of 3.0 is fixed), the presence of a QTL is inferred. For estimating the LOD score, one has to have the odds ratio, which is the ratio between chance of QTL at a given site and chance of no QTL at a given site.

Threshold statistic adopted for different methods may not have the same strength, resulting in differences in detecting the QTL. Under the circumstance, the threshold
Establishing Marker-QTL Linkage: Principles, Requirements and Methodologies

The statistic of each method has to be evolved to eliminate the discrepancies between methods. Churchill and Deorge (1994) evolved a method to relate the LOD score and F statistic of ANOVA.

$$\text{LOD} = \left[ \frac{n_1 + n_2}{2} \right] \log_{10} \left[ 1 + \frac{T^2}{n_1 + n_2 - 2} \right]$$

where, $T^2$ is F statistic of ANOVA and $(n_1 + n_2)$ is sample size.

When using any statistic of any method, as a criterion in model selection for QTL detection, it is very important to determine the appropriate critical value or threshold value for claiming QTL detection such that correct statistical interference about QTL parameters can be made.

Lander and Botstein (1989) suggested using the Bonferroni argument for the sparse map case and Orenstein-uhlenback diffusion for the dense map case to determine the critical value. Generally, it has been pointed out that the critical value might need to be adjusted for the number and size of interval, different levels of heritability, different number of multiple linked or unlinked and unlinked in the same or opposite direction (Lander and Botstein, 1989; Jansen, 1993; Zeng, 1994). Visscher and Haley (1996) suggested that the critical value should be reduced after a QTL of large effect has been detected. However, most of this information is not available before mapping and consequently the answers to most of the above questions remain unknown. Churchill and Deorge (1994) therefore suggested using permutation test for determining an appropriate critical value for specific data sets.

The permutation test (Churchill and Doerge, 1994; Deorge and Churchill, 1996) is a method for establishing the significance of the LRS generated by single locus association or interval mapping. In this test, the trait values are randomly permitted among the progeny, destroying the relationship between the trait values and the genotypes of the marker loci in the observed data, QTL parameters and LRS value are estimated for each permuted data set at regular intervals throughout the genome (or some part of the genome) and the maximum LRS is recorded. This procedure is repeated numerous times, giving a distribution of LRS values expected if there were no QTL linked to any of the marker loci. An empirical $p$-value can be obtained for a given LRS by computing the proportion of permuted data sets for which LRS exceeds the LRS for the observed data. Alternatively, values at appropriate percentile points of the empirical distribution can be used as LRS threshold values to establish significance of the observed LRS. For example, the 95th percentile value is that which would establish significance corresponding to the usual criterion of $\alpha=0.05$. Churchill and Doerge (1994) recommended at least 1000 permutations for establishing a threshold for $\alpha=0.05$. Permutation tests, therefore, can be time consuming and may be impractical on some computers.

Fixing a correct critical value to detect QTL is still a debatable issue in QTL mapping. Having a uniform stringent standard such as a critical value based on a whole genome search or a critical value based on an infinitely dense map is not acceptable since some times QTL mapping involves few markers (or few chromosomal regions) or a sparse map. Under the circumstance, a hierarchical search – in which one performs a genome
scan with a sparse map and then follows up interesting regions with a denser map as suggested by Lander and Kruglyak (1999) is an efficient study design.

Lander and Kruglyak (1995) proposed the following classification based on the number of times that one would expect to see a result at random in a dense, complete genome scan:

- **Suggestive linkage:** statistical evidence that would be expected to occur one time at random in a genome scan.
- **Significant linkage:** statistical evidence expected to occur 0.05 times in a genome scan (that is, with probability 5 percent).
- **Highly significant linkage:** statistical evidence expected to occur 0.001 time 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.

### Software

Compared to general statistical analysis of biological data, statistical analysis for the study of genes controlling complex traits has the following characteristics: 1) many repeated analysis in one task, 2) lack of standard distribution for some test statistics and 3) complexity of models used in QTL mapping. For using these software packages a known linkage map is needed for either running the programmes or interpreting results. Several companion packages are also available for linkage map construction.

These packages have some similarities such as: 1) interface is not user friendly compared to some commercial software, 2) user support is also limited due to their non-commercial status, 3) statistical models which can be built using the software are limited and 4) speed of model building is high for the models which the software can build. The details on some of the software routinely employed in QTL mapping are given below.

**MAPMAKER/QTL** (Lincoln et al., 1992b) is a widely used program for UNIX or DOS operating systems and is the original QTL mapping program intended for distribution. It can perform composite interval mapping, although the documentation does not use that term; but it cannot perform permutation tests. It requires the companion program MAPMAKER/EXP (Lander et al., 1987; Lincoln et al., 1992a) to format data and to calculate marker maps.

**QTL Cartographer** (Basten et al. 1994, 1997) is a suite of programs for DOS, UNIX, or Mac OS. They are designed to be used in sequence, each accepting input in the form of text files and storing its output in text files for the next program. This suite offers several variations of CIM with automatic selection of background loci. It also has provision for estimating confidence intervals by resampling. QTL Cartographer, MapQTL, and PLABQTL are similar in many respects. QTL Cartographer is distinguished by its menu-driven interface, more detailed documentation, resampling methods and the lack of a licensing fee.
Map Manager QT (Manly and Elliott, 1991; Manly, 1997) is a program for Mac OS distinguished by its graphical user interface for data entry, editing, manipulation, and display. It is designed to be used either as a mapping program itself or as a data-preparation program for other mapping programs.

QGene (Nelson, 1997) is a commercial program for Mac OS whose strength is a variety of graphics for displaying trait data and relationships among marker genotypes and between traits and marker genotypes. These functions make it uniquely useful for rapid exploration of data. However, it does not perform CIM.

MapQTL (van Ooijen and Maliepaard, 1996) is a commercial program for several operating systems that is distinguished by its ability to map QTL in populations derived from non-inbred parents, in which both markers and QTL may have more than two alleles. It also offers a nonparametric form of single-locus association, the Kruskal-Wallis rank sum test, appropriate for data with distributions far from normal.

PLABQTL (Utz and Melchinger, 1996) is a script-driven program for DOS or AIX that is designed to analyze automatically a dataset at increasing levels of complexity in successive runs. The final level is capable of evaluating the effect of different environments and the effect of interactions between QTL and environmental effects.

MQTL (Tinker and Mather, 1995a, 1995b) is a program for DOS or Sun OS that uses a simplified form of composite interval mapping (sCIM) for mapping QTL in large data sets derived from multiple environments. Like PLABQTL, it will estimate environmental effects and QTL-environment interactions.

Multimapper (Sillanpaa, 1998) is a program for UNIX that implements a Bayesian method for building multi-QTL models automatically. Multimapper is designed to map QTL within a single linkage group and it produces a plot of QTL probability as a function of map distance. This type of plot seems intuitively more interpretable than the plot of the likelihood ratio statistic or LOD score produced by other programs. However, it seems to be the most suited to the analysis of single chromosomes for which other programs have indicated the possibility of multiple QTL. Multimapper is designed to work with QTL Cartographer as a companion program.

Epistat (Chase et al., 1997) is a program for DOS designed primarily for the detection and analysis of interactions between QTL. It does not perform interval mapping and therefore does not require mapped markers. It is an interactive program, displaying graphic results in response to single-keystroke commands.

The QTL Cafe is a program being developed in Java to make it available for multiple computer platforms. It is currently available as an applet that runs in a Java-enabled World Wide Web browser.
Available public domain software packages for studying genes controlling complex traits are not adequate for development of genomic research on complex traits in terms of user interface, flexibility and user support. Software packages with commercial quality are needed to accommodate the growing needs of data analysis and management in genomic research. This is especially true for study of genes controlling complex traits.

**Reliable phenotypic screening and generation of phenotypic data**

To adequately explore the QTL during the mapping phase, the phenotype must be evaluated in replicated trials in different environments. Moreover, phenotypic screening should be done based on reliable and reproducible screening methods. Large data sets can be generated by the coordinated efforts of several groups, providing valuable information about genes governing quantitative characters in a range of environments. Such data will provide information about the magnitude of the effect of different QTL and whether there is interaction between QTL and environment.

The issues related to population development and construction of linkage maps do not pose many problems with the existing level of knowledge. Though, issues related to methods for detecting QTL, software for QTL analysis are having problems such as inaccurate detection of QTL (occurrence of false positives and false negatives), the issues associated with phenotypic screening pose severe threat to an emerging tool of plant breeding.

Most agronomically important characters involve multiple genes that interact with each other and with the environment in complex ways. This creates a situation wherein QTL can be detected only some of the time. This necessitates designing of experiments to qualify the value of specific QTL. To adequately explore the value of QTL, the phenotypes must be evaluated in well-replicated trials in different environments. The conduct of replicated trials warrants a suitable population which can be effectively replicated. Here comes the problem of developing an immortal population such as RILs and DHLs. Developing both kinds of populations and their maintenance is a cumbersome process even by a well established breeding institute. QTL are hypothetical genes based on statistical inference. Genetic effects used in QTL mapping could have very little biological meaning. To have a biological meaning of QTL mapping the selection of traits to be phenotyped is very important. More over, phenotype of several traits is not amenable for QTL dissection.

**Conclusion**

QTL, otherwise described as hypothetical genes based on statistical inferences, have very little biological meaning. To date the knowledge on QTL mapping is enormous. However, the accrued knowledge does not have immediate solutions to the problems associated with QTL mapping. Some of the approaches such as adopting metabolic genetic model (Byrne *et al.*, 1996; Mitchell-Olds and Pedersen, 1998) and candidate gene concept (Long and Longly, 1999) in conjunction with Single Nucleotide Polymorphism (SNP) may make the QTL mapping approach as biologically meaningful one. Issues

References


Sax K (1923). The association of size differences with seed coat pattern and pigmentation in *Phaseolus vulgaris*. Genetics 8: 552 – 556.


