11 Computing Strategies and Software for Gene Mapping

Brian S. Yandell[1] **and Peter Bradbury**[2]

[1]University of Wisconsin, 1675, Observatory Drive, Biometry Program, Madison, WI 53706, USA;
E-mail: byandell@wisc.edu

[2]USDA-ARS, 741 Rhodes Hall, Ithaca, NY 14853, USA

1 INTRODUCTION

This chapter focuses on computing strategies and software for gene mapping. We separately address software strategies for experimental crosses, known as quantitative trait loci (QTL) mapping, from those used in natural populations for association analysis. Both of these approaches look for correlations between genotypes and phenotypes. For most of the development in this chapter, we focus on a single phenotype, but we briefly note strategies that can examine multiple correlated phenotypes.

The goal of gene mapping is model selection for the genetic architecture of a phenotypic trait. That is, we wish to infer what genomic regions, or **genetic loci**, are associated with a phenotype, and what mode of gene action is involved. Genetic loci typically cover several megabases of DNA containing many closely linked genes. **Gene action** is often interpreted in terms of the additive and/or dominance effect of single loci, and epistatic interaction among two or more loci. A well estimated genetic architecture for a trait of interest can be used in disease prognosis, marker-assisted selection or studies of the evolution of the trait.

Depending on the method of analysis, the genomic region for a locus may cover a single genetic marker or a set of markers. **Genetic markers** are short segments of DNA, or in some cases qualitative traits, that can

be scored to partially or fully identify the inheritance of parental alleles. The markers are ideally arranged in a genetic linkage map corresponding to the physical map or sequence of the genome under study. While these previously determined maps are now routinely available for model organisms, such as *Arabidopsis* and rice, there are many taxa with no marker map, or only with maps based on the marker data in hand.

This chapter does not address the problem of building marker maps, except to point out that there are standard tools to build maps for experimental crosses, including MapMaker (Lander and Green 1987; Lander et al. 1987), R/qtl (Broman et al. 2003), MultiPoint (Mester et al. 2004) and JoinMap (Jansen 1993). Genetic maps are typically in units of centi-Morgan (cM), with two loci separated by 1 cM having an expected recombination frequency of 1%. Roughly speaking, 1 cM is about 1 megabase of DNA, depending on taxa. However, recombination frequency is not linear with **genetic distance**. The choice of recombination model, defining the relationship between genetic distance and recombination fraction, plays a minor role in gene mapping for experimental crosses once a linkage map is built. Most gene mapping methods for experimental crosses assume no crossover interference. Typically there are further assumptions of independent crossovers with equal likelihood across the genome (Haldane map function), although most packages allow other options. More complicated experimental cross designs, particularly in outbred populations, may require multipoint mapping. There are many subtle issues about the relationship of recombination to distance that are beyond the scope of this chapter.

Gene mapping involves multiple tests for correlation between a set of genetic markers and a trait of interest. Typically for an individual experiment, a large number of markers are tested against a phenotype, which leads to multiple testing issues when each marker-trait combination is tested individually. We will address these issues in context of specific software strategies.

QTL mapping in experimental crosses usually begins with two inbred lines. The individuals in such an experimental cross, are created, nurtured and measured under uniform conditions so that the primary differences among individuals are due to their genetic 'treatment'. In contrast, **association analysis** considers the relationship between genotype and phenotype in a natural population. It uses the extent of linkage disequilibrium (LD) between a trait and markers to infer the location of QTLs.

The existence of an association between a marker and a trait in an experimental cross implies that either the marker itself is the cause of the observed phenotypic variation or that it is linked to a causal

polymorphism. There are two problems with interpreting correlation as causation in an association study of a natural population: (1) a linked marker might not be in LD with the trait; and (2) a marker that is in LD with a trait might not be linked to it.

Several distinctions between QTL mapping with experimental crosses and association mapping with natural populations are worth mentioning beyond the issue of causal inference. The resolution of QTL mapping in experimental crosses is typically fairly coarse, on the order of 5-20 cM, whereas association mapping can lead to much finer resolution maps. Experimental crosses from two inbred lines have at most two alleles at any genomic locus, while natural populations may have multiple alleles. As a result, heterozygosity is fairly uniform in experimental crosses, and markers are usually informative, or not, of parentage for the whole population. However, in natural populations, markers may have quite different heterozygosity and may only be informative for a certain subset of individuals. All individuals in an experimental cross have equal genetic correlation on average, but this does not hold in natural populations. Finally, missing data presents much more difficult problems in natural populations, where one may need to infer the phase of inheritance to estimate haplotypes.

2 QTL ANALYSIS WITH INBRED LINES

Modern methods for QTL analysis in experimental crosses derived from inbred lines largely employ the **interval mapping** framework developed by Lander and Botstein (1989). This fundamental paper viewed the relationship between phenotype and genotype as a genomic question, providing a visual LOD score map to profile evidence for association across the genome. The linkage map inference inherent in this work (Lander and Green 1987) provided an algorithmic approach to model missing genotype information between markers.

Good expositions of QTL methods for experimental crosses can be found in Broman (2001) or Hackett (2003). Doerge et al. (1997) reviewed the statistical issues, while Mackay (2001) placed QTL mapping in the context of identifying underlying mechanisms. Several recent papers have addressed the difficult issue of moving from QTL intervals to confirmed genes (Guo and Lange 2000; Nadeau and Frankel 2000; Glazier et al. 2002; Korstanje and Paigen 2002; Page et al. 2003; Darvasi 2005).

This section gives an overview of QTL analysis for inbred lines, contrasting the algorithms commonly used. Rather than show screen

shots of individual packages, we primarily use graphs developed in R (R Core Development Team 2006) to illustrate concepts. We point out which packages use what methods as we go along. We begin with a detailed investigation for mapping a single QTL. This leads to questions of assessing significance (thresholds, support intervals), which leads to model selection for multiple QTL. We finish with a brief overview of packages.

2.1 Single QTL Mapping

QTL mapping models the relationship between phenotype and genotype at each location across the genome. We briefly review methods employed to assess this relationship profiled across the genome. We begin with a complete data situation with 100 individuals and 201 markers spaced every 1 cM, with a single QTL at 100 cM. Here all methods agree exactly with each other.

Marker regression (MR) examines the association between each marker and a phenotype. This was the available method until 1989, performed marker by marker using *t*-tests for backcross or ANOVA for intercross. When markers are arranged in a linkage map, *p*-value summaries provide a crude genome-wide profile. However, it was widely recognized that regression with a single marker confounds the allele substitution effect and linkage between each marker and the pertinent genetic locus. Further, any missing data reduces power, as those individuals must be dropped for that marker. This method is still sometimes used as a quick initial examination, particularly in genomes with no linkage map yet available.

Simple interval mapping (SIM) models the relationship between phenotype and genotype by testing for a QTL at each location across the genome (Lander and Botstein 1989). This involves a likelihood ratio test, rescaled as a familiar LOD score. That is, interval mapping (IM) states that the phenotype has a normal, or bell-shaped, histogram for any given genotype, with a different mean depending on the genotype. When the genotype is not known, IM assigns probabilities to missing genotype data based on informative flanking markers. The full likelihood mixes over all possible missing genotype values. That is, the distribution for an individual is a weighted average of normal distributions, with the weights being the probabilities for QTL genotypes given flanking markers. Individuals with the same flanking marker genotypes would have the same mixture distribution. The profile likelihood is the product of these distributions at the phenotype values maximized for unknown effects using the **expectation-maximization (EM)** algorithm (see Lander

and Botstein 1989). The **log odds (LOD)** is the log base 10 of this likelihood profile divided by the null likelihood for the no QTL model.

Regression mapping, also known as **Haley-Knott (HK) regression** (Haley and Knott 1992; Martinez and Curnow 1992), considers regression of the phenotype on the expected genotype, which approximates the more correct mixture detailed above. That is, they agree in mean value, but differ in variance and in distribution shape. This is not a serious problem if markers are closely spaced and there are only a few missing marker genotypes. However, with selective genotyping, this Haley-Knott regression can be seriously biased (Kao 2000). Xu (1995) showed that this method can overestimate residual variance when QTL effects are large or markers are widely spaced.

When there are no missing data, interval mapping and Haley-Knott regression are *exactly* identical. Marker regression agrees as well at every fully informative marker with these curves once it is rescaled in terms of LOD scores. That is, there is no approximation involved when we have complete data at markers, and all methods agree. Consider the following simple example with 100 individuals in a backcross fully genotyped at 201 markers spaced every 1 cM. The QTL is located at 100 cM, with a substitution effect of 2 relative to a standard deviation of 1. The LOD curve peaks near 100 cM, but slowly trails off (Figure 1a). The attenuated substitution effect at a locus that has a recombination rate of r with the QTL is $(1\text{-}2r)a$, where a is the substitution effect at the QTL (Figure 1b). The long attenuation of the LOD curve is due to this confounding of QTL substitution effect and linkage of nearby markers (Wright and Kong 1997). The expected LOD at the QTL is $(n/2) * \log10(1 + a^2/2)$, which is attenuated to $(n/2) * \log10((1 + a^2/2)/(1 + 2a^2r(1 - r))$ at the linked marker. The attenuation can only be relieved by increasing sample size or considering multiple QTL models.

2.1.1 Missing Genotypes

Linkage maps for some time have had markers every 5 to 20 cM, requiring some way to fill in for missing genotypes between markers. Thus genotype information between markers is completely missing. Further, some markers may have missing values due to technical reasons unrelated to phenotype. In other situations, genotype data may be missing in a pattern associated with the phenotype, either by design or chance. This subsection examines the impact of various types of missing data on QTL mapping.

Marker data **missing at random** does not introduce any appreciable bias to QTL mapping. However, missing data must be replaced by assumptions about what that data might have been. This introduces

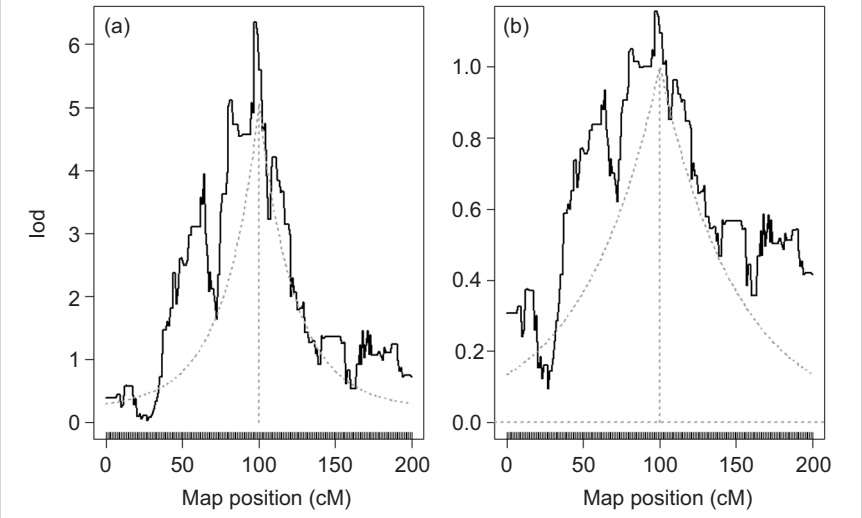

Fig. 1 Simulation with 1 QTL at 100 cM and complete marker information. Markers every 1 cM. (a) LOD profile scan for simple interval mapping. Black line is interval mapping; gray dashed line is expected LOD profile. Note long-term attenuation of peak away from 100 cM. (b) Substitution effect attenuated by linkage (1-2r) to nearby markers. Estimated effect from interval mapping in black; idealized effect in gray dotted.

uncertainty into LOD maps, and the impact of that uncertainty depends on the algorithm employed. Several ways to address missing genotype data have been implemented in QTL software. It is important to understand these because some methods have known bias when data are not missing at random.

It is useful here to introduce a third method of QTL mapping known as **multiple imputation (IMP)**. The basic idea is to fill in the missing genotype at each 1 cM step using the assumed map function. This is done multiple times, recognizing that any particular realization is flawed. These multiple imputations are then averaged in a careful way to produce a **log posterior density (LPD)** that is very close to the LOD score (Sen and Churchill 2001).

What is the impact of marker spacing on LOD profiles? Simple IM fills in missing data between markers using the EM algorithm and the map function. This leads to a smooth parabola shape for the LOD between partially informative markers. HK regression tends to dampen those parabolas slightly. Multiple imputation averages as well, but may occasionally wiggle due to sampling variation. Figure 2a shows the effect

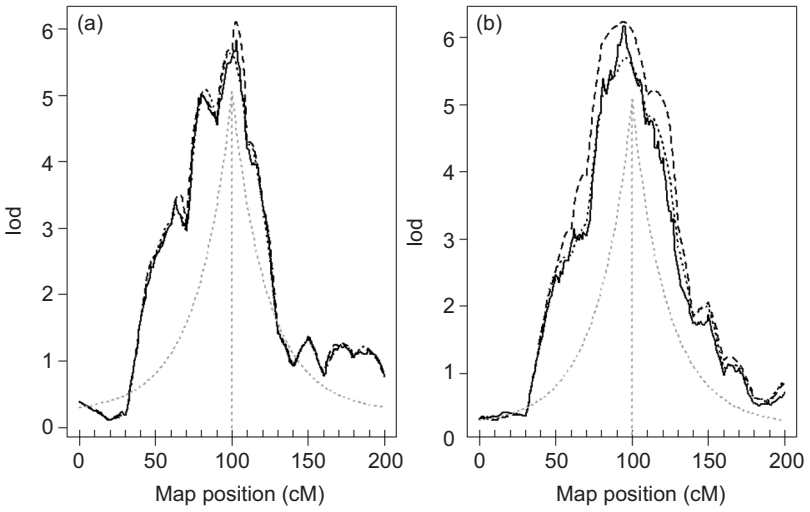

Fig. 2 Impact of marker spacing and genotype data missing at random. Markers every 10 cM, with profile evaluated every 1 cM. (a) Complete marker data. (b) Half of marker data missing at random. Dashed line = interval mapping, dotted = Haley-Knott regression, solid = multiple imputation; gray dotted is expected substitution effect.

of 10 cM spacing on the single QTL example introduced earlier. When we in addition remove half the marker data at random, the LOD curves keep the same basic shape (Figure 2b).

Markers may have missing data due to a design decision, such as **selective genotyping**. Selective genotyping implies a biased pattern of missing marker information. Typically, some fraction (10-25%) of extreme high and extreme low phenotype individuals are fully genotyped, while those in the middle range are not genotyped. All phenotype data are used for analysis, even those with no genotype information. Figure 3a shows how HK regression greatly overestimates the strength of the QTL signal when only extreme quartiles are genotyped. The other two methods tend to have reduced peaks relative to the figures seen earlier, which is not surprising since much data has been lost, leading to a reduction in power. However, they capture the same essential strength of relationship between phenotype and genotype and are not inflated by the pattern of missing genotype data.

The impact of selective genotyping on Haley-Knott regression can be more or less ameliorated by having a framework map of fully informative markers. Figure 3b shows a situation with fully informative

markers every 20 cM starting at 10 cM. The peak for Haley-Knott regression is still biased upwards. Note that the addition of these fully informative framework markers drops the proportion of missing data per marker from 50% to at most 20% in this situation. For more information on designing experiments with selective genotyping (see Sen et al. 2005).

It should be pointed out that the EM method can introduce artifacts in the presence of some patterns of missing data. Sometimes we see an unusual spike in the LOD map between partially informative markers. This indicates a region where the EM method is achieving 'too good' a fit to the data. Multiple imputation and HK regression tend to dampen such effects. The bottom line is that missing data is 'filled in' by assumptions in one way or another. No way is perfect, and artifacts can emerge. Always use caution interpreting QTL analyses in the presence of much missing data.

2.1.2 Detection of QTL

LOD maps as shown above provide a sense of where strong correlation between phenotype and genotype lie. But how large is large enough to say a QTL is detected with confidence? Theoretical guidelines based on high-powered math suggest a LOD threshold of about 3 (Lander and Botstein 1989; Lander and Kruglyak 1995), with some adjustment for design. In practice, it is wise to use resampling methods to assess the strength of the LOD signal.

A **permutation threshold** can be computed with most packages. The idea is to permute, or shuffle, the phenotypes independent of the genotypes. For each permutation, construct the LOD profile and record the maximum. The distribution of maximum LOD under the 'null' model of no QTL is approximated by a histogram of such values. Typically, 1,000 permutations are recommended for genome-wide purposes (Churchill and Doerge 1994). For the 10 cM data spacing shown in Figure 3, the EM thresholds at 1%, 5% and 10% are, respectively, 2.61, 1.81, and 1.53. Similar thresholds are found for the Haley-Knott and multiple imputation methods (not shown). Figure 4 shows these thresholds superimposed on the LOD maps. There is strong evidence to support a QTL, but there is also a wide region of the LOD curve that exceeds the 1% threshold. Note that permutation thresholds for the sex chromosome may need to constructed separately (Broman et al. 2006).

Common practice involves constructing a **LOD support interval** that spans a genomic region where LOD values are within 1 to 2 LOD of the peak. Strictly speaking a LOD support interval is *not* analogous to a

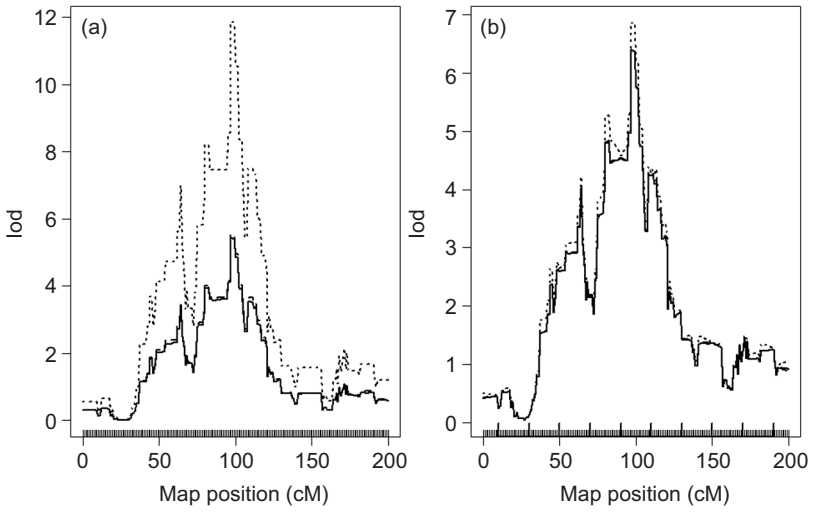

Fig. 3 Impact of selective genotyping. (a) Only extreme quartiles are genotyped while the genotypes for the middle 50% of individuals are missing. (b) Complete genotyping at framework map (markers every 20 cM indicated by larger tics); selective genotyping at all other markers. Note different vertical scales on the two figures.

confidence interval, although it is commonly misinterpreted as such. Its properties, in terms of percent of times it covers the true QTL location, depend on the marker spacings, pattern of missing data, and presence of other linked and epistatic QTLs. Still, it is a useful guide to the uncertainty in the QTL location. The 1.5 LOD support interval shown in Figure 4 is 76 to 109 cM for 10 cM marker spacing, 80 to 101 cM for 1 cM spacing. Permutation thresholds for single QTL scans are available in most QTL packages, although some only offer them for single QTL scans. Permutation thresholds for more complicated models, with multiple QTL and/or covariates, require more care (Doerge and Churchill 1996). Recently, some faster methods for permutation thresholds have emerged (Zou et al. 2004; Jin et al. 2007) and are gradually being incorporated into QTL packages.

Another resampling approach, bootstrapping, has been applied to QTL mapping. Several packages (Seaton et al. 2002; Mester et al. 2004; Wang et al. 2003) offer bootstrap of the distribution of the probability of the presence of a QTL across intervals. This tool is not well studied in this context, can be misleading, and should be approached with caution (Manichaikul et al. 2006).

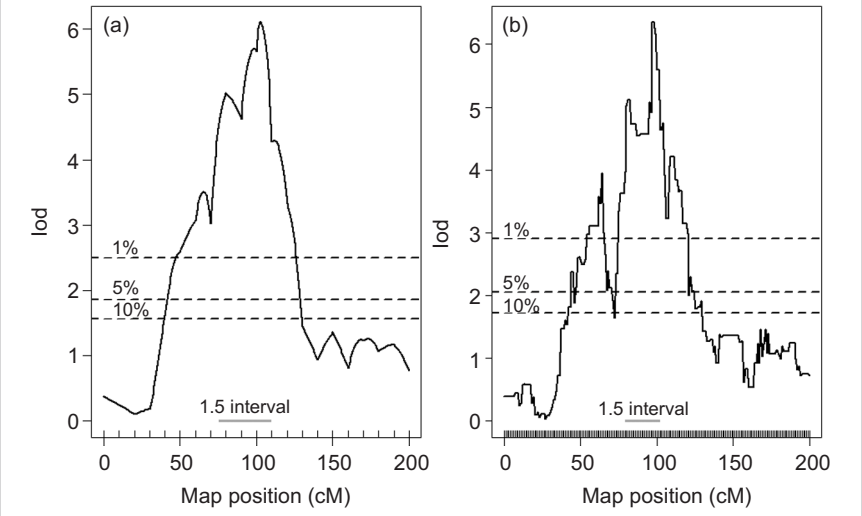

Fig. 4 Permutation-based thresholds for maximum LOD. Dotted horizontal lines at 1%, 5% and 10% threshold using EM method. Gray horizontal line is 1.5 LOD support interval from 76 to 109 cM. (a) Markers every 10 cM; (b) markers every 1 cM. Note slightly higher thresholds for 1 cM spacing.

2.1.3 Model Selection for Multiple QTL

Detection of a single QTL is an important first step. However, most complex traits are likely influenced by several, if not hundreds of genetic loci. We cannot hope to uncover the 'true model' in any given experimental cross, but we can infer major aspects of the genetic architecture that are supported by the data. Extension from a single QTL to multiple QTLs has been implemented in several distinct ways. We illustrate this with the one QTL example above and with another simulation having several QTLs. We focus on R/qtl (Broman et al. 2003), R/qtlbim (Yandell et al. 2007) and QTLCart (Basten et al. 1999) in this demonstration, as they have the key features of the methods found in most packages. Further, their graphics can be readily annotated and prepared for publication. We show graphs of one- and two-dimensional scans, as well as model selection tools. We briefly discuss model selection criteria, though that is beyond the scope of this chapter.

We need to briefly state that **epistasis** herein refers to the effect on a phenotype of statistical interaction among distinct genetic loci. Model-based epistasis is related to biological epistasis, which W. Bateson defined in 1907: "The allelic state at one locus can mask or uncover the

effects of allelic variation at another" (*cf.* Hollander 1955). We model epistatic interaction for complex traits with QTL as we do in ANOVA. Thus marker regression (MR) can be extended to multiple QTL using standard statistical packages. This quick and dirty method should only be considered a first pass, and epistasis uncoverd with MR is in general not to be trusted.

The process of **model selection** typically involves finding a balance between models that are too simple, missing key features and introducing bias, models that are overly complicated, inflating the variance of parameter estimates. A more complicated model with more QTLs *will* fit better and have a higher likelihood. But we pay a price for this: interpretation of a larger model is more complicated, and its components—including loci and genotypic effects—are each less precisely estimated. Further, the utility of a model is in its ability to predict effects of genoytpe on phenotype in a new experiment. An overly simple model will give biased predictions, missing linked QTLs and important epistasis. However, a model that is too complicated can be biased in other ways, being constrained to particulars of the current experiment. Thus, apparent evidence of subtle QTLs and epistasis may be artifacts of the data at hand, and may not generalize to other settings. These problems are not new to QTLs, and they have been well studied in stepwise regression. The basic idea in comparing models of different sizes, varying by the number of QTLs and the degree of epistasis, is to use an information criterion that equals the likelihood less some penalty that measures model 'complexity'. No one criterion is 'best', as each involves reducing a complicated, multidimensional comparison to a single number. We often compare models based on their maximum LOD scores. This criterion has no penalty for complexity, and is most appropriate when doing a few comparisons, say one vs. two QTL, with or without epistasis.

Broman and Speed (2002) compare various methods of model selection for multiple QTL that are located only at markers spaced every 10 cM. This is one of the only simulation studies to date comparing multiple QTL strategies. We refer the reader to this paper for a discussion of information criteria that measure the bias/variance tradeoff.

2.1.4 Multiple QTL Estimation Approaches

This subsection reviews the approaches to multiple QTL model fitting commonly used in software. These include regression mapping approximations, maximum likelihood, and Bayesian posteriors. A hybrid between regression and maximum likelihood is also highlighted, as it is

found in several packages. Model fitting is typically coupled with a model selection procedure, as we indicate in particular for the recommended maximum likelihood and Bayesian approaches.

As pointed out above, HK regression mapping is challenged by missing genotype data. This problem is exacerbated with multiple QTL (Xu 1995; Kao 2000). Still, the method is fast and works well when there are few missing data subject to selective genotyping. QTL Express (Seaton et al. 2002) is a popular package for this method. Tools for HK regression with multiple interacting QTL are available as well in R/qtl (Broman et al. 2003).

Hybrid methods, known as **composite interval mapping (CIM)** (Zeng 1994) or multiple QTL mapping (Jansen 1993) were proposed as ways to scan the genome with interval mapping while approximately adjusting for other QTL using nearby markers, or co-factors. This method is fast and fairly easy to implement, hence it has been incorporated into several packages, including PLABQTL (Utz and Melchinger 1996), MapQTL (van Ooijen and Maliepaard 1996) and MapManager/QTX (Meer et al. 2004). However, the approximation can be problematic, and its properties depend on the minimal spacing window to linked co-factors as well as the number of co-factors. An early analysis of morphological shape (Liu et al. 1996) using CIM was later revised using MIM described below (Zeng et al. 2000). Broman and Speed (2002) showed that CIM is effective when there are 'enough' co-factors, but misses linked QTL if there are too few co-factors. Returning to the fully informative simulation of a single QTL, Figure 5 shows how CIM can narrow the support interval for a QTL, but at the same time the peak is slightly elevated as the variance is artificially deflated by the co-factors. The 5% permutation threshold for CIM is estimated at 2.04 (300 permutations, default settings), compared to 2.09 for IM (1000 permutations). CIM should be used with caution as an exploratory tool, in conjunction with other methods described below.

Methods that estimate all QTLs together began emerging in the late 1990s and into this century. These methods include an extension of EM for maximum likelihood (Kao et al. 1999; Kao and Zeng 2002), known as **multiple interval mapping (MIM)**. MIM is available in QTLCart/WinQTLCart (Basten et al. 1999) and in MultiQTL (Mester et al. 2004). It turns out there are a number of technical issues that arise, making this a difficult problem. Basically, it is hard to know when you have actually reached maximum for a model with many QTLs! On top of that, there is the issue of deciding among models. Again, Broman and Speed (2002) provide simulation studies of the stepwise regression model selection strategy adapted to multiple QTL mapping. MIM applied to the earlier

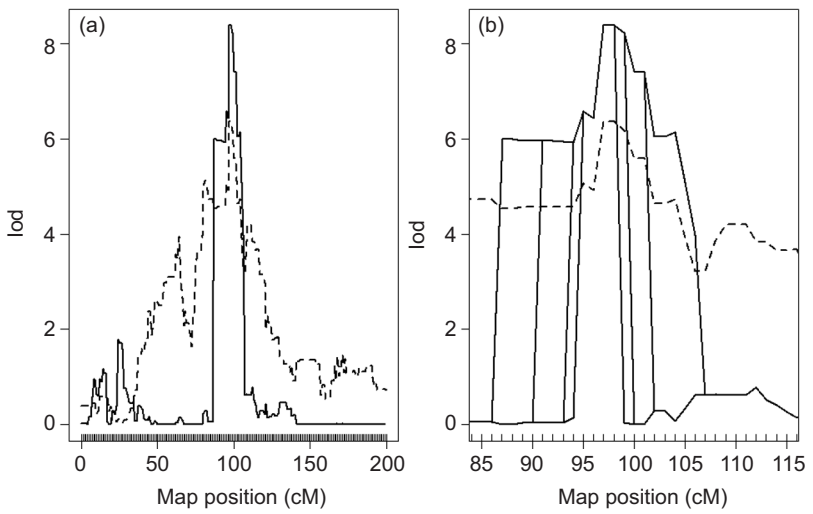

Fig. 5 Composite interval mapping (CIM) LOD profile. IM profile is dashed, CIM solid. Right panel zooms in on peak area, with CIM curves for 5 co-factors and window widths 10 (default), 5, 2 and 1 cM.

example definitively concludes there is exactly one QTL, thus reducing to simple interval mapping in this instance.

Two primary **Bayesian interval mapping (BIM)** methods have emerged, and they have been incorporated into available packages: multiple imputation and Markov chain Monte Carlo (MCMC). The Bayesian approach focuses on studying random samples from the posterior distribution, which is basically the likelihood weighted by a prior distribution on unknowns, rescaled to have area 1 to make it a distribution. The posterior is a useful device to examine the entire likelihood, rather than focusing only on the maximum peak. Priors play the role of formally incorporating uncertainty about unknowns into our models. These include positions of QTLs, their genotypic effects and epistatic effects, and even the complexity of the genetic architecture.

Multiple imputation (IMP), introduced above, is available in R/qtl (Broman et al. 2003) and the Matlab application, Pseudomarker (Sen and Churchill 2001). This method profiles the **log posterior density (LPD)**, yielding curves markedly similar to the LOD curves, as shown in Figures 2-4 above. The LOD maximizes the genotypic effects at each locus, while the LPD averages over the effects; these are essentially the same when

the assumed phenotype model is normal. The advantage of the multiple imputation method lies in its simplicity. We fill in (impute) missing genotype data based on flanking markers, then compute the LPD; repeat this several times and (carefully) average the results. Thus, it can readily be extended to multiple QTL using standard linear model tools for model building and model selection. Many useful tools along this line are incorporated into R/qtl (Broman et al. 2003).

A second BIM method was developed using **Markov chain Monte Carlo (MCMC)**. The MCMC method in itself is not Bayesian—it has been used for maximum likelihood in human QTL studies (Heath 1997)—but it can be used to obtain random samples from the posterior. MCMC methods have proven quite useful for complex models in a variety of settings where a Bayesian perspective, modeling the uncertainty about many relationships, is important (*cf.* Gelman et al. 2003). Initially handling a fixed number of QTLs, (Satagopan et al. 1996) MCMC methods for QTL now incorporate uncertainty about the genetic architecture (Satagopan and Yandell 1996; Sillanpää and Arjas E 1998; Stephens and Fisch 1998; Gaffney 2001; Yi 2004; Wang et al. 2005; Yi et al. 2005; Yandell et al. 2007). This allows us to use the Bayesian approach for model selection, in which we allow the number, position and genotypic effects of QTL to be unknown.

Figure 6 shows MCMC applied to the fully informative one QTL example used to this point. The peaks for LOD and LPD are nearly the same, but the MCMC curve drops off more quickly. This is also apparent with the substitution effect. The dropoff is more dramatic with larger sample sizes (here we have 100) and/or larger substitution effects. The reason for this drop-off is that the question has changed somewhat. Up until now, the LOD or LPD profiles compared a model with one QTL at the locus under consideration against the null model of no QTL. The MCMC samples allow us to compare models with and without a particular locus while allowing other QTL to be present. Thus, away from the peak, the question is about a *second* QTL allowing for a major QTL found near 100 cM. This type of comparison is analogous to type III ANOVA, in which we test for a second predictor adjusting for the effect of the first being predictor. Simple IM profiles shown earlier are analogous to type I ANOVA, asking about one predictor at a time. CIM approximates this type III ANOVA idea by using co-factors (Figure 5).

MIM provides formal inference on genetic architecture, following a stepwise approach to model building that compares simpler models to more complicated models by adding or dropping main QTLs and/or epistatic QTLs. Thus, at each point, we consider one model at a time. Multiple imputation fits a model with a set number of QTL, again to be

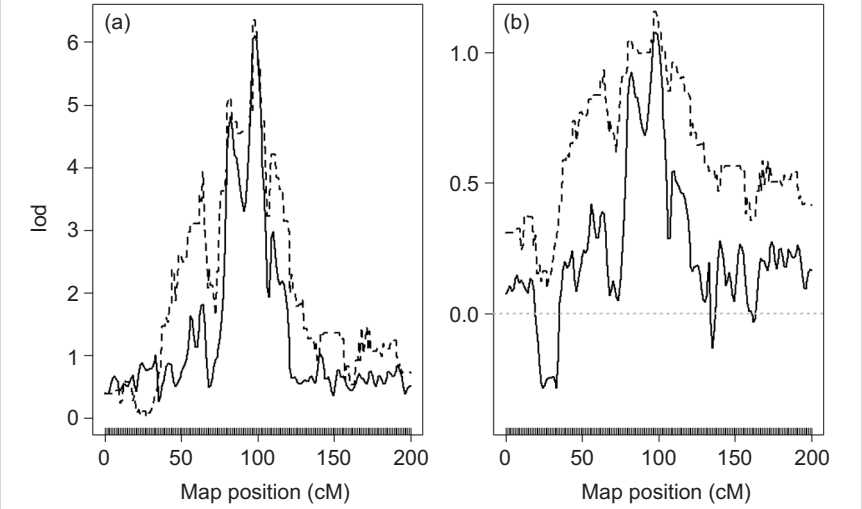

Fig. 6 Bayesian LPD allowing for 2 QTL. Solid curve is from Bayesian posterior, simple interval mapping is dashed. (a) LOD and LPD; (b) Substitution effect.

compared with other genetic architectures through profiles or other summaries. The MCMC approach allows us to sample all possible models, or more exactly the more probably models of the genetic architecture. Thus we have information immediately about a variety of plausible models and can construct summaries to explore these in detail.

2.1.5 Detailed Analysis of Multiple QTL Simulated Cross

We now consider a simulation with four QTL on three chromosomes, including two pairs of epistatic loci (Table 1). We draw a sample of 100 individuals from a backcross, with markers spaced roughly every 10 cM on chromosomes that are 60 cM in length. There is a small amount of data missing at random. The goal is to recover the genetic architecture as much as possible. Our strategy is to consider models with one, two or an arbitrary number of QTL and examine how strong the data are to support them. We give detailed analysis using IM in conjunction with multiple imputation, MIM and BIM via MCMC.

Simple IM picks up strong evidence for a QTL on chromosome 1 and weak evidence on chromosome 2 (Figure 7a and Table 2). The effects are underestimated (Figure 7b). None of this is surprising, since IM only considers one effect at a time. The QTL on chromosome 2 is suggestive

Table 1 Backcross simulation with 4 QTL on 3 chromosomes, and two pairs of epistatic loci, both with QTL 2. Standard deviation was 1; effects were supplied while heritabilities were estimated from simulated sample of 100 individuals.

qtl	*chr*	*pos*	*effect*	*herit*	*qtl2*	*effect2*	*herit2*
1	1	15	1.5	25.6%			
2	1	45	0.0	0.0%			
3	2	12	−1.0	11.4%	2	−2.0	11.4%
4	3	15	0.0	0.0%	2	3.0	26.5%

Table 2 IM one-dimensional summary. LOD scores for single main QTL at best position on each chromosome. The notation "c2.loc15" means chromosome 2, location 15 cM.

	chr	*pos*	*lod*
C1M2	1	15.9	4.156
c2.loc15	2	29.7	2.298
C3M5	3	45.9	0.816

Author: (a) missing?

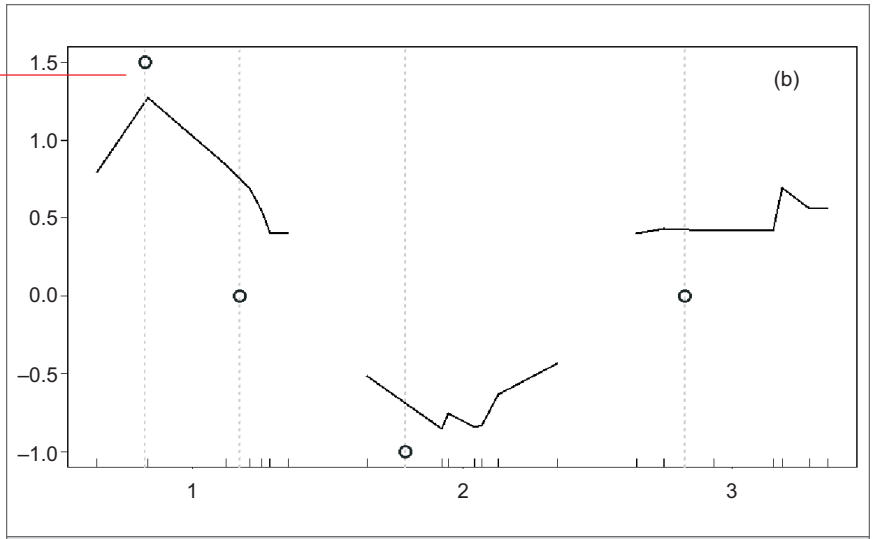

Fig. 7 Simple interval mapping on multiple QTL example. (a) LOD scan with 1000 permutation-based thresholds constructed with method HK; (b) effect scans (line = estimate, circle = true). Vertical gray dashed lines at true location of QTL.

(1% and 5% critical values based on 1000 HK permutations are 2.49 and 1.89, respectively).

A two-dimensional EM scan of the genome shows some of the epistatic effects (Figure 8). Similar scans are obtained with HK and IMP for this multiple QTL simulation. Note the strong evidence for epistasis between chromosomes 1 and 3, and little apparent evidence for any other QTL. All methods pick up the epistasis between chromosomes 1 and 3, but they show little indication of the other epistatic pair (1 and 2). Normally, a 'zscale' would appear to the right, but that was suppressed so that the red lines at true values could be added. Summaries shown in

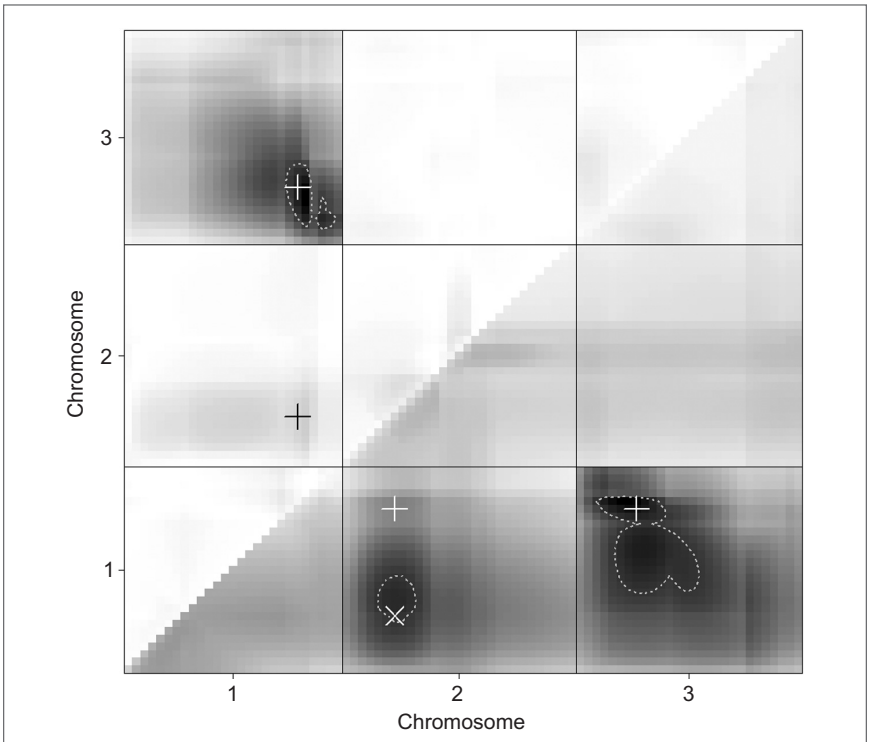

Fig. 8 Two-dimensional profile of simulated data with multiple QTL. Scan is based on EM, but HK and IMP are similar. Contours are 1.5 down from peak. Lower triangle shows LOD comparing full model with epistasis to no QTL. Upper triangle showing LOD of full model to additive model assesses epistasis. Diagonal has LOD for single QTL model. Notice that chromosome 1 and 2 epistasis is barely visible in upper triangle, but appears to contribute in lower triangle. Crosses at true loci pairs; X at mistaken locus pair.

Table 3 support one of the two epistatic pairs (LODs of 8.06 and 1.95, respectively). The 1% critical value based on 1000 permutations (HK method) is 3.27; the *p*-value for the lesser epistatic pair is 0.17. Thus there is strong evidence for the 1 by 3 epistasis, but very weak evidence for the 1 by 2 epistasis. The latter would be kept for exploratory purposes, but probably not believed to be real.

Given the one-dimensional and two-dimensional scans, we now have a hypothetical model. An ad-hoc approach available in R/qtl involves ANOVA averaged over imputed samples (Table 4). This strongly supports a three QTL model with one epistatic pair, with the suggestion for a fourth QTL, on chromosome 2, possibly epistatic to the distal QTL on chromosome 1.

Now let's consider a strategy for building the genetic architecture using MIM, to be performed after these initial one QTL and two QTL investigations. We use WinQTLCart (Basten et al. 1999), as it has a good graphical interface and is free. We fit a new model using the MIM

Table 3 IM two-dimensional summaries. LOD scores for "full" model with two QTL (lod.full), "additive" model with two QTL (lod.add) or epistatic pair adjusted (type II) for main QTL effects (lod.int). Comparisons of full vs. best single QTL (lod.fv1) and additive vs. best single QTL (lod.av1) are also provided. Only entries with LOD > 1 shown.

2-QTL "best" summary evaluated at best full model per pair:

	pos1f	*pos2f*	*lod.full*	*lod.fv1*	*lod.int*	*pos1a*	*pos2a*	*lod.add*	*lod.av1*
c1:c1	15.93	40.6	4.65	0.497	0.247	6.83	11.4	4.41	0.249
c1:c2	20.05	12.8	**8.71**	**4.550**	1.497	17.99	15.0	7.21	3.053
c1:c3	48.12	13.0	**9.78**	**5.620**	4.249	15.93	52.1	5.53	1.371
c2:c2	29.70	36.1	3.45	1.148	0.541	31.80	33.9	2.91	0.607
c2:c3	29.70	60.0	3.04	0.742	0.000	29.70	60.0	3.04	0.741
c3:c3	2.12	24.3	1.53	0.709	0.536	42.88	45.9	0.99	0.173

2-QTL "int" summary evaluated at best epistasis per pair:

	pos1	*pos2*	*lod.full*	*lod.fv1*	*lod.int*	*lod.add*	*lod.av1*	
c1:c1	40.60	51.75	2.73	–1.4263	0.863	1.867	–2.289	
c1:c2	48.12	12.82	5.45	1.2936	**1.954**	3.495	–0.661	
c1:c3	48.12	13.00	9.78	5.6200	**8.060**	1.716	–2.440	
c2:c2	12.82	21.36	3.24	0.9394	1.325	1.913	–0.385	
c2:c3	6.41	4.23	2.25	–0.0508	0.511	1.737	–0.562	
c3:c3	2.12	24.35	1.53	0.7092	0.956	0.569	–0.247	

Table 4 ANOVA on best main QTL and epistatic pairs inferred from 1-QTL and 2-QTL scans. ANOVA table and Type III tests of effects based on 128 imputations. Performed using sim.geno, makeqtl and fitqtl in R/qtl (Broman et al. 2003).

	df	*SS*	*MS*	*LOD*	*%var*	*Pvalue (F)*
Model	6	115.64912	19.2748536	17.54648	56.90287	1.909584e-14
Error	89	87.59038	0.9841616			
Total	95	203.23951				

	df	*Type III SS*	*LOD*	*%var*	*F value*	*Pvalue(F)*	
Chr1@15.9	1	26.155	5.447	12.869	26.576	1.52e-06	***
Chr1@48.1	3	61.268	11.055	30.146	20.751	2.79e-10	***
Chr2@12.8	2	16.559	3.610	8.147	8.413	0.000450	***
Chr3@13	2	51.000	9.565	25.094	25.911	1.36e-09	***
Chr1@48.1:Chr2@12.8	1	6.266	1.440	3.083	6.367	0.013403	*
Chr1@48.1:Chr3@13	1	50.220	9.448	24.710	51.028	2.39e-10	***

forward search method, which is similar to the forward selection from regression applied to markers followed by CIM. We then refine the model by alternatively optimizing QTL positions, searching to add or testing to delete main QTLs and/or QTL interactions between pairs of QTLs. This is somewhat an art form, and can take many steps. There is no guarantee that different paths will lead to the same final model. The model achieved actually included two closely linked QTLs on chromosome 1 near 45 cM with opposite main effects. The BIC criterion (see WinQTLCart or Broman and Speed 2002) accepted all five QTLs, but the two closely linked QTLs were not really believable. For instance, they had high negative correlation of effects, and the one without epistatic effects had a very modest LOD. Therefore, we dropped this fifth QTL and obtained a model with four QTLs and two epistatic pairs, which is quite close to the truth (Table 5). Preserving hierarchy, the least significant effect is the epistasis between chromosomes 1 and 2, which agrees with the design of this simulation. In short, MIM using the BIC criterion almost recovered the correct model, with some subjective intervention during model search and model selection.

Figure 9 shows some one-dimensional summaries of the MIM fit. Each LOD profile is the added contribution of a QTL conditional on the maximum likelihood estimates of the three other QTLs in the model. The red curves are for the three QTLs picked up by MIM on its own. The blue

Table 5 MIM inferred QTL model. Compare estimates and heritabilities with values in Table 1. LOD scores test main QTL or epistatic pair adjusted (type III) for other effects.

qtl	*chr*	*pos*	*effect*	*herit*	*LOD*	*qtl2*	*effect2*	*herit2*	*LOD*
1	1	16	1.16	15.9%	3.26				
2	1	48	−0.05	0.0%	0.01				
3	2	18	−0.96	10.9%	2.42	2	−1.43	6.0%	1.46
4	3	10	−0.09	0.0%	0.03	2	3.00	26.5%	5.39

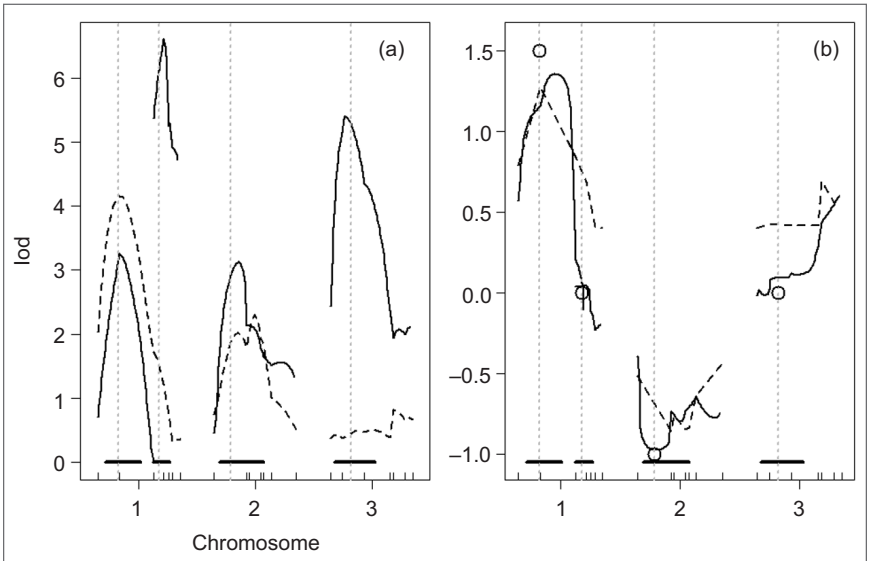

Fig. 9 MIM one-dimensional profiles. Detected QTL as solid lines. IM profile as dashed for comparison. True QTL indicated by gray vertical lines. Horizontal lines at base of figures specify 1.5 LOD support intervals. (a) LOD profiles for contribution of QTL in presence of other QTL (adjusted type III test from MIM, solid) or on its own (single QTL type I, IM, dashed); (b) main effects profile, with true values added as circles.

curve corresponds to the second QTL on chromosome 1. The IM profile is included for comparison (black). Estimates of main effects of QTL are shown in Figure 9b, with true values in purple. Currently, there is no two-dimensional graphic for MIM fit.

The strategy for model selection with BIM using MCMC samples is somewhat different. We first draw many samples (default is 120,000, saving every 40th) from the more probably genetic architectures. We then

use Bayes factors and Bayesian model averaging to uncover evidence for the better models. The one-dimensional and two-dimensional summaries are different as well. They measure the contribution of a particular locus (or pair of loci for 2-D scans) *after adjusting for all other possible QTL.* That is, these are adjusted LPD, allowing for multiple QTL and averaging over all possible genetic architectures. This is distinct from IM scans, which have no adjustment for other QTLs; it is also distinct from MIM scans, which fix the genetic architecture for a set number of QTLs when scanning a particular QTL. The LPD peaks on Figure 10a are higher than for MIM, because other QTL effects are adjusted by model averaging rather than conditioning on the maximum likelihood estimates. The properties of these BIM scans are an area of active research.

The R/qtlbim software can separate linked effects, although effects for linked QTLs are averaged together in the 1-D projections of Figure 10. The estimates of genotypic effects for the main QTL (Figure 9b) are again close to the true values. Estimates of epistatic effects projected onto each QTL in Figure 10b should be interpreted with caution—better to view them with a 2-dimensional scan (not shown). For instance, the epistatic

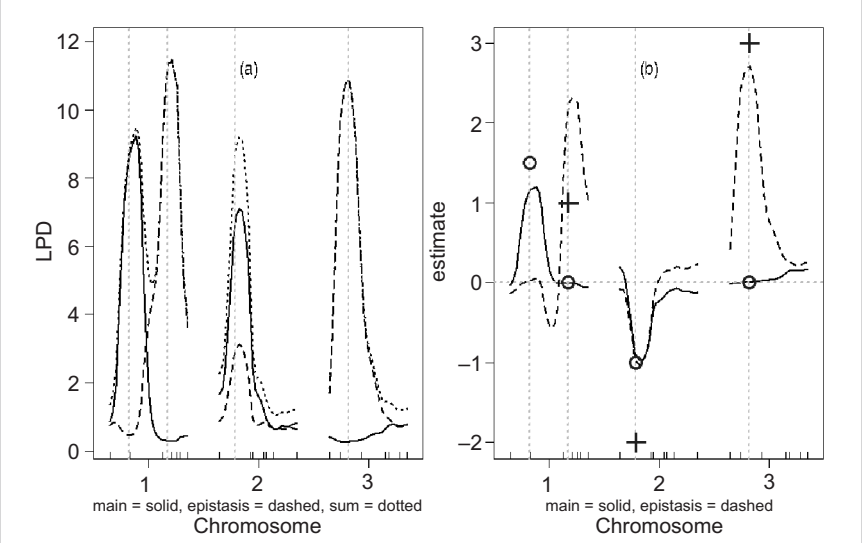

Fig. 10 BIM one dimensional profiles. (a) log posterior density (LPD) for combined effects (dotted), main QTL (solid) and epistatic effects (dashed); (b) main QTL (solid line = estimate, circle = true) and projected epistatic effects (dashed = estimate, cross = true). Contributions of epistatic pairs are shown at both loci, although it is not possible with this representation to determine which loci are paired.

effect on chromosome one distal end is an average of the two pairs of epistatic effects.

Table 6 shows one dimensional BIM summaries. The expected number of QTLs (n.qtl) is estimated from the posterior distribution of the number of QTLs (Yandell et al. 2007). That is, MCMC samples had on average 2.66 QTLs on chromosome 1 and almost two on chromosomes 2 and 3. It appears from Figure 10 that these extra QTLs are not major contributors, as the peaks for contributions from main QTL and epistatic QTLs are unimodal.

Figure 11 and Table 7 show a two-dimensional summary of LPD. Note how adjustment for other QTLs leads to a tightening of peaks relative to Figure 8. Once again, the contribution of the epistatic interaction between chromosomes 1 and 2 is not very strong, although it is more evident in Figure 11 than in Figure 8.

Other summaries can be useful. Figure 12 profiles Bayes factors, rescaled as 2log(BF), and means by genotype. Values of 2log(BF) above 2.1 are considered significant; values below zero are truncated. Other possibilities include the posterior intensities, variance estimates, or heritabilities.

Figure 13 shows posteriors and Bayes factor ratios for several important summaries. The posterior mode for number of QTLs is 7, but there is only weak to moderate evidence for more than 4 QTLs (BF ratios of ~3 comparing 4 to 5-9). Several chromosome patterns are equally

Table 6 BIM one dimensional summaries. Separate tables for log posterior density (LPD) and estimate of genotype effects. n.qtl is expected number of QTL on chromosome; pos are positions of the peak total effects per chromosome; m.pos are positions of main peaks in posterior; e.pos is position of epistatic peak.

LPD of pheno.normal for main, epistasis, sum

	n.qtl	*pos*	*m.pos*	*e.pos*	*main*	*epistasis*	*sum*
c1	2.66	48.1	15.9	48.1	8.86	11.78	11.81
c2	1.93	15.0	15.0	15.0	7.14	3.15	9.27
c3	1.99	10.7	60.0	10.7	0.79	10.70	10.72

estimate of pheno.normal for main, epistasis

	n.qtl	*pos*	*m.pos*	*e.pos*	*main*	*epistasis*	
c1	2.66	48.1	15.9	48.1	1.173	2.36	
c2	1.93	15.0	15.0	15.0	–0.977	–1.04	
c3	1.99	10.7	60.0	10.7	0.165	2.65	

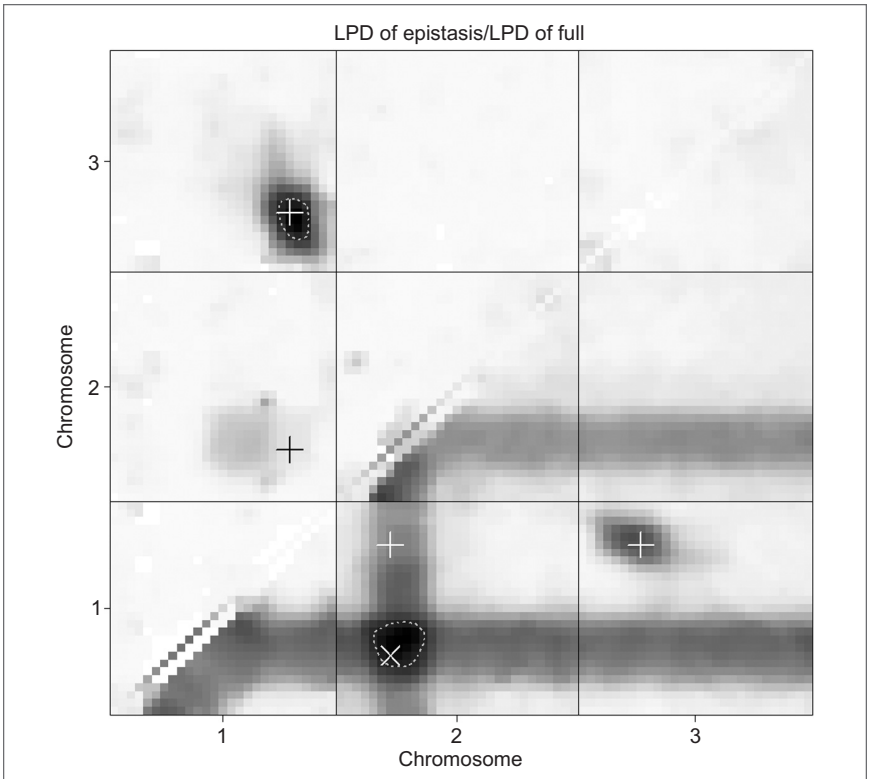

Fig. 11 BIM two-dimensional summary. Lower triangle has LPD contribution of both QTL plus epistasis; upper triangle has contribution of epistasis only. Crosses indicate locations of true loci pairs; X at mistaken loci pair. Gray contours are 1.5 LPD support interval.

Table 7 BIM two-dimensional summaries. n.qtl is expected number of QTL on chromosome; l.pos are positions of main peaks (lower triangle) in posterior; u.pos are positions of epistatic peaks (upper triangle).

uupper: LPD of pheno.normal for epistasis

lower: LPD of pheno.normal for full

	n.qtl	*l.pos1*	*l.pos2*	*lower*	*u.pos1*	*u.pos2*	*upper*
c1:c1	2.34	24.2	32.38	10.91	6.83	57.19	1.398
c1:c2	5.18	20.0	17.09	14.51	38.55	25.50	4.598
c1:c3	5.37	48.1	13.00	11.32	48.12	13.00	11.292
c2:c2	1.23	0.0	10.68	11.89	4.27	36.13	2.846
c2:c3	3.97	15.0	60.00	8.19	6.41	4.23	0.856
c3:c3	1.32	0.0	6.35	2.17	0.00	6.35	1.840

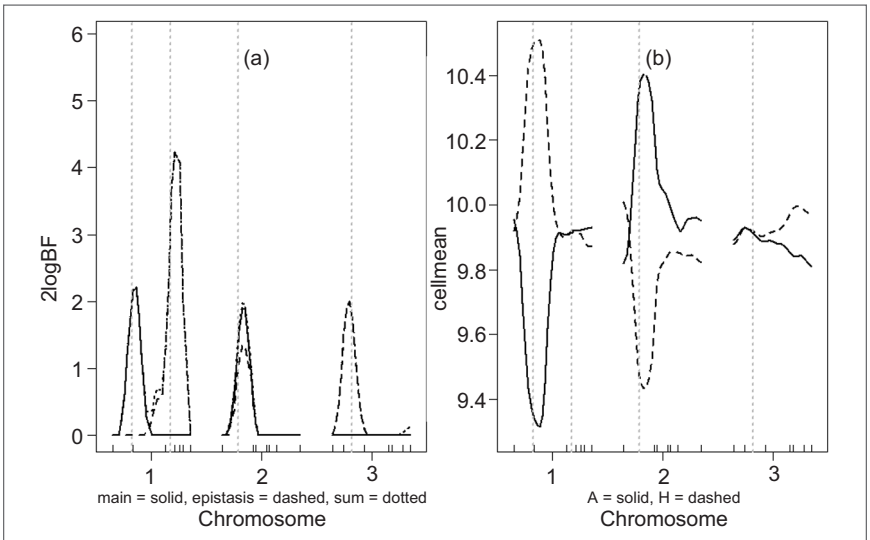

Fig. 12 Further BIM one dimensional profiles. (a) Bayes factors as 2log(BF), with main QTL (solid), epistatic effect (dashed) and combined effects (dotted, often hidden behind main or epistatic). Contributions of epistatic pairs are shown at both loci, although it is not possible with this representation to determine which loci are paired. (b) Means by genotype (solid = AA genotype; dashed = AB genotype).

favored, but the simplest has two on chromosome 1, one on chromosome 2, and one on chromosome 3 (coded as 2*1,2,3). Finally, the epistatic pairs 1.3 and 1.2 have the highest posteriors and the highest Bayes factors relative to any other epistatic pairs. In short, these summaries support the true model.

2.1.6 Covariates and Gene-Environment Interactions

Rarely is an experimental cross conducted in a single environment with all individuals handled identically, leading to measurements of just one phenotype. Complications arise, planting times vary, or there are broader scientific questions about differences across environments. We often do not have an ideal measurement of the characteristic we are most interested in studying. Instead, we measure multiple traits that are correlated, hoping that one of them will show strong heritability. It can be useful to think of such multiple correlated traits as covariates, measurements that covary with each other. Covariates are very important in understanding the effect of genotype on phenotype. First, including covariates can reduce residual variation and potentially enhance the power to detect QTL. Second, there may be important

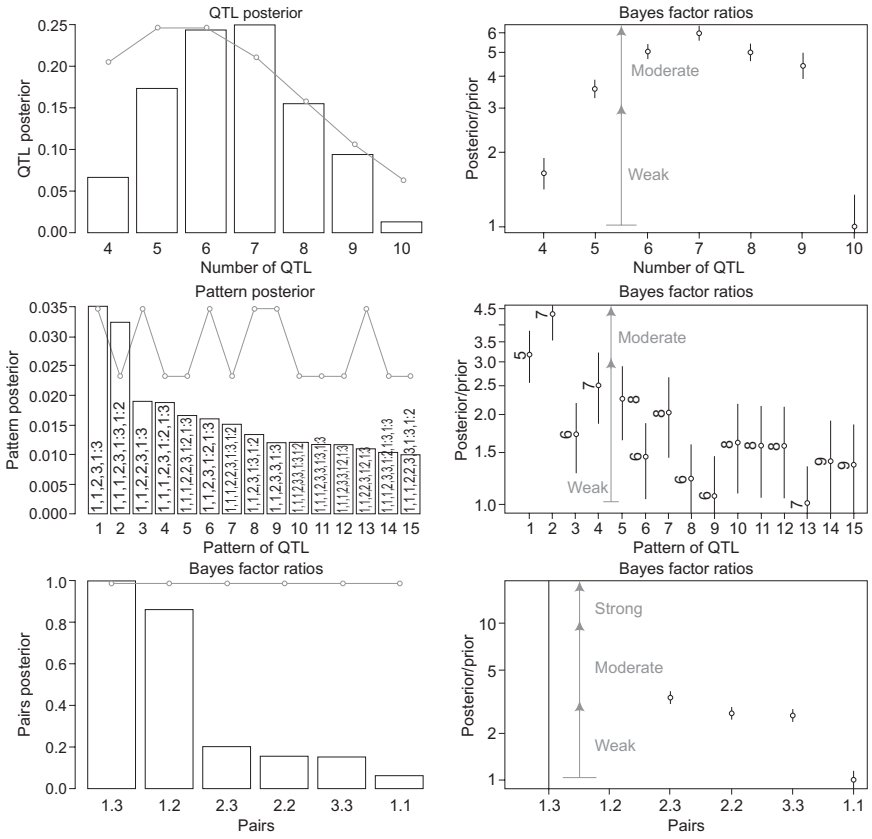

Fig. 13 Bayes factors for (a-b) number of QTL, (c-d) chromosome pattern of QTL and (e-f) epistatic pairs. Prior is rescaled and overlaid on posterior in left panels; vertical arrows in right panels indicate 3/10/30-fold ratios for Bayes factor comparisons.

differences in genotypic effects that depend on covariates or the environment. It is beyond the scope of this chapter to examine GxE interactions and multiple trait analysis in detail. Instead, we highlight a few important issues concerning **adjustment for covariates** and **genotype by environment (GxE)** interaction, indicating how they may be assessed with current software.

The most important covariates for plant breeding typically involve aspects of the environment. That is, different genotypes might perform best (in terms of the phenotype) in different environments. There are basically two types of experiments with crosses over different

environment (Paterson et al. 1991; Stuber et al. 1992). Design I has individuals from one cross evaluated in two or more environments, while Design II has different genetic material in each environment. Many experiments have multiple traits measured on each individual. We might view these multiple traits in some sense as evaluating the same genetic material in different environments. That is, measurements of multiple traits per individual and measurements of a single trait in multiple environments for an individual (or clones or progeny) can be analyzed in much the same way as a Design I experiment.

Design I experiments with measurements on two or more environments, or correlated traits, for each individual in a cross can be examined in a variety of ways. For instance, we can first scan each trait on its own, noticing where peaks appear in common. However, we should not be surprised to find the same genomic regions cropping up, as we probably started with correlated traits. We can subsequently consider one trait as a covariate for the other, particularly when traits are over time or there is some other suggestion of causation. A dramatic reduction of peaks after adjusting for a covariate provides evidence that the QTL has an indirect effect through that covariate (although recall that statistical evidence in itself does not imply causation!). Alternatively, additional QTL can be detected after adjusting for covariates that reduce residual variation (*cf*. Li et al. 2006; Stylianou et al. 2006;). The primary advantage of joint analysis of multiple phenotypes is the ability to distinguish between **pleiotropy**, where one gene affects many traits, and **close linkage** of QTL that independently affects separate traits (Jiang and Zeng 1995; Vieira et al. 2000; Li et al. 2006). Another approach is to combine correlated traits using some multivariate approach such as principal components (Liu et al. 1996).

Any category that divides a cross up into groups can lead to a Design II situation. Sex in animals and dioecious plants can be viewed as an example of Design II. Stratifying by age (young, old) or on experiments done over time are other ways. The key for Design II is that different, independent individuals are evaluated. While it is helpful to examine each 'environment' separately, this leads to a reduction in power to detect QTL: no evidence of QTL is inconclusive regarding GxE interaction. We must conduct a combined analysis adjusting for environment using all the data to properly assess GxE interaction. Be sure to adjust for the ***interaction* of genotype and environment** (known as GxE, or interacting covariates) rather than just for the main effect of environment. An excellent example of this is found in the work of Solberg et al. (2004, 2006).

Thus, the effect of QTL may depend on the value of the covariate for both Design I and Design II. The formal LOD (or LPD) score for a QTL allowing for GxE assesses both the effect of QTL on phenotype and the interaction of QTL with the covariate. Separately we can test if the GxE effect adds anything to assess the influence of the covariate. The papers cited below give examples using available software, primarily Pseudomarker or R/qtl (Solberg et al. 2004, 2006; Li et al. 2006; Stylianou et al. 2006), WinQTLCart (Jiang and Zeng 1995; Vieira et al. 2000) and R/qtlbim (Yi et al. 2005). Many packages back to MapMaker/QTL (Lander et al. 1987; Lander and Botstein 1989) have allowed some form of adjustment for covariates. However, only a few packages appear to have full interacting covariate GxE adjustments. Pseudomarker (Sen and Churchill 2001), R/qtl (intcov option to scans; Broman et al. 2003) and R/qtlbim (intcov for MCMC sampling; Yandell et al. 2007) allow adjustment for covariates to individual phenotypes. WinQTLCart (Basten et al. 1999) and MultiQTL (Mester et al. 2004) conduct multiple trait mapping of a modest number of correlated traits, providing joint LODs that formally test a QTL for *any* considered trait.

Expression QTL (eQTL) studies are now appearing with thousands of traits per individual in an experimental cross. WebQTL (Wang et al. 2003) is a handy, intuitive tool designed for expression traits, although it is largely focused on HK regression and correlation among traits. See Lan et al. (2006) for one way to extend this approach.

2.1.7 Overview of Available Packages

A number of packages were created for QTL analysis of inbred lines in the late 1980s and early 1990s. A handful of those survive today, some of them static and some of them under continual development. We focus attention here on the more current packages, with occasional reference to historical packages. Terms used here to describe methods and properties are explained in more detail above.

Marker regression can be used in any statistical package, and in the QTL packages QTLCart (Basten et al. 1999) and R/qtl (Broman et al. 2003).

The package MapMaker/QTL (Lander et al. 1987; Lander and Botstein 1989) greatly modified the conceptual framework for gene mapping, and its ideas are central to all other packages found today. MapMaker/QTL is still available as the original source (the Windows exe file does not appear to work under Windows/XP). Most users seeking to conduct simple interval mapping via the EM algorithm now rely on currently maintained packages that have incorporated the interval mapping algorithm, with slight variations, including

WinQTLCart or QTL Cartographer (Basten et al. 1999), R/qtl (Broman et al. 2003), MapQTL (van Ooijen and Maliepaard 1996) and MultiQTL (Mester et al. 2004).

Haley-Knott regression is used in QTL Express (Seaton et al. 2002), PLABQTL (Utz and Melchinger 1996), MapManager/QTX (Meer et al. 2004), and WebQTL (Wang et al. 2003). It is also available as an option in R/qtl (Broman et al. 2003) and R/qtlbim (Yandell et al. 2007).

Table 8 shows a summary of key features in packages for inbred lines. All except MultQTL and MapQTL are free. Most of the free packages have an open source, which means you can examine and

Table 8 Comparison of packages for inbred lines. Platform: W=Windows, L=Linux, M=MacOSX. Analysis method: B =Bayesian interval mapping, EM = expectation-maximation of likelihood, HK = Haley-Knott regression (marker regression only for R/bqtl). Platform is standalone (solo), R statistical system, Matlab, the Web (Java), or another package on the list. Most packages have a graphics user interface (GUI), typically coming from Windows, the Web (Java), or a platform application (R or Matlab). Out indicates capability to handle some outbred populations; * use blocking factors or a limited outbred option (e.g. 4-way cross in R/qtl). Most software is Free; some providing source, others providing applications only. GxE for interacting covariates or multiple trait mapping. X for presence, O for absence, * for limited ability, ? for unknown.

Package	*W*	*L*	*M*	*B*	*EM*	*HK*	*Platform*	*GUI*	*Out*	*Free*	*GxE*
MapMaker/ QTL	X	X	X	O	X	O	solo	O	O	X	*
Pseudomarker	X	X	X	X	O	O	Matlab	X	*	X	X
R/qtl	X	X	X	O	X	X	R	J/qtl	*	X	X
R/qtlbim	X	X	X	X	O	X	R/qtl	O	*	X	X
R/bim	X	X	X	X	O	O	R/qtl	O	*	X	O
R/bqtl	X	X	X	X	O	*	R	O	O	X	O
QTLCart	X	X	X	O	X	O	solo	WinQTL	O	X	X
WinQTL	X	O	O	X	X	O	QTLCart	X	O	X	X
Webqtl	X	X	X	O	O	X	Java	X	O	X	O
PLABQTL	X	O	O	O	O	X	solo	O	O	X	*
MapManager/ QTX	X	O	O	O	O	X	solo	X	O	X	O
MultiQTL	X	O	X	O	X	?	solo	X	*	O	X
MapQTL	X	O	O	O	X	?	solo	X	X	O	*
QTLExpress	X	X	X	O	O	X	Java	X	X	X	O
QTLCafe	X	X	X	O	O	X	QTLExpress	X	X	X	O

modify the code used for the QTL analysis. To our knowledge, only R/qtl properly handles the X chromosome (Broman et al. 2006). Some packages have extensive manuals, with screen shots available at their web sites. See the references for current URLs to individual packages.

Some software only available in source form is not included in Table 8 Multimapper (Sillanpää and Arjas 1998) conducts Bayesian interval mapping and model selection, with summaries in terms of the posterior intensity per locus. DIRECT (Ljungberg et al. 2004) is a very fast algorithm for solving linear models, and it has been incorporated into R/qtl and WebQTL. QTLNetwork (Yang et al. 2005) is a recent package with appealing graphics, but it is poorly documented to date, and the underlying methods of analysis are unclear.

Of the packages listed in Table 8 only QTLCart, MultiQTL, and R/qtlbim fully consider model selection with an arbitrary number of QTLs and epistasis. The former two use MIM while the latter uses Bayesian model averaging over possible genetic architectures. The R/qtl and Pseudomarker packages have some tools for arbitrarily large genetic architectures, but primarily focuses on two QTL with epistasis. R/bim allows for multiple QTL in a Bayesian model averaging framework, but cannot handle epistasis. QTLExpress does a limited investigation of epistasis for pairs of linked QTL. PLABQTL and MapQTL employ CIM to adjust for other QTL when conducting a 1-QTL profile. MapManager/QTX and WebQTL rely on user-supplied markers to manually adjust for other QTL, somewhat analogous to CIM.

Several packages now employ Bayesian methods for interval mapping, most built on the R system (R Core Development Team 2006). R/qtl (Broman et al. 2003) includes multiple imputation, in addition to classical methods mentioned above. R/bqtl (Borevitz et al. 2002) was an early entrant, using marker regression in a Bayesian framework. The packages R/bim (Satagopan et al. 1996; Satagopan and Yandell 1996; Gaffney 2001) and R/qtlbim (Yi et al. 2005: Yandell et al. 2007) estimate the full posterior for models involving an arbitrary number of loci that may be in intervals between markers. R/qtlbim allows for epistasis and gene by environment interaction (see section 1.4). Pseudomarker and R/qtl both incorporate multiple imputation (Sen and Churchill 2001), with graphics for two QTLs and some tools for examining more that two QTLs.

The packages MultiQTL, R/qtl, Pseudomarker, and R/qtlbim all handle gene by environment (GxE) interactions, also known as interacting covariates. Other packages such as PLABQTL and QTLExpress handle covariates in a limited way.

Each package for QTL analysis of inbred lines has its own data input format. Several packages allow multiple importing formats, or can export data in a few different forms. This step is the biggest headache about using packages—figuring how to get your data in. Fortunately, most packages include well-documented examples. Further, most authors are open to email questions about package use. The MapMaker/QTL format is widely used, but it is not easy to set up the first time. The CSV format used in R/qtl allows one to build data in a spreadsheet, in a format that can be opened by Excel or by R. The R/qtl package has several other input and output formats, and it is not that difficult (with the help of a programmer) to customize output from R to most other packages.

Nothing has been said yet about non-normal phenotypes. There are some papers in this area, but only modest availability in packages to date. Nonparametric analysis (Kruglyak and Lander 1995) basically involves replacing trait values by their ranks; it is available in several packages. R/qtl includes binary traits, and R/qtlbim can handle ordinal traits (qualitative rankings such as poor/fair/good/excellent). Semi-parametric methods have been developed but are not broadly available yet (see Jin et al. 2007). Other approaches, such as Poisson regression, have been used in specialized software that, to our knowledge, has not been released.

2.1.8 QTL Analysis with Outbred Lines

This section is quite brief. Some packages such as QTLExpress, MapQTL and PLABQTL handle certain types of outbred crosses, including full sibs, half sibs and other relatedness designs. Some of these can combine different inbred crosses (*cf.* QTLExpress). SOLAR (Almasy and Blangero 1998) is a general purpose linkage and QTL mapping package using identity by descent (IBD) that works well for modest sized pedigrees. HAPPY (Mott et al. 2000) is specifically designed for heterogeneous stocks created from known founders.

In a way, QTL mapping for inbred lines can be adapted to experimental crosses (e.g. backcross or intercross) based on outbred founders. The easiest way is to use markers that distinguish among the founders. There is some loss of precision, as the QTL genotype can have more than two alleles. If your experiment is an F1 resulting from crossing two outbred founders, and the phases (haplotypes) of the founders are known, one can use the '4way' cross type in R/qtl for analysis and follow methods detailed above.

The QTL analysis methods with inbred lines detailed in this section are in theory extendable to outbred population, taking care of IBD and

multiple alleles. All the problems and subtleties encountered above carry over and become harder. Further investigation of outbred crosses is beyond the scope of this chapter.

3 ASSOCIATION ANALYSIS

While an in-depth description of association analysis is beyond the scope of this chapter, several recent reviews describe association analysis in some detail (Flint-Garcia et al. 2003; Gupta et al. 2005; Hirschhorn and Daly 2005; Breseghello and Sorrells 2006; Yu and Buckler 2006). The review by Gupta et al. also provides a list of software as supplemental electronic material.

Association analysis is also called linkage disequilibrium (LD) mapping because it uses the extent of LD between a trait and markers to identify and find the location of QTLs. Such an association might imply that either the marker or some polymorphism linked to it is the cause of the observed phenotypic variation. However, two problems exist with this reasoning. First, a linked marker might not be in LD with the trait. Second, a marker that is in LD with a trait might not be linked to it. The story of the development and refinement of association analysis, and of the software for performing it, is largely a story of how these two problems have been addressed.

In the first case, LD between a linked marker and a quantitative trait locus (QTL) will be difficult to detect when the frequency of the marker is very different from that of the QTL. Presumably, sequence polymorphisms of some type underlie both markers and QTLs. A marker is simply a polymorphism that can be detected by an assay. A QTL is a polymorphism that causes a measurable change in phenotype. In the ideal situation, the marker and QTL are the same polymorphism and consequently have the same frequency. Otherwise, a closely linked marker may have a very different frequency in a population and as a result not be very useful for detecting a QTL.

This problem can be addressed by increasing marker density to make it more likely that at least one linked marker will be in LD with the QTL or be the QTL itself. For example, using the candidate gene approach to association mapping, an entire gene may be resequenced and all the sequence polymorphisms identified. For genome-wide scans, of course, that approach is not feasible. Another approach has been to use haplotypes instead of single markers to look for associations. Since the number of haplotypes will generally be greater than the number of individual markers, there may be more opportunities to match the frequency of the QTL.

The second problem with association analysis is that unlinked markers may be in LD with a trait of interest. LD structure can be affected by a number of factors including mutation, recombination, selection, mating patterns, and population admixture. Strategies for dealing with these problems include genomic control (GC), structured association (SA), family-based studies, and the use of marker data to correct for polygenic background effects.

In spite of these drawbacks, association analysis has been gaining interest rapidly in the plant genetics community. The key advantages of association analysis that are driving interest in this approach include the ability to use existing germplasm without the need to develop special populations, the ability to survey the diversity of alleles present in a broad-based population instead of being restricted to those present in the two parents of a mapping population, and the ability to map with high resolution. Resolution is determined by how rapidly LD decays in the population being sampled and can vary greatly depending on the species being sampled. For example, LD has been found to span just 1 kb in a diverse maize population, over 100 kb in a population of US elite maize inbred lines, and about 10cM in sugarcane, a vegetatively propagated species (Flint-Garcia et al. 2003).

As the use of association analysis, especially in plants, is a relatively recent development, the methodology and software is still undergoing development. No standard, accepted methodology exists and, consequently, standard software does not exist either. Nonetheless, software has been released that is useful for association analysis and related tasks, though knowledge is required on the part of the user to be sure that the analysis is appropriate. Related tasks include inference of population structure, derivation of measures of relatedness between individuals, haplotype inference, and analysis and visualization of LD structure.

Most of the software developed to date for association analysis has been developed for human genetics. Often that software is not directly useful for plant genetics. First, two tools widely used by plant geneticists, planned crosses and inbreeding, are not available in human genetics. As a result, family structures in human studies tend to be quite different from those in plant studies. The result is that the methods of analysis best suited for human genetics studies are often not optimal for plant genetics. Second, a lot of human genetics studies involve case-control studies for diseases. Phenotype data from case control studies is binary, affected versus unaffected. Relatively little plant phenotype data is binary.

While some very useful software has been developed for association analysis of case-control studies, it has been left out of this review since it is not likely to find much use by plant geneticists. On the other hand, some of the family-based methods could be used for plant species that are both naturally cross-pollinated and difficult to inbreed. As a result some software that uses family-based methods has been included.

Association analysis without some form of correction for population structure is straightforward and can be run using a variety of general statistical software packages. For example, Czika and Yu (2004) describe how to use SAS to perform marker-trait association tests for unrelated individuals or for populations with known family structure. In addition, freely available software has been written for association analysis that uses various strategies to cope with population structure. This software includes TASSEL, Powermarker, QTDT, MTDFREML, GC, BAMA, and TreeLD.

TASSEL (Yu et al. 2006) is written using Java and, as a result, runs on most computing platforms. It has an elaborate graphic user interface with a large number of functions for data management, analysis, and visualization. It accepts input either as text files or by way of the GDPC (Genomic Diversity and Phenotype Connection) browser (Casstevens and Buckler 2004), middleware providing a web connection to databases that have been made available through a GDPC server. Data imported independently from different sources can be combined for analysis. In addition, TASSEL will extract SNPs and indels from aligned sequence using flexible filtering criteria.

TASSEL has a number of analysis routines. Most notably, it implements a mixed model approach to association analysis (Yu et al. 2006) that uses both large-scale population structure and pair-wise kinship coefficients derived from marker data to correct for population stratification. The mixed model function, called MLM, requires a matrix of kinship coefficients to correct for population substructure and can optionally incorporate a population structure matrix or Q-matrix. TASSEL can calculate kinship coefficients for homozygous inbred lines. For heterozygotes, the kinship matrix can be calculated from marker data using the program SpaGeDi (Hardy and Vekemans 2002). The Q-matrix can be calculated using the program STRUCTURE (Pritchard et al. 2000). Structured association analysis, which uses the Q-matrix but not the kinship matrix, can be carried out using a fixed-effect linear model using the GLM function or logistic regression. In addition, TASSEL can be used to calculate and display LD measures for pairs of markers, calculate an evolutionary tree, or cladogram, and calculate population diversity statistics.

Powermarker (Liu and Muse 2005) is another example of multi-function software with a well-designed graphic user interface. Written in MS Visual C++ under the Microsoft .NET framework, Powermarker requires the MS Windows operating system. Data can be input from either text or Excel files. Powermarker provides data management functions and a number of descriptive genetic statistics, including measures of LD, population heterozygosity, inbreeding coefficients, tests of Hardy-Weinberg equilibrium, and Wright's F-statistics. It calculates genotype and allele frequencies and estimates haplotype frequencies. For association analysis, Powermarker will calculate an F-test for association between individual markers and traits and will perform Zaykin's haplotype trend regression (HTR) (Zaykin et al. 2002). Both of the association analyses assume a homogenous population without underlying structure or stratification.

QTDT (Abecasis et al. 2000) is software that implements the quantitative transmission disequilibrium test, which uses family structure to correct for population stratification. The analysis method uses a maximum likelihood approach to partition the genetic effects into within and between family components. The within family component provides an estimate of the genetic effect of a marker that is free of any population structure effect. As it makes use of family structure, QTDT, in effect, combines association and linkage analysis. QTDT analyzes data from nuclear families. At a minimum, it requires trios of parents plus one offspring or full sib pairs. Larger sibships can be analyzed as well. Other software, such as Merlin (Abecasis et al. 2002), must be used to calculate the IBD matrix required by the analysis. It also requires map positions of the genetic markers. QTDT has a command line interface, uses text files for input and output and has been compiled for Windows, Linux, and SunOS.

MTDFREML (Boldman et al. 1995) is software designed to set up and solve mixed model equations with individuals treated as random effects and an additive genetic relationship matrix used to define the covariance between individuals, similar to the model used in TASSEL's MLM function. The original program calculates the inverse of the additive genetic relationship matrix directly from pedigrees. It was recently modified (Zhang et al. 2006) to use a relationship matrix calculated directly from markers. MTDFREML can be used to estimate variance components, predict breeding values, and estimate associations between markers and phenotypes. The software is quite flexible, can incorporate covariates, and analyze multivariate models. Written in FORTRAN, it must be compiled by the user. In addition, it requires either FSPAK or SPARSPAK, which are sparse matrix libraries available from other

sources. The interface is command-line. Text files are used for both input and output. The paper by Zhang et al. (2006) gives an example using data for canine hip dysplasia that uses MTDFREML to compute marker effects and other genetic parameters.

Genomic control (Bacanu et al. 2002; Devlin and Roeder 1999) is a method of correcting for unknown population structure by using bi-allelic markers distributed across the genome, that are not expected to be linked to QTLs. This method calculates a correction factor, λ, based on the values of a test statistic at the null loci. The test statistics for the loci being tested for association with a trait are then divided by λ. The programs, GC and GCF, are implemented in the R statistical programming language. Both programs estimate λ, then use that to construct a test of either a single locus or a pair of loci plus interaction. Devlin et al. 2004 recommend using GCF when testing a large number of candidate loci or when the required a-level is small.

BAMA provides a Bayesian solution which tests multiple loci simultaneously (Kilpikari and Sillanpaa 2003). As such, it avoids problems with multiple testing and over-estimation of the effect size when using the same data for detection and estimation. It assumes that the population being investigated has no substructure. The program is distributed as C source code, which must be compiled by the user, is designed for Linux or Unix operating systems, uses a command-line interface and text files for input and output.

TreeLD (Zollner and Pritchard 2005) provides an approach to association analysis that is different from the other programs mentioned. First, the method uses marker data to model the ancestry as a set of coalescent trees. Next, association between markers and phenotypes is evaluated by looking at the distribution of phenotypes among the tips of the trees. The authors demonstrate the effectiveness of the method using both simulated and real datasets. The method should not be used when population structure issues exist. The software requires phased genotypes and marker positions as input, though PHASE, described below, has been used to infer haplotypes from diplotype data with unknown phase. TreeLD is very computationally intensive. Documentation at the TreeLD website (Pritchard) notes that a thorough analysis of a data set with 250 individuals and 130 markers required 48 hours on a 10 processor Linux cluster.

As suggested above, haplotypes may have advantages over individual SNPs for conducting association analyses. In fact, haplotypes could be regarded as multi-allelic markers. Consequently, they have greater information content than individual SNPs. Buntjer et al. 2005 discuss the use of haplotypes in plant association analysis. For inbred

lines, haplotypes can be identified directly. For homozygous individuals, TASSEL has a function that derives haplotypes from SNPs using a sliding window. Those haplotypes can then be used as markers in association analysis. For heterzygotes, however, haplotypes must be inferred. Separate software can be used to infer haplotypes, and the resulting haplotypes can be used as input for another analysis platform. Powermarker, mentioned earlier, uses an EM algorithm for inferring haplotypes. PHASE and Haplotyper use two different Bayesian methods to identify haplotypes.

The LD viewers, a related category of software, can be helpful for visualizing haplotypes or for interpreting the results of association analysis. Examples of LD viewers include Haploview (Barrett et al. 2005;Wu et al. 2006), MIDAS (Gaunt et al. 2006), JLIN (Carter et al. 2006), LDA (Ding et al. 2003), and GOLD (Abecasis and Cookson 2000; Ding et al. 2003). Haploview is easily the most versatile of these programs. It accepts phased chromosomes or unphased diplotypes as input. Family structure information can be incorporated but is optional. It will calculate and graph several pairwise LD measures, including D' and r^2. The user can select groups of markers for haplotype analysis or have Haploview automatically generate haplotypes. Haplotypes and haplotype frequencies can be output, but the user will have to recode genotypes using that information for further analysis. As Haploview's association testing is restricted to case-control and TDT trios, it will be of limited value for analyzing plant genetics data. GOLD provides nice graphics but is designed for use with human genetics data and may be challenging to adapt to other types of data. Both LDA and JLIN take SNP data with unknown phase as input and calculate and graph pairwise LD measures but provide no other functionality. MIDAS (Multiallelic Interallelic Disequilibrium Analysis Software) is unique in a couple of respects. First, it is designed to evaluate multiallelic markers rather than SNPS only. Second, it is a Python program and uses Tkinter for its graphic user interface. Haploview, JLIN, and LDA are all java applications. As mentioned earlier, both TASSEL and Powermarker calculate pairwise LD statistics and display the results graphically.

A final category of software which is required for SA assigns members of a population to subpopulations or mixtures of them. By far the most widely used of these is STRUCTURE (Pritchard et al. 2000). It uses a Bayesian model to construct subpopulations that minimize LD within a set of unlinked markers. While it requires the user to choose the number of subpopulations, the software can be used to estimate the number of subpopulations by running the analysis multiple times with different numbers of subpopulations then choosing the number that

produces the best fit. PSMIX (Wu et al. 2006) is an R package that uses a maximum likelihood approach that is solved using an EM algorithm. It is computationally less intensive than STRUCTURE and, the authors find that it gives comparable results. The PSMIX article provides a good review of some additional programs for finding population substructure.

3.1 Association Analysis Examples

We illustrate the use of some of the association analysis software and highlight some of the issues that arise when doing this type of analysis with examples taken from studies of maize. In the following examples, the trait being analyzed is days to silk for maize inbred lines. The lines used were chosen to represent as much of the diversity present in the species as possible while restricting them to a manageable range of flowering dates. The trait 'days to silk' was chosen because flowering time often associates strongly with population structure. As plants must flower at roughly the same time in order to be cross-pollinated, populations tend to become stratified into flowering time or maturity groups.

Measurements were taken in 1999, 2000, and 2001 in Clayton, NC. The days to silk data and sequence for the dwarf8 gene was downloaded from the Panzea database (www.panzea.org) using the middleware application GDPC (Casstevens and Buckler 2004). The matrix of kinship coefficients and the population parameters used are part of the TASSEL tutorial. The population parameters were derived from SSR data using the program STRUCTURE (Pritchard et al. 2000). The kinship coefficient matrix was calculated using the program SPAGeDi (Hardy and Vekemans 2002) using data for 553 random SNPs. The random SNP data is also available as part of the TASSEL tutorial. Statistical analyses were run using TASSEL.

Using good quality phenotypic data with reasonably high heritability is critical to association analysis. While beyond the scope of this chapter, using principles of experiment design, checking for outliers and unreliable data, making certain that important assumptions are not violated, and following accepted statistical procedures are important steps in producing good phenotype data. For example, in each year that the days to silk data were taken, two or three replicates of data were taken but in some cases much data were missing from some of the replicates. Consequently, using least square means as estimates of days to silk within each year or across years was better than using simple averages.

The effectiveness of the analysis depends in part on how the kinship coefficients are derived. For this data, using SPAGeDi to calculate kinship coefficients is relatively straightforward. Information describing the data must be included in the input dataset. The data file specified that there were zero categories, zero spatial coordinates, and that the ploidy level was 1 since the data was for inbreds. Ritland's method was used to create a matrix of kinship coefficients. Negative values in the output matrix obtained from SPAGeDi were set to zero before running the subsequent association analysis.

As the results described below show, this method of calculating a kinship (K) matrix works well for this data. In part, this may be a result of the fact that maize is a natural outcrosser and lacks strongly differentiated subpopulations. Applying the K matrix method to self-pollinated species with major population substructure, such as rice, may require a modified approach.

The first set of examples uses the candidate gene approach to association analysis. This method entails using different lines of evidence to develop a list of genes which could be important in controlling the expression of a trait of interest. The genes identified are then resequenced for each of the individual taxa in the study. The resulting sequences are aligned, and sequence polymorphisms identified. The results below use data previously analyzed and reported by Thornsberry et al. (2001). In this study, the dwarf8 gene was chosen because QTL studies and mutagenesis had suggested that it affects maize flowering time and plant height. For simplicity, the results below only look at SNPs. As the published analysis shows, including indels is critical for proper interpretation.

The software TASSEL was used to extract SNPs from the aligned sequence and to perform the analyses. Three related analytical methods were examined. In each case, each SNP was analyzed individually. First, a fixed effect linear model (GLM) was solved using the SNP as a classification variable. Second, the population parameters (Q-matrix) derived from STRUCTURE were added to the model as covariates (GLM + Q). As STRUCTURE was used to assign each line to three populations, each line had three parameters which added to 1. As a result, the three sets of parameters were linearly dependent and only two of the three needed to be included in the model as covariates. The F-test of SNP is the same regardless of which two are actually used. Third, a mixed linear model (MLM) solution (Yu et al. 2006) was used which treated the SNP and the population parameters as fixed effects and taxa as a random effect with the kinship matrix (K-matrix) defining the additive genetic covariance structure.

An important difference between the models is the way in which taxa (the inbred lines) are treated. In the case of the two GLM analyses, taxa do not enter the model explicitly. Each line can be thought of as a random sample from two underlying populations, that differ only by the value of the SNP. When the model contains only the SNP main effect, it is equivalent to a t-test of the difference between the means of the two populations. If additional terms are added to the model, such as replications or environments, the SNP effect must be tested using the taxa nested within SNP mean square, not the model residual. In the case of MLM, taxa enter the model as random effects with a known covariance structure. This structure is specified by the K-matrix. The K-matrix supplies more information about relationships between lines than the Q-matrix and does a better job of removing spurious effects due to population substructure (Yu et al. 2006).

The probability values derived from the F-tests of selected SNPs for association with the days to silk data for each of the analysis methods is shown in Table 9. As expected, the different analysis methods give similar results. The p-values are generally lowest for GLM, especially in the case of site 2625. As several hypotheses are being tested, some form of multiple test correction should be used to evaluate the results. While

Table 9 Probability values for F-tests of selected SNP sites in the dwarf8 gene using different analysis methods.

Site	*GLM*	*GLM + Q*	*MLM*
184	0.3921	0.5684	0.4528
677	0.0012	0.0019	0.0007
680	0.0157	0.057	0.0298
695	0.8967	0.3757	0.5568
699	0.0006	0.0013	0.0004
713	0.7852	0.3555	0.4713
736	0.8501	0.4321	0.5638
741	0.0031	0.0422	0.0185
756	0.0101	0.0841	0.0746
1663	0.0000	0.0004	0.0003
2511	0.6236	0.5742	0.5632
2625	0.0007	0.0515	0.051
2880	0.5412	0.4678	0.511
3000	0.6979	0.254	0.213
3459	0.0014	0.0017	0.007
3490	0.0001	0.0003	0.0002
3570	0.0001	0.0003	0.0002

conservative, a Bonferroni correction is easy to use. If an a-level of 0.01 is desired for rejection of the null hypothesis of no association, the Bonferroni corrected a-level is 0.01/17 = .0006. Using that cut-off would lead us to reject the null hypothesis for sites 1663, 3490, and 3570 for all three methods and 699 for GLM and MLM. However, the choice of a-level is clearly arbitrary and should only serve as a guide to our interpretation of the results. Sites 677 in the MLM results and 2625 in the GLM results are close to the cutoff and could be considered as well. Even in this small example, the results of the three analyses vary, but without additional information, there is no way to decide which is best.

The same days to silk data used to generate the results in Table 9 were also tested for association with 553 random unlinked SNPs from maize genes. As linkage disequilibrium in maize decays very rapidly with distance, randomly chosen SNPs are not expected to be linked to any individual trait and should not be associated with it. The SNPs with the lowest p-values from the GLM analysis are shown in Table 10. For

Table 10 Probability values for F-tests of a subset of 553 random SNPs from maize genes. All 553 SNPs were tested. Those with the smallest p-values for the GLM method are shown here.

SNP	*GLM*	*GLM+Q*	*MLM*
514	0.00000	0.0212	0.0025
10	0.00000	0.0447	0.2464
429	0.00000	0.0031	0.0148
469	0.00000	0.0111	0.0103
319	0.00002	0.0308	0.076
318	0.00002	0.0000	0.0004
368	0.00003	0.0502	0.0954
45	0.00003	0.2761	0.6529
203	0.00003	0.0498	0.0909
46	0.00004	0.0431	0.1814
464	0.00005	0.3304	0.3111
478	0.00005	0.1382	0.175
388	0.00007	0.0527	0.2574
398	0.00014	0.6201	0.2859
307	0.00018	0.1304	0.2586
157	0.00020	0.0324	0.0827
526	0.00022	0.0001	0.0008
173	0.00022	0.2148	0.0969
1	0.00027	0.0061	0.0072

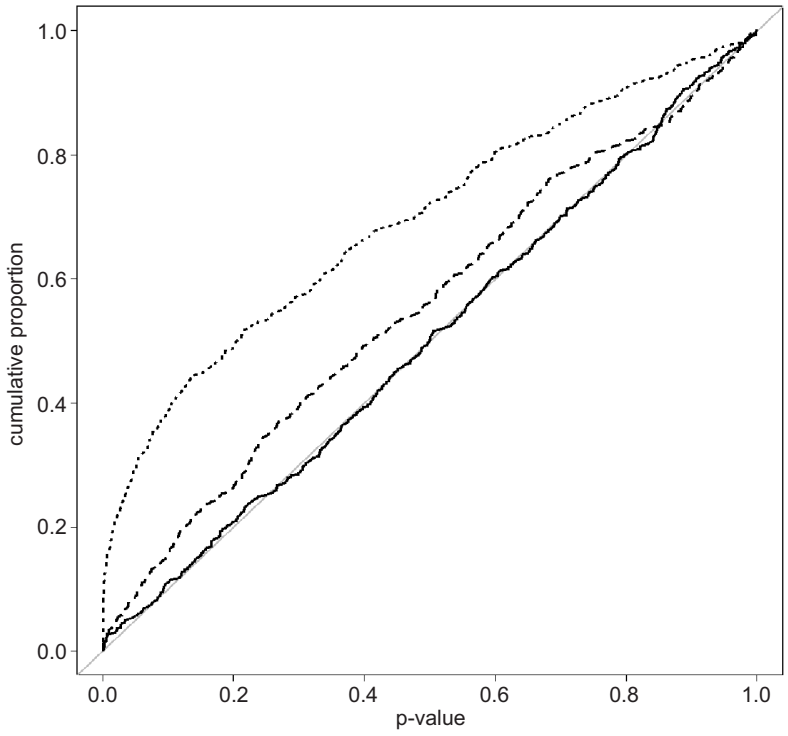

Fig. 14 Cumulative probability curve from F-tests of the association of 553 random SNPs with days to silk for three analysis methods. A straight line with a slope of 1 is expected under the null hypothesis of no association. Solid = MLM, dashed = GLM + covariate, dotted = GLM.

this set the Bonferroni corrected a-level corresponding to an overall desired a-level of 0.01 is 0.01/553 ≅ 2E-5. At that level, using GLM several SNPs appear to be associated with flowering date. Only two of those associations remain using GLM+Q. None of them are close to our chosen significance level using MLM.

A more effective way to summarize the data from this example is shown in Figure 14, a graph of the cumulative distribution functions. To generate the graph, the p-values from the F-tests for SNP were sorted in ascending order individually for each method. An order statistic was assigned to each value with 1 for the lowest p-value and 553 for the highest. The order statistic divided by 553 is plotted on the y-axis and the actual p-value on the x-axis. Under the null hypothesis of no association,

COLOUR FIGURE

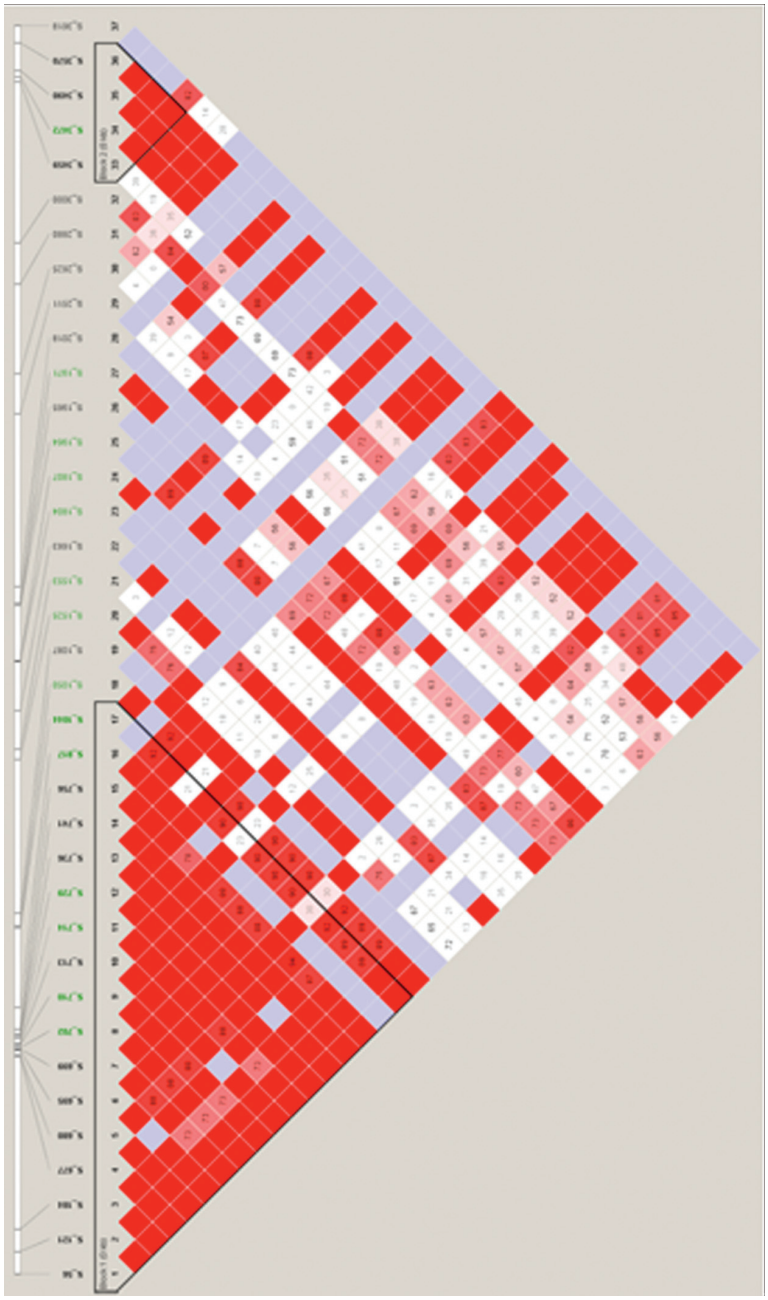

Fig. 15 LD graph for dwarf8 polymorphisms generated by Haploview software. Shades of red indicate a LOD score > 2. White and blue indicate LOD score < 2. Blue and dark red indicate D' = 1. Light red and white indicate D' < 1. Black lines show haplotype blocks.

the points are expected to lie on a straight line with a slope of 1. A deviation from that line is an indication of false positives probably resulting from underlying population substructure. As was obvious from Table 10, the GLM method has a number of false positives. Adding population covariates to the model helps, but the cumulative probability curve still deviates from a straight line. The MLM results lie almost exactly on the line expected under the null hypothesis, indicating that in this example the MLM method effectively eliminated association due to population substructure alone. Other traits show the same trend, though the strength of that trend varies depending on how strongly a trait is associated with population structure (Yu et al. 2006).

Graphs showing the extent of LD between markers can help to interpret the output from an association analysis. As described earlier, a number of software packages can be used to visualize the pattern of linkage disequilibrium between markers. The LD graph shown in Figure 15 was generated by Haploview. This figure graphs the relationships between SNPs and indels from the dwarf8 sequence alignment used in the first example. It reveals two haplotype blocks defined by high internal levels of LD. In addition, those two blocks are seen to be in LD with each other. The graph shows why, using this dataset, several polymorphisms are almost equally likely to be the cause of phenotypic variation. Relative positions of the polymorphisms are shown in the bar above the graph. Gray lines indicate SNPs and green lines indicate indels because the indels were identified in the input file of marker names. The color coding is based on values of D' and its associated LOD scores or tests of significance. Values of D' are displayed in the sqaures. In addition, Haploview provides other color schemes, can display r^2 values as well, has a display which shows haplotype values, but only analyzes data for bi-allelic markers.

Acknowledgements

Yandell was supported in part by National Institutes of Health (NIH) Grants R01 GM069430, NIDDK 5803701 and NIDDK 66369-01. The authors wish to thank Karl Broman, Natalia de Leon and Nengjun Yi for helpful comments on early drafts of this chapter.

References

Abecasis GR, Cardon LR, Cookson WOC (2000) A general test of association for quantitative traits in nuclear families. Am J Hum Genet 66: 279-292

Abecasis GR, Cookson WO (2000) GOLD—Graphical Overview of Linkage Disequilibrium. Bioinformatics 16: 182-183: www.sph.umich.edu/csg/abecasis/GOLD

Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2002) Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30: 97-101: www.sph.umich.edu/csg/abecasis/Merlin

Almasy L, Blangero J (1998) Multipoint quantitative trait linkage analysis in general pedigrees. Am J Hum Genet 62: 1198-1211: www.sfbr.org/solar

Bacanu S, Devlin B, Roeder K (2002) Association studies for quantitative traits in structured populations. Genet Epidemiol 22: 78-93

Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21: 263-265: www.broad.mit.edu/mpg/haploview

Basten CJ, Weir BS, Zeng ZB (1999) QTL Cartographer: A Reference Manual and Tutorial for QTL Mapping. Center for Quantitative Genetics, NCSU: statgen.ncsu.edu/qtlcart

Boldman KG, Kriese LA, VanVleck LD, Van Tassell CP, and Kachman SD (1995) A manual for use of MTDFREML. A set of programs to obtain estimates of variance and covariance. USDA, Agriculture Research Service, Clay Center, NE: aipl.arsusda.gov/curtvt/mtdfreml.html

Borevitz JO, Maloof JN, Lutes J, Dabi T, Redfern JL, Trainer GT, Werner JD, Asami T, Berry CC, Weigel D, Chory J (2002) Quantitative trait loci controlling light and hormone response in two accessions of *Arabidopsis thaliana*. Genetics 160(2): 683-96: hacuna.ucsd.edu/bqtl

Breseghello F, Sorrells ME (2006) Association analysis as a strategy for improvement of quantitative traits in plants. Crop Sci 46: 1323-1330

Broman KW (2001) Review of statistical methods for QTL mapping in experimental crosses. Lab Anim 30(7): 44-52

Broman KW, Sen Ś, Owens SE, Manichaikul A, Southard-Smith EM, Churchill GA (2006) The X chromosome in quantitative trait locus mapping. Genetics 174: 2151-2158

Broman KW, Speed TP (2002) A model selection approach for the identification of quantitative trait loci in experimental crosses (with discussion). J Roy Stat Soc B 64: 641-656, 731-775

Broman KW, Wu H, Sen Ś, Churchill GA (2003) R/qtl: QTL mapping in experimental crosses. Bioinformatics 19: 889-890: www.rqtl.org

Buntjer JB, Sorensen AP, Peleman JD (2005) Haplotype diversity: the link between statistical and biological association. Trends Plant Sci 10: 466-471

Carter K, McCaskie P, Palmer L (2006) JLIN: A java based linkage disequilibrium plotter. BMC Bioinformatics 7: 60: www.genepi.org.au/jlin

Casstevens TM, Buckler ES (2004) GDPC: connecting researchers with multiple integrated data sources. Bioinformatics 20: 2839-2840: www.maizegenetics.net/gdpc

Churchill GA, Doerge RW (1994) Empirical threshold values for quantitative trait mapping. Genetics 138: 963-971

Czika W, Yu X (2004) Gene frequencies and linkage disequilibrium. In: AM Saxton (ed) Genetic Analysis of Complex Traits Using SAS. Cary, NC, USA, SAS Institute Inc: 179-200

Darvasi A (2005) Dissecting complex traits: the geneticists' 'Around the world in 80 days'. Trends Genet 21: 373-376

Devlin B, Roeder K (1999) Genomic control for association studies. Bioinformatics 55: 997-1004

Devlin B, Bacanu S, Roeder K (2004) Genomic Control to the extreme. Nat Genet 36: 1129-1130

Doerge RW, Churchill GA (1996) Permutation tests for multiple loci affecting a quantitative character. Genetics 142: 285-294

Doerge RW, Zeng ZB, Weir BS (1997) Statistical issues in the search for genes affecting quantitative traits in experimental populations. Statist Sci 12: 195-219

Ding K, Zhou K, He F, Shen Y (2003) LDA—a java-based linkage disequilibrium analyzer. Bioinformatics 19: 2147-2148: www.chgb.org.cn/lda/lda.htm

Flint-Garcia SA, Thornsberry JM, Buckler ES (2003) Structure of linkage disequilibrium in plants. Annu Rev Plant Biol 54: 357-374

Gaffney PJ (2001) An efficient reversible jump Markov chain Monte Carlo approach to detect multiple loci and their effects in inbred crosses. PhD Dissertation, Dept of Statistics, UW-Madison, WI, USA

Gaunt TR, Rodriguez S, Zapata C, Day INM (2006) MIDAS: software for analysis and visualisation of interallelic disequilibrium between multiallelic markers. BMC Bioinformatics 7: 227-237: www.genes.org.uk/software/midas

Gelman A, Carlin JB, Stern HS, Rubin DB (2003) Bayesian Data Analysis. 2nd edn. CRC Press, London, UK

Glazier AM, Nadeau JH, Aitman TJ (2002) Finding genes that underlie complex traits. Science 298: 2345-2349

Guo SW, Lange K (2000) Genetic mapping of complex traits: promises, problems, and prospects. Theor Pop Biol 57: 1-11

Gupta P, Rustgi S, Kulwal P (2005) Linkage disequilibrium and association studies in higher plants: Present status and future prospects. Plant Mol Biol 57: 461-485

Hackett CA (2003) Statistical methods for QTL mapping in cereals. Plant Mol Biol 48: 585-599

Haley C, Knott S (1992) A simple regression method for mapping quantitative trait loci of linked factors. J Genet 8: 299-309

Hardy OJ, Vekemans X (2002) Spagedi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. Mol Ecol Notes 2: 618-620: www.ulb.ac.be/sciences/lagev/spagedi.html

Heath SC (1997) Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. Am J Hum Genet 61: 748-760

Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6: 95-108

Hollander WF (1955) Epistasis and hypostasis. J Hered 46: 222-225

Jansen RC (1993) Interval mapping of multiple quantitative trait loci. Genetics 135: 205-211

Jiang C, Zeng ZB (1995) Multiple trait analysis of genetic mapping for quantitative trait loci. Genetics 140: 1111-1127

Jin C, Fine JP, Yandell BS (2007) A unified semiparametric framework for QTL analyses, with application to spike phenotypes. J Am Statist Assoc 102: 56-67

Kao CH (2000) On the differences between maximum likelihood and regression interval mapping in the analysis of quantitative trait loci. Genetics 156: 855-865

Kao CH, Zeng ZB (2002) Modeling epistasis of quantitative trait loci using Cockerham's model. Genetics 160: 1243-1261

Kao CH, Zeng ZB, Teasdale RD (1999) Multiple interval mapping for quantitative trait loci. Genetics 152: 1203-1216

Kilpikari R, Sillanpaa MJ (2003) Bayesian analysis of multilocus association in quantitative and qualitative traits. Genet Epidemiol 25: 122-135

Korstanje R, Paigen B (2002) From QTL to gene: the harvest begins. Nat Genet 31: 235-236

Kruglyak L, Lander ES (1995) A nonparametric approach for mapping quantitative trait loci. Genetics 139: 1421-1428

Lan H, Chen M, Byers JE, Yandell BS, Stapleton DS, Mata CM, Mui ETK, Flowers MT, Schueler KL, Manly KF, Williams RW, Kendziorski C, Attie AD (2006) Combined expression trait correlations and expression quantitative trait locus mapping. PLoS Genetics 2: e6

Lander E, Abrahamson J, Barlow A, Daly M, Lincoln S, Newburg L, Green P (1987) MapMaker: A computer package for constructing genetic linkage maps. Cytogenet Cell Genet 46: 642-642

Lander ES, Botstein D (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121: 185-199

Lander ES, Green P (1987) Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci USA 84: 2363-2367: www.broad.mit.edu/genome_software

Lander ES, Kruglyak L (1995) Genetic dissection of complex traits- guidelines for interpreting and reporting linkage results. Nat Genet 11: 241-247

Li R, Tsaih SW, Shockley K, Stylianou IM, Wergedal J, Paigen B, Churchill GA (2006) Structural model analysis of multiple quantitative traits. PLoS Genetics 2: e114

Liu J, Mercer JM, Stam LF, Gibson GC, Zeng ZB, Laurie CC (1996) Genetic analysis of a morphological shape difference in the male genitalia of *Drosophila simulans* and *D. mauritiana*. Genetics 142: 1129-1145

Liu K, Muse SV (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. Bioinformatics 21: 2128-2129: www.powermarker.net

Ljungberg K, Holmgren S, Carlborg Ö (2004) Simultaneous search for multiple QTL using the global optimization algorithm DIRECT. Bioinformatics 20: 1887-1895: user.it.uu.se/~kl

Mackay TFC (2001) The genetic architecture of quantitative traits. Annu Rev Genet 35: 303-339

Manichaikul A, Dupuis J, Sen Ś, Broman KW (2006) Poor performance of bootstrap confidence intervals for the location of a quantitative trait locus. Genetics 174:481-489

Martinez O, Curnow RN (1992) Estimation the locations and the sizes of the effects of quantitative trait loci using flanking markers. Theor Appl Genet 85: 480-488

Meer JM, Cudmore Jr RH, Manly KF (2004) MapManager/QTX: software for complex trait analysis. www.mapmanager.org/mmQTX.html

Mester DI, Ronin YI, Nevo E, Korol AB (2004) Fast and high precision algorithms for optimization in large-scale genomic problems. Comp Biol Chem 28: 281-290 www.multiqtl.com

Mott R, Talbot CJ, Turri MG, Collins AC, Flint J (2000) A new method for fine-mapping quantitative trait loci in outbred animal stocks. Proc Natl Acad Sci USA 97(23):12649-12654: www.well.ox.ac.uk/~rmott/happy.html

Nadeau JH, Frankel WN (2000) The roads from phenotypic variation to gene discovery: mutagenesis versus QTLs. Nat Genet 25: 381-384

Page GF, George V, Go RC, Page PZ, Allison DB (2003) "Are we there yet?": Deciding when one has demonstrated specific genetic causation in complex diseases and quantitative traits. Am J Hum Genet 73: 711-719

Paterson AH, Damon S, Hewitt JD, Zamir D, Rabinowitch HD, Lincoln SE, Lander ES, Tanksley SD (1991) Mendelian factors underlying quantitative traits in tomato: comparison across species, generations and environments. Genetics 127: 181-197

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155: 945-959: pritch.bsd.uchicago.edu/software.html

Satagopan JM, Yandell BS (1996) Estimating the number of quantitative trait loci via Bayesian model determination. Special Contributed Paper Session on Genetic Analysis of Quantitative Traits and Complex Diseases, Biometric Section, Joint Statistical Meetings, Chicago, IL, USA

Satagopan JM, Yandell BS, Newton MA, Osborn TC (1996) Markov chain Monte Carlo approach to detect polygene loci for complex traits. Genetics 144*:* 805-816

Seaton G, Haley CS, Knott SA, Kearsey M, Visscher PM (2002) QTL Express: mapping quantitative trait loci in simple and complex pedigrees. Bioinformatics 18: 339-340: qtl.cap.ed.ac.uk

Sen Ś, Churchill GA (2001) A statistical framework for quantitative trait mapping. Genetics 159*:* 371-387: www.jax.org/staff/churchill/labsite/software/pseudomarker

Sen Ś, Satagopan JM, and Churchill GA (2005) Quantitative trait locus study design from an information perspective. Genetics 170*:* 447-464: www.jax.org/staff/churchill/labsite/software/pseudomarker

Sillanpää MJ, Arjas E (1998) Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. Genetics 148: 1373-1388: www.rni.helsinki.fi/~mjs

Solberg LC, Baum AE, Ahmadiyeh N, Shimomura K, Li R, Turek FW, Churchill GA, Takahashi JS, Redei EE (2004) Sex- and lineage-specific inheritance of depression-like behavior in the rat. Mamm Genom 15: 648-662

Solberg LC, Baum AE, Ahmadiyeh N, Shimomura K, Li R, Turek FW, Takahashi JS, Churchill GA, Redei EE (2006) Genetic analysis of the stress-responsive adrenocortical axis. Physiol Genom 00: 000-000: dx.doi.org/10.1152/physiolgenomics.00052.2006

Stephens DA, Fisch RD (1998) Bayesian analysis of quantitative trait locus data using reversible jump Markov chain Monte Carlo. Biometrics 54: 1334-1347

Stuber CW, Lincoln SE, Wolff DW, Helentjaris T, Lander ES (1992) Identification of genetic factors contributing to heterosis in a hybrid from two elite maize inbred lines using molecular markers. Genetics 132: 823-839

Stylianou IM, Tsai SW, diPetrillo K, Ishimori N, Li R, Paigen B, Churchill G (2006) Complex genetic architecture revealed by analysis of HDL in chromosome substitution strains in F2 crosses. Genetics 174: 999-1007

Thornsberry JM, Goodman MM, Doebley J, Kresovich S, Nielsen D, Buckler ES 4th (2001) Dwarf8 polymorphisms associate with variation in flowering time. Nat Genet 28: 286-289

Utz HF, Melchinger AE (1996) PLABQTL: a program for composite interval mapping of QTL. J Quant Trait Loci 2 : www.uni-hohenheim.de/ipspwww/soft.html

van Ooijen JW, Maliepaard C (1996) MapQTL version 3.0: software for the calculation of QTL positions on genetic maps. CPRO-DLO, Wageningen, The Netherlands: www.kyazma.nl

Vieira C, Pasyukova EG, Zeng ZB, Hackett JB, Lyman RF, Mackay TFC (2000) Genotype-environment interaction for quantitative trait loci affecting life span in *Drosophila melanogaster*. Genetics 154: 213-227

Wang H, Zhang YM, Li X, Masinde GL, Mohan S, Baylink DJ, Xu S (2005) Bayesian shrinkage estimation of quantitative trait loci parameters. Genetics 170: 465-480

Wang J, Williams RW, Manly KF (2003) WebQTL: Web-based complex trait analysis. Neuroinformatics 1: 299-308: www.genenetwork.org

Wright FA, Kong A (1997) Linkage mapping in experimental crosses: the robustness of single-gene models. Genetics 146: 417-425

Wu B, Nianjun L, Hongyu Z (2006) PSMIX: an R package for population structure inference via maximum likelihood method. BMC Bioinformatics 7: 317: bioinformatics.med.yale.edu/PSMIX

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

Yandell BS, Mehta T, Banerjee S, Shriner D, Venkataraman R, Moon JY, Neely WW, Wu H, von Smith R, Yi N (2007) R/qtlbim: QTL with Bayesian interval mapping in experimental crosses. Bioinformatics 23: 641-643: www.rqtlbim.org

Yang J, Hu CC, Ye XZ, Zhu J (2005) QTLNetwork 2.0. Institute of Bioinformatics, Zhejiang University, Hangzhou, China: ibi.zju.edu.cn/software/qtlnetwork

Yi N (2004) A unified Markov chain Monte Carlo framework for mapping multiple quantitative trait loci. Genetics 167: 967-975

Yi N, Yandell BS, Churchill GA, Allison DB, Eisen EJ, Pomp D (2005) Bayesian model selection for genome-wide epistatic quantitative trait loci analysis. Genetics 170: 1333-1344

Yu J, Buckler ES (2006) Genetic association mapping and genome organization of maize. Curr Opin Biotechnol 17: 155-160

Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38: 203-208: www.maizegenetics.net/bioinformatics/tassel

Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, Ehm MG (2002) Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. Hum Hered 53: 79-91

Zeng ZB (1994) Precision mapping of quantitative trait loci. Genetics 136: 1457-1468

Zeng ZB, Liu J, Stam LF, Kao CH, Mercer JM, Laurie CC (2000) Genetic architecture of a morphological shape difference between two drosophila species. Genetics 154: 299-310

Zhang Z, Todhunter RJ, Buckler ES, van Vleck LD (2006) Technical note: use of marker based relationships with multiple-trait derivative-free restricted maximal likelihood. J Anim Sci 85: 881-885

Zollner S, Pritchard JK (2005) Coalescent-based association mapping and fine mapping of complex trait loci. Genetics 169: 1071-1092

Zou F, Fine JP, Hu J, Lin DY (2004) An efficient resampling method for assessing genome-wide statistical significance in mapping quantitative trait loci. Genetics 168: 2307-2316