

Introduction

Many phenotypes (traits) of biomedical, agricultural, or evolutionary importance are quantitative in nature. Examples include blood pressure (to study hypertension), milk output (in dairy breeding), and number of seeds produced per plant (to study evolutionary fitness). Many phenotypes such as coat color of mice, or cancer tumor aggressiveness, may not be strictly quantitative, but may be studied by a derived quantitative measure. We may classify mice by whether or not they have an agouti coat color, a 0/1 measure, or grade tumors by aggressiveness on a scale of 1 to 4 by examining tumor biopsies.

Variation in such quantitative traits is often due to the effects of multiple genetic loci as well as environmental factors. Knowledge of the number, locations, effects, and identities of such genetic loci (called quantitative trait loci, QTL) can lead to new biological insights. The information from QTL can be used to develop new therapeutic drugs, assist the selection for improved agricultural crops or breeds, and improve our understanding of natural selection.

Studies to identify, or “map,” QTL may be undertaken in humans or in nonhuman species including model organisms such as mice or *Drosophila* (fruit flies). In these studies, we assemble or create a genetically diverse population. Then we associate genetic variation with phenotypic variation in the study population. Regions of the genome that show convincing evidence of association are flagged as QTL.

In this book, we consider the problem of mapping QTL in an experimental cross formed from two inbred lines. Such crosses are the simplest populations in which we can perform QTL mapping. They are the easiest to understand biologically, as well as mathematically. For this reason, they are the staples of experimental geneticists, and the most common QTL study population. QTL mapping in more complex populations, including in humans, may be viewed as generalizations. Thus, for both biologists and quantitative methodologists, understanding QTL mapping in experimental crosses is an excellent launching point for more ambitious investigations.

We focus on QTL mapping with the R/*qtl* software, an add-on package for the R statistical software, for the analysis of QTL experiments. In the following

three sections, we elaborate on the idea of a QTL experiment, describe the basic crosses and data, and introduce the central statistical problems in QTL mapping. Subsequently, we describe R and R/qtl, as well as some of the alternative programs for QTL mapping. We conclude the chapter with a brief description of the general work flow of QTL data analysis.

1.1 Why perform a QTL experiment?

As mentioned above, the fundamental idea underlying QTL mapping is to associate genotype and phenotype in a population exhibiting genetic variation. Conceptually, the most straightforward study would be in a natural population of the organism of interest. For example, to understand hypertension, we may study genetic associations with hypertension in a large cohort such as the Nurses Health Study (London *et al.*, 1989). While extremely useful, this approach presents a few problems. First, human studies are expensive; second, the phenotypic characterization may be noisy because we cannot control the subjects' environment and life history; and finally, associations do not necessarily imply causation because of possible confounding due to population structure.

QTL mapping in experimental crosses provides an excellent alternative. By suitably choosing a model organism we can home in a particular aspect of the phenotype of interest. For example, we may study hypertension in mice, or alcohol addition in *Drosophila*, by appealing to the evolutionary connection between humans, mice, and *Drosophila*. We have greater control over the phenotyping accuracy, environment, and life history than in natural populations. For example, we can feed all mice the same diet, and keep the mouse rooms climate-controlled, and phenotype the mice at the same day in identical experimental conditions. We can perform phenotyping that may be invasive, impractical, or unethical in humans (such as examining the liver in mice on a high-fat diet). We also have greater control over the genetic composition of the population in experimental crosses. This allows us to magnify the genetic effect of a putative QTL by judiciously choosing the strains to cross. By crossing two (or more) strains, we also effectively randomize genetic variation in the progeny population. This allows us to conclude that genetic associations detected in a cross are causal. Because experimental crosses segregate genetic variation that is simpler relative to a natural population, we can perform more complex statistical modeling to identify epistasis (QTL $\times$ QTL interactions), and QTL $\times$ environment interactions.

Experimental crosses do have some disadvantages. If the cross is in a model organism, the degree to which the conclusions apply to the target organism depends on how good the model is. For example, if we use *Arabidopsis* to study drought tolerance, the conclusions may not be directly applicable to grapes, although there is a good chance that the pathways implicated in *Arabidopsis* are also relevant to grapes. Identification of a QTL does not necessarily

help us identify a gene, since the region spanned by a QTL may contain tens, and sometimes thousands, of genes. A QTL may not be successfully replicated in a congenic strain (identical to one strain at all but the QTL region derived from a different strain) because of epistatic interactions with the genetic background. Developing statistical and experimental solutions to these shortcomings is an active research area.

QTL mapping in experimental crosses is an excellent first step towards more expensive investigations. The simple genetic structure of experimental crosses provides a relatively tractable framework to study the conceptual and statistical principles of genetic mapping.

1.2 Crosses and data

We focus on experimental crosses between inbred lines. An inbred line is formed by repeated sibling mating (or, in many plants, selfing) to obtain individuals who are completely homozygous: both chromosomes are identical at all positions. Inbred lines are essentially “immortal” since all individuals in an inbred line are genetically identical to one another, and to their progeny (apart from sex).

If two inbred lines show a consistent phenotypic difference, despite being raised in a common environment, one may be confident that the strain difference has a genetic basis. The identity of QTL underlying such phenotypic differences may be revealed analyzing crosses between the strains. By crossing the strains, we obtain progeny whose genomes are shuffled versions of the parental genomes.

The simplest cross to describe is the backcross (Fig. 1.1). Two inbred strains, denoted A and B, are crossed to obtain the first filial (F_1) generation. F_1 individuals receive a copy of each chromosome from each of the two parental strains; wherever the parental strains differ, the F_1 generation is heterozygous. The F_1 individuals are crossed to one of the two parental strains. For example, if an F_1 individual is crossed with its A strain parent, the backcross progeny receive one chromosome from the A strain and one from the F_1 . Thus, at each autosomal locus, they have genotype AA or AB. The chromosome received from the F_1 parent may be one of the original parental chromosomes intact but is generally a mosaic of the two parental chromosomes, as a result of recombination at meiosis (the process of cell division that gives rise to sex cells). The points of exchange are called crossovers.

While the backcross is the simplest possible experimental cross, the intercross (Fig. 1.2) is also commonly used. In an intercross, one crosses F_1 siblings (or, in many plants, one may self an F_1 individual) to obtain the F_2 generation, who receive a recombinant chromosome from each parent and so, at any autosomal locus, have genotype either AA, AB, or BB. The intercross allows the detection of QTL for which one allele is dominant, while in a backcross to the A strain, one may detect a QTL only if the A allele is not dominant.

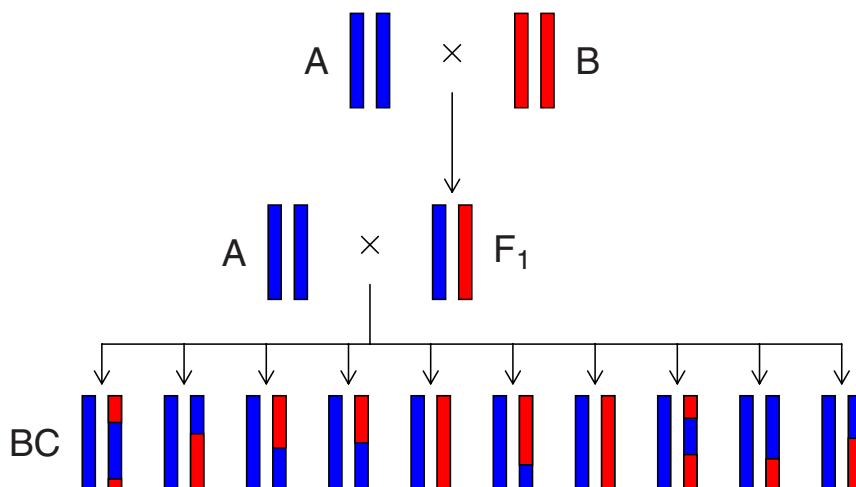


Figure 1.1. Schematic representation of the autosomes in a backcross experiment. The two inbred strains, A and B, are represented by blue and pink chromosomes, respectively. The F_1 generation, obtained by crossing the two strains, receives a single chromosome from each parent, and all individuals are genetically identical. If we cross an individual from the F_1 generation “back” to one of the parental strains (the A strain, in this example), we obtain a population exhibiting genetic variation. The backcross individuals receive an intact A chromosome from their A parent. The chromosome received from their F_1 parent may be an intact A or B chromosome, but is generally a mosaic of the A and B chromosomes as a result of recombination at meiosis. Any given locus has a 50% chance of being heterozygous and a 50% chance of being homozygous. This figure represents the autosomes only. When considering the X chromosome, four backcross populations (“directions”) are possible (see Fig. 4.22 on page 110).

Moreover, the intercross allows one to estimate the degree of dominance at a QTL.

Another strategy would be to use recombinant inbred lines (Fig. 1.3). These are constructed by beginning with an intercross, and then mating pairs of F_2 siblings, followed by a parallel series of repeated sibling mating to construct a new panel of inbred lines whose genomes are a mosaic of the two initial lines. In organisms that allow selfing, one may following the initial cross by repeated selfing, multiple times in parallel; the progress to inbreeding in recombinant inbred lines by selfing is more rapid.

Recombinant inbred lines (RILs) have a number of advantages. Since they are immortal, we need genotype each line only once; we can phenotype multiple individuals from each line to reduce individual, environmental and measurement variability; we can obtain multiple invasive phenotypes on the same set of genomes; and, as the breakpoints in RILs are more dense than those

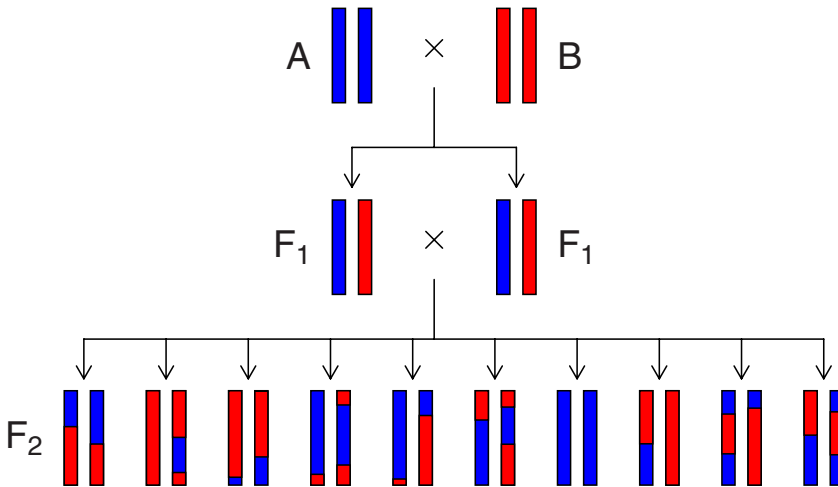


Figure 1.2. Schematic representation of the autosomes in an intercross experiment. The two inbred strains, A and B, are represented by blue and pink chromosomes, respectively. As with the backcross strategy, the F_1 generation, obtained by crossing the two strains, has a chromosome from each parent, and all F_1 individuals are genetically identical. By crossing two individuals in the F_1 generation (or by selfing when possible), we create genetic variation in the resulting F_2 population. The three possible genotypes AA, AB, and BB appear in a 1:2:1 ratio. This figure represents the autosomes only; the behavior of the X chromosome is displayed in Fig. 4.23 on page 111.

that occur in any one generation, we can achieve better mapping resolution. However, constructing and maintaining RILs is expensive. For this reason, they are more commonly used by plant biologists whose costs are lower relative to animal biologists. Most of the examples in this book will focus on the backcross and intercross. However, we will further consider the RIL design in the chapter on experimental design (Chap. 6).

Before embarking on a QTL experiment, the experimenter will usually phenotype several individuals from the two parental strains, and often some F_1 hybrid individuals. Hypothetical phenotype data for two inbred strains, the F_1 hybrid, and a backcross population are shown in Fig 1.4. Individuals within each of the parental strains are genetically identical, and so any variation in the phenotype within the strains is nongenetic (due to a combination of measurement error, environmental variation, and individual developmental noise). One generally chooses parental strains that show systematic phenotypic differences. Note that the within-strain variation is not necessarily the same in the two parental strains. The F_1 individuals are also genetically identical, and so any phenotypic variation in the F_1 is again nongenetic. The phenotype distribution of the F_1 individuals is often intermediate between the parental

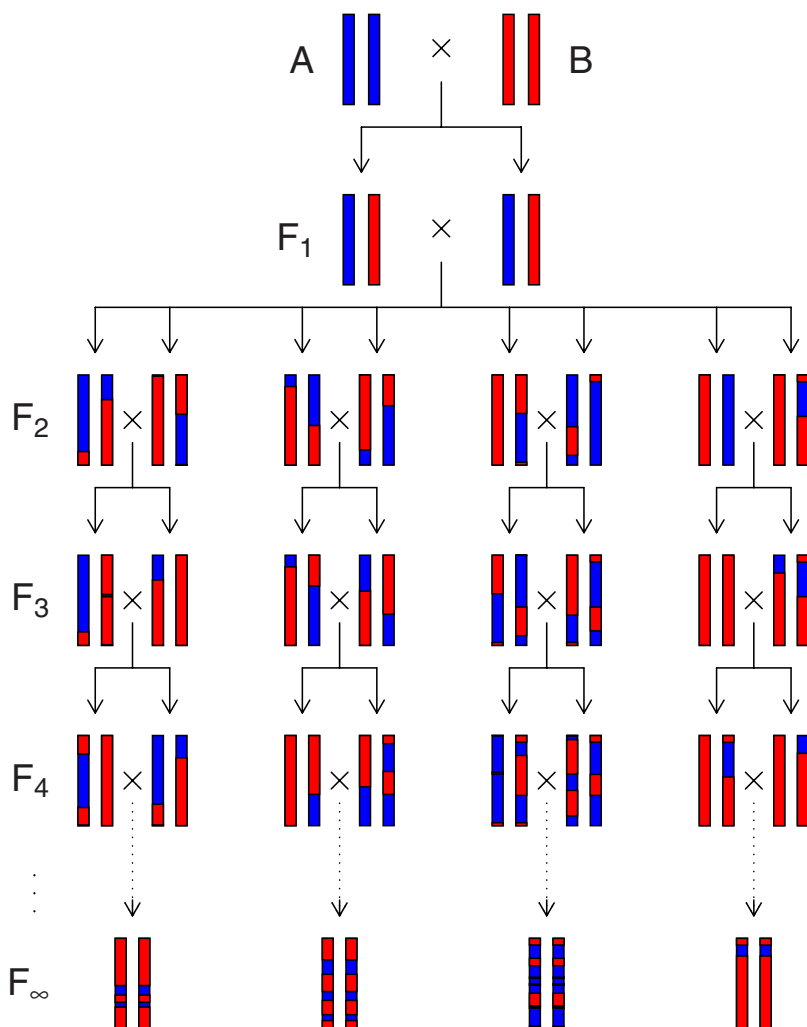


Figure 1.3. Schematic representation of the autosomes during the breeding of recombinant inbred lines by sibling mating. The first two generations are identical to an F₂ intercross. In subsequent generations, siblings are mated producing progeny that are less and less heterozygous. If continued indefinitely, this process will produce individuals that are completely homozygous at every locus, but with chromosomes that are a mosaic of the parental chromosomes. The frequency of the breakpoints between the AA and BB genotypes is determined by the breeding scheme (sib-mating or selfing, or some other scheme). In practice, 10–20 generations of inbreeding are actually performed.

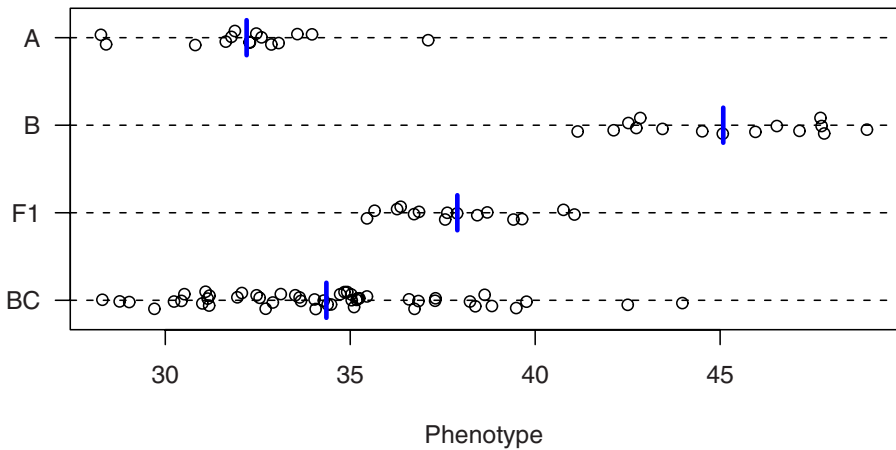


Figure 1.4. (Hypothetical) phenotype data for the two parental strains, the F_1 hybrid, and a backcross population. Vertical line segments are plotted at the within-group averages. In this simulated example, the difference in the phenotype means between strains is quite large relative to the within-strain variation. The mean phenotype in the F_1 generation is intermediate between the two strains, while the backcross progeny exhibit a wider spectrum of phenotypic variation and resemble the A strain more than the B strain.

lines' distributions, but this is not necessarily the case. Individuals in the backcross population are not genetically identical, and so they should exhibit greater phenotypic variation.

The key idea in QTL mapping is to obtain phenotype data on a number of backcross or intercross progeny and then identify regions in the genome where genotype is associated with the phenotype. However, the genotype is not observed at every possible position along the chromosomes, but only at a set of discrete landmarks called genetic markers. These are biochemical assays that reveal the identity of a defined genomic region. Microsatellite and SNP markers are the most common types of markers used for QTL mapping.

QTL mapping data has three interrelated data structures: the phenotypes, the genotypes, and the marker map. The phenotype data consists of observable characteristics of each individual in the population. These would include traits of interest such as blood pressure, or body weight, but also covariates such as sex, cross direction, and environmental conditions such as diet. Typically one has data on 100–1000 individuals. The genotype data consists of a set of genetic markers spanning the genome. In a typical mouse experiment, one may start with about 100 markers approximately evenly spaced along the genome. The genetic map specifies the markers locations on the chromosomes in terms of genetic distance. If a genetic map is not available, one would infer marker order and position with the available data.

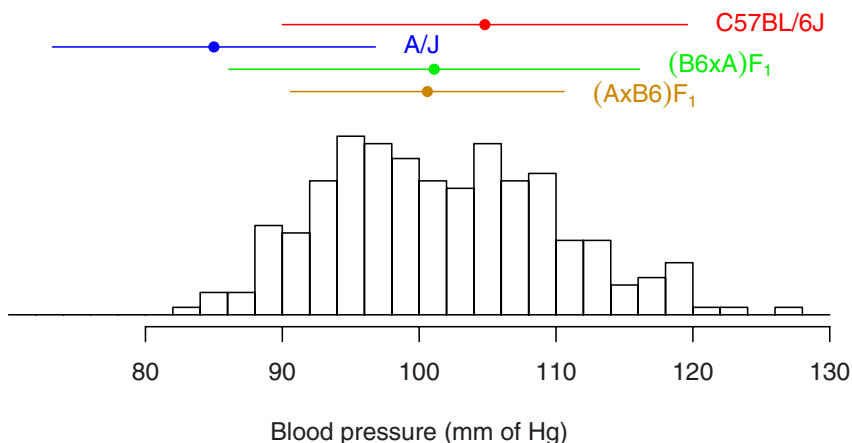


Figure 1.5. Histogram of systolic blood pressure for 250 backcross mice from Sugiyama *et al.* (2001). Also shown are the phenotypic ranges of parental and F_1 hybrid strains (mean $\pm$ 2 SD). The A/J (or A) strain is normotensive, while the C57BL/6J (or B6) strain is hypertensive. The F_1 hybrids and the backcross resemble the hypertensive B6 strain. Notice that the range of the phenotype is not unlike humans.

While a physical map specifies the physical position of markers on the chromosomes, in a genetic map distance is measured by the rate of crossover events at meiosis. Two markers are d centiMorgans (cM) apart if there is an average of d crossovers in the intervening interval in every 100 products of meiosis.

1.2.1 Mouse hypertension data as an example

As an example, we consider the data of Sugiyama *et al.* (2001) on salt-induced hypertension in the mouse. They measured systolic blood pressure in 250 male mice from a backcross between the hypertensive C57BL/6J (B6) and normotensive A/J (A) strains. $(B6 \times A)F_1$ mice were mated to B6 mice to produce a total of 250 male mice. The data are included with R/qtl, and will be referred to as the **hyper** data. (Further details will be provided in Sec. 2.3.) A histogram of the blood pressures of these backcross mice is shown in Fig. 1.5.

A total of 173 markers were genotyped; the genetic map of the markers is shown in Fig. 1.6. For most regions of the genome, markers were placed at a spacing of 10–20 cM. In regions for which an initial analysis indicated some evidence for a QTL (for example, chromosomes 1 and 4), additional markers were added.

The actual genotype data are shown in Fig. 1.7; individuals have been sorted by their phenotype: mice with lower blood pressure are at the bottom, and mice with higher blood pressure are at the top. By careful study of Fig. 1.7,

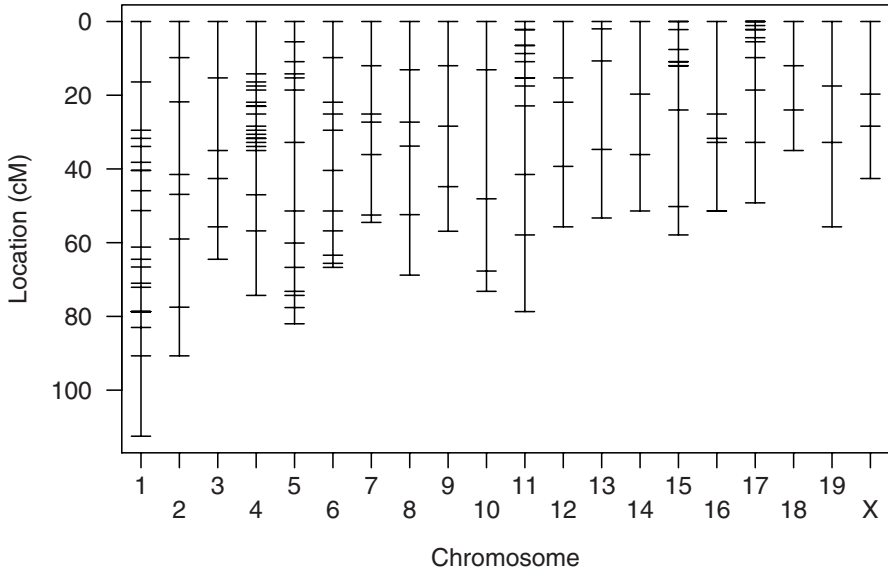


Figure 1.6. The genetic map of markers typed in the data from Sugiyama *et al.* (2001). Almost all marker intervals are less than 20cM. Some regions, most notably on chromosomes 1 and 4, have a higher density of markers.

one can guess the identity of a few putative QTL. For example, on each of chromosomes 1 and 4, there is mostly blue at the bottom of the figure and mostly red at the top. Homozygotes (red) tend to have higher blood pressure than heterozygotes (blue). On chromosome 6, the opposite pattern is seen: the bottom is mostly red and the top is mostly blue. This may indicate a QTL with an effect of the opposite sign. Visual examination of the raw data has an important role in detecting data quality issues, and can also provide informal evidence for QTL. However, for objective evidence for QTL formal statistical methods are required.

1.3 Central statistical problems

Investigators perform QTL experiments with a variety of goals in mind, and so the details of the appropriate statistical methods to meet those goals will also vary. For example, an evolutionary biologist may be particularly interested in the number and effects of QTL, while for a biomedical researcher the goal is to identify the gene (or genes) underlying at least one QTL; the number and effects of QTL may be of minor interest.

The principal goals of QTL mapping are nevertheless well defined. First, we seek to detect QTL (and, potentially, interactions among QTL). Second, we seek confidence regions for the locations of the QTL. Finally, we seek to

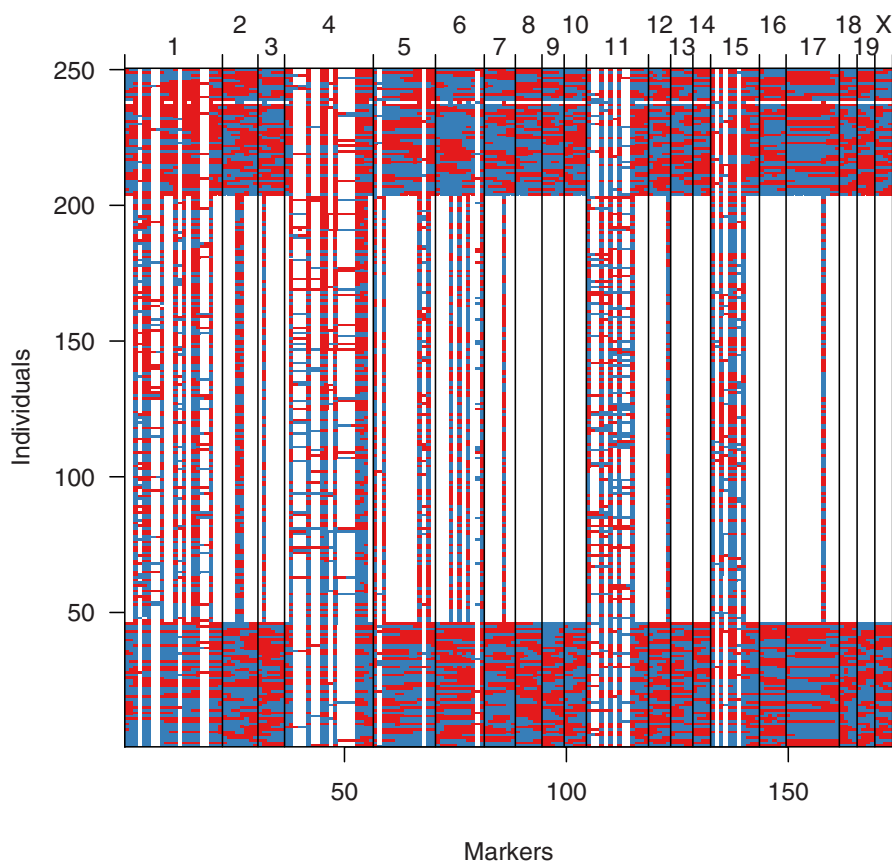


Figure 1.7. Genotype data for the backcross from Sugiyama *et al.* (2001). Red and blue pixels correspond to homozygous and heterozygous genotypes, respectively. White pixels indicate missing genotype data. Black vertical lines indicate the boundaries between chromosomes. Individuals are sorted by their phenotype (low blood pressure at the bottom, high blood pressure at the top). A three-step selective genotyping strategy was used for this cross. Individuals at the extremes of the phenotype distribution were genotyped for a set of framework markers spanning the genome. On selected chromosomes all individuals were typed for framework markers; dense genotyping was performed on recombinant marker intervals to narrow down the locations of recombination events.

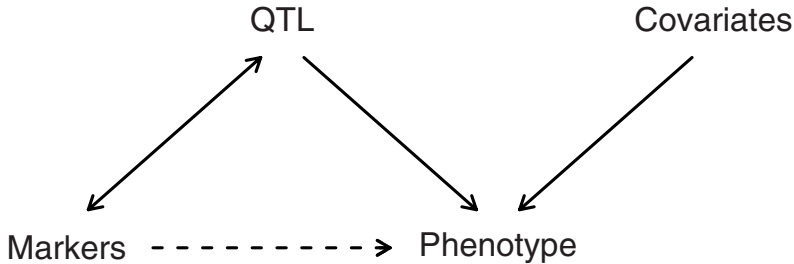


Figure 1.8. The statistical structure of the QTL mapping problem. The QTL and covariates are responsible for phenotypic variation (indicated by the directed solid arrows). The markers and the QTL are correlated with each other due to linkage (indicated by the bidirectional solid arrow). The markers do not directly cause the phenotype; some markers may be associated with the phenotype via linkage to the QTL (indicated by the directed dashed arrow).

estimate the effects of the QTL (that is, the effect, on the phenotype, of substituting one allele for another). These are generally viewed to be of decreasing importance. (For example, there is little need for a confidence interval for a QTL that has not been clearly detected). In this book, we will focus primarily on detecting QTL and their interactions.

The QTL mapping problem is best split into two distinct parts: the *missing data problem* and the *model selection problem*.

As illustrated schematically in Fig. 1.8, the phenotype is influenced by the genotype at QTL plus possible covariates such as sex, treatment, or environmental effects. However, we generally do not observe the genotype at the QTL, but only at marker loci. The genotypes at the markers and the QTL are associated due to linkage, which results in an association between the phenotype and the marker genotypes.

If one knew the genotype of each individual at each position in the genome, QTL mapping would reduce to the identification of the set of sites in the genome that matter in producing the phenotype and of how these sites combine together (and with other covariates) to produce the phenotype. This is the *model selection problem*. However, since we observe individuals' genotypes only at a discrete set of genetic markers and wish to consider positions between markers as the possible locations of QTL, we must use the marker genotype data to infer the genotypes at intervening locations. This is the *missing data problem*.

There are several solutions to the missing data problem, which are all satisfactory when the markers are reasonably dense. Although solutions to the model selection problem have been proposed, it remains challenging. The general problem of variable selection in regression is an active area of research for which no generally acceptable solution is known.

We complete this section with a more detailed discussion of probability models for the processes that give rise to QTL data. For the missing data problem, in which one uses marker genotype data to infer the genotype at intervening positions, one requires a model for the recombination process. More important, however, are models that connect the genotype and phenotype.

1.3.1 Models for recombination

Solutions to the missing data problem mentioned in the previous subsection rely on models for recombination. These models help us probabilistically connect unobserved genotypes to observed genotypes. Genotypes are missing for all individuals at locations between typed markers; they may be missing for untyped individuals at typed markers. All approaches for dealing with missing data rely on the calculation of the genotype probabilities at a putative QTL on the basis of the available multipoint marker data. To do so, one must have a model for the recombination process.

The most convenient model is that of no crossover interference: that the locations of crossovers on a meiotic product are according to a Poisson process. Informally, this means that crossover locations are random. Under the no interference model, recombination events in disjoint intervals are independent. As a result, the genotypes at markers along a chromosome form a Markov chain. In other words, conditional on the genotype at a particular locus, the genotypes at positions to the left are independent of the genotypes at positions to the right. We should emphasize that we also assume that there is no segregation distortion (i.e., that the frequencies of genotypes at an autosomal locus are in the ratios 1:1 in a backcross and 1:2:1 in an intercross).

The convenience of the no crossover interference assumption is best illustrated by an example. In Fig. 1.9, we display hypothetical genotype data for three different backcross individuals at a set of six markers along a chromosome; each row corresponds to a different individual. (The dashes are meant to indicate missing data.) We seek the probability that an individual is AA or AB at the locus (perhaps a putative QTL) indicated by the triangle, given its available marker data.

If no crossover interference is assumed to hold, one need only consider the genotypes at the nearest flanking typed markers. For the first individual, we need only consider the genotypes at markers M_3 and M_4 ; we can ignore the genotypes at the other markers. If r_{iQ} is the recombination fraction between marker M_i and the putative QTL, and if r_{ij} is the recombination fraction between markers M_i and M_j , then the probability that the individual has genotype AA at the putative QTL is $(1-r_{3Q})(1-r_{4Q})/(1-r_{34})$; the probability that it is AB is $r_{3Q}r_{4Q}/(1-r_{34})$.

Note that the fact that the individual showed a recombination event between markers M_4 and M_5 is irrelevant here, as under the no interference model, recombination events in disjoint intervals are independent. However, most organisms exhibit positive crossover interference (crossovers tend not to

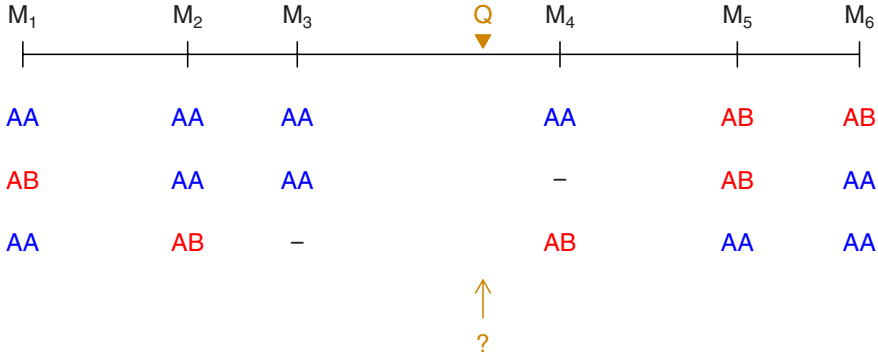


Figure 1.9. Illustration of the problem of inferring missing genotype data. Each row is the marker genotype data for a different backcross individual. Dashes indicate missing data. We seek the probability that each individual has genotype AA or AB at the putative QTL indicated by the triangle.

occur too close together). In particular, the mouse exhibits extremely strong crossover interference, with crossovers seldom being separated by less than 20 cM. In the presence of positive crossover interference, the recombination event between markers M₄ and M₅ would result in a decreased chance for a double recombinant in the interval between M₃ and M₄, and so there would be a greater probability that the individual is AA at the putative QTL. However, calculations are much simpler under the no interference model, and the difference has been seen to have little influence on QTL mapping results.

For the second individual, we consider the genotypes at markers M₃ and M₅, as the genotype at marker M₄ is missing. The probability that the individual is AA at the putative QTL is $(1 - r_{3Q})r_{5Q}/r_{35}$. The probability that it is AB is $r_{3Q}(1 - r_{5Q})/r_{35}$.

Finally, for the third individual, we consider the genotypes at markers M₂ and M₄. The probability that it is AA at the putative QTL is $r_{2Q}r_{4Q}/(1 - r_{24})$. The probability that it is AB at the putative QTL is $(1 - r_{2Q})(1 - r_{4Q})/(1 - r_{24})$.

In summary, an essential task in QTL mapping concerns the reconstruction of missing genotype data conditional on the observed marker genotypes. This requires a model for the recombination process. While meiosis generally exhibits positive crossover interference (with crossovers not occurring close together), we generally assume no crossover interference, as calculations are then greatly simplified. We have illustrated the value of no crossover interference with some simple examples. In general, one may use algorithms for hidden Markov models (HMMs) for these sorts of calculations, as one can then allow for the presence of genotyping errors and more simply deal with partially informative genotypes (such as the case of dominant markers in an intercross). HMM algorithms form the core of R/qtl, and are described in detail in Appendix D.

One last point: it may be worthwhile to discuss the role of map functions. A map function relates the genetic length of an interval (which is generally not estimable) to its recombination fraction (which generally is estimable); that is, it relates the expected number of crossovers in an interval to the probability of an odd number of crossovers (as a recombination event implies an odd number of crossovers). A model for crossover interference may imply a map function (though not necessarily, as if there is variation in the level of interference along a chromosome, an interval of a given genetic length would not correspond to a fixed recombination fraction), though the converse is not true. Common map functions include the Haldane map function (corresponding to no interference), the Kosambi map function (corresponding approximately to the level of interference in humans) and the Carter–Falconer map function (corresponding approximately to the level of interference in mice).

In calculating the genotype probabilities at a putative QTL given the available marker data, and using the no interference assumption, a map function is used to convert genetic lengths to recombination fractions. But if one's own data is used to estimate the genetic map, and if that is done (as is typical) under the no interference assumption, it is the recombination fractions that are actually estimated; the estimated genetic distances are derived via a map function. In this case, it will not matter what map function is used, provided that the same map function used to derive genetic distances is also used to convert back to recombination fractions. The only role of the map function concerns the scale on which results are plotted, as the results are generally plotted as a function of genetic distance. One should not treat estimated genetic distances with much reverence; most important are the markers themselves, which tie one's results back to the DNA sequence.

1.3.2 Models connecting genotype and phenotype

Most important are models connecting genotype and phenotype. Consider a single backcross individual, and let y denote its phenotype and $\mathbf{g}$ its whole-genome genotype (that is, its genotype at all polymorphisms between the parental strains).

Imagine that there are just p sites that matter in producing the phenotype, where p is some small proportion of the total number of polymorphisms, and let $g_1, \dots, g_p$ denote the individual's genotype at these p QTL. We then have $E(y|\mathbf{g}) = \mu_{g_1 \dots g_p}$ and $\text{var}(y|\mathbf{g}) = \sigma_{g_1 \dots g_p}^2$. That is, the expected value (i.e., mean or average) and variance of an individual's phenotype, given its whole-genome genotype, depends only on its genotype at the p QTL; all other polymorphisms are inconsequential.

Thus we may split individuals into 2^p groups (3^p groups in an intercross) that are genetically identical at all polymorphisms contributing to the phenotype. The residual variation in each group, $\sigma_{g_1 \dots g_p}^2$, will be entirely nongenetic (measurement error, environmental variation, and individual developmental noise).

We often make a number of simplifying assumptions. First, we may assume constant variance (homoscedasticity): that $\sigma_{g_1 \dots g_p}^2 \equiv \sigma^2$. In other words, the individual, residual variation (which includes measurement error and environmental variation) is constant within each of the 2^p groups. To the contrary, we often see that such variation increases with the average phenotype, but the constant variance assumption is convenient.

Second, we may assume that the residual variation follows a normal distribution: $y|\mathbf{g} \sim N(\mu_{g_1 \dots g_p}, \sigma^2)$. That is, the phenotype distribution within each of the 2^p groups follows a normal curve, though with different averages. Note that this is not the same as to say that the marginal phenotype distribution follows a normal distribution; rather, the marginal distribution follows a *mixture* of normal distributions, the components of the mixture being the 2^p groups with distinct genotypes at the p QTL.

Finally, we often assume that the QTL act *additively*, so that

$$E(y|\mathbf{g}) = \mu_{g_1 \dots g_p} = \mu + \sum_j \Delta_j z_j$$

where $z_j = 0$ or 1 , according to whether g_j is AA or AB. That is, the effect of QTL j is Δ_j , no matter the genotype at the other loci.

Any deviation from additivity is called epistasis. The term *epistasis* is often reserved for a particular type of interaction, in which the effect of a mutation at a locus may be masked by the presence of a mutation at a second locus. However, statistical geneticists have come to use the term more generally.

As an illustration of epistasis, consider the hypothetical data plotted in Fig. 1.10. Focus first on Fig. 1.10A; we split backcross individuals into four groups according to their joint genotypes at two QTL. Dots are plotted at the average phenotype for each of the two-locus genotypes; line segments are drawn between the averages for a fixed genotype at QTL 2.

The average phenotype for individuals with genotype AA at both QTL is 10, while the average phenotype for individuals with genotype AB at QTL 1 and AA at QTL 2 is 40, and so the effect of QTL 1 is 30 in individuals with genotype AA at QTL 2. The average phenotype for the individuals with genotype AA at QTL 1 and AB at QTL 2 is 60, while the average phenotype for individuals with genotype AB at both QTL is 90, and so the effect of QTL 1 is 30 in individuals with genotype AB at QTL 2. Thus the effect of QTL 1 is the same, no matter the genotype at QTL 2; similarly, the effect of QTL 2 is the same, no matter the genotype at QTL 1. Thus the QTL are said to be additive. (Sometimes, in this case, the QTL are said to be *independent*, but this can be confused with whether or not the QTL are linked on a chromosome, and so we prefer the term *additive*.)

In Fig. 1.10B, on the other hand, the effect of QTL 1 is 30 when the genotype at QTL 2 is AA, but is 65 when the genotype at QTL 2 is AB; similarly the effect of QTL 2 depends on the genotype at QTL 1. Thus the QTL are said to interact (or to be *epistatic*): the effect of one QTL depends on the genotype at the other QTL.

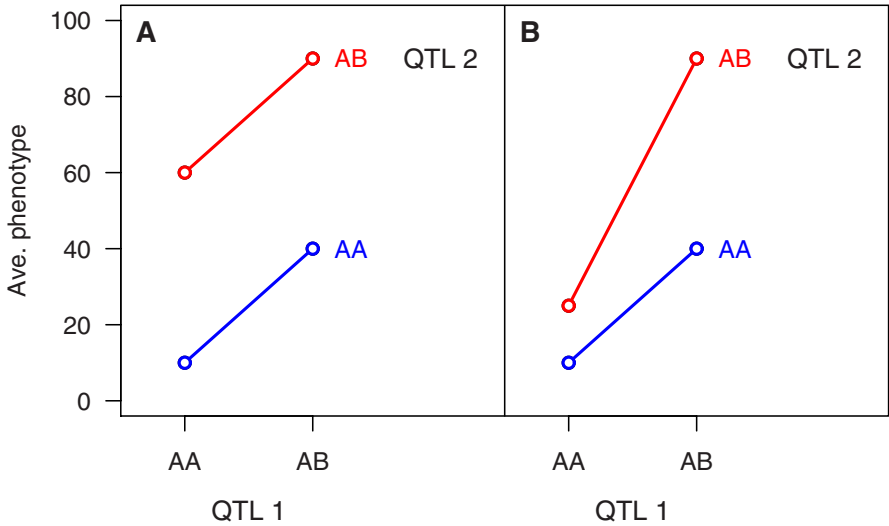


Figure 1.10. Illustration of possible effects of two QTL in a backcross. **A.** Additive QTL. **B.** Interacting QTL. Points are located at the average phenotype for a given two-locus genotype. Line segments connect the averages for a given genotype at the second QTL.

Figure 1.11 provides the analogous illustration for an intercross. The position of the average phenotype for the AB group between the averages for the AA and BB groups concerns additivity at a locus (versus dominance or recessivity). Two loci are additive (as in panel A) if the pattern of effect at one QTL is the same no matter the genotype at the other QTL: that the three curves in Fig. 1.11A are parallel. In Fig. 1.11B, the pattern of effect of QTL 1 is different for different genotypes at QTL 2, and vice versa, and so the two QTL are said to interact.

Sometimes a distinction is made between epistasis that concerns simply differences in the sizes of effects (as in Fig. 1.10B) versus changes in the sign of effect (as in Fig. 1.11B), as the former might be eliminated by a change in the scale on which the phenotype is measured, while the latter cannot be. We should emphasize further: strict additivity of QTL is rare and would be lost with a transformation of the phenotype. Thus, the question is not whether two loci interact but by how much. Further, deviation from additivity does not necessarily imply physical interaction or even a shared pathway. Polymorphisms in multiple genes may underly each QTL, and so conclusions regarding potential biological interactions are quite tenuous.

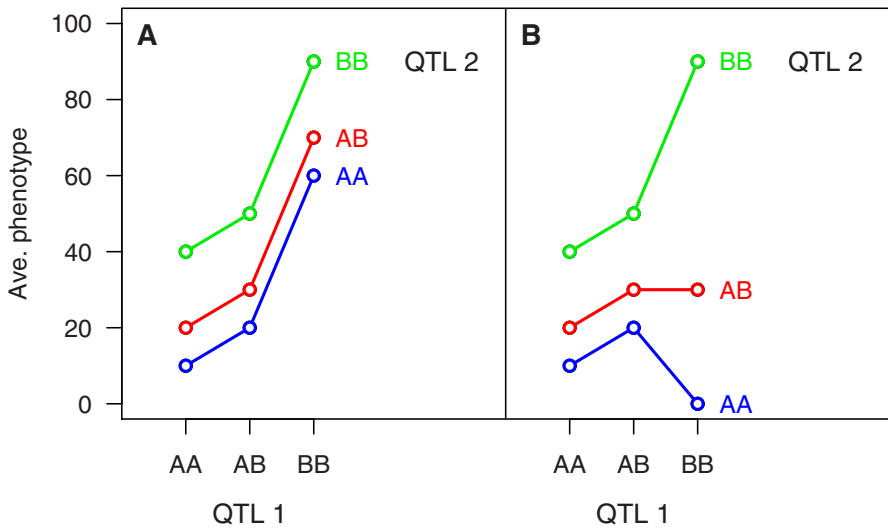


Figure 1.11. Illustration of possible effects of two QTL in an intercross. **A.** Additive QTL. **B.** Interacting QTL. Points are located at the average phenotype for a given two-locus genotype. Line segments connect the averages for a given genotype at the second QTL.

1.4 About R and R/qlt

The development of the R/qlt software was begun at the suggestion of Gary Churchill. Our primary goal was to make complex QTL mapping methods widely accessible and allow users to focus on modeling rather than computing. We further sought to develop an extensible platform for QTL mapping: to have a fast implementation of the hidden Markov model technology for dealing with the problem of missing genotype information, which forms the core of all QTL mapping methods, and to make these intermediate calculations readily accessible to the sophisticated user, so that specially tailored mapping methods can be more easily implemented.

R/qlt has been implemented as an add-on package to the general statistical software, R. R is an open source implementation of the S language, is widely used by academic statisticians, and is extensively used for microarray analyses (see the Bioconductor Project, <http://www.bioconductor.org>). As described on the R project homepage (<http://www.r-project.org>):

R is a system for statistical computation and graphics. It consists of a language plus a run-time environment with graphics, a debugger, access to certain system functions, and the ability to run programs stored in script files.

The core of R is an interpreted computer language which allows branching and looping as well as modular programming using

functions. Most of the user-visible functions in R are written in R. It is possible for the user to interface to procedures written in the C, C++, or FORTRAN languages for efficiency. The R distribution contains functionality for a large number of statistical procedures. Among these are: linear and generalized linear models, nonlinear regression models, time series analysis, classical parametric and nonparametric tests, clustering and smoothing. There is also a large set of functions which provide a flexible graphical environment for creating various kinds of data presentations. Additional modules are available for a variety of specific purposes.

The development of R/qtl as an add-on to R allows us to take advantage of the basic mathematical and statistical functions, and powerful graphics capabilities, that are provided with R. Further, the user benefits by the seamless integration of the QTL mapping software into a general statistical analysis program.

Much of the source code for R/qtl is written in R (particularly the portion that concerns data manipulation and graphics). However, most functions that require fast computation (such as those concerning hidden Markov models) were written in C.

R and R/qtl are freely available for Windows, Unix and Mac OS X, and may be downloaded from the Comprehensive R Archive Network (CRAN, <http://cran.r-project.org>). Also see the R/qtl web site, <http://www.rqtl.org>. For instructions regarding downloading and installing R and R/qtl, see Appendix A.

Much can be accomplished with R/qtl (and much of this book may be understood) with little detailed knowledge of R. Learning R may require a formidable investment of time, but it will definitely be worth the effort, both for increased facility in the use of R/qtl and for more general statistical analysis. Numerous free documents on getting started with R are available at CRAN. In addition, a growing list of books on R are available [for example, see Dalgaard (2002) or Venables and Ripley (2002)].

In learning R and R/qtl, as with any computer language or program, it is important to fiddle about: try out the example code, and explore what happens when the code is modified. In addition, one should refer to the extensive documentation included with both R and R/qtl. See Sec. A.5 for details on accessing the documentation.

1.5 Other software

There are numerous other computer programs for QTL mapping. We do not attempt to cover these exhaustively.

MapMaker/QTL was the first computer program for QTL mapping, but it has not been updated since 1994, and only allows the fit of single-QTL

models. QTL Cartographer provides more extensive facilities for the fit of multiple-QTL models, and has a graphical user interface (GUI) for Windows. Map Manager QTX is a GUI-based program that some users find most intuitive, but QTL mapping is performed exclusively with Haley–Knott regression, which can be inefficient and prone to artifacts. Commercial software programs include MapQTL and MultiQTL.

R/qtl is one of the few open source QTL mapping programs; its extensible structure, which simplifies the implementation of specially tailored mapping methods, is unique. R/qtl includes many of the important diagnostics present in MapMaker, but also allows the fit of multiple-QTL models, as in QTL Cartographer.

1.6 Work flow

In this section, we briefly describe the general work flow in QTL mapping. One must start with the design of the experiment: the choice of strains, of phenotypes to measure, of whether to perform a backcross or an intercross, what markers should be genotyped and which individuals should be genotyped. Aspects of design are discussed in Chap. 6.

After the data have been obtained, the first task is to assemble them into a computer file or files and import them into software. The import of QTL mapping data into R/qtl is discussed in Chap. 2. Second, one performs a variety of diagnostic checks on the data to identify possible errors, such as mistakes in data entry, errors in marker order, and genotyping errors. This task is discussed in Chap. 3, and is particularly important, as our ability to map QTL will be eroded by low-quality data. Aspects of the phenotype distribution may lead one to consider phenotype transformations or the use of special phenotype models, such as the two-part model discussed in Sec. 5.3.

Next, one uses interval mapping (or one of its variants), performing a genome scan with a single-QTL model, to detect loci with important marginal effects. Interval mapping is discussed in Chap. 4 and 5. The statistical significance of putative QTL is established, taking into account the genome-wide scan, generally by a permutation test (Sec. 4.3). Interval estimates for the locations of QTL may also be obtained (Sec. 4.5).

One next turns to two-dimensional, two-QTL scans of the genome (discussed in Chap. 8). Such two-dimensional scans provide the first opportunity to identify interactions between QTL, including the possibility of detecting QTL with limited marginal effects, whose importance may be seen only by considering their interaction with other loci. In addition, evidence for two linked QTL (versus a single QTL on a chromosome) is best obtained by considering an explicit two-QTL model.

Finally, one will bring all of the putative QTL and QTL $\times$ QTL interactions together into an overall multiple-QTL model (Chap. 9). In the context of a global model, some QTL may then be omitted, while the exploration

of additional QTL or interactions may lead to the addition of further terms to the model. The fit of the global model may allow some refinement in the location of QTL, and provides the most reliable estimates of the QTL effects.

The result of the analysis is some set of inferred QTL, with some understanding of their effects, locations, and possible interactions. These results will be used to guide further experiments, perhaps with the aim of fine-mapping the QTL and ultimately identifying the underlying gene or genes.

1.7 Further reading

There are numerous review articles on QTL mapping (e.g., Doerge *et al.*, 1997; Broman and Speed, 1999; Jansen, 2007; Broman, 2001). We particularly like the review of Jansen (2007). Broman (2001) is one of the few reviews written for nonstatisticians.

A number of books have some discussion of QTL mapping: Falconer and Mackay (1996) has a chapter on QTL mapping, and Lynch and Walsh (1998) and Liu (1998) each have several. Silver (1995) is a very nice book on mouse genetics, and it is freely available online at <http://www.informatics.jax.org/silver>. Also see the recent book by Wu *et al.* (2007).

McPeck (1996) provides a useful introduction to recombination and crossover interference. See also McPeck and Speed (1995). For a discussion of map functions, see Speed (1996) and Zhao and Speed (1996). Broman *et al.* (2002) studied crossover interference in the mouse. Strickberger (1985) contains an excellent chapter on epistasis.

Ihaka and Gentleman (1996) is the paper introducing R. Broman *et al.* (2003) is the original article reporting R/qtl. The most important book on R is Venables and Ripley (2002); every user of R should have a copy. Dalgaard (2002) provides a more gentle introduction.

Lander *et al.* (1987) is the original paper on MapMaker; Manly *et al.* (2001) described Map Manager. The only clear reference on QTL Cartographer is the manual (Basten *et al.*, 2002), available online at <http://statgen.ncsu.edu/qtlcart/manual>.