Statistical Methods for Mapping Quantitative Trait Loci

Zhao-Bang Zeng
Program in Statistical Genetics
Department of Statistics
North Carolina State University
Raleigh NC 27695-8203

August 21, 2000

Contents

Chapter 1

Introduction

1.1 Some basic genetic concepts

Genes are fundamental units of genetic information that are transmitted from parent to offspring in reproduction and influence hereditary traits of organisms. Genes are composed by the sequence of nucleotides in a segment of **DNA** (deoxyribonucleric acid). A molecule of DNA consists of two strands wound around each other in the form of right-hand helix. Each strand is composed by a sequence of **nucleotides** of four types, either adenine (A), thymine (T), guanine (G) or cytosine (C). The strands are held together by pairing between the bases A and T and between G and C in opposite strands. The whole set of DNA sequences for an organism is called **genome**. For human, the DNA sequences are linearly arranged in 23 chromosomes (22 autosomes and a pair of sex-chromosomes) with the total nucleotide base pairs about 3×10^9. Genes can have different forms or states. These alternative forms of a gene are called **alleles**.

Most genes code for the polypeptide chains that constitute **proteins**. Genes carry the information, while proteins provide the means of executing it. The sequence of DNA is related to the sequence of protein by the **genetic code**. A coding sequence of DNA consists of a series of **codons**, read as nonoverlapping triplet from a fixed starting point. **Mutations** in the sequence of DNA, *i.e.* errors in DNA replication, can change the sequence of amino acids in the protein. Of the 64 triplets, 61 code for 20 **amino acids** and 3 provide termination signals. Synonym codons that represent the same amino acid are related, often by a change in the third base of the codon. This third base degeneracy, coupled with the arrangement in which related amino acids tend to be coded by related codons, minimizes the effects of mutations. The genetic code is universal.

Genetic information carried by DNA is expressed in two stages: **transcription** of one DNA strand to mRNA (messenger ribonucleic acid)) and **translation** of mRNA into protein. The coding regions of DNA in a gene are often interrupted by one or more non-coding regions known as intervening sequences or **introns**. Every gene also includes one or more regulatory regions that determine when transcription takes place, the type of cells in which it takes place, and which strand that is to be transcribed. In the first step of gene expression

(transcription), a molecule of RNA is produced that is complementary in base sequence to one of the strands of DNA including the sequence in the introns. The second step in gene expression is RNA processing in which the beginning and end of the RNA transcript are chemically modified and the introns are removed by splicing. RNA processing results in a molecular called mRNA in which the coding regions have been made contiguous. The region in the original RNA that are retained in the mature mRNA are called **exons**. The mRNA also includes an upstream region (the 5′ untranslated region) and a downstream region (the 3′ untranslated region) besides of the protein-coding region. The final step in gene expression is translation, in which the mRNA molecule combines with ribosomes and other types of RNA molecules in the cytoplasm to produce the final peptide to form protein. The information is translated according to genetic code.

The **genotype** consists of the complete set of genetic information inherited by an organism; its expression is responsible for generating the **phenotype**, the physical form of the organism which is also affected by environmental factors. The genotype includes many genes, organized into **chromosomes**. Genes far apart on a chromosome behave like genes on different chromosomes and obey **Mendel's laws**, which treat genes as discrete factors. Alleles segregate and genes assort independently.

Genes on the chromosome form a linear linkage group, in which those genes near one another tend to inherit together. Linkage between genes can be used to construct a **linkage map**, which provides a linear representation of the locations of the genes on a chromosome.

1.2 Quantitative traits

Quantitative genetics is concerned with the inheritance of those differences between individuals that are of degree rather than of kind, quantitative rather than qualitative. These individual differences are called **quantitative traits**. They are a major part of the variation of individuals for any species, and are the materials for natural and artificial selection. An understanding of the inheritance of these differences is of fundamental significance in the study of evolution and in the application of genetics to animal and plant breeding and also to medicine.

Virtually every biological aspect of any species shows individual differences of this nature. They do not show simple Mendelian transmission, a characteristic for qualitative traits. These differences are usually considered as resulting from the combined effects of many causal factors, some genetic in origin and some environmental. Because of these complex genetic and environmental effects, study in quantitative genetics involves a wealth of statistical methodology and tools.

There are three types of quantitative traits:

1. **Continuous traits**. There is a continuum of possible phenotypes for these traits in a population. The phenotype of an individual can take on any one of a continuous range of values. There are numerous traits of this kind, such as height, weight, corn yield, and growth rate.

2. **Meristic traits**. The phenotype is expressed in discrete, integral classes, such as number of offspring or litter size, number of ears on a stalk of corn, and number of bristles on a fruit fly. When the number of possible phenotypes of a meristic trait is large, the difference between continuous traits and meristic traits becomes small.

3. **Discrete traits** or **threshold traits**: These traits are usually expressed as either presence or absence of a characteristic in an individual. In these cases, it is considered that the multiple genetic and environmental factors combine to determine an underlying risk or **liability** toward the trait. Liability values are not directly observable. However, an individual that actually expresses the trait is assumed to have liability value greater than some threshold value. Examples in human genetics include diabetes and schizophrenia. Meristic traits can also be considered as threshold traits in multiple categories.

Genetical studies in quantitative traits are not restricted to morphological traits (such as height, weight, yield and bristle number). There have been more and more studies on physiological traits and biochemical traits (such as blood pressure and cholesterol level), complex genetic diseases (such as heart disease and diabetes) and also behavioral traits (such as reaction to novel environment and alcoholism).

1.3 QTL mapping

Genes that affect quantitative trait variation in a population are called **quantitative trait loci (QTL)**. A very important study in quantitative genetics is to localize QTL on a genetic linkage map and further through more detailed genetic studies to characterize QTL, which may include the identification of DNA sequence polymorphisms that cause the quantitative trait variation.

QTL mapping is basically a genome-wide inference of the relationship between phenotypic values of quantitative traits and genotypes of QTL. This relationship includes the number and genomic positions of QTL. It also includes the effects of QTL, the interaction of QTL alleles within (**dominance**) and between (**epistasis**) loci, pleiotropic effects of QTL, and QTL by environment interaction. This relationship is also called the **genetic architecture of quantitative traits**. Depending on the data and the nature of molecular markers used for mapping analysis, we will see, however, that what is usually identified as a QTL is a segment of chromosome that affects a quantitative trait, not necessarily a single locus.

Identification of QTL are important for our understanding of genetic nature of quantitative trait variation within a population and between populations or species. Biologically, it is important to know many genes are involved for a quantitative trait within and between populations. What is the distribution of effects of QTL? How much genetic variation is due to additive effects, dominance effects and epistatic effects for QTL? Are QTL effects dependent significantly on environments? How many QTL have pleiotropic effects on multiple traits? Are the effects in a chromosome region on multiple traits due to a common

pleiotropic QTL or multiple non-pleiotropic QTL in close linkage?

Indentification of QTL can lead to several useful applications. First, it could improve the efficacy of selective breeding of animals and plants, particularly for traits with low heritability or that can only be measured in one sex. Second, it could facilitate transgression of QTL alleles from one population to another. Third, in medicine, the identification of alleles causing predisposition to complex genetic diseases could lead to improved methods of prevention. It is also the first step toward to functional genetic analysis of quantitative traits.

Data for QTL mapping constitute of two parts. One is the scores of genotypes or phenotypes of a number of molecular markers for a number of individuals, and the other is the measurements of one or more quantitative traits of the individuals. This is briefly discussed in Chapter 2. Chapter 3 introduces the statistical methods for marker analysis, which include the marker segregation analysis, testing and estimation of linkage between markers, map functions, and marker linkage map construction. In Chapter 4, we will discuss quantitative genetic models. These models provide the statistical basis and specify the parameters to interprete the genetic architecture of quantitative traits. In the next several chapters, we will introduce several statistical methods used for mapping QTL (in a single trait), starting with the single marker analysis (Chapter 5), then interval mapping (Chapter 6), composite interval mapping (Chapter 7), and finally to multiple interval mapping (Chapter 8). The QTL mapping analyzes with multiple traits and environments are discussed in Chapter 9 in the framework of composite interval mapping. Chapter 10 deals with missing and dominant marker analysis for several experimental designs commonly used for inbred lines. Chapter 11 introduces the full-sib family, particularly on the inference of multiple-marker linkage phases.

A computer software, called **QTL CARTOGRAPHER** (Basten, Weir and Zeng, 1995-2000), has been developed to accompany this course. Many statistical methods (not all yet) discussed in this course have been implemented to the software. The software and manual can be downloaded from the web site (http://statgen.ncsu.edu/qtlcart/cartographer.html).

Chapter 2

QTL Mapping Data

2.1 Sample data

- Data for mapping QTL have two components:

    1. Marker data: categorical data; contain information about segregation of a genome at various positions in a population.

       Examples: RFLP (Restriction Fragment Length Polymorphism), SSR (Simple Sequence Repeats), RAPD (Random Amplified Polymorphic DNA), AFLP (Amplified Fragment Length Polymorphism), SNP (Single Nucleotide Polymorphism).

    2. Trait data: continuous or discrete data; contain information about segregation and effects of QTL in the same population.

       Examples: 12 week body weight of mouse, Grain yield of maize, Litter size of pigs, Blood pressure, Disease resistant score.

- For most part of discussion, we are going to use two data sets as examples to illustrate some statistical analyses discussed.

    1. Mouse data set: The data contain typing results of 189 markers (covering all 20 chromosomes of mouse genome) and 12 week body weight of 103 individuals from a backcross mouse population. Only 14 markers on X chromosome are shown in Table 2.1, however.
    2. Maize data set: This is a F_2 data set. The sample size of the population is 171. The total number of genotyped markers is 184, covering all 10 chromosomes of maize genome. However, only 12 markers on chromosome 5 are shown in Table 2.2.

Table 2.1: Mouse mapping data

		Markers on X chromosome																Markers on X chromosome													
Ind	BW	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Ind	BW	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	50	1	1	1	1	1	1	1	1	1	1	1	1	1	1	53	62	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	54	1	1	1	1	1	1	1	1	1	1	1	1	1	0	54	49	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3	49	0	1	1	1	1	1	1	1	1	1	1	1	1	1	55	59	0	0	1	1	1	1	1	1	1	1	1	1	1	1
4	41	0	0	0	0	0	0	0	0	0	0	0	0	0	0	56	35	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	36	1	1	1	1	1	1	1	1	1	1	1	1	1	1	57	43	0	0	0	0	0	0	0	0	1	1	1	1	1	1
6	48	0	0	0	0	0	0	0	0	0	0	0	0	0	0	58	45	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	37	0	0	0	0	0	0	1	1	1	1	1	1	1	1	59	44	0	0	0	0	1	1	1	1	1	1	0	1	1	0
8	55	1	1	1	1	1	1	1	1	1	1	1	1	1	0	60	47	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	42	0	0	0	0	0	0	0	0	0	0	0	0	0	0	61	51	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	46	0	0	0	0	0	0	0	0	0	0	0	0	0	0	62	50	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	39	0	0	0	0	1	1	1	1	1	1	1	1	1	1	63	44	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	58	1	1	1	1	1	1	1	1	1	1	1	1	0	0	64	44	0	0	0	0	0	0	0	0	0	0	0	0	1	1
13	56	1	1	1	1	1	1	1	1	1	1	1	1	1	0	65	49	0	1	1	1	1	1	1	1	1	1	1	1	1	0
14	44	0	0	0	0	0	0	0	0	0	0	0	0	0	0	66	43	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	60	1	1	1	1	1	1	1	1	1	1	1	1	1	1	67	45	0	0	0	0	0	0	0	0	0	0	0	0	1	1
16	70	1	1	1	1	1	1	1	1	1	1	1	0	1	1	68	53	0	0	0	0	0	0	0	0	0	0	0	0	0	1
17	62	0	0	0	0	0	1	1	1	1	1	1	1	1	1	69	42	0	0	0	0	0	0	0	0	0	0	1	1	1	1
18	48	0	0	0	0	0	0	0	0	0	0	0	0	0	0	70	43	1	1	1	1	1	1	1	1	1	1	1	1	1	1
19	51	0	0	0	0	0	0	0	0	0	0	0	0	0	0	71	58	1	1	1	1	1	1	1	1	1	1	1	1	1	1
20	48	0	0	0	0	0	0	0	0	0	0	0	0	0	0	72	36	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	44	0	0	0	0	0	0	0	0	0	0	0	0	0	0	73	51	1	0	0	0	0	0	0	0	0	0	0	0	0	0
22	71	1	1	1	1	1	1	1	1	1	1	1	1	1	1	74	47	1	1	1	1	1	1	1	1	1	1	1	1	1	1
23	49	0	0	0	0	0	1	1	1	1	1	1	1	1	1	75	68	1	1	1	1	1	1	1	1	1	1	1	1	1	0
24	40	0	0	0	0	0	0	0	0	0	0	0	0	1	1	76	56	1	1	1	1	1	1	1	1	1	1	1	1	1	1
25	51	1	1	1	1	1	1	0	0	0	0	0	0	0	0	77	56	1	1	1	1	1	1	1	1	1	1	1	1	1	1
26	43	0	0	0	0	0	0	0	0	0	0	0	0	0	0	78	69	1	1	1	1	1	1	1	1	1	1	1	1	1	1
27	63	1	1	1	1	1	1	1	1	1	1	1	1	1	1	79	46	1	1	1	1	1	1	0	0	0	0	0	0	0	0
28	62	1	1	1	1	1	1	1	1	1	1	1	1	1	1	80	52	1	1	1	1	1	1	1	1	1	1	1	1	1	1
29	71	1	1	1	1	1	1	1	1	1	1	1	1	1	1	81	43	1	0	0	0	0	0	0	0	0	0	1	1	1	1
30	52	1	1	1	1	1	1	1	1	1	1	1	1	1	1	82	40	0	0	0	0	0	0	0	0	0	0	0	0	1	1
31	47	0	0	0	0	0	0	0	0	0	0	0	0	0	0	83	79	1	1	1	1	1	1	1	1	1	1	1	1	1	1
32	56	0	0	0	0	0	0	0	1	1	1	1	1	1	1	84	57	0	0	0	0	0	0	0	1	1	1	1	1	1	1
33	53	0	0	0	0	0	0	0	0	0	1	1	1	1	1	85	43	1	1	1	1	1	1	1	1	1	1	1	1	1	1
34	50	1	1	1	1	1	1	1	1	1	0	0	0	0	0	86	56	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	71	1	1	1	1	1	1	1	1	1	1	1	1	1	1	87	40	0	0	0	0	0	0	0	0	0	0	0	0	0	0
36	53	0	0	0	0	0	0	0	0	0	0	0	0	0	0	88	44	0	0	0	0	0	0	0	0	0	0	0	0	0	0
37	57	0	1	1	1	1	1	1	1	1	1	1	1	1	0	89	40	0	0	0	0	0	0	0	0	0	0	0	0	0	0
38	41	0	0	0	0	0	0	0	0	0	0	0	0	0	0	90	66	0	0	0	0	0	0	0	0	0	0	0	0	1	1
39	30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	91	32	0	0	0	0	0	0	0	0	0	0	0	0	0	0
40	56	1	1	1	1	1	1	1	1	1	1	1	1	1	1	92	45	0	0	0	0	0	0	0	0	0	0	0	0	0	0
41	59	1	1	1	1	1	1	1	1	1	1	1	1	1	1	93	45	1	1	1	1	1	0	0	0	0	0	0	0	0	0
42	60	0	0	0	0	0	0	1	1	1	1	1	1	1	1	94	53	0	0	0	0	0	0	0	0	0	0	0	0	0	0
43	46	1	1	1	1	1	1	1	1	1	1	1	1	1	1	95	48	0	0	0	0	0	0	0	1	1	1	1	1	1	0
44	50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	96	36	0	0	0	0	0	0	1	1	1	1	1	1	1	1
45	49	1	1	1	1	1	1	1	1	1	1	1	1	1	1	97	50	0	0	0	0	0	0	0	0	0	0	0	0	0	0
46	71	0	1	1	1	1	1	1	1	1	1	1	1	1	1	98	56	0	0	0	0	0	0	0	0	0	0	0	0	0	0
47	55	0	0	0	0	0	0	0	0	0	0	0	0	0	0	99	50	0	0	0	0	0	0	0	0	0	0	0	0	0	0
48	52	0	0	0	0	0	0	0	1	1	1	1	1	1	1	100	45	1	1	1	1	0	0	0	0	0	0	0	0	0	0
49	44	0	0	0	0	0	0	0	0	0	0	0	0	0	0	101	43	1	1	1	0	0	0	0	0	0	0	0	0	0	0
50	35	1	1	1	1	1	1	1	1	1	1	1	1	1	1	102	37	0	0	0	0	0	0	0	0	0	0	0	0	0	1
51	57	1	1	1	1	1	1	1	1	1	1	1	1	1	0	103	35	1	0	0	0	0	0	0	0	0	0	0	0	0	0
52	54	0	0	0	0	0	0	0	0	0	0	0	0	0	0																

Table 2.2: Maize mapping data

		Markers														Markers											
Ind	Trait	1	2	3	4	5	6	7	8	9	10	11	12	Ind	Trait	1	2	3	4	5	6	7	8	9	10	11	12
1	6.25	2	1	0	0	0	0	0	1	1	1	0	0	51	6.50	1	1	1	1	1	1	1	2	1	1	1	1
2	3.00	1	1	1	1	1	2	2	2	2	0	0	1	52	2.75	2	1	1	0	0	0	0	1	0	0	0	0
3	3.00	1	2	2	2	2	1	1	1	1	2	2	2	53	3.50	1	2	2	2	2	2	2	2	2	2	2	2
4	4.00	1	0	0	0	0	0	0	0	0	1	2	2	54	4.25	1	2	2	2	2	2	2	2	2	2	2	2
5	3.00	0	0	1	1	1	1	1	1	1	1	1	1	55	3.00	0	1	2	2	2	2	2	0	1	1	1	1
6	3.75	1	0	0	0	0	1	1	1	1	0	0	0	56	6.25	1	0	1	1	1	2	2	2	2	1	1	1
7	8.25	2	2	2	1	1	0	0	0	0	1	2	2	57	1.50	0	0	0	0	0	0	1	1	1	0	0	0
8	2.50	0	0	0	0	0	0	0	1	1	1	1	2	58	3.25	1	1	0	1	1	1	1	1	1	1	1	1
9	4.25	1	0	1	1	1	1	1	0	0	1	1	1	59	4.75	1	0	0	0	0	1	2	2	2	1	1	1
10	4.50	0	1	1	1	1	1	1	1	0	0	0	0	60	2.25	0	1	1	1	1	1	1	1	0	0	0	2
11	6.00	1	1	2	2	2	1	1	1	1	1	1	0	61	2.75	1	1	1	1	1	1	1	1	1	2	2	2
12	3.25	2	1	1	1	1	1	1	1	1	1	1	1	62	5.00	2	2	2	2	2	2	2	1	1	1	1	1
13	5.50	2	1	2	1	1	2	2	1	1	1	1	1	63	3.00	2	1	0	0	0	0	0	0	0	1	1	1
14	7.25	1	1	1	2	2	2	2	2	1	1	1	1	64	5.00	0	1	1	1	1	1	1	1	0	0	1	1
15	5.50	1	2	2	2	2	2	2	2	2	2	2	2	65	3.25	1	0	0	0	0	1	1	2	2	2	2	2
16	4.00	1	1	1	1	1	1	1	1	1	1	1	1	66	6.75	1	1	1	1	1	1	1	1	1	1	1	1
17	2.75	1	1	1	1	1	1	0	0	0	0	0	0	67	6.25	1	1	1	1	1	1	1	1	1	0	0	0
18	7.50	2	2	2	2	2	2	2	1	0	0	0	0	68	1.25	0	2	2	2	2	2	2	1	1	1	2	2
19	6.00	0	0	0	0	0	0	0	1	1	1	1	1	69	2.75	1	2	2	2	2	1	1	1	1	1	1	1
20	4.75	2	1	1	1	1	1	1	1	1	1	0	0	70	3.00	0	1	1	2	2	2	1	0	0	1	1	1
21	2.75	0	1	1	1	1	0	1	1	1	0	0	0	71	8.25	2	2	2	1	1	2	2	2	2	1	1	1
22	5.50	2	1	1	1	1	0	0	0	0	0	0	0	72	3.50	1	2	1	1	1	2	2	2	2	1	1	0
23	1.75	1	1	1	1	1	1	2	2	2	2	2	2	73	2.75	1	0	0	2	0	1	1	1	1	1	1	2
24	5.00	1	1	1	2	2	1	1	0	0	0	0	0	74	7.25	1	1	1	0	2	1	0	0	0	0	0	1
25	4.00	2	2	1	1	1	1	1	1	1	1	1	1	75	5.75	2	1	0	0	0	1	1	1	1	1	1	1
26	4.00	1	2	2	2	2	2	0	0	0	0	0	0	76	5.25	1	1	1	1	1	1	1	1	1	0	0	0
27	2.00	0	1	1	1	1	1	1	2	2	2	2	2	77	6.75	2	2	1	1	1	0	0	1	1	2	2	2
28	3.00	0	1	2	2	2	1	1	0	0	1	1	1	78	6.00	2	1	1	1	1	1	1	2	2	2	2	2
29	4.75	2	1	1	1	1	0	0	0	0	0	1	1	79	1.50	0	1	1	1	1	1	1	1	1	2	2	2
30	3.75	2	0	0	0	0	1	1	1	1	1	1	1	80	3.50	2	1	0	1	1	1	2	2	2	2	2	2
31	5.75	1	2	2	2	2	2	2	2	2	2	2	2	81	5.00	0	1	2	2	2	1	1	0	0	0	0	0
32	8.00	2	1	1	1	1	0	1	2	1	1	1	1	82	5.00	0	0	0	1	1	1	0	1	1	1	1	1
33	3.00	1	0	0	1	1	2	2	2	1	1	1	0	83	3.75	1	2	1	1	1	2	2	2	2	2	2	2
34	2.25	1	1	1	0	0	0	0	0	1	1	1	1	84	4.00	1	1	1	1	1	1	1	1	1	1	1	1
35	3.50	1	1	1	1	1	2	2	2	2	2	2	2	85	4.50	2	2	1	1	1	2	2	2	2	1	1	1
36	6.75	0	1	1	1	1	0	0	0	0	0	0	0	86	1.50	0	1	1	1	1	0	0	0	0	1	1	1
37	2.25	1	2	2	2	2	1	1	1	1	1	1	2	87	7.50	1	1	1	0	0	0	0	0	0	0	0	0
38	3.50	1	1	1	1	1	1	1	1	1	2	2	2	88	3.25	1	1	1	1	1	1	1	2	2	1	1	1
39	6.25	1	1	1	0	0	0	0	0	0	0	0	0	89	7.50	1	1	1	0	0	0	0	1	1	1	2	2
40	1.50	0	1	1	1	1	0	0	0	0	1	1	1	90	2.75	1	0	0	0	0	0	0	0	0	0	0	0
41	3.50	1	1	1	1	1	1	1	1	1	1	1	1	91	5.50	1	1	1	1	1	1	1	1	1	2	2	1
42	4.00	2	2	2	2	2	2	2	2	2	2	2	2	92	2.50	2	1	1	1	0	0	0	0	0	0	0	1
43	6.25	2	2	2	2	2	2	2	2	2	1	1	1	93	3.00	1	1	1	1	1	1	1	2	2	2	2	1
44	3.00	1	2	2	2	2	1	1	1	1	1	2	2	94	5.00	1	1	1	1	1	1	0	0	0	1	1	2
45	7.25	2	2	2	1	1	0	1	1	1	1	1	1	95	3.00	1	1	2	2	2	2	2	2	2	1	1	0
46	1.75	0	0	0	0	0	0	1	1	1	2	1	1	96	2.75	2	1	0	0	0	0	1	2	2	2	2	2
47	7.25	0	0	0	0	0	1	1	1	1	1	2	1	97	2.75	1	0	0	1	1	1	1	1	1	0	0	0
48	1.25	0	0	0	0	0	1	1	1	1	1	1	1	98	3.50	0	2	2	2	2	1	1	2	1	0	0	1
49	7.25	1	2	2	1	1	1	1	1	1	1	2	1	99	7.00	1	1	1	1	1	1	1	1	1	1	1	1
50	6.75	2	2	2	2	2	2	2	2	2	2	2	2	100	1.75	0	0	0	1	1	1	1	1	2	2	2	2

Table 2.2 (continue)

		Markers															Markers											
Ind	Trait	1	2	3	4	5	6	7	8	9	10	11	12	Ind	Trait	1	2	3	4	5	6	7	8	9	10	11	12
101	2.75	0	0	0	0	0	1	1	1	1	1	1	1	137	7.50	1	2	1	1	1	1	1	1	1	1	1	0
102	3.50	2	2	2	2	2	2	2	2	2	1	1	0	138	2.50	0	1	1	1	1	0	1	1	1	1	1	0
103	6.75	2	2	2	2	2	2	2	2	2	2	2	2	139	2.00	0	0	0	0	0	0	0	0	0	1	1	1
104	2.75	0	0	0	0	0	1	1	2	2	1	1	1	140	2.50	1	2	1	1	1	2	2	2	2	2	2	2
105	3.00	1	1	1	1	1	1	1	1	1	1	0	0	141	2.50	1	1	1	1	1	1	1	1	0	1	1	1
106	2.75	0	1	1	1	1	1	2	2	2	2	1	1	142	7.50	1	1	2	2	2	2	1	0	1	2	2	2
107	2.75	1	1	1	2	2	0	0	0	0	1	1	1	143	2.00	0	1	1	1	1	0	0	0	0	0	1	1
108	5.75	2	1	1	1	1	1	1	2	1	1	1	1	144	4.00	2	0	0	0	0	0	0	0	1	1	0	0
109	2.75	0	1	1	2	2	2	2	2	2	1	1	1	145	2.75	1	1	1	1	1	1	1	1	1	2	2	2
110	6.50	2	2	2	2	2	1	1	0	1	1	1	1	146	3.00	1	1	1	1	1	2	2	2	2	2	2	2
111	1.75	1	0	0	0	0	1	1	1	1	1	1	1	147	3.75	1	1	1	2	2	2	2	2	2	2	2	2
112	4.75	2	1	1	0	0	1	1	1	1	1	1	1	148	6.50	0	0	0	0	0	0	0	1	1	1	1	1
113	4.75	2	2	2	2	2	2	2	2	2	1	1	1	149	4.25	1	1	1	1	1	0	0	1	1	1	1	1
114	3.50	2	1	1	0	0	0	0	1	1	1	1	1	150	3.50	1	1	1	1	1	0	0	0	0	0	0	0
115	5.50	1	0	0	0	0	0	0	0	0	0	0	0	151	1.75	1	1	0	0	0	0	0	0	1	1	2	1
116	6.50	1	1	2	2	2	0	0	0	0	0	0	1	152	4.00	0	1	1	1	1	1	1	1	1	2	2	1
117	5.00	2	2	1	1	1	1	1	1	1	1	1	1	153	5.25	1	1	1	1	0	0	1	1	1	1	1	2
118	4.75	0	2	2	2	2	2	2	2	2	2	2	2	154	2.50	1	1	1	1	1	1	0	0	0	1	1	1
119	2.25	1	2	1	1	1	2	2	1	1	1	1	0	155	7.25	2	2	2	2	2	0	0	0	0	2	2	2
120	2.00	0	1	1	1	1	2	2	2	1	2	2	2	156	2.75	1	2	1	1	1	1	1	1	1	0	0	0
121	4.25	0	2	2	2	2	2	2	2	2	2	2	2	157	1.00	0	0	0	0	0	0	0	0	0	0	0	0
122	4.25	1	2	1	1	1	1	1	1	1	1	1	1	158	7.50	1	1	1	1	1	1	1	1	1	1	1	1
123	6.00	1	0	0	1	1	1	1	1	1	0	0	0	159	4.50	1	1	2	2	2	2	0	0	0	1	1	1
124	5.25	2	2	1	1	1	2	2	2	2	1	1	1	160	5.75	1	1	1	1	1	1	1	1	1	2	1	1
125	6.00	2	2	2	1	1	2	2	1	1	1	1	1	161	4.50	1	1	2	2	1	1	1	0	0	0	0	0
126	3.50	0	0	0	0	0	0	0	0	0	1	1	1	162	4.50	2	1	1	1	1	1	1	0	0	0	0	0
127	2.25	1	2	2	1	1	1	1	1	1	2	1	1	163	3.50	1	1	1	1	1	1	1	2	2	2	2	2
128	3.00	1	2	1	1	1	1	1	1	1	1	1	1	164	5.50	1	2	2	2	2	2	2	2	2	1	1	1
129	4.50	1	2	2	2	2	2	2	2	2	2	2	2	165	1.75	1	1	1	1	1	1	2	1	1	1	1	1
130	4.00	1	1	1	0	0	0	0	0	0	0	0	0	166	5.75	1	0	0	0	0	0	0	0	0	1	2	2
131	8.25	1	1	1	1	1	1	0	0	0	1	1	1	167	4.00	2	2	1	1	1	0	0	0	0	1	1	1
132	4.75	2	2	1	0	0	1	1	1	1	2	2	2	168	3.50	1	0	0	0	0	1	1	1	1	1	1	2
133	1.50	0	1	1	1	1	1	1	1	1	1	1	1	169	5.50	0	1	1	1	1	0	1	1	1	1	1	1
134	1.25	0	0	1	2	2	2	2	1	1	1	1	1	170	3.00	1	0	0	1	1	1	1	1	1	1	1	1
135	2.25	0	1	1	2	2	1	1	1	1	1	1	1	171	5.75	2	2	1	1	0	1	1	1	1	0	0	0
136	6.75	2	2	2	2	2	2	2	2	2	1	1	0														

2.2 Experimental designs

Traditional experimental designs for locating QTL start with two parental inbred lines, P_1 and P_2, differing both in trait values and in the marker variants they carry. In practice, markers are sought that have different variants in parents.

Advantages: F_1 is heterozygote for all loci which differ in P_1 and P_2, and has the maximum linkage disequilibrium. For mapping QTL, this type of experimental design has the maximum power.

For most part of discussion, we will consider backcross and F_2 designs to illustrate statistical analyses discussed. In the end, we will discuss methods to extend the analyses to other types of experimental designs. A schematic diagram of the designs is as follows:

$$
\begin{array}{ccccc}
\mathrm{P}_1 & & \times & & \mathrm{P}_2 \\
\frac{AB}{AB} & & \downarrow & & \frac{ab}{ab} \\
\downarrow & & \mathrm{F}_1 & & \downarrow \\
\downarrow & & \frac{AB}{ab} & & \downarrow \\
AB & \times & \begin{bmatrix} AB & Ab & aB & ab \\ \frac{1-r}{2} & \frac{r}{2} & \frac{r}{2} & \frac{1-r}{2} \end{bmatrix} & \times & ab \\
\downarrow & & \downarrow & & \downarrow \\
\mathrm{B}_1 & & \downarrow & & \mathrm{B}_2 \\
\begin{bmatrix} \frac{AB}{AB} & \frac{AB}{Ab} & \frac{AB}{aB} & \frac{AB}{ab} \\ \frac{1-r}{2} & \frac{r}{2} & \frac{r}{2} & \frac{1-r}{2} \end{bmatrix} & & \downarrow & & \begin{bmatrix} \frac{ab}{AB} & \frac{ab}{Ab} & \frac{ab}{aB} & \frac{ab}{ab} \\ \frac{1-r}{2} & \frac{r}{2} & \frac{r}{2} & \frac{1-r}{2} \end{bmatrix} \\
 & & \downarrow & & \\
 & & \mathrm{F}_2 & &
\end{array}
$$

$$
\begin{bmatrix}
\frac{AB}{AB} & \frac{AB}{Ab} & \frac{Ab}{Ab} & \frac{AB}{aB} & \frac{AB}{ab}/\frac{Ab}{aB} & \frac{Ab}{ab} & \frac{aB}{aB} & \frac{aB}{ab} & \frac{ab}{ab} \\
\frac{(1-r)^2}{4} & \frac{r(1-r)}{2} & \frac{r^2}{4} & \frac{r(1-r)}{2} & \frac{(1-r)^2+r^2}{2} & \frac{r(1-r)}{2} & \frac{r^2}{4} & \frac{r(1-r)}{2} & \frac{(1-r)^2}{4}
\end{bmatrix}
$$

Another commonly used experimental design in experimental animals and plants is recombinant inbred lines, which has the following structure:

$$
\begin{array}{ccc}
\mathrm{P}_1\text{: } \frac{AB}{AB} & \times & \mathrm{P}_2\text{: } \frac{ab}{ab} \\
 & \downarrow & \\
 & \mathrm{F}_1\text{: } \frac{AB}{ab} & \\
 & \downarrow & \\
 & \mathrm{F}_2\text{:} &
\end{array}
$$

$$\begin{bmatrix} \frac{AB}{AB} & \frac{AB}{Ab} & \frac{Ab}{Ab} & \frac{AB}{aB} & \frac{AB}{ab}/\frac{Ab}{aB} & \frac{Ab}{ab} & \frac{aB}{aB} & \frac{aB}{ab} & \frac{ab}{ab} \\ \frac{(1-r)^2}{4} & \frac{r(1-r)}{2} & \frac{r^2}{4} & \frac{r(1-r)}{2} & \frac{(1-r)^2+r^2}{2} & \frac{r(1-r)}{2} & \frac{r^2}{4} & \frac{r(1-r)}{2} & \frac{(1-r)^2}{4} \end{bmatrix}$$

$$\downarrow$$

$$\vdots$$

$$\downarrow$$

$$\mathrm{F}_\infty\text{:}$$

$$\begin{bmatrix} \frac{AB}{AB} & \frac{Ab}{Ab} & \frac{aB}{aB} & \frac{ab}{ab} \\ \frac{1}{1+2r} & \frac{2r}{1+2r} & \frac{2r}{1+2r} & \frac{1}{1+2r} \end{bmatrix}$$

Chapter 3

Linkage Map Analysis

3.1 Testing Mendelian segregation

3.1.1 Backcross population

A cross between A/A and A/a produces the following zygotes

	A/A	A/a
Frequency under H_0	$1/2$	$1/2$
Expected number	$n/2$	$n/2$
Observed number	n_1	n_2

A test statistic can be constructed by using χ^2 under the null hypothesis $p(A/A) = p(A/a) = 1/2$ (Mendelian segregation).

$$\chi^2 = \sum \frac{(\text{Obs.\#} - \text{Exp.\#})^2}{\text{Exp.\#}} = \frac{(n_1 - n/2)^2}{n/2} + \frac{(n_2 - n/2)^2}{n/2} = \frac{(n_1 - n_2)^2}{n} \sim \chi_1^2$$

Under the null hypothesis, this statistic is chi-square distributed with 1 degree of freedom.

3.1.2 F_2 population

For F_2 which is a cross between A/a and A/a, the distribution of zygotes is as follows:

	A/A	A/a	a/a
Frequency under H_0	$1/4$	$1/2$	$1/4$
Expected number	$n/4$	$n/2$	$n/4$
Observed number	n_1	n_2	n_3

Table 3.1: Example of testing Mendelian segregation: Mouse data

Marker	n_1	n_0	χ^2	P value
1 Hmg1-rs13	41	62	4.282	0.038
2 DXMit57	42	61	3.505	0.061
3 Rps17-rs11	43	60	2.806	0.094
4 Rps18-rs17	42	61	3.505	0.061
5 DXMit48	43	60	2.806	0.094
6 DXNds1	44	59	2.184	0.142
7 DXMit109	45	58	1.641	0.20
8 Hmg14-rs6	49	54	0.243	0.61
9 DXMit60	50	53	0.087	0.77
10 DXMit16	50	53	0.087	0.77
11 DXMit97	50	53	0.087	0.77
12 Hmg1-rs14	51	52	0.010	0.92
13 DXMit3	56	47	0.786	0.38
14 Tpm3-rs9	49	54	0.243	0.61

Under the null hypothesis (Mendelian segregation) $p(A/A) = p(a/a) = 1/4$ and $p(A/a) = 1/2$,

$$\chi^2 = \frac{(n_1 - n/4)^2}{n/4} + \frac{(n_2 - n/2)^2}{n/2} + \frac{(n_3 - n/4)^2}{n/4} \sim \chi^2_2$$

3.2 Linkage analysis

3.2.1 Backcross population

	AB/AB		$\times$	AB/ab	
			$\downarrow$		
	AB/Ab	AB/aB		AB/AB	AB/ab
Frequency	$r/2$	$r/2$		$(1-r)/2$	$(1-r)/2$
Observed number	n_2	n_3		n_1	n_4
	Recombinant			Non-recombinant	
	$n_R = n_2 + n_3$			$n_{NR} = n_1 + n_4$	

Under the null hypothesis $r = 1/2$ (no linkage), the test statistic can be constructed as

$$\chi^2 = \frac{(n_{NR} - n_R)^2}{n} = \frac{(n_1 + n_4 - n_2 - n_3)^2}{n} \sim \chi^2_1$$

The estimate of recombination frequency is $\hat{r} = n_R/n$ with $n = n_1 + n_2 + n_3 + n_4$.

Table 3.2: Example of testing Mendelian segregation: Maize data

Marker	n_2	n_1	n_0	χ^2	P value
1	43	86	42	0.018	0.99
2	48	89	34	2.579	0.28
3	42	92	37	1.281	0.52
4	44	89	38	0.708	0.70
5	43	87	41	0.099	0.95
6	43	83	45	0.193	0.91
7	44	83	44	0.146	0.93
8	47	81	43	0.661	0.72
9	41	86	44	0.111	0.95
10	40	94	37	1.795	0.40
11	45	89	37	1.035	0.61
12	46	85	40	0.427	0.80

Table 3.3: Example of linkage analysis: Mouse data

Markers		n_{NR}	n_R	χ^2	r	cM(H)	cM(K)
1 Hmg1-rs13	2 DXMit57	96	7	76.903	0.068	7.3	6.8
2 DXMit57	3 Rps17-rs11	102	1	99.039	0.010	1.0	1.0
3 Rps17-rs11	4 Rps18-rs17	102	1	99.039	0.010	1.0	1.0
4 Rps18-rs17	5 DXMit48	100	3	91.350	0.029	3.0	2.9
5 DXMit48	6 DXNds1	100	3	91.350	0.029	3.0	2.9
6 DXNds1	7 DXMit109	98	5	83.971	0.049	5.1	4.9
7 DXMit109	8 Hmg14-rs6	99	4	87.621	0.039	4.0	3.9
8 Hmg14-rs6	9 DXMit60	102	1	99.039	0.010	1.0	1.0
9 DXMit60	10 DXMit16	101	2	95.155	0.019	2.0	1.9
10 DXMit16	11 DXMit97	101	2	95.155	0.019	2.0	1.9
11 DXMit97	12 Hmg1-rs14	100	3	91.350	0.029	3.0	2.9
12 Hmg1-rs14	13 DXMit3	96	7	76.903	0.068	7.3	6.8
13 DXMit3	14 Tpm3-rs9	92	11	63.699	0.107	12.0	10.8

cM(H) is estimated genetic distance (cM) using Haldane map function. cM(K) is estimated genetic distance (cM) using Kosambi map function.

Table 3.4: Estimated pairwise recombination frequency: Mouse data

	2	3	4	5	6	7	8	9	10	11	12	13	14
1	0.07	0.08	0.09	0.12	0.15	0.19	0.23	0.24	0.26	0.26	0.27	0.32	0.35
2		0.01	0.02	0.05	0.08	0.13	0.17	0.17	0.19	0.19	0.22	0.27	0.34
3			0.01	0.04	0.07	0.12	0.16	0.17	0.18	0.18	0.21	0.26	0.33
4				0.03	0.06	0.11	0.15	0.16	0.17	0.17	0.20	0.25	0.32
5					0.03	0.08	0.12	0.13	0.15	0.17	0.17	0.22	0.31
6						0.05	0.09	0.10	0.12	0.14	0.15	0.19	0.28
7							0.04	0.05	0.07	0.09	0.10	0.15	0.23
8								0.01	0.03	0.05	0.06	0.11	0.21
9									0.02	0.04	0.05	0.10	0.20
10										0.02	0.03	0.08	0.18
11											0.03	0.08	0.17
12												0.07	0.17
13													0.11

3.2.2 F2 population

A mating between AB/ab and AB/ab can produce ten genotypes, but only nine observable genetic classes (two double heterozygotes are generally not distinguishable) with the following expected frequencies:

		Frequency			
Genetic class	Code	$H_0 : r = 1/2$	H_1: $r < 1/2$	Rec. Event	Obs. Number
$\frac{AB}{AB}$	2 2	$\frac{1}{16}$	$\frac{1}{4}(1-r)^2$	0	n_1
$\frac{AB}{Ab}$	2 1	$\frac{2}{16}$	$\frac{1}{2}r(1-r)$	1	n_2
$\frac{Ab}{Ab}$	2 0	$\frac{1}{16}$	$\frac{1}{4}r^2$	2	n_3
$\frac{AB}{aB}$	1 2	$\frac{2}{16}$	$\frac{1}{2}r(1-r)$	1	n_4
$\left\{ \begin{array}{l} \frac{AB}{ab} \\ \frac{Ab}{aB} \end{array} \right.$	1 1	$\frac{4}{16}$	$\frac{1}{2}[(1-r)^2 + r^2]$	$\left\{ \begin{array}{ll} 0 & 1-q \\ 2 & q \end{array} \right.$	n_5
$\frac{Ab}{ab}$	1 0	$\frac{2}{16}$	$\frac{1}{2}r(1-r)$	1	n_6
$\frac{aB}{aB}$	0 2	$\frac{1}{16}$	$\frac{1}{4}r^2$	2	n_7
$\frac{aB}{ab}$	0 1	$\frac{2}{16}$	$\frac{1}{2}r(1-r)$	1	n_8
$\frac{ab}{ab}$	0 0	$\frac{1}{16}$	$\frac{1}{4}(1-r)^2$	0	n_9

with

$$q = \frac{r^2}{(1-r)^2 + r^2} \tag{3.1}$$

In this case the analysis is a little more complicated largely because of genetic class 5. In estimating recombination frequency, we can utilize the above classification of recombination

Table 3.5: Example of linkage analysis: Maize data

Markers		n_{0R}	n_{1R}	n_{2R}	n_5	LOD	r	cM(H)	cM(K)
1	2	39	77	6	49	6.08	0.31	47.7	35.8
2	3	62	37	0	72	30.87	0.12	13.1	11.7
3	4	63	33	1	74	32.08	0.11	12.3	11.0
4	5	79	4	2	86	60.60	0.02	2.4	2.4
5	6	54	58	3	56	15.92	0.21	27.0	22.2
6	7	75	22	2	72	41.55	0.08	8.6	8.0
7	8	68	40	1	62	29.52	0.13	15.2	13.4
8	9	78	19	0	74	48.92	0.06	6.1	5.7
9	10	50	58	2	61	15.73	0.20	26.1	21.5
10	11	70	19	0	82	46.53	0.06	6.1	5.8
11	12	70	26	1	74	38.86	0.09	9.4	8.6

$n_{0R} = n_1 + n_9$, $n_{1R} = n_2 + n_4 + n_6 + n_8$, $n_{2R} = n_3 + n_7$

events and estimate r as

$$\hat{r} = \frac{1}{2n}[(n_2 + n_4 + n_6 + n_8) + 2(n_3 + n_7 + qn_5)] \tag{3.2}$$

as the genetic classes 2, 4, 6 and 8 contain one recombination event, and 3, 7 and 5 (with probability q) contain two recombination events. However, as q is a function of r, the analysis has to be performed in a loop and updated between (3.1) and (3.2). This is the so-called EM algorithm with (3.1) being the E-step (Expectation step) and (3.2) being the M-step (Maximization step).

The statistical test for linkage can be performed by LOD score (a likelihood ratio test statistic):

$$\text{LOD} = \log_{10} \frac{L(r)}{L(r = 1/2)}$$

with

$$L(r) \propto [\frac{1}{4}(1-r)^2]^{n_1+n_9}[\frac{1}{2}r(1-r)]^{n_2+n_4+n_6+n_8}[\frac{1}{4}r^2]^{n_3+n_7}[\frac{1}{2}(1-r)^2 + \frac{1}{2}r^2]^{n_5}$$

and

$$L(r = 1/2) \propto [\frac{1}{16}]^{n_1+n_3+n_7+n_9}[\frac{2}{16}]^{n_2+n_4+n_6+n_8}[\frac{4}{16}]^{n_5}$$

3.2.3 Three point analysis

- Consider three loci ABC (in no particular order) in a triple backcross: $ABC/abc \times abc/abc$. Let n_{ij} be the number of recombinants $(i, j = 1)$ or nonrecombinants $(i, j = 0)$ between A and B and between B and C and g_{ij} be the corresponding joint recombination fraction.

Table 3.6: Estimated pairwise recombination frequency: Maize data

	2	3	4	5	6	7	8	9	10	11	12
1	0.31	0.39	0.49	0.50	0.46	0.46	0.45	0.45	0.49	0.50	0.51
2		0.12	0.22	0.22	0.31	0.33	0.40	0.39	0.40	0.39	0.42
3			0.11	0.12	0.28	0.32	0.44	0.44	0.43	0.42	0.43
4				0.02	0.22	0.26	0.39	0.39	0.41	0.41	0.42
5					0.21	0.27	0.39	0.40	0.41	0.41	0.43
6						0.08	0.21	0.22	0.34	0.37	0.40
7							0.13	0.14	0.31	0.34	0.39
8								0.06	0.26	0.30	0.35
9									0.20	0.25	0.31
10										0.06	0.13
11											0.09

	Loci B and C		
Loci A and B	R	NR	Total
R	g_{11}	g_{10}	r_{AB}
NR	g_{01}	g_{00}	$1 - r_{AB}$
Total	r_{BC}	$1 - r_{BC}$	1

- Maximum likelihood estimate of g_{ij} is $\hat{g}_{ij} = n_{ij}/n$. The pairwise recombination frequency can be estimated from $\hat{g}_{ij}$ as

$$\begin{aligned} r_{AB} &= g_{11} + g_{10} \\ r_{BC} &= g_{11} + g_{01} \\ r_{AC} &= g_{01} + g_{10} \end{aligned} \qquad \Longrightarrow \qquad \begin{aligned} \hat{r}_{AB} &= \hat{g}_{11} + \hat{g}_{10} \\ \hat{r}_{BC} &= \hat{g}_{11} + \hat{g}_{01} \\ \hat{r}_{AC} &= \hat{g}_{01} + \hat{g}_{10} \end{aligned}$$

  A relation:

$$\begin{aligned} 0 \le g_{11} &= \frac{1}{2}(r_{AB} + r_{BC} - r_{AC}) \Longrightarrow r_{AC} \le r_{AB} + r_{BC} \\ 0 \le g_{10} &= \frac{1}{2}(r_{AB} - r_{BC} + r_{AC}) \Longrightarrow r_{BC} \le r_{AB} + r_{AC} \\ 0 \le g_{01} &= \frac{1}{2}(-r_{AB} + r_{BC} + r_{AC}) \Longrightarrow r_{AB} \le r_{BC} + r_{AC} \\ g_{00} &= 1 - g_{11} - g_{10} - g_{01} = 1 - \frac{1}{2}(r_{AB} + r_{BC} + r_{AC}) \end{aligned}$$

- In several respects, three-point (and generally multipoint) analysis yields more information than does two-point analysis.

Three-point (and generally multipoint) analysis works directly with $\hat{g}_{ij}$, whereas two-point analysis translates $\hat{g}_{ij}$ into $\hat{r}$'s and works with $\hat{r}$.

"Three-way data provide very much more information, particularly with regard to the problem of ordering the loci" (E. Thompson, 1984), where individuals informative for all other loci are observed. Information is lost (or overestimated) by summarizing the data in pairwise.

3.2.4 Multilocus likelihood

Consider three loci ABC (in no particular order) in a triple backcross: $ABC/abc \times abc/abc$. Assume there is no crossover interference. Let

Genotype	$\frac{ABC}{abc}$ or $\frac{abc}{abc}$	$\frac{ABc}{abc}$ or $\frac{abC}{abc}$	$\frac{Abc}{abc}$ or $\frac{aBC}{abc}$	$\frac{AbC}{abc}$ or $\frac{aBc}{abc}$
Observed number	n_1	n_2	n_3	n_4
Frequency under order ABC	$(1-r_1)(1-r_2)$	$(1-r_1)r_2$	$r_1(1-r_2)$	r_1r_2
Frequency under order ACB	$(1-r_3)(1-r_2)$	r_3r_2	$r_3(1-r_2)$	$(1-r_3)r_2$
Frequency under order BAC	$(1-r_1)(1-r_3)$	$(1-r_1)r_3$	r_1r_3	$r_1(1-r_3)$

$$\begin{array}{lll} r_1 = \text{recombination frequency between } A \text{ and } B & & \hat{r}_1 = (n_3+n_4)/n \\ r_2 = \text{recombination frequency between } B \text{ and } C & \text{with} & \hat{r}_2 = (n_2+n_4)/n \\ r_3 = \text{recombination frequency between } A \text{ and } C & & \hat{r}_3 = (n_2+n_3)/n \end{array}$$

Likelihoods for the three linkage orders:

$$\begin{aligned}
L_{ABC} &\propto (1-r_1)^{n_1+n_2}(1-r_2)^{n_1+n_3}(r_1)^{n_3+n_4}(r_2)^{n_2+n_4} \\
&\longrightarrow \left(1-\frac{n_3+n_4}{n}\right)^{n_1+n_2}\left(1-\frac{n_2+n_4}{n}\right)^{n_1+n_3}\left(\frac{n_3+n_4}{n}\right)^{n_3+n_4}\left(\frac{n_2+n_4}{n}\right)^{n_2+n_4} \\
L_{ACB} &\propto (1-r_3)^{n_1+n_4}(1-r_2)^{n_1+n_3}(r_3)^{n_2+n_3}(r_2)^{n_2+n_4} \\
&\longrightarrow \left(1-\frac{n_2+n_3}{n}\right)^{n_1+n_4}\left(1-\frac{n_2+n_4}{n}\right)^{n_1+n_3}\left(\frac{n_2+n_3}{n}\right)^{n_2+n_3}\left(\frac{n_2+n_4}{n}\right)^{n_2+n_4} \\
L_{BAC} &\propto (1-r_1)^{n_1+n_2}(1-r_3)^{n_1+n_4}(r_1)^{n_3+n_4}(r_3)^{n_2+n_3} \\
&\longrightarrow \left(1-\frac{n_3+n_4}{n}\right)^{n_1+n_2}\left(1-\frac{n_2+n_3}{n}\right)^{n_1+n_4}\left(\frac{n_3+n_4}{n}\right)^{n_3+n_4}\left(\frac{n_2+n_3}{n}\right)^{n_2+n_3}
\end{aligned}$$

Under the maximum likelihood principle, the linkage order that gives the maximum likelihood for a data set is the best linkage order supported by the data.

The same principle and procedure can be applied to multiple markers for searching for the best supported linkage order. However, as the number of markers increases, the number of possible linkage orders increases very dramatically. A number of numerical algorithms have been devised to reduce the number of linkage orders evaluated during the search for the best supported linkage order.

By the maximum likelihood property, this method is *asymptotically* most powerful method in ordering markers. Although the above analysis assumes independent crossover event in different marker intervals (*i.e.* there is no crossover interference), Speed, McPeek and Evans (1992) showed that the ordering of the loci that maximizes the likelihood under the assumption of no interference is, in fact, a consistent estimator of the true order even when there is interference. This means that for sufficiently large sample size with probability 1, the linkage order that gives the maximum likelihood will be the true order, regardless whether there is crossover interference. However, with small sample size, there is no guarantee that the linkage order with the maximum likelihood will be the true order due to statistical sampling.

3.2.5 Ordering markers

Linkage map analysis includes grouping markers into different linkage groups (by testing linkage) and ordering markers in the same linkage group.

The problem of ordering a set of markers is equivalent to the traveling salesman problem. A number of numerical algorithms have been devised to reduce the number of linkage orders evaluated during the search for the best supported linkage order by guiding the search process toward to certain directions.

1. Branch and Bound (Thompson 1984)
2. Simulated Annealing (Weeks and Lange 1987)
3. Seriation (Buetow and Chakravarti 1987a,b)
4. Rapid Chain Delineation (Doerge 1993)

Method 1 and 2 are two commonly used strategies in the numerical optimization. Method 3 and 4 are based on a set of rules to guild the search process. There is no guarantee that the best supported linkage order can be found by these algorithms, short of the exhaustive search.

Objective functions that have been used to evaluate linkage orders include:

- Multilocus likelihood (discussed above): computationally more intensive, a more appropriate measure.
- Sum of Adjacent Recombination Fractions (SAR): easy to compute, can be less reliable.

3.3 Map Functions

Map function is a function that translates recombination frequency into a measure which measures the genetic distance between loci. One property of this genetic distance should possess is the additivity.

3.3.1 Haldane map function

- Let r_1 and r_2 be the recombination frequency between AB and BC. Under the assumption of no interference, *i.e.*, independence of crossover, we would expect a triple heterozygous organism to produce gametes in the following proportions (given ABC order):

Event	Gametes	Frequency
No crossover	ABC or abc	$(1-r_1)(1-r_2)$
Crossover between A & B	aBC or Abc	$r_1(1-r_2)$
Crossover between B & C	ABc or abC	$(1-r_1)r_2$
Crossover between A & B and B & C	AbC or aBc	r_1r_2

  The recombination frequency between A and C are expected to be

$$r_{12} = r_1(1-r_2) + (1-r_1)r_2 = r_1 + r_2 - 2r_1r_2$$
$$\Longrightarrow (1-2r_{12}) = (1-2r_1)(1-2r_2) \tag{3.3}$$

- One map function is Haldane's map function which is based on the assumption of no interference and Poisson distribution of crossing-over event.

  Map distance: A genetic length (map distance) x of a chromosome is defined as the mean number of crossovers.

  Poisson distribution (x = genetic length):

Crossover event	0	1	2	3	$\cdots$	t	$\cdots$
Probability	e^{-x}	xe^{-x}	$\frac{x^2}{2!}e^{-x}$	$\frac{x^3}{3!}e^{-x}$	$\cdots$	$\frac{x^t}{t!}e^{-x}$	$\cdots$

  Hence the value of r (recombination frequency) for a genetic length of x is the sum of the probabilities of all odd numbers of crossovers.

$$r = e^{-x}\left(\frac{x}{1!} + \frac{x^3}{3!} + \frac{x^5}{5!} + \frac{x^7}{7!} + \cdots\right) = \frac{1}{2}(1-e^{-2x}) \quad \Longrightarrow \quad x = -\frac{1}{2}\ln(1-2r)$$

- The unit of genetic distance x in a chromosome is measured by Morgan (M) after geneticist T. H. Morgan. One Morgan length of a chromosome has an expected number of crossover 1. One centiMorgan (cM) is one hundredth of a Morgan.

- It can be checked that, given $r_1 = \frac{1}{2}(1-e^{-2x_1})$ and $r_2 = \frac{1}{2}(1-e^{-2x_2})$,

$$\begin{aligned} r_{12} &= r_1 + r_2 - 2r_1r_2 \\ &= \frac{1}{2}(1-e^{-2x_1}) + \frac{1}{2}(1-e^{-2x_2}) - 2\frac{1}{2}(1-e^{-2x_1})\frac{1}{2}(1-e^{-2x_2}) \\ &= \frac{1}{2}\left[1 - e^{-2x_1} + 1 - e^{-2x_2} - 1 + e^{-2x_1} + e^{-2x_2} - e^{-2x_1}e^{-2x_2}\right] \end{aligned}$$

$$= \frac{1}{2}(1 - e^{-2(x_1+x_2)})$$
$$= \frac{1}{2}(1 - e^{-2x_{12}}) \quad \text{where } x_{12} = x_1 + x_2$$

- Although in the above derivation and in Haldane (1919) as well, Poisson distribution of crossing-over event is assumed, this assumption is actually not necessary. Haldane's mapping function requires only the assumption of no interference. If we take a logarithm in (3.3), we have

$$\ln(1 - 2r_{12}) = \ln(1 - 2r_1) + \ln(1 - 2r_2)$$
$$\Longrightarrow -\frac{1}{2}\ln(1 - 2r_{12}) = -\frac{1}{2}\ln(1 - 2r_1) - \frac{1}{2}\ln(1 - 2r_2)$$
$$\Longrightarrow x_{12} = x_1 + x_2$$

The map function is established.

3.3.2 Kosambi map function

- Kosambi map function is an extension of Haldane map function. It is based on the empirical observation that

when	r_1 and r_2 are small	$r_{12} \approx r_1 + r_2$
	r_1 and r_2 are median	$r_{12} \approx r_1 + r_2 - r_1 r_2$
	r_1 and r_2 are large	$r_{12} \approx r_1 + r_2 - 2r_1 r_2$

We want to seek a map function $r = f(x)$ that reflects closely the above observation. Let

$$f(x + h) = f(x) + f(h) - pf(x)f(h)$$

We hope to choose a p to satisfy the condition that when x is small, $p \to 0$ so that $f(x)/x \to 1$ and when x is large, $p \to 2$.

Let

$$\frac{f(x+h)}{h} - \frac{f(x)}{h} = \frac{f(h)}{h} - pf(x)\frac{f(h)}{h}$$
$$\Longrightarrow f'(x) = 1 - pf(x) \quad \text{as} \quad h \to 0$$

Thus

$$\frac{dr}{dx} = 1 - pr$$

If we take

$$p = 4r \to \begin{cases} 0 & \text{as } r \to 0 \\ 2 & \text{as } r \to \frac{1}{2} \end{cases} \tag{3.4}$$

Then we have

$$\frac{dr}{dx} = 1 - 4r^2$$

By solving this differential function, we obtain Kosambi map function

Kosambi Map Function	Haldane Map Function
$x = \frac{1}{4} \ln \frac{1+2r}{1-2r}$	$x = -\frac{1}{2} \ln(1-2r)$
$r = \frac{1}{2} \frac{e^{2x}-e^{-2x}}{e^{2x}+e^{-2x}}$	$r = \frac{1}{2}(1-e^{2x})$
$r_{12} = \frac{r_1+r_2}{1+4r_1r_2}$	$r_{12} = r_1 + r_2 - 2r_1r_2$

- Note that equation (3.4) is rather arbitrary. By choosing a different equation, one may obtain a different map function.

- Multilocus feasibility of map function (Liberman and Karlin, 1984, TPB 25:331-346): In order for $r = f(x)$ to provide a valid map function in a multilocus structure, it is *necessary* that its derivative functions $f^{(n)}$ obey the inequalities

$$(-1)^n f^{(n)}(0) \leq 0, \quad n = 1, 2, \cdots$$

  That is the successive derivatives of $f(x)$ evaluated at $x = 0$ alternative in sign.

- By this criterion, Kosambi map function is actually not a valid multilocus map function, whereas Haldane map function is a valid multilocus map function. The sufficient condition for a multilocus map function is not known.

- Kosambi map function gives a measure of genetic distance that generally fits to data better than Haldane map function does, because it takes interference into account in the approximation. Because of that, Kosambi map function is quite popular in literature in translating r to x.

Chapter 4

Quantitative Genetic Models

4.1 Least squares genetic model

- The basic quantitative genetic model is based on the least-square model that partitions the total variance into different components. In the simplest setting,

$$P = G + E + G \times E$$

where P is phenotypic effect, G is genetic effect and E is environmental effect, $G \times E$ is genotype by environment interaction effect with

$$Var(P) = Var(G) + 2COV(G, E) + Var(E) \quad \text{ignoring } G \times E$$

$$Var(P) = Var(G) + Var(E) \quad \text{ignoring } G \times E \text{ and } Cov(G, E)$$

- Partition of genetic variance component for a locus.

Genotype	Genotypic value (α)	Frequency (f)
BB	$G_{11} = m + a$	$p^2 = p_1^2$
Bb	$G_{12} = G_{21} = m + d$	$2pq = 2p_1p_2$
bb	$G_{22} = m - a$	$q^2 = p_2^2$

To simplify the analysis, we set $m = 0$. Thus

$$\mu = \sum_i f_i\alpha_i = \sum_{i,j} p_ip_jG_{ij} = (p_1 - p_2)a + 2p_1p_2d$$

$$\begin{aligned}
\sigma_G^2 &= \sum_i f_i\alpha_i^2 - (\sum_i f_i\alpha_i)^2 = \sum_{i,j} p_ip_jG_{ij}^2 - \mu^2 \\
&= p_1^2a^2 + 2p_1p_2d^2 + p_2^2(-a)^2 - [(p_1 - p_2)a + 2p_1p_2d]^2 \\
&= 2p_1p_2[a + (p_2 - p_1)d]^2 + 4p_1^2p_2^2d^2 \\
&= \sigma_a^2 + \sigma_d^2
\end{aligned}$$

where σ_a^2 is the additive variance component and σ_d^2 is the dominant variance component as shown below.

- Least square partition of genetic variance component.

  Model: $G_{ij} = \mu + a_i + a_j + d_{ij}$

  Define:

$$\begin{aligned} G_{..} &= \sum_{i,j} p_i p_j G_{ij} = \mu \\ a_i &= \sum_j p_j G_{ij} - \mu = G_{i.} - G_{..} \\ a_j &= G_{.j} - G_{..} \\ d_{ij} &= G_{ij} - a_i - a_j - \mu = G_{ij} - G_{i.} - G_{.j} + G_{..} \end{aligned}$$

  which satisfy the conditions

$$\sum_i p_i a_i = 0, \quad \sum_{i,j} p_i p_j d_{ij} = 0, \quad \text{and} \quad \sum_i p_i d_{ij} = \sum_j p_j d_{ij} = 0$$

  The genetic variance is defined as

$$\begin{aligned} \sigma_G^2 &= \sum_{i,j} p_i p_j (G_{ij} - \mu)^2 \\ &= \sum_{i,j} p_i p_j (a_i + a_j + d_{ij})^2 \\ &= 2\sum_i p_i a_i^2 + \sum_i p_i p_j d_{ij}^2 \\ &= \sigma_a^2 + \sigma_d^2 \end{aligned}$$

  For the two allele model:

$$\begin{aligned} \mu &= p_1^2 G_{11} + 2p_1 p_2 G_{12} + p_2^2 G_{22} \\ a_1 &= p_2[p_1(G_{11} - G_{12}) + p_2(G_{12} - G_{22})] \\ a_2 &= -p_1[p_1(G_{11} - G_{12}) + p_2(G_{12} - G_{22})] \end{aligned}$$

  Thus

$$\begin{aligned} \sigma_a^2 &= 2\sum_i p_i a_i^2 = 2[p_1 a_1^2 + p_2 a_2^2] \\ &= 2p_1 p_2 [p_1(G_{11} - G_{12}) + p_2(G_{12} - G_{22})]^2 \\ &= 2p_1 p_2 [p_1(a - d) + p_2(a + d)]^2 \\ &= 2p_1 p_2 [a + (p_2 - p_1)d]^2 \end{aligned}$$

Similarly,

$$\begin{aligned}
d_{11} &= G_{11} - 2p_2[p_1(G_{11} - G_{12}) + p_2(G_{12} - G_{22})] - [p_1^2 G_{11} + 2p_1p_2G_{12} + p_2^2 G_{22}] \\
&= -p_2^2[2G_{12} - G_{11} - G_{22}] \\
d_{12} &= d_{21} = G_{12} - (p_2 - p_1)[p_1(G_{11} - G_{12}) + p_2(G_{12} - G_{22})] - [p_1^2 G_{11} + 2p_1p_2G_{12} + p_2^2 G_{22}] \\
&= p_1p_2[2G_{12} - G_{11} - G_{22}] \\
d_{22} &= G_{22} + 2p_1[p_1(G_{11} - G_{12}) + p_2(G_{12} - G_{22})] - [p_1^2 G_{11} + 2p_1p_2G_{12} + p_2^2 G_{22}] \\
&= -p_1^2[2G_{12} - G_{11} - G_{22}]
\end{aligned}$$

Thus

$$\begin{aligned}
\sigma_d^2 &= \sum_{i,j} p_i p_j d_{ij}^2 = p_1^2 d_{11}^2 + 2p_1p_2d_{12}^2 + p_2^2 d_{22}^2 \\
&= p_1^2 p_2^2 [2G_{12} - G_{11} - G_{22}]^2 \\
&= 4p_1^2 p_2^2 d^2
\end{aligned}$$

4.2 Hardy-Weinberg disequilibrium

For a locus with two alleles:

Genotype	AA	Aa	aa
Frequency	P_{AA}	P_{Aa}	P_{aa}
H.-W. Equi.	p_A^2	$2p_Ap_a$	p_a^2

with $p_A = P_{AA} + \frac{1}{2}P_{Aa}$ and $p_a = P_{aa} + \frac{1}{2}P_{Aa}$.

Define disequilibrium coefficients as:

$$\begin{aligned}
P_{AA} &= p_A^2 + D_{AA} \\
P_{Aa} &= 2p_Ap_a - 2D_{Aa} \\
P_{aa} &= p_a^2 + D_{aa}
\end{aligned}$$

However, because of bounds of gene and genotype frequencies

$$P_{AA} + P_{Aa} + P_{aa} = 1 \Longrightarrow D_{AA} + D_{Aa} + D_{aa} = 0$$

$$p_A^2 + D_{AA} + p_Ap_a - D_{Aa} = p_A \Longrightarrow D_{AA} = D_{Aa} \Longrightarrow D_{AA} = D_{Aa} = D_{aa},$$

$$\begin{aligned}
P_{AA} &= p_A^2 + D_A \\
P_{Aa} &= 2p_Ap_a - 2D_A \\
P_{aa} &= p_a^2 + D_A
\end{aligned}$$

where D_A measures Hardy-Weinberg disequilibrium for the locus.

4.3 Linkage disequilibrium

- Let us consider the distribution of gametes for a two-allele and two-locus model:

Gamete	AB	Ab	aB	ab
Frequency	P_{AB}	P_{Ab}	P_{aB}	P_{ab}
L.-E.	$p_A p_B$	$p_A p_b$	$p_a p_B$	$p_a p_b$

  Define disequilibrium coefficients as:

$$\begin{aligned} P_{AB} &= p_A p_B + D_{AB} \\ P_{Ab} &= p_A p_b + D_{Ab} \\ P_{aB} &= p_a p_B + D_{aB} \\ P_{ab} &= p_a p_b + D_{ab} \end{aligned}$$

  Also, because of bounds of gene and genotype frequencies

$$D_{AB} + D_{Ab} + D_{aB} + D_{ab} = 0 \text{ and } p_A = P_{AB} + P_{Ab} = p_A + D_{AB} + D_{Ab}$$

$$\Longrightarrow D_{AB} = -D_{Ab} \text{ and } D_{ab} = -D_{aB},$$

  In general,

$$\sum_i D_{ij} = \sum_j D_{ij} = 0.$$

  Thus

$$\begin{aligned} P_{AB} &= p_A p_B + D_{AB} \\ P_{Ab} &= p_A p_b - D_{AB} \\ P_{aB} &= p_a p_B - D_{AB} \\ P_{ab} &= p_a p_b + D_{AB} \end{aligned}$$

- Correlation coefficient between two loci:

$$\gamma = \frac{D_{AB}}{\sqrt{p_A p_B p_a p_b}}$$

- Relationship between linkage disequilibrium and recombination frequency: For both backcross and F_2 populations, segregating gametes are produced from F_1 so the linkage disequilibrium in these populations are determined by the gametes of F_1. F_1 produces four gametes in the following proportion:

Gamete	AB	Ab	aB	ab
Frequency	$\frac{1-r}{2}$	$\frac{r}{2}$	$\frac{r}{2}$	$\frac{1-r}{2}$

r is the recombination frequency ($0 < r < 0.5$).

As $p_A = p_B = p_a = p_b = \frac{1}{2}$, $D_{AB} = P_{AB} - p_A p_B = \frac{1-r}{2} - \frac{1}{4} = \frac{1}{4}(1-2r)$,

$$\gamma = \frac{D_{AB}}{\sqrt{p_A p_B p_a p_b}} = \frac{(1-2r)/4}{\sqrt{1/16}} = 1 - 2r$$

4.4 Linkage disequilibrium and partition of genetic variance

Consider two loci, ignoring epistasis for the moment. The least square genetic model is

$$G_{jl}^{ik} = \mu + a^i + a^k + a_j + a_l + d_j^i + d_l^k$$

Let P_{jl}^{ik} be the frequency of genotype G_{jl}^{ik} in a population. Assuming Hardy-Weinberg equilibrium gives $P_{jl}^{ik} = P_{ik}P_{jl}$. Define $P_{ik} = p_i p_k + D_{ik}$. The mean effect of the population is

$$\begin{aligned}
\bar{G}_{\cdot\cdot}^{\cdot\cdot} &= \sum_{i,k}\sum_{j,l} P_{jl}^{ik} G_{jl}^{ik} = \sum_{i,k}\sum_{j,l} P_{ik}P_{jl}G_{jl}^{ik} \\
&= \sum_{i,k}\sum_{j,l} (p_i p_k + D_{ik})(p_j p_l + D_{jl})(\mu + a^i + a^k + a_j + a_l + d_j^i + d_l^k) \\
&= \mu = G_{\cdot\cdot}^{\cdot\cdot} = \sum_{i,k}\sum_{j,l} p_i p_k p_j p_l G_{jl}^{ik}
\end{aligned}$$

as $\sum_i p_i a_i = 0$, $\sum_i p_i d_j^i = \sum_j p_j d_j^i = 0$, $\sum_i D_{ik} = \sum_k D_{ik} = 0$. So linkage disequilibrium does not affect the mean of the trait. (However, if there is epistasis such as when $(aa)_{ik} \neq 0$, the mean will be affected as $\sum_{i,k}\sum_{j,l} D_{ik}p_j p_l (aa)_{ik} = \sum_{i,k} D_{ik}(aa)_{i,k} \neq 0$.)

The variance is

$$\begin{aligned}
\sigma_G^2 &= \sum_{i,k}\sum_{j,l} P_{ik}P_{jl}(G_{jl}^{ik} - \mu)^2 \\
&= \sum_{i,k}\sum_{j,l} (p_i p_k + D_{ik})(p_j p_l + D_{jl})(a^i + a^k + a_j + a_l + d_j^i + d_l^k)^2 \\
&= \sum_i p_i a^{i2} + \sum_k p_k a^{k2} + \sum_j p_j a_j^2 + \sum_l p_l a_l^2 + 2\sum_{i,k} D_{ik} a^i a^k + 2\sum_{j,l} D_{jl} a_j a_l \\
&\quad + \sum_{i,j} p_i p_j d_j^{i2} + \sum_{k,l} p_k p_l d_l^{k2} + 2\sum_{i,k}\sum_{j,l} D_{ik} D_{jl} d_j^i d_l^k
\end{aligned}$$

The variance contains covariances between additive effects and between dominance effects due to linkage disequilibrium. There is, however, no covariance between a^i and d_j^i. This is because of the assumption of Hardy-Weinberg equilibrium. With Hardy-Weinberg disequilibrium (*e.g.* under inbreeding), there will be some covariance between a^i and d_j^i.

4.5 Genetic model for backcross and F_2 populations

- Let the effects of three genotypes at a locus B_i be defined as

Genotype	B_iB_i	B_ib_i	b_ib_i
Effect	a_i	d_i	$-a_i$

- Assume P_1 and P_2 populations fixed with allele B_i and b_i, respectively, for every one of m loci. Ignoring epistasis, phenotypic value of an individual k in P_1 population can be defined as

$$y_k = \mu + \sum_{i=1}^{m} a_i + e_k$$

where $e_k \sim N(0, \sigma_e^2)$. This gives the mean and variance of the trait in P_1 population as

$$\mu_{P_1} = \mu + \sum_{i=1}^{m} a_i, \quad \sigma_{P_1}^2 = \sigma_e^2$$

- Similarly for P_2 population

$$y_k = \mu - \sum_{i=1}^{m} a_i + e_k, \quad \mu_{P_2} = \mu - \sum_{i=1}^{m} a_i, \quad \sigma_{P_2}^2 = \sigma_e^2$$

- For F_1 population

$$y_k = \mu + \sum_{i=1}^{m} d_i + e_k, \quad \mu_{F_1} = \mu + \sum_{i=1}^{m} d_i, \quad \sigma_{F_1}^2 = \sigma_e^2$$

- For B_1 which is a cross between P_1 and F_1, the trait values of individuals can be defined as

$$y_k = \mu + \sum_{i=1}^{m} [x_{ik} a_i + (1 - x_{ik}) d_i] + e_k = \mu + \sum_{i=1}^{m} d_i + \sum_{i=1}^{m} x_{ik}(a_i - d_i) + e_k$$

$$x_{ik} = \begin{cases} 1 & \text{if individual } k \text{ receives a } B_i \text{ allele in the gamete from } F_1 \\ 0 & \text{if individual } k \text{ receives a } b_i \text{ allele in the gamete from } F_1 \end{cases}$$

$$\mathcal{E}(x_{ik}) = \frac{1}{2}, \quad Var(x_{ik}) = \frac{1}{4}, \quad Cov(x_{ik}, x_{jk}) = \frac{1}{4}(1 - 2r_{ij})$$

Thus the mean and variance of the trait in B_1 population are

$$\mu_{B_1} = \mu + \frac{1}{2} \sum_{i=1}^{m} (a_i + d_i)$$

$$\sigma_{B_1}^2 = \frac{1}{4} \sum_{i=1}^{m} (a_i - d_i)^2 + \frac{1}{4} \sum_{i=1}^{m} \sum_{j=1, i \neq j}^{m} (1 - 2r_{ij})(a_i - d_i)(a_j - d_j) + \sigma_e^2$$

- Similarly, for B_2 $(= P_1 \times F_1)$

$$y_k = \mu + \sum_{i=1}^{m}[x_{ik}d_i - (1-x_{ik})a_i] + e_k = \mu - \sum_{i=1}^{m} a_i + \sum_{i=1}^{m} x_{ik}(a_i + d_i) + e_k$$

$$\mu_{B_2} = \mu + \frac{1}{2}\sum_{i=1}^{m}(-a_i + d_i)$$

$$\sigma^2_{B_2} = \frac{1}{4}\sum_{i=1}^{m}(a_i + d_i)^2 + \frac{1}{4}\sum_{i=1}^{m}\sum_{j=1,i\neq j}^{m}(1-2r_{ij})(a_i + d_i)(a_j + d_j) + \sigma^2_e$$

- For F_2 population, let

$$\begin{aligned} y_k &= \mu + \sum_{i=1}^{m}[\xi_{ik1}\xi_{ik2}a_i + [\xi_{ik1}(1-\xi_{ik2}) + (1-\xi_{ik1})\xi_{ik2}]d_i \\ &\quad +(1-\xi_{ik1})(1-\xi_{ik2})(-a_i)] + e_k \\ &= \mu - \sum_{i=1}^{m} a_i + \sum_{i=1}^{m}[(\xi_{ik1} + \xi_{ik2})(a_i + d_i) - 2\xi_{ik1}\xi_{ik2}d_i] + e_k \end{aligned}$$

$$\xi_{ik1} = \begin{cases} 1 & \text{if individual } k \text{ receives a } B_i \text{ allele in a gamete from } F_1 \\ 0 & \text{if individual } k \text{ receives a } b_i \text{ allele in a gamete from } F_1 \end{cases}$$

$$\mathcal{E}(\xi_{ik1}) = \frac{1}{2}, \quad Var(\xi_{ik1}) = \frac{1}{4}, \quad Cov(\xi_{ik1}, \xi_{jk1}) = \frac{1}{4}(1-2r_{ij})$$

$$\xi_{ik2} = \begin{cases} 1 & \text{if individual } k \text{ receives a } B_i \text{ allele in the other gamete from } F_1 \\ 0 & \text{if individual } k \text{ receives a } b_i \text{ allele in the other gamete from } F_1 \end{cases}$$

$$\mathcal{E}(\xi_{ik2}) = \frac{1}{2}, \quad Var(\xi_{ik2}) = \frac{1}{4}, \quad Cov(\xi_{ik2}, \xi_{jk2}) = \frac{1}{4}(1-2r_{ij})$$

Under the assumption of Hardy-Weinberg equilibrium, ξ_{ik1} and ξ_{ik2} are independent. With these conditions, the mean and variance of individuals in F_2 population become

$$\mu_{F_2} = \mu + \frac{1}{2}\sum_{i=1}^{m} d_i$$

$$\sigma^2_{F_2} = \frac{1}{2}\sum_{i=1}^{m} a_i^2 + \frac{1}{4}\sum_{i=1}^{m} d_i^2 + \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1,i\neq j}^{m}(1-2r_{ij})a_i a_j + \frac{1}{4}\sum_{i=1}^{m}\sum_{j=1,i\neq j}^{m}(1-2r_{ij})^2 d_i d_j + \sigma^2_e.$$

The first term on the right side of above equation is the additive variance within loci, the second is the dominance variance within loci, the third is the additive covariance between loci and the fourth is the dominance covariance between loci.

4.6 Modelling epistasis

Fisher (1918) first partitioned the genetic variance into components corresponding to additive, dominance, and epistatic variances using the least squares principle. Cockerham (1954) further partitioned the two-gene epistatic variance into four variance components which are additive×additive, additive×dominance, dominance×additive, and dominance×dominance.

4.6.1 Gene effects and variances with epistasis under the least square genetic model (adapted from lecture note of C. Clark Cockerham)

Let G_{jl}^{ik} be the genotypic value for the genotype formed by the union of male gamete A_iB_k with the female gamete A_jB_l, and $G_{jl}^{ik} = G_{ik}^{jl}$. The effects of the model below for these two loci are classified in two ways. By columns they are grouped according to locus effect, g_j^i and g_l^k, and loci interaction effect $(g_j^i g_l^k)$. By rows they are grouped according to gametic effects, g^{ik} and g_{jl}, and gametic interaction effect $(g^{ik}g_{jl})$.

$$\begin{array}{rcccl} G_{jl}^{ik} = \mu & & +g_j^i & +g_l^k & +(g_j^i g_l^k) \\ & & = & = & = \\ +g^{ik} & = & +a^i & +a^k & +(a^i a^k) \\ +g_{jl} & = & +a_j & +a_l & +(a_j a_l) \\ +(g^{ik}g_{jl}) & = & +d_j^i & +d_l^k & +(a^i a_l)+(a_j a^k)+(a^i d_l^k)+(d_j^i a^k) \\ & & & & +(a_j d_l^k)+(d_j^i a_l)+(d_j^i d_l^k) \end{array}$$

These methods of classifying the effects illustrate the connections between gametic effects and loci effects. All epistatic effects, (aa)'s = additive by additive, (ad)'s = additive by dominance, and (dd) = dominance by dominance, represent interactions among nonallelic genes, and thus constitute interactions among loci. Each of the gametic effects contains in addition to average effects, a's, an additive by additive effect. Interactions among the gametes include dominance effects and the remaining epistatic effects.

With random mating and linkage equilibrium, and allowing for different frequencies in male and female gametes, the effects in terms of the means, corresponding to a four factor factorial design, are

$$a^i = G_{..}^{i\cdot} - G_{..}^{\cdot\cdot}, \quad a^k = G_{..}^{\cdot k} - G_{..}^{\cdot\cdot}, \quad a_j = G_{j.}^{\cdot\cdot} - G_{..}^{\cdot\cdot}, \quad a_l = G_{.l}^{\cdot\cdot} - G_{..}^{\cdot\cdot}$$

$$d_j^i = G_{j.}^{i\cdot} - G_{..}^{i\cdot} - G_{j.}^{\cdot\cdot} + G_{..}^{\cdot\cdot}, \quad d_l^k = G_{.l}^{\cdot k} - G_{..}^{\cdot k} - G_{.l}^{\cdot\cdot} + G_{..}^{\cdot\cdot}$$

$$(a^i a^k) = G_{..}^{ik} - G_{..}^{i\cdot} - G_{..}^{\cdot k} + G_{..}^{\cdot\cdot}, \quad (a^i a_l) = G_{.l}^{i\cdot} - G_{..}^{i\cdot} - G_{.l}^{\cdot\cdot} + G_{..}^{\cdot\cdot}, \quad \cdots$$

$$(a^i d_l^k) = G_{.l}^{ik} - G_{..}^{ik} - G_{.l}^{i\cdot} - G_{.l}^{\cdot k} + G_{..}^{i\cdot} + G_{..}^{\cdot k} + G_{.l}^{\cdot\cdot} - G_{..}^{\cdot\cdot}, \quad \cdots$$

$$(d_j^i a^k) = G_{j.}^{ik} - G_{..}^{ik} - G_{j.}^{i\cdot} - G_{j.}^{\cdot k} + G_{..}^{i\cdot} + G_{..}^{\cdot k} + G_{j.}^{\cdot\cdot} - G_{..}^{\cdot\cdot}, \quad \cdots$$

$$(d_j^i d_l^k) = G_{jl}^{ik} - G_{j.}^{ik} - G_{.l}^{ik} - G_{jl}^{i\cdot} - G_{jl}^{\cdot k} + G_{..}^{ik} + G_{jl}^{\cdot\cdot} + G_{j.}^{i\cdot} + G_{.l}^{\cdot k} + G_{j.}^{\cdot k} + G_{.l}^{i\cdot} - G_{..}^{i\cdot} - G_{..}^{\cdot k} - G_{j.}^{\cdot\cdot} - G_{.l}^{\cdot\cdot} + G_{..}^{\cdot\cdot}$$

The effects average to zero over any index. Some examples are

$$\sum_i p^i(a^i a^k) = 0, \quad \sum_k p^k(a^i a^k) = 0, \quad \sum_k p^k(d^i_j a^k) = 0, \quad \sum_k p^k(a^i d^k_l) = 0, \quad \sum_j p_j(d^i_j d^k_l) = 0$$

The total variance is $\sum_{i,j,k,l} p^i p_j p^k p_l (G^{ik}_{jl} - \mu)^2$, and all product terms go to zero.

$$\begin{aligned}
\sigma^2_G &= \left[\sum_i p^i(a^i)^2 + \sum_k p^k(a^k)^2 + \sum_j p_j(a_j)^2 + \sum_l p_l(a_l)^2\right] \\
&+ \left[\sum_{i,j} p^i p_j (d^i_j)^2 + \sum_{k,l} p^k p_l (d^k_l)^2\right] + \left[\sum_{i,k} p^i p^k (a^i a^k)^2\right. \\
&+ \left.\sum_{j,l} p_j p_l (a_j a_l)^2 + \sum_{i,l} p^i p_l (a^i a_l)^2 + \sum_{j,k} p_j p^k (a_j a^k)^2\right] \\
&+ \left[\sum_{i,k,l} p^i p^k p_l (a^i d^k_l)^2 + \sum_{i,j,k} p^i p_j p^k (d^i_j a^k)^2\right. \\
&+ \left.\sum_{j,k,l} p_j p^k p_l (a_j d^k_l)^2 + \sum_{i,j,l} p^i p_j p_l (d^i_j a_l)^2\right] \\
&+ \left[\sum_{i,j,k,l} p^i p_j p^k p_l (d^i_j d^k_l)^2\right] \\
&= \sigma^2_a + \sigma^2_d + \sigma^2_{aa} + \sigma^2_{ad} + \sigma^2_{dd}
\end{aligned}$$

4.6.2 Orthogonal partition of genetic variance

Consider one locus with two alleles with three genotypes.

Genotype	AA	Aa	aa
Genotypic value	G_2	G_1	G_0
Frequency	f_2	f_1	f_0
w_1	$2f_0 + f_1$	$f_0 - f_2$	$-(2f_2 + f_1)$
w_2	$-1/(8f_2)$	$1/(4f_1)$	$-1/(8f_0)$

For three genotypes, there are two degrees of freedom, and we can choose two orthogonal scales w_1 and w_2 such that

$$\sum_{i=0}^{2} f_i w_{1i} = \sum_{i=0}^{2} f_i w_{2i} = \sum_{i=0}^{2} f_i w_{1i} w_{2i} = 0$$

$$\sigma^2_G = \sigma^2_1 + \sigma^2_2$$

The partition of the variance is

$$\sigma_1^2 = \left(\sum_{i=0}^{2} f_i G_i w_{1i}\right)^2 / \left(\sum_{i=0}^{2} f_i w_{1i}^2\right)$$

$$\sigma_2^2 = \left(\sum_{i=0}^{2} f_i G_i w_{2i}\right)^2 / \left(\sum_{i=0}^{2} f_i w_{2i}^2\right)$$

This corresponds to

$$\sigma_t^2 = [Cov(G, w_t)]^2 / \sigma_{w_t}^2 = \beta_{Gw_t}^2 \sigma_{w_t}^2$$

and

$$G = \mu + \beta_{Gw_1} w_1 + \beta_{Gw_2} w_2$$

Example: For F_2 population, let

Genotype	AA	Aa	aa
Genotypic value	G_2	G_1	G_0
Frequency	$p^2 = \frac{1}{4}$	$2pq = \frac{1}{2}$	$q^2 = \frac{1}{4}$
w_1	1	0	-1
w_2	$-\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$

$$\sigma_1^2 = 2pq[p(G_2 - G_1) + q(G_1 - G_0)]^2 = \frac{1}{8}[G_2 - G_0]^2$$

$$\sigma_2^2 = p^2 q^2 [G_2 - 2G_1 + G_0]^2 = \frac{1}{16}[G_2 - 2G_1 + G_0]^2$$

$$\beta_{Gw_1} = \frac{p^2 G_2 - q^2 G_0}{p^2 + q^2} = \frac{1}{2}(G_2 - G_0) = a$$

$$\beta_{Gw_2} = \frac{-\frac{1}{2}p^2 G_2 + \frac{1}{2}2pqG_1 - \frac{1}{2}q^2 G_0}{\frac{1}{4}p^2 + \frac{1}{4}2pq + \frac{1}{4}q^2} = (G_1 - \frac{1}{2}G_2 - \frac{1}{2}G_0) = d$$

For this analysis, we obtain an orthogonal genetic model for F_2 population
Model I: $G_2 = \mu + a - \frac{1}{2}d, \quad G_1 = \mu + \frac{1}{2}d, \quad G_0 = \mu - a - \frac{1}{2}d$
with $\mu = p^2 G_2 + 2pqG_1 + q^2 G_0$.

This model (Model I) is different from the usual model that we use in quantitative genetics
Model II: $G_2 = \mu + a, \quad G_1 = \mu + d, \quad G_0 = \mu - a$

$$v_1 = \begin{cases} 1 & \text{for } AA \\ 0 & \text{for } Aa \\ -1 & \text{for } aa \end{cases} \qquad v_2 = \begin{cases} 0 & \text{for } AA \\ 1 & \text{for } Aa \\ 0 & \text{for } aa \end{cases}$$

For F_2 population, $Cov(v_1, v_2) = \sum_{i=0}^{2} f_i v_{1i} v_{2i} = 0$. But $E(v_2) = \sum_{i=0}^{2} f_i v_{2i} = 1/2 \neq 0$.

This, however, does not affect the single locus (or single marker) analysis in the F_2 population. That is whether we use Model I or Model II for a single marker analysis, we will have the same estimates of a and d. This does not affect multiple loci (or markers) analysis as well *as long as epistatic terms are not included in the model*. When epistatic terms are included in the model, Model I and Model II will have different estimates of a and d, and Model II is not appropriate for analysis in an F_2 population. This is discussed below.

4.6.3 An F_2 based epistatic model (Cockerham model)

There are 9 genotypes in an F_2 population, so we need 8 genetic parameters to give a complete description of values of 9 genotypes. Under the assumption of Hardy-Weinberg and linkage equilibrium, Cockerham (1954)'s orthogonal partition of genetic variance leads to the definition of the genotypic value G_{ij}

$$G_{ij} = \beta_0 + \sum_{t=1}^{8} \beta_{Gw_t} w_{tij}$$

by eight orthogonal scales or contrasts w_t's. Four are marginal scales and four are interaction scales. Marginal scales (defined by Model I for an F_2 population) are called linear and quadratic scales (additive and dominance scales in genetic terms). Correspondingly, the interaction scales are

$w_5 = w_1 \times w_3$	linear × linear	additive × additive
$w_6 = w_1 \times w_4$	linear × quadratic	additive × dominance
$w_7 = w_2 \times w_3$	quadratic × linear	dominance × additive
$w_8 = w_2 \times w_4$	quadratic × quadratic	dominance × dominance

Thus the model is

$$G_{ij} = \beta_0 + \beta_{Gw_1} w_{1ij} + \beta_{Gw_2} w_{2ij} + \beta_{Gw_3} w_{3ij} + \beta_{Gw_4} w_{4ij}$$

$$+\beta_{Gw_5} w_{1ij} w_{3ij} + \beta_{Gw_6} w_{1ij} w_{4ij} + \beta_{Gw_7} w_{2ij} w_{3ij} + \beta_{Gw_8} w_{2ij} w_{4ij}$$

with

$$w_1 = \begin{cases} 1 & \text{for } AA \\ 0 & \text{for } Aa \\ -1 & \text{for } aa \end{cases} \qquad w_2 = \begin{cases} -\frac{1}{2} & \text{for } AA \\ \frac{1}{2} & \text{for } Aa \\ -\frac{1}{2} & \text{for } aa \end{cases}$$

$$w_3 = \begin{cases} 1 & \text{for } BB \\ 0 & \text{for } Bb \\ -1 & \text{for } bb \end{cases} \qquad w_4 = \begin{cases} -\frac{1}{2} & \text{for } BB \\ \frac{1}{2} & \text{for } Bb \\ -\frac{1}{2} & \text{for } bb \end{cases}$$

$$w_5 = w_1 w_3, \;\; w_6 = w_1 w_4, \;\; w_7 = w_2 w_3, \;\; w_8 = w_2 w_4$$

Table 4.1: Genotypic values, frequencies, and eight orthogonal scales for F_2 populations

Scale	AABB	AABb	AAbb	AaBB	AaBb	Aabb	aaBB	aaBb	aabb
G	G_{22}	G_{21}	G_{20}	G_{12}	G_{11}	G_{10}	G_{02}	G_{01}	G_{00}
f	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{16}$
W_1	1	1	1	0	0	0	-1	-1	-1
W_2	$-\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$
W_3	1	0	-1	1	0	-1	1	0	-1
W_4	$-\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$
W_5	1	0	-1	0	0	0	-1	0	1
W_6	$-\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$	0	0	0	$\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{2}$
W_7	$-\frac{1}{2}$	0	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$	$-\frac{1}{2}$	0	$\frac{1}{2}$
W_8	$\frac{1}{4}$	$-\frac{1}{4}$	$\frac{1}{4}$	$-\frac{1}{4}$	$\frac{1}{4}$	$-\frac{1}{4}$	$\frac{1}{4}$	$-\frac{1}{4}$	$\frac{1}{4}$

In matrix notation,

$$\mathbf{G} = \mathbf{W}\beta$$

$$\begin{bmatrix} G_{22} \\ G_{21} \\ G_{20} \\ G_{12} \\ G_{11} \\ G_{10} \\ G_{02} \\ G_{01} \\ G_{00} \end{bmatrix} = \begin{bmatrix} 1 & 1 & -\frac{1}{2} & 1 & -\frac{1}{2} & 1 & -\frac{1}{2} & -\frac{1}{2} & \frac{1}{4} \\ 1 & 1 & -\frac{1}{2} & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & -\frac{1}{4} \\ 1 & 1 & -\frac{1}{2} & -1 & -\frac{1}{2} & -1 & -\frac{1}{2} & \frac{1}{2} & \frac{1}{4} \\ 1 & 0 & \frac{1}{2} & 1 & -\frac{1}{2} & 0 & 0 & \frac{1}{2} & -\frac{1}{4} \\ 1 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 & \frac{1}{4} \\ 1 & 0 & \frac{1}{2} & -1 & -\frac{1}{2} & 0 & 0 & -\frac{1}{2} & -\frac{1}{4} \\ 1 & -1 & -\frac{1}{2} & 1 & -\frac{1}{2} & -1 & \frac{1}{2} & -\frac{1}{2} & \frac{1}{4} \\ 1 & -1 & -\frac{1}{2} & 0 & \frac{1}{2} & 0 & -\frac{1}{2} & 0 & -\frac{1}{4} \\ 1 & -1 & -\frac{1}{2} & -1 & -\frac{1}{2} & 1 & \frac{1}{2} & \frac{1}{2} & \frac{1}{4} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \\ \beta_7 \\ \beta_8 \end{bmatrix}$$

$$\beta = \mathbf{W}^{-1}\mathbf{G}$$

$$\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \\ \beta_7 \\ \beta_8 \end{bmatrix} = \begin{bmatrix} \frac{1}{16} & \frac{1}{8} & \frac{1}{16} & \frac{1}{8} & \frac{1}{4} & \frac{1}{8} & \frac{1}{16} & \frac{1}{8} & \frac{1}{16} \\ \frac{1}{8} & \frac{1}{4} & \frac{1}{8} & 0 & 0 & 0 & -\frac{1}{8} & -\frac{1}{4} & -\frac{1}{8} \\ -\frac{1}{16} & -\frac{1}{8} & -\frac{1}{16} & \frac{1}{8} & \frac{1}{4} & \frac{1}{8} & -\frac{1}{16} & -\frac{1}{8} & -\frac{1}{16} \\ \frac{1}{8} & 0 & -\frac{1}{8} & \frac{1}{4} & 0 & -\frac{1}{4} & \frac{1}{8} & 0 & -\frac{1}{8} \\ -\frac{1}{16} & \frac{1}{8} & -\frac{1}{16} & -\frac{1}{8} & \frac{1}{4} & -\frac{1}{8} & -\frac{1}{16} & \frac{1}{8} & -\frac{1}{16} \\ \frac{1}{4} & 0 & -\frac{1}{4} & 0 & 0 & 0 & -\frac{1}{4} & 0 & \frac{1}{4} \\ -\frac{1}{4} & \frac{1}{2} & -\frac{1}{4} & 0 & 0 & 0 & \frac{1}{4} & -\frac{1}{2} & \frac{1}{4} \\ -\frac{1}{4} & 0 & \frac{1}{4} & \frac{1}{2} & 0 & -\frac{1}{2} & -\frac{1}{4} & 0 & \frac{1}{4} \\ \frac{1}{4} & -\frac{1}{2} & \frac{1}{4} & -\frac{1}{2} & \frac{1}{4} & -\frac{1}{2} & \frac{1}{4} & -\frac{1}{2} & \frac{1}{4} \end{bmatrix} \begin{bmatrix} G_{22} \\ G_{21} \\ G_{20} \\ G_{12} \\ G_{11} \\ G_{10} \\ G_{02} \\ G_{01} \\ G_{00} \end{bmatrix}$$

$$\beta = \mathbf{W}^{-1}\mathbf{G}$$

Table 4.2: Definition of genetic parameters

Parameters	Definition
$\mu = \beta_0$	mean
$a_1 = \beta_1$	additive effect of locus A
$d_1 = \beta_2$	dominance effect of locus A
$a_2 = \beta_3$	additive effect of locus B
$d_2 = \beta_4$	dominance effect of locus B
$i_{aa} = \beta_5$	additive × additive effect of locus A and B
$i_{ad} = \beta_6$	additive × dominance effect of locus A and B
$i_{da} = \beta_7$	dominance × additive effect of loci A and B
$i_{dd} = \beta_8$	dominance × dominance effect of loci A and B

$$
\begin{aligned}
\beta_0 &= \frac{G_{22}}{16} + \frac{G_{21}}{8} + \frac{G_{20}}{16} + \frac{G_{12}}{8} + \frac{G_{11}}{4} + \frac{G_{10}}{8} + \frac{G_{02}}{16} + \frac{G_{01}}{8} + \frac{G_{00}}{16} = G_{..} \\
\beta_1 &= \frac{G_{22}}{8} + \frac{G_{21}}{4} + \frac{G_{20}}{8} - \frac{G_{02}}{8} - \frac{G_{01}}{4} - \frac{G_{00}}{8} = \frac{G_{2.} - G_{0.}}{2} \\
\beta_2 &= \frac{G_{11}}{4} + \frac{G_{12}}{8} + \frac{G_{10}}{8} - \frac{G_{22}}{16} - \frac{G_{21}}{8} - \frac{G_{20}}{16} - \frac{G_{02}}{16} - \frac{G_{01}}{8} - \frac{G_{00}}{16} = \frac{2G_{1.} - G_{2.} - G_{0.}}{2} \\
\beta_3 &= \frac{G_{22}}{8} + \frac{G_{12}}{4} + \frac{G_{02}}{8} - \frac{G_{20}}{8} - \frac{G_{10}}{4} - \frac{G_{00}}{8} = \frac{G_{.2} - G_{.0}}{2} \\
\beta_4 &= \frac{G_{11}}{4} + \frac{G_{21}}{8} + \frac{G_{01}}{8} - \frac{G_{22}}{16} - \frac{G_{12}}{8} - \frac{G_{02}}{16} - \frac{G_{20}}{16} - \frac{G_{10}}{8} - \frac{G_{00}}{16} = \frac{2G_{.1} - G_{.2} - G_{.0}}{2} \\
\beta_5 &= \frac{(G_{22} - G_{02}) - (G_{20} - G_{00})}{4} = \frac{(G_{22} - G_{20}) - (G_{02} - G_{00})}{4} \\
\beta_6 &= \frac{(2G_{21} - G_{22} - G_{20}) - (2G_{01} - G_{02} - G_{00})}{4} \\
\beta_7 &= \frac{(2G_{12} - G_{22} - G_{02}) - (2G_{10} - G_{20} - G_{00})}{4} \\
\beta_8 &= \frac{2(2G_{11} - G_{21} - G_{01}) - (2G_{12} - G_{22} - G_{02}) - (2G_{10} - G_{20} - G_{00})}{4} \\
&= \frac{2(2G_{11} - G_{12} - G_{10}) - (2G_{21} - G_{22} - G_{20}) - (2G_{01} - G_{02} - G_{00})}{4}
\end{aligned}
$$

An important property: By orthogonality, the regression coefficient β_{Gw_t} associated with each scale w_t is the corresponding genetic effect, and each effect contributes to its corresponding genetic variance component.

$$\sigma^2_G = \frac{1}{2}a_1^2 + \frac{1}{4}d_1^2 + \frac{1}{2}a_2^2 + \frac{1}{4}d_2^2 + \frac{1}{4}i_{aa}^2 + \frac{1}{8}i_{ad}^2 + \frac{1}{8}i_{da}^2 + \frac{1}{16}i_{dd}^2$$

There is no covariance between different effects.

4.6.4 A comparison with Mather and Jinks model

If we extend Model II to include epistasis as

$$G_{ij} = \beta_0 + \sum_{t=1}^{8} \beta_{Gw_t} v_{tij}$$

with $v_1, \cdots, v_4$ defined by Model II and $v_5 = v_1 v_3$, $v_6 = v_1 v_4$, $v_7 = v_2 v_3$, $v_8 = v_2 v_4$, we then have Mather and Jinks (1982 *Biometrical Genetics*) model:

$$G_{ij} = \beta_0 + \beta_{Gv_1} v_{1ij} + \beta_{Gv_2} v_{2ij} + \beta_{Gv_3} v_{3ij} + \beta_{Gv_4} v_{4ij}$$

$$+\beta_{Gv_5} v_{1ij} v_{3ij} + \beta_{Gv_6} v_{1ij} v_{4ij} + \beta_{Gv_7} v_{2ij} v_{3ij} + \beta_{Gv_8} v_{2ij} v_{4ij}$$

$$v_1 = \begin{cases} 1 & \text{for } AA \\ 0 & \text{for } Aa \\ -1 & \text{for } aa \end{cases} \qquad v_2 = \begin{cases} 0 & \text{for } AA \\ 1 & \text{for } Aa \\ 0 & \text{for } aa \end{cases}$$

$$v_3 = \begin{cases} 1 & \text{for } BB \\ 0 & \text{for } Bb \\ -1 & \text{for } bb \end{cases} \qquad v_4 = \begin{cases} 0 & \text{for } BB \\ 1 & \text{for } Bb \\ 0 & \text{for } bb \end{cases}$$

$$v_5 = v_1 v_3, \quad v_6 = v_1 v_4, \quad v_7 = v_2 v_3, \quad v_8 = v_2 v_4$$

$$\mathbf{G} = \mathbf{V}\beta$$

$$\begin{bmatrix} G_{22} \\ G_{21} \\ G_{20} \\ G_{12} \\ G_{11} \\ G_{10} \\ G_{02} \\ G_{01} \\ G_{00} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & -1 & 0 & -1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & -1 & 0 & 0 & 0 & -1 & 0 \\ 1 & -1 & 0 & 1 & 0 & -1 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 1 & 0 & -1 & 0 & 0 \\ 1 & -1 & 0 & -1 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \mu \\ a_1 \\ d_1 \\ a_2 \\ d_2 \\ \text{ı}_{aa} \\ i_{ad} \\ i_{da} \\ i_{dd} \end{bmatrix}$$

$$\beta = \mathbf{V}^{-1}\mathbf{G}$$

$$\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \\ \beta_7 \\ \beta_8 \end{bmatrix} = \begin{bmatrix} \frac{1}{4} & 0 & \frac{1}{4} & 0 & 0 & 0 & \frac{1}{4} & 0 & \frac{1}{4} \\ \frac{1}{4} & 0 & -\frac{1}{4} & 0 & 0 & 0 & \frac{1}{4} & 0 & -\frac{1}{4} \\ -\frac{1}{4} & 0 & -\frac{1}{4} & \frac{1}{2} & 0 & \frac{1}{2} & -\frac{1}{4} & 0 & -\frac{1}{4} \\ \frac{1}{4} & 0 & \frac{1}{4} & 0 & 0 & 0 & -\frac{1}{4} & 0 & -\frac{1}{4} \\ -\frac{1}{4} & \frac{1}{2} & -\frac{1}{4} & 0 & 0 & 0 & -\frac{1}{4} & \frac{1}{2} & -\frac{1}{4} \\ \frac{1}{4} & 0 & -\frac{1}{4} & 0 & 0 & 0 & -\frac{1}{4} & 0 & \frac{1}{4} \\ -\frac{1}{4} & \frac{1}{2} & -\frac{1}{4} & 0 & 0 & 0 & \frac{1}{4} & -\frac{1}{2} & \frac{1}{4} \\ -\frac{1}{4} & 0 & \frac{1}{4} & \frac{1}{2} & 0 & -\frac{1}{2} & -\frac{1}{4} & 0 & \frac{1}{4} \\ \frac{1}{4} & -\frac{1}{2} & \frac{1}{4} & -\frac{1}{2} & \frac{1}{4} & -\frac{1}{2} & \frac{1}{4} & -\frac{1}{2} & \frac{1}{4} \end{bmatrix} \begin{bmatrix} G_{22} \\ G_{21} \\ G_{20} \\ G_{12} \\ G_{11} \\ G_{10} \\ G_{02} \\ G_{01} \\ G_{00} \end{bmatrix}$$

Table 4.3: Mean value and sample size of two unlinked QTL genotypes for trait 1 (John Doebley)

	bb	Bb	BB
aa	17.98 (10)	54.57 (21)	47.80 (11)
Ab	40.94 (24)	47.55 (42)	83.62 (20)
AA	61.11 (3)	66.50 (22)	101.65 (8)

This model was actually derived by starting the analysis with the four homozygote genotypes (G_{22}, G_{20}, G_{02}, G_{00}) (which can be composed by a population of recombinant inbred lines, F_∞) and defining μ, a_1, a_2 and i_{aa} there. So

$$\mu = \frac{1}{4}G_{22} + \frac{1}{4}G_{20} + \frac{1}{4}G_{02} + \frac{1}{4}G_{00}$$

Then by considering heterozygote genotypes, d_1, d_2, i_{aa}, i_{ad}, i_{da} and i_{dd} are added to the model one by one. Because that the stating base is F_∞ population for four homozygote genotypes, the model is usually called F_∞ metric model (Va der Veen 1959 *Genetica* 30:201-232). So the model is not based on a particular population where the nine genotypes are observed and is not based on the orthogonal partitions of genetic effect and variance.

If the model is applied to an F_2 population, v_2 and v_8 are not independent, neither are v_4 and v_8. Because of it, some genetic effects defined by the model do not only contribute to their own genetic variance component, but others also. For this reason, we prefer Cockerham model, also called F_2 metric model, for analyzing QTL epistasis in an F_2 population.

Note: Estimates of i_{aa}, i_{ad}, i_{da}, and i_{dd} are the same under both models, but those of a_1, d_1, a_2 and d_2 are different.

4.6.5 Examples of analysis

As an example, we showed the epistatic analysis of two QTL in a series of experiments of introgressing maize allele into teosinte genetic background (Doebley, Stec, and Gustus 1995 *Genetics* 141:333-346). Two QTL, one located on chromosome arm 1L and one on 3L, were introgressed by repeated backcross to teosinte and then intercrossed to produce an F_2 population to have these two QTL segregating simultaneously in teosinte background. We showed the mean values and sample sizes of the nine genotypes and the analyses of gene effects on three traits.

4.6.6 Issues on detecting QTL epistasis

- How to detect and analyze QTL epistasis is a very important issue for QTL mapping analysis. If we know where QTL are located, we can build appropriate genetic and statistical models to test and estimate the effects and interactions of QTL at the

Table 4.4: Estimate and test of gene effects: Trait 1

Effect	Estimate	S.E.	t	P
a_1	15.11	4.47	3.41	0.0008
d_1	-3.92	5.84	-0.67	0.5035
a_2	19.46	4.42	4.40	0.0001
d_2	-5.66	5.84	-0.97	0.3336
i_{aa}	2.68	7.07	0.38	0.7054
i_{ad}	-18.28	8.87	-2.06	0.0411
i_{da}	3.75	8.85	0.42	0.6725
i_{dd}	-18.13	11.68	-1.55	0.1227

Table 4.5: Mean value and sample size of two unlinked QTL genotypes for trait 4 (John Doebley)

	bb	Be	BB
aa	4.46 (12)	3.05 (22)	2.09 (11)
Aa	3.23 (26)	2.02 (46)	1.12 (21)
AA	1.43 (7)	1.08 (25)	1.00 (9)

Table 4.6: Estimate and test of gene effects: Trait 4

Effect	Estimate	S.E.	t	P
a_1	-1.01	0.09	-10.67	0.0001
d_1	-0.06	0.13	-0.43	0.6712
a_2	-0.88	0.09	-9.29	0.0001
d_2	-0.17	0.13	-1.29	0.1994
i_{aa}	0.48	0.14	3.45	0.0007
i_{ad}	0.05	0.19	0.25	0.8017
i_{da}	-0.36	0.19	-1.89	0.0606
i_{dd}	0.03	0.26	0.11	0.9126

Table 4.7: Mean value and sample size of two unlinked QTL genotypes for trait 9 (John Doebley)

	bb	Bb	BB
aa	7.32 (12)	0.41 (21)	1.12 (10)
Aa	1.34 (26)	0.22 (46)	0.28 (21)
AA	0.79 (7)	0.00 (25)	0.00 (9)

Table 4.8: Estimate and test of gene effects: Trait 9

Effect	Estimate	S.E.	t	P
a_1	-1.06	0.29	-3.68	0.0003
d_1	-0.74	0.39	-1.88	0.0621
a_2	-1.14	0.28	-3.98	0.0001
d_2	-1.35	0.39	-3.42	0.0008
i_{aa}	1.35	0.43	3.16	0.0019
i_{ad}	1.71	0.57	2.98	0.0033
i_{da}	1.22	0.57	2.13	0.0346
i_{dd}	1.52	0.79	1.93	0.0550

locations. But the problem is we do not know where QTL are, and the analyses of QTL effects and interactions are complicated by the search for QTL.

- A common practice to detect QTL epistasis in many published QTL mapping analyses is to perform pairwise marker interaction analysis and to compare the percentage of significant marker pairs on interaction at a given significance level with that expected at the null hypothesis of no interaction.

  This analysis, of course, does not analyze interaction of individual QTL. It generally has low power to detect QTL epistasis because of segregation between markers and QTL and possible aggregation and *cancellation* effects from multiple QTL. Also some interactions at some regions can be swamped by the huge number of possible pair-wise tests of markers.

- An appropriate method to analyze QTL epistasis is to integrate QTL epistatic effects in the analysis of QTL and to search and map multiple QTL including epistasis simultaneously. Multiple interval mapping discussed below is such a method.

Chapter 5

One Marker Analysis

5.1 Backcross population

The simplest method of associating markers with quantitative trait variation is to test for trait value differences between different marker groups of individuals for a particular marker. For example, if we let $\tilde{\mu}_1$ and $\tilde{\mu}_0$ be the observed trait means of the groups of individuals with marker genotypes M_i/M_i and M_i/m_i for marker i in a backcross population, we can test for significance between means $\tilde{\mu}_1$ and $\tilde{\mu}_0$ using the usual t test with the statistic

$$t = \frac{\tilde{\mu}_1 - \tilde{\mu}_0}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_0}\right)}}.$$

where s^2 is the pooled sampling variance given by

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_0 - 1)s_0^2}{n_1 + n_0 - 2}$$

n_1, n_0 and s_1^2, s_0^2 are corresponding sample size and variance in each marker class, respectively.

The hypotheses under the test can be

$$H_0 : \ \mu_1 = \mu_0 \ \text{ and } \ H_1 : \ \mu_1 \neq \mu_0$$

Statistically, this is equivalent to the simple regression analysis with a model

$$y_j = \mu + bx_j + e_j \qquad j = 1, 2, \cdots, n \tag{5.1}$$

where

$$\begin{aligned} y_j &= \text{the trait value of individual } j \\ \mu &= \text{mean of the model} \end{aligned}$$

Table 5.1: One marker analysis (t test): Mouse data

	Marker class 1			Marker class 0				
Marker	n_1	$\hat{\mu}_1$	s_1^2	n_0	$\hat{\mu}_0$	s_0^2	t	P value
1 Hmg1-rs13	41	54.20	111.81	62	47.32	63.67	3.754	0.0001
2 DXMit57	42	55.21	104.12	61	46.51	56.12	4.994	0.000001
3 Rps17-rs11	43	55.30	101.98	60	46.30	54.38	5.231	<0.000001
4 Rps18-rs17	42	55.60	100.69	61	46.25	53.66	5.467	<0.000001
5 DXMit48	43	55.19	105.20	60	46.38	53.60	5.085	<0.000001
6 DXNds1	44	55.43	102.25	59	46.05	50.12	5.538	<0.000001
7 DXMit109	45	55.00	114.05	58	46.22	44.77	5.103	<0.000001
8 Hmg14-rs6	49	54.86	105.83	54	45.70	43.19	5.431	<0.000001
9 DXMit60	50	54.62	106.49	53	45.75	43.88	5.218	<0.000001
10 DXMit16	50	54.68	106.10	53	45.70	43.22	5.306	<0.000001
11 DXMit97	50	54.64	107.05	53	45.74	43.01	5.248	<0.000001
12 Hmg1-rs14	51	53.90	104.61	52	46.29	54.88	4.333	0.000008
13 DXMit3	56	53.50	112.25	47	45.96	41.17	4.266	0.00001
14 Tpm3-rs9	49	53.02	126.06	54	47.37	50.01	3.085	0.001

$$
\begin{aligned}
x_j &= \begin{cases} 1 & \text{if individual } j \text{ has } M_i/M_i \text{ genotype} \\ 0 & \text{if individual } j \text{ has } M_i/m_i \text{ genotype} \end{cases} \\
b &= \mu_1 - \mu_0 \text{ (the simple regression coefficient)} \\
e_j &\sim N(0, \sigma^2) \text{ (a normal distributed random residual variable)}
\end{aligned}
$$

A test can be performed on $\hat{b}$ under $H_0: \ b = 0$ and $H_1: \ b \neq 0$.

It is clear that all markers are significantly associated with QTL as t tests are very significant for all the markers. From this analysis alone, however, it is not clear how many QTL are on the chromosome and where are those QTL located. Note that in this sample, sample variances are very different for different marker classes.

5.2 F_2 population

In an F_2 population, there are three marker genotypes. For this population we can construct two t tests to test *marker* additive and dominance effects separately. Let $\tilde{\mu}_2$, $\tilde{\mu}_1$ and $\tilde{\mu}_0$ be the observed trait means of the groups of individuals with marker genotypes M_i/M_i, M_i/m_i and m_i/m_i for marker i in a F_2 population with corresponding sample sizes n_2, n_1, n_0 and variances s_2^2, s_1^2, s_0^2. To test marker additive effect, the test statistic is

$$
t_1 = \frac{\tilde{\mu}_2 - \tilde{\mu}_0}{\sqrt{s^2\left(\frac{1}{n_2} + \frac{1}{n_0}\right)}} \quad \text{with} \quad s^2 = \frac{(n_2 - 1)s_2^2 + (n_0 - 1)s_0^2}{n_2 + n_0 - 2}
$$

Table 5.2: One marker analysis (t test): Maize data

	Marker class 2			Marker class 1			Marker class 0						
M	n_2	$\hat{\mu}_2$	s_2^2	n_1	$\hat{\mu}_1$	s_1^2	n_0	$\hat{\mu}_0$	s_0^2	t_1	P value	t_2	P value
1	43	5.24	2.44	86	4.27	2.93	42	3.11	2.76	6.10	<0.000001	0.38	0.704
2	48	4.82	3.15	89	4.17	3.26	34	3.54	2.84	3.28	0.001	-0.05	0.958
3	42	5.01	3.23	92	4.14	3.18	37	3.57	2.68	3.71	0.0002	-0.57	0.567
4	44	4.47	2.96	89	4.21	3.36	38	3.99	3.61	1.20	0.230	-0.05	0.958
5	43	4.57	3.13	87	4.21	3.37	41	3.91	3.28	1.68	0.093	-0.13	0.897
6	43	4.48	3.03	83	4.06	2.85	45	4.29	4.43	0.46	0.646	-1.18	0.238
7	44	4.28	3.09	83	4.09	3.01	44	4.44	4.14	-0.39	0.698	-0.96	0.337
8	47	4.36	2.73	81	4.09	3.35	43	4.34	3.93	0.06	0.952	-0.92	0.358
9	41	4.16	2.36	86	4.21	3.62	44	4.32	3.68	-0.41	0.684	-0.10	0.920
10	40	3.94	2.75	94	4.30	3.68	37	4.34	2.99	-1.04	0.298	0.57	0.567
11	45	4.25	3.64	89	4.14	3.32	37	4.41	2.97	-0.38	0.704	-0.66	0.509
12	46	4.04	3.25	85	4.25	3.54	40	4.40	2.94	-0.94	0.347	0.09	0.928

and to test marker dominance effect, the test statistic is

$$t_2 = \frac{\tilde{\mu}_1 - \tilde{\mu}_2/2 - \tilde{\mu}_0/2}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{4n_2} + \frac{1}{4n_0}\right)}} \quad \text{with} \quad s^2 = \frac{(n_2-1)s_2^2 + (n_1-1)s_1^2 + (n_0-1)s_0^2}{n_2 + n_1 + n_0 - 3}$$

For this data, only t_1 for marker 1, 2 and 3 are significant. This indicates that there could be at least one, maybe two, QTL on the chromosome, and QTL effects are largely additive.

Test can also be performed through ANOVA to analyze between marker class variance for each marker for this design.

Point: Although the t test assumes normal trait distributions within marker classes, the test is quite robust to violations of the assumption. It can be argued that trait distributions within marker classes are not normals, but mixtures of normals because of segregation of genes. Doerge (1993), however, argued that the t test generally performs well even though the distributions are mixtures. Departures of the normal distributions of the test statistic for the t test are likely to be anticipated only for parental populations with large differences between means, but this is the condition for which it is most likely there will be departures from the null hypothesis and the power of the test is not significantly affected.

5.3 Genetical meaning of the analysis

To understand the relevance of this test to QTL mapping, we need to know what is being tested in genetic terms. Consider a QTL Q linked to a marker M. For a backcross $\frac{MQ}{MQ} \times \frac{MQ}{mq}$,

we have

	$\frac{MQ}{MQ}$	$\frac{MQ}{Mq}$	$\frac{MQ}{mQ}$	$\frac{MQ}{mq}$
Frequency	$(1-r)/2$	$r/2$	$r/2$	$(1-r)/2$
Mean effect	$\mu+a$	$\mu+d$	$\mu+a$	$\mu+d$

		QQ	Qq	mean effect
MM	Cond. Freq.	$1-r$	r	$(1-r)(\mu+a)+r(\mu+d)$
	Effect	$\mu+a$	$\mu+d$	
Mm	Cond. Freq.	r	$1-r$	$r(\mu+a)+(1-r)(\mu+d)$
	Effect	$\mu+a$	$\mu+d$	

Thus

$$\begin{aligned}\mu_{MM}-\mu_{Mm} &= [(1-r)(\mu+a)+r(\mu+d)]-[r(\mu+a)+(1-r)(\mu+d)] \\ &= (1-2r)a-(1-2r)d=(1-2r)(a-d)\end{aligned}$$

For multiple QTL linked to marker M_i (ignoring epistasis):

$$\mu_{M_iM_i}-\mu_{M_im_i}=\sum_{k=1}^{m}(1-2r_{ik})(a_k-d_k)$$

This means that we are testing a composite parameter that is comprised of QTL effects and recombination frequencies for (potentially) a number of QTL. Of course, many QTL may not be linked to the marker and thus have 0.5 recombination frequency. The above hypotheses are then equivalent to

$$H_0:\ \text{all } r_{ik}=0.5 \quad \text{and} \quad H_1:\ \text{at least one } \ r_{ik}<0.5,$$

because δ_k's are assumed to be non-zero (*i.e.*, by experiment design we know that there are some genes segregating in the population). If $\tilde{\mu}_1$ and $\tilde{\mu}_0$ are found to be significantly different, it is indicated that the marker is linked to one or possibly more QTL.

5.4 Likelihood analysis

Likelihood analysis can also be performed on a single marker. Consider a marker M and a QTL Q.

	$\frac{MQ}{MQ}$	$\frac{MQ}{Mq}$	$\frac{MQ}{mQ}$	$\frac{MQ}{mq}$
Frequency	$(1-r_{MQ})/2$	$r_{MQ}/2$	$r_{MQ}/2$	$(1-r_{MQ})/2$
Dist. of y	$N(\mu+\delta,\sigma^2)$	$N(\mu,\sigma^2)$	$N(\mu+\delta,\sigma^2)$	$N(\mu,\sigma^2)$

	QQ	Qq	Distribution of y
MM	$1-r_{MQ}$	r_{MQ}	$(1-r_{MQ})N(\mu+\delta,\sigma^2)+r_{MQ}N(\mu,\sigma^2)$
Mm	r_{MQ}	$1-r_{MQ}$	$r_{MQ}N(\mu+\delta,\sigma^2)+(1-r_{MQ})N(\mu,\sigma^2)$

Likelihood:

$$L(\mu, \delta, \sigma^2, r_{MQ}) = \prod_{j=1}^{n_{MM}} \left[(1 - r_{MQ}) \phi \left(\frac{y_{1j} - \mu - \delta}{\sigma} \right) + r_{MQ} \phi \left(\frac{y_{1j} - \mu}{\sigma} \right) \right]$$

$$\prod_{j=1}^{n_{Mm}} \left[r_{MQ} \phi \left(\frac{y_{2j} - \mu - \delta}{\sigma} \right) + (1 - r_{MQ}) \phi \left(\frac{y_{2j} - \mu}{\sigma} \right) \right]$$

where $\phi(z) = \frac{1}{\sqrt{2\pi}} \exp[-z^2/2]$.

The hypothesis $H_0 : r_{MQ} = 1/2$ can be tested by a likelihood ratio

$$LR = -2 \ln \frac{L(\hat{\mu}, \hat{\delta}, \hat{\sigma^2}, r_{MQ} = 1/2)}{L(\hat{\mu}, \hat{\delta}, \hat{\sigma^2}, \hat{r}_{MQ})}$$

Note that this is a very specific model and specific test. Basically we assume that there is one and only one QTL Q segregating in the population and we are asking whether this Q is linked to M or not. If linked, estimate r_{MQ}.

5.5 Problems of the analysis

Although simple, this analysis captures the basic ideas of QTL mapping. Clearly there are many problems with this simple approach (McMillan and Robertson 1974; Lander and Botstein 1989), such as:

1. The method cannot tell whether the markers are associated with one or more QTL;
2. The method does not estimate the likely positions of the QTL;
3. The effects of QTL are likely to be underestimated because they are confounded with the recombination frequencies;
4. Because of the confounding effects, the method is not very powerful and many individuals are required for the test.

Chapter 6

Interval Mapping

Interval mapping was introduced by Lander and Botstein (1989) as a systematical way to scan the genome for evidence of QTL. Statistically, it is an extension of one marker analysis, extending the analysis from a marker to any genomic location flanked by two markers. However, conceptually, it represents a leap from the old thinking of one marker at a time to genome-wise search. The concept of using complete marker linkage maps for genomic scanning of QTL is revolutionary, and the idea of viewing QTL genotypes as missing data and using a mixture model for maximum likelihood analysis is influential. This concept and idea have been maintained in various extensions of interval mapping analysis.

In this part of lecture, we review the basic and practice of interval mapping, and then proceed to composite interval mapping and multiple interval mapping, some extensions of interval mapping analysis. For most discussion, we use a backcross design for illustration. The argument and analysis can be readily extended to other experimental designs (such as F_2) (*e.g.,* Luo and Williams 1993; Jansen 1996; Jiang and Zeng 1997; Kao and Zeng 1997). For a more detailed coverage of the subject, consult Lynch and Walsh (1998) and Doerge, Zeng and Weir (1997).

6.1 Model

In one marker analysis, we test and estimate quantitative trait effect associated to each marker. If the effect of a marker is tested to be significant, we say that the marker is linked to one or more QTL. However, the analysis can not directly map the QTL because the marker effect is a function of r (recombination frequency between a marker and a QTL) and a (effect of a QTL). To solve this problem, Lander and Botstein (1989) proposed to use a pair of markers to disentangle r and a from the test statistic and implemented it by using maximum likelihood procedures.

Specifically, for a backcross design they proposed the following linear model to test for a QTL (Q) located on an interval flanked by markers i and $i+1$ (M_i and M_{i+1}) (assuming the order M_iQM_{i+1})

$$y_j = \mu + b^* x_j^* + e_j \qquad j = 1, 2, \cdots, n \tag{6.1}$$

where

$$\begin{aligned} b^* &= \text{the effect of the putative QTL} \\ x_j^* &= \begin{cases} 1 & \text{if the QTL genotype is } QQ \\ 0 & \text{if the QTL genotype is } Qq \end{cases} \\ e_j &\sim N(0, \sigma^2) \end{aligned}$$

Statistically this is a mixture model as x_j^*, which is unobserved, can take different values with probabilities depending on the type of the flanking markers (M_i, M_{i+1}) of the jth individual and the testing position $(\theta = r_{M_iQ}/r_{M_iM_{i+1}})$. Let

$$p_{kj} = Prob(x_j^* = k | M_i, M_{i+1}, \theta) \quad k = 0, 1.$$

which is specified below for the backcross population.

		QTL genotype	
Marker genotype	Sample size	QQ (1)	Qq (0)
$M_iM_iM_{i+1}M_{i+1}$	n_1	$\frac{(1-r_{M_iQ})(1-r_{QM_{i+1}})}{1-r_{M_iM_{i+1}}} \approx 1$	$\frac{r_{M_iQ}r_{QM_{i+1}}}{1-r_{M_iM_{i+1}}} \approx 0$
$M_iM_im_{i+1}M_{i+1}$	n_2	$\frac{(1-r_{M_iQ})r_{QM_{i+1}}}{r_{M_iM_{i+1}}} \approx 1-\theta$	$\frac{r_{M_iQ}(1-r_{QM_{i+1}})}{r_{M_iM_{i+1}}} \approx \theta$
$m_iM_iM_{i+1}M_{i+1}$	n_3	$\frac{r_{M_iQ}(1-r_{QM_{i+1}})}{r_{M_iM_{i+1}}} \approx \theta$	$\frac{(1-r_{M_iQ})r_{QM_{i+1}}}{r_{M_iM_{i+1}}} \approx 1-\theta$
$m_iM_im_{i+1}M_{i+1}$	n_4	$\frac{r_{M_iQ}r_{QM_{i+1}}}{1-r_{M_iM_{i+1}}} \approx 0$	$\frac{(1-r_{M_iQ})(1-r_{QM_{i+1}})}{1-r_{M_iM_{i+1}}} \approx 1$

Approximation is obtained by assuming that double recombination can be ignored. The likelihood function of (6.1) is given by

$$\begin{aligned} L(\mu, b^*, \sigma^2, \theta) &= \prod_{j=1}^{n} \left[p_{1j}\phi\left(\frac{y_j - \mu - b^*}{\sigma}\right) + p_{0j}\phi\left(\frac{y_j - \mu}{\sigma}\right)\right] \\ &= \prod_{j=1}^{n_1} \phi\left(\frac{y_j - \mu - b^*}{\sigma}\right) \prod_{j=1}^{n_2} \left[(1-\theta)\phi\left(\frac{y_j - \mu - b^*}{\sigma}\right) + \theta\phi\left(\frac{y_j - \mu}{\sigma}\right)\right] \\ &\quad \prod_{j=1}^{n_3} \left[\theta\phi\left(\frac{y_j - \mu - b^*}{\sigma}\right) + (1-\theta)\phi\left(\frac{y_j - \mu}{\sigma}\right)\right] \prod_{j=1}^{n_4} \phi\left(\frac{y_j - \mu}{\sigma}\right) \end{aligned}$$

where $\phi(z) = \frac{1}{\sqrt{2\pi}} \exp[-z^2/2]$ is the standard normal density function.

6.2 Maximum likelihood analysis

The maximum likelihood analyses of interval mapping have been discussed extensively in the literature (*e.g.*, Lander and Botstein 1989; Carbonell and Gerig 1991; Luo and Kearsey

1992; van Ooijen 1992; Carbonell et al. 1992; Luo and Williams 1993; Jansen 1992, 1993, 1994, 1996; Jansen and Stam 1994). Here, we give a derivation based on an EM algorithm.

Differentiating the log of likelihood with respect to b^* and setting the differentiation to zero give

$$\frac{\partial \ln L}{\partial b^*} = \sum_{j=1}^{n} \frac{p_{1j}\phi([y_j - \mu - b^*]/\sigma)}{p_{1j}\phi([y_j - \mu - b^*]/\sigma) + p_{0j}\phi([y_j - \mu]/\sigma)} \frac{[y_j - \mu - b^*]}{\sigma^2} = 0$$

$$\Longrightarrow \sum_{j=1}^{n} P_j(y_j - \mu - b^*) = 0$$

$$\Longrightarrow \hat{b}^* = \sum_{j=1}^{n}(y_j - \mu)P_j / \sum_{j=1}^{n} P_j \tag{6.2}$$

where

$$P_j = \frac{p_{1j}\phi([y_j - \mu - b^*]/\sigma)}{p_{1j}\phi([y_j - \mu - b^*]/\sigma) + p_{0j}\phi([y_j - \mu]/\sigma)}. \tag{6.3}$$

Differentiating the log-likelihood with respect to μ gives:

$$\frac{\partial \ln L}{\partial \mu} = \sum_{j=1}^{n}[P_j(y_j - \mu - b^*) + (1 - P_j)(y_j - \mu)]/\sigma^2 = 0$$

$$\Longrightarrow \hat{\mu} = \sum_{j=1}^{n}(y_j - P_j\hat{b}^*)/n. \tag{6.4}$$

Differentiating the log-likelihood with respect to σ^2 gives

$$\frac{\partial \ln L}{\partial \sigma^2} = \sum_{j=1}^{n}[P_j(y_j - \mu - b^*)^2 + (1 - P_j)(y_j - \mu)^2]/(2\sigma^4) - n/(2\sigma^2) = 0$$

$$\Longrightarrow \hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n}[(y_j - \mu)^2 - P_j b^{*2}]. \tag{6.5}$$

P_j can be called the posterior probability of $x_j^* = 1$, whereas p_{1j} is the prior probability and is a function of θ. So far, these solutions have been derived under the assumption that θ was known. If θ is regarded as unknown, the maximum likelihood estimate of θ is given by

$$\begin{aligned}
\frac{\partial \ln L}{\partial \theta} &= \sum_{j=1}^{n_2}\left[\frac{-\phi\left(\frac{y_j-\mu-b^*}{\sigma}\right) + \phi\left(\frac{y_j-\mu}{\sigma}\right)}{(1-\theta)\phi\left(\frac{y_j-\mu-b^*}{\sigma}\right) + \theta\phi\left(\frac{y_j-\mu}{\sigma}\right)}\right] \\
&\quad + \sum_{j=1}^{n_3}\left[\frac{\phi\left(\frac{y_j-\mu-b^*}{\sigma}\right) - \phi\left(\frac{y_j-\mu}{\sigma}\right)}{\theta\phi\left(\frac{y_j-\mu-b^*}{\sigma}\right) + (1-\theta)\phi\left(\frac{y_j-\mu}{\sigma}\right)}\right] \\
&= \sum_{j=1}^{n_2}\left[-\frac{P_j}{1-\theta} + \frac{1-P_j}{\theta}\right] + \sum_{j=1}^{n_3}\left[\frac{P_j}{\theta} - \frac{1-P_j}{1-\theta}\right] = 0
\end{aligned}$$

$$\Longrightarrow \hat{\theta} = \frac{\sum_{j=1}^{n_2}(1-\hat{P}_j) + \sum_{j=1}^{n_3}\hat{P}_j}{n_2+n_3} \tag{6.6}$$

These solutions are not in the closed form and each estimate depends on estimates of other parameters. So, numerical methods have to be used. This can be achieved by iterating the above equations via an EM algorithm (Expectation/Maximization) beginning with the initial estimate $\hat{b}^* = 0$ or the least squares estimates of b^* and μ using $x_j^* = p_{1j}$. In each iteration, the algorithm consists of one E-step, Equation (6.3), and three M-steps, Equations (6.2), (6.4), (6.5) and (6.6). This process is iterated until convergence of estimates.

6.3 Likelihood ratio test statistic

The test statistic is constructed using the LOD score

$$\text{LOD} = -\log_{10}\frac{L(\hat{\hat{\mu}}, b^* = 0, \hat{\hat{\sigma}}^2)}{L(\hat{\mu}, \hat{b^*}, \hat{\sigma}^2)} \tag{6.7}$$

under the hypotheses

$$H_0: \ \ b^* = 0 \ \text{ and } \ H_1: \ \ b^* \neq 0$$

assuming that the putative QTL was located at the point (θ) of consideration, where $\hat{\mu}$, $\hat{b^*}$, $\hat{\sigma}^2$ are the maximum likelihood estimates of μ, b^*, σ^2 under H_1, and $\hat{\hat{\mu}}$, $\hat{\hat{\sigma}}^2$ are the estimates of μ, σ^2 under H_0 with b^* constrained to zero.

Note that the LOD score test is essentially the same test as the usual likelihood ratio test

$$\text{LR} = -2\ln\frac{L(\hat{\hat{\mu}}, b^* = 0, \hat{\hat{\sigma}}^2)}{L(\hat{\mu}, \hat{b^*}, \hat{\sigma}^2)}$$

Thus

$$\text{LOD} = \frac{1}{2}(\log_{10} e)\text{LR} = 0.217\text{LR}$$

This test can be performed at any position covered by markers and thus the method creates a systematic strategy of searching for QTL. The amount of support for a QTL at a particular map position is often displayed graphically through the use of likelihood maps or profile, which plot the likelihood ratio test statistic (or a closely related quantity) as a function of map position of the putative QTL. If the LOD score at a region exceeds a pre-defined critical threshold, a QTL is indicated at the neighborhood of the maximum of the LOD score with the width of the neighborhood defined by one or two LOD support interval (Lander and Botstein 1989). By the property of the maximum likelihood analysis, the estimates of locations and effects of QTL are asymptotically unbiased if the assumption that there is at most one QTL on a chromosome is true.

6.4 Threshold determination

The test statistic LR for a given position is expected to be asymptotically chi-square distributed with one degree of freedom under the null hypothesis for the backcross design, *i.e.* LR $\sim \chi_1^2$ thus LOD $\sim 0.217\chi_1^2$, and with two degrees of freedom for the F_2 design (Lander and Botstein 1989; van Ooijen 1992; Zeng 1994). However, because the test is usually performed in the whole genome, there is a multiple testing problem, and the distribution of the maximum LR or LOD score over the whole genome under the null hypothesis becomes very complicated. An asymptotic theory based on an Orenstein-Uhlenbeck diffusion process for determining appropriate genome-wise critical values has been developed by Lander and Botstein (1989), Feingold et al. (1993) and Lander and Schork (1994). For loosely linked markers, the analysis of Zeng (1994) and Rebai et al (1994) is relevant. Lander and Botstein (1989) suggested that a typical LOD score threshold should be between 2 and 3 to ensure a 5% overall false positive error for detecting QTL.

- Taking backcross design as an example: At a particular marker, the square of t test statistic corresponds to

$$t^2 = \left[\frac{\tilde{\mu}_1 - \tilde{\mu}_0}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_0}\right)}}\right]^2 \longrightarrow \chi_1^2 = z^2 \longleftarrow \text{LR} \quad \text{as } n \to \infty$$

  where $z \sim N(0,1)$.

- At a particular point within a marker interval (*i.e.* for fixed θ)

$$\text{LR} \longrightarrow \chi_1^2 \quad \text{as } n \to \infty$$

  So the threshold of LR approaches to

$$\text{LR}_\alpha \longrightarrow \chi_{1,\alpha}^2 \quad \text{as } n \to \infty$$

- The maximum of LR for the whole marker interval: The distribution of max[LR($\theta, 0 < \theta < 1$)] is between χ_1^2 and χ_2^2, more close to χ_1^2 for relatively small interval (say $<$ 10cM).

$$\chi_{1,\alpha}^2 < \text{LR}_\alpha < \chi_{2,\alpha}^2$$

- What is the distribution and appropriate threshold for max[LR($x, x \in$ genome)] ? This threshold is called experimental-wise threshold.

  Need to find the value of T_α such that if there is no QTL in the whole genome, the chance of max[LR] exceeding T_α *somewhere* in the genome is α *i.e.*

$$\alpha = Prob[\max[\text{LR}(x, x \in \text{genome})] > T_\alpha] \quad \text{under } \text{H}_0$$

  Two special cases:

1. The sparse-map case: If markers are sparse and widely separately, we can regard marker intervals approximately independent. For M independent intervals,

$$\begin{aligned} 1-\alpha &= Prob(\text{no error in the genome}) \\ &= \prod_{i=1}^{M} Prob(\text{no error in interval } i) \\ &= \prod_{i=1}^{M}(1-p) = (1-p)^M \approx 1-Mp \end{aligned}$$

$$\Longrightarrow p \approx \alpha/M$$

$$\chi^2_{1,\alpha/M} < T_\alpha < \chi^2_{2,\alpha/M}$$

2. Dense-map case: Suppose we have markers everywhere, at each position

$$LR \to z^2 \quad \text{with } z \sim N(0,1)$$

For two positions, LR(x_1) and LR(x_2) are correlated, as

$$Corr(z(x_1), z(x_2)) = 1-2r$$

Lander and Botstein's Proposition:

Consider an organism with C chromosomes and genetic length G, measured in Morgans. When no QTL are present, the probability that the LOD score exceeds a high level T is approximately $(C+2Gt)\chi^2[t]$, where $T=(2\ln 10)^{-1}t$ and $\chi^2[t]$ denotes the inverse cumulative distribution function of the χ^2 distribution with 1 d.f. In order to make the probability less than α that a false positive occurs somewhere in the genome, the appropriate LOD threshold is thus approximately $T_\alpha=(2\ln 10)^{-1}t_\alpha$ where t_α solves the equation $\alpha=(C+2Gt_\alpha)\chi^2[t_\alpha]$.

- **Bottom line**: For many organisms, LOD threshold is between 2 and 3. For mouse with 20 chromosomes, it is close to 3. For maize with 10 chromosomes, it is close 2.7. The threshold for F_2 design should be higher than that for backcross design.

6.5 Permutation test

Churchill and Doerge (1994) and Doerge and Churchill (1996) proposed a data-based numerical method, based on the concept of a permutation test, to estimate empirical critical values for mapping QTL for a given data set. The method proceeds as follows:

1. Randomly pair an individual marker genotypes with another individual trait phenotype to generate a permuted sample of the data (to simulate the null hypothesis of no association between genotype and phenotype).

2. Perform interval mapping analysis on the permuted sample.

3. Do it for a number of times to obtain an empirical distribution of the test statistic at the null hypothesis and from it to determine a 95% significance value for the test in the original data analysis.

4. If the test statistic in the original data in a genomic region is higher than this critical value, a QTL is declared.

The critical values obtained by permutation reflect the specifics of the experiment (*e.g.* sample size, experimental design, missing data, etc.). This method is flexible and can be used to construct empirical critical values for other mapping methods discussed below as well.

More specifically, let $\mathbf{X} = \{X_1, X_2, \cdots, X_n\}$ denote a random sample of size n, which is the data, from a population. For QTL mapping, each observation point, X_j, is composed by a (or multiple) trait value, y_j, and marker genotypes, $\mathbf{M}_j = \{M_{1j}, M_{2j}, \cdots, M_{tj}\}$, for t markers, *i.e.*

$$X_j = \{y_j, \mathbf{M}_j\}$$

A permuted sample of size n, denoted as $\mathbf{X}' = \{X'_1, X'_2, \cdots, X'_n\}$, is a sample of random match between y_i and $\mathbf{M}_j$ from the original sample without replacement, such that

$$X'_k = \{y'_k, \mathbf{M}'_k\} = \{y_i, \mathbf{M}_j\}$$

with $k = i$, $i = 1, 2, \cdots, n$, with probability $1/n$ and $k = j$, $j = 1, 2, \cdots, n$, with probability $1/n$ independently.

In permuted samples $\mathbf{X}'$, y' and $\mathbf{M}'$ do not have any intrinsic relationship, thus simulating the null hypothesis of no QTL.

For N permuted samples, let LR'_p be the maximum likelihood test statistic for a particular genomic position, or maximum value of the test statistic for a marker interval or for the whole genome in the pth permuted sample. The $\alpha \times 100\%$ threshold of the test statistic under the null hypothesis can be estimated empirically as

$$\hat{T}_\alpha = \alpha \times 100 \text{ percentile of } \{LR'_1, LR'_2, \cdots, LR'_N\}$$

This value will be used to compare with the test statistic to declare the mapping of a QTL.

How large should N be for practical data analysis? Churchill and Doerge (1994) suggest that for $\alpha = 0.05$, N should be at least 1000, and for $\alpha = 0.01$, N should be at least 5000.

6.6 Estimating sampling variance

Given an observed random sample $\mathbf{X}$, we wish to estimate a parameter of interest, θ, on the basis of $\mathbf{X}$. For this purpose, we calculate an estimate $\hat{\theta} = S(\mathbf{X})$ from $\mathbf{X}$. Now we want to know how accurate $\hat{\theta}$ is, its standard error (SE) or confidence interval (CI).

For estimating a confidence interval of QTL position, Lander and Botstein (1989) proposed to use the one LOD support interval (a standard procedure in human genetic analysis), which is based on the asymptotical χ^2 distribution of the test statistic under the null hypothesis that the maximum likelihood estimate of QTL position is the correct position. Mangin, Goffinet and Rebai (1994) argued that this method is appropriate when the QTL effect is large. When the QTL effect is small, the test statistic does not follow a χ^2 distribution, and as a result the one LOD support interval underestimates the confidence interval. They then devised an approximate method to take QTL effect into account in constructing a confidence interval. Kao and Zeng (1997) have worked out the procedures of using Fisher's information matrix to estimate sampling variances of estimates of QTL positions and effects under the mixture model framework for both interval mapping and composite interval mapping.

6.7 Bootstrap estimate of sampling variance

Bootstrap samples are generally created by sampling with replacement n individual observations. An observation consists of a trait phenotype and marker genotypes. At each bootstrap sample, n observations are sampled with replacement out of the pool of the n original observations. Some records can appear more than once in a bootstrap sample, while others are not included at all. Estimation on the parameters is performed for each bootstrap sample, and the mean, standard error, confidence interval of estimates can be estimated from N bootstrap samples. For example, the empirical central 95% confidence interval of a QTL position can be determined by ordering the N estimates and taking the bottom and top 2.5th percentile of the N bootstrap estimates, respectively. Visscher, Thompson and Haley (1996) studied the use of bootstrap samples to estimate confidence interval of QTL position.

More specifically, a bootstrap sample is a random sample of size n, denoted as $\mathbf{X}^* = \{X_1^*, X_2^*, \cdots, X_n^*\}$, drawn with replacement from the original sample $\mathbf{X}$, such that

$$X_i^* = X_j, \quad \text{for } i = j, j = 1, 2, \cdots, n \text{ with probability } 1/n$$

The method proceeds as follows:

1. Select N independent bootstrap samples, $\mathbf{X}_1^*, \mathbf{X}_2^*, \cdots, \mathbf{X}_N^*$, each of size n drawn with replacement from $\mathbf{X}$.

2. Estimate parameter θ from the bth bootstrap sample

$$\hat{\theta}_b^* = S(\mathbf{X}_b^*) \quad \text{for } b = 1, 2, \cdots, N$$

3. Estimate the standard error, $SE(\hat{\theta})$, of $\hat{\theta}$ by the sample standard deviation of the bootstrap estimates, $\hat{\theta}_b^*$, from the N replications

$$SE(\hat{\theta})_B = \left[\sum_{b=1}^{N} [\hat{\theta}_b^* - \hat{\bar{\theta}}^*]^2/(N-1)\right]^{1/2}$$

where $\hat{\bar{\theta}}^* = \sum_{b=1}^{N} \hat{\theta}_b^*/N$.

4. Estimate the 95% confidence interval, $CI(\hat{\theta})$, of $\hat{\theta}$ by

$$CI(\hat{\theta})_B = \text{2.5th and 97.5th percentiles of } \{\hat{\theta}_1^*, \hat{\theta}_2^*, \cdots, \hat{\theta}_N^*\}$$

6.8 Variance explained by QTL

Sometimes the magnitude of a QTL is also reported as the proportion of the variance explained by the QTL (σ^2_{exp}). This is usually estimated as

$$\hat{\sigma}^2_{exp} = \frac{\hat{\sigma}^2_{tot} - \hat{\sigma}^2_{reduce}}{\hat{\sigma}^2_{tot}}$$

where $\hat{\sigma}^2_{tot}$ is an estimate of the total phenotypic variance ($\hat{\bar{\sigma}}^2$ of (6.7) at the null hypothesis) and $\hat{\sigma}^2_{reduce}$ is an estimate of the residual variance of the IM model ($\hat{\sigma}^2$ of (6.7) at the alternative hypothesis).

However, this estimate is not additive for multiple QTL indicated by the analysis and usually overestimates the variance explained by a QTL, because the mapping analysis is not independent for different regions of the genome. Sometimes, $\hat{\sigma}^2_{exp}$ for different QTL can add up more than 100%. A more appropriate estimate of variances and covariances explained by different QTL can be obtained through multiple interval mapping discussed below.

6.9 Haley-Knott regression approximation

A simplified approximation of model (6.1) has been proposed by Haley and Knott (1992) and Martinez and Curnow (1992). Instead of treating x_j^* as missing data and using a mixture model via maximum likelihood for missing data analysis, this approximation uses $p_{1j} = Prob(x_j^* = 1|M_i, M_{i+1}, \theta)$ in the place of x_j^* and simplifies model (6.1) to

$$y_j = \mu + b^* p_{1j} + e_j \qquad j = 1, 2, \cdots, n \tag{6.8}$$

Since this is a simple regression model, statistical analysis becomes straightforward. Haley and Knott (1992) and Rebai et al. (1995) have shown that this procedure gives a very good approximation of the likelihood profile for ML interval mapping. Xu (1995) notes that this regression approach tends to overestimates the residual variance, and presents a correction.

6.10 Advantages and disadvantages

Compared with one marker analysis, the interval mapping method has several advantages. These include:

1. The probable position of the QTL can be inferred by the support interval;

2. The estimated locations and effects of QTL tend to be asymptotically unbiased if there is only one segregating QTL on a chromosome;

3. The method requires fewer individuals than one marker analysis for the detection of QTL.

There are, however, still many problems with interval mapping. These include:

1. The test is not an interval test (a test which could distinguish whether or not there is a QTL within a defined interval and should be independent of the effects of QTL that are outside a defined region). Even when there is no QTL within an interval, the likelihood profile on the interval can still exceed the threshold significantly if there is a QTL at some nearby region on the chromosome. If there is only one QTL on a chromosome, this effect, though undesirable, may not matter because the QTL is more likely to be located at the region which shows the maximum likelihood profile. However, the number of QTL on a chromosome is unknown.

2. If there is more than one QTL on a chromosome, the test statistic at the position being tested will be affected by all those QTL and the estimated positions and effects of "QTL" identified by this method are likely to be biased.

3. It is not efficient to use only two markers at a time to do the test, as the information from other markers is not utilized.

Recognizing these problems, Lander and Botstein (1989) proposed to extend the method to analyze multiple markers for multiple QTL simultaneously by introducing more b^* and x_j^* in the model (6.1), which is the approach that will be discussed later. There are, of course, a number of issues that are involved with this approach and need to be solved for the successful implementation and application of the method.

6.11 Examples

- Figure 6.1 shows the result of interval mapping on the mouse data. The result of interval mapping at each marker should correspond to one marker analysis. Basically, interval mapping analysis just extends one marker analysis to a marker interval bracketed by two markers.

- Figure 6.2 shows the result of interval mapping on the maize data.

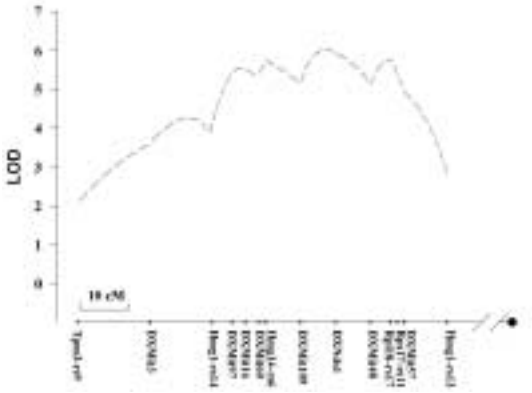

Figure 6.1: Interval mapping of the mouse data

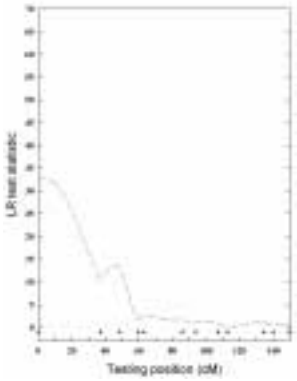

Figure 6.2: Interval mapping of the maize data

Chapter 7

Composite Interval Mapping

Most of the single-QTL methods developed can be extended to multiple QTL by conditioning additional marker loci and using conditional probabilities for multilocus genotypes. This approach has been used to develop explicit models for two or three linked QTL (*e.g.,* Knapp 1991; Haley and Knott 992; Martinez and Curnow 1992, 1994; Jansen 1996; Satagopan et al. 1996). Kearsey and Hyne (1994), Hyne and Kearsey (1995), Wu and Li (1994, 1996) also proposed a very simple regression based method that simultaneously considers all the markers on a single chromosome for locating multiple linked QTL. Wright and Mowers (1994) and Whittaker et al. (1996) also have showed how position information for linked QTL can be extracted from the regression coefficients of a standard multiple regression incorporating several linked markers. See Lynch and Walsh (1998) for detailed discussion on these topics. Here we introduce a composite interval mapping discussed in Zeng (1993, 1994) and Jansen and Stam (1994).

7.1 Properties of multiple regression analysis

Ideally, when we test an interval for a QTL, we would like to have our test statistic independent of the effects of possible QTL at other regions of the chromosome. If such a test can be formulated, we can simplify mapping for multiple QTL from a multiple dimensional search problem to a one dimensional search problem, as the test for each interval is independent and for each marker interval we can effectively consider the possibility of the presence of only a single QTL. This test can be constructed by using a combination of interval mapping with multiple regression. Largely because of linear structures of locations of genes on chromosomes, multiple regression analysis has a very important property that the partial regression coefficient of a trait on a marker is expected to depend only on those QTL which are located on the interval bracketed by the two neighboring markers and to be independent of any other QTL, if there is no crossing-over interference and no epistasis. Interference and epistasis will introduce non-linearity in the model. This is the basis of the composite interval mapping method proposed by Zeng (1994) to improve the precision and efficiency of mapping multiple QTL.

Before we go to the composite interval mapping method, let us briefly review some relevant theory in multiple regression analysis for QTL mapping (Zeng 1993). Suppose we regress trait values y on t markers observed in B_1 populations

$$y_j = \mu + \sum_{k=1}^{t} b_k x_{jk} + e_j, \tag{7.1}$$

where x_{jk} is the value (1 or 0) of the kth marker in the jth individual, and b_k (also denoted by $b_{yk.s_k}$ where s_k denotes a set which includes all markers except the kth marker) is the partial regression coefficient of the phenotype y on the kth marker conditional on all other markers. Since x_{jk} takes a value of 1 or 0 with equal probability, the variance of the kth marker in the population is

$$\sigma_k^2 = 1/4.$$

It is easy to show that the covariance between the ith and kth markers is

$$\sigma_{ik} = (1 - 2r_{ik})/4.$$

and the covariance between the trait value, y, and the kth marker is

$$\sigma_{yk} = \sum_{u=1}^{m} (1 - 2r_{uk})\delta_u/4.$$

With these basic equations, any conditional variance and covariance can be derived. For example, the variance of marker k conditional on marker i is

$$\begin{aligned} \sigma_{k\cdot i}^2 &= \sigma_k^2 - \sigma_{ik}^2/\sigma_i^2 \\ &= [1 - (1 - 2r_{ik})^2]/4 \\ &= r_{ik}(1 - r_{ik}). \end{aligned}$$

The covariance between markers i and k conditional on marker l is

$$\begin{aligned} \sigma_{ik\cdot l} &= \sigma_{ik} - \sigma_{il}\sigma_{kl}/\sigma_l^2 \\ &= [(1 - 2r_{ik}) - (1 - 2r_{il})(1 - 2r_{kl})]/4 \\ &= \begin{cases} 0 & \text{for order } ilk \text{ or } kli \\ r_{kl}(1 - r_{kl})(1 - 2r_{ik}) & \text{for order } ikl \text{ or } lki \\ r_{il}(1 - r_{il})(1 - 2r_{ik}) & \text{for order } lik \text{ or } kil \end{cases} \end{aligned}$$

because without interference

$$(1 - 2r_{ik}) = (1 - 2r_{il})(1 - 2r_{kl}) \quad \text{for order } ilk \text{ or } kli.$$

This shows that, conditional on an intermediate marker, the covariance between two flanking markers (or QTL) is expected to be zero. This is the foundation for composite interval mapping. From this property, Zeng (1993) showed that

$$b_{yk\cdot s_k} = \sum_{k-1<u\le k} \frac{r_{(k-1)u}(1 - r_{(k-1)u})(1 - 2r_{uk})}{r_{(k-1)k}(1 - r_{(k-1)k})} a_u + \sum_{k<u<k+1} \frac{r_{u(k+1)}(1 - r_{u(k+1)})(1 - 2r_{ku})}{r_{k(k+1)}(1 - r_{k(k+1)})} a_u$$

where the first summation is for all QTL located between markers $k-1$ and k and the second summation is for all QTL located between markers k and $k+1$. This regression coefficient depends only on those QTL which are located between markers $k-1$ and $k+1$. This is a very desirable property. By using this property we can create an *interval test* in which we can test whether there is a QTL within a marker interval.

There are also other properties of the multiple regression that have direct relevance to QTL mapping. These are summarized as follows:

Property 1 *In the multiple regression analysis, assuming additivity of QTL effects between loci (i.e., ignoring epistasis), the expected partial regression coefficient of the trait on a marker depends only on those QTL which are located on the interval bracketed by the two neighboring markers, and is unaffected by the effects of QTL located on other intervals.* This property essentially says that a conditional (interval) test can be constructed based on the partial regression coefficient and such a test would test the linkage effect of only those QTL which are located within the defined interval.

Property 2 *Conditioning on unlinked markers in the multiple regression analysis will reduce the sampling variance of the test statistic by controlling some residual genetic variation and thus will increase the power of QTL mapping.* This means that even unlinked markers contain useful information which can be used to increase the statistical power of the test and the efficiency of the genetic mapping. This useful information has not been utilized in the current QTL mapping methods.

Property 3 *Conditioning on linked markers in the multiple regression analysis will reduce the chance of interference of possible multiple linked QTL on hypothesis testing and parameter estimation, but with a possible increase of sampling variance.* The first part of the sentence restates Property 1, and the second part of the sentence says that an interval test may entail a loss in the statistical power of the test because the test is a conditional test. This summarizes the advantage and disadvantage of the interval test: that is, there is a trade-off between precision and efficiency of mapping by using an interval test. Effective balance on these two issues will be the major consideration in practical mapping of QTL.

Property 4 *Two sample partial regression coefficients of the trait value on two markers in a multiple regression analysis are generally uncorrelated unless the two markers are adjacent markers.* This is related to the correlation between two test statistics in two intervals for an interval test. It has been shown that, for an interval test, a test statistic on an interval is generally asymptotically uncorrelated to the test statistic on another interval unless two intervals are adjacent intervals. Even when the two intervals are adjacent intervals, the correlation between two test statistics in two intervals is usually very small. This property is related to the issue of determining an appropriate critical value of a test statistic under a null hypothesis for an overall test covering a whole genome.

7.2 Composite interval mapping Model

Composite interval mapping (CIM) is an extension of interval mapping with some selected markers also fitted in the model as cofactors to control the genetic variation of other possibly linked or unlinked QTL. Specifically, to test for a QTL on an interval between adjacent markers M_i and M_{i+1}, we extend model (6.1) to

$$y_j = \mu + b^* x_j^* + \sum_k b_k x_{jk} + e_j \tag{7.2}$$

where x_j^* refers to the putative QTL and x_{jk} refers to those markers selected for genetic background control. Appropriate selection of markers as cofactors is important and discussed below.

7.3 Likelihood analysis

In this case, the likelihood function is specified as

$$L(b^*, \mathbf{B}, \sigma^2) = \prod_{j=1}^{n} \left[p_{1j} \phi \left(\frac{y_j - \mathbf{X}_j\mathbf{B} - b^*}{\sigma} \right) + p_{0j} \phi \left(\frac{y_j - \mathbf{X}_j\mathbf{B}}{\sigma} \right) \right] \tag{7.3}$$

where $\mathbf{X}_j\mathbf{B} = \mu + \sum_k b_k x_{jk}$. The maximum likelihood estimates of the various parameters can be found in a similar way as for interval mapping. These are given below. For b^*:

$$\frac{\partial \ln L}{\partial b^*} = \sum_{j=1}^{n} \frac{p_{1j}\phi([y_j - \mathbf{X}_j\mathbf{B} - b^*]/\sigma)}{p_{1j}\phi([y_j - \mathbf{X}_j\mathbf{B} - b^*]/\sigma) + p_{0j}\phi([y_j - \mathbf{X}_j\mathbf{B}]/\sigma)} \frac{[y_j - \mathbf{X}_j\mathbf{B} - b^*]}{\sigma^2}.$$

Setting this derivative to zero provides

$$\sum_{j=1}^{n} P_j (y_j - \mathbf{X}_j\mathbf{B} - b^*) = 0$$

where

$$P_j = \frac{p_{1j}\phi([y_j - \mathbf{X}_j\mathbf{B} - b^*]/\sigma)}{p_{1j}\phi([y_j - \mathbf{X}_j\mathbf{B} - b^*]/\sigma) + p_{0j}\phi([y_j - \mathbf{X}_j\mathbf{B}]/\sigma)}. \tag{7.4}$$

This leads to the solution given by Zeng (1994) as

$$\begin{aligned} \hat{b}^* &= \sum_{j=1}^{n} (y_j - \mathbf{X}_j\hat{\mathbf{B}}) P_j / \sum_{j=1}^{n} P_j \\ &= (\mathbf{Y} - \mathbf{X}\mathbf{B})' \mathbf{P} / c \end{aligned} \tag{7.5}$$

where $c = \sum_{j=1}^{n} P_j$, $\mathbf{Y} = \{y_j\}_{n\times 1}$, $\mathbf{P} = \{P_j\}_{n\times 1}$, and a prime denotes transposition.

Differentiating the log-likelihood with respect to $\mathbf{B}$:

$$\frac{\partial \ln L}{\partial \mathbf{B}} = \sum_{j=1}^{n} [P_j \mathbf{X}_j'(y_j - \mathbf{X}_j\mathbf{B} - b^*) + (1 - P_j)\mathbf{X}_j'(y_j - \mathbf{X}_j\mathbf{B})]/\sigma^2.$$

Expressed in matrix notation, the equation $\partial \ln L/\partial \mathbf{B} = 0$ is

$$\mathbf{X}'(\mathbf{Y} - \mathbf{X}\mathbf{B}) = \mathbf{X}'\mathbf{P}b^*$$

$$\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{Y} - \mathbf{P}\hat{b}^*). \tag{7.6}$$

Differentiating the log-likelihood with respect to σ^2:

$$\frac{\partial \ln L}{\partial \sigma^2} = \sum_{j=1}^{n} [P_j(y_j - \mathbf{X}_j\mathbf{B} - b^*)^2 + (1 - P_j)(y_j - \mathbf{X}_j\mathbf{B})^2]/(2\sigma^4) - n/(2\sigma^2).$$

Setting this derivative to zero leads to the solution

$$n\hat{\sigma}^2 = (\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})'(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}) - \hat{b}^{*2}c. \tag{7.7}$$

7.4 Hypothesis test

The hypotheses to be tested are $H_0 : b^* = 0$ and $H_1 : b^* \neq 0$. The likelihood function under the null hypothesis is

$$L(b^* = 0, \mathbf{B}, \sigma^2) = \prod_{j=1}^{n} \phi\left(\frac{y_j - \mathbf{X}_j\mathbf{B}}{\sigma}\right)$$

with the maximum likelihood estimates

$$\hat{\hat{\mathbf{B}}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

$$\hat{\hat{\sigma}}^2 = (\mathbf{Y} - \mathbf{X}\hat{\hat{\mathbf{B}}})'(\mathbf{Y} - \mathbf{X}\hat{\hat{\mathbf{B}}})/n.$$

The likelihood ratio (LR) test statistic is

$$\text{LR} = -2\ln\frac{L(b^* = 0, \hat{\hat{\mathbf{B}}}, \hat{\hat{\sigma}}^2)}{L(\hat{b}^*, \hat{\mathbf{B}}, \hat{\sigma}^2)} \quad \text{or} \quad \text{LOD} = -\log_{10}\frac{L(b^* = 0, \hat{\hat{\mathbf{B}}}, \hat{\hat{\sigma}}^2)}{L(\hat{b}^*, \hat{\mathbf{B}}, \hat{\sigma}^2)} \tag{7.8}$$

Like Lander and Botstein's interval mapping, this test can be performed at any position in a genome covered by markers. Thus it gives a systematic strategy to search for QTL in a genome. As the test statistic is almost independent for each interval, a test on each interval is more likely to test for a single QTL only.

7.5 Analysis in an F_2 population

For an F_2 population, the mixture model equivalent to (7.2) can be specified as

$$y_j = \mu + b^* x_j^* + d^* z_j^* + \sum_k b_k x_{jk} + \sum_k d_k z_{jk} + e_j \qquad j = 1, 2, \cdots, n$$

where

$$\begin{aligned}
b^* &= \text{additive effect of the putative QTL} \\
d^* &= \text{dominance effect of the putative QTL} \\
x_j^* &= \begin{cases} 2 & \text{if the QTL genotype is } QQ \\ 1 & \text{if the QTL genotype is } Qq \\ 0 & \text{if the QTL genotype is } qq \end{cases} \\
z_j^* &= \begin{cases} 1 & \text{if the QTL genotype is } Qq \\ 0 & \text{if the QTL genotype is } QQ \text{ or } qq \end{cases}
\end{aligned}$$

Let

$$p_{kj} = Prob(x_j^* = k | M_i, M_{i+1}, \theta) \qquad k = 0, 1, 2.$$

which is specified below.

	QTL genotype		
Marker genotype	QQ (2)	Qq (1)	qq (0)
$M_1M_1M_2M_2$	1	0	0
$M_1M_1M_2m_2$	$1-\theta$	θ	0
$M_1M_1m_2m_2$	$(1-\theta)^2$	$2\theta(1-\theta)$	θ^2
$M_1m_1M_2M_2$	θ	$1-\theta$	0
$M_1m_1M_2m_2$	$\eta\theta(1-\theta)$	$1-2\eta\theta(1-\theta)$	$\eta\theta(1-\theta)$
$M_1m_1m_2m_2$	0	$1-\theta$	θ
$m_1m_1M_2M_2$	θ^2	$2\theta(1-\theta)$	$(1-\theta)^2$
$m_1m_1M_2m_2$	0	θ	$1-\theta$
$m_1m_1m_2m_2$	0	0	1

$\theta = r_{M_1Q}/r_{M_1M_2}$, $\eta = r_{M_1M_2}^2/[(1-r_{M_1M_2})^2 + r_{M_1M_2}^2]$, where r_{M_1Q} is the recombination frequency between marker M_1 and QTL Q, and $r_{M_1M_2}$ is the recombination frequency between markers M_1 and M_2. Double recombination is ignored.

The likelihood function is then given by

$$L(b^*, d^*, \mathbf{B}, \sigma^2) = \prod_{j=1}^{n} \left[p_{2j}\phi\left(\frac{y_j - \mathbf{X}_j\mathbf{B} - 2b^*}{\sigma}\right) + p_{1j}\phi\left(\frac{y_j - \mathbf{X}_j\mathbf{B} - b^* - d^*}{\sigma}\right) + p_{0j}\phi\left(\frac{y_j - \mathbf{X}_j\mathbf{B}}{\sigma}\right)\right]$$

The maximum likelihood estimates of various parameters are

$$\hat{b}^* = (\mathbf{Y} - \mathbf{XB})'\mathbf{P}_2/(2\mathbf{1}'\mathbf{P}_2)$$

$$\hat{d}^* = (\mathbf{Y} - \mathbf{XB})'\mathbf{P}_1/(\mathbf{1}'\mathbf{P}_1) - \hat{b}^*$$

$$\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{Y} - (2\mathbf{P}_2 + \mathbf{P}_1)\hat{b}^* - \mathbf{P}_1\hat{d}^*)$$

$$\hat{\sigma}^2 = \frac{1}{n}[(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})'(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}) - 4(\mathbf{1}'\mathbf{P}_2)\hat{b}^{*2} - (\mathbf{1}'\mathbf{P}_1)(\hat{b}^* + \hat{d}^*)^2]$$

$$P_{2j} = \frac{p_{2j}\phi([y_j - \mathbf{X}_j\mathbf{B} - 2b^*]/\sigma)}{p_{2j}\phi([y_j - \mathbf{X}_j\mathbf{B} - 2b^*]/\sigma) + p_{1j}\phi([y_j - \mathbf{X}_j\mathbf{B} - b^* - d^*]/\sigma) + p_{0j}\phi([y_j - \mathbf{X}_j\mathbf{B}]/\sigma)}$$

$$P_{1j} = \frac{p_{1j}\phi([y_j - \mathbf{X}_j\mathbf{B} - b^* - d^*]/\sigma)}{p_{2j}\phi([y_j - \mathbf{X}_j\mathbf{B} - 2b^*]/\sigma) + p_{1j}\phi([y_j - \mathbf{X}_j\mathbf{B} - b^* - d^*]/\sigma) + p_{0j}\phi([y_j - \mathbf{X}_j\mathbf{B}]/\sigma)}$$

The test statistic is

$$\text{LOD} = -\log_{10}\frac{L(b^* = 0, d^* = 0, \hat{\hat{\mathbf{B}}}, \hat{\hat{\sigma}}^2)}{L(\hat{b}^*, \hat{d}^*, \hat{\mathbf{B}}, \hat{\sigma}^2)}$$

under the hypotheses

$$H_0: \ \ b^* = 0, d^* = 0 \ \text{ and } \ H_1: \ \ b^* \neq 0, d^* \neq 0$$

The maximum likelihood analysis looks like complicated and depends on the genetic model and experimental design specified. However, general formulae that apply to different genetic models and experimental designs have been derived by Kao and Zeng (1996) and are particularly useful for practical data analysis.

7.6 A simulation example

Some features of CIM are illustrated in Figure 7.1 which is taken from Zeng (1994). This figure shows some results from a simulated data set. Four "chromosomes" each with sixteen markers separated in fifteen 10 cM intervals were simulated for a backcross population. The trait is affected by 10 QTLs with positions and effects depicted in Figure 7.1. Together the QTLs account for 70% of the phenotypic variance in a backcross population. Sample size is 300. The trait value of an individual is determined by the sum of effects of the QTLs which the individual possesses, plus a random (environmental) variable which is normally distributed with mean zero and variance scaled to give the expected 0.7 heritability of the population.

Three models (I, II and III) were fitted to the data. In model I, all other markers are combined with interval mapping in the model for genetic background control. In model II, only unlinked markers are used for genetic background control. Model III is interval mapping analysis. It can be seen that the test statistic under model I (composite interval mapping) on an interval is almost independent of that on other intervals. Because of this property model I correctly identified six largest QTL with relatively high precision. Like model II, model III does not have the property of the interval test, so that the test statistics under models II and III, affected by all those linked QTL, tend to be biased on mapping individual QTL. However, model II has higher power on identifying marker-QTL association than models I and III (because of Property 2].

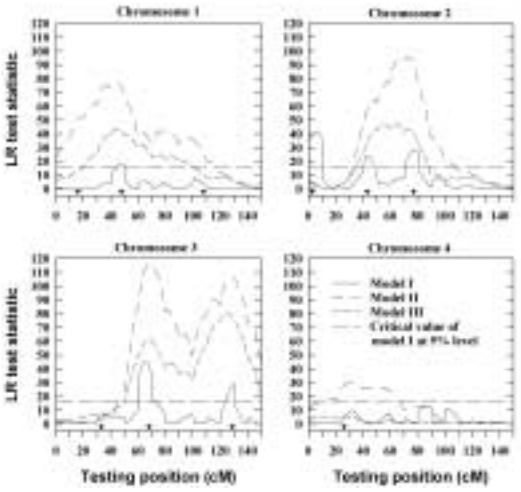

Figure 7.1: Comparison of interval mapping and composite interval mapping in a simulated data set

7.7 Marker selection

Which markers should be added? There is no simple solution for this question, as the question depends on the number and positions of underlying QTL, the information that is not available *a prior*. Too few markers selected may not achieve the purpose of reducing the most residual genetic variation, and too many markers selected may reduce the power of the analysis.

In QTL Cartographer (http://statgen.ncsu.edu/qtlcart/cartographer.html), we implemented a two step procedure for practical data analysis using composite interval mapping method. In the first step, n_p markers that are significantly associated with the trait are selected by (forward or backward) stepwise regression. In the second step (mapping step), for each testing interval, except of the putative QTL two markers that are at least W_s cM away from the testing interval (one for each direction) are first picked up to fit in the model to define a testing window for blocking other possible linked QTL effects on the test. Then, those selected n_p markers that are outside of the testing window are also fitted into the model to reduce the residual variance.

By changing the values of n_p and W_s, different conditions for composite interval mapping can be created. Generally n_p should be much significantly smaller than n, not exceeding $2\sqrt{n}$ (Jansen and Stam 1994), or alternatively it can be determined automatically by F-to-enter or F-to-drop criterion in the forward or backward stepwise regression analysis. W_s should at least 10 or 15 cM depending on sample size.

Rule of thumb:

- n_p can be decided based on the stepwise regression analysis of *SRmapqtl* using F-to-enter (forward) or F-to-drop (backward) statistic with significance level $\alpha = 0.01$.

- w_s should be as large as possible when there is no indication of other linked QTL; otherwise w_s can be gradually decreased as long as the test statistic for a putative QTL is significant.

7.8 Examples

7.8.1 Example 1

Interval mapping and composite interval mapping on chromosome X of the mouse data (Dragani *et al.* 1995 *Mammalian Genome* 6:778-781).

- The experiment is a backcross design and has 181 microsatellite markers (SSR, simple sequence repeats) typed on 103 individuals. The markers are distributed in 20 chromosomes, including 14 markers in chromosome X. The trait is 12 weeks body weight.

- The procedure of composite interval mapping: The boundary markers x^L and x^R are chosen to be the closest markers which are at least 10cM away from the testing interval. Besides x^L and x^R, 20 other linked or unlinked markers are also selected as cofactors in analysis by stepwise regression to absorb the effects of other QTL.

- The analysis of interval mapping indicates the existence of QTL on the chromosome because the LOD score is significant for a wide region (the threshold is 3.3 for the experimental design). However, not all significant peaks could be interpreted as QTL because of linkage effects, the "ghost" gene phenomenon and statistical sampling effects. The fact that a very wide region shows significant and comparable effects could suggest multiple QTL.

- The LOD score of the composite interval mapping analysis shows two distinct major peaks. This suggests that there are at least two body weight QTL on chromosome X in the mouse genome. One named as *Bw1* is mapped near marker *Rp18-rs11* and the other *Bw2* near *DXMIT60* (Dragani *et al.* 1995 *Mammalian Genome* 6:778-781). The two QTL together explain 25% of the phenotypic variance in the mapping population. In this case, the composite interval mapping analysis achieved a much better resolution in mapping QTL.

7.8.2 Example 2

Interval mapping and composite interval mapping on the maize data.

- The procedure of composite interval mapping: The boundary markers x^L and x^R are chosen to be the closest markers which are at least 10cM away from the testing interval. For each boundary marker, two variables (additive and dominance effects)

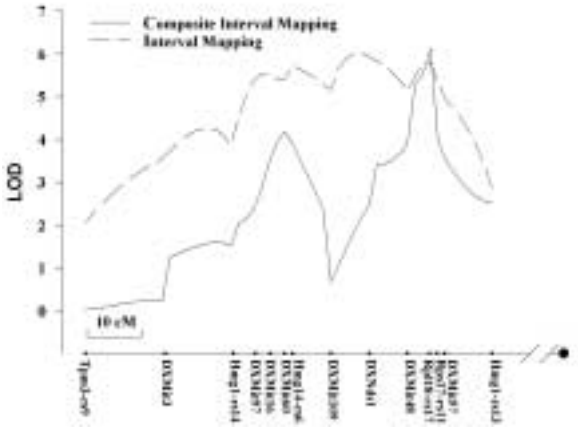

Figure 7.2: Composite interval mapping of the mouse data

are fitted in the model. Besides x^L and x^R, 27 markers from other 9 chromosomes are also selected as cofactors in analysis (each with one variable, additive effect) by stepwise regression to absorb the effects of other QTL.

- The interval mapping analysis shows a very significant peak on the first interval which indicates that there is a QTL there or on the left of the first marker.
- The composite interval mapping analysis strongly supports the mapping on the first interval, and also indicates that there is a second QTL around the third marker. The interval mapping analysis also shows a bump in the same region, but fails to reach the significance. However, even if LR under interval mapping reached significance, it may not be regarded as evidence for the second QTL because of possible linkage effect from the first QTL in the analysis. In this case, the composite interval mapping analysis effectively separated the effects in two regions, magnified and established the evidence of two QTL in the two regions.

7.8.3 Example 3

Composite interval mapping analysis of a Drosophila data set (Zeng *et al.* 2000 *Genetics* 154:299-310).

- The experiment is a two-way backcross (B_1 and B_2) between two Drosophila species, *Drosophila. simulans* and *D. mauritiana*.
- The sample size is 491 for B_1 and 474 for B_2.
- The marker number is 45, distributed on the three major chromosomes: X, II and III.

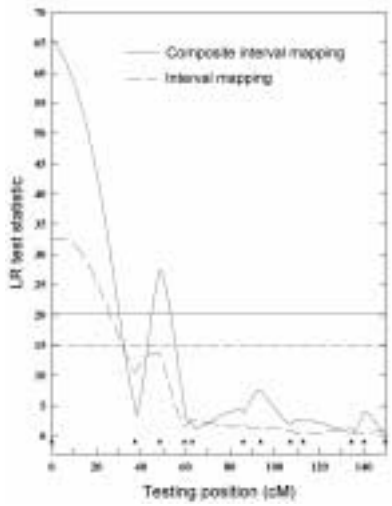

Figure 7.3: Composite interval mapping of the maize data

- The trait is morphological shape in the male genitalia measured by a morphometric descriptor based on elliptical Fourier and principal components analyses. The two species differ more than 30 environmental standard deviations on the measurement.

- Procedure of composite interval mapping: Because of large sample size, well separated marker coverage, and potentially large number of QTL, the analysis on an interval simply includes all other linked and unlinked markers, except of the flanking markers of the interval, as cofactors. Analysis was performed on the two backcrosses separately and also jointly (see the section of multiple trait analysis in Mapping Quantitative Trait Loci II).

- The composite interval mapping analysis reveals: 2 QTL on chromosome X, 4 on chromosome II, 7-8 on chromosome III (two backcrosses show different mapping results on the 70-100 cM region and indicate 2-3 QTL in the region). Together these QTL explain about 90% of the phenotypic variances in the populations.

  The interval mapping analysis (not shown) is not informative at all on the number and positions of QTL.

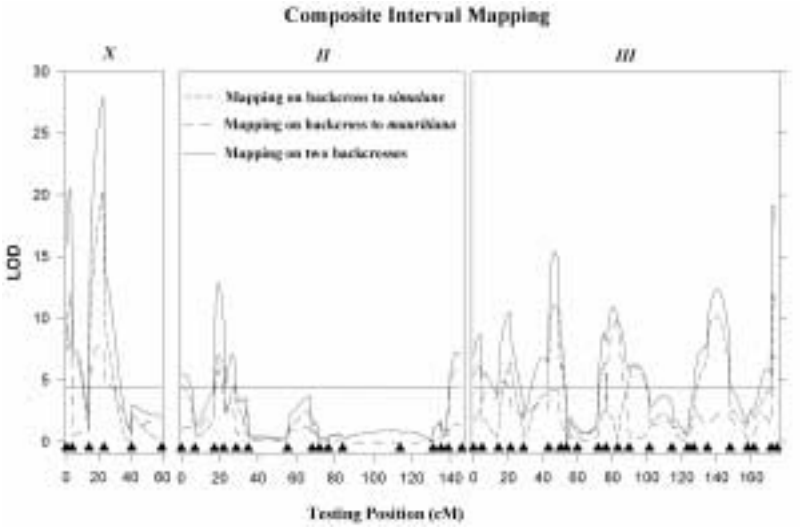

Figure 7.4: Composite interval mapping of a Drosophila morphological shape difference

Chapter 8

Multiple Interval Mapping

Multiple interval mapping (Kao and Zeng 1997; Kao, Zeng and Teasdal 1998) is a multiple QTL oriented method combining QTL mapping analysis with the analysis of genetic architecture of quantitative traits through a search algorithm to search for number, positions, effects and interaction of significant QTL. Using markers for simultaneous multiple QTL analysis has been suggested by Lander and Botstein (1989), although the idea was pursued only with a very limited scope. Recent developments on using Bayesian statistics via Markov chain Monte Carlo for mapping QTL (Satagopan et al. 1996; Uimari and Hoeschele 1997; Sillanpaa and Arjas; Heath 1997) are also multiple QTL based, particularly when it is combined with a reversible jump process (Green 1995).

Multiple interval mapping (MIM) consists of four components:

1. **An evaluation procedure** to analyze the likelihood of the data given a genetic model (number, positions and epistasis of QTL);
2. **A search strategy** to search and select the best genetic model (among those sampled) in the parameter space;
3. **An estimation procedure** to estimate all interested parameters of the genetic architecture of quantitative traits (number, positions, effects and epistasis of QTL; genetic variances and covariances explained by QTL effects) given the selected genetic model;
4. **A prediction procedure** to estimate or predict the genotypic values of individuals based the selected genetic model and estimated genetic parameter values for marker assisted selection.

8.1 Multiple interval mapping model and likelihood analysis

For m putative QTL, the model of multiple interval mapping is specified as

$$y_i = \mu + \sum_{r=1}^{m} \alpha_r x_{ir}^* + \sum_{r \neq s \subset (1,\cdots,m)}^{t} \beta_{rs}(x_{ir}^* x_{is}^*) + e_i \tag{8.1}$$

where

- y_i is the phenotypic value of individual i;
- i indexes individuals of the sample: $i = 1, 2, \cdots, n$;
- μ is the mean of the model;
- α_r is the marginal effect of putative QTL r,
- x^*_{ir} is an indicator variable denoting genotype of putative QTL r (defined by $1/2$ or $-1/2$ for the two genotypes), which is unobserved but can be inferred from maker data in sense of probability;
- β_{irs} is the epistatic effect between putative QTL r and s;
- $r \neq s \subset (1, \cdots, m)$ denotes a subset of QTL pairs that each shows a significant epistatic effect, because if all pairs of m QTL are fitted in the model, the model can be over parameterized;
- m is the number of putative QTL chosen based on either their significant marginal effects or significant epistatic effects;
- t is the number of significant pairwise epistatic effects;
- e_i is a residual effect of the model assumed to be normally distributed with mean zero and variance σ^2.

As the genotypes of an individual at many genomic locations are not observed (but marker genotypes are), the model contains missing data. So the likelihood function of the data given the model is a mixture of normal distributions

$$L(\mathbf{E}, \mu, \sigma^2) = \prod_{i=1}^{n} \left[\sum_{j=1}^{2^m} p_{ij} \phi \left(y_i | \mu + \mathbf{D}_{ij}\mathbf{E}, \sigma^2 \right) \right]. \tag{8.2}$$

The term in bracket is the weighted sum of a series of normal density functions, one for each of 2^m possible multiple-QTL genotypes. p_{ij} is the probability of each multilocus genotype conditional on marker data; $\mathbf{E}$ is a vector of QTL parameters (α's and β's); $\mathbf{D}_{ij}$ is a vector of the genetic model design specifying the configuration of x^*'s associated with each α and β for the jth QTL genotype (see Kao and Zeng 1997); and $\phi(y|\mu, \sigma^2)$ denotes a normal density function for y with mean μ and variance σ^2.

Thus the probability density of each individual is a mixture of 2^m possible normal densities with different means $\mu + \mathbf{D}_{ij}\mathbf{E}$ and mixing proportions p_{ij} which is calculated through marker information.

The procedure to obtain maximum likelihood parameter estimates using an EM algorithm has been described by Kao and Zeng (1997). EM is an iterative procedure involving

an E-step (Expectation) and M-step (Maximization) in each iteration. In the $[t+1]$th iteration, the E-step is

$$\pi_{ij}^{[t+1]} = \frac{p_{ij}\phi(y_i|\mu^{[t]} + \mathbf{D}_{ij}\mathbf{E}^{[t]}, \sigma^{2[t]})}{\sum_{j=1}^{2^m} p_{ij}\phi(y_i|\mu^{[t]} + \mathbf{D}_{ij}\mathbf{E}^{[t]}, \sigma^{2[t]})}. \tag{8.3}$$

The M-step is

$$E_r^{[t+1]} = \frac{\sum_i \sum_j \pi_{ij}^{[t+1]} D_{ijr}[(y_i - \mu^{[t]}) - \sum_{s=1}^{r-1} D_{ijs}E_s^{[t+1]} - \sum_{s=r+1}^{w} D_{ijs}E_s^{[t]}]}{\sum_i \sum_j \pi_{ij}^{[t+1]} D_{ijr}^2} \tag{8.4}$$

$$\mu^{[t+1]} = \frac{1}{n}\sum_i \left(y_i - \sum_j \sum_r \pi_{ij}^{[t+1]} D_{ijr} E_r^{[t+1]} \right) \tag{8.5}$$

$$\sigma^{2[t+1]} = \frac{1}{n}\left[\sum_i (y_i - \mu^{[t+1]})^2 - 2\sum_i (y_i - \mu^{[t+1]}) \sum_j \sum_r \pi_{ij}^{[t+1]} D_{ijr} E_r^{[t+1]} + \sum_r \sum_s \sum_i \sum_j \pi_{ij}^{[t+1]} D_{ijr} D_{ijs} E_r^{[t+1]} E_s^{[t+1]} \right] \tag{8.6}$$

where E_r is the rth element of $\mathbf{E}$ and D_{ijr} is the rth element of $\mathbf{D}_{ij}$.

These equations can be expressed in a general form in matrix notation as (Kao and Zeng 1997)

$$\mathbf{E}^{(t+1)} = diag(\mathbf{V})^{-1}[\mathbf{D}'\mathbf{\Pi}'(\mathbf{Y} - \mu) - nondiag(\mathbf{V})\mathbf{E}^{(t)}] \tag{8.7}$$

$$\mu = \frac{1}{n}\mathbf{1}'[\mathbf{Y} - \mathbf{\Pi}\mathbf{D}\mathbf{E}] \tag{8.8}$$

$$\sigma^2 = \frac{1}{n}[(\mathbf{Y} - \mu)'(\mathbf{Y} - \mu) - 2(\mathbf{Y} - \mu)'\mathbf{\Pi}\mathbf{D}\mathbf{E} + \mathbf{E}'\mathbf{V}\mathbf{E}] \tag{8.9}$$

with

$$\mathbf{V} = \{\mathbf{1}'\mathbf{\Pi}(D_r \# D_s)\}_{r,s=1,\cdots,w} \tag{8.10}$$

where # denotes Hadamard product, which is the element-by-element product of corresponding elements of two same order matrices, and $'$ denotes transposition of a matrix or vector.

Note on the meaning and difference between p_{ij} and π_{ij}: p_{ij} is the probability of each multilocus QTL genotype conditional on marker genotype, and π_{ij} is the probability of each multilocus QTL genotype conditional on marker genotype and also phenotypic value. If we let g denote a QTL genotype, M denote a marker genotype and y denote a phenotype, $\pi_{ij} = Prob(g|M,y) = Prob(g|M)Prob(y|g)/\sum_g Prob(g|M)Prob(y|M)$ as shown in (8.3) with $p_{ij} = Prob(g|M)$ and $Prob(y|g) = \phi\left(y_i|\mu + \mathbf{D}_{ij}\mathbf{E}, \sigma^2\right)$ being the probability of observing the phenotype given a genotype, specified by model (8.1).

It should also be pointed out that when m is large, the number of possible mixture components (QTL genotypes) can become a prohibitive large number for efficient numerical analysis. However, since the probabilities of different genotypes for each observation have to be summed up to one and as the number of genotypes increases, an increasingly large proportion of genotypes have zero or very small probabilities and do not need to be evaluated. In the practical implementation of the algorithm, we have adopted a selection procedure to select a subset of "significant" mixture components for each individual for evaluation. The procedure is that each mixture component that will be evaluated needs to have $p_{ij} > \delta$ (default $\delta = 0.001$) and the sum of those "significant" p_{ij} needs to be larger than 0.95 (otherwise the criterion (δ) for each p_{ij} will be lower). After the selection, the "significant" p_{ij}'s will be normalized to give a sum of 1. By this selection procedure, we found out that the number of those "significant" mixture components is usually on the order of tens and occasionally hundreds (depending on marker density, number and positions of the putative QTL selected), with little or almost no loss on the accuracy of likelihood evaluation as compared to no selection. The burden of numerical analysis is very significantly alleviated.

The test for each QTL effect, say E_r, is performed by a likelihood ratio test conditional on other selected QTL effects

$$LOD = \log_{10} \frac{L(E_1 \neq 0, \cdots, E_{m+t} \neq = 0)}{L(E_1 \neq 0, \cdots, E_{r-1} \neq 0, E_r = 0, E_{r+1} \neq 0, \cdots, E_{m+t} \neq 0)}. \tag{8.11}$$

For given positions of m putative QTL and $m + t$ QTL effects, the likelihood analysis can proceed as outlined above. Now the task is to search and select the best genetic model (number, positions and interaction of QTL) that fits the data well.

8.2 Model selection

8.2.1 Premodel selection

As the evaluation of MIM model is computationally intensive, it is important to select a good premodel for MIM analysis. There are several ways to select a premodel. Currently, we have adopted the following procedure for premodel selection.

First, select a subset of significant markers using a backward stepwise regression (if the number of markers is not larger than the number of sample size); otherwise a forward or a combined forward/backward stepwise regression. We have found that using a stopping rule based on F-to-drop or F-to-enter statistic with $\alpha = 0.01$ is satisfactory in most cases.

Then, use the results from marker selection to perform composite interval mapping to scan the genome for candidate positions.

To identify candidate epistatic terms for the premodel, we find the following procedure useful. This procedure pools markers and marker pairs together in a combined forward stepwise regression analysis. This combined analysis treats marker marginal effects and pairwise interaction effects equally in the selection. After the selection, it is, however,

appropriate to compare the results of this analysis with CIM results to reach a consensus premodel that includes m marginal effects in m positions and t epistatic effects.

Finally, evaluate and test each parameter in the premodel under MIM and drop any non-significant estimate stepwisely.

8.2.2 Model selection under multiple interval mapping

After the first evaluation of the premodel, perform the following stepwise selection analysis to finalize the search for a genetic model under MIM.

1. Begin with a model that contains m QTL and t epistatic effects.

2. Scan the genome to search for the best position of an $(m+1)$th QTL, and then perform a likelihood ratio test for the marginal effect of this putative QTL. If the test statistic exceeds the critical value, this effect is retained in the model.

3. Search for the $t+1$ epistatic effect among those pairwise interaction terms not yet included in the model, and perform the likelihood ratio test on the effect. If LOD exceeds the critical value, the effect is retained in the model. Repeat the process until no more significant epistatic effects are found.

4. Re-evaluate significance of each QTL effect currently fitted in the model. If LOD for a QTL (marginal or epistatic) effect falls below the significant threshold conditional on other fitted effects, the effect is removed from the model. However, if the marginal effect of a QTL that has significant epistatic effect with other QTL falls below the threshold, this marginal effect is still retained. This process is performed stepwisely until test statistic for each effect is above significance threshold.

5. Optimize estimates of QTL positions based on the currently selected model. Instead of performing a multi-dimensional search around the regions of current estimates of QTL positions (which is an option), estimates of QTL positions are updated in turn for each region. For the ith QTL in the model, the region between its two neighbor QTL is scanned to find the position that maximizes the likelihood (conditional on the current estimates of positions of other QTL and QTL epistasis). This refinement process is repeated sequentially for each QTL position until there is no change on estimates of QTL positions.

6. Return step 2 and repeat the process until no more significant QTL effect can be added into the model and estimates of QTL positions are optimized.

It may be worthwhile to attempt to search for significant epistatic effects between selected QTL positions and unselected QTL positions. This may be done stepwisely by searching the largest epistatic effect between a current QTL position and an unselected genomic position at 1 or 2 cM interval and testing it for significance. Of course, numerical calculation is very intensive for this analysis.

Analysis of model selection in a high and unknown dimension is a very complicated and difficult analysis, particularly when it is performed on the whole genome (not just on the observed markers). There could be numerous peaks separated by valleys or connected by ridges in a likelihood landscape with different dimensions. Any model selected, including the current one, from an analysis may well be just a local peak, and there is no guarantee that a global peak can be found from an analysis. There are also issues about appropriate criteria used in model selection for an analysis and appropriate strategies to search for epistatic QTL and to estimate QTL epistasis. Clearly, more detailed and in-depth analyses about these issues are critically needed. These analyses must be globally (i.e. genome wide) and architecturally (i.e. multiple components and multiple levels of genetic effects) oriented to be informative for the study of genetic architecture of quantitative traits.

8.3 Stopping rules

An important issue associated to model selection is stopping rule for a model search algorithm or criterion for comparing different models. In regression analysis with model selection, the stopping rules are usually decided based on minimizing the final prediction error (FPE) criterion or information criteria (IC) (Stuart and Ord 1991; Miller 1990). As pointed out by Broman (1997), QTL mapping analysis is also a model selection analysis. However, unlike many model selection problems in regression analysis, the independent variables (QTL genotypes) are not observed, but markers are. Model selection practiced on markers only (Broman 1997) is very informative, but insufficient in achieving the main objective of QTL mapping, locating QTL positions. Current many statistical analyses for mapping QTL use likelihood ratio or F statistic to test each genetic effect fitted in the model as a basis of model selection with adjustment on significance value for each test to account for multiple tests practiced in searching for QTL (Lander and Botstein 1987; Haley and Knott 1992; Zeng 1994; Jansen and Stem 1994).

The final prediction error (FPE) criterion is

$$S_k = (n+k)RSS_k/(n-k) \tag{8.12}$$

where $RSS(k)$ is residual sum of squares and k is the number of parameters fitted in the model. The information criteria of the general form is

$$\text{IC} = -2(\log L_k - kc(n)/2) \tag{8.13}$$

where L_p is the likelihood of data (8.2) given a genetic model with k parameters. This is approximately equivalent to

$$\text{IC} = \log[RSS_k/n] + kc(n)/n \tag{8.14}$$

in regression analysis. Akaike (1969) suggested $c(n) = 2$, whereas Schwarz (1978) recommends $c(n) = \log(n)$ and Hannan and Quinn (1979) consider $c(n) = 2\log(\log n)$. It has been shown that S_k and Akaike's IC produce equivalent results asymptotically (Shibata

1981, 1984). Breiman and Freedman (1983) showed that S_k is asymptotically optimal in the sense of minimizing the prediction error. The Schwarz and Hannan-Quinn criteria produce consistent estimators in the sense that the probability of selecting the true model approaches one as $n \to \infty$; the other measures do not achieve this and typically include too many terms. These asymptotic results give no indication of the behavior for finite samples sizes however.

The IC criteria can be related to F-to-enter statistic (for regression analysis) or LR-to-enter statistic (for likelihood analysis) in the stepwise selection procedure. It was shown (Miller 1990, p.208) that (8.13) leads to the F-to-enter statistic for regression analysis at the minimum

$$\frac{RSS_k - RSS_{k+1})}{RSS_{k+1}/(n-k-1)} \leq (n-k-1)(e^{c(n)/n} - 1) \approx 2c(n)\left(1 - \frac{k+1}{n}\right) \tag{8.15}$$

provided that $c(n)/n$ is small. As $LR = n\log(RSS_k/RSS_{k+1})$ in the setting of regression analysis, (8.13) and (8.15) imply that the LR-to-enter statistic for likelihood analysis at minimum is

$$LR_k = -2\log\frac{L_k}{L_{k+1}} \leq n\log(c(n)/n + 1) \approx c(n). \tag{8.16}$$

The criterion is basically defined by $c(n)$. Using $c(n) = 2$ as suggested by Akaike (1969) would mean that the final threshold in LOD score is 0.43.

In reference to QTL analysis on markers, Broman (1997) suggested to use $c(n) = \delta\log n$ and recommended δ be between 2 and 3. For $n = 100 \sim 500$, the threshold in LOD would be $2 \sim 2.7$ for $\delta = 2$ and $3 \sim 4$ for $\delta = 3$. This, on the surface, appears to be in line with some current practice in interval mapping (Lander and Botstein 1989; Zeng 1994) based on different arguments.

However, this argument is still rather arbitrary and does not relate it to the genetic length of linkage map, number of markers and linkage groups, and distribution of markers. Clearly more studies on stopping rules are needed.

8.4 Other estimations and prediction

Given estimates of the QTL parameters, one can estimate genotypic values of an individual. This estimation is complicated by the fact that QTL genotypes are not observed directly. Only marker genotypes are observed. Thus, the estimation for an individual is the weighted mean of all possible genotypic values, weighted by the probability ($\hat{\pi}_{ij}$) of each QTL genotype conditional on the marker and phenotypic data. From (8.5), this estimation equation is

$$\hat{y}_i = \hat{\mu} + \sum_{j=1}^{2^m}\sum_{r=1}^{m+t}\hat{\pi}_{ij}D_{ijr}\hat{E}_r \tag{8.17}$$

where the first summation is over all possible 2^m QTL genotypes (in numerical analysis only those "significant" QTL genotypes, see above) and the second summation is over all

effects of the model (m main effects and t epistatic effects). $\hat{\mu}$ is the maximum likelihood estimate (MLE) of μ obtained from 8.5 at the equilibrium of the final model, and $\hat{E}_r$ is MLE of QTL effect E_r obtained from (8.4). $\hat{\pi}_{ij}$ is MLE of π obtained from (8.3).

To predict the genotypic values of quantitative traits based on marker information only (e.g. in cross-prediction; early selection), we need to use

$$\hat{y}_i = \hat{\mu} + \sum_j \sum_r p_{ij} D_{ijr} \hat{E}_r \tag{8.18}$$

as $\hat{\pi}_{ij}$ is a function of phenotype y_i which is unavailable in early selection.

The genetic variances and covariances explained by each QTL effect can be estimated directly from the likelihood analysis. At the convergence of the EM algorithm, equation (8.7) leads to

$$\hat{\mathbf{E}} = \hat{\mathbf{V}}^{-1}\mathbf{D}'\hat{\mathbf{\Pi}}'(\mathbf{Y} - \hat{\mu}) \tag{8.19}$$

This means that

$$\hat{\sigma}^2 = \frac{1}{n}[(\mathbf{Y} - \hat{\mu})'(\mathbf{Y} - \hat{\mu}) - \hat{\mathbf{E}}'\hat{\mathbf{V}}\hat{\mathbf{E}}] \tag{8.20}$$

or

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n}\left[\sum_{i=1}^{n}(y_i - \hat{\mu})^2 - \sum_{r=1}^{m+t}\sum_{s=1}^{m+t}\sum_{i=1}^{n}\sum_{j=1}^{2^m}\hat{\pi}_{ij}D_{ijr}D_{ijs}\hat{E}_r\hat{E}_s\right] \\ &= \frac{1}{n}\left[\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{r=1}^{m+t}\sum_{s=1}^{m+t}\sum_{i=1}^{n}\sum_{j=1}^{2^m}\hat{\pi}_{ij}(D_{ijr} - \bar{D}_r)(D_{ijs} - \bar{D}_s)\hat{E}_r\hat{E}_s\right] \end{aligned} \tag{8.21}$$

where $\bar{y} = \sum_{i=1}^{n} y_i/n$ and $\bar{D}_r = \sum_{i=1}^{n}\sum_{j=1}^{2^m}\hat{\pi}_{ij}D_{ijr}/n$. In this form $\hat{\sigma}^2$ is expressed as a difference between MLE of total phenotypic variance $\hat{\sigma}_p^2$ (the first part of (8.21)) and that of genetic variance $\hat{\sigma}_g^2$ (the second part of (8.21)).

$\hat{\sigma}_g^2$ can be further partitioned into

$$\begin{aligned} \hat{\sigma}_g^2 &= \sum_{r=1}^{m+t}\left[\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{2^m}\hat{\pi}_{ij}(D_{ijr} - \bar{D}_r)^2\hat{E}_r^2\right] \\ &\quad + \sum_{r=2}^{m+t}\sum_{s=1}^{r-1}\left[\frac{2}{n}\sum_{i=1}^{n}\sum_{j=1}^{2^m}\hat{\pi}_{ij}(D_{ijr} - \bar{D}_r)(D_{ijs} - \bar{D}_s)\hat{E}_r\hat{E}_s\right] \\ &= \sum_{r=1}^{m+t}\hat{\sigma}_{E_r}^2 + \sum_{r=2}^{m+t}\sum_{s=1}^{r-1}\hat{\sigma}_{E_r,E_s} \end{aligned} \tag{8.22}$$

$\hat{\sigma}_{E_r}^2$ estimates genetic variance due to QTL effect E_r and $\hat{\sigma}_{E_r,E_s}$ estimates genetic covariance between QTL effects E_r and E_s.

It is convenient and also informative to combine the variance due to each QTL effect with half of the covariances between this QTL effect and other effects and report this variance

component as the variance component explained by this QTL effect

$$\hat{\sigma}_r^2 = \hat{\sigma}_{E_r}^2 + \frac{1}{2}\sum_{s \neq r} \hat{\sigma}_{E_r,E_s} \tag{8.23}$$

Whereas $\hat{\sigma}_{E_r}^2$ estimates the variance of the rth QTL effect in linkage equilibrium (in which $\sigma_{E_r,E_s} = 0$), $\hat{\sigma}_r^2$ estimates the contribution to the total variance in the current population with linkage disequilibrium. Estimates of these variances, covariances and variance components are given in Table 8.2 expressed as a ratio of the total phenotypic variance. Note that $\hat{\sigma}_g^2/\hat{\sigma}_p^2$ is the coefficient of determination (R^2) of the MIM model. Note also whereas $\hat{\sigma}_{E_r}^2$ is always positive, $\hat{\sigma}_r^2$ is not necessary to be positive.

8.5 Genetic architecture of a morphological shape difference

We show as an example the mapping results of an experiment in *Drosophila* (Zeng et al. 1998). Two *Drosophila* species, *D. simulans* and *D. mauritiana*, were crossed to make F_1 hybrids. Because F_1 males are sterile, females of F_1 population were backcrossed to each of the parental lines. Two independent samples (of size 200 and 300) were drawn from each backcross population and genotyped and phenotyped at two different times. Therefore the total sample size of the experiment is about 1000. We refer to the two samples from backcross to *D. simulans* as BS-S1 and BS-S2, and those to *D. mauritiana* as BM- S1 and BM-S2. The trait is the morphology of the posterior lobe of the male genital arch analyzed as the first principal component in an elliptical Fourier analysis (Liu et al. 1995). The results of MIM analysis are shown in Figures 8.1 and 8.2 and Tables 8.1 and 8.2. The final model selected contains 19 QTL (based on the joint analysis of the samples in two backcrosses) distributed on the three *Drosophila* major chromosomes, X, II and III. Figure 8.1b depicts the likelihood profile (LOD score) for each QTL that spans from one QTL to its neighbors. The threshold used in analysis is also shown in the figure. As a comparison, Figure 8.1a shows the mapping result based on CIM. Forty five markers were genotyped and their map positions are indicated by filled triangles in the figure.

Table 8.1 shows the estimates of positions and effects of these 19 QTL as percentage of the observed difference of the trait means between the respective parental populations and the F_1 hybrid. The sum of the 19 QTL effects explain 99% of the observed differences. Also, because the estimates of substitution effects are estimates of $a + d$ in BM and $a - d$ in BS, where a is the additive effect of a QTL and d is the dominance effect, a and d can be jointly estimated and are expressed as percentage of half the observed difference between two parental populations in Table 8.1. Again the additive effects of these 19 QTL explain 99% of the observed difference. There are substantial dominance effects, but overall the dominance effects are marginal compared to the additive effects. Six QTL pairs show significant epistatic effects in BM (Table 8.2), and none in BS according to the threshold adopted for the study. Together, these 19 QTL explain 93.2% of the total variance in BS and 91.6% (plus epistatic variances in Table 8.2) in BM. These are the coefficients of determination (R^2) of the MIM model in the respective populations, an estimate of

heritability of the trait. With these estimates, the genetic architecture of the trait difference between *D. simulans* and *D. mauritiana* becomes clear.

Because this experiment contains two independent samples for each backcross, we asked whether the mapping results obtained from one sample can be adequately applied to the other sample. The estimated (using estimates of model parameters from the sample) and cross-predicted (using estimates from the other sample) phenotypic values of the two samples in BS are shown in Figure 8.2. Although there is some reduction in R^2 in cross-prediction as compared to direct estimation, the reduction is not as profound as one might expect. This shows the power of molecular marker information and multiple interval mapping. Even estimates of model parameters from an independent sample can be used to predict phenotypes of another sample with 89% accuracy.

An analysis of the data using a composite interval mapping analysis finds only 14 of the 19 QTL (Figure 8.1a). In this case, the use of MIM and a consideration of the genetic complexity in mapping analysis (two backcrosses and epistasis) helped greatly in the identification of QTL, and the improved identification of QTL in turn helped greatly to uncover the genetic architecture of the trait.

8.6 Genetic architecture of Drosophila wing shape

Genetic architecture of the wing size of *Drosophila melanogaster* on chromosome 3 (Weber *et al.* 1999 *Genetics* 153: 773-786).

- Population: 519 recombinant inbred lines originating from a cross between high and low selected lines on wing size. Only QTL on chromosome 3 are segregating in the population, and other chromosomes are identical for all RIL.
- Trait: wing size measured in radian in an allometric analysis.
- 11 QTL are identified by MIM analysis. There is a good agreement between the sum of estimated additive effects of QTL and the observed parental genotype difference.
- There are some significant additive by additive interaction effects between QTL. The interaction pattern is complex.
- Together, 11 additive and 9 additive by additive QTL effects explain 96% of the total variance in the population.

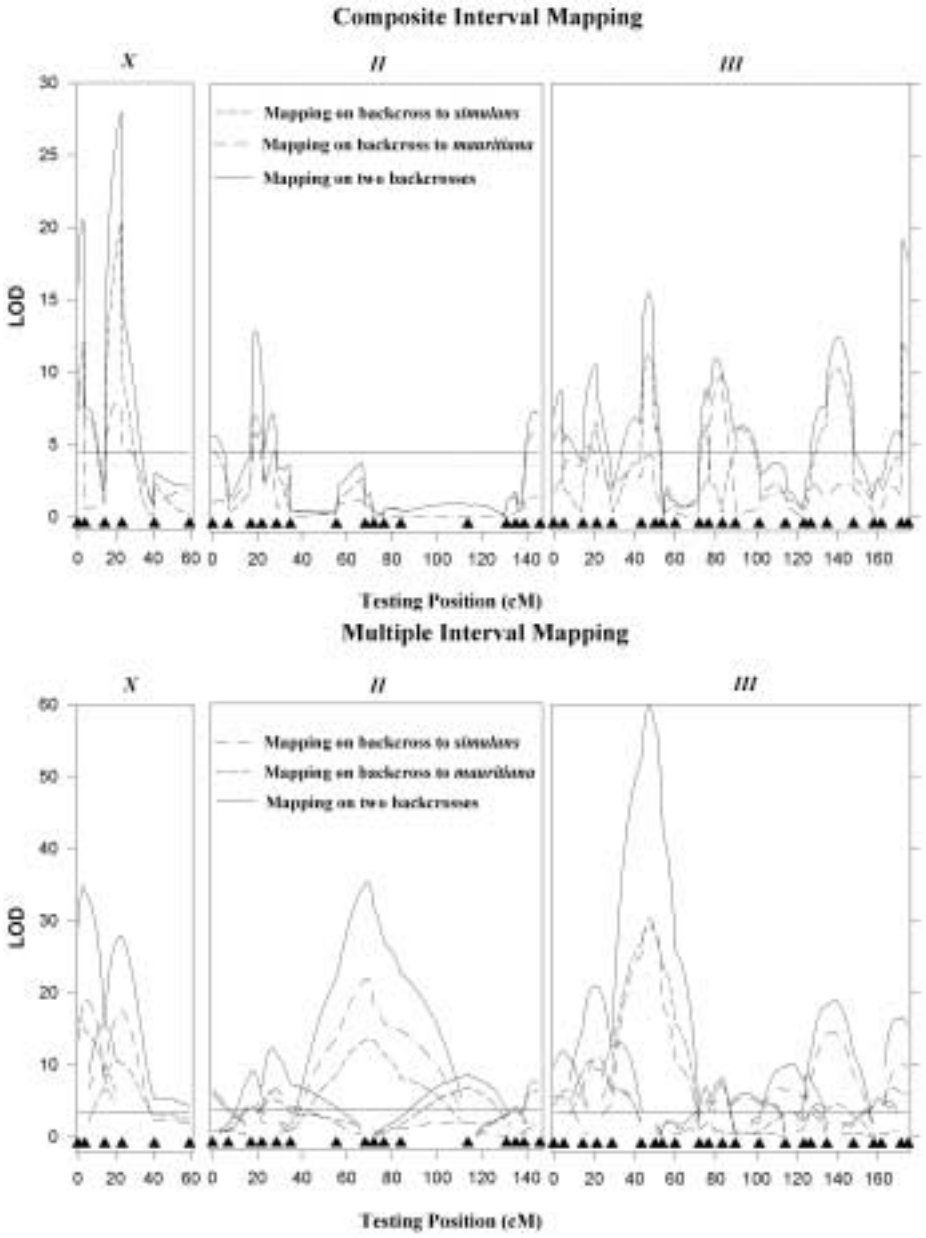

Figure 8.1: Comparison of composite interval mapping and multiple interval mapping on PC1 in two *Drosophila* backcrosses

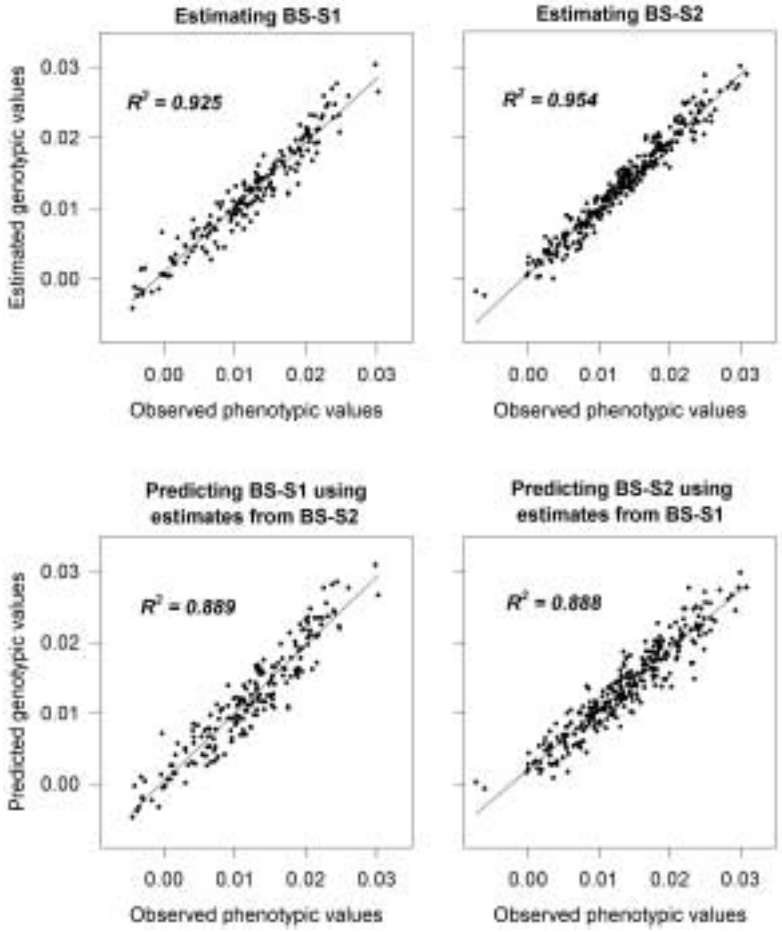

Figure 8.2: Prediction and cross-prediction of genotypic values in the backcross to *D. simulans*

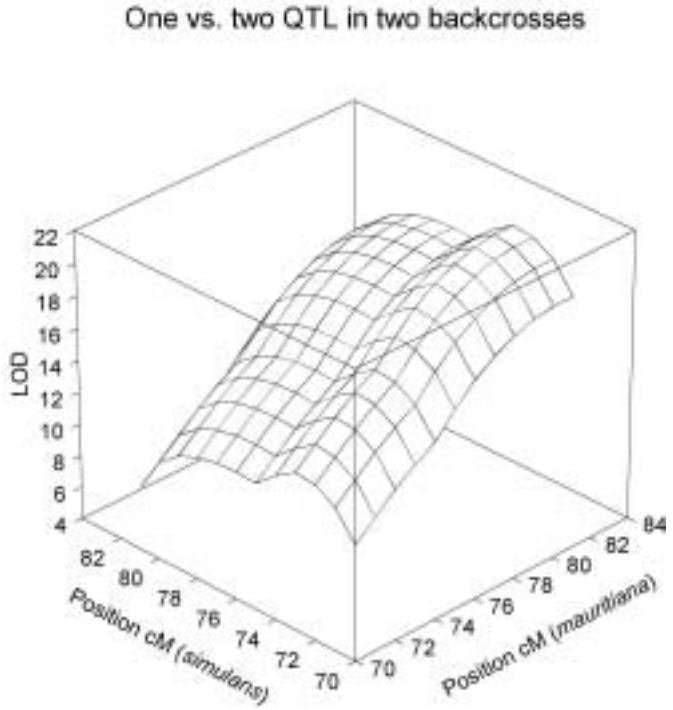

Figure 8.3: Test between one common QTL vs. two different QTL in two backcrosses in the region of III:70–84 cM, QTL 13 and 14

Table 8.1: Estimates of QTL positions, effects and variance components by MIM on PC1 in the two *Drosophila* backcrosses

QTL	Posi (Chro:cM)	BM (%) $(a+d)^a$	BS (%) $(a-d)^b$	a^c	d^c	BM (%) $\hat{\sigma}_r^2/\hat{\sigma}_p^2$	BS (%) $\hat{\sigma}_r^2/\hat{\sigma}_p^2$
1	*X*:3	8.4	–[d]	3.8[e]	–[f]	4.4	4.4
2	*X*:23	8.3	–[d]	3.7[e]	–[f]	4.5	3.0
3	*II*:0	-0.6	4.3	2.1	-2.6	0.1	2.8
4	*II*:17	5.1	6.5	5.9	-1.2	3.8	5.9
5	*II*:27	9.0	7.0	7.9	0.3	6.7	5.7
6	*II*:69	4.6	7.9	6.4	-2.2	3.3	5.0
7	*II*:114	4.7	2.4	3.5	0.8	2.5	0.9
8	*II*:135	-2.6	0.3	-1.0	-1.4	-0.7	0.3
9	*II*:143	5.9	3.1	4.4	1.0	3.2	0.9
10	*III*:5	5.0	5.1	5.0	-0.5	4.5	3.5
11	*III*:21	8.0	7.7	7.8	-0.5	7.7	6.8
12	*III*:47	10.2	12.3	11.4	-2.0	12.7	11.6
13	*III*:75	0.7	8.4	4.9	-4.3	0.7	9.1
14	*III*:83	12.4	-1.2	5.0	6.3	14.9	-0.3
15	*III*:94	1.7	7.0	4.6	-3.0	2.6	7.6
16	*III*:117	4.4	5.6	5.1	-1.1	4.3	6.4
17	*III*:139	4.8	8.3	6.8	-4.7	4.2	8.9
18	*III*:160	1.6	7.1	4.6	-3.2	1.3	5.5
19	*III*:172	7.5	7.2	7.3	-0.5	4.4	5.2
Total		99.1	99.0	99.2	-18.8	85.1	93.2

[a] As percentages of the phenotypic difference between F_1 and *D. mauritiana*.
[b] As percentages of the phenotypic difference between *D. simulans* and F_1.
[c] As percentages of half the difference between *D. simulans* and *D. mauritiana*.
[d] QTL in chromosome *X* does not contribute to the observed difference.
[e] Only half of the additive effect contributes to the observed difference.
[f] There is no dominance effect for QTL in chromosome *X*.

Table 8.2: Estimates of QTL epistatic effects and variance components in *D. mauritiana* backcross

QTL 1	QTL 2	LOD	Epis. Effect	$\hat{\sigma}^2_r/\hat{\sigma}^2_p$ (%)
3	*12*	3.29	0.89	2.2
8	*15*	3.44	1.48	1.0
1	*17*	7.32	2.01	0.8
3	*17*	3.01	1.17	0.8
6	*17*	4.22	1.41	0.6
12	*17*	7.57	2.00	1.1
Total				6.5

8.7 Advantages of multiple interval mapping

As pointed out in Kao, Zeng and Teasdale (1998), there are several advantages of using multiple interval mapping for QTL mapping study. First, by directly using multiple QTL components and QTL epistasis in the model analysis for the search of individual QTL, MIM helps greatly in the identification of QTL. It helps to improve statistical power to identify more and complex QTL, and it helps to improve the precision of estimating QTL positions.

Second, with multiple genetic components and epistasis in a single model analysis and with the improved identification of QTL, MIM helps greatly to uncover the genetic architecture of quantitative traits. It helps to identify patterns and individual elements of QTL epistasis, and it helps to provide appropriate and integral estimation of individual QTL effects, variance and covariance contribution. As shown in Tables 8.1 and 8.2, these estimates make much better sense than those by other mapping methods, both biologically and statistically about the partition of genetic effects and genetic variance. This helps us greatly to assess relative contribution and importance of different genetic components, and in short to understand the genetic architecture of quantitative trait values and quantitative trait variation in a population.

Third, these improvements on the identification and estimation of QTL components, positions and epistasis are directly used to help to estimate genotypic values of individuals for marker assisted selection.

Multiple interval mapping helps to bring the three important studies together, and surely will be the direction for future research and the basis for QTL mapping data analysis.

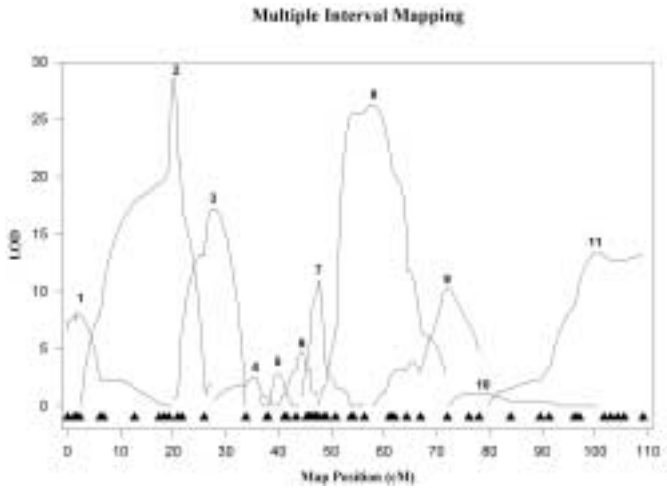

Figure 8.4: Likelihood profiles of the 11 QTL identified by multiple interval mapping in LOD score. The likelihood profile for each QTL spans from near the estimated position of one neighbor QTL to the other or the end of the chromosome. Triangles indicate the map positions of markers.

Table 8.3: Estimates of QTL positions and effects

QTL	Posi (cM)	LOD	Effect ($\times 10^{-2}$)	Effect (%)[a]
1	3	8.1	0.41	6.7
2	20	28.6	1.15	18.9
3	28	17.2	0.91	14.9
4	35	2.5	0.53	8.7
5	40	2.8	0.22	3.6
6	44	4.8	0.62	10.3
7	48	11.0	0.81	13.3
8	58	26.3	0.86	14.1
9	72	10.3	0.43	7.0
10	78	1.2	-0.14	-2.3
11	100	13.3	0.47	7.7
Total				102.9

[a] The effects are in percentage of the phenotypic difference between the high and low nonrecombinant lines

Table 8.4: Estimates of QTL interactions

QTL (pair)	LOD	Effect ($\times 10^{-2}$)
(1&2)	1.53	-0.33
(2&4)	1.73	-0.36
(3&8)	2.85	-0.81
(3&9)	7.65	1.38
(4&5)	1.47	-0.56
(5&8)	2.85	0.88
(5&9)	5.60	-1.29
(6&10)	1.14	0.31
(8&11)	4.08	-0.47

Table 8.5: Estimated variances and covariances of QTL main effects in ratio of total phenotypic variance

	1	2	3	4	5	6	7	8	9	10	11	Sum
1	0.009	0.018	0.009	0.003	0.001	0.001	0.000	-0.002	-0.002	0.001	-0.004	0.034
2	0.018	0.075	0.044	0.018	0.006	0.011	0.011	0.006	-0.002	0.001	-0.009	0.179
3	0.009	0.044	0.047	0.020	0.007	0.014	0.016	0.011	0.002	0.000	-0.005	0.163
4	0.003	0.018	0.020	0.016	0.005	0.012	0.013	0.010	0.003	0.000	-0.002	0.097
5	0.001	0.006	0.007	0.005	0.003	0.006	0.007	0.006	0.002	0.000	0.000	0.042
6	0.001	0.011	0.014	0.012	0.006	0.022	0.025	0.020	0.007	-0.001	0.000	0.117
7	0.000	0.011	0.016	0.013	0.007	0.025	0.037	0.028	0.009	-0.002	0.000	0.145
8	-0.002	0.006	0.011	0.010	0.006	0.020	0.028	0.042	0.015	-0.004	0.005	0.136
9	-0.002	-0.002	0.002	0.003	0.002	0.007	0.009	0.015	0.011	-0.003	0.005	0.045
10	0.001	0.001	0.000	0.000	0.000	-0.001	-0.002	-0.004	-0.003	0.001	-0.002	-0.010
11	-0.004	-0.009	-0.005	-0.002	0.000	0.000	0.000	0.005	0.005	-0.002	0.012	-0.001
Total												0.947

Table 8.6: Estimated variances and covariances of QTL epistatic effects in ratio of total phenotypic variance

	1-2	2-4	3-8	3-9	4-5	5-8	5-9	6-10	8-11	Sum
1-2	0.001	0.000	0.000	0.001	0.000	0.000	-0.001	0.000	0.000	0.001
2-4	0.000	0.001	0.000	0.000	0.000	0.001	-0.001	0.000	0.000	0.001
3-8	0.000	0.000	0.009	-0.011	0.001	-0.005	0.004	0.000	-0.001	-0.004
3-9	0.001	0.000	-0.011	0.027	-0.001	0.005	-0.014	0.002	0.001	0.009
4-5	0.000	0.000	0.001	-0.001	0.002	0.000	-0.001	0.000	0.000	0.001
5-8	0.000	0.001	-0.005	0.005	0.000	0.008	-0.008	0.001	0.001	0.004
5-9	-0.001	-0.001	0.004	-0.014	-0.001	-0.008	0.021	-0.003	0.000	-0.003
6-10	0.000	0.000	0.000	0.002	0.000	0.001	-0.003	0.001	-0.001	0.001
8-11	0.000	0.000	-0.001	0.001	0.000	0.001	0.000	-0.001	0.003	0.002
Total										0.012

Chapter 9

A General View and Directions for Extension

9.1 A general view of QTL mapping analysis

QTL mapping analysis is about linking observed trait phenotypes (Y) to unobserved QTL genotypes (and other non-genetic factors as well) (θ) through observed marker phenotypes (X).

For a given mapping population, the likelihood of data for a genetic model can be represented symbolically by

$$L(\theta|Y,X) = \prod_{j=1}^{n} \sum_{g_1} \cdots \sum_{g_t} P_j(g|X,\lambda) f_j(\theta \mapsto Y|g)$$

- g specifies the genotypes of QTL.
- $\theta \mapsto y$ specifies the model or mapping from QTL genotypes to trait phenotypes. Thus $f_j(\theta \mapsto Y|g)$ is a function of the mapping given the QTL genotypes.
- $P_j(g|X,\lambda) = Prob(x^*_{j1} = g_1, \cdots, x^*_{jt} = g_t|X,\lambda)$ specifies the probability of QTL genotypes at specified genomic locations (λ) given the observed marker phenotypes.

Thus, basically, QTL mapping analysis consists of two parts:

1. The analysis of QTL genotypes (missing data) given the observed marker phenotypes at various genomic locations for various experimental designs and data structures, *i.e.* analyzing $P_j(g|X,\lambda) = Prob(x^*_{j1} = g_1, \cdots, x^*_{jt} = g_t|X,\lambda)$.
2. Search and evaluate the genotype-phenotype mapping relationship $\theta \mapsto Y$ for the given QTL genotypes based on likelihood.

For many experimental designs and data structures, complications can arise on both parts. Fortunately, for many analyses, we can separate the analyses of the two parts and study them independently.

For missing marker analysis and analysis of different experimental designs, we are mostly concerned with the analysis of

$$P_j(g|X,\lambda) = Prob(x^*_{j1} = g_1, \cdots, x^*_{jt} = g_t|X,\lambda)$$

For testing different genetic hypotheses, we are mostly concerned with the analysis of

$$f_j(\theta \mapsto Y|g)$$

and related likelihood.

9.2 Some complications in analysis

- Marker data and trait data may not be recorded on the same individuals. For example, if the marker data is recorded on individuals in generation u and the trait data on individuals in generation v, we need to analyze $P_j(g^v|X^u,\lambda)$.
- Marker data may also be available on relatives (X^R), thus we need to analyze $P_j(g|X,X^R,\lambda)$.
- There could be a hierarchical structure in the data (and thus a hierarchical model of genetic effects as well). For example, suppose we have a group of related or unrelated families (F) and each family with a number of offsprings (n_J). We may extend the likelihood symbolically to

$$L(\theta|Y,X) = \int \prod_{J=1}^{F} \prod_{j=1}^{n_J} \sum_{g} P_j(g|X,\lambda) f_j([\phi_J|\theta] \mapsto Y|g) p(\phi) d(\phi)$$

- If we want to utilize the prior information about θ ($p(\theta)$) in mapping analysis, we may further extend the analysis into a Bayesian framework and analyze $p(\theta|Y,X) \propto p(\theta)p(Y,X|\theta)$ as $L(\theta|Y,X) = p(Y,X|\theta)$.

9.3 Some extensions of QTL mapping analysis

1. Beyond backcross and F_2 populations
   - Combining QTL mapping with breeding programs: Mapping from advanced cross populations (F_t, $t \geq 2$) by selfing or random mating; recombinant inbred lines (RI). [plant and experimental animal populations]
   - Combining QTL mapping with gene introgression: Mapping from repeated backcross or test cross populations (B_t, $t \geq 1$). [plant and experimental animal populations]

- Mapping from outbreed crosses (segregating parental populations). [farm animal populations]
- Mapping from full-sib, half-sib, mixed sib families, grand-daughter designs. [farm animal populations]
- Mapping in extended families; closed populations; general populations. [human populations]

2. Beyond co-dominant markers

   - Missing or uninformative markers.
   - Dominant or partially informative markers.

3. Beyond additive and dominance model of QTL

   - Test and estimate gene interaction.
   - Study QTL by environment interaction.

4. Beyond single trait analysis

   - Multiple trait/environment analysis: Study the basis of genetic correlations between different traits; pleiotropy or linkage; QTL by enviornment interaction.

5. Beyond quantitative traits with a normal distribution

   - Study threshold traits. (This is not discussed in the current course, and will be included in the future.)

Chapter 10

Dominant and Missing Marker Analysis

Many PCR based markers are dominant in nature. There has been some concern about the use of dominant markers in QTL mapping because of partial missing information. Although the problem of dominant markers can be avoided through experimental designs, such as using recombinant inbred lines and double haploids to remove heterozygote class or just using segregating markers in a backcross, it is common to have dominant markers in other populations such as F_2. When some markers show dominant phenotype, it is appropriate to utilize all available data in the analysis. Here, we discuss an efficient algorithm to use dominant markers in an F_2 population and also extend it to other types of populations stemming from two inbred lines.

10.1 A Markov chain algorithm for F_2 population

Consider an F_2 population from a cross between two inbred lines, P_1 and P_2. Suppose there are m markers on a chromosome whose map positions are known and arranged in the order of $M_1, \cdots, M_m$. In this population, each marker or QTL has three possible genotypes. Let x_k denote the genotype of marker (or QTL) M_k for an individual, which takes a value 1, 0 or -1 if M_k is homozygote of P_1 type, heterozygote or homozygote of P_2 type respectively.

To facilitate the following discussion, we let z_k denote the phenotype of marker (or QTL) M_k for the same individual. When a marker is fully observed, the phenotype equals the genotype, *i.e.* $z_k = x_k = \{1\}, \{0\}$, or $\{-1\}$. When a marker is unobserved (*i.e.* missing), the genotype is unknown with $z_k = \{1,\ 0,\ -1\} = M$ for missing. When a marker is partially observed, the phenotype includes two possible genotypes. In this paper, a dominant phenotype represents homozygote of P_1 type or heterozygote (*i.e.*, $z_k = \{1,\ 0\} = D$ for dominance or $z_k \neq -1$), and a recessive phenotype represents heterozygote or homozygote of P_2 type (*i.e.*, $z_k = \{0,\ -1\} = R$ for recessive or $z_k \neq 1$). [Another subset $z_k = \{1,\ -1\}$ can also be included in analysis if necessary].

If M_k is a putative QTL whose genotype may or may not be observed (depending on

the testing position), a very important analysis in QTL mapping is to calculate, for each individual, the conditional probability of x_k being in different values given observed marker phenotypes on the chromosome. We denote this probability as $P(x_k|z_1, \cdots, z_m)$.

If M_{k-1} and M_{k+1}, the two flanking markers of M_k, are both fully observed for the individual, the conditional probability depends entirely on the phenotype of M_k, the genotypes of M_{k-1} and M_{k+1} and the recombination frequencies between M_{k-1} and M_k and between M_k and M_{k+1} and is independent of other markers on the chromosome under the assumption of *no crossing-over interference*, *i.e.*

$$P(x_k|z_1, \cdots, x_{k-1}, z_k, x_{k+1}, \cdots, z_m) = P(x_k|x_{k-1}, z_k, x_{k+1}).$$

However, if one (or both) of the flanking markers is unobserved or only partially observed, the genotype or phenotype at the next marker away from the flanking marker can provide some information about the genotype of the flanking marker and this in turn will improve the estimation of the probability distribution of the genotype at the testing position for the QTL for the individual. This dependence can be extended further in each direction until a marker locus which is fully observed or to the terminal marker of the chromosome.

Let M_i and M_l ($i \leq k \leq l$) are two most adjacent fully observed markers. If there is no fully observed marker in one or both directions, take $M_i = M_1$ or $M_l = M_m$ or both. The task is to calculate $P(x_k|z_i, \cdots, z_l)$. By Bayes' theorem,

$$P(x_k|z_i, \cdots, z_l) = \frac{P(x_k)P(z_i \cdots z_l|x_k)}{\sum_{x_k} P(x_k)P(z_i \cdots z_l|x_k)}. \tag{10.1}$$

Note that under the assumption of no crossing-over interference, for a given specific value of x_k

$$\begin{aligned} P(z_i \cdots z_l|x_k) &= P(z_i \cdots z_k|x_k, z_{k+1}, \cdots, z_l)P(z_{k+1} \cdots z_l|x_k) \\ &= P(z_i \cdots z_k|x_k)P(z_{k+1} \cdots z_l|x_k). \end{aligned}$$

In (10.1), $P(x_k)$ is the unconditional or prior probability of x_k in a population. Let

$$\mathbf{q}_k = \{P(x_k)\}_{(3\times 1)}$$

denote a row vector for the prior probability $P(x_k)$, *i.e.*

$$\mathbf{q}_k' = [P(x_k = 1), \;\; P(x_k = 0), \;\; P(x_k = -1)]$$

where $'$ denotes transposition. Similarly, let also

$$\begin{aligned} \mathbf{p}_k^R &= \{P(z_{k+1} \cdots z_l|x_k)\}_{(3\times 1)} \\ \mathbf{p}_k^L &= \{P(z_i \cdots z_k|x_k)\}_{(3\times 1)} \\ \mathbf{p}_k &= \{P(x_k|z_i, \cdots, z_l)\}_{(3\times 1)}. \end{aligned}$$

Then, equation (10.1) can be expressed as

$$\mathbf{p}_k = \frac{\mathbf{q}_k \circ (\mathbf{p}_k^L \circ \mathbf{p}_k^R)}{\mathbf{q}_k'(\mathbf{p}_k^L \circ \mathbf{p}_k^R)} \tag{10.2}$$

where $\circ$ denotes componentwise product of vectors.

For an F_2 population,

$$\mathbf{q}_k' = [\frac{1}{4},\ \frac{1}{2},\ \frac{1}{4}]. \tag{10.3}$$

$\mathbf{p}_k^R$ and $\mathbf{p}_k^L$ can be calculated via a Markov chain process. To show how to calculate them, let us first consider some simple situations. If an individual has the dominant phenotype at M_{k+1} (*i.e.* $z_{k+1} = \{1,\ 0\}$),

$$P(z_{k+1}|x_k) = \sum_{x_{k+1} \in z_{k+1}} P(x_{k+1}|x_k) = P(x_{k+1} = 1|x_k) + P(x_{k+1} = 0|x_k).$$

If the individual also has the recessive phenotype at M_{k+2} (*i.e.* $z_{k+2} = \{0,\ -1\}$),

$$\begin{aligned} & P(z_{k+1}z_{k+2}|x_k) \\ = & \sum_{x_{k+1} \in z_{k+1}} \sum_{x_{k+2} \in z_{k+2}} P(x_{k+1}x_{k+2}|x_k) \\ = & \sum_{x_{k+1} \in z_{k+1}} \sum_{x_{k+2} \in z_{k+2}} P(x_{k+2}|x_{k+1})P(x_{k+1}|x_k) \\ = & [P(x_{k+2} = 0|x_{k+1} = 1) + P(x_{k+2} = -1|x_{k+1} = 1)]P(x_{k+1} = 1|x_k) + \\ & [P(x_{k+2} = 0|x_{k+1} = 0) + P(x_{k+2} = -1|x_{k+1} = 0)]P(x_{k+1} = 0|x_k). \end{aligned} \tag{10.4}$$

Then if we let

$$\begin{aligned} \mathbf{H}(r_k) &= \begin{bmatrix} P(x_{k+1} = 1|x_k = 1) & P(x_{k+1} = 0|x_k = 1) & P(x_{k+1} = -1|x_k = 1) \\ P(x_{k+1} = 1|x_k = 0) & P(x_{k+1} = 0|x_k = 0) & P(x_{k+1} = -1|x_k = 0) \\ P(x_{k+1} = 1|x_k = -1) & P(x_{k+1} = 0|x_k = -1) & P(x_{k+1} = -1|x_k = -1) \end{bmatrix} \\ &= \begin{bmatrix} (1-r_k)^2 & 2r_k(1-r_k) & r_k^2 \\ r_k(1-r_k) & (1-r_k)^2 + r_k^2 & r_k(1-r_k) \\ r_k^2 & 2r_k(1-r_k) & (1-r_k)^2 \end{bmatrix} \end{aligned} \tag{10.5}$$

which denotes a transition probability matrix from M_k to M_{k+1} (and is also a transition probability matrix from M_{k+1} to M_k), where r_k is the recombination frequency between M_k and M_{k+1}, the equation (10.4) can be expressed in matrix form as

$$\mathbf{p}_k^R = \mathbf{H}_D(r_k)\mathbf{H}_R(r_{k+1})\mathbf{c}$$

with $\mathbf{H}_D(r_k) = \mathbf{H}(r_k)\mathbf{I}_D$, $\mathbf{H}_R(r_{k+1}) = \mathbf{H}(r_{k+1})\mathbf{I}_R$,

$$\mathbf{I}_D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \qquad \mathbf{I}_R = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \mathbf{c} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

The function of matrices $\mathbf{I}_D$ and $\mathbf{I}_R$ is to make appropriate column elements of $\mathbf{H}$ zero.

Thus, in general,

$$\mathbf{p}_k^R = \mathbf{H}_{z_{k+1}}(r_k)\mathbf{H}_{z_{k+2}}(r_{k+1})\cdots\mathbf{H}_{z_l}(r_{l-1})\mathbf{c} \tag{10.6}$$

where $z_j = M$, D, R, 1, 0 or -1, depending on the information content of the phenotype of marker M_j. As specified above, $\mathbf{H}_{z_j} = \mathbf{H}\mathbf{I}_{z_j}$ with $\mathbf{I}_M = \mathbf{I}$ the identity matrix, $\mathbf{I}_1$, $\mathbf{I}_0$ and $\mathbf{I}_{-1}$ having 1 in one corresponding diagonal element and 0 everywhere else. The chain specified by (10.6) connects and sums together all relevant paths of joint probabilities of genotypes based on observed marker phenotypes in the right side of x_k. Similarly,

$$\mathbf{p}_k^L = \mathbf{I}_{z_k}\mathbf{H}_{z_{k-1}}(r_{k-1})\mathbf{H}_{z_{k-2}}(r_{k-2})\cdots\mathbf{H}_{z_i}(r_i)\mathbf{c}. \tag{10.7}$$

The function of $\mathbf{I}_{z_k}$ is to make appropriate row elements of $\mathbf{p}_k^L$ and thus $\mathbf{p}_k$ as well zero. Similarly, if $\mathbf{p}_k^R$ is defined to include z_k as well for some applications, *i.e.* $\mathbf{p}_k^R = \{P(z_k\cdots z_l|x_k)\}_{(3\times1)}$ (see below), $\mathbf{p}_k^R$ can also be calculated by (10.6) and then premultiplied by $\mathbf{I}_{z_k}$ just like (10.7). Usually, the testing position M_k is between markers and z_k is missing. In this case, the individual has non zero probability at all of the three genotypes. When the testing position M_k for a QTL is at a marker and the individual has the dominant phenotype at the position, $z_k = D$ and the individual has zero probability for $x_k = -1$.

Note that $\mathbf{H}_M(r_{k,k+1}) = \mathbf{H}_M(r_k)\mathbf{H}_M(r_{k+1})$ with $r_{k,k+1} = r_k + r_{k+1} - 2r_kr_{k+1}$. Also $\mathbf{H}_D(r_{k,k+1}) = \mathbf{H}_M(r_k)\mathbf{H}_D(r_{k+1})$ and $\mathbf{H}_R(r_{k,k+1}) = \mathbf{H}_M(r_k)\mathbf{H}_R(r_{k+1})$. Thus the operation of the chains can be shortened for intervals with missing markers.

In practice, $\mathbf{p}_k^R$ and $\mathbf{p}_k^L$ can be calculated first for all markers, which can be used later to obtain the conditional probabilities of genotypes for any position covered by markers in mapping QTL. For example, assuming $M_{k'}$ is a testing position for a QTL in an interval flanked by M_k and M_{k+1}, then $\mathbf{p}_{k'}^R = \mathbf{H}(r_{k'}^R)\mathbf{p}_{k+1}^R$ and $\mathbf{p}_{k'}^L = \mathbf{H}(r_{k'}^L)\mathbf{p}_k^L$ where $r_{k'}^R$ is the recombination frequency between $M_{k'}$ and M_{k+1} and $r_{k'}^L$ between M_k and $M_{k'}$.

Efficiency of the algorithm: Simulations were performed to investigate the effects of missing and dominant markers on mapping QTL. We simulated a chromosome of 80 cM in length with marker coverage at every 5 cM, 10 cM or 20 cM (three marker coverages) for an F_2 population. The linkage map is assumed to be known. Five marker compositions were simulated and compared: (a) all markers are codominant with no missing marker data; (b) all markers are codominant with 15% random missing marker data; (c) markers are codominant and dominant in alternate order; (d) markers are codominant, dominant and recessive in alternate order; and (e) markers are dominant and recessive in alternate order. One QTL was considered and simulated at 47.5 cM position. Analysis was performed by using simple interval mapping (Lander and Botstein 1989; Model III of Zeng 1994) with the conditional probability of the putative QTL genotype at a testing position calculated by (10.2).

The threshold used in reporting the power of the test was chosen to be LOD=2.3 for all the marker compositions. Although it is not strictly appropriate, this value was chosen merely for the convenience of comparison. The sample size is 150 and the replicates of simulation were 100. Results are presented in Table 10.1.

Results show that both statistical power of QTL detection and precision of QTL estimation generally decrease as more markers become missing or partially missing as expected. The power, however, does not change significantly when the marker density is 5 cM. There is also relatively little difference on the estimated standard deviations (SD) of estimates of QTL additive and dominance effects for different marker compositions. The standard deviations of estimates of QTL position increase noticeably only for case (e) when the marker density is 5 cM and for cases (d) and (e) when 10 and 20 cM. Significant decrease on the proportion of QTL mapped to the correct interval occurs also mostly for cases (d) and (e). Overall, the effect of different marker compositions on the power and precision of QTL mapping is small. This basically reflects the efficiency of the algorithm in utilizing all available marker information to infer the probability of QTL genotype.

It is also interesting to compare case (c) of 5 cM interval with case (a) of 10 cM interval. The later case approximately corresponds to the former case when the dominant marker data are not utilized in analysis. The results of case (c) of 5 cM are consistently close to those of case (a) of 5 cM than those of case (a) of 10 cM. Similar results can also be found when we compare case (c) of 10 cM with case (a) of 20 cM.

10.2 Extension to several experimental designs

We now generalize the above Markov chain analysis to many other commonly used experimental designs stemming from two inbred lines. We first outline a general algorithm and then specify it for different experimental designs.

Recently, Fisch, Ragot and Gay (1996) derived the conditional genotypic probability distribution of a testing position given two fully observed flanking markers for F_t populations by selfing and backcrosses of F_t to parental lines by using a Markov chain to link the transition of crossing-over events in different generations. They, however, employed two intervals involving genotypes of three loci in their analysis, and also did not analyze the dominant marker situation. As demonstrated above, the analysis can be performed only in one interval in different generations and multiple intervals can be linked by another Markov chain. This approach is particularly important when we consider multiple intervals involving dominant markers.

In the above F_2 analysis, marker genotyping and trait phenotyping are assumed to be performed on the same individual. In some QTL mapping experiments or commercial breeding programs, traits can be, however, measured on some progeny of the individuals whose marker composition is genotyped (*e.g.* Stuber *et al.* 1992; Beavis *et al.* 1994). In this analysis, we also take this situation into account as Fisch, Ragot and Gay (1996) did. Let $\{z\}^u$ denote a phenotypic observation of the set of markers $M_i, \cdots, M_l$ for an individual in population u, and x_k^v denote the putative QTL genotype at M_k in population v for the same individual (when $v = u$) or for the progeny of the individual (when $v \neq u$). Then

$$P(x_k^v|\{z\}^u) = \sum_{x_k^u} P(x_k^v x_k^u|\{z\}^u)$$

Table 10.1: Simulation results on estimated power, QTL additive effect (a), dominance effect (d), position (θ), and proportion of QTL mapped into the correct interval (p_{int}) for different marker compositions and densities with parameters $a = 0.5$, $d = 0.25$ and $\theta = 47.5$ cM

Parameter estimated	Marker composition[a]	Marker density (interval size)		
		5 cM	10 cM	20 cM
Power	(a)	0.95	0.94	0.88
	(b)	0.96	0.91	0.84
	(c)	0.95	0.91	0.88
	(d)	0.94	0.84	0.77
	(e)	0.94	0.84	0.84
a(SD)	(a)	0.51(0.12)	0.52(0.13)	0.51(0.13)
	(b)	0.51(0.12)	0.51(0.13)	0.51(0.13)
	(c)	0.51(0.12)	0.52(0.13)	0.51(0.13)
	(d)	0.51(0.12)	0.50(0.13)	0.49(0.14)
	(e)	0.51(0.12)	0.51(0.13)	0.53(0.14)
d(SD)	(a)	0.26(0.20)	0.24(0.21)	0.23(0.21)
	(b)	0.27(0.20)	0.23(0.21)	0.22(0.23)
	(c)	0.26(0.20)	0.24(0.22)	0.20(0.25)
	(d)	0.27(0.20)	0.21(0.22)	0.21(0.25)
	(e)	0.26(0.21)	0.21(0.22)	0.25(0.27)
θ(SD)	(a)	47.4(6.7)	49.5(5.6)	48.7(11.0)
	(b)	47.3(6.8)	49.4(6.4)	48.2(11.7)
	(c)	47.5(6.5)	49.5(5.8)	48.1(11.7)
	(d)	47.4(6.4)	48.2(7.6)	49.8(12.6)
	(e)	47.0(8.6)	49.0(8.2)	47.9(11.3)
p_{int}	(a)	0.63	0.65	0.73
	(b)	0.58	0.66	0.71
	(c)	0.60	0.64	0.75
	(d)	0.58	0.57	0.66
	(e)	0.56	0.59	0.69

[a] Marker composition: (a) all markers are codominant; (b) markers are codominant with 15% data missing at random; (c) markers are codominant and dominant in alternate order; (d) markers are codominant, dominant and recessive in alternate order; (e) markers are dominant and recessive in alternate order.

$$\begin{aligned} &= \sum_{x_k^u} P(x_k^u|\{z\}^u)P(x_k^v|x_k^u,\{z\}^u) \\ &= \sum_{x_k^u} P(x_k^u|\{z\}^u)P(x_k^v|x_k^u) \end{aligned} \tag{10.8}$$

Let

$$\begin{aligned} \mathbf{p}_k^v &= \{P(x_k^v|\{z\}^u)\}_{(3\times 1)} \\ \mathbf{q}_k^u &= \{P(x_k^u)\}_{(3\times 1)} \\ \mathbf{p}_k^{Ru} \circ \mathbf{p}_k^{Lu} &= \{P(\{z\}^u|x_k^u)\}_{(3\times 1)} \\ \mathbf{M}_{u\to v} &= \{P(x_k^v|x_k^u)\}_{(3\times 3)}. \end{aligned}$$

Equation (10.8) can be expressed as

$$\mathbf{p}_k^v = \frac{\mathbf{M}'_{u\to v}[\mathbf{q}_k^u \circ (\mathbf{p}_k^{Ru} \circ \mathbf{p}_k^{Lu})]}{\mathbf{q}_k^{u\prime}(\mathbf{p}_k^{Ru} \circ \mathbf{p}_k^{Lu})}. \tag{10.9}$$

Now it remains to determine $\mathbf{q}_k^u$, $\mathbf{p}_k^{Ru}$, $\mathbf{p}_k^{Lu}$ and $\mathbf{M}_{u\to v}$ for different experimental designs.

10.2.1 Selfed $\mathbf{F}_t$

We first extend the analysis from F_2 to selfed F_t for an arbitrary t. When $u = v$, $\mathbf{M}_{u\to v} = \mathbf{I}$. For $u = F_t$, it is known that

$$\mathbf{q}_k^{F_t\prime} = [\frac{1}{2} - \frac{1}{2^t},\ \ \frac{1}{2^{t-1}},\ \ \frac{1}{2} - \frac{1}{2^t}]. \tag{10.10}$$

To obtain $\mathbf{p}_k^{Ru}$ and $\mathbf{p}_k^{Lu}$, we need to derive $\mathbf{H}$ for $u = F_t$. For two loci, there are ten genotypes including two double heterozygotes. These ten genotypes are

$$[\frac{AB}{AB},\ \frac{AB}{Ab},\ \frac{Ab}{Ab},\ \frac{AB}{aB},\ \frac{AB}{ab},\ \frac{Ab}{aB},\ \frac{Ab}{ab},\ \frac{aB}{aB},\ \frac{aB}{ab},\ \frac{ab}{ab}].$$

Let $\mathbf{p}'_{F_t}$ be a row vector of frequencies of these ten genotypes in an F_t population. For F_1 which is a cross from two inbred lines,

$$\mathbf{p}'_{F_1} = [0,\ \ 0,\ \ 0,\ \ 0,\ \ 1,\ \ 0,\ \ 0,\ \ 0,\ \ 0,\ \ 0].$$

The transition probability matrix of these genotypes in two generations by selfing is

$$\mathbf{T} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{4} & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & \frac{1}{4} & 0 & 0 \\ \frac{w^2}{4} & \frac{rw}{2} & \frac{r^2}{4} & \frac{rw}{2} & \frac{w^2}{2} & \frac{r^2}{2} & \frac{rw}{2} & \frac{r^2}{4} & \frac{rw}{2} & \frac{w^2}{4} \\ \frac{r^2}{4} & \frac{rw}{2} & \frac{w^2}{4} & \frac{rw}{2} & \frac{r^2}{2} & \frac{w^2}{2} & \frac{rw}{2} & \frac{w^2}{4} & \frac{rw}{2} & \frac{r^2}{4} \\ 0 & 0 & \frac{1}{4} & 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & \frac{1}{4} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

where $w = 1 - r$. Then,

$$\mathbf{p}'_{F_t} = \mathbf{p}'_{F_1}\mathbf{T}^{t-1}. \tag{10.11}$$

The transition probability matrix equivalent to (10.5) for F_t is then

$$\mathbf{H}_{F_t} = \begin{bmatrix} \frac{\mathbf{p}_{F_t}[1]}{\mathbf{p}_{F_t}[1]+\mathbf{p}_{F_t}[2]+\mathbf{p}_{F_t}[3]} & \frac{\mathbf{p}_{F_t}[2]}{\mathbf{p}_{F_t}[1]+\mathbf{p}_{F_t}[2]+\mathbf{p}_{F_t}[3]} & \frac{\mathbf{p}_{F_t}[3]}{\mathbf{p}_{F_t}[1]+\mathbf{p}_{F_t}[2]+\mathbf{p}_{F_t}[3]} \\ \frac{\mathbf{p}_{F_t}[4]}{\mathbf{p}_{F_t}[4]+\mathbf{p}_{F_t}[5]+\mathbf{p}_{F_t}[6]+\mathbf{p}_{F_t}[7]} & \frac{\mathbf{p}_{F_t}[5]+\mathbf{p}_{F_t}[6]}{\mathbf{p}_{F_t}[4]+\mathbf{p}_{F_t}[5]+\mathbf{p}_{F_t}[6]+\mathbf{p}_{F_t}[7]} & \frac{\mathbf{p}_{F_t}[7]}{\mathbf{p}_{F_t}[4]+\mathbf{p}_{F_t}[5]+\mathbf{p}_{F_t}[6]+\mathbf{p}_{F_t}[7]} \\ \frac{\mathbf{p}_{F_t}[8]}{\mathbf{p}_{F_t}[8]+\mathbf{p}_{F_t}[9]+\mathbf{p}_{F_t}[10]} & \frac{\mathbf{p}_{F_t}[9]}{\mathbf{p}_{F_t}[8]+\mathbf{p}_{F_t}[9]+\mathbf{p}_{F_t}[10]} & \frac{\mathbf{p}_{F_t}[10]}{\mathbf{p}_{F_t}[8]+\mathbf{p}_{F_t}[9]+\mathbf{p}_{F_t}[10]} \end{bmatrix},$$

where $\mathbf{p}_{F_t}[i]$ is the ith element of $\mathbf{p}_{F_t}$. A general solution of $\mathbf{p}_{F_t}$ in terms of r and t can be found in Bulmer (1985 pp. 33). By using this result, it is shown that

$$\begin{aligned}
\mathbf{H}_{F_t}[1,1] = \mathbf{H}_{F_t}[3,3] &= \frac{1}{2^{t-1}-1}\left[\frac{2^{t-1}}{1+2r} - 1 - \frac{2^{t-1}(\frac{1}{2}-r)^t}{1+2r} + 2^{t-2}[\frac{1}{2} - r(1-r)]^{t-1}\right] \\
\mathbf{H}_{F_t}[1,2] = \mathbf{H}_{F_t}[3,2] &= \frac{1}{2^{t-1}-1}\left[1 - 2^{t-1}[\frac{1}{2} - r(1-r)]^{t-1}\right] \\
\mathbf{H}_{F_t}[1,3] = \mathbf{H}_{F_t}[3,1] &= \frac{1}{2^{t-1}-1}\left[\frac{2^t r}{1+2r} - 1 + \frac{2^{t-1}(\frac{1}{2}-r)^t}{1+2r} + 2^{t-2}[\frac{1}{2} - r(1-r)]^{t-1}\right] \\
\mathbf{H}_{F_t}[2,1] = \mathbf{H}_{F_t}[2,3] &= \frac{1}{2} - 2^{t-2}[\frac{1}{2} - r(1-r)]^{t-1} \\
\mathbf{H}_{F_t}[2,2] &= 2^{t-1}[\frac{1}{2} - r(1-r)]^{t-1}
\end{aligned} \tag{10.12}$$

where $\mathbf{H}_{F_t}[i,j]$ is the ith row and jth column element of $\mathbf{H}_{F_t}$.

It is easy to check that when $t = \infty$ (recombinant inbred lines),

$$\mathbf{H}_{F_\infty} = \frac{1}{1+2r}\begin{bmatrix} 1 & 0 & 2r \\ 0 & 0 & 0 \\ 2r & 0 & 1 \end{bmatrix}$$

which is the classical result given by Haldane and Waddington (1931). When $t = 2$, $\mathbf{H}_{F_2}$ reduces to (10.5) for a corresponding interval.

Then, given the transition probability matrix (10.12), $\mathbf{p}_k^{Ru}$ and $\mathbf{p}_k^{Lu}$ are given by (10.6) and (10.7), respectively, and together with the prior probability vector (10.10) give $\mathbf{p}_k^u$ by using (10.9).

When $v \neq u$ and $v = F_{\tilde{t}}$ for $\tilde{t} > t$, we need to derive $\mathbf{M}_{u\to v}$. Let

$$\mathbf{S} = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ 0 & 0 & 1 \end{bmatrix}$$

be the transition probability matrix between generations for three genotypes of a locus by selfing.

$$\mathbf{M}_{F_t\to F_{\tilde{t}}} = \mathbf{S}^{\tilde{t}-t} = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} - \frac{1}{2^{\tilde{t}-t+1}} & \frac{1}{2^{\tilde{t}-t}} & \frac{1}{2} - \frac{1}{2^{\tilde{t}-t+1}} \\ 0 & 0 & 1 \end{bmatrix}.$$

10.2.2 Random mating F_t

The random mating F_t population is another quite commonly used experimental design for gene mapping. This design has an advantage in separating closely linked QTL. The effect of further random mating on the conditional genotypic distribution $\mathbf{H}$ is to increase the recombination fraction between linked markers. For F_t, the transition probability matrix is still that defined by (10.5) with r replaced by

$$r^{(t)} = \frac{1}{2} - \frac{1}{2}(1-r)^{t-2}(1-2r) \tag{10.13}$$

(Falconer 1989; Darvasi and Soller 1995). For random mating populations, it is usually necessary to have $u = v$. $\mathbf{q}_k^u$ should remain the same as (10.3) for large random mating population.

10.2.3 Backcross from selfed F_t

When a selfed F_t population is backcrossed to P_1 or P_2 or test crossed to another inbred line, we need to infer only the gametic type produced by F_t. Let $\mathbf{g}'_{F_t}$ be a row vector of frequencies of the four gametic types $[AB,\ Ab,\ aB,\ ab]$ produced by F_t. Given (10.11),

$$\mathbf{g}'_{F_t} = \mathbf{p}'_{F_t}\mathbf{C} \tag{10.14}$$

where

$$\mathbf{C} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ \frac{w}{2} & \frac{r}{2} & \frac{r}{2} & \frac{w}{2} \\ \frac{r}{2} & \frac{w}{2} & \frac{w}{2} & \frac{r}{2} \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 1 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Because there are only two genotypes in a backcross at a locus, the transition probability matrix equivalent to (10.12) is

$$\begin{aligned} \mathbf{G}_{F_t} &= \begin{bmatrix} \frac{\mathbf{g}_{F_t}[1]}{\mathbf{g}_{F_t}[1]+\mathbf{g}_{F_t}[2]} & \frac{\mathbf{g}_{F_t}[2]}{\mathbf{g}_{F_t}[1]+\mathbf{g}_{F_t}[2]} \\ \frac{\mathbf{g}_{F_t}[3]}{\mathbf{g}_{F_t}[3]+\mathbf{g}_{F_t}[4]} & \frac{\mathbf{g}_{F_t}[4]}{\mathbf{g}_{F_t}[3]+\mathbf{g}_{F_t}[4]} \end{bmatrix} \\ &= \begin{bmatrix} \frac{1-(\frac{1}{2}-r)^t}{1+2r} + (\frac{1}{2}-r)^t & \frac{2r+(\frac{1}{2}-r)^t}{1+2r} - (\frac{1}{2}-r)^t \\ \frac{2r+(\frac{1}{2}-r)^t}{1+2r} - (\frac{1}{2}-r)^t & \frac{1-(\frac{1}{2}-r)^t}{1+2r} + (\frac{1}{2}-r)^t \end{bmatrix}. \end{aligned} \tag{10.15}$$

The definition of the two genotypes, however, depends on whether the backcrossed parental population is P_1 or P_2. Assuming $u = v$, $\mathbf{p}_k^u$ is calculated from (10.9) using $\mathbf{G}_{F_t}$

in the place of $\mathbf{H}$ in (10.6) and (10.7) and with $\mathbf{q}_k^u = [\frac{1}{2}, \quad \frac{1}{2}]$. Since missing and dominant (or recessive) phenotypes are equivalent in backcrosses as far as information provided, this calculation is equivalent to that by Martinez and Curnow (1994) when the backcross population is derived from F_1.

10.2.4 Backcross from random mating F_t

For the backcrosses from random mating F_t, the transition probability matrix is given by

$$\mathbf{G}_{F_t} = \begin{bmatrix} 1 - r^{(t+1)} & r^{(t+1)} \\ r^{(t+1)} & 1 - r^{(t+1)} \end{bmatrix} \tag{10.16}$$

with $r^{(t+1)}$ defined by (10.13). All other specifications are the same as for the backcrosses from selfed F_t.

10.2.5 Design III

Another special design which has been used extensively in QTL mapping in plants is Design III (*e.g.* Stuber *et al.* 1992; Xiao *et al.* 1995). This design was originally introduced by Comstock and Robinson (1952) for estimating the average degree of dominance for quantitative trait loci. In this design, random F_2 or F_t individuals were each backcrossed to both original parental inbred lines P_1 and P_2. Adapted for QTL mapping, markers are usually genotyped on (selfed) F_t individuals and quantitative traits are measured on the individuals of backcrosses $B_{t1} = F_t \times P_1$ and $B_{t2} = F_t \times P_2$.

Thus the design involves $u = F_t$ and $v = B_{t1}$ or B_{t2}, or more generally $v = B_{\tilde{t}1}$ or $B_{\tilde{t}2}$ for $\tilde{t} \geq t$. Let

$$\mathbf{C}_1 = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 1 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{C}_2 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 1 \end{bmatrix}.$$

$\mathbf{M}_{F_t \to B_{\tilde{t}1}} = \mathbf{S}^{\tilde{t}-t}\mathbf{C}_1$ and $\mathbf{M}_{F_t \to B_{\tilde{t}2}} = \mathbf{S}^{\tilde{t}-t}\mathbf{C}_2$. Other terms are the same as for selfed F_t.

As emphasized by Cockerham and Zeng (1996), when both backcrosses are available, QTL mapping analysis should be performed on both the backcrosses *simultaneously*.

Chapter 11

Multiple Trait Analysis

Many data for mapping quantitative trait loci (QTL) contain observations on multiple traits, or on one or several traits in multiple environments. With such data, we can ask questions like: Does a QTL have pleiotropic effects on multiple traits? Does a QTL show genotype-environment interaction? What is the nature of genetic correlation between different traits? Is the correlation due to pleiotropy or linkage in certain regions of a genome? Statistically this involves multiple trait analysis, as the expression of a trait in different environments can be regarded as different traits or different trait states (Falconer 1952).

In this section, we discuss statistical methods for analyzing multiple trait data in mapping QTL, following (Jiang and Zeng 1995).

11.1 Statistical models and likelihood analyses

11.1.1 Composite interval mapping model for multiple traits

We first formulate statistical models and likelihood analyses.

Suppose that we have a sample of n individuals from an F_2 population crossed from two inbred lines, with observations on m quantitative traits and on a number of codominant genetic markers. Let the value of each marker be recorded as 2, 1 and 0 for the homozygote in one parental line, heterozygote and homozygote in the other parental line, respectively. These markers can be mapped in linkage groups or mapped on chromosomes if the locations of some of them are known.

Further, let y_{jk} denote the value of the kth trait in the jth individual. To test for a QTL on a marker interval $(i,\ i+1)$, the statistical model for mapping QTL for one trait (Zeng 1994) can be readily extended for mapping QTL for multiple traits as

$$\begin{array}{l} y_{j1} = b_{01} + b_1^* x_j^* + d_1^* z_j^* + \sum_l^t (b_{l1} x_{jl} + d_{l1} z_{jl}) + e_{j1} \\ y_{j2} = b_{02} + b_2^* x_j^* + d_2^* z_j^* + \sum_l^t (b_{l2} x_{jl} + d_{l2} z_{jl}) + e_{j2} \\ \vdots \qquad\qquad\qquad\qquad \vdots \\ y_{jm} = b_{0m} + b_m^* x_j^* + d_m^* z_j^* + \sum_l^t (b_{lm} x_{jl} + d_{lm} z_{jl}) + e_{jm} \end{array} \tag{11.1}$$

$$j = 1, \cdots, n$$

where

y_{jk} is the phenotypic value for trait k in individual j,

b_{0k} is the mean effect of the model for trait k,

b_k^* is the additive effect of the putative QTL on trait k,

x_j^* counts the number of the allele at the putative QTL from one of the two parental lines, say parent P_1 [taking values of 2, 1 and 0 with probabilities depending on genotypes of the markers i and $i+1$ flanking the putative QTL and the recombination frequencies between the QTL and the markers (Section 7)],

d_k^* is the dominance effect of the putative QTL on trait k,

z_j^* is an indicator variable of the heterozygosity at the QTL taking values 1 and 0 for heterozygote and homozygotes (Section 7),

x_{jl} and z_{jl} are corresponding variables for marker l, assuming t markers are selected for controlling residual genetic variation,

b_{lk} and d_{lk} are partial regression coefficients of y_{jk} on x_{jl} and z_{jl},

e_{jk} is the residual effect on trait k for individual j.

It is assumed that the residual effects e_{jk}'s (the error terms) are correlated among traits within individuals with covariance $\text{Cov}(e_{jk}, e_{jl}) = \sigma_{kl} = \rho_{kl}\sigma_k\sigma_l$, but are independent among individuals. For likelihood analysis, it will be further assumed that e_{jk}'s are multivariate normally distributed among individuals with means zero and covariance matrix

$$\mathbf{V} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m1} & \sigma_{m2} & \cdots & \sigma_m^2 \end{pmatrix}. \tag{11.2}$$

In matrix notation, model (11.1) can be expressed as

$$\underset{n\times m}{\mathbf{Y}} = \underset{n\times 1}{\mathbf{x}^*}\ \underset{1\times m}{\mathbf{b}^*} + \underset{n\times 1}{\mathbf{z}^*}\ \underset{1\times m}{\mathbf{d}^*} + \underset{n\times(2t+1)}{\mathbf{X}}\ \underset{(2t+1)\times m}{\mathbf{B}} + \underset{n\times m}{\mathbf{E}} \tag{11.3}$$

where $\mathbf{Y}$ is a $(n \times m)$ matrix of y_{jk}, $\mathbf{x}^*$ is a $(n \times 1)$ column vector of x_j^*, $\mathbf{z}^*$ is a $(n \times 1)$ column vector of z_j^*, $\mathbf{b}^*$ is a $(1 \times m)$ row vector of b_k^*, $\mathbf{d}^*$ is a $(1 \times m)$ row vector of d_k^*, $\mathbf{X}$ is a $[n \times (2t+1)]$ matrix of data on t markers, x_{jl}'s and z_{jl}'s, fitted in the model as background control, including also the mean effect, $\mathbf{B}$ is a $[(2t+1) \times m]$ matrix of b_{lk}, d_{lk}, and b_{0k}, and $\mathbf{E}$ is a $(n \times m)$ matrix of e_{jk}.

In model (11.1), for simplicity it is assumed that t markers are each fitted with additive and dominance effects for genetic background control. Since many different markers can be fitted in the model, the question of selecting how many and what markers to be fitted in the model becomes a major issue for mapping QTL. A number of considerations have to be taken into account for such a selection process, and many relevant issues involved have been discussed in Zeng (1994). Here, we concentrate our discussion on the extension of the method to multiple trait analysis and avoid lengthy discussion on this issue.

11.1.2 Likelihood analysis

Given model (11.1), which is defined as a mixture model, and the assumption of multivariate normal distribution of error terms, the likelihood function of the data is defined as

$$L_1 = \prod_{j=1}^{n} \left[p_{2j} f_2(\mathbf{y}_j) + p_{1j} f_1(\mathbf{y}_j) + p_{0j} f_0(\mathbf{y}_j) \right] \tag{11.4}$$

where p_{2j}, p_{1j} and p_{0j} denote the prior probability of x_j^* taking values 2, 1 and 0, respectively, for the three genotypes of the putative QTL, and $f_2(\mathbf{y}_j)$, $f_1(\mathbf{y}_j)$ and $f_0(\mathbf{y}_j)$ represent the multivariate normal density functions of the vector variable $\mathbf{y}_j$ (the jth row of $\mathbf{Y}$) with means $\mathbf{u}_{j2} = \mathbf{x}_j\mathbf{B} + 2\mathbf{b}^*$, $\mathbf{u}_{j1} = \mathbf{x}_j\mathbf{B} + \mathbf{b}^* + \mathbf{d}^*$, and $\mathbf{u}_{j0} = \mathbf{x}_j\mathbf{B}$, respectively, and the covariance matrix (11.2).

Then, by applying standard maximum likelihood procedures, the maximum likelihood estimates of parameters can be found as in the following. As in Zeng (1994), these maximum likelihood estimates can be computed by iteration through an ECM algorithm (Meng and Rubin 1993) which is a special version of general EM algorithms. In the $(v+1)$th iteration, the E-step calculates the posterior probabilities of individual j being a particular genotype at the putative QTL as

$$\begin{aligned}
q_{2j}^{(v+1)} &= p_{2j} f_2^{(v)}(\mathbf{y}_j) / \left[p_{2j} f_2^{(v)}(\mathbf{y}_j) + p_{1j} f_1^{(v)}(\mathbf{y}_j) + p_{0j} f_0^{(v)}(\mathbf{y}_j) \right], \\
q_{1j}^{(v+1)} &= p_{1j} f_1^{(v)}(\mathbf{y}_j) / \left[p_{2j} f_2^{(v)}(\mathbf{y}_j) + p_{1j} f_1^{(v)}(\mathbf{y}_j) + p_{0j} f_0^{(v)}(\mathbf{y}_j) \right], \\
q_{0j}^{(v+1)} &= p_{0j} f_0^{(v)}(\mathbf{y}_j) / \left[p_{2j} f_2^{(v)}(\mathbf{y}_j) + p_{1j} f_1^{(v)}(\mathbf{y}_j) + p_{0j} f_0^{(v)}(\mathbf{y}_j) \right],
\end{aligned} \tag{11.5}$$

where $f_2^{(v)}(\mathbf{y}_j)$, $f_1^{(v)}(\mathbf{y}_j)$ and $f_0^{(v)}(\mathbf{y}_j)$ are the corresponding normal density functions with parameters replaced by estimates in the vth iteration. In the CM-step (*i.e.* Conditional Maximization step), parameters in $f_2(\mathbf{y}_j)$, $f_1(\mathbf{y}_j)$ and $f_0(\mathbf{y}_j)$ are divided into three groups, ($\mathbf{b}^*$, $\mathbf{d}^*$), $\mathbf{B}$, $\mathbf{V}$, and estimated consecutively between groups, but simultaneously within each group. These estimators can be shown to be

$$\mathbf{b}^{*(v+1)} = \mathbf{q}_2^{(v+1)\prime}(\mathbf{Y} - \mathbf{X}\mathbf{B}^{(v)}) / (2\mathbf{q}_2^{(v+1)\prime}\mathbf{1}) \tag{11.6}$$

$$\mathbf{d}^{*(v+1)} = \left[\mathbf{q}_1^{(v+1)\prime} / (\mathbf{q}_1^{(v+1)\prime}\mathbf{1}) - \mathbf{q}_2^{(v+1)\prime} / (2\mathbf{q}_2^{(v+1)\prime}\mathbf{1}) \right] (\mathbf{Y} - \mathbf{X}\mathbf{B}^{(v)}) \tag{11.7}$$

$$\mathbf{B}^{(v+1)} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \left[\mathbf{Y} - (2\mathbf{q}_2^{(v+1)} + \mathbf{q}_1^{(v+1)})\mathbf{b}^{*(v+1)} - \mathbf{q}_1^{(v+1)}\mathbf{d}^{*(v+1)} \right] \tag{11.8}$$

$$\begin{aligned}\mathbf{V}^{(v+1)} &= \Big[(\mathbf{Y}-\mathbf{X}\mathbf{B}^{(v+1)})'(\mathbf{Y}-\mathbf{X}\mathbf{B}^{(v+1)}) - 4(\mathbf{q}_2^{(v+1)\prime}\mathbf{1})\mathbf{b}^{*(v+1)\prime}\mathbf{b}^{*(v+1)} \\ &\quad -(\mathbf{q}_1^{(v+1)\prime}\mathbf{1})(\mathbf{b}^{*(v+1)}+\mathbf{d}^{*(v+1)})'(\mathbf{b}^{*(v+1)}+\mathbf{d}^{*(v+1)})\Big]/n \qquad (11.9)\end{aligned}$$

where $\mathbf{q}_2^{(v+1)}$ and $\mathbf{q}_1^{(v+1)}$ are $(n\times 1)$ vectors of $q_{2j}^{(v+1)}$ and $q_{1j}^{(v+1)}$, and $\mathbf{1}$ is a column vector of ones. A prime represents the transpose of a matrix or a vector.

The calculation begins with $q_{2j}^{(0)}=p_{2j}$, $q_{1j}^{(0)}=p_{1j}$, $q_{0j}^{(0)}=p_{0j}$, and some starting values for $\mathbf{b}^{*(0)}$ and $\mathbf{d}^{*(0)}$ (one possible choice is to set them to zero). Iterations are then made between (11.5), (11.6), (11.7), (11.8), and (11.9), and terminated when a predetermined criterion is satisfied. The criterion for termination is set to be that the changes of the parameter estimates, or the increment of the log-likelihood value, at each iteration become less than ε (a small positive number, say, 10^{-8}). The final estimates are denoted as $\hat{\mathbf{b}}^*$, $\hat{\mathbf{d}}^*$, $\hat{\mathbf{B}}$ and $\hat{\mathbf{V}}$, which will then be used for the calculation of the maximum likelihood value for hypothesis testing.

The log-likelihood of (11.4) is calculated, with the parameters replaced by the estimates, as

$$\begin{aligned}\ln(L_1) &= K-(n/2)\ln(|\hat{\mathbf{V}}|)+\sum_{j=1}^{n}\ln\Big\{p_{2j}\exp\Big[-(1/2)(\mathbf{y}_j-2\hat{\mathbf{b}}^*-\mathbf{x}_j\hat{\mathbf{B}})\hat{\mathbf{V}}^{-1}(\mathbf{y}_j-2\hat{\mathbf{b}}^*-\mathbf{x}_j\hat{\mathbf{B}})'\Big] \\ &\quad +p_{1j}\exp\Big[-(1/2)(\mathbf{y}_j-\hat{\mathbf{b}}^*-\hat{\mathbf{d}}^*-\mathbf{x}_j\hat{\mathbf{B}})\hat{\mathbf{V}}^{-1}(\mathbf{y}_j-\hat{\mathbf{b}}^*-\hat{\mathbf{d}}^*-\mathbf{x}_j\hat{\mathbf{B}})'\Big] \\ &\quad +p_{0j}\exp\Big[-(1/2)(\mathbf{y}_j-\mathbf{x}_j\hat{\mathbf{B}})\hat{\mathbf{V}}^{-1}(\mathbf{y}_j-\mathbf{x}_j\hat{\mathbf{B}})'\Big]\Big\} \\ &= K-(n/2)\ln(|\hat{\mathbf{V}}|)-(1/2)\sum_{j=1}^{n}(\mathbf{y}_j-\mathbf{x}_j\hat{\mathbf{B}})\hat{\mathbf{V}}^{-1}(\mathbf{y}_j-\mathbf{x}_j\hat{\mathbf{B}})' \\ &\quad +\sum_{j=1}^{n}\ln\Big\{p_{2j}\exp\Big[2\hat{\mathbf{b}}^*\hat{\mathbf{V}}^{-1}(\mathbf{y}_j-\hat{\mathbf{b}}^*-\mathbf{x}_j\hat{\mathbf{B}})'\Big] \\ &\quad +p_{1j}\exp\Big[(\hat{\mathbf{b}}^*+\hat{\mathbf{d}}^*)\hat{\mathbf{V}}^{-1}(\mathbf{y}_j-\hat{\mathbf{b}}^*/2-\hat{\mathbf{d}}^*/2-\mathbf{x}_j\hat{\mathbf{B}})'\Big]+p_{0j}\Big\} \qquad (11.10)\end{aligned}$$

where $|\hat{\mathbf{V}}|$ is the determinant of the covariance matrix, and $K=-nm\ln(2\pi)/2$.

11.2 Hypothesis tests of QTL effects

In hypothesis testing, model (11.1) is usually called the full model. With some parameter values constrained to some specific values, a number of null hypotheses can be constructed and tested. For mapping QTL, we are mostly concerned with testing hypotheses about the additive and dominance effects of QTL. However, before we test other hypotheses, we need first to test for the presence of QTL. Without losing generality, we restrict our discussion to two traits in the following.

11.2.1 Joint mapping for QTL on two traits

With phenotypic observations on two traits, mapping for QTL can be performed for each trait individually or jointly on both traits. Under the joint mapping, the hypotheses to be tested are

$$\begin{array}{l} H_0: \ b_1^* = 0, \ \ d_1^* = 0, \ \ b_2^* = 0, \ \ d_2^* = 0 \\ H_1: \ \text{At least one of them is not zero.} \end{array} \tag{11.11}$$

The log-likelihood under H_0 is then

$$\ln(L_0) = \ln\left[\prod_{j=1}^{n} f_0(\mathbf{y}_j)\right] = K - (n/2)\ln(|\hat{\mathbf{V}}_0|) - nm/2 \tag{11.12}$$

where $\hat{\mathbf{V}}_0 = (\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}_0)'(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}_0)/n$ and $\hat{\mathbf{B}}_0 = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. The test is performed with a likelihood ratio statistic

$$LR_1 = -2\ln(L_0/L_1). \tag{11.13}$$

Under H_0, the likelihood ratio LR_1 will be approximately chi-square distributed. The determination of the critical value for the test is, however, very complicated. The complication is due largely to the fact that the test is usually performed for the whole genome – a situation of multiple tests. With multiple tests, the critical level for each test has to be adjusted. It has been shown (Zeng, 1994) that, for the composite interval mapping, tests in different intervals are close to be independent except for those in adjacent intervals which are slightly correlated. Thus, if we choose α' as the genome-wise error rate, the error rate of the test per interval, α, can be approximated by using the Bonferroni correction as $\alpha = 1 - (1-\alpha')^{1/M}$, or to a good approximation as $\alpha = \alpha'/M$, where M is the number of intervals involved in the test. The maximum of the test statistic within an interval can be roughly approximated by a chi-square distribution with a degree of freedom $2m+1$ (the number of parameters under the test including one for the position of the putative QTL). Thus, in practice, we may use $\chi^2_{\alpha/M,2m+1}$ to approximate the critical value of the test in this situation (for two traits $m = 2$). However, we emphasize that the correct determination of appropriate critical value for this test is a very complicated statistical issue. The above recommendation is suggested only for a very rough approximation. Recently, Churchill and Doerge (1994) proposed to use a permutation test to empirically estimate the genome-wise critical value for a given data set and a given test. Their method can be extended for multiple trait analysis.

Why do we need to perform joint mapping? First, the joint analysis on two traits can provide formal procedures for testing a number of biologically interesting hypotheses, such as pleiotropic effects of QTL, QTL by environment interaction and pleiotropy vs. close linkage, as shown below. Second if the putative QTL has pleiotropic effects on both traits, the joint mapping on both traits may perform better than mapping on each trait separately.

11.2.2 Testing pleiotropic effects

Given a genome position or region where the presence of a QTL is indicated by joint mapping, statistical tests can proceed to test whether the QTL has pleiotropic effects on

both traits. On the assumption that there is only one QTL in the relevant region, which has effects on either one or both of two traits, hypotheses can be formulated as

$$\begin{array}{l} H_{10}: \ b_1^* = 0, d_1^* = 0, b_2^* \neq 0, d_2^* \neq 0 \ \text{ given a position for a QTL} \\ H_{11}: \ b_1^* \neq 0, d_1^* \neq 0, b_2^* \neq 0, d_2^* \neq 0 \ \text{at the position given by } H_{10}; \end{array} \tag{11.14}$$

and

$$\begin{array}{l} H_{20}: \ b_1^* \neq 0, d_1^* \neq 0, b_2^* = 0, d_2^* = 0 \ \text{at the position given by } H_{10} \\ H_{21}: \ b_1^* \neq 0, d_1^* \neq 0, b_2^* \neq 0, d_2^* \neq 0 \ \text{at the position given by } H_{10}. \end{array} \tag{11.15}$$

A test for pleiotropic effects is then equivalent to the test of both (11.14) and (11.15). Only rejecting both the null hypotheses (*i.e.*, H_{10} and H_{20}) will suggest the presence of pleiotropic effects.

Although each of (11.14) and (11.15) shows restriction only on one trait, the tests will not be the same as for each trait separately since two traits are correlated. When the two traits are correlated, this test will have more power than separate analyses. The estimates of model parameters under H_{10} and H_{20} can be obtained as in joint mapping of (11.5)–(11.9) except that some estimates in (11.6) and (11.7) are set to zero. The likelihood ratio test statistics for (11.14) and (11.15) can then be calculated correspondingly from (11.10) in a ratio of the likelihoods with and without constraints. Fixing a testing position or region by the joint mapping for this test is consistent with the pleiotropic hypothesis. Since the testing position is fixed, the likelihood ratio test statistics under the null hypotheses of (11.14) and (11.15) will each be asymptotically chi-square distributed with two degrees of freedom.

11.2.3 Testing pleiotropic effects against close linkage

Although rejecting both H_{10} of (11.14) and H_{20} of (11.15) is consistent with the hypothesis of pleiotropic effects of a QTL, the test itself does not distinguish whether the significant effect is due to one QTL having pleiotropic effects on both traits, or possibly two (or more) closely linked QTL each having a predominant effect on one trait only. Two closely linked QTL each with an effect on only one trait may behave like one pleiotropic QTL in joint mapping. Also one pleiotropic QTL may be estimated as two QTL at two nearby but different positions if each trait is analyzed separately. Thus in mapping for QTL, for some regions there may exist sufficient interests to distinguish these two possibilities. Undoubtedly, distinguishing these two cases has important implications in genetics and breeding.

Clearly, this test of pleiotropy vs. close linkage is for some specific genome regions only. The regions to be tested are first determined by joint mapping. Only those genome regions which are significant under joint mapping may be suitable for this test. Relatively loosely linked pleiotropic or non-pleiotropic QTL (*i.e.,* separated by several relatively large, say 10 cM, marker intervals) may be detected by joint mapping, and may not be necessary to this test to distinguish them, although this test can be applied to those situations.

To test the hypotheses of pleiotropy vs. close linkage at some significant regions, the likelihood analysis has to be reformulated. Let two QTL, each with an effect on one trait

only, have positions symbolically specified by $p(1)$ for the QTL having an effect on trait 1 and $p(2)$ for the QTL having an effect on trait 2 (if the two positions are in the same marker interval, $p(1)$ and $p(2)$ are then defined as the ratios of the recombination frequencies between a marker and the two positions, respectively, and between the two flanking markers). The hypotheses can then be formulated as

$$\begin{array}{l} H_0: \ p(1) = p(2) \\ H_1: \ p(1) \neq p(2). \end{array} \tag{11.16}$$

The H_1 here is a special case of many possible alternatives. A more general alternative may be that both QTL have pleiotropic effects. This alternative is, however, the hypothesis for mapping two closely linked pleiotropic QTL. Although this hypothesis can be tested, we confine our attention here to the alternative of two non-pleiotropic QTL.

The log-likelihood of H_0 in (11.16) is given by (11.10). The statistical model for H_1 in (11.16) is, however, given by

$$\begin{array}{l} y_{j1} = b_{01} + b_1^* x_{1j}^* + d_1^* z_{1j}^* + \sum_l^t (b_{l1} x_{jl} + d_{l1} z_{jl}) + e_{j1} \\ y_{j2} = b_{02} + b_2^* x_{2j}^* + d_2^* z_{2j}^* + \sum_l^t (b_{l2} x_{jl} + d_{l2} z_{jl}) + e_{j2}. \end{array} \tag{11.17}$$

Model (11.17) should be the same as model (11.1) for $m = 2$, except that $(x_{1j}^*,\ z_{1j}^*)$, and $(x_{2j}^*,\ z_{2j}^*)$ are now defined for two QTL at two different positions in one or some nearby marker intervals. Note that when the test and search cover several marker intervals, the markers inside the search region should not be used for background control, otherwise the models under the null and alternative hypotheses can be inconsistent on the markers used for background control.

Model (11.17) is actually a mixture model with nine components since recombination can result in nine possible QTL genotypes in an F_2 population for two QTL. Let $p_{i_1 i_2 j}$ (i_1, $i_2 = 0$, 1 and 2) be the probability of individual j having genotype i_1 for a putative QTL affecting trait 1 and i_2 for another QTL affecting trait 2 for given two testing positions for two QTL. The likelihood function is given by

$$L_2 = \prod_{j=1}^{n} \sum_{i_1=0}^{2} \sum_{i_2=0}^{2} p_{i_1 i_2 j} f_{i_1 i_2}(\mathbf{y}_j) \tag{11.18}$$

where $f_{i_1 i_2}(\mathbf{y}_j)$ is a bivariate normal density function for $\mathbf{y}_j$ with a mean vector

$$\mathbf{u}'_{i_1 i_2 j} = \begin{pmatrix} \mathbf{x}_j \mathbf{b}_1 + i_1 b_1^* + \delta(i_1) d_1^* \\ \mathbf{x}_j \mathbf{b}_2 + i_2 b_2^* + \delta(i_2) d_2^* \end{pmatrix}$$

and covariance matrix (11.2), where the indicator function $\delta(i_1) = 1$ if $i_1 = 1$ and 0 otherwise.

The probability $p_{i_1 i_2 j}$ can be inferred from the observed genotypes of the flanking markers. If the two putative QTL are tested in different marker intervals, the probability of

Table 11.1: Probability of QTL genotype given flanking marker genotype for two QTL within a marker interval

Marker genotype	QTL genotype						
	$Q_1Q_1Q_2Q_2$	$Q_1Q_1Q_2q_2$	$Q_1Q_1q_2q_2$	$Q_1q_1Q_2Q_2$	$Q_1q_1Q_2q_2$	$Q_1q_1q_2q_2$	q_1q
	(22)	(21)	(20)	(12)	(11)	(10)	
$M_1M_1M_2M_2$	1	0	0	0	0	0	
$M_1M_1M_2m_2$	p_3	p_2	0	0	p_1	0	
$M_1M_1m_2m_2$	p_3^2	$2p_2p_3$	p_2^2	0	$2p_1p_3$	$2p_1p_2$	
$M_1m_1M_2M_2$	p_1	0	0	p_2	p_3	0	
$M_1m_1M_2m_2$	δp_1p_3	δp_1p_2	0	δp_2p_3	$\delta(p_1^2+p_2^2+p_3^2)+(1-\delta)$	δp_2p_3	
$M_1m_1m_2m_2$	0	0	0	0	p_3	p_2	
$m_1m_1M_2M_2$	p_1^2	$2p_1p_2$	0	$2p_2p_3$	$2p_1p_3$	0	
$m_1m_1M_2m_2$	0	0	0	0	p_1	0	
$m_1m_1m_2m_2$	0	0	0	0	0	0	

It is assumed that the order of markers and QTL are $M_1Q_1Q_2M_2$ and the recombination frequencies between M_1Q_1, Q_1Q_2, Q_2M_1, M_1M_2 are r_1, r_2, r_3, r respectively. Double recombination is ignored. $p_1 = r_1/r$, $p_2 = r_2/r$, $p_3 = r_3/r$, and $\delta = r^2/[(1-r)^2 + r^2]$.

QTL genotype can be calculated independently for each QTL, *i.e.*, $p_{i_1i_2j} = p_{i_1j}p_{i_2j}$, assuming that there is no crossing-over interference. If the two putative QTL are tested in the same marker interval, $p_{i_1i_2j}$ can be calculated from Table 11.1.

The E-step in this case is to calculate the posterior probabilities of individual j having genotype i_1 for QTL 1 affecting trait 1 at position $p(1)$ and i_2 for QTL 2 affecting trait 2 at position $p(2)$

$$q_{i_1i_2j}^{(v+1)} = p_{i_1i_2j}\hat{f}_{i_1i_2}^{(v)}(\mathbf{y}_j) / \sum_{k_1=0}^{2}\sum_{k_2=0}^{2} p_{k_1k_2j}\hat{f}_{k_1k_2}^{(v)}(\mathbf{y}_j) \quad \text{for } i_1, i_2 = 0, 1, 2. \tag{11.19}$$

The CM-step is to calculate

$$\begin{aligned} b_1^{*(v+1)} &= \left\{\mathbf{q}_{2\cdot}^{(v+1)\prime}\left[(\mathbf{y}_1 - \mathbf{X}\mathbf{b}_1^{(v)}) - (\rho^{(v)}\sigma_1^{(v)}/\sigma_2^{(v)})(\mathbf{y}_2 - \mathbf{X}\mathbf{b}_2^{(v)})\right]\right. \\ &\quad \left. +(\rho^{(v)}\sigma_1^{(v)}/\sigma_2^{(v)})\left[(2\mathbf{q}_{22}^{(v+1)} + \mathbf{q}_{21}^{(v+1)})'\mathbf{1}b_2^{*(v)} + \mathbf{q}_{21}^{(v+1)\prime}\mathbf{1}d_2^{*(v)}\right]\right\} / \\ &\quad (2\mathbf{q}_{2\cdot}^{(v+1)\prime}\mathbf{1}) \end{aligned} \tag{11.20}$$

$$\begin{aligned} d_1^{*(v+1)} &= \left\{\mathbf{q}_{1\cdot}^{(v+1)\prime}\left[(\mathbf{y}_1 - \mathbf{X}\mathbf{b}_1^{(v)}) - (\rho^{(v)}\sigma_1^{(v)}/\sigma_2^{(v)})(\mathbf{y}_2 - \mathbf{X}\mathbf{b}_2^{(v)})\right]\right. \\ &\quad \left. +(\rho^{(v)}\sigma_1^{(v)}/\sigma_2^{(v)})\left[(2\mathbf{q}_{12}^{(v+1)} + \mathbf{q}_{11}^{(v+1)})'\mathbf{1}b_2^{*(v)} + \mathbf{q}_{11}^{(v+1)\prime}\mathbf{1}d_2^{*(v)}\right]\right\} / \\ &\quad (\mathbf{q}_{1\cdot}^{(v+1)\prime}\mathbf{1}) - b_1^{*(v+1)} \end{aligned} \tag{11.21}$$

Table 11.1: Probability of QTL genotype given flanking marker genotype for two QTL within a marker interval

Marker genotype	QTL genotype								
	$Q_1Q_1Q_2Q_2$	$Q_1Q_1Q_2q_2$	$Q_1Q_1q_2q_2$	$Q_1q_1Q_2Q_2$	$Q_1q_1Q_2q_2$	$Q_1q_1q_2q_2$	$q_1q_1Q_2Q_2$	$q_1q_1Q_2q_2$	$q_1q_1q_2q_2$
	(22)	(21)	(20)	(12)	(11)	(10)	(02)	(01)	(00)
$M_1M_1M_2M_2$	1	0	0	0	0	0	0	0	0
$M_1M_1M_2m_2$	p_3	p_2	0	0	p_1	0	0	0	0
$M_1M_1m_2m_2$	p_3^2	$2p_2p_3$	p_2^2	0	$2p_1p_3$	$2p_1p_2$	0	0	p_1^2
$M_1m_1M_2M_2$	p_1	0	0	p_2	p_3	0	0	0	0
$M_1m_1M_2m_2$	δp_1p_3	δp_1p_2	0	δp_2p_3	$\delta(p_1^2+p_2^2+p_3^2)+(1-\delta)$	δp_2p_3	0	δp_1p_2	δp_1p_3
$M_1m_1m_2m_2$	0	0	0	0	p_3	p_2	0	0	p_1
$m_1m_1M_2M_2$	p_1^2	$2p_1p_2$	0	$2p_2p_3$	$2p_1p_3$	0	p_2^2	0	p_3^2
$m_1m_1M_2m_2$	0	0	0	0	p_1	0	0	p_2	p_3
$m_1m_1m_2m_2$	0	0	0	0	0	0	0	0	1

It is assumed that the order of markers and QTL are $M_1Q_1Q_2M_2$ and the recombination frequencies between M_1Q_1, Q_1Q_2, Q_2M_1, M_1M_2 are r_1, r_2, r_3, r respectively. Double recombination is ignored. $p_1 = r_1/r$, $p_2 = r_2/r$, $p_3 = r_3/r$, and $\delta = r^2/[(1-r)^2 + r^2]$.

$$
\begin{aligned}
b_2^{*(v+1)} &= \left\{\mathbf{q}_{\cdot 2}^{(v+1)\prime}\left[(\mathbf{y}_2-\mathbf{X}\mathbf{b}_2^{(v)})-(\rho^{(v)}\sigma_2^{(v)}/\sigma_1^{(v)})(\mathbf{y}_1-\mathbf{X}\mathbf{b}_1^{(v)})\right]\right. \\
&\quad \left.+(\rho^{(v)}\sigma_2^{(v)}/\sigma_1^{(v)})\left[(2\mathbf{q}_{22}^{(v+1)}+\mathbf{q}_{12}^{(v+1)})'\mathbf{1}b_1^{*(v+1)}+\mathbf{q}_{12}^{(v+1)\prime}\mathbf{1}d_1^{*(v+1)}\right]\right\}/ \\
&\quad (2\mathbf{q}_{\cdot 2}^{(v+1)\prime}\mathbf{1}) \qquad (11.22)\\
d_2^{*(v+1)} &= \left\{\mathbf{q}_{\cdot 1}^{(v+1)\prime}\left[(\mathbf{y}_2-\mathbf{X}\mathbf{b}_2^{(v)})-(\rho^{(v)}\sigma_2^{(v)}/\sigma_1^{(v)})(\mathbf{y}_1-\mathbf{X}\mathbf{b}_1^{(v)})\right]\right. \\
&\quad \left.+(\rho^{(v)}\sigma_2^{(v)}/\sigma_1^{(v)})\left[(2\mathbf{q}_{21}^{(v+1)}+\mathbf{q}_{11}^{(v+1)})'\mathbf{1}b_1^{*(v+1)}+\mathbf{q}_{11}^{(v+1)\prime}\mathbf{1}d_1^{*(v+1)}\right]\right\}/ \\
&\quad (\mathbf{q}_{\cdot 1}^{(v+1)\prime}\mathbf{1})-b_2^{*(v+1)} \qquad (11.23)\\
\mathbf{B}^{(v+1)} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{(v+1)} \qquad (11.24)\\
\mathbf{V}^{(v+1)} &= (\mathbf{W}^{(v+1)}-\mathbf{X}\mathbf{B}^{(v+1)})'(\mathbf{W}^{(v+1)}-\mathbf{X}\mathbf{B}^{(v+1)})/n \qquad (11.25)
\end{aligned}
$$

where

$$
\begin{aligned}
\mathbf{B}^{(v+1)} &= (\mathbf{b}_1^{(v+1)} \quad \mathbf{b}_2^{(v+1)}) \\
\mathbf{W}^{(v+1)\prime} &= \begin{pmatrix} \mathbf{y}_1'-(2\mathbf{q}_{2\cdot}^{(v+1)}+\mathbf{q}_{1\cdot}^{(v+1)})'b_1^{*(v+1)}-\mathbf{q}_{1\cdot}^{(v+1)\prime}d_1^{*(v+1)} \\ \mathbf{y}_2'-(2\mathbf{q}_{\cdot 2}^{(v+1)}+\mathbf{q}_{\cdot 1}^{(v+1)})'b_2^{*(v+1)}-\mathbf{q}_{\cdot 1}^{(v+1)\prime}d_2^{*(v+1)} \end{pmatrix} \\
\mathbf{q}_{2\cdot}^{(v+1)} &= \mathbf{q}_{22}^{(v+1)}+\mathbf{q}_{21}^{(v+1)}+\mathbf{q}_{20}^{(v+1)} \\
\mathbf{q}_{1\cdot}^{(v+1)} &= \mathbf{q}_{12}^{(v+1)}+\mathbf{q}_{11}^{(v+1)}+\mathbf{q}_{10}^{(v+1)} \\
\mathbf{q}_{\cdot 2}^{(v+1)} &= \mathbf{q}_{22}^{(v+1)}+\mathbf{q}_{12}^{(v+1)}+\mathbf{q}_{02}^{(v+1)} \\
\mathbf{q}_{\cdot 1}^{(v+1)} &= \mathbf{q}_{21}^{(v+1)}+\mathbf{q}_{11}^{(v+1)}+\mathbf{q}_{01}^{(v+1)}.
\end{aligned}
$$

After the convergence of the ECM algorithm, the log-likelihood value of (11.18) will be calculated from

$$
\ln(L_2)=K-(n/2)\ln(|\hat{\mathbf{V}}|)+\sum_{j=1}^{n}\ln\left\{\sum_{i_1=0}^{2}\sum_{i_2=0}^{2}p_{i_1i_2j}\exp\left[-(1/2)(\mathbf{y}_j-\hat{\mathbf{u}}_{i_1i_2j})\hat{\mathbf{V}}^{-1}(\mathbf{y}_j-\hat{\mathbf{u}}_{i_1i_2j})'\right]\right\}. \qquad (11.26)
$$

In theory, the search can be made in the two-dimensional space of possible locations of the two putative non-pleiotropic QTL in the testing region. The test statistic is then given by the likelihood ratio

$$
LR_2=-2\ln(L_{20}/L_2) \qquad (11.27)
$$

where L_2 is the maximum of the likelihoods in the two-dimensional space, and L_{20} is the maximum of the likelihoods on the diagonal of the two-dimensional space which corresponds to the null hypothesis of $p(1)=p(2)$. In practice, however, the search in the two-dimensional space is unnecessary. It is expected that the likelihood under the alternative hypothesis is maximized in the region near the peak indicated by the separate mappings, whereas the maximum likelihood under the null hypothesis corresponds to the joint mapping under the same model. So instead of searching in the two-dimensional space, the search can be safely concentrated in the areas suggested by the joint and separate mappings. As the hypotheses

in (11.16) are nested hypotheses, the test statistic under H_0 will be asymptotically chi-square distributed with 1 degree of freedom. The performance of this test will be investigated by simulations below.

11.2.4 QTL by environment interaction

Genes expressed in different environments can show different effects. This differential expression is usually called genotype by environment interaction. There are generally two types of experimental designs used in practice to study QTL × environment interaction (Paterson *et al.* 1991; Stuber *et al.* 1992). In one design, the same set of genotypes recorded on markers is evaluated phenotypically in different environments, which may be called paired comparison or Design I. In the other design, different random sets of genotypes (or individuals) from a common population are evaluated phenotypically in different environments, which may be called group comparison or Design II.

In Design I, since the same set of genotypes recorded on markers are evaluated phenotypically on multiple environments, the same $\mathbf{X}$ (marker data) matrix is paired to multiple phenotypic vectors in $\mathbf{Y}$, and the statistical model for analysis is just the same as model (11.1). Essentially, we regard different expressions of the same trait in different environments as different traits or different trait states, a concept originally introduced by Falconer (1952). However, by testing the QTL × environment interaction, we test the hypotheses

$$\begin{array}{l} H_0: \ b_1^* = b_2^* = b^*, \ \ d_1^* = d_2^* = d^* \\ H_1: \ b_1^* \neq b_2^*, \ \ d_1^* \neq d_2^*. \end{array} \tag{11.28}$$

This test, of course, is performed only at the chromosome regions where QTL have been suggested by the joint mapping (*i.e.*, the null hypothesis H_0 of (11.11) has been rejected). Under H_1 of (11.28), the model is the full model of (11.1).

Under H_0, the E-step will be similar to the full model except that b^* is substituted for b_1^* and b_2^* and d^* for d_1^* and d_2^*. In the CM-step,

$$b^{*(v+1)} = \mathbf{q}_2^{(v+1)\prime}(\mathbf{Y} - \mathbf{X}\mathbf{B}^{(v)})(\mathbf{V}^{(v)})^{-1}\mathbf{1} / \left[2c^{(v)}\mathbf{q}_2^{(v+1)\prime}\mathbf{1}\right] \tag{11.29}$$

$$d^{*(v+1)} = \left[\mathbf{q}_1^{(v+1)\prime}/(c^{(v)}\mathbf{q}_1^{(v+1)\prime}\mathbf{1}) - \mathbf{q}_2^{(v+1)}/(2c^{(v)}\mathbf{q}_2^{(v+1)\prime}\mathbf{1})\right](\mathbf{Y} - \mathbf{X}\mathbf{B}^{(v)})(\mathbf{V}^{(v)})^{-1}\mathbf{1} \tag{11.30}$$

with $c^{(v)} = \mathbf{1}'(\mathbf{V}^{(v)})^{-1}\mathbf{1} = (\sigma_1^{2(v)} - 2\rho^{(v)}\sigma_1^{(v)}\sigma_2^{(v)} + \sigma_2^{2(v)})/[\sigma_1^{2(v)}\sigma_2^{2(v)}(1-\rho^{2(v)})]$. Here, $b^{*(v+1)}$ and $d^{*(v+1)}$ are just the weighted averages of (11.6) and (11.7) respectively, weighted by the residual variance-covariance matrix $\mathbf{V}^{(v)}$. $\mathbf{B}^{(v+1)}$ and $\mathbf{V}^{(v+1)}$ are given by (11.8) and (11.9) with again b^* substituted for b_1^* and b_2^* and d^* for d_1^* and d_2^*. The log-likelihood is calculated from

$$\begin{aligned} \ln(L_3) \ &= \ K - (n/2)\ln(|\hat{\mathbf{V}}|) - (1/2)\sum_{j=1}^{n}(\mathbf{y}_j - \mathbf{x}_j\hat{\mathbf{B}})\hat{\mathbf{V}}^{-1}(\mathbf{y}_j - \mathbf{x}_j\hat{\mathbf{B}})' \\ &\quad + \sum_{j=1}^{n}\ln\left\{p_{2j}\exp\left[2\hat{b}^*\mathbf{1}'\hat{\mathbf{V}}^{-1}(\mathbf{y}_j - \mathbf{1}'\hat{b}^* - \mathbf{x}_j\hat{\mathbf{B}})'\right]\right. \end{aligned}$$

$$+p_{1j}\exp\left[(\hat{b}^*+\hat{d}^*)\mathbf{1}'\hat{\mathbf{V}}^{-1}(\mathbf{y}_j-\mathbf{1}'\hat{b}^*/2-\mathbf{1}'\hat{d}^*/2-\mathbf{x}_j\hat{\mathbf{B}})'\right]+p_{0j}\Big\}. \quad (11.31)$$

The test is performed by a likelihood ratio

$$LR_3 = -2\ln(L_3/L_1) \quad (11.32)$$

which is asymptotically chi-square distributed with 2 degrees of freedom under the null hypothesis.

In Design II, the statistical model for analysis can be specified as

$$\begin{array}{ll} y_{1j} = x^*_{1j}b^*_1 + z^*_{1j}d^*_1 + \mathbf{x}_{1j}\mathbf{b}_1 + e_{1j} & j=1,2,\cdots,n_1 \\ y_{2j} = x^*_{2j}b^*_2 + z^*_{2j}d^*_2 + \mathbf{x}_{2j}\mathbf{b}_2 + e_{2j} & j=1,2,\cdots,n_2. \end{array} \quad (11.33)$$

This can be expressed in matrix notation as

$$\begin{array}{l} \mathbf{y}_1 = \mathbf{x}^*_1 b^*_1 + \mathbf{z}^*_1 d^*_1 + \mathbf{X}_1\mathbf{b}_1 + \mathbf{e}_1 \\ \mathbf{y}_2 = \mathbf{x}^*_2 b^*_2 + \mathbf{z}^*_2 d^*_2 + \mathbf{X}_2\mathbf{b}_2 + \mathbf{e}_2. \end{array} \quad (11.34)$$

We will assume that e_{1j} and e_{2j} are independently normally distributed with means zero and variances σ^2_1 and σ^2_2, respectively. Under H_1 of (11.28), since e_{1j} and e_{2j} are independent, parameters in each environment can be estimated separately, and in each environment the analysis is on one trait. The log-likelihood is then just the sum of the log-likelihoods in each environment and can be expressed as

$$\ln(L_4) = \sum_{j=1}^{n_1}\ln\left[\sum_{i=0}^{2}p_{1ij}f_i(y_{1j})\right] + \sum_{j=1}^{n_2}\ln\left[\sum_{i=0}^{2}p_{2ij}f_i(y_{2j})\right] = \ln(L_1)_1 + \ln(L_1)_2 \quad (11.35)$$

where $\ln(L_1)_1$ and $\ln(L_1)_2$ are the log-likelihoods of (11.10) (with $m=1$) in the first and second environments respectively.

Under H_0 of (11.28), the parameters have to be estimated jointly. In each ECM iteration, the E-step constitutes

$$q^{(v+1)}_{kij} = p_{kij}f^{(v)}_i(y_{kj}) \Big/ \sum_{i=0}^{2}\left(p_{kij}f^{(v)}_i(y_{kj})\right) \quad (11.36)$$

for the kth environment, the ith genotype and the jth individual. The CM-step constitutes

$$\begin{aligned} b^{*(v+1)} &= \left[\mathbf{q}^{(v+1)\prime}_{12}(\mathbf{y}_1-\mathbf{X}_1\mathbf{b}^{(v)}_1)/\sigma^{2(v)}_1 + \mathbf{q}^{(v+1)\prime}_{22}(\mathbf{y}_2-\mathbf{X}_2\mathbf{b}^{(v)}_2)/\sigma^{2(v)}_2\right] / \\ &\quad \left[2(\mathbf{q}^{(v+1)\prime}_{12}\mathbf{1}/\sigma^{2(v)}_1 + \mathbf{q}^{(v+1)\prime}_{22}\mathbf{1}/\sigma^{2(v)}_2)\right] \quad (11.37) \\ d^{*(v+1)} &= \left[\mathbf{q}^{(v+1)\prime}_{11}(\mathbf{y}_1-\mathbf{X}_1\mathbf{b}^{(v)}_1)/\sigma^{2(v)}_1 + \mathbf{q}^{(v+1)\prime}_{21}(\mathbf{y}_2-\mathbf{X}_2\mathbf{b}^{(v)}_2)/\sigma^{2(v)}_2\right] / \\ &\quad \left[\mathbf{q}^{(v+1)\prime}_{11}\mathbf{1}/\sigma^{2(v)}_1 + \mathbf{q}^{(v+1)\prime}_{21}\mathbf{1}/\sigma^{2(v)}_2\right] - b^{*(v+1)} \quad (11.38) \\ \mathbf{b}^{(v+1)}_1 &= (\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1\left[\mathbf{y}_1 - (2\mathbf{q}^{(v+1)}_{12} + \mathbf{q}^{(v+1)}_{11})b^{*(v+1)} - \mathbf{q}^{(v+1)}_{11}d^{*(v+1)}\right] \quad (11.39) \end{aligned}$$

$$\mathbf{b}_2^{(v+1)} = (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\left[\mathbf{y}_2 - (2\mathbf{q}_{22}^{(v+1)} + \mathbf{q}_{21}^{(v+1)})b^{*(v+1)} - \mathbf{q}_{21}^{(v+1)}d^{*(v+1)}\right] \quad (11.40)$$

$$\sigma_1^{2(v+1)} = \Big[(\mathbf{y}_1 - \mathbf{X}_1\mathbf{b}_1^{(v+1)})'(\mathbf{y}_1 - \mathbf{X}_1\mathbf{b}_1^{(v+1)}) - 4\mathbf{q}_{12}^{(v+1)\prime}\mathbf{1}b^{*2(v+1)} - \mathbf{q}_{11}^{(v+1)\prime}\mathbf{1}(b^{*(v+1)} + d^{*(v+1)})^2\Big]/n_1 \quad (11.41)$$

$$\sigma_2^{2(v+1)} = \Big[(\mathbf{y}_2 - \mathbf{X}_2\mathbf{b}_2^{(v+1)})'(\mathbf{y}_2 - \mathbf{X}_2\mathbf{b}_2^{(v+1)}) - 4\mathbf{q}_{22}^{(v+1)\prime}\mathbf{1}b^{*2(v+1)} - \mathbf{q}_{21}^{(v+1)\prime}\mathbf{1}(b^{*(v+1)} + d^{*(v+1)})^2\Big]/n_2. \quad (11.42)$$

The log-likelihood, $\ln(L_5)$, under H_0 will be similar to (11.35) in form with the constraint that $\hat{b}_1^* = \hat{b}_1^* = \hat{b}^*$ and $\hat{d}_1^* = \hat{d}_2^* = \hat{d}^*$. The likelihood ratio test statistic is given by

$$LR_4 = -2\ln(L_5/L_4) \quad (11.43)$$

which is asymptotically chi-square distributed with two degrees of freedom under the null hypothesis.

It is interesting to compare the relative efficiency of the two experimental designs on mapping QTL and testing QTL × environment interaction. When $n_1 = n_2 = n$ and n is large, the test statistic under Design II can be treated approximately as a special case of Design I with $\rho = 0$. With this assumption, it can be shown that, with the same sample size for phenotyping in the two Designs (the sample size of marker genotyping is doubled in Design II), Design II is likely to have more statistical power for mapping QTL, whereas Design I is likely to have more statistical power for testing QTL × environment interaction.

In this analysis, we treat the QTL × environment effects as fixed effects. This may be appropriate for those environments which are distinctively different and the inference is applied to those environments. For many experiment designs, the QTL × environment effects may be modeled as random effects and the method of variance components may have to be used.

11.3 Examples from simulation studies

11.3.1 Joint mapping vs. separate mapping

To further explore some properties of joint mapping, simulation experiments were performed. For simplicity, one chromosome with 16 uniformly distributed markers was simulated for an F_2 population. Each marker interval is 10 cM in length with a total length of 150 cM. Three QTL were simulated to affect three traits with effects and positions listed in Table 11.2 (with no epistasis). Heritabilities of the three traits are all assumed to be 0.3. The sample size is 150. The trait values of an individual on three traits are determined by the sum of effects of QTL sampled, plus a random vector of environmental effects which are sampled from a trivariate normal distribution with means zero, and variances and covariances given in Table 11.3. Simulation and analyses were performed on 100 replicates.

Given the positions and effects of QTL, the genetic correlations between traits are expected to be 0.54, −0.22, 0.68 between traits 1 and 2, 1 and 3, 2 and 3, respectively.

(See, for example, APPENDIX C of Zeng (1992) for the calculation of the expected genetic variance in an F_2 population. The genetic covariance can be calculated similarly.) However, despite the substantial genetic correlations among traits, phenotypic correlations are low after adding environmental effects (Table 11.3). Sample means, variances and correlation coefficients averaged over 100 replicates are also listed in Table 11.3. It is interesting to observe that the observed (averaged) residual variances and correlations are very close to the expected environmental variances and correlations, as in the analysis most of the genetic variation is absorbed by markers fitted in the model.

Seven methods of QTL mapping were performed on each simulated data set at every 1 cM on the chromosome. These include: joint mapping for three traits (J-123); joint mapping for each pair of traits (J-12, J-13 and J-23); and separate mapping for each trait (S-1, S-2 and S-3). Simply for the convenience of discussion, in mapping except for the flanking markers, all other markers are fitted in the model to control the genetic background because markers are evenly distributed and widely separated. We used $\chi^2_{0.05/15,7} = 21.4$, $\chi^2_{0.05/15,5} = 17.7$, and $\chi^2_{0.05/15,3} = 13.6$ as the critical values of the test for the three levels of mapping.

Summary estimates of QTL positions and effects by the seven mapping methods are given in Table 11.2, and the observed power of detection of QTL are given in Table 11.4. The statistics given in Tables 11.2 and 11.4 are summarized from 100 replicates for three QTL regions separately. Although three QTL are assumed to be located on the same chromosome, they are widely separated. Also because of the composite interval mapping used in analysis, tests in different regions are statistically independent (Zeng 1993), so that the statistical power and sampling variance of estimates can be calculated separately for three QTL at and around the intervals surrounding the QTL. In Table 11.4, S-123 denotes the overall performance of the three separate mappings, and its power was calculated as the frequency of the detection of the QTL by at least one of the three separate mappings. It is seen that the power of J-123 in this case is higher than that of S-123 for all three QTL. This shows that some QTL with relatively small effects may be missed by separate mappings on different traits, but detected by joint mapping which combined information from different traits. The power of J-12 is very close to that of J-123 on QTL *1*. This is because QTL *1* has effects mainly on traits 1 and 2, and just a small effect on trait 3. The exclusion of trait 3 in J-12 only slightly affects the power of detection of the QTL. This also shows that small pleiotropic effects on additional traits included in the joint mapping may be large enough to compensate the lose of power due to the increase of the critical value. The powers of QTL detection by J-12, J-13 and J-23 are generally comparable to the sizes of pleiotropic QTL effects involved. J-23, however, tends to have relatively higher power. This is because the pleiotropic effects of three QTL on traits 2 and 3 are all in the same direction and the environment correlation is negative (see APPENDIX A).

Means and standard deviations of estimates (over all replicates) of QTL positions and effects by different methods are also given in Table 11.2. All estimates seem to be relatively unbiased. In general, the precision of estimation of QTL positions and effects by J-123, as indicated by standard deviations, is better than other methods. Particularly, the joint analysis has a significant effect on improving the sampling variance of estimates of QTL

positions. Sampling variances of estimates of QTL positions by J-123 are consistently smaller than those by other analyses, except of that for QTL 3 by J-13 and for QTL 2 by J-23 in which cases two trait analyses have some particular advantages as noticed above.

11.3.2 Pleiotropy vs. close linkage

A simulation study was also performed to test close linkage of two non-pleiotropic QTL against pleiotropy of a common QTL. In this example, one chromosome with 11 markers in 10 marker intervals each with 15 cM was simulated. Three QTL, one pleiotropic and two non-pleiotropic, were assumed to affect two traits, with parameters given in Table 11.5. The heritability for each trait is 0.4. (The effects of QTL are undoubtedly very large.) The sample genetic, environmental and phenotypic correlations are 0.42, 0.2 and 0.29 respectively. The sample size is 300.

The results of joint mapping (J-12) and separate mappings (S-1 and S-2) in one replicate are presented in Table 11.5 and Figure 11.1. At least two major QTL are indicated by the analyses. There is some evidence from separate mappings that there might be two non-pleiotropic QTL in the region between 105 and 135 cM with each showing a significant effect on one trait only. To test this hypothesis, the test of pleiotropy vs. close linkage was performed in the two major regions which show significant effects on the traits for comparison: one between 45 and 75 cM and one between 105 and 135 cM. The results are plotted in Figure 11.2.

In Figure 11.2, the two-dimensional log-likelihood surface (as deviations from the maximum of the log-likelihoods on the diagonal) is presented. In this analysis, unlike joint mapping and the separate mappings which used all but flanking markers for background control, only markers 1-3 and 7-11 are used in the model for background control in Figure 11.2A, and only markers 1-7 and 11 are used for background control in Figure 11.2B. On this two-dimensional surface, the diagonal elements represents null hypotheses of one pleiotropic QTL, and the off-diagonal elements represent alternative hypotheses of two non-pleiotropic QTL. The likelihood ratio test is performed at the maximum of the surface against the maximum of the diagonals. This likelihood ratio is 0.53 for Figure 11.2A at the position 56 cM for trait 1 and 57 cM for trait 2 (the maximum diagonal is at 57 cM), which is clearly not significant, and 7.26 for Figure 11.2B at the position 111 cM for trait 1 and 126 cM for trait 2 (the maximum diagonal is at 125 cM), which is significant at 0.01 level (for chi-square distribution with one degree of freedom). There is clear evidence to support the presence of two QTL with one located around 111 cM showing a significant effect on trait 1 and one located around 126 cM showing a significant effect on trait 2.

The maximum positions of diagonals in Figure 11.2 are the same maximum positions under the joint mapping in Figure 11.1 and Table 11.5. The maximum positions of off-diagonals in Figure 11.2 are also very close to the maximum positions indicated by the separate mappings (Table 11.5). As discussed above, in practice this two-dimensional search for the best fit of two non-pleiotropic QTL is unnecessary, although the two-dimensional surface is more illuminating. The test can be constructed at the peak suggested by joint mapping for the null hypothesis and at or around the peak suggested by separate mappings

Table 11.2: Parameters of QTL positions and effects used in simulations and their mean estimates (standard deviations), over 100 replicates, by the joint mapping on three traits (J-123) and on two traits at a time (J-12, J-13 and J-23) and by the separate mapping on each trait (S-1, S-2 and S-3)

QTL	Position	Additive effect			Dominance effect		
	(cM)	Trait 1	Trait 2	Trait 3	Trait 1	Trait 2	Trait 3
				Parameters			
1	21.0	1.00	1.00	0.30	0.43	0.43	0.13
2	84.0	-0.30	-1.00	-1.00	-0.09	-0.30	-0.30
3	142.0	-1.00	0.30	1.00	0.19	0.06	0.19
				Estimates by J-123			
1	21.0 (4.3)	1.00 (0.44)	1.00 (0.42)	0.35 (0.36)	0.43 (0.44)	0.42 (0.28)	0.13 (0.25)
2	84.9 (4.7)	-0.24 (0.51)	-1.00 (0.38)	-1.01 (0.40)	-0.01 (0.51)	-0.27 (0.25)	-0.25 (0.23)
3	142.4 (3.9)	-1.03 (0.38)	0.27 (0.27)	1.02 (0.31)	0.17 (0.38)	0.05 (0.29)	0.07 (0.22)
				Estimates by J-12			
1	20.9 (5.2)	1.03 (0.43)	1.02 (0.41)		0.44 (0.32)	0.42 (0.29)	
2	83.7 (9.6)	-0.25 (0.56)	-0.97 (0.43)		0.03 (0.36)	-0.24 (0.30)	
3	141.3 (6.8)	-1.08 (0.36)	0.26 (0.34)		0.17 (0.30)	0.02 (0.30)	
				Estimates by J-13			
1	22.4 (7.8)	1.05 (0.48)		0.30 (0.39)	0.47 (0.32)		0.14 (0.25)
2	85.6 (6.0)	-0.26 (0.52)		-1.03 (0.40)	-0.01 (0.33)		-0.25 (0.24)
3	143.1 (2.8)	-1.00 (0.38)		1.03 (0.30)	0.18 (0.29)		0.08 (0.23)
				Estimates by J-23			
1	21.8 (6.5)		1.05 (0.41)	0.34 (0.40)		0.41 (0.30)	0.12 (0.26)
2	84.9 (4.4)		-1.01 (0.38)	-1.04 (0.37)		-0.27 (0.25)	-0.25 (0.23)
3	142.4 (4.5)		0.28 (0.30)	1.04 (0.35)		0.03 (0.29)	0.07 (0.23)
				Estimates by S-1, S-2 and S-3			
1	21.9,21.6,25.8 (7.7,5.7,13.6)	1.08 (0.42)	1.12 (0.34)	0.38 (0.49)	0.48 (0.33)	0.41 (0.30)	0.11 (0.33)
2	84.9,84.3,86.0 (17.1,8.7,6.7)	-0.29 (0.69)	-1.03 (0.38)	-1.06 (0.37)	0.06 (0.42)	-0.26 (0.29)	-0.24 (0.26)
3	141,136,143 (6.2,11.7,4.1)	-1.13 (0.31)	0.33 (0.45)	1.06 (0.30)	0.17 (0.30)	-0.01 (0.37)	0.08 (0.24)

Table 11.3: Summary parameters and mean statistics of traits for the simulation example in Table 11.2

Trait	Genetic variance	Phenotypic variance	Environmental variance	Genetic correlation		Phenotypic correlation		Environ
				Trait 2	Trait 3	Trait 2	Trait 3	Trait 2
				Parameters				
1	1.02	3.40	2.38	0.54	-0.22	0.30	-0.07	0.20
2	0.76	2.53	1.77		0.68		0.06	
3	0.71	2.35	1.64					
		Sample variance	Residual variance			Sample correlation		Resid
						Trait 2	Trait 3	Trait 2
				Mean estimates				
1		3.47	2.25			0.30	-0.07	0.21
2		2.51	1.67				0.07	
3		2.28	1.51					

Table 11.4: Observed statistical power (proportion of significant replicates over all replicates) of seven methods of QTL mapping from 100 replicates of simulations

QTL	J-123	J-12	J-13	J-23	S-1	S-2	S-3	S-123
1	0.80	0.78	0.51	0.64	0.46	0.64	0.04	0.78
2	0.79	0.37	0.36	0.84	0.00	0.39	0.41	0.64
3	0.89	0.51	0.84	0.64	0.42	0.00	0.64	0.79

Table 11.3: Summary parameters and mean statistics of traits for the simulation example in Table 9.2

	Genetic	Phenotypic	Environmental	Genetic correlation		Phenotypic correlation		Environmental correlation	
Trait	variance	variance	variance	Trait 2	Trait 3	Trait 2	Trait 3	Trait 2	Trait 3
				Parameters					
1	1.02	3.40	2.38	0.54	-0.22	0.30	-0.07	0.20	0.00
2	0.76	2.53	1.77		0.68		0.06		-0.20
3	0.71	2.35	1.64						
		Sample	Residual			Sample correlation		Residual correlation	
		variance	variance			Trait 2	Trait 3	Trait 2	Trait 3
				Mean estimates					
1		3.47	2.25			0.30	-0.07	0.21	-0.01
2		2.51	1.67				0.07		-0.14
3		2.28	1.51						

Table 11.5: Parameters of QTL positions and effects and their estimates in a single replicate of simulation by the joint mapping on two traits (J-12) and the separate mapping on each trait (S-1 and S-2)

	Position	Additive effect		Dominance effect		Likelihood
QTL	(cM)	Trait 1	Trait 2	Trait 1	Trait 2	ratio
			Parameters			
1	54	-1.36	-1.44	1.28	1.35	
2	114	-1.16		0.75		
3	128		1.30		0.49	
			Estimates by J-12			
1	57	-1.03	-1.34	1.36	1.42	66.73
2/3	125	-0.82	1.31	-0.35	0.52	37.49
			Estimates by S-1			
1	55	-1.05		1.42		28.07
2	110	-0.75		1.16		15.01
			Estimates by S-2			
1	61		-1.22		1.24	51.95
3	127		1.34		0.45	24.24

for the alternative hypothesis. This is clearly supported by the results of Figure 11.2.

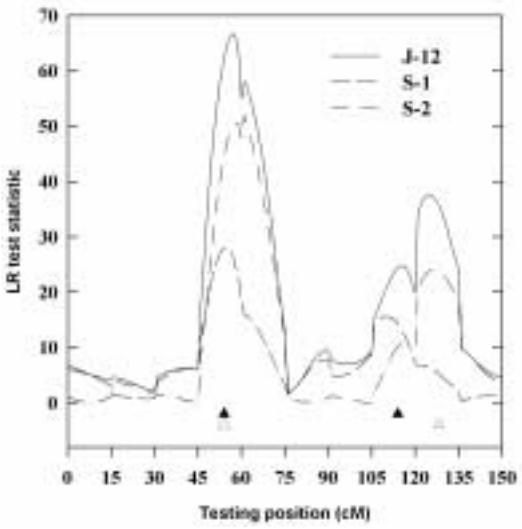

Figure 11.1: A simulation example of QTL mapping on two traits from an F_2 population. Likelihood ratio test statistics are calculated and plotted at every 1 cM position of a "chromosome" for three mapping methods. J-12 is the joint mapping on two traits. S-1 is the separate mapping on trait 1 and S-2 is the separate mapping on trait 2. The genetic length of the "chromosome" is 150 cM with markers at every 15 cM. Three QTL, one pleiotropic and two non-pleiotropic, were simulated with positions and effects given in Table 6 and indicated by the triangles. The filled triangles indicate for QTL effects on trait 1 and the open triangles indicate for QTL effects on trait 2.

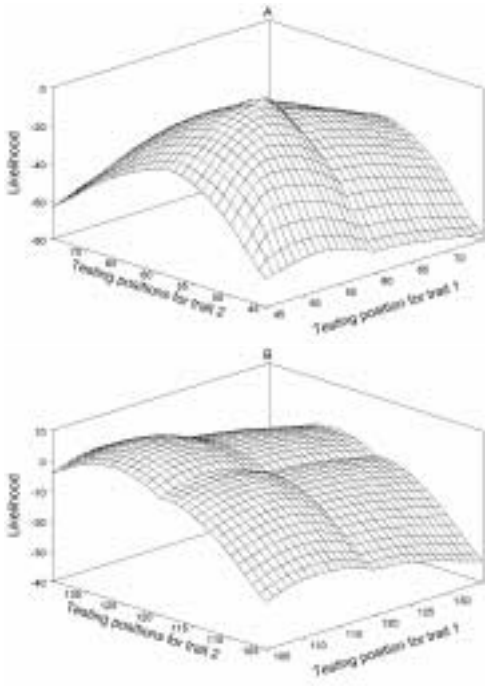

Figure 11.2: Two dimensional log-likelihood surfaces (expressed as deviations from the maximum of the log-likelihoods on the diagonal) for the test of pleiotropy vs. close linkage are presented for two regions. Figure 2A is for the region between 45 and 75 cM of Figure 1 and Figure 2B is for the region between 105 and 135 cM. X is the testing position for a QTL affecting trait 1 and Y is the testing position for a QTL affecting trait 2. On the diagonal of X-Y plane, two QTL are located in the same position, and statistically are treated as one pleiotropic QTL. Z is the likelihood ratio test statistic scaled to zero at the maximum point of the diagonal.

Chapter 12

References

Akaike, H., 1969 Fitting autoregressive models for prediction. Ann. Inst. Statist. Math. **21:** 243-247.

Basten, C. J., B. S. Weir, and Z.-B. Zeng, 1995-1998 *QTL Cartographer: A Reference Manual and Tutorial for QTL Mapping*. Department of Statistics, North Carolina State University, Raleigh NC.

Beavis, W. D., O. S. Smith, D. Grant and R. Fincher, 1994 Identification of quantitative trait loci using a small sample of topcrossed and F_4 progeny from maize. Crop Sci. **34**: 882-896.

Breiman, L., and D. Freedman, 1983 How many variables should be entered in a regression? J. Amer. Statist. Ass. **78:** 131-

Broman, K. W., 1997 Identifying quantitative trait loci in experimental crosses. Ph.D. thesis, Department of Statistics, University of California, Berkeley.

Bulmer, M. G., 1985 *The Mathematical Theory of Quantitative Genetics*. Oxford University Press, Oxford.

Carbonell, E. A. and T. M. Gerig, 1991 A program to detect linkage between genetic markers and non-additive quantitative trait loci. J. Heredity **82:** 435.

Carbonell, E. A., T. M. Gerig, E. Balansard, and M. J. Asin, 1992 Interval mapping in the analysis of non-additive quantitative trait loci. Biometrics **48:** 305-315.

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

Cockerham, C. C., and Z.-B. Zeng, 1996 Design III with marker loci. Genetics .

Comstock, R. E., and H. F. Robinson, 1952 Estimation of average dominance of genes. pp. 495-516 in *Heterosis*, edited by J. W. Gowen, Iowa Sate College Press, Ames, Iowa.

Darvasi, A. and M. Soller, 1995 Advanced intercross lines, an experimental population to interval mapping of quantitative trait loci. Genetics **141**: 1199-1207.

Doerge, R. W., 1993 *Statistical Methods for Locating Quantitative Trait loci with Molecular Markers*. Ph.D. Thesis, Department of Statistics, North Carolina State University, Raleigh, NC.

Doerge, R. W., and G. A. Churchill, 1996 Permutation tests for multiple loci affecting a quantitative characters. Genetics **142:** 285-294.

Doerge, R. W., Z.-B. Zeng, and B. S. Weir, 1997 Statistical issues in the search for genes affecting quantitative traits in experimental populations. Statistical Science **12**: 195-219.

Dragani, T. A., Z.-B. Zeng, F. Canzian, M. Gariboldi, G. Manenti, and M A. Pierotti, 1995 Molecular mapping of body weight loci on mouse chromosome X. Mammalian Genome **6**: 778-781.

Falconer, D. S., 1952 The problem of environment and selection. Amer. Natur. **86**: 293-298.

Falconer, D. S., and T. F. C. Mackay 1996 *Introduction to Quantitative Genetics*. Ed. 4. Longman, New York.

Feingold, E., P. O. Brown, and D. Siegmund, 1993 Gaussian models for genetic linkage analysis using complete high-resolution maps of identity by descent. Am. J. Hum. Genet. **53**: 234-251.

Fisch, R. D., M. Ragot and G. Gay, 1996 A generalization of the mixture model in the mapping of quantitative trait loci for progeny from a bi-parental cross of inbred lines. Genetics **143**: 571-577.

Fisher, R. A. 1918 The correlation between relatives on the supposition of Mendelian inheritance. Proc. Roy. Soc. Edin. 52 (Part II): 399-433.

Green, P. J., 1995 Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82: 711-732.

Haldane, J. B. S., 1919 The combination of linkage values and the calculation of distance between the loci of linked factors. J. Genet. **8**: 299-309.

Haldane, J. B. S., and C. H. Waddington, 1931 Inbreeding and linkage. Genetics **16**: 357-374.

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

Haley, C. S., S. A. Knott, and J.-M. Elsen, 1994 Mapping quantitative trait loci in crosses between outbred lines using least squares. Genetics **136**: 1195-1207.

Hannan, E. J., and B. G. Quinn, 1979 The determination of the order of an autoregression. J. Roy. Statist. Soc. B, **41:** 190-195.

Health, S. C., 1997 Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. Am. J. Hum. Genet. 61: 748-760.

Hyne, V., and M. J. Kearsey, 1995 QTL analysis: further use of 'marker regression'. Theor. Appl. Genet. **91:** 471-476.

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

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

Jansen, R. C., 1994 Controlling the type I and type II errors in mapping quantitative trait loci. Genetics **138**: 871-881.

Jansen, R. C., 1996 A general Monte Carlo method for mapping multiple quantitative trait loci. Genetics **142:** 305-311.

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

Jiang, C., and Z.-B. Zeng, 1995 Multiple trait analysis of genetic mapping for quantitative trait loci. Genetics **140**: 1111-1127.

Jiang, C., and Z.-B. Zeng, 1997 Mapping quantitative trait loci with dominant and missing markers in various crosses from two inbred lines. Genetica **101**: 47-58.

Kao, C.-H., and Z.-B. Zeng, 1997 General formulae for obtaining the MLEs and the asymptotic variance-covariance matrix in mapping quantitative trait loci when using the EM algorithm. Biometrics **53:** 653-665.

Kao, C.-H., Z.-B. Zeng, and R. Teasdale, 1998 Multiple interval mapping for quantitative trait loci. Genetics (submitted).

Kearsey, M. J., and V. Hyne, 1994 QTL analysis: a simple 'marker regression' approach. Theor. Appl. Genet. **90:** 698-702.

Knapp, S. J., 1991 Using molecular markers to map multiple quantitative trait loci: model for backcross, recombinant inbred, and double-haploid progeny. Theor. Appl. Genet. **81:** 333-338.

Knott, S. A., and C. S. Haley, 1992 Aspects of maximum likelihood methods for the mapping of quantitative trait loci in line crosses. Genet. Res. **60**:139-151.

Lander, E., P. Green, J. Abrahamson, A. Barlow, M. Daley, S. Lincoln, and L. Newburg, 1987 MAPMAKER: An interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics 1: 174-181.

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

Lander, E. S., and N. J. Schork, 1994 Genetic dissection of complex traits. Science **265**: 2037-2049.

Liu, J., J. M. Mercer, L. F. Stam, G. C. Gibson, Z.-B. Zeng, and C. C. Laurie, 1996 Genetic analysis of a morphological shape difference in the male genitalia of *Drosophila simulans* and *D. mauritiana*. Genetics 142: 1129-1145.

Luo, Z. W., and M. J. Kearsey, 1992 Interval mapping of quantitative trait loci in an F_2 population. Heredity **66:** 117-124.

Luo, Z. W., and J. A. Williams, 1993 Estimation of genetic parameters using linkage between a marker gene and a locus underlying a quantitative character in F_2 populations. Heredity **70:** 245-253.

Lynch, M., and B. Walsh, 1998 *Genetics and Analysis of Quantitative Traits* Sinauer Associates, Inc. Sunderland, Massachusetts.

McLachlan, G. J., and K. E. Basford, 1988 *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, New York.

McMillan, I., and A. Robertson, 1974 The power of methods for the detection of major gene affecting quantitative characters. Heredity **32**: 349-356.

Mangin, B., B. Goffinet, and A. Rebai, 1994 Constructing confidence interval for QTL location. Genetics **138**: 1301-1308.

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

Martinez, O. and R. N. Curnow, 1994 Missing markers when estimating quantitative trait loci using regression mapping. Heredity **73**: 198-206.

Mather, K. and J. L. Jinks, 1982 *Biometrical Genetics*) 3rd Edn. Chapman and Hall.

Meng, X.-L., and D. B. Rubin, 1993 Maximum likelihood estimation via the ECM algorithm: A general framework. Biometrika **80**: 267-268.

Miller, A. J., 1990 *Subset Selection in Regression.* Chapman and Hall, London.

Ott, J., 1991 *Analysis of Human Genetic Linkage*. Revised edition, John Hopkins University Press, Baltimore.

Paterson, A. H., E. S. Lander, J. D. Hewitt, S. Peterson, S. E. Lincoln, and S. D. Tanksley, 1988 Resolution of quantitative traits into Mendelian factors by using a complete RFLP linkage map. Nature **335**: 721-726.

Paterson, A. H., S. Damon, J. D. Hewitt, D. Zamir, H. D. Rabinowitch, S. E. Lincoln, E. S. Lander, and S. D. Tanksley, 1991 Mendelian factors underlying quantitative traits in tomato: comparison across species, generations and environments. Genetics **127**: 181-197.

Penrose, L. S., 1938 Genetic linkage in graded human character. Ann. Eugen. 8: 233-237.

Rebai, A., B. Goffinet, and B. Mangin, 1994 Approximate thresholds of interval mapping tests for QTL detection. Genetics **138**: 235-240.

Rebai, A., B. Goffinet, and B. Mangin, 1995 Comparing power of different methods of QTL detection. Biometrics **51:** 87-99.

Rodolphe, F., and M. Lefort, 1993 A multi-marker model for detecting chromosomal segments displaying QTL activity. Genetics **134**: 1277-1288.

Satagopan, J. M., B. S. Yandell, M. A. Newton, and T. C. Osborn, 1996 A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. Genetics **144:** 805-816.

Sax, K., 1923 The association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris*. Genetics **8**: 552-560.

Schwarz, G., 1978 Estimating the dimension of a model. Ann. Statist. **6:** 461-464.

Shibata, R., 1981 An optimal selection of regression variables. Biometrika **68:** 45-

Shibata, R., 1984 Approximate efficiency of a selection procedure for the number of regression variables. Biometrika **71:** 43-

Soller, M., T. Brody, and A. Genizi, 1976 On the power of experimental design for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines. Theor. Appl. Genet. **47**: 35-39.

Speed, T. P., M. S. McPeek and S. N. Evans, 1992 Robustness of the no-interference model for ordering genetic markers. Proc. Natl. Acad. Sci. USA **89**: 3103-3106.

Stam, P., 1991 Some aspects of QTL analysis. *Proceedings of the Eighth Meeting of the*

Eucarpia Section Biometrics on Plant Breeding, Brno, Czechoslovakia. Pp. 24-32.

Stuart, A., and J. K. Ord, 1991 *Kendall's Advanced Theory of Statistics* Vol. 2. Fifth Ed. Oxford University Press, New York.

Stuber, C. W., S. E. Lincoln, D. W. Wolff, T. Helentjaris, and E. S. Lander, 1992 Identification of genetic factors contributing to heterosis in a hybrid from two elite maize inbred lines using molecular markers. Genetics **132**: 823-839.

Tanksley, S. D., 1993 Mapping polygenes. Annu. Rev. Genet. **27**: 205-233.

Thoday, J. M., 1961 Location of polygenes. Nature **191**: 368-370.

Thoday, J. M., 1977 Effects of specific genes. in *Proc. Int. Conf. Quantitative Genetics* ed. E. Pollak, O. Kempthorne, and T. B. Bailey, pp.141-159, Iowa State Univ., Ames, Iowa.

Titterington, D. M., A. F. M. Smith, and U. E. Markov, 1985 *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York.

Uimari, P., I. Hoeschele, 1997 Mapping two linked quantitative trait loci with Bayesian analysis and Markov chain Monte Carlo algorithms. Genetics (in press).

van Ooijen, J. W., 1992 Accuracy of mapping quantitative trait loci in autogamous species. Theor. Appl. Genet. **84**: 803-831.

Visscher, P. M., R. Thompson and C. S. Haley, 1996 Confidence intervals in QTL mapping by bootstrapping. Genetics **143**: 1013-1020.

Weller, J. I., 1986 Maximum likelihood techniques for the mapping and analysis of quantitative trait loci with the aid of genetic markers. Biometrics **42**: 627-640.

Weller, J. I., Y. Kashi, and M. Soller, 1990 Power of daughter and granddaughter designs for determining linkage between marker loci and quantitative trait loci in dairy cattle. J. Dairy Sci. **73**: 2525-2537.

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

Wright, A. J., and R. P. Mowers, 1994 Multiple regression for molecular- marker, quantitative trait data from large F_2 populations. Theor. Appl. Genet. **89:** 305-312.

Wu, W. R., and W. M. Li, 1994 A new approach for mapping quantitative trait loci using complete genetic marker linkage maps. Theor. Appl. Genet. **89:** 535-539.

Wu, W. R., and W. M. Li, 1996 Model fitting and model testing in the method of joint mapping of quantitative trait loci. Theor. Appl. Genet. **92:** 477-482.

Xiao, J., J. Li, L. Yuan and S. D. Tanksley, 1995 Dominance is the major genetic basis of heterosis in rice as revealed by QTL analysis using molecular markers. Genetics **140**: 745-754.

Xu, S., 1997 A comment on the simple regression method for interval mapping. Genetics **141:** 1657-1659.

Xu, S., and W. R. Atchley, 1995 A random model approach to interval mapping of quantitative genes. Genetics **141**: 1189-1197.

Zeng, Z.-B., 1993 Theoretical basis of separation of multiple linked gene effects on mapping quantitative trait loci. Proc. Natl. Acad. Sci. USA **90**: 10972-10976.

Zeng, Z.-B., 1994 Precision mapping of quantitative trait loci. Genetics **136**: 1457-1468.

Zeng, Z.-B., J. Liu, L. F. Stam, C.-H. Kao, J. M. Mercer, and C. C. Laurie, 1998 Genetic architecture of a morphological shape difference between two *Drosophila* species. Proc. Natl. Acad. Sci. USA (submitted).