Plant Molecular Breeding

Biological Sciences Series

A series which provides an accessible source of information at research and professional level in chosen sectors of the biological sciences.

Series Editors:

Titles in the series:

Stress Physiology in Animals
Edited by P. H. M. Balm

Seed Technology and its Biological Basis
Edited by M. Black and J. D. Bewley

Leaf Development and Canopy Growth
Edited by B. Marshall and J. A. Roberts

Environmental Impacts of Aquaculture
Edited by K. D. Black

Herbicides and their Mechanisms of Action
Edited by A. H. Cobb and R. C. Kirkwood

The Plant Cell Cycle and its Interfaces
Edited by D. Francis

Meristematic Tissues in Plant Growth and Development
Edited by M. T. McManus and B. E. Veit

Fruit Quality and its Biological Basis
Edited by M. Knee

Pectins and their Manipulation
Edited by G. B. Seymour and J. P Knox

Wood Quality and its Biological Basis
Edited by J. R. Barnett and G. Jeronimidis

Plant Molecular Breeding
Edited by H. J. Newbury

Biogeochemistry of Marine Systems
Edited by K. D. Black and G. Shimmield

Plant Molecular Breeding

Edited by

H. JOHN NEWBURY
School of Biosciences
University of Birmingham
Edgbaston, Birmingham
UK

Blackwell
Publishing

CRC Press

Editorial offices:
Blackwell Publishing Ltd, 9600 Garsington Road, Oxford OX4 2DQ, UK
Tel: +44 (0)1865 776868
Blackwell Publishing Asia Pty Ltd, 550 Swanston Street, Carlton, Victoria 3053, Australia
Tel: +61 (0)3 8359 1011

ISBN 1-84127-321-X
Originated as Sheffield Academic Press

Published in the USA and Canada (only) by CRC Press LLC, 2000 Corporate Blvd., N.W., Boca Raton, FL 33431, USA
Orders from the USA and Canada (only) to CRC Press LLC

USA and Canada only:
ISBN 0-8493-2813-X

This book contains information obtained from authentic and highly regarded sources. Reprinted material is quoted with permission, and sources are indicated. Reasonable efforts have been made to publish reliable data and information, but the author and the publisher cannot assume responsibility for the validity of all materials or for the consequences of their use.

Trademark notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation, without intent to infringe.

First published 2003

Library of Congress Cataloging-in-Publication Data:
A catalog record for this title is available from the Library of Congress

British Library Cataloguing-in-Publication data:
A catalogue record for this title is available from the British Library

Set in 10/12 pt Times
by Sparks Computer Solutions Ltd
http://www.sparks.co.uk
Printed and bound in Great Britain using acid-free paper by MPG Books Ltd, Bodmin, Cornwall

For further information on Blackwell Publishing, visit our website:
www.blackwellpublishing.com

Contents

Contributors

James A. Anderson
Department of Agronomy and Plant Genetics, University of Minnesota, St Paul, Minnesota, USA

M. W. Bonierbale
Centro Internacional de la Papa (CIP), Apartado 1558, Lima 12, Peru

Keith J. Edwards
School of Biological Sciences, University of Bristol, Woodland Road, Bristol BS8 1UG, UK

M. Ghislain
Centro Internacional de la Papa (CIP), Apartado 1558, Lima 12, Peru

Ian Godwin
School of Land and Food Sciences, University of Queensland, Brisbane QLD 4072, Australia

Frédéric Hospital
INRA, Station de Génétique Végétale, Ferme du Moulon, 91190 Gif sur Yvette, France

Michael J. Kearsey
School of Biosciences, University of Birmingham, Birmingham B15 2TT, UK

X-Q. Li
Potato Research Centre, Agriculture and Agri-Food Canada, PO Box 20280, Fredericton, New Brunswick EB3 4Z7, Canada

Zewei W. Luo
School of Biosciences, University of Birmingham, Birmingham B15 2TT, UK

C. Mba

Centro Internacional de Agricultura
Tropical (CIAT), A. A. 6713, Cali,
Colombia

H. John Newbury

School of Biosciences, University of
Birmingham, Edgbaston, Birmingham,
UK

Andy H. Paterson

College of Agricultural and
Environmental Sciences, University of
Georgia, Athens, GA 30602, USA

Ian Puddephat

Horticulture Research International,
Wellesbourne, Warwickshire, CV35
9EF, UK
Now at: Syngenta, Jealott's Hill
International Research Centre,
Bracknell, Berkshire RG42 6EY, UK

R. Simon

Centro Internacional de la Papa (CIP),
Apartado 1558, Lima 12, Peru

Donal M. O'Sullivan

Project Leader, Molecular Research
Group, NIAB, Huntingdon Road,
Cambridge CB3 0LE, UK

D. P. Zhang

Centro Internacional de la Papa (CIP),
Apartado 1558, Lima 12, Peru

Preface

The past few years have seen an explosion of new information and resources in the area of plant molecular genetics and genomics. As a result of developments in high-throughput sequencing and user-friendly databases with easy access via the internet, we now have available huge amounts of information on plant genes and physical and genetic maps. The milestone whole genome sequencing of both *Arabidopsis* and rice provides the most obvious evidence of progress in this area. For these model species, we have entered the era of functional genomics, which aims to determine the functions of all the genes identified in their genomes. A community-wide effort is underpinning studies in this area, and the insights provided by these analyses make this an exciting time to be a plant biologist. But how does the growing mountain of information on gene structure, organization and function help people charged with the task of improving crop species? This is one of the central themes of this book, in which researchers from leading laboratories around the world provide insights into their specialized areas to provide an overview of the state of molecular plant breeding at the beginning of the twenty-first century.

It is usual to classify attempts to improve plants into one of two strategies: by plant breeding, or by plant genetic engineering. In fact, improvement programmes frequently use a combination of these approaches. In any case, a thorough understanding of both classical and molecular genetics is required for either technique to be successful and efficient. Most of the traits that are the subject of improvement programmes are quantitative in nature. The chapter on the mapping, characterization and deployment of quantitative genetic trait loci (QTL) introduces the reader to the convergence between the statistical approaches of quantitative genetics and genomic sequence data that together allow the proposal of candidate genes for QTL controlling important crop traits. The chapter on genomic colinearity describes how an understanding of the conserved organization of genes between plant taxa facilitates the exploitation of genetic information derived from model species in less-heavily studied crops. Exploitation of wild relatives of crop species has also been reviewed, in order to explain the value of plant genetic resources as a supply of novel alleles for crop improvement. All three chapters demonstrate the importance of an array of molecular marker technologies for the location and characterization of plant genes.

In many cases, genes identified using the approaches just described can be deployed in breeding programmes, and the chapter describing marker-assisted breeding analyses the efficiency of different marker-based methods for following the inheritance of target genes in progeny plants. In other cases, isolated genes can be transferred using genetically modified (GM) technology, and an extensive review of recent advances in plant genetic engineering has been included. Finally, three crop types have been selected for a detailed analysis of the effects of developments in molecular genetics on improvement programmes. The choice of wheat and maize as examples is based upon their relative global importance as crop plants. The chapter on root and tuber crop improvement has a more commodity-based perspective and offers the opportunity to compare the advantages and disadvantages of working with several starch crops.

H.J. Newbury

1 Mapping, characterization and deployment of quantitative trait loci

Michael J. Kearsey and Zewei W. Luo

1.1 Introduction

Most of the very considerable progress in genetics over the last century has focused on using single gene mutants that produce relatively clear-cut effects on the phenotype. They have involved comparisons between normal and disfunctional alleles, so effects are very large compared with the background variation at other gene loci and the environment. As a result, their inheritance can easily be seen to follow Mendelian laws. However, most natural variation of importance to plant breeders (yield, emergence time, stress tolerance, etc.) is not of this sort. The phenotypes in an F_2 do not fall into clear-cut Mendelian ratios, but most commonly show a continuous, approximately normal, range of variation. The analysis and deployment of quantitative trait loci is therefore of enormous importance in breeding programmes. Although it is possible to undertake breeding programmes using only phenotypic selection, an understanding of the number and location of quantitative trait loci (QTL) controlling performance for a target trait can markedly enhance the efficiency of breeding. In this chapter, we will be reviewing the theoretical background to quantitative trait analysis and explaining how the concepts and statistical approaches that arise from these theoretical considerations can be employed to analyse the genetic basis of quantitative traits in plants. Most of the progress in QTL analysis over the past decade has occurred because of the availability of a range of informative molecular marker techniques. We will explain the methods by which associations between the inheritance of alleles at marker and trait loci allow the identification and mapping of QTL.

Information about QTL location, effects and even the sequence can be exploited in a number of ways in crop improvement programmes, and these are explained in later chapters in this book. In Chapter 2, the reader is shown how map location information can be used in a range of marker-assisted selection protocols. In Chapter 3, there is a review of the value of synteny between plant genomes. It is clear from the information presented that knowledge of the map location of an important QTL in one species allows the accurate prediction of the map location of that QTL in related species. This means, for example, that knowledge of the location of an important flowering time QTL in rice allows one to target specific chromosome intervals in less well-studied cereals in breeding programmes concerned with earliness. Recent advances in plant molecular genetics have eventually allowed the cloning of QTL largely because of the information produced by the sequencing of the entire genome of the model species *Arabidopsis.* Hence, QTL – which were once largely theoretical genetic determinants that helped explain trait performance in statistical analyses – are now being sequenced and the allelic variations that account for their different effects are being elucidated. The isolation of QTL as cloned DNA fragments opens the possibility of the transfer of QTL alleles by genetic engineering techniques and GM technologies are reviewed in Chapter 4.

1.2 Genetic basis of quantitative trait performance

It has been generally assumed, ever since the pioneering work of Nilsson-Ehle (1909) almost 100 years ago, that continuous variation in trait performance is due to the joint segregation of several genes, all of which have a small but quasi-additive effect on the phenotype, together with a major effect of the environment. It has been surmised that the allelic differences for these genes are small. This is because all alleles are presumed to be functional but have slightly different efficiencies in their contribution to the trait in question. The genes responsible for such traits were originally called polygenes by Mather (1941), but are now generally referred to as QTL (Gelderman 1975). Their existence was indicated from the fact that selection for five or ten generations from an F_2 population normally resulted in extreme lines that transgressed the range of the original population. This indicated that the allelic variation was already present in the F_2 because the time scale was too short and the responses too repeatable for new mutations to have been responsible. Also, the transgressive segregation indicated that new genotypes had been formed that were not present in the F_2 and this was only possible if two or more genes were involved. Their existence was confirmed by elegant experiments by Breese and Mather (1957, 1960) and Thoday (1961, 1979) using fruit flies. Their approach was to construct breeding lines that contained various combinations of chromosomal segments from two different parental lines. They showed that several regions of the chromosome could be associated with any given trait and that these segments did, indeed, contribute almost additively. However, at that time, they only had access to major gene mutants to identify and manipulate the segments of chromosome during the construction of the lines used to analyse their material, and these mutants sometimes also affected the traits concerned. It was not a feasible approach to be easily adapted to plant breeding. However, several workers have capitalized on the use of aneuploids in wheat to achieve the location of QTL first to whole chromosomes and then to parts of chromosomes in wheat (Sears 1953; Law 1967; Law *et al.* 1983).

Because of the difficulties of identifying the individual QTL, little progress was made in actually studying their detailed nature until the advent of molecular, DNA markers in the late 1980s (Lander & Botstein 1989). Thus, it was not – and still is not – generally known whether QTL are structural or regulatory genes, and we have no clear idea of how many there are nor the precise nature of the allelic variation. Nonetheless, quantitative geneticists have been very successful in obtaining useful information to predict response to selection, understand heterosis, optimize breeding strategies and obtain general information on the type of gene action and interaction underlying the traits (Falconer & Mackay 1996; Kearsey & Pooni 1996; Lynch & Walsh 1998). This has all been achieved simply by studying the correlations in phenotypes between relatives in a range of family structures. Quantitative geneticists adopted the view that it was easier to assume that there were many QTL of roughly equal effect and to manipulate their combined effects rather than to try and dissect their individual components. Even if the individual components were known, it was argued, it would still be difficult accurately to predict how they might work in combination.

1.3 Basic modelling of quantitative traits

It was assumed that the individual QTL follow all the basic laws of Mendelian inheritance. Thus, they segregate independently at meiosis, sometimes exhibit linkage to other QTL, show some degree of dominance though probably not over-dominance, and they could show gene interaction. Based on this basic, yet simple structure, it is possible to construct models to explain the means and variances of various types of family. The essential principles were set down by Fisher (1918) but have been considerably developed subsequently (Mather & Jinks 1982; Falconer & Mackay 1996; Kearsey & Pooni 1996; Lynch & Walsh 1998).

Consider an F_2 derived from two parental inbred lines that were identical at all genes except one, gene A. If we call the two alleles of gene A, A^+ and A^-, and the + and – indicate whether the allele increases or decreases the trait, then A will segregate in the F_2 with frequencies and genetic values as shown below:

F_2 genotypes	A^-A^-	A^+A^-	A^+A^+	mean
Frequency	¼	½	¼	
Genetic value	$m-a$	$m+d$	$m+a$	$m+½d$

It is customary to define parameters such as a and d as deviations from the mean of the two homozygotes m, although there are other approaches. The genetic value is the model of the parameters contributing to the mean phenotype of each genotype. Thus, replacing the one homozygous allele with the other causes the mean to change by $2a$. The homozygous effect, a is referred to as the additive genetic effect as opposed to d the dominance effect. The value of d is assumed to be between +/– a in size; that is, an allele may show zero, partial or complete dominance but, based on most known alleles for major genes, over-dominance, that is where $d > a$, is **not** expected to occur.

The genetic variation among the F_2 individuals around the F_2 mean for this single gene is:

$$\Sigma(\mathrm{x} - \bar{\mathrm{x}})^2 = \tfrac{1}{4}(-a - \tfrac{1}{2}d)^2 + \tfrac{1}{2}(d - \tfrac{1}{2}d)^2 + \tfrac{1}{4}(a - \tfrac{1}{2}d)^2$$

$$= \tfrac{1}{4}\,(-a)^2 + \tfrac{1}{2}\,(d)^2 + \tfrac{1}{4}\,(a)^2 - (\tfrac{1}{2}d)^2$$

$$= \tfrac{1}{2}\,a^2 + \tfrac{1}{4}\,d^2$$

This simply states that, for a single gene, A, the genetic variation between individuals in an F_2 population will consist of the squared additive and dominance effects in the ratio ½: ¼. This will be true for all other genes providing that they are independent in action and inheritance; that is, they do not interact and they are not linked. If these assumptions hold, then the combined variance of all the genes controlling the trait is simply the sum of their individual components. Thus the combined genetic variance, V_G will be:

$$V_G = \tfrac{1}{2}\,\Sigma a^2 + \tfrac{1}{4}\,\Sigma d^2 = V_A + V_D \qquad (1.1)$$

where V_A and V_D are the variances due to additive and dominance variation respectively. The overall phenotypic variation (V_P) will also include environmental variation (V_E) and so

$$V_P = V_A + V_D + V_E \qquad (1.2)$$

This is a fundamental equation of quantitative genetics and applies to all populations not just to F_2s, although the exact formulations of V_A and V_D will depend on the nature of the population being studied.

The relative contributions of genetic and additive genetic variation to a particular trait in a particular environment are referred to as the broad (h^2_b) and narrow (h^2_n) heritabilities of the trait, respectively. The former ($h^2_b = V_G/V_P$) indicates the overall contribution of genetical variation to the trait and so puts a limit on the extent to which the combined effects of individual genes can contribute as we will see later. The latter ($h^2_n = V_A/V_P$) indicates the amount of variation available for selection and can be used to predict the response to selection.

Equation 1.1 above indicates that a few large genes can have a disproportionate effect on V_G because it is their squared effects that are important. Thus, two genes with additive effects, a, of 3 and 1 units each will contribute 9 and 1 units to the variance. As we will see later, when the individual QTL are located and their effects, a, estimated, it is normally their relative contribution to the variance that is quoted. Of course, to a plant breeder, it is their relative contribution to the mean that may be most relevant because varieties are sold on the basis of their means, not their variances. However, if those genes with the greatest effect on a particular trait could be located and selection focused on them alone, it would provide an obvious benefit to breeders.

In what follows we discuss how the individual QTL can be located and their additive (*a*) and dominance (*d*) effects estimated. This will lead into how they can be manipulated in breeding programmes and, if located accurately enough, cloned.

1.4 Statistical principles and methods for mapping QTL

In order to obtain a better understanding of quantitative genetic, polygenic variation it is necessary to ask fundamental questions about the number, genomic positions, genetic effects and interactions of quantitative trait loci, QTL (Mather & Jinks 1982). The central approach to QTL location and analysis in the era of structural genomics is to attempt to correlate the genetic variation in a given quantitative trait with polymorphic genomic regions identified by molecular markers. This is an essential primary step for the ultimate identification of candidate genes and also to pursue our understanding of the molecular basis underlying the variation.

The basic idea behind QTL mapping is no more than that for mapping genes controlling morphological traits that show a simple pattern of Mendelian segregation, as in classical linkage studies. The degree of co-segregation of genes at different loci reflects the genetic distance between the loci under question, but the co-segregation has to be modelled and analysed using different approaches for simple Mendelian traits and quantitative traits. For the former, the pattern of phenotypic variation provides full information about the genotypic segregation at the loci because there is a one-to-one correspondence between phenotype and genotype. Thus, the gene-mapping problem can be based entirely on surveying the frequency of recombinant genotypes observed from experimental trials. For quantitative traits, on the other hand, such a one-to-one relationship no longer exists because the phenotype is the result of several genes and the environment. Therefore, the phenotypic variation provides only partial information about the segregation of the underlying genes, so the key problem for mapping quantitative trait loci is to uncover genotypic information about each individual QTL from relevant marker mapping data, using appropriate statistical methods.

Data for mapping QTL consist of three resources: trait phenotype; polymorphic genetic markers; and genetic structure of mapping populations. The phenotypic record of an individual for a trait reflects the genetic effects of QTL alleles that the individual carries as well as environmental contributions to the development of the character. Because markers have individually recognizable effects, they can be tracked and mapped like major genes. The genetic structure of a mapping population defines the domain in which genes at individual QTL segregate and the pattern of recombination between genes at linked loci. The statistical task of QTL mapping analysis is essentially to bridge the relationship between the trait phenotype with the genotype at the genomic regions specified by the marker loci. In this section, we review the development of the major statistical tools used in QTL analysis and explore their properties and utilities in analysing mapping experiments. Because these methods were developed for different mapping populations, the following discussion is organized on the basis of population type.

1.4.1 Molecular markers for QTL mapping

QTL can only be mapped by following their co-segregation with other markers, and it has been the proliferation of simple, reliable molecular marker methods that has been responsible for much of the progress in this area over recent years. There is not space here to present a full review of molecular marker technologies, but some basic points will be made. For a more detailed coverage of this area, the reader is referred to Staub *et al.* (1996) and Westman and Kresovich (1997).

In the approximate chronological order of their development, the major molecular marker types have included isoenzymes, RFLPs, SSRs, AFLPs and SNPs (see below). Isoenzyme methods depend upon the electrophoretic separation of proteins in a non-denaturing gel. This is followed by enzyme-specific staining which allows the visualization of bands of coloured reaction products (Hamrick & Godt 1990). The technique is robust and the marker is co-dominant (both alleles can be scored in a heterozygote), but the number of markers that can be employed is severely limited by the number of enzyme-specific stains that are available.

The scoring of restriction fragment length polymorphisms (RFLPs) requires the availability of sets of DNA sequences that can be used as (normally radioactive) probes (Tanksley *et al.* 1989). These are often cDNAs, selected from a library produced for the species under study. They are used to hybridize to homologous sequences on DNA fragments that have been produced by the digestion of the genome of test genotypes by a restriction enzyme (such as *Eco*RI). Genomic fragments, which may vary in size in different genotypes, are separated by gel electrophoresis and then transferred to a membrane filter by blotting before probe hybridization. Again, the technique is robust and the marker is co-dominant. RFLP technology remains extremely useful in genetic mapping, but in recent years has often been replaced by polymerase chain reaction (PCR)-based methods that require smaller quantities of DNA (and hence tissue for extraction) and are usually faster.

Simple sequence repeats (SSRs) are a widely used marker type that relies upon the high rate of polymorphism observed at microsatellite loci (Goldstein & Sclotterer 1999; Morgante & Olivieri 1993). These are tandem repeats of short units (usually one to four bases) that are widespread within eukaryotic genomes. Variation in the numbers of repeats is observed by developing locus-specific primers that anneal to sequences flanking the repeat region, and then using PCR to amplify the intervening DNA fragment. Alleles are visualized as bands with differing mobilities on a gel, with the marker again being co-dominant in nature. The hyper-variability in the microsatellite repeat numbers means that one is very likely to detect different alleles when one genotypes two parents used in a cross. However, considerable work has to be carried out in a species to obtain sequence data at exploitable loci before routine genotyping can be achieved.

Amplified fragment length polymorphisms (AFLPs), on the other hand, require no such preliminary work (Vos *et al.* 1997). The same kits of oligonucleotides can be used with any plant species. In this technique, DNA is digested with two different restriction enzymes (e.g. *Eco*R1 and *Mse*1) creating differing 'sticky ends'. Different adapters (short double-stranded DNA sequences) are added to the different sticky ends, after which primers specific to the two adapters are used to direct amplification of the fragments. As described, this would lead to the amplification of every restriction fragment leading to an uninformative smear of bands on a gel. However, the critical characteristic of the AFLP technique is that the primers used for amplification carry short extensions (or 'anchors', that are typically three bases long) at their 3' ends so that only a small sub-set of adapter-ligated restriction fragments is selectively amplified. One of the primers used is usually radioactively labelled, and the amplification products are typically separated on a large polyacrylamide sequencing gel. This is dried and 50 to 100 bands can normally be scored when it is subjected to autoradiography. The large numbers of markers obtained are offset by the disadvantages that the technique is more complex than most other PCR-based methods and that the markers are dominant: the scored alleles are 'band present' and 'band absent'. Increasingly, automated DNA sequencers – which are in essence DNA fragment analysers – are being used for the separation and scoring of amplification products produced by the SSR and AFLP procedures. Automated DNA sequencers require the use of fluorescent labels for fragment detection. By attaching different fluorescent labels to individual primers, one can distinguish amplification fragments obtained for different loci in the same sample. This so-called 'multiplexing' – along with the fact that many sequencers can analyse several 96-well plates of samples in a few hours and that liquid-handling robots can be used to set up the necessary PCRs – has allowed many commercial breeding programmes to make use of high-throughput genotyping.

The most recent molecular marker method to be used for plant genotyping is single nucleotide polymorphism (SNP) technology (Schafer & Hawkins 1998; Chicurel 2001). This name is somewhat misleading, since the difference between alleles at polymorphic RFLP or AFLP loci may be due to single nucleotide changes. However, SNP technology has become the generic term used for a series of high-throughput methods that each directly scores alleles that differ by a single point mutation at specific loci. The methods by which differing alleles are scored vary considerably. In some cases this involves the use of a sequencer following a primer extension protocol, while in other cases it requires the use of mass spectrometry to distinguish allelic DNA fragments on the basis of their exact mass. In either case, routine application of SNP technology for molecular marker studies requires that the DNA sequences of sets of loci are known for both parents of a cross. For the current user in an academic environment, this largely restricts genome-wide SNP genotyping to studies of *Arabidopsis* for which SNPs between the Columbia and Landsberg erecta genotypes have been made publicly available. However, in a commercial context, where extensive sequencing data may be held on particular crop species, this approach offers the opportunity for rapid and reliable genotyping of large numbers of progeny plants.

1.4.2 QTL mapping in segregating populations

The most commonly used populations for QTL analysis in crop and some model animal species are segregating populations created from two inbred lines or strains. These strains are usually assumed to be homozygous with different alleles at both QTL and genetic markers. These refer mainly to F_2 and backcross populations but also to their inbred derivatives (see section 1.5.1). There are several major advantages of using such segregating populations for QTL mapping analysis:

1. Simple bi-allelic segregation at each of these loci.
2. Full information about the linkage phase of genes at the marker loci and QTL.
3. Ease in creating a large full-sib family size.
4. Versatility of the experimental design for both detecting marker-QTL linkage and estimating genetic parameters defining genetic effects at the QTL.

The most powerful statistical method for modelling segregating populations for QTL mapping can be traced back to the benchmark paper by Lander and Botstein (1989). The method has come to be known as ‘interval mapping’ because it systematically searches all possible QTL locations within every chromosomal interval flanked by a pair of adjacent marker loci. Interval mapping analysis considers the following linear model to test for and to localize a putative QTL on an interval flanked by markers M_i and M_{i+1}. Under the model, the phenotypic record (y_j) of the j^{th} individual within a random sample of size n from a segregating population is given as

$$y_j = u + u_k + e_j$$

where: u is the population mean; u_k is the genetic value of the genotype of that QTL where k represents the particular QTL genotype (of which there are three possible in an F_2); e_j reflects the residual random variation in the model which is assumed to follow a normal distribution with mean zero and the variance σ^2.

The difficulty of statistical inference based on the above model lies in the fact that the genotype at the QTL is unknown, and so its genetic effect, u_k will have to be predicted from analysing the likelihood given below. The formula states the likelihood of there being a QTL at a particular location, its genetic effect and residual variance due to environmental variation and genetic segregation at other QTL, given the data of trait phenotype and genotype at the flanking marker loci. Statistical analysis of QTL mapping essentially involves a search of the likelihood function for values of the parameters which maximize the likelihood.

$$L_q(u;u_1,u_2,..u_S;\sigma^2) = \prod_{j=1}^{n} \{\sum_{k=1}^{s} h_{fk}\,\phi([y_j - u - u_k]/\sigma)\}$$

In the likelihood function, S is the number of possible genotypes at the QTL within the population (for example, it takes a value of 2 for a backcross population and 3 for an F_2 population); h_{fk} is the conditional probability of the individual having the k^{th} genotype at the QTL $(1 \leq k \leq S)$ given its genotype $f (1 \leq f \leq R)$ at the flanking marker loci, M_i and M_{i+1} (R = 4 or 9 for a backcross or an F_2 population respectively); and $\phi(\bullet)$ stands for the probability density function of a standard normal distribution. Calculation of the conditional probability h_{fk} depends on the genetic structure of the segregating population in question (i.e. F_2, backcross, etc.) and the location of the QTL within the flanking marker interval, and has been illustrated elsewhere (Lander & Botstein 1989; Luo & Kearsey 1992).

Differentiating the logarithm of the likelihood function with respect to each of the unknown parameters and setting the corresponding derivative to zero gives

$$\frac{\partial \ln L_q}{\partial x} = \sum_{j=1}^{n} \frac{1}{\sum_{l=1}^{S} h_{fl}\, \varphi([y_j - u - u_l]/\sigma)} \sum_{k=1}^{S} h_{fk} \frac{\partial}{\partial x} \varphi([y_j - u - u_k]/\sigma) = 0$$

Taking $x = u, u_k (k=1, 2,..., S)$ and σ^2 in order, and solving the differential equations, yields the maximum likelihood estimates (MLE) of the model parameters given by

$$\hat{u} = \sum_{j=1}^{n} \sum_{k=1}^{n} w_{jk} (y_j - u_k)/n$$

$$\hat{u}_k = \sum_{j=1}^{n} w_{jk} (y_j - \hat{u}) / \sum_{j=1}^{n} w_{jk}$$

in which

$$w_{jk} = \frac{h_{fk}\, \varphi([y_j - u - u_k]/\sigma)}{\sum_{l=1}^{S} h_{fl}\, \varphi([y_j - u - u_l]/\sigma)}$$

represents the posterior probability of individual j having the k^{th} genotype at the QTL given its trait phenotype y_j and flanking marker genotype f. In terms of the basic model given earlier for the additive and dominance effects of a QTL these parameters translate to: u = m; u_1 = m – a; u_2 = m + d; u_3 = m + a.

The algorithm demonstrated above for calculating the MLEs consists of two steps. The first step calculates the posterior probability distribution of the missing information about the QTL genotype; it results in producing an expected value for the missing data. Thus, this step is termed E-step for 'expectation' step. The second step involves calculating the MLEs of the unknown parameters. It is achieved by making use of the posterior probabilities and is termed the M-step for 'maximization' step. These two steps of numerical calculation are iterated starting with initial guesses for the parameter values (e.g. u, u_k and σ^2) being, for example, the corresponding sample estimates, until the likelihood function converges at any given prior criterion. A rigorous mathematical treatment of the algorithm can be found in Dempster *et al.* (1977).

The test statistic for the presence of the putative QTL at the given chromosomal position within the flanking interval is the likelihood ratio

$$LR = 2\ln\frac{L_q(\hat{u};\hat{u}_1,\hat{u}_2,..,\hat{u}_S;\hat{\sigma}^2)}{L_o(\hat{u},\hat{\sigma}^2)}$$

where $L_o(\bullet,\bullet)$ is the likelihood of the null hypothesis and is calculated at the sample mean and variance of the trait phenotypic records. The likelihood ratio is converted into the log-odds score (LOD) by $LOD = (log_{10}e)LR/2 = 0.217LR$, which is asymptotically distributed as a chi-square distribution with $df = S - 1$ under the null hypothesis.

The test above can be carried out at any given chromosomal position that is bracketed by a pair of marker loci. Thus, the method represents a systematic procedure to search for QTL over the whole genome and reduces a multi-dimensional search problem for multiple QTL to one involving just a single dimension. Because the test is performed repeatedly at multiple locations, a practical problem arises of determining an appropriate threshold for declaring the presence of QTL at a genome scanning level. In other words, how to avoid false positives. Several methods have been proposed in the literature to determine the threshold. Lander and Botstein (1989) developed an asymptotic formulation, which was based on an Orenstein-Uhlenbeck diffusion process, for a genome-wide LOD score threshold. They suggested the value of the threshold should be between 2 and 3 to ensure a 5% overall false-positive error. Alternatively, Churchill and Doerge (1994) proposed a data-based numerical method, based on the theory of the permutation test, to determine the critical value of the significance test in the QTL analysis. A permutation test is a general numerical approach for calculating a significance threshold for a test statistic. The test statistic is calculated based on random permutations of the data, simulating random sampling of the data under the null hypothesis that there is no QTL. This procedure is repeated a large number of times, resulting in a series of values of the test statistic under the null hypothesis. The $1 - \alpha$ percentile of these observed values of the test statistic gives the critical value (i.e. the threshold) at significance level α. The theoretical basis behind this approach was explained in detail in Lehmann (1986). The permutation test models the null hypothesis by essentially de-coupling the trait and marker data and deriving an empirical distribution of the test statistic.

The above analysis provides point estimates of QTL map locations, genetic effects and residual variances of the QTL. In practice, it is very important to know the sampling variances associated with these estimates, that is, how accurate they are. In order to estimate the confidence interval of a QTL map location, Lander and Botstein (1989) proposed the use of a one LOD support interval based on the asymptotic chi-square distribution of the likelihood ratio test statistic under the null hypothesis. Mangin *et al.* (1994) pointed out that this method is appropriate only when the QTL effect is large. However, when the QTL effect is small, the test statistic may not follow a chi-square distribution and, as result, the one LOD support interval underestimates the confidence interval. They developed a novel statistic with asymptotic properties that do not depend on the QTL effect. In this, they were the first to provide an unbiased and feasible method for estimating the confidence interval of a QTL map location. In addition, Visscher *et al.* (1996) suggested calculating the sampling variances of the parameter estimates by making use of a bootstrap sampling method. Bootstrapping simulates the estimation of the QTL parameters from many repeats of the experiment. It achieves this by repeatedly sampling from the existing experimental trait and marker data with replacement, every sample mimicking a repeat of the experiment (Davison & Hinkley 1998). The statistical analysis is similarly performed on each replicate data set, yielding repeated estimates of the parameters. Variances of these repeated estimates provide the sampling variances of the estimates. More recently, Kao and Zeng (1997) formulated the observed information matrix of the MLEs of the model parameters in an interval mapping analysis. In theory, the inverse of the matrix provides the corresponding estimates of the sampling variances. However, Luo *et al.* (2000) showed that this approach might fail to produce stable estimates of the sampling variance when convergence of the 'expectation and maximization' (EM) algorithm was slow, and an approximate but robust method for calculating the stable estimates was proposed.

The interval mapping theory is built as the kernel of the statistical framework for QTL mapping. Since its publication, many modified versions based on the basic principle have been proposed for improving various aspects of this method. A simple regression model was suggested in Haley and Knott (1992). Instead of treating the quantitative trait as a mixture model with missing information, the regression analysis models individual phenotypic record y_j by

$$y_j = u + \xi_{jk} u_k + e_j$$

in which ξ_{jk} is the conditional probability of the individual having the k^{th} QTL genotype given its genotype at the flanking marker loci and the test position of the putative QTL. Statistical analysis of the model is straightforward, and it provides the flexibility to fit different fixed effects in the model such as site, sex and maternal effects. Numerical analyses based on simulation studies have shown that this procedure gives a very good approximation to the likelihood-based analysis even though it was pointed out by Xu (1997) that the regression approach tends to overestimate the residual variance.

One practical problem of interval mapping with a pair of flanking markers is that the test statistic at any test position will be affected by other QTL linked or unlinked to the test position. To overcome this problem, Zeng (1994) and Jansen and Stam (1994) suggested the use of other markers in addition to the flanking markers as a background control in the interval mapping analysis. The composite interval mapping for such analysis combines interval mapping with multiple regression and fits the following linear model for quantitative trait value of the j^{th} individual as

$$y_j = u + u_k + \sum_{i=1}^{m} b_i x_{ji} + e_j$$

where x_{ji} is the type of the i^{th} marker in the individual and b_i is the partial regression coefficient of phenotype y on the marker i conditional on all other markers. The consequences of incorporating additional m markers as cofactors into the model depend on the relationship between these markers and the flanking markers. The role played by the cofactor markers is similar to that of covariates in multiple regression analysis. When the additional markers are linked to the flanking markers, the effect of those QTL located outside the boundaries defined by these markers will be effectively controlled. This leads to a higher mapping precision, but at the same time the statistical power for detecting the QTL may be reduced under the conditional test (Zeng 1993). However, use of unlinked markers as cofactors in the model may be effective in reducing residual variation of the model, and in turn leads to increase in the test power.

A common feature of the interval mapping and the composite interval mapping protocols is their one-at-a-time strategy, whereby one evaluates the association of a single QTL with a marker interval while ignoring the interaction of the tested QTL and other QTL segregating in the mapping population. Because epistatic effects between different QTL are found to be a common phenomenon for most quantitative complex traits (Mackay 2001a), these one-at-a-time approaches are limited in detecting such effects. To meet this requirement, Kao *et al.* (1999) proposed a multiple QTL interval mapping approach, which considers all possible parameters defining the genetic architecture of polygenic inheritance in a likelihood-based statistical analysis. These include the number, effects and epistasis of QTL, genetic variance and covariances explained by QTL effects. Implementation of the multiple interval mapping is computationally much more demanding than other interval mapping approaches but its dynamic search for all significant genetic components makes the model fitting more close to the real genetic architecture of quantitative traits.

It must be pointed out that the interval mapping and its modified versions share the common problem that estimates of QTL map locations and effects are highly model-dependent. The use of a misleading model may result in severely biased prediction of the genetic parameters. To avoid this malpractice, QTL mapping analysis should not be built solely on a simple additive/dominance model but integrate all aspects of genetic analysis of quantitative traits (Mackay 2001b).

Efficiency of QTL mapping analysis may be influenced by many factors, predominant among which are population size and trait heritability. The density of markers scored in a mapping experiment defines the scale of map information, but there exists an upper limit to improvement in the efficiency of QTL mapping through increasing marker density. The use of an extremely dense marker map will be of little use if the mapping population fails to provide sufficient recombination between the markers and QTL (Hyne *et al.* 1995). Moreover, the segregating populations derived from different mating designs can be characterized with different genetic structures and thus exhibit varying utilities in QTL mapping. For example, use of backcross populations is more powerful than use of F_2 populations for detecting QTL effects, but the latter are preferred for achieving better estimates of the effects (Darvasi 1997).

1.4.3 QTL mapping in pedigree populations

In many out-breeding species the establishment of inbred lines is not practical. Mapping QTL in these species cannot be performed by use of simple segregating population as discussed above. Populations of these species (e.g. most trees) exist in a pedigree structure. In sharp contrast to the segregating populations, mapping genes segregating in the pedigree population is much more problematic. First, the size of the nucleus family in most pedigree populations is substantially smaller than commonly used segregating populations in QTL analysis. The whole analysis involves a large number of independent pedigree families to ensure an adequate statistical power for detecting linkage between markers and QTL. Second, information about the linkage phase of genes at the marker loci and QTL is no longer directly extractable from these pedigrees. To achieve this requires the development of complicated statistical tools for modelling the inheritance of genes within a multiple generation pedigree and sophisticated computational algorithms to assess the likelihood of all possible configurations of linkage phases at a finite number of loci. Guo and Thompson (1992) developed a Gibbs sampling-based approach to combine conventional segregation analysis at individual loci with linkage analysis. Gibbs sampling is a numerical approach to calculate marginal distributions from joint distribution (Casella & George 1992). In the linkage analysis setting, this technique was used to calculate the likelihood function by integrating over the polygenic additive effects. This provides a test for, and estimates of, the linkage between a single marker locus and a locus underlying quantitative genetic variation in a large complicated pedigree. Third, the fact that different families may bear different alleles at QTL poses another severe problem of genetic heterogeneity in the linkage analysis. All of these make QTL mapping in pedigree populations a challenging topic in both theoretical and experimental studies.

There are several approaches of QTL analysis that make use of multiple pedigrees with a simple consanguineous relationship among members. The basic idea of the interval mapping of QTL was extended by Fulker and Cardon (1994) to model a series of independently collected sib-pairs. Instead of working with the probability of QTL genotype conditional on flanking marker genotype, they modelled the proportion of genes IBD (identical by descendent) shared by a sib-pair at a putative QTL (Π_q) in terms of the IBD genes shared at two flanking marker loci (Π_l and Π_2) as

$$\Pi_q = \alpha + \beta_1 \Pi_1 + \beta_2 \Pi_2$$

in which

$$\beta_1 = [(1-2c_1)^2 - (1-2c_2)^2(1-2c)^2]/[1-(1-2c)^4]$$

$$\beta_2 = [(1-2c_1)^2 - (1-2c_2)^2(1-2c)^2]/[1-(1-2c)^4]$$

$$\alpha = (1-\beta_1 - \beta_2)/2$$

where c_1, c_2 and c are respectively recombination frequencies between the left flanking marker and the QTL, between the QTL and the right flanking marker, and between the flanking marker loci. Regression of the difference in trait phenotype record between the sib-pair on the IBD proportion at the putative QTL (Π_q) creates a test statistic for detecting the presence of the QTL. The analysis can be performed at any position within the flanking interval, and thus the method essentially provides a systematic search for QTL. Simulation studies suggested that thousands of sib-pairs are needed in order to obtain adequate power to detect the QTL and meaningful estimates of the QTL parameters. However, the efficiency of the design should be improved substantially if the sib-pairs with extreme phenotype are selectively used in the QTL mapping analysis (Risch & Zhang 1995).

1.4.4 QTL analysis in natural populations

The precision with which a single QTL can be localized relative to a marker locus is directly proportional to the number of informative meioses provided by a mapping population. A large number of such informative meioses indicate a large number of recombinations between the marker and trait locus. The linkage analyses demonstrated in the previous two sections are typically restricted in the number of such useful meioses, and thus have yielded poor mapping resolution. It was observed that the 95% confidence interval for QTL map locations inferred from many plant experiments were often in a range of 20 to 30 cM and seldom less than 5 cM (Kearsey & Farquhar 1998). This is too coarse for utility of the mapping information in marker-assisted selection for genetic improvement of quantitative traits or in targeting candidate genes affecting the traits. Use of historically accumulated recombinations in the populations from a well-designed breeding scheme (Xiong & Guo 1997) or in natural populations (Lander & Schork 1994) has been shown to be an effective way to improve resolution of QTL mapping. Methodologically, use of natural populations for mapping genes underlying quantitative traits requires modelling linkage disequilibrium between genes segregating at marker loci and trait loci.

Linkage disequilibrium is a central concept of population genetics and is defined as the non-random association of alleles at different loci in a given population. The degree of the non-random association is referred to as the coefficient of linkage disequilibrium (Crow & Kimura 1970). The basic idea behind the linkage disequilibrium analysis for gene mapping has been demonstrated in Terwilliger (1995) and Kaplan and Weir (1997). Suppose that a mutation at a gene affecting a character occurred many years ago and – possibly through a founder effect – was propagated in the population. Thus, it is possible that the marker alleles on the original mutant haplotype may still be in linkage disequilibrium with the mutant allele in the