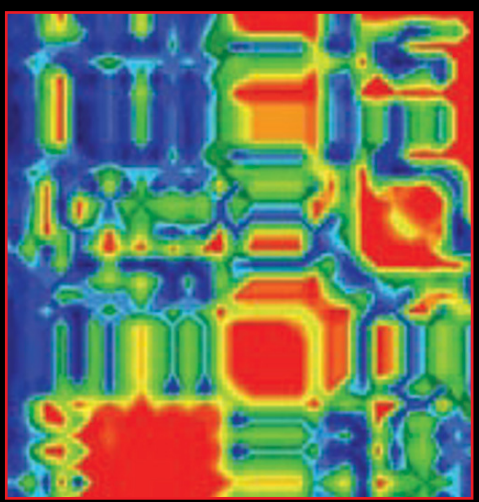
ASSOCIATION MAPPING IN PLANTS



Edited by

Nnadozie C. Oraguzie
Erik H. A. Rikkerink
Susan E. Gardiner
H. Nihal De Silva

Association Mapping in Plants

Association Mapping in Plants

Edited by

Nnadozie C. Oraguzie
The Horticulture and Food Research Institute of New Zealand Ltd (HortResearch)
Havelock North, New Zealand

Erik H.A. Rikkerink
The Horticulture and Food Research Institute of New Zealand Ltd (HortResearch)
Auckland, New Zealand

Susan E. Gardiner
The Horticulture and Food Research Institute of New Zealand Ltd (HortResearch)
Palmerston North, New Zealand

H. Nihal De Silva
The Horticulture and Food Research Institute of New Zealand Ltd (HortResearch)
Auckland, New Zealand

Springer

Dr Nnadozie C. Oraguzie
HortResearch
Hawkes Bay Research Centre
Private Bag 1401
Havelock North
New Zealand

Dr Erik H.A. Rikkerink
HortResearch
Mt. Albert Research Centre
Private Bag 92169
Auckland
New Zealand

Dr Susan E. Gardiner
HortResearch
Palmerston North Research Centre
Private Bag 11030
Palmerston North
New Zealand

Dr H. Nihal De Silva
HortResearch
Mt. Albert Research Centre
Private Bag 92169
Auckland
New Zealand

Library of Congress Control Number: 2006928327

ISBN-10: 0-387-35844-7 e-ISBN-10: 0-387-36011-5
ISBN-13: 978-0387-35844-4 e-ISBN-13: 978-0-387-36011-9

Printed on acid-free paper.

9 8 7 6 5 4 3 2 1

springer.com

Preface

The approach taken for locating the genes that underlie human diseases has shifted from pedigree-based linkage studies to population-based association studies. In both cases the proximity of a genetic marker to a susceptibility locus is inferred from statistical measures that reflect the number of recombination events between them: in a disease pedigree there are no more than a few hundred opportunities for recombination so that recombination rates less than about one percent cannot be estimated and genes can be located only coarsely on a genetic map with that approach. The linkage disequilibrium detected in an association study, however, reflects the actions of many thousands of recombination events since the initial disease mutation and the expectation is that susceptibility genes can then be mapped more accurately.

The editors of this volume have recognized the need for parallel activity in plant species. For the past 20 years, the genes that affect plant economic traits have usually been mapped with data collected from "pedigrees" of populations formed by crossing inbred lines. These Quantitative Trait Loci have been mapped on a coarse scale, and a QTL is likely to refer to several genes in a region. The move to population-based association studies was therefore as necessary in plants as it was in humans, and readers will find this book to be a useful review of the marker technology, statistical methodology, and progress to date. Although one of the authors fears that "plant genetics can be considered as less advanced than human genetics" the chapters suggest that if that is the case it will not be so for long.

The recent increased activity in association mapping in humans has rested on the development of efficient and affordable methods for discovering and employing Single Nucleotide Polymorphism markers. Plant geneticists cannot command the resources available to their human geneticist colleagues, but they can anticipate benefiting from the success of the International HapMap Project. The improvement in marker technology from such large projects will inevitably be imported to plant studies. The editors have provided helpful guides to the use of SNPs in association studies.

Along with the substantial increase in the volume of data when large numbers of individuals are typed at millions of SNPs there are substantial challenges in the statistical interpretation of the data. This book contains a valuable account of the issues of multiple testing and an accessible account of False Discovery Rates. The more basic concepts of linkage disequilibrium and case-control versus family-based association tests are also discussed. It is often the case that geneticists do not receive extensive statistical training and the coverage of the theory of estimation and testing is therefore welcome. Readers will notice a greater use of Bayesian methods than is usually found in statistical genetics books. Such methods are appearing more frequently in scientific papers.

I congratulate the editors and all the authors on this timely and comprehensive treatment of association mapping in plants. The importance of food and fiber for human welfare cannot be overstated, and progress in plant improvement will rest in no small part on the work described in these pages. On a personal level, I am delighted by the leadership shown by my fellow antipodeans.

B.S. Weir
Professor and Chair
Department of Biostatistics
University of Washington

Acknowledgements

The motivation to write this book came from numerous correspondences and finally a meeting with Keri Witman, former senior editor in plant sciences at Springer Science+Business Media, New York, USA (who continually stressed the need for a book in this field of research), in Tulln Austria in 2004 at an EUCARPIA Quantitative genetics conference. The recommendation and suggestions by anonymous referees consulted by Springer to review the original proposal kept up the initial enthusiasm. This book would not have been possible without HortResearch's support. In particular, we would like to thank the following people; Drs Vincent Bus (Science Leader, Pipfruit and Summerfruit breeding), Andrew Granger (Future Fruit Group Leader) and Bruce Campbell (General manager science operations) for their encouragement and moral support, Stuart Ritchie for assistance with contract negotiation, Sharlene Cookson for assistance with book cover design, Dr David Chagné for his immense contribution right from the proposal stage of the book till it went into press, and the SPU team especially, Dr Anne Gunson and Christine Lamont, for editorial and technical assistance. We are grateful to the following for permission to reproduce copyright material: Trustees of the Royal Botanic Gardens, Kew; Elsevier (Figure 1 in Trends in Genetics 11: 83-90(2002) and Tables 1 & 2 in Trends in Genetics 20(2): 105(2004)); Swedish National Biobanking program, Wallenberg Consortium North (http://www.meb.ki.se/genestat.htm/-pairwise D` for 45 SNPs within a linked region); The University of Chicago press (Table 3, American Journal of Human Genetics 60(3): 676-690, 1997); Annual Reviews (Table 1 in Annual Review of Plant Biology 54, 2003); Nature Publishing group (Table 3 in Nature Genetics 28:286-289); and MacMillan publishing Inc. (Tables 2, 3, 4, 5 & 6 in Genetics 170:859-873, 2005). Finally we thank the following colleagues who reviewed the manuscripts and made suggestions that significantly improved the quality of the chapters: Drs David B. Neale (Professor of Forest Genetics, Dept of Plant Sciences, UCLA), Mark E. Sorrells (Professor, Dept of Plant Breeding and Genetics, Cornell University, Ithaca, NY), Trudy F.C. Mackay (WNR Distinguished Professor of Genetics, North Carolina State

University), Christophe Plomion (Molecular Geneticist, INRA, Cedex, France), David Pot (Molecular Geneticist, CIRAD, Montpellier, Cedex, France), Pauline Garnier-Géré (Forest Genetics Group, INRA, Cedex, France), Fred van Eeuwijk (Associate Professor in Statistics, Laboratory of Plant Breeding, Wageningen University and Research center, Wageningen, The Netherlands), Rasmus Neilsen (Ole RØmer Fellow/Professor in Statistical Genomics, University of Copenhagen, Denmark), Diane Mather (Professorial Research Fellow, School of Agriculture and Wine Science, University of Adelaide/Program Leader-New Molecular technologies, Molecular plant breeding cooperative Research Center, Plant Genomics Center, Glen Osmond SA, Australia), Martin Lascoux (Professor-Program in Evolutionary Functional Genomics, Evolutionary Biology Center, Uppsala University, Sweden), Hans van Buijtenen (Professor emeritus-Genetics, Dept of Forest Science, Texas A & M University), and Slade Lee (Associate Professor/Deputy Director, Center for Conservation Genetics/Sub-Program Leader (Phenomics), Grain Foods CRC Limited, Southern Cross University, Lismore NSW, Australia.

Contents

Contributors

Roderick D. Ball
Ensis Wood Quality, Scion (New Zealand Forest Research Institute Limited), 49 Sala Street, Private Bag 3020, Rotorua, New Zealand

Jacqueline Batley
Primary Industries Research Victoria, Victorian AgriBiosciences Centre, La Trobe R&D Park, Bundoora, Victoria 3083, Australia

Rowland D. Burdon
Ensis Genetics, Scion (New Zealand Forest Research Institute Limited), 49 Sala Street, Private Bag 3020, Rotorua, New Zealand

David Chagné
HortResearch, Plant Gene Mapping group, Private Bag 11030, Palmerston North, New Zealand

Noel O.I. Cogan
Primary Industries Research Victoria, Victorian AgriBiosciences Centre, La Trobe R&D Park, Bundoora, Victoria 3083, Australia

Craig E. Echt
USDA Forest Service, Southern Institute of Forest Genetics 23332 MS Highway 67, Saucier, MS 39574, USA

David Edwards
Primary Industries Research Victoria, Victorian AgriBiosciences Centre,
La Trobe R&D Park, Bundoora, Victoria 3083, Australia

John W. Forster
Primary Industries Research Victoria, Victorian AgriBiosciences Centre,
La Trobe R&D Park, Bundoora, Victoria 3083, Australia

Mark P. Dobrowolski
Primary Industries Research Victoria, Plant Genetics and Genomics Platform, Hamilton Centre, Mt Napier Road Hamilton, Victoria 3300, Australia

Susan E. Gardiner
HortResearch, Palmerston North Research Centre, Private Bag 11030, Palmerston North, New Zealand

Nnadozie C. Oraguzie
HortResearch, Hawkes Bay Research Centre, Cnr. Crosses and St. George's Roads, Private Bag 1401, Havelock North, New Zealand

Erik H.A. Rikkerink
HortResearch, Mt Albert Research Centre, 120 Mt Albert Road, Private Bag 92169, Auckland, New Zealand

H. Nihal De Silva
HortResearch, Mt Albert Research Centre, 120 Mt Albert Road, Private Bag 92169, Auckland, New Zealand

Phillip L. Wilcox
Cell wall Biotechnology Centre, Scion (New Zealand Forest Research Institute Limited), 49 Sala Street, Private Bag 3020, Rotorua, New Zealand

Introduction

Most traits we deal with on a daily basis have complex inheritance patterns that complicate the ability of existing mapping technologies to detect the underlying genetic factors. In the last decade or so, we have seen the successful use of conventional map-based strategies in identification and cloning of quantitative trait loci (QTLs) in model plant species including tomato and Arabidopsis. However, efficient gene discovery with this method will probably continue to be largely limited to those loci that have large effects on quantitative trait variation. Techniques are also needed to more rapidly identify genes that play a modest role in regulating quantitative trait variation. Association mapping via linkage disequilibrium or LD (non-random association of alleles at different loci) offers promise in this area. The traditional approach of linkage/QTL mapping reliant on developing large mapping populations continues to suffer from lack of mapping resolution inherent in samples with limited meiotic cross-over events. These problems are exacerbated in tree crops, where very large populations are impractical from a plant management point of view. In association mapping, there may not be any need to make crosses initially to generate segregating populations. The natural variation that exists in the available germplasm can be utilized for mapping straightaway.

Association genetics via LD mapping is an emerging field of genetic mapping that has the potential for resolution to the level of individual genes (alleles) underlying quantitative traits. LD mapping is a technology that can take full advantage of the phenomenal leaps and bounds in technology development in the area of molecular biology and marry it with our increasing understanding of the molecular basis of inheritance and molecular tools recently developed in terms of molecular markers and genetic maps in a way that could have a significant practical impact on breeding. The convergence of improved statistical methods, availability of growing plant genomics databases and improvements in the affordability and potential scale of sequencing and

genotyping, suggests that this technology will probably be more widely adopted for mapping and gene discovery in plants in the near future.

This book provides a basic understanding of association mapping and an awareness of population genomics tools available to facilitate mapping and identification of the underlying causes of quantitative trait variation, as well as an analysis of the prospects of applying this technology to plants. In the book, we discuss how technological advances have recently brought association mapping into the realm of possibility for plants, particularly, the second and third tier crops (which include a number of long-lived tree species), which normally lag some way behind the first tier of crops (including annuals - cereals mainly) in technology development. For convenience, the book can be divided into 4 sections; there are two chapters in Section 1 which introduce association mapping and present the basic principles of association genetics in relation to the concept of linkage disequilibrium (LD); Section 2 comprises four chapters which deal with technology development in relation to SNP discovery through to SNP applications; Section 3 consists of 2 chapters providing a detailed discussion on statistical methodology and experimental design issues necessary for the successful application of association genetics; and lastly Section 4 contains 3 chapters which deal with specific issues and applications of association genetics using the crop groupings of forage, forestry and horticultural species as case examples. Application of LD mapping in model organisms including humans, Drosophila, Arabidopsis and maize is discussed in chapter 2 (in section 1).

The book brings together all the information on association genetics and linkage disequilibrium published in different journals in one volume and will be of interest to advanced breeding/genetics students, researchers, professional plant breeders and university lecturers. Breeders will find it particularly useful as a guide for making decisions on breeding strategies that will facilitate identification of ‘superior’ parents for development of new improved varieties. Difficult statistical concepts and tools are presented with detailed illustrations in a way readers can comprehend hence; the book could also serve as a teaching aid for postgraduate students. A very comprehensive comparison of statistical approaches/methodologies and guidelines on optimal study design, as well as the comparison of the relative benefits of association mapping and conventional QTL mapping will be particularly useful to geneticists wanting to set up studies on gene/genome mapping.

Association mapping now stands at the cross-roads of application to a large number of species and situations. Over the next decade it will become more apparent just how much influence this technology will have on increasing our fundamental understanding of the genetic basis of variation in plants and the practical outcomes of plant breeding in the future.

Chapter 1

AN OVERVIEW OF ASSOCIATION MAPPING

Nnadozie C. Oraguzie[1] and Phillip L. Wilcox[2]

1.1 WHAT IS ASSOCIATION MAPPING?

As the fundamental aim of genetics is to connect genotype to phenotype (Botstein and Risch 2003), association mapping seeks to identify specific functional variants (i.e., loci, alleles) linked to phenotypic differences in a trait, to facilitate detection of trait-causing DNA sequence polymorphisms and/or selection of genotypes that closely resemble the phenotype. Association mapping has been variously defined (Chakraborty and Weiss 1988; Kruglyak 1999), and has also been referred to as "association genetics," "association studies," and "linkage disequilibrium mapping"– although the latter term is also used to reflect studies detecting associations among loci. The general characteristics of this field of genetics involve the use of unstructured or loosely structured populations – usually intraspecific – that are both phenotypically and genotypically characterized to detect statistical associations between genetic polymorphisms and heritable trait variation. Some experimental designs involve use of progenies (Chapters 7 and 8). The actual polymorphisms causing trait variation are usually not known, and therefore are not directly observed, but rather, are detected via statistical inference. The predicating condition for detection of such associations is nonrandom association of causative trait polymorphisms with observed polymorphisms, i.e., linkage disequilibrium (LD). While LD can arise for a number of reasons, the primary focus of association genetics is to identify polymorphism(s) that are located physically close to the causative trait polymorphism(s). Association genetics therefore also encapsulates analytical methods aimed at determining if reasons other than close physical linkage give rise to such associations.

Association genetics shares much in common with the field of what is commonly known as quantitative trait loci (QTL) mapping. Both attempt – via statistical inference – to detect co-segregation of polymorphic genetic markers with genes underpinning trait variation. However, the two differ in terms of some key properties (Table 1.1), which have implications for the applications of each of these areas of study. QTL mapping

[1] The Horticulture and Food Research Institute of New Zealand Limited (HortResearch), Cnr Crosses and St George's Roads, P.B. 1401, Havelock North, New Zealand

[2] Scion (New Zealand Forest Research Institute Limited), 49 Sala Street, P.B. 3020, Rotorua, New Zealand

usually involves structured populations – in short generation plant species, mapping populations derived from homozygous inbred lines are commonly used. In out-crossing species single or multiple pedigrees of known relationships can be used. The net result of using such populations that are usually relatively few generations from a common ancestor is to maximize LD per base pair. Therefore, relatively distant markers can co-segregate with QTL. In contrast, unstructured populations are usually many generations descended from common ancestors, and therefore have been subjected to many more recombination events. As a result, there is much less LD between segregating markers and causative variants. Furthermore, in QTL mapping populations co-segregation occurs in a manner consistent with Mendelian expectations. In unstructured populations this is not the case as populations are not defined pedigrees and causative polymorphisms are not usually known *a priori* so therefore could be segregating at different frequencies from nearby markers, but still in disequilibrium. Polymorphisms chosen for screening could come from whole genome scans, selectively chosen but phenotypically neutral sequences, or preselected candidate genes.

Table 1.1. A comparison of association genetics and conventional QTL mapping.

Attribute	QTL mapping	Association genetics
Detection goal	Quantitative trait *locus*, i.e., wide region within specific pedigrees within which a QTN is located	Quantitative trait *nucleotide*, i.e., physically as close as possible to causative sequence(s)
Resolution of causative trait polymorphism	Low – moderate density linkage maps only required	High – disequilibrium within small physical regions requiring many markers
Experimental populations for detection	Defined pedigrees, e.g., backcross, F_2, RI, three and two generation pedigrees/families, half-sib families, etc.	Linkage disequilibrium experiments: unrelated individuals ("unstructured" populations), large numbers of small unrelated families (e.g., transmission disequilibrium tests, TDT)
Marker discovery costs	Moderate	Moderate for few traits, high for many traits
Extent of inference	Pedigree specific, except where species has high extant LD	Species or subspecies wide
Number of markers required for genome coverage	10^2–low 10^3	10^5 for small genomes –~10^9 for large genomes

Association genetics is a multidisciplinary field, involving components of genomics, statistical genetics, molecular biology, and bioinformatics which together form the basis for selecting, evaluating, and associating genomic regions for correlation with trait variation. Other disciplines are also required, particularly population genetics as well as a detailed knowledge of trait variability in the species of interest.

1.2 WHY ASSOCIATION GENETICS?

There are a number of generic applications of association genetics, which we review briefly here, and will be described in more detail throughout this book. Firstly, the higher resolution afforded by use of unstructured populations allows the intriguing possibility of identifying the genes – or even the specific nucleotides underpinning trait variation.

Secondly, the opportunity to use molecular markers to enhance rates of genetic gain, including the utilization of specific genes from non-elite germplasm in a more directed and efficient manner than was hitherto possible. A further application is the generation of fundamental knowledge around the genetic architecture of extant variation in populations, and the opportunity to determine evolutionary phenomena that have led to existing population structures.

So why is association genetics now becoming more widely used? A number of key factors have contributed to the recent interest in association genetics, including methods for high throughput gene discovery, polymorphism detection, and genotyping (see Chapters 3–5 for more discussion). The prevalence of many complex human diseases such as asthma, cardiovascular disease, bipolar disorder, and diabetes, has increased over the past two decades in developed countries (reviewed in Risch 2000). During the same period, the genetic causes of such disorders have been increasingly emphasized as a means to better understand their parthenogenesis, with the ultimate goal of improvement of preventative strategies, diagnostic tools, and treatment. Geneticists wanting to identify the genetic causes of these disorders through conventional map-based strategies including linkage analysis, QTL mapping, and positional cloning have constantly been met with only limited success. However, these map-based approaches have been instrumental in the identification and cloning of genes responsible for less common and simply inherited human disorders, as well as traits controlled by major genes in plants. Examples of such traits and the genes responsible for them in humans are breast cancer (BRCA-1 and -2), Alzheimer's disease (β-amyloid precursor protein (APP) and presenilin-1 and -2), diabetes (maturity-onset diabetes of youth (MODY)-1, -2), colon cancer (familial adenomatous polyposis (FAP)) and hereditary nonpolyposis colorectal cancer (HNPCC) (FPC), heart disease (LDL receptor genes). The best examples of simply inherited traits in plants controlled mostly by a single locus apart from Mendel's well-known examples in peas include resistance to certain pests and diseases, flower and fruit color, plant growth habit, reproductive mechanisms (such as self incompatibility), and aspects of genetic load. The map-based strategies have also been utilized for positional cloning of genes that underlie QTL in plants (reviewed in Yano 2001). For example, the morphological differences between maize and its wild relative teosinte have been studied through the analysis of QTL. As a result of such studies, one of the major QTL involved in maize domestication (*teosinte branch 1*) has been cloned. Other examples of cloned genes underlying QTL in other smaller plants are: in tomato, *Brix9-2-2* which encodes *Lycopersicon* apoplastic invertase (*Lin5*), responsible for soluble acid content, and *fw2.2*, responsible for fruit size; in rice, *Heading date 1* which encodes a protein with high similarity to that encoded by the *Arabidopsis* gene *Constans*, responsible for photoperiod sensitivity; in *Arabidopsis*, a QTL at the *Frigida* locus, responsible for vernalization response to flowering (Johanson *et al*. 2000), and an allele of *Cryptochrome 2* (*Cry2*), responsible for variation in flowering time (El-Assal *et al*. 2001).

Despite the successes of conventional QTL mapping strategies, efficient gene discovery with these methods will probably continue to be largely limited to those loci that have large effects on quantitative trait variation. These loci have large effects compared with the environmental effect. Furthermore, individuals in segregating populations can usually be assigned to discrete groups corresponding directly to their genotypes. Unlike these Mendelian traits for which (usually) alleles at single loci

determine the phenotype in a predictable manner, complex trait phenotypes are determined by alleles at many loci. Not only is the number of loci unknown, the phenotypic effects of alleles at each locus may also be influenced by the environment. Also, the relative importance of alleles at different loci could vary from family to family, and some proportion of these loci may have relatively modest effects on the phenotype. Borrowing the illustration of Risch (2000), the gene mutations that control the inheritance of simple traits (easily identified with map-based strategies) could be regarded as the "low hanging fruit that are easy to harvest" while complex traits are the "great majority of fruit at the top of the tree with no obvious way to reach." In genetic terms, these are the numerous genes of smaller effect that are likely to underlie most common familial traits and diseases in humans, and most agronomic and horticultural traits in plants. Thus, identifying genes that influence the expression of complex traits would require novel approaches and analytic strategies.

The objective of genetic mapping is to identify simply inherited markers that are physically close to the genes underlying quantitative traits. The localization of these genes relies on processes that create a statistical association between marker and QTLs and processes that selectively reduce that association as a function of the marker distance from the QTL (Jannink and Walsh 2002). For example, in a typical cross between two parents, the number of recombinant hybrids determines the distance between the marker and QTL. The more recombinants, the further the distance, and vice versa. In a cross between inbred parents mostly used to map QTL in self-pollinated crops, we create in the F_1 hybrid complete association between all marker and QTL alleles that has been derived from the same parent. Recombination in the meioses that lead to doubled haploid, F_2, or recombinant inbred lines, reduces the association between a given QTL and markers distant from it (Jannink and Walsh 2002). Because these generations of progeny have undergone relatively few meioses, even markers distant from the QTL may still be strongly associated with it. The use of advanced intercross lines first proposed by Darvasi and Soller (1995), such as F_6 or higher generational lines derived by continual generations of out-crossing the F_2, may seem to be useful for fine mapping QTL, because of the higher number of meiotic crossovers that have occurred in the populations. However, when these advance generation lines are created through selfing, the reduction in disequilibrium will not be as great as that under random mating (Jannink and Walsh 2002). Therefore, the main problem with the current approaches for fine mapping in plants is the limited number of meioses that have occurred (in the case of advanced intercross lines, recombinant inbred lines and near isogenic lines) and the cost of propagating lines for a sufficient number of meioses. Association mapping however, is an alternative approach that can take advantage of events that created association in the relatively distant past in natural populations. Assuming many generations, and therefore meioses, have elapsed since these events, recombination will have removed association between a QTL and any marker not tightly linked to it. Association mapping thus allows for much finer mapping than standard biparental cross approaches. In species that are limited to two growing seasons per year, it can take up to five years to produce the population needed for fine scale mapping with traditional linkage analysis. With long-lived perennial crops (particularly those that take up to 10 years to become reproductively viable) this could take up to 100 years. The typical resolution observed in plant genetics studies where recombinant inbred lines have been used to map QTL is 10–30 cM (Alpert and Tanksley 1996; Stuber *et al*. 1999). At this resolution (equivalent to 20–30 million base pairs) hundreds of genes within each QTL will still be left unidentified. Association

studies based on LD may allow the identification of the actual genes represented by these QTLs. Only polymorphisms with extremely tight linkage to a locus with phenotypic effects are likely to be significantly associated with a trait in populations typically used for association mapping, thus providing much finer resolution than QTL mapping based on pedigreed populations (Remington *et al*. 2001).

1.3 HOW IS ASSOCIATION GENETICS IMPLEMENTED?

There are two main approaches to association studies particularly in humans namely, case-control design and family-based design. These approaches are discussed in detail in Chapters 7 and 8. In case-control studies, marker frequencies are determined in a group of affected individuals (individuals with disease state) and compared with allele frequencies in a control population (i.e., individuals without disease). This design however, is very susceptible to population structure which could lead to spurious associations. Methods developed to control population structure in such populations are discussed in Chapter 8. The family-based design commonly referred to as the transmission disequilibrium test (TDT) generally uses family trios involving two parents (one of which is heterozygous) and an affected child. It is based on unequal transmission of alleles to the single affected child in each family, and associations are summed up over many unrelated families. The TDT approach was originally designed for dichotomous traits in human genetics studies but nowadays, several variants are available (discussed in detail in Chapter 7) which can be applied to both plants and animals and for the study of continuous traits. Another approach to association studies in other organisms, particularly plants, involves the use of unstructured populations (without progenies) which include unrelated individuals from diverse genetic backgrounds, different selection histories, diverse geographic origins, and so on, representing a range of phenotypes for the trait of interest. The methodology aims to identify as many allelic variants as possible which could potentially correlate with the trait of interest. Like the case-control studies in humans, genetic associations with such heterogeneous populations can be influenced by admixture or population stratification.

Association mapping has since been used to examine the role of candidate genes in human diseases and to refine the location of disease genes in regions previously identified by linkage analysis. Improvements in techniques for DNA sequencing and high throughput genotyping of polymorphisms particularly, single nucleotide polymorphisms (SNPs), have necessitated extending the technology to studies of an entire genome and this has resulted in the recent completion of the first phase of the human HapMap (The International HapMap Consortium 2005). Biallelic SNPs are especially attractive as genetic markers in association studies because of their high frequency, low mutation rate, and amenability to automation (see Chapters 3–6). Fast and efficient generation of these SNPs has been facilitated by high throughput genotyping methods, including DNA chips, allele-specific PCR, and primer extension approaches (discussed in Chapter 5), thus making SNPs a marker of choice for association genetics studies. In case-control studies, for example, differences in disease frequencies between groups (or in trait levels for continuously varying characters) are compared with differences in allele frequencies at an SNP to check for statistically significant correlation between the SNP and the disease. Thus the frequencies of the two variant forms (alleles) of an SNP are of primary interest for identification of genes affecting traits of interest.

The key advantages of association tests include their speed, because mapping populations may not be necessary, particularly in crops that are limited to no more than one generation per year. Controlled breeding is lacking in humans, as are large numbers of progeny per family. These among other reasons may well be why association genetics approaches have been exploited better and to a higher degree in human genetics studies. The other advantage is high resolution as already mentioned. The resolution and cost of association approaches depend largely on the nature and extent of LD, i.e., the nonrandom association of alleles in test populations. LD can result from population structure, selection, drift, or physical linkage. The physical extent of LD around a gene determines the effectiveness of association mapping, and this could result from many factors, including rate of out-crossing, the degree of artificial or natural selection on the region or regions of the genome, recombination rate, chromosomal location, population size and structure, and the age of the allele under study. In cultivated species, the extent of LD will also be shaped by human selection and the bottlenecks associated with crop dispersal beyond the center of origin (see Rafalski and Morgante 2004). The concept and factors that influence LD are discussed in detail in Chapters 2 and 7. Estimates of LD are important as an indicator of how useful LD-based association genetics approaches can be when compared with other available mapping methods, on the basis of the trade-off between population size and informativeness (Rafalski and Morgante 2004). In a situation where LD is large, genome wide scans may be possible albeit with poor resolution. Conversely, if there is a rapid decline of LD, examination of candidate genes may be a more viable option for association studies, as genome wide scans will require excessively large numbers of markers – the cost of which will be too prohibitive for many applied breeding programs. For example, the rapid decay of LD at 100 kb around the *Xa5* locus in rice (*Oryza sativa* L.), would require an average of one marker per centiMorgan (where 1 cM = 200–300 kb) (Garris *et al.* 2003), suggesting that candidate gene-based LD mapping could provide greater resolution than conventional QTL mapping (as a result of more recombination events). This is even more apparent in plants such as conifers and onions which have many megabase pairs per centiMorgan, but relatively rapid decay of disequilibrium over short distances (Chapter 10). It is important therefore to gain an understanding of the patterns of LD in different regions of the genome and in different populations in one organism, to make an informed choice of a methodology for association genetics studies. According to Rafalski and Morgante (2004), one of the first tasks to undertake will be to identify populations with different amounts of LD, including high LD populations for high resolution mapping. In most cases, existing germplasm collections can be exploited for this purpose. The plant research community can take advantage of being able to create these by crossing populations with the required amount of LD and diversity (Rafalski and Morgante 2004). Because of limited genomic resources in most crops, the candidate gene approach may continue to be used widely in plant association genetics studies irrespective of the extent of LD (Buckler and Thornsberry 2002). The number of markers needed to scan the whole genome remains to be determined, but it will differ between populations as well as between species. Suggestions on the possible number of markers and sample size for association studies are provided in Chapter 8.

Although, LD-based association genetics methods hold promise for speeding up the fine mapping and identification of genes responsible for variation in agronomic traits, the traditional linkage mapping methods will continue to be useful, particularly when trying to "mendelise" QTLs and assessing the effects of a QTL in isolation (Paran and Zamir

2003). According to Rafalski and Morgante (2004), plant biologists could potentially create experimental populations of unlimited size for the purpose of high resolution genetic mapping. Also, association mapping methods could be adapted to utilize pre-existing populations. Such populations should offer improved mapping resolution because of the many opportunities for recombination that will have been realized over many generations. In addition, the many years of testing that have been carried out on some breeding lines could result in a more accurate assessment of phenotypic traits that are difficult to score (Rafalski and Morgante 2004). In general, association genetics approaches may be more suited to organisms with little or no pedigree information; large effective population sizes resulting in less differentiation in trait values and little or no structure in the population; populations with rich allelic diversity, moderate to high nucleotide diversity; and traits with little or no selection history and controlled by many loci with small effects, and low frequency older alleles. On the other hand, linkage based fine mapping methods may be more efficient for marker assisted breeding in inbred crops than in some out-breeding perennial species. Also, a functional understanding of QTLs will require positional cloning and complementation tests and this will be more feasible in organisms with small genomes, mutants with well-defined effects, efficient transformation systems and near complete genomic sequence. See Neale and Savolainen (2004) for a detailed discussion on these factors and how they will impinge on the choice of a mapping strategy. A comparison of the relative power and cost of association genetics and conventional QTL mapping approaches is presented in Chapter 8.

A number of factors have been identified that can potentially limit the application of association mapping via the LD approach. These are population structure, i.e., the presence of subgroups with an unequal distribution of alleles within a population, population stratification resulting from the complex breeding history of many agronomically important crops and the limited gene flow in most wild plants (Sharbel *et al*. 2000), pleiotropic and epistatic interactions, genotype × environment interactions, poor experimental design, weak statistical tests/inferences, small sample size, complexity of the trait under study, as well as the quality of the phenotypic data. Many of these factors individually and in combination have been suggested to lead to spurious associations in association genetics studies. Methods to control these factors are discussed in Chapter 8.

1.4 CONCLUSION

Over the past decades, we have seen the successful use of map-based strategies including linkage analysis, QTL mapping, and positional cloning for the dissection of the mechanism of trait inheritance. These approaches have facilitated the identification of major genes and QTL in human, plant, and animal species particularly in model organisms. However, efficient gene discovery with these approaches will probably continue to be largely limited to loci that have a large effect on quantitative variation. Techniques are also needed to more rapidly identify genes that play a modest role in regulating quantitative trait variation. Current procedures are time consuming and it can take several years to develop populations for fine scale mapping. Apart from inherently poor resolution resulting from limited meiotic crossover events in pedigreed populations, developing large full sib families for each major gene may be impractical from a plant management view point, particularly in tree crops (Chapters 10 and 11). A more efficient

approach that may not need the generation of large pedigreed mapping populations with higher resolution is therefore needed to complement conventional QTL mapping strategies. Currently, a population genomics tool termed *association mapping* seems to be the method of choice and has been used more extensively in human genetics studies than in any other species.

To design appropriate association genetics studies we need to understand the structure of LD within and among populations as well as in different regions of the genome in an organism. Depending on the extent of LD, a candidate gene approach or genome-wide association study may be carried out. For most plant species, at least in the near future, pre-existing mapping populations and germplasm collections may be used as a starting point because of limited genomic resources and increased precision in phenotypic assessments resulting from repeated measurements in such populations. Also, there are situations (depending on species, populations, domestication/selection history, etc.) under which conventional QTL mapping methods may work better than association genetics methods, and vice versa. Note also that association studies could result in spurious associations if factors such as population structure, experimental design, statistical tests, and inferences are not adequately addressed.

1.5 REFERENCES

Alpert, K.B., Tanksley, S.D., 1996, High-resolution mapping and isolation of a yeast artificial chromosome contig containing *fw2.2*: a major fruit weight quantitative trait locus in tomato. Proc Natl Acad Sci USA 93: 15503–15507.

Botstein, D., Risch, N., 2003, Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nature Genetics 33: 228–237.

Buckler, E.S., Thornsberry, J.M., 2002, Plant molecular diversity and applications to genomics. Curr Opin Plant Biol 5: 107–111.

Chakraborty, R., Weiss, K.M., 1988, Admixture as a tool for finding genes and detecting that difference from allelic association between loci. Proc Natl Acad Sci 85: 9119–9123.

Darvasi, A., Soller, M., 1995, Advanced intercross lines, an experimental population for fine genetic mapping. Genetics 141: 1199–1207.

El-Assal, S., Alonso-Blanco, C., Peeters, A., Raz, V., Koornneef, M., 2001, A QTL for flowering time in *Arabidopsis* reveals a novel allele of CRY2. Nat Genet. 29: 435–440.

Garris, A.J., McCouch, S.R., Kresovich, S., 2003, Population structure and its effect on haplotype diversity and linkage disequilibrium surrounding the *xa5* locus of rice (*Oryza sativa* L.). Genetics 165: 759–769.

Jannink, J.-L., Walsh, B., 2002, Association mapping in plant populations. In: Kang M.S. (ed.). Quantitative Genetics, Genomics and Plant Breeding, CABI, Wallingford, UK.

Johanson, U., West, J., Lister, C., Michaels, S., Amasino, R., Dean, C., 2000, Molecular analysis of *FRIGIDA*, a major determinant of natural variation in *Arabidopsis* flowering time. Science 290: 344–347.

Kruglyak, L., 1999, Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat Genet 22: 139–144.

Neale, D.B., Savolainen, O., 2004, Association genetics of complex traits in conifers. Trends Plant Sci 9(7): 325–330.

Paran, I., Zamir, D., 2003, Quantitative traits in plants: beyond the QTL. Trends Genet 19: 303–306.

Rafalski, A., Morgante, M., 2004, Corn and humans: recombination and linkage disequilibrium in two genomes of similar size. Trends Genet 20(2): 103–111.

Remington, D.L., Thornsberry, J.M., Matsuoka, Y., Wilson, L.M., Whitt, S.R., Doebley, J., Kresovich, S., Goodman, M.M., Buckler, E.S., 2001, Structure of linkage disequilibrium and phenotypic associations in maize genome. PNAS 98(20): 11479–11484.

Risch, N.J., 2000, Searching for genetic determinants in the new millennium. Nature 405: 847–856.
Sharbel, T.F., Haubold, B., Mitchell-Olds, T., 2000, Genetic isolation by distance in *Arabidopsis thaliana*: Biogeography and postglacial colonization of Europe. Molecular Ecology 9:2109-2118.
Stuber, C.W., Polacco, M., Senior, M.L., 1999, Synergy of empirical breeding, marker-assisted selection, and genomics to increase crop yield potential. Crop Sci 39: 1571–1583.
The International HapMap Consortium, 2005, A haplotype map of the human genome. Nature 437: 1299–1320.
Yano, M., 2001, Genetic and molecular dissection of naturally occurring variation. Curr Opin Plant Biol 4: 130–135.

Chapter 2
LINKAGE DISEQUILIBRIUM

Nnadozie C. Oraguzie[1], Phillip L. Wilcox[2], Erik H.A. Rikkerink[3], and H. Nihal de Silva[3]

2.1 INTRODUCTION

The recent surge of interest in linkage disequilibrium (LD) mapping stems in part from pioneering work in humans in which LD testing is a convenient means of examining genetic polymorphisms in different genetic backgrounds, taking advantage of generations of recombination present in such samples. LD mapping is appealing due to the potential to identify a large number of haplotypes at many genetic loci across a large collection of phenotypically well-characterized germplasm, either by DNA sequencing or by high-throughput single nucleotide polymorphism (SNP) analysis. LD mapping exploits the phenotypic and genetic variation present across a natural population and draws inferences on the basis of past recombination events that have shaped the haplotype structure of that species (Nordborg and Tavare 2002; Borevitz and Nordborg 2003). On the other hand, conventional quantitative trait locus (QTL) mapping or linkage analysis usually considers only variation among offspring of relatively few genotypes (most often between two crossed individuals) and relies solely on recombination events observed in their progeny (note also that only allelic variation present in the parental genotypes can be evaluated, limited to up to four alleles in outbred families and two in inbred families). The resolution of mapping using crosses or pedigrees depends on the amount of recombination which is determined mostly by the number of meiotic crossover events. Typical rates of recombination as estimated in humans are in the order of 10^{-8} per base pair per meiosis (Hagenblad and Nordborg 2002). This is in the same order of magnitude in *Drosophila melanogaster* (10^{-6} to 10^{-7} per map distance per meiosis) as shown by intragenic recombination at the *rosy* locus on chromosome 3 (Chovnick *et al.* 1964) – indicating that the best resolution achievable in a single generation is always going to be low. Since conventional mapping studies cannot be easily performed with very large

[1] The Horticulture and Food Research Institute of New Zealand Limited (HortResearch), Cnr Crosses and St George's Roads, P.B. 1401, Havelock North, New Zealand.
[2] Scion (New Zealand Forest Research Institute Limited), 49 Sala Street, P.B. 3020, Rotorua, New Zealand.
[3] The Horticulture and Food Research Institute of New Zealand Limited (HortResearch), Mt Albert Research Centre, 120 Mt Albert Road, P.B. 92169, Auckland, New Zealand.

numbers of individuals or very many generations, their resolution is generally poor. They provide a good way of localizing genes to individual chromosomes, or if sample size is adequate, specific genomic regions, but typically do not provide sufficient resolution to locate the gene or functional polymorphism. They are also inefficient at finding alleles at low frequencies in the population. In contrast, LD mapping takes advantage of historical recombination in the ancestry of a lineage and may be more efficient for detecting contributions of rare alleles, and for localizing the genes of interest.

In order to use association genetics most effectively, we need to understand the structure of LD in a genome. In the presence of significant LD, it may be possible to identify genetic regions that are associated with a trait of interest by a systematic scan of individuals from an existing population using polymorphisms from either well chosen genomic regions, or full genome scans where affordable. LD mapping plays a fundamental role in human gene mapping and has been used extensively to dissect complex diseases including Alzheimer (Corder *et al*. 1994) and cystic fibrosis (Kerem *et al.* 1989). However, many of the initial associations detected have not been consistently replicated and may well have been spurious, particularly because the tests could not take sufficient account of the effect of population structural problems such as admixture (see below). Nonrepeatable results could also be due to inadequate experimental design (Altshuler *et al*. 2000; Ball 2005). LD mapping is of further interest as it may shed light on the origins and evolutionary history of an organism since the distribution of LD is determined, in part, by population history (Tishkoff *et al*. 1996). Moreover, knowledge of the level of disequilibrium in a population may enable us to learn more about the biology of recombination in that species (Pritchard and Przeworski 2001). Potentially, it could also provide information on intraspecific lineages carrying genetic factors (for example, insertions or inversions that generate large scale differences between chromosomes and presumably reduce crossovers) capable of modulating rates of recombination, allowing subsequent characterization.

There is still a lot to learn about genomic patterns of LD in plants. In addition, knowledge of LD at the chromosomal level is relatively small. LD mapping in plants will be useful to identify allelic variants that potentially relate to a trait(s) of interest to complement QTL mapping and for general application of molecular markers to germplasm characterization. In this review, we will provide some background information on the theory of LD, measures of LD in a population and factors that influence LD. We will conclude the discussion with some empirical examples of LD testing in model organisms including humans, *Drosophila melanogaster* and plants, particularly the two most advanced model plant systems with respect to LD studies, maize and *Arabidopsis thaliana.*

2.2 THE CONCEPT AND DEFINITION OF LD

Linkage equilibrium (LE) and LD are population genetics terms used to describe the likelihood of co-occurrence of alleles at different loci in a population. Generally, linkage refers to the correlated inheritance of loci through physical connection on a chromosome. While LE refers to random association of alleles at different loci (that is, the chance of finding one allele at one locus that is independent of an allele at another locus), LD refers to nonrandom association of alleles at different loci. That is, when a particular allele at

one locus is found together with a specific allele at a second locus more often than expected if alleles at the loci were combining independently in a population, the loci are said to be in LD (see Figure 2.1). LD does not automatically imply linkage. Tight linkage may result in high levels of LD and it is in this sense that LD is also a powerful mapping tool. This occurs when the correlation of allelic states of loci in different parts of the genome is caused by the physical proximity of the loci. LD may also be influenced by other factors which will be discussed in the later sections of this chapter.

Although the concept of LD dates to the early part of the 20th century (Jennings 1917), the first commonly used LD measure, *D*, was developed about four decades ago (Lewontin 1964). The digenic *D* (in common with most other measures of LD), quantifies disequilibrium, as the difference between the observed frequency of co-occurrence of an allele of locus *A* with an allele of another locus *B*, and the expected frequency of co-occurrence under LE (i.e., if the two alleles were combining at random). For two loci *A/a* and *B/b*, let the frequency of the observed haplotype with alleles *A* and *B* be P_{AB}. Assuming independence, the expected haplotype frequency is the product of the corresponding two allele frequencies, i.e., $p_A \times p_B$. Therefore, $D = P_{AB} - p_A p_B$. If *D* differs significantly from zero, LD is said to exist. We will provide a full discussion of different measures of LD in Section 2.3. If loci, *A* and *B* are both biallelic, four different haplotypes are possible. Under LD, some of these two-locus haplotype frequencies will be over-represented and others under-represented. Figure 2.1 illustrates two scenarios where DNA sequences of haplotypes are in complete LD or LE (i.e., no LD).

A)

1 2

AAGCTGTCACTG…/intervening DNA sequence/…TCATCGTACTCA
AGGCTGTCACTG…/intervening DNA sequence/…TCATCGTACTCA

A	…/intervening DNA sequence/…	C
A	…/intervening DNA sequence/…	C
G	…/intervening DNA sequence/…	T
A	…/intervening DNA sequence/…	C
G	…/intervening DNA sequence/…	T
G	…/intervening DNA sequence/…	T
G	…/intervening DNA sequence/…	T
A	…/intervening DNA sequence/…	C
A	…/intervening DNA sequence/…	C
A	…/intervening DNA sequence/…	C
G	…/intervening DNA sequence/…	T
G	…/intervening DNA sequence/…	T

		Site 1	
		A	G
Site 2	C	6	0
	T	0	6

Complete LD

B)

1 2

AAGCTGTCACTG…/intervening DNA sequence/…TCATCGTACTCA
AGGCTGTCACTG…/intervening DNA sequence/…TCATCGTACTCA

A	…/intervening DNA sequence/…	C
G	…/intervening DNA sequence/…	C
G	…/intervening DNA sequence/…	C
A	…/intervening DNA sequence/…	C
G	…/intervening DNA sequence/…	T
G	…/intervening DNA sequence/…	C
A	…/intervening DNA sequence/…	T
A	…/intervening DNA sequence/…	T
A	…/intervening DNA sequence/…	C
A	…/intervening DNA sequence/…	T
G	…/intervening DNA sequence/…	T
G	…/intervening DNA sequence/…	T

		Site 1	
		A	G
Site 2	C	3	3
	T	3	3

No LD

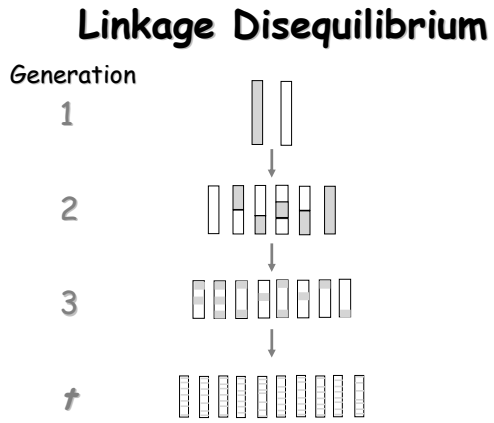
Figure 2.1. Hypothetical scenarios of LD between linked polymorphisms caused by different mutational and recombinational histories. The starting population has only two haplotypes; AG at locus 1 and TT at locus 2. Mutation later occurs at locus 2 with "T" being replaced by "C" in some cases. (A) shows maintenance of LD due to lack of recombination between loci 1 and 2 in generations following mutation, and (B) is a situation where LE is attained due to recombination breaking down the initial disequilibrium. The corresponding contingency table shows the haplotype counts. Absolute LD exists when two loci share a similar mutational history with no recombination, LE is attained when there is recombination between loci regardless of mutational history. The influence of recombination and mutational history on LD will be discussed in subsequent paragraphs.

LD is commonly found in natural populations between loci for which recombination has not had sufficient time to dissipate the initial disequilibrium. When a population in LD mates at random, the amount of disequilibrium is progressively reduced with each succeeding generation. The degree of LD between two loci is therefore dependent on both the recombination fraction, θ, and time in generations, t, since the origin of a new mutation at time = 0. Theoretically, LD decays with time and recombination distance according to the following formula (Falconer and Mackay 1996):

$$D_t = (1-\theta)^t D_0, \tag{2.1}$$

where, θ is recombination fraction while D_0 and D_t represent LD in time at generations 0 and t, respectively. Thus, LD will tend to be smaller when two loci are located further apart and D will decrease through time as a result of recombination (see Figure 2.2).

A



B

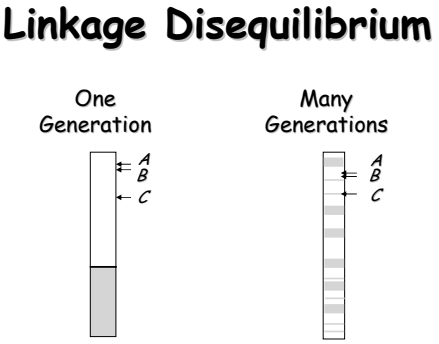


Figure 2.2. Hypothetical diagrams showing decay of LD after many generations following recombination. (A) shows complete LD in generation one to almost complete dissipation of LD in generation *t*, (B) After a **few** generations, alleles of moderately distant genes still cosegregate (e.g., A & C), but after **many** generations only alleles of very close genes cosegregate (e.g., A & B).

Figure 2.3 shows an example of how *D* is reduced with time (in this case after 100 generations) under different degrees of recombination fraction, θ. For example, if $\theta = 0.10$ (10% recombination), it will take 6.5 generations for *D* to be cut in half and 28.4 generations for *D* to drop by 95%, according to the equation obtained by rearranging (2.1):

$$t = \frac{In\left(D_t / D_0\right)}{In\left(1-\theta\right)}.$$

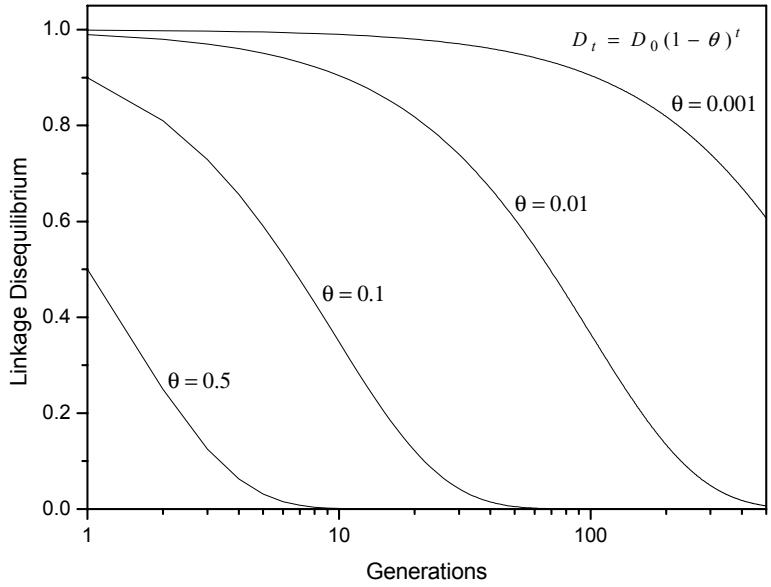


Figure 2.3. Decay of LD with generation time.

For DNA sequence variants, assuming recombination rates of approximately 10^{-8}/bp (equivalent to about 1 cM/Mb) i.e., $\theta = 1 \times 10^{-5}$/kb, for sites 100 kb apart it will take 693 generations for *D* to be cut in half, while for sites 1 kb apart, it will take 69,315 generations for *D* to be cut in half (Nachman MW, personal communication).

2.3 MEASURES OF LD

A variety of measures of LD have been developed, and there are some good reviews that compare these different measures (Hendrick 1987; Devlin and Risch 1995; Jorde 1995). A good general account of LD is given by Weir (1996). The basis for many LD measures is the deviation of observed haplotype frequency from their expectation assuming independence. Consider two biallelic loci A and B. Let p_1 and $p_2 (=1-p_1)$ be the frequencies of A_1 and A_2 alleles at locus A. Likewise, q_1 and $q_2 (=1-q_1)$ are the frequencies of B_1 and B_2 alleles at the second locus. These two bi-allelic loci will then produce four different haplotypes, A_1B_1, A_1B_2, A_2B_1 and A_2B_2. These haplotypes can be represented in a 2×2 contingency table as in Table 2.1. We use the notation *P* (uppercase), with a subscript indicating the two alleles it carries, to represent the haplotype frequency. Lower case letters are used to represent allele frequencies. Note that haplotype frequencies as given in Table 2.1 are the unconditional probabilities. Sometimes we are interested in conditional probabilities, e.g., the probability of a haplotype having A_1 allele given that allele B_1 is present. This can be calculated using the haplotype and marginal allele frequencies as P_{11}/q_1. Similarly, the probability of having A_2 allele in the presence of allele B_1 is P_{21}/q_1.

Table 2.1. Notation for haplotype and allele frequencies

Allele	Allele		
	B_1	B_2	Total
A_1	P_{11}	P_{12}	p_1
A_2	P_{21}	P_{22}	p_2
Total	q_1	q_2	1

As pointed out, LD is commonly measured as deviations of haplotype frequencies from their expectations, given the alleles at the two loci are independent. A common measure of LD, therefore, would be

$$\begin{aligned} D &= \Pr(A_1, B_1) - \Pr(A_1)\Pr(B_1) \\ &= P_{11} - p_1 q_1 = P_{22} - p_2 q_2 \\ &= -P_{12} + p_1 q_2 = -P_{21} + p_2 q_1 \\ &= P_{11}P_{22} - P_{12}P_{21}. \end{aligned} \tag{2.2}$$

In the first three lines, the first term of each expression is the observed haplotype frequency and the second term the expected frequency under independence. Note that all expressions in equation (2.2) follow from the first because of inequalities in Table 2.1; i.e., $p_1 = P_{11} + P_{12}$; $q_1 = P_{11} + P_{21}$ etc. For instance,

$$\begin{aligned} D &= P_{11} - p_1 q_1 \\ &= P_{11} - (1 - p_2)(1 - q_2) \\ &= P_{11} - (1 - p_2 - q_2 + p_2 q_2) \\ &= P_{11} - (P_{11} + P_{12} - P_{12} - P_{22} + p_2 q_2) \\ &= P_{22} - p_2 q_2 \end{aligned}$$

Likewise,

$$\begin{aligned} D &= P_{11} - p_1 q_1 \\ &= P_{11} - (P_{11} + P_{12})(P_{11} + P_{21}) \\ &= P_{11} - (P_{11}P_{11} + P_{11}P_{21} + P_{11}P_{12} + P_{12}P_{21}) \\ &= P_{11}(1 - P_{11} - P_{21} - P_{12}) - P_{12}P_{21} \\ &= P_{11}P_{22} - P_{12}P_{21} \end{aligned}$$

The ordering of alleles into rows and columns of Table 2.1 is arbitrary and often D is reported without any sign as $|D|$. The measure D is dependent on allele frequencies,

so some standardized measure would be useful for comparisons across loci with different frequencies. Lewontin (1964) defined a standardized measure of D, called D' as follows:

$$D' = \begin{cases} \dfrac{D}{\min(p_1 q_2, p_2 q_1)} & D > 0 \\ \dfrac{D}{\min(p_1 q_1, p_2 q_2)} & D < 0 \,. \end{cases} \tag{2.3}$$

The denominator of the expression is the maximum absolute value of D that could be achieved for given marginal totals, which of course are the allele frequencies at the two loci (Table 2.1). The absolute value of D is scaled for the observed allele frequencies; hence the resulting value is bounded between 0 and 1. The case of $|D'|=1$ is known as complete LD. This occurs when the two loci are in complete LD or if there are less than four of the possible haplotypes as described above. Values of $|D'|<1$ indicate that the complete ancestral LD has been disrupted presumably due to recombination, resulting in all four possible haplotypes being observed.

According to Hill and Weir (1994), D is in fact more frequently used in the standardized form as:

$$r = \frac{D}{(p_1 p_2 q_1 q_2)^{1/2}} \tag{2.4}$$

or its squared value, r^2 (also described as Δ^2). The r^2 is considered as the square of the correlation coefficient between the two loci. It assumes a value of 1 if only two haplotypes are present.

D' is useful for comparisons across loci with different frequencies. For low allele frequencies, r^2 has more reliable sampling properties than $|D'|$. A key difference between r^2 and D' is that the latter is affected more by mutational histories and the former by a combination of mutation and recombination. The difference can be explained thus: consider the two loci, *A* and *B* as above, with alleles *A/a*, at locus *A* and only one allele, *B*, at the other locus. A mutation then occurs at locus *B* (to create allele *b*), but only within the "*a*" allele lineage, thus there will be disproportionately higher frequencies of the "*ab*" genotype compared to the "*Ab*" genotype if the mutant haplotype is favored. In such a case, D' is higher than r^2. Simple diagrammatic explanations of this phenomenon are presented in Gaut and Long (2003) and Flint-Garcia *et al*. (2003). Observed differences between D' and r^2 can therefore reflect events such as recent admixture of two or more populations, or emergence from a recent population bottleneck, where new mutations have arisen but not had sufficient time to fully recombine.

Estimates of D' can be strongly inflated in small samples, although both measures are subject to sampling error. Therefore, statistically significant values of D' that are

near unity provide a useful indication of minimal historical recombination, but intermediate values should not be used for comparisons of the strength of LD between studies, or to measure the extent of LD – at least not without also assessing r^2.

The measure r^2 is in some ways complementary to D'. r^2 is equal to D^2 divided by the product of the allele frequencies at the two loci (Equation (2.4)). Unless the two loci have identical allele frequencies, a r^2 value of 1 is often not possible. So what measure is most appropriate for association genetics? In general, the aim of association genetics is to infer the presence of something not directly observed or known (the quantitative trait nucleotide (or QTN)) via correlation of a phenotypic effect of the QTN with something that is observed (DNA polymorphisms). Thus r^2 is a more informative measure, as it measures the overall departure from complete independence between pairwise combinations of polymorphisms. D' may also have some utility, although this may be restricted to fine mapping once initial associations have been detected. As an initial indicator of LD r^2 is the preferred initial proxy for assessing extent of LD for association genetics applications.

Apart from D' and r^2, there are several other measures of LD, which originated from epidemiology. The "population attributable risk" used in epidemiology has been re-derived and used as a measure of LD, and represented as δ (Levin and Bertell 1978):

$$\delta = \frac{D}{q_1 P_{22}}. \tag{2.5}$$

This is the same measure that Terwilliger (1995) referred to as λ. Yet another epidemiological measure recommended for LD by Kaplan and Weir (1992) is based on the difference in conditional frequencies:

$$d = \frac{P_{11}}{q_1} - \frac{P_{12}}{q_2} = \frac{D}{q_1 q_2}. \tag{2.6}$$

The final LD measure that Devlin and Risch (1995) included in their list of commonly used ones also originated from epidemiology. It is based on the odds ratio (Equation (2.7)). The odds-ratio for an event is the probability of that event happening divided by the probability of it not happening, i.e., if P is the probability of event, then the odds-ratio is $P/(1-P)$. The first line of Equation (2.7) is then read as odds for B_1 allele to be present in the haplotype given that A_1 is present, divided by odds for B_1 allele to be present in the haplotype in the presence of A_2.

$$Odds\ Ratio\ (OR) = \frac{P_{11}P_{22}}{P_{12}P_{21}}$$

$$Q = \frac{OR-1}{OR+1} = \frac{P_{11}P_{22} - P_{12}P_{21}}{P_{11}P_{22} + P_{12}P_{21}} = \frac{D}{P_{11}P_{22} + P_{12}P_{21}}. \quad (2.7)$$

Unlike the odds ratio, which ranges from zero to infinity, Q is bounded between $(-1,+1)$.

As Devlin and Risch (1995) pointed out these five measures of LD differ only in their denominator, which only serves to standardize D. One might expect then that all these measures provide the same information for simple disequilibrium mapping. Devlin and Risch (1995) illustrated with simple examples that this wasn't the case. We advise readers to refer to their paper for a full account of the comparison of the different LD measures.

To measure LD experimentally, we take a random sample from the population of interest. LD calculated on the sample is an estimate of the corresponding parameter. Assuming that all haplotypes are observed, we make counts of the different haplotypes. Let the sample contain n total number of diploid individuals. The composition of different haplotypes for two biallelic loci can be represented in a 2×2 contingency table as follows:

Table 2.2. Sample counts of haplotypes of two loci in a 2×2 contingency table

Allele	B_1	B_2	Total
A_1	a	b	$a+b$
A_2	c	d	$c+d$
Total	$a+c$	$b+d$	$2n=a+b+c+d$

The marginal row and column totals are allelic counts at A and B loci, respectively. Cell counts can vary independently with the only constraint being they add up to the total count $2n$. Dividing these cell and marginal counts by the total gives estimates of corresponding haplotypes and allele frequencies as shown in Table 2.1. For data such as in Table 2.2, the hypothesis of association between alleles at the two loci (i.e., LD) can be tested by either Chi-square or Fisher's exact test. These tests are discussed in detail in Chapter 7 (Section 7.6.1). In concluding this section, we will show how different measures of LD are calculated using some hypothetical data as in Example 2.1.

Example 2.1. Calculation of LD

The following is a hypothetical example showing frequencies of haplotypes at two SNP sites in four different populations. The SNPs are due to single nucleotide difference at the marker locus, hence are biallelic. Note that populations (a) and (b) contain only two haplotypes, whereas (d) and (c) have three and all four possible haplotypes, respectively. We want to estimate LD based on the different measures.

(a)
50% —T——A—
50% —C——G—

(b)
80% —T——A—
20% —C——G—

(c)
25% —T——A—
25% —T——G—
25% —C——G—
25% —C——A—

(d)
10% —T——A—
20% —T——G—
70% —C——G—

The haplotype frequencies can be tabulated in the following table for population (d).

	Allele		
Allele	A	G	Total
T	0.1	0.2	0.3
C	0	0.7	0.7
Total	0.1	0.9	

Using the equations described in the text the different measures of LD are calculated for the four populations, and these values are tabulated below:

	LD measure					
Population	$\|D\|$	$\|D'\|$	r	δ	d	Q
a	0.25	1	1	1	1	1
b	0.16	1	1	1	1	1
c	0	0	0	0	0	0
d	0.07	1	0.51	1	0.78	1

It is noted that markedly different values are obtained with the different measures, keeping in mind the range for each measure. $|D|=1$ if only two or three haplotypes are present. A close look at populations (a) and (b) shows that $|D|$ is dependent on the range of allele frequencies.

2.3.1 Haplotype Blocks

When haplotypes are known, it is easy to estimate LD from sample data (Example 2.1). In practice, however, when unrelated individuals are sampled it is not possible to determine the phase of the double heterozygote, $A_1A_2B_1B_2$ of two marker loci. The double heterozygote can produce two different pairs of haplotypes depending on the phase configuration; i.e., A_1B_1 / A_2B_2 or A_1B_2 / A_2B_1. The genotypes with unknown phase need to be determined to infer on the haplotypes. In the case of two biallelic markers, maximum likelihood estimates (MLE) of haplotype frequencies can be obtained analytically by solving a cubic equation (SAS 2004). For multiple loci or markers with more than two alleles an iterative process could be used.

One of the earliest methods used for haplotyping was Clark's algorithm (Clark 1990). Today more efficient alternatives are available. One such alternative is the Expectation Maximization (EM) algorithm of Excoffier and Slatkin (1995). It is a combination of two algorithms: an EM statistical algorithm for handling missing data, and a counting algorithm for frequencies. As the process is iterative, it starts by guessing haplotype frequencies. It then uses the current allele estimates to replace the ambiguous phased genotypes. Given the phase configurations of unphased genotypes it then goes to estimate the frequency of each haplotype by counting. The process is repeated until the frequencies converge. As the number of markers increases, the process can be computationally very demanding. For a set of m unphased biallelic marker loci there will be 2^m possible haplotypes. Typically, the algorithm is appropriate for < 25 SNPs. This maximum likelihood method makes the assumption that the population is in Hardy–Weinberg equilibrium. Haplotype inference from genotype data is becoming more important in association studies. Intuitively, one would expect an analysis based on haplotypes to be more powerful because of simultaneous use of multiple marker infor–mation. But as discussed here haplotypes often need to be estimated based on assumptions made on the populations. This process leads to loss of some information.

The extent of LD can be highly variable across genomes in many of the species studied to date. Within a given region, LD will decrease with the distance between marker sites (Figure 2.4). Genome-wide patterns of pairwise LD values can often show regions of high LD separated by regions of low LD. This scenario is often referred to as "haplotype block." Regions that are high LD and low in recombination are also referred to as "LD hot spots" in the literature. A "hot spot" for LD also implies a "cold spot" for recombination. There are two common approaches to haplotype blocking. One method defines a block whenever LD is greater than some threshold value. The second method defines a block when a smaller number of haplotypes make up a high proportion of observed haplotypes. There is an ongoing debate regarding whether haplotype blocks truly exist as our understanding of genomic patterns of recombination and disequilibrium is still limited (Cardon and Abecasis 2003).

When a large number of markers are considered it is useful to graphically display estimated LD values. It is common to visualize LD patterns in the form of color-coded matrices (Pettersson *et al*. 2004). This way we can identify blocks within a genome area in which there is a strong LD. Graphical Overview of Linkage Disequilibrium (GOLD) is a computer program, which can graphically display LD structures (Abecasis and Cookson 2000). The software can be downloaded from http://www.sph.umich.edu/csg/abecasis/GOLD/. A sample output from GOLD is shown in Figure 2.5 (figure from GENESTAT, http://www.meb.ki.se/genestat/). Pettersson *et al*. (2004) have developed GOLDsurfer, which is an extension of the 2D view in GOLD to a 3D package. D' values are represented by colors, with hotter colors representing high D'. Note the several large red blocks on the diagonal, indicating haplotype blocks of maximal disequilibrium where there has been no recombination since the LD was formed.

2.4 FACTORS THAT INFLUENCE LD

A wide variety of mechanisms generate LD and several of these can operate simultaneously. Table 2.5, presents a summary of the genetic and demographic factors that affect LD in a population. Out of these factors, mutation and recombination may

seem to have the most evident impact on LD. Mutation provides the raw material for producing polymorphisms that will be in LD (Flint-Garcia *et al.* 2003). LD is created when a new mutation occurs on a chromosome that carries a particular allele at a nearby locus. Recombination is the main mechanism that breaks down LD. Meiotic crossing-over weakens intrachromosomal LD while independent assortment is particularly responsible for breaking down interchromosomal LD. Recombination rates are known to vary by more than an order of magnitude across the genome. Because breakdown of LD is primarily driven by recombination, the extent of LD is expected to vary in inverse relation to the local recombination rate (Nachman 2002). Recurrent mutations can also lessen the association between alleles at adjacent loci. Some SNPs, such as those at CpG dinucleotides, might have high mutation rates (due to decay of methylated cytosine – 5MeC – by deamination to thymidine over evolutionary time, leading to CpG suppression) and therefore show little or no LD with nearby markers, even in the absence of historical recombination. On the other hand in places where the levels of DNA methylation are generally lower (e.g., within genes), this effect would tend to be mitigated.

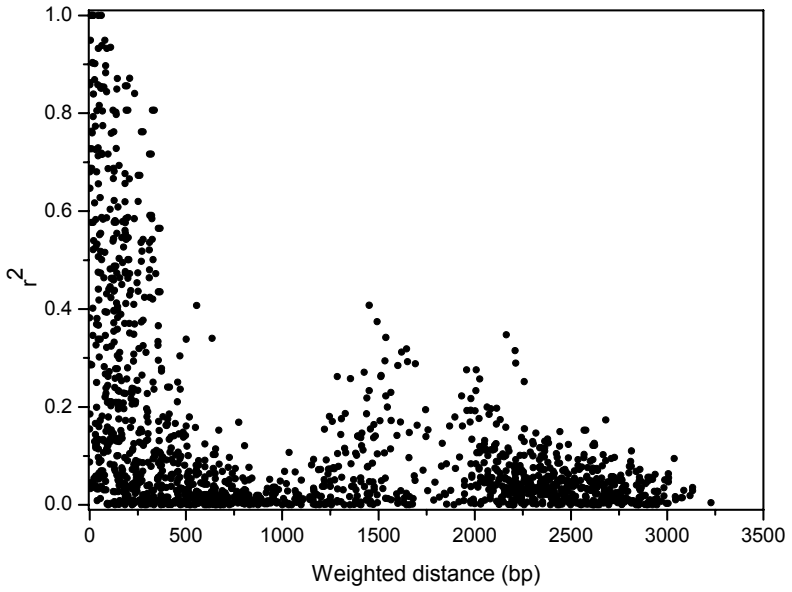


Figure 2.4. Plot of LD (in r^2) against weighted distance between polymorphic sites in the candidate gene "d3" in maize. (Data from Remington *et al*. 2001.)

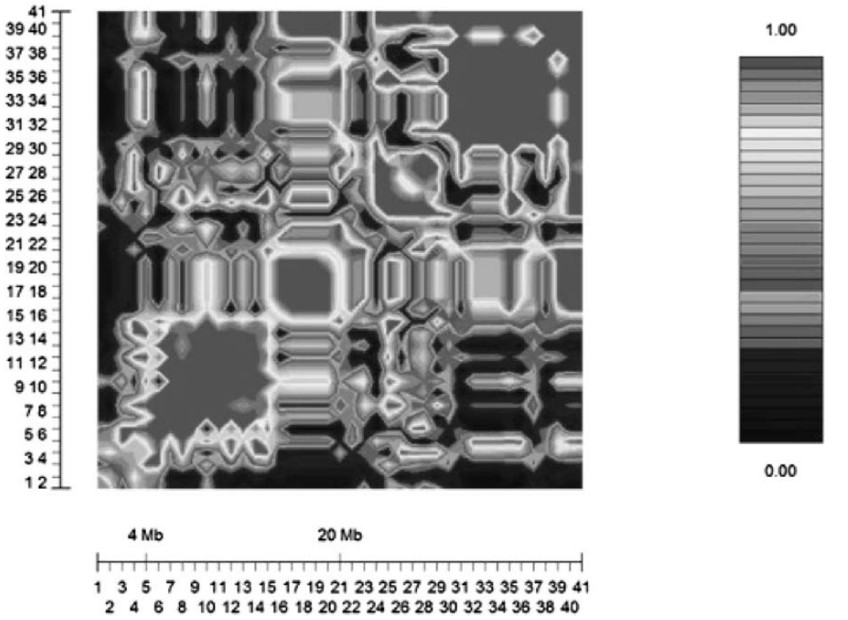


Figure 2.5. Pairwise $|D'|$ for 45 SNPs within a linked region (figure from GENESTAT, http://www.meb.ki.se/genestat/, courtesy of the Swedish National Biobanking program, Wallenberg consortium north). (see color plate)

In a gene conversion event (i.e., the nonreciprocal transfer of genetic information – note that most available evidence for gene conversion has come from fungal systems in which each of the products of meiosis can be recovered and studied individually), a short stretch of heteroduplex DNA is created during meiosis (see Figure 2.6). The subsequent correction of any mismatches resulting from this heteroduplex DNA can appear to be equivalent to two very closely spaced recombination events, and can break down LD in a manner similar to typical recombination (i.e., where there is the usual meiotic crossover) or recurrent mutation. Since gene conversion acts only on short segments of the genome its importance (when compared to crossing over which affects large chromosomal segments) would be expected to increase as genetic analysis concentrates on smaller and smaller regions. Indeed it has recently been shown that rates of gene conversion in humans are high and are important in LD between very tightly linked markers (Ardlie *et al*. 2001). Gene conversion has also played a role in the analysis and interpretation of *Drosophila* data. Langley *et al*. (2000) recorded polymorphism data at the tip of chromosome X in different populations of *D. melanogaster*. They found evidence for a surprisingly large amount of recombination, even though crossing-over rates are known to be extremely low. They inferred that this must be the result of a high local rate of gene conversion. This suggests that areas where crossing-over is suppressed may have normal rates of gene conversion. Similar phenomena could operate in plants but may be difficult to prove due to the difficulty of classical tetrad analysis. The quartet mutation in *Arabidopsis* where each of the products of meiosis develops as viable pollen grains thus representing an unordered tetrad, appears to offer a good avenue for gene conversion studies in plants.

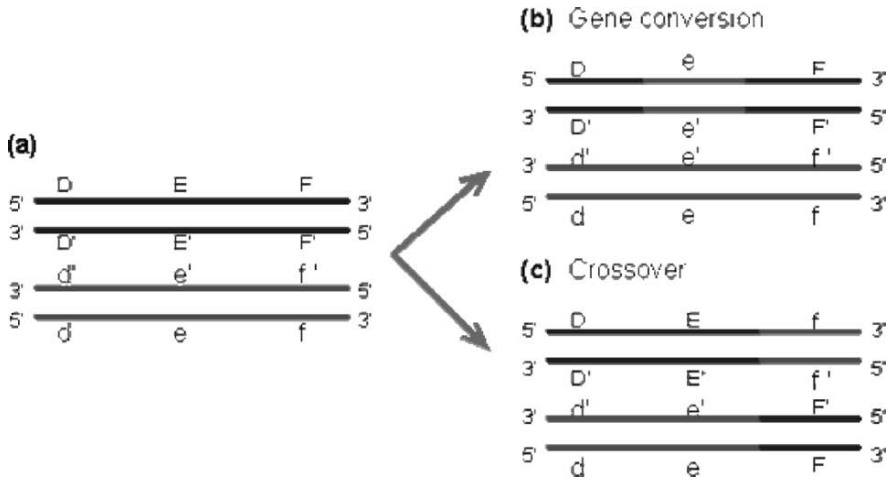


Figure 2.6. A simplistic diagram showing the major difference between gene conversion and crossover. (A) Two DNA molecules. (B) Gene conversion after mismatch correction – the red DNA donates part of its genetic information (e–e' region) to the blue DNA. (C) DNA crossover – the two DNAs exchange part of their genetic information (f–f' and F–F'). (see color plate)

Evolutionary forces such as mutation, genetic drift, migration (resulting in admixture), selection, etc. can shape LD. The extent of separation of clines and the amount of LD generated by migration depend very strongly on the allele frequencies in the initial populations, the recombination fraction between the two loci and the rate of migration.

LD between two segregating loci may be built up in a population by random drift due to random sampling of possible progeny and possible matings with the effect being subject to a combination of time and population size. Even though the expected value of the LD between alleles at two unlinked loci is zero in small populations, the variance of the LD may be large. So, although the average of D may be close to zero, the actual observed values could be quite different. Rapid population growth for example, decreases LD by reducing genetic drift. Conversely, for populations expanding from a small number of founders (i.e., bottlenecks), the haplotype present in the founders will be more frequent than expected under equilibrium. In small populations, the effects of genetic drift result in the consistent loss of rare allelic combinations. When genetic drift and recombination are in equilibrium,

$$r^2 = \frac{1}{1+4Nc}, \tag{2.8}$$

where N is the effective population size and c is the recombination fraction between sites (Weir 1996).

Selection (natural or artificial) at a locus is expected to reduce diversity and increase LD in the surrounding region (often colloquially termed “hitchhiking”). Also, selection for or against a phenotype controlled by alleles at two unlinked loci that show epistatic interaction may result in LD despite the fact that the loci are not physically linked. Strong selection for a particular allele limits genetic diversity around a locus

(producing a "signature of selection"), resulting in a short-term increase in LD around the selected gene. The *y1* locus in maize could be used as a classic example of how artificial selection can influence diversity and LD (Palaisa *et al*. 2003). The dominant allele, *Y1*, of the gene encoding for phytoene synthase (PS) is responsible for the yellow endosperm in maize, while the recessive *y1* allele is responsible for the white endosperm particularly among maize used for human consumption. The white endosperm appears to be the ancestral condition since the progenitor of maize (*teosinte*) carries white endosperm. Yellow varieties have been selected recently by breeders for their high carotenoid content, which is advantageous for animal nutrition. Sequence analysis of the PS gene in many white and yellow varieties have shown that the yellow alleles, *Y1*, are more than 20 times less diverse than the recessive white alleles, *y1*, confirming white endosperm as the ancestral state. Moreover, this depletion of diversity can be detected 500 kb from the gene itself. This long range effect is facilitated by the relative isolation of white corn germplasm from yellow germplasm. The resulting LD from the mutation for the yellow endosperm is yet to be broken down due to continued selection and lineage isolation. This example also demonstrates nicely how these types of events caused by human intervention can be utilized to map important domestication genes to limited regions of the genome.

There are two primary routes by which selection can affect the extent of disequilibrium. The first is a hitchhiking effect, in which an entire haplotype that flanks a favored variant can be rapidly swept to high frequency or even fixation (Wang *et al*. 2002; Parsch *et al*. 2001). Although the effect is generally milder, selection against deleterious variants can also inflate LD, as the deleterious haplotypes are swept from the population. Genetic hitchhiking is expected to affect the frequency of distribution of variants at segregating sites such that derived variants will be in higher frequency than expected under a neutral equilibrium model. Genetic hitchhiking is also expected to skew the frequency distribution of variants at the segregating site toward rare alleles, resulting in a significantly negative value of measures such as Tajima's D (a statistic used to examine the presence of selection, see Tajima 1989 for details). It is unknown to what extent this mode of selection increases pairwise LD between high frequency alleles. The second way in which selection can affect LD is through epistatic selection for combinations of alleles at two or more loci on the same chromosome (Cannon 1963), or even different chromosomes. Co-adapted gene complexes are an example of this. This form of selection leads to an association of particular alleles at different loci. This has provided a major motivation for historical studies of LD in *Drosophila* genetics as a means of detecting the action of natural selection. However, epistatic selection would have to be very strong to maintain allelic associations at the scale of megabases, in the face of substantial recombination. One obvious example of this would be the generation of linked "super-gene" complexes under conditions of disruptive selection, such as those controlling floral morphology – pin and thrum types – in *Primula* spp. Of course, the stability of such complexes (and consequent extensive LD) is generally consolidated by chromosomal inversions or other mechanisms to minimize recombination. In this context, the Y chromosome of mammals could be seen as a very good example of extensive LD arising from suppressed recombination possibly promoted by ancient epistasis.

Various aspects of population structure are thought to influence LD. LD arises in structured populations when allelic frequencies differ at two loci across subpopulations, irrespective of the linkage status of the loci. Admixed populations formed by the union of previously separate populations into a single panmictic one, can be considered as a case of

structured population where substructuring has recently ceased. Admixture results in the introduction of chromosomes of different ancestry and allele frequencies. Often the resulting LD extends to unlinked sites, on the same and different chromosomes but breaks down rapidly with random mating (Pritchard and Przeworski 2001). With regard to LD detection studies, population admixture is one of the major factors that causes spurious associations between marker alleles and the phenotype. Although our interest in LD is because it is likely to be caused by tightly linked loci, spurious associations due to population admixture can often lead to incorrect conclusions. Population admixture can generate LD even though the individual populations forming the mixture do not show any such disequilibrium. We will illustrate this point using a hypothetical example. Let us consider a locus with a disease allele, D and a second unlinked marker locus with allele, M. Take two populations (I and II) of equal size, but with different frequencies of alleles D and M, as shown in Table 2.4. Since we assume the loci to be independent the expected frequency of individuals carrying both D and M alleles in the population is the product of their individual frequencies (Table 2.4). For our hypothetical example we take these to be the observed frequencies. Consider now the admixture of these two populations in equal proportions. The new observed allele frequencies of the mixed population would simply be the average of the two as shown in the table. The observed frequency of individuals with D and M alleles in the mixed population is greater than what would be expected for independent loci, i.e., the mixing has resulted in a spurious association between D and M alleles. The use of a population, which is likely to have resulted from admixture, is therefore not recommended for association studies without proper genomic control (Table 2.3).

Table 2.3. A hypothetical scenario showing how population admixture can lead to spurious associations

	Frequency		
Population	I	II	Admixture
M -Disease allele	0.7	0.1	0.4
M -allele	0.8	0.2	0.5
$D \& M$	0.56	0.02	0.29

An example based on empirical data of how admixture and selection can influence LD resulting in significant association between genotype and phenotype is the oat study carried out by Beer *et al*. (1997). In this study, the authors used 64 North American oat varieties and landraces that have been phenotyped for 13 quantitative characters and grouped based on their restriction fragment length polymorphism (RFLP) genotype at 48 loci. They found significant associations between RFLP fragments and the group means for 11.2% of the fragments at 1% significance level. The authors, however, did not take into consideration in their data analysis the fact that both spring and winter varieties were represented in the germplasm pool they used for the study (Souza and Sorrells 1991). These groups differed in both phenotype means and marker frequencies. Also, the germplasm had undergone four decades of selection and improvement with some genotypes older than others. Hence, the germplasm may be considered as an admixture of old and modern subpopulations, with one having undergone less selection than the other. In which case, one would expect to find fewer associations between the marker alleles within each subpopulation than in the combined pool. This indeed was the case when the data was re-analyzed, with only 6% and 4.9% of allele-trait associations significant in the

subpopulations of old and modern varieties, respectively. A similar example from humans of this population stratification effect can be seen in Knowler *et al*. (1988). The authors examined candidate haplotypes for Type-2 diabetes in members of the Pima and Tohono O'odham Native American tribes of southern Arizona. Individuals with one particular haplotype had only an 8% rate of diabetes, while those lacking this haplotype had a 30% rate of diabetes. However, this particular haplotype is much more common in Caucasian populations than in full-heritage Native American populations. When correcting for this population difference by only considering individuals of full-heritage, 59% of individuals with the haplotype had diabetes, while 60% of the individuals lacking the haplotype had diabetes.

Population mating pattern can also have a strong influence on LD. Humans and animals are out-crossers while plants can either be autogamous (inbreeding) or allogamous (out-breeding). Generally, there is more rapid decline in LD in out-crossing species compared to selfing species (Nordborg 2000). For example, in predominantly selfing species such as *Arabidopsis thaliana* or soybean, LD persists over tens to hundreds of kilobases, whereas in out-crossing species such as human, maize or conifers (Dvornyk *et al*. 2002) a much more rapid decline has been observed. In maize, at least in some populations, LD declines over a few hundred base pairs (Rafalski and Morgante 2004). Similarly, in conifers, LD declines rapidly over 2–4 kb (Brown *et al*. 2004; Krutovsky and Neale 2005) and this is also likely to be the case in some angiosperm forest trees (e.g., Thumma *et al*. 2005). Selfing species may show increased recombination rates per meiosis: for example, the recombination rate per base pair is estimated to be approximately twofold and sixfold higher in selfing *Arabidopsis* than in Drosophila and maize, respectively. However, selfing increases homozygosity, thereby limiting the number of double heterozygotes that can be shuffled by recombination. Nordborg and Donnelly (1997) used the equation: $c = (1 - s)/(2 - s)$, to describe the relationship between effective recombination rate and the degree of selfing, where c is the recombination rate and s is the selfing fraction. Following this equation, the effective recombination rate for a species that undergoes selfing half of the time will be reduced to 1/3 that of an obligate out-crossing species. Hence, the effective rate of recombination is low in selfing species; genetic polymorphisms tend to remain correlated, and LD is expected to be maintained over long physical distances. However, although selfing increases homozygosity, genetic diversity can be quite high at the population level as has been reported in the case of *Arabidopsis* (Nordborg *et al*. 2002). At the other extreme, where there is presumed lack of recombination as has been demonstrated in the nonpseudoautosomal region of the mammalian Y chromosome and presumably occurs in the nonrecombining region of similar plant dimorphic sex chromosome systems, complete LD would be expected.

Table 2.4. Factors that affect LD in a population

Factor	Effect
Recombination rate	Higher recombination lowers LD
Mating systems: selfing species	High LD
Mating systems: out-crossing species	Low LD
Genetic isolation between lineages	Increases LD
Population subdivision	Increases LD
Population admixture	Increases LD
Natural and artificial selection	Locally increases LD
Population size	Small populations have more LD
Balancing selection	Increases LD
Mutation rate	High mutation rate decreases overall LD but LD around newly created mutated allele remains high until dissipated by recombination
Genomic rearrangements	Rearrangements suppress local recombination
Stochastic effects (chance)	Increase or decrease LD
Epistatic interactions with significant phenotypic effects	Increase LD

Modified from Rafalski and Morgante (2004).

2.5 EMPIRICAL EXAMPLES OF LD MAPPING IN VARIOUS ORGANISMS

Due to the different genetic and environmental factors that affect LD, the extent and pattern of LD are expected to vary within and between species and even between different regions of the genome of the same species. Below, we discuss the species from which we currently derive most of our information about LD.

2.5.1 LD in Humans

In humans the initial interest was in accurate characterization of LD prompted in part by the question of what marker density will be needed to attain reasonable power in genome-wide association studies (Kruglyak 1999; Gabriel *et al*. 2002; Phillips *et al*. 2003). Tremendous variation in the amount of LD within the species among different regions of the genome, both on a large and on a small scale has been reported. Much of this variation is deduced to derive from variation in recombination rate (Nachman 2002). The observation of a block-like structure of LD with long stretches of strong allelic associations followed by shorter segments with weak associations, has led researchers to suggest that this structure reflects extensive recombination rate variation in the human genome (Daly *et al*. 2001; Gabriel *et al*. 2002).

Jeffreys *et al*. (2001) showed that recombination rate can vary on a kb scale at the class II region of the major histocompatibility complex (MHC). Using sperm typing to measure recombination rates over small distances (~2 kb), they found that recombination rate varied by more than three orders of magnitude, within distances of 10 kb. The highest rate was 130 cM/Mb while the lowest was <0.1 cM/Mb. Most but not all recombination events were restricted to hot spots. LD was found mainly in extended domains (haplotype blocks), interrupted by areas of LD breakdown. These areas correspond precisely to meiotic crossover hot spots at the MHC. It is not known however, to what extent this example can be generalized to other loci. Gabriel *et al*. (2002) observed that the human genome in general is characterized by haplotype blocks and by

hot spots of recombination. The authors examined haplotype patterns across 51 autosomal regions (spanning 13 Mb of the human genome) in samples from Africa, Europe and Asia and reported that the human genome can be parsed objectively into haplotype blocks; including sizable regions over which there is little evidence of historical recombination, and within which only a few common haplotypes are observed. The boundaries of blocks and specific haplotypes they contained were highly correlated across populations. The study demonstrated that such haplotype frameworks provide substantial statistical power in association studies of common genetic variation across each region and they highlight the need to develop this detail of data for other species such as plant species.

It has also been documented that regions of high LD, in general, correspond to regions of low recombination. For example, McVean *et al*. (2004) developed and validated a method for estimating recombination rates from patterns of genetic variation. From extensive SNP surveys in European and African populations, the authors found evidence for extreme local rate variation spanning four orders in magnitude, in which 50% of all recombination events take place in less than 10% of the sequence. They demonstrated that recombination hot spots are a ubiquitous feature of the human genome, occurring on average every 200 kb or less, but recombination occurs preferentially outside genes.

Recent human patterns of LD have also highlighted the importance of a second feature of recombination: homologous gene conversion (Frisse *et al*. 2001; Przeworski and Wall 2001). Ptak *et al*. (2004) estimated local recombination rates indirectly from patterns of LD in 84 genomic regions in a sample of individuals of European origin and of African–American descent. They found that LD based estimates are significantly positively correlated with map-based estimates. Also, using LD based estimators, the authors found evidence for homologous gene conversion in patterns of polymorphism. Frisse *et al*. (2001) also identified significant differences between the African and non-African populations which will impact on the design of future association studies in these populations. In general, there is less LD in African populations than in non-African populations. The half-length of D in the Utah population in USA is about 60 kb, whereas the half-length is considerably less than 5 kb for the Yoruba tribe from the southwestern part of Nigeria. Although it is generally believed that these results could be attributed to major human historical events particularly population bottlenecks associated with geographical expansion and population isolation, it would be worthwhile estimating inbreeding coefficients in these populations to see what role (if at all) it might have played in shaping the LD. Nevertheless, these studies highlight the importance of developing an understanding of the distribution of LD in any particular population as a prerequisite for subsequent experimental design. In a high LD population, genome-wide scans could be conducted to minimize the number of markers needed, and this could be followed by high resolution mapping in a low LD population (Reich *et al*. 2001).

Much attention is now focused on the identification of susceptibility genes underlying complex diseases, such as diabetes, schizophrenia and hypertension. Parametric linkage analysis narrowed the diastrophic dysplasia (DTD) gene to a ~2 Mb interval, but an LD study in Finnish patients pinpointed the gene to a ~40 kb interval and made its positional cloning possible (Häsbacka *et al*. 1992). The study also showed that the DTD gene lies within 0.06 cM (about 60 kb) of the colony stimulating factor 1 receptor (CSF1R) gene. Positional cloning of both the Huntington disease (HD), cystic fibrosis (CF) genes (Kerem *et al*. 1989) and one of the major Alzheimer factors (Corder *et al*. 1994) was helped greatly by LD mapping followed by association analysis. Several

studies have demonstrated a significant association between angiotensinogen (AGT) polymorphisms and hypertension, with a combined relative risk of ~1.2 for the T235 allele (Kunz *et al*. 1997; Kato *et al*. 1999; Staessen *et al*. 1999). The T235 allele is in nearly complete LD with the allele A(-6) and is associated with higher plasma AGT levels than are the M235 and G(-6) alleles (Inoue *et al*. 1997; Jeunemaitre *et al*. 1997; Iso *et al*. 2000; Pan *et al*. 2000; Rankinen *et al*. 2000; Rice *et al*. 2000; Sato *et al*. 2000).

A recent publication (The International Human HapMap Consortium 2005) outlines the considerable progress made in the human HapMap project which has recently reached the end of its first phase. The project will likely provide the raw data and design insights that will make association genetics studies possible over the next two decades. A substantial section of the genome (ten regions of 500 kb) has been sequenced from 48 individuals from four geographically defined populations in order to develop a base level of understanding of polymorphisms across the whole genome. This data supports previous findings discussed above, based on much smaller and less systematic analyses of the human genome, that had already postulated the existence of phenomena such as recombination hot spots, large block-like segments in LD and limited haplotype diversity in humans in regions other than these recombination hot spots. The sequence information also indicates that there are SNPs on average every 279 bases for the 48 diverse individuals analyzed. This data is already beginning to reveal some intriguing insights into human gene variation that make obvious biological sense. For example, genes under diversifying selection pressure (such as those involved in immune responses) are in regions of low LD, whereas genes involved in important core biological processes that are highly conserved across the living world (e.g., cell cycle, DNA repair) tend to reside in regions of high LD. Thus, as hypothesized for a long time, LD may well leave behind characteristic signatures of important natural selection events that have occurred in the past history of the population. At the same time, this analysis has led to a number of previously undetected observations that are still open to interpretation in terms of identifying a likely biological cause. For example, both regions of high and low LD in the genome show an association with gene dense regions of the genome. As the first comprehensive project of its type, Human HapMap is likely to be the forerunner of similar projects in other organisms.

2.5.2 LD in Drosophila

Much of our understanding of how LD is shaped in natural populations initially came from research on *Drosophila* species. *Drosophila* population history is still not well understood. *Drosophila melanogaster* and *Drosophila simulans* are human commensals; as with humans, they are thought to have originated in Africa, and only recently spread to other continents. The levels of diversity seem to be higher in African populations than non-African ones for *D. melanogaster* (David and Capy 1988).

The most detailed analysis of LD has been made in *D. melanogaster*, in which allelic combinations can readily be determined for individual chromosomes that have been extracted from wild populations through inbred lines. Most studies have focused on in-depth comparisons of single gene loci and/or single populations, and the principal finding is one of regional variations in LD among loci. A total of 3,143 pairwise comparisons involving 206 polymorphic restriction variants or eight gene regions of *Drosophila melanogaster* were included in one analysis (Zapata and Alvarez 1983). It was found that heterogeneity is mostly explained by large differences in the intensity of sample disequilibrium among regions. Langley *et al*. (2000) found evidence for a

surprisingly large amount of recombination at the tip of the X chromosome even though crossing over rates are known to be extremely low. They inferred that this must be the result of a high rate of gene conversion. Schaeffer *et al*. (2001) presented an analysis of protein variation at the ADH and ADH-related (ADHR) loci in the alcohol dehydrogenase (*Adh*) region in 139 strains of *Drosophila pseudoobscura*. Several conclusions can be drawn from the LD analysis of SNPs and ADHR haplotypes. First, recombination reduces the fraction of polymorphic loci that show associations with a disease-causing gene, but significant LD can be observed as a result of mutation and random genetic drift. Second, LD studies will be most effective in detecting allele-phenotype associations when the alleles are at moderate frequencies and the authors suggest that their model system conclusions may be applicable to other organisms. In-depth studies of how several forces (for example, mutation, recombination and selection) act to increase or decrease LD in a given region indicate that the balance of these forces should result in strongest disequilibrium around alleles at frequencies of ~10%. However, even adjacent regions can experience quite different evolutionary histories. A recent chromosome-wide study of the fourth chromosome (Wang *et al*. 2002), previously believed to be nonrecombining and invariable, found polymorphic regions interspersed with regions of little to no variation. Therefore, recombination was shown to occur on the chromosome, and although at very low rate consistent with previous findings, this has been sufficient to affect the structure of genetic variation on the chromosome, allowing different regions to have different evolutionary histories.

Recombination rates per physical length are well known to show marked regional variation, and much research on LD in *Drosophila* has used this factor to focus on understanding the effects of selection and other forces on the degree of LD. Over the past decade, numerous surveys of DNA sequence variation in natural populations of several *Drosophila* species have established that polymorphism levels are positively correlated with the regional rate of crossing over, and are not generally explained by variation in mutation rates (Wang *et al*. 2002; Begun and Aquadro 1992; 1994). This correlation has been proposed to result from the hitchhiking that is associated with fixation of advantageous mutants: in a region of low recombination, if directional selection drives an advantageous mutation through a population to fixation, much of the variations at linked sites will be eliminated during the process (Parsch *et al*. 2001). Selection on a region will therefore also increase the strength of LD observed: that is, significant allelic associations over large genetic distances might result from the action of natural selection. For example, strong geographical clinal variation in many enzyme loci around the phosphogluconate mutase (*Pgm*) locus is likely to be explained by clinal selection at *Pgm* and pervasive low levels of recombination in the region, so that the other loci are forced to hitchhike along with it (Verrelli and Eanes 2001). Selection against deleterious mutations can also reduce variation at linked sites. A recent analysis of multiple loci in *D. melanogaster* and *D. simulans* showed that both species have greater within locus LD than expected theoretically. This could be due to a departure from the demographic assumption of a panmitic equilibrium in *Drosophila* and/or the action of natural selection on many loci.

2.5.3 LD in Plants

Genetic diversity at the sequence level has been studied in only a handful of plant taxa, with maize and *Arabidopsis thaliana*, the most commonly studied species. These

two species have evolved different mating systems (out-crossing and selfing, respectively – although maize can be readily inbred and many cultivated varieties have been derived by this process, so maize might be more accurately described as a facultative inbreeder, in contrast to species like perennial ryegrass, which can accurately be described as an obligate out-breeder) and provide contrasting views of LD in plant genomes.

2.5.3.1. Maize

Maize is a good candidate for DNA sequence polymorphism survey because of its long history as a model genetic system and because of its agricultural importance. Maize was domesticated in Mexico about 7,500 years ago and dispersed throughout the Americas shortly thereafter. As a result of dispersal, there are now hundreds of maize landraces representing worldwide geographic locales. However, most of these have contributed little to modern maize breeding programs, and virtually all elite US inbred germplasm is derived from only a few landraces. The first published LD study on maize was based on a survey of 21 loci distributed along chromosome 1 of maize (Tenaillon *et al*. 2001). Each locus was sampled in 25 individuals representing a "species-wide" sample of maize that included US and exotic landraces. Although the length of these genes was short (1.5 kb), the rate of decay in LD was surprisingly rapid. On average, LD declined below nominal levels, which we arbitrarily define here as $r^2 = 0.20$, within 400 bp. By contrast, a subsample that included only US inbred lines demonstrated a lower rate of decay over distance, reaching nominal levels in ~1 kb. Higher LD in the US germplasm is consistent with the recent formation of these inbred lines and their relatively narrow genetic base.

A second study surveyed six genes of longer length (1.2–10 kb) in 102 inbred lines (Remington *et al*. 2001). These lines included tropical and semitropical lines and thus are more genetically diverse than samples of US inbred lines alone but probably less diverse than a species-wide sample. In this study, LD again declined rapidly; for five of six genes, LD was below the nominal level in 200 to 1,500 bp. In four of six genes sampled, predicted r^2 values declined to less than 0.1 within 2,000 bp, much less than the 50 kb observed for the same degree of LD decay in Northern European human population. However, LD did not decay to nominal levels in 10 kb for one gene, *shrunken* (*sh*1). Selection can also maintain elevated LD in localized regions. A subsequent study showed that *sh1*, an enzyme in the starch biosynthesis pathway, was under directional selection during either domestication or breeding (Whitt *et al*. 2002). This may provide an explanation for the persistence of LD at *sh1*. Although LD decays rapidly in a gene after selection for a particular allele (Przeworski 2002), maize is believed to have arisen from a single domestication event in southern Mexico about 9,000 years ago (Matsuoka *et al*. 2002). Based on this supposition, an appreciable selective effect on LD may still remain. Another surprising aspect of this study was that a genome-wide sample of 47 simple sequence repeats (SSRs) demonstrated higher levels of LD than SNPs in candidate genes. The reason for the apparent difference between SNPs and SSRs is unclear at present, but it may reflect differences in the type of historical information captured by markers with different mutation rates (Remington *et al*. 2001). Thornsberry *et al*. (2001) measured disequilibrium in and around the *Dwarf8* locus in maize, and found examples of disequilibrium spanning in excess of 3 kb in this region. An interesting feature was that within this region were regions in equilibrium, indicating a nonlinear decay of

disequilibrium, and the potential for more complex patterns of LD than simply being restricted to small regions of localized disequilibrium.

Longer stretches of LD have also been observed in maize: Jung *et al*. (2004) found stretches of disequilibrium of up to 500 kb in the vicinity of the *adh1* locus. Longer range LD has been reported in other plants, for example by Yin *et al*. (2004) in *Populus trichocarpa*, another out-crossing angiosperm, where LD was observed in the vicinity of a resistance gene at a distance of 34 and 16 kb, respectively. These results also indicated region-specific LD differences. Although these observations could be due to phenomena such as selective sweeps, they also raise the intriguing possibility, hypothesized by Rafalski and Morgante (2004), of nonuniform recombination between genic and nongenic regions, where less crossover occurs between the lower homology intergenic DNA (or alternatively, preferential pairing in regions of high sequence homology – such as expressed genes). Such a phenomenon may be restricted to species where there is less homology in intergenic sequences, such as in regions of the maize genome. Or it could also operate in regions where there are clusters of genes under more extreme diversifying selection such as resistance genes where the birth and death process can sometimes eliminate pairing gene partners and can result in a rather abrupt localized end to the region of homology along otherwise homologous chromosomes. Further re-sequencing of large stretches of gDNA such as BAC libraries will be informative in revealing the extent and frequency of longer range LD in out-crossing species.

2.5.3.2. Arabidopsis

Arabidopsis thaliana is believed to be 99% selfing and characterized by a patchy distribution of highly inbred populations (Abott and Gomes 1989; Bergelson *et al.* 1998; Todokoro *et al.* 1995). Therefore, it is expected to show extensive amounts of LD. Studies based on individual loci, for example, the alcohol dehydrogenase (Adh) locus (Hansfstingl *et al*. 1994; Innan *et al*. 1996), and the FAH1 and F3H loci (Aguadé 2001), have reported the presence of extensive LD in this species. However, the authors also observed some intralocus recombination which suggests a possible role of recombination in the evolutionary history of the species. To gain some insight into the relationship between recombination and the scale of LD decay in this species, Nordborg *et al.* (2002) sequenced 13 short segments (0.5–1.0 kb) from a 250-kb region surrounding the flowering time locus *FRI* in a global sample of 20 *Arabidopsis thaliana* accessions. These authors observed strong LD in the samples which however, decayed with distance up to ~1 cM (i.e., 250 kb). To compare this genome-wide LD decay with local LD decay, the same authors surveyed several local Michigan populations of *A. thaliana* species using markers surrounding the disease resistance locus *RPM1*, and observed extensive LD which decayed on a scale over 50–100 cM. The much stronger LD in the local populations was attributed to founder effect by the authors since *A. thaliana* was introduced into North America only about 200 years ago. The *FRI* locus may have contributed to local adaptation resulting in increased levels of LD. Two recent studies suggest that this may be the case. The first is a study of diversity in the region surrounding the disease resistance gene *rps5* (Tian *et al*. 2002). For this region, LD breaks down in as little as 10 kb in a species-wide sample. Similarly, LD decays within 10–50 kb around the *CLAVATA2* region (Shepard and Purugganan 2003). However, these two regions, like *FRI*, also may be atypical; both regions appear to have been subjected to balancing selection, which retains distinct alleles within populations for long periods of

time. When alleles are retained within populations for substantially longer periods than expected for neutral genes, there is time to accumulate relatively high levels of diversity and ample opportunity for recombination among alleles. Because *FRI*, *rps5*, and *CLAVATA2* may be atypical, additional studies of the genomic patterns of *Arabidopsis* LD are merited. Nonetheless, one can draw two conclusions. The first is that LD in species-wide samples decays far more slowly over physical distance in this selfing species relative to out-crossing maize; this difference is consistent with the low effective recombination rate and the demographic consequences of selfing. The second is that selection on a particular gene, such as *rps5*, affects the distribution of genetic diversity in neighboring genes through genetic associations. The extent of these effects likely is stronger in selfing than in out-crossing taxa.

Table 2.5. Linkage disequilibrium in different species

Species	LD	Criterion
Human	60 kb	D' half-length, North Europeans
Human	5 kb	D', half-length, Yoruba-Nigerians
Cattle	>10 cM	D' half-length
A. thaliana	50–100 kb	r^2, half-length
Soybean	>50 kb	Little LD decay found
Norway Spruce	~100 bp	r^2, half-length
Norway Spruce	~200 bp	$r^2 = 0.2$
Grape	>500 bp	r^2, half-length
Maize	~400 bp	$r^2 = 0.2$
Maize (inbreds from USA)	~1 kb	$r^2 = 0.2$
Maize	200–1,500 bp	$r^2 = 0.2$

Reprinted from Trends in Genetics 20(2), Rafalski and Morgante; Corn and humans: recombination and LD in two genomes of similar size, pp. 103–111, 2004, with permission from Elsevier.

2.6 CONCLUSION

LD mapping potentially has two advantages over conventional linkage or QTL mapping. The first is that it may be logistically easier. In theory, breeding schemes such as backcrosses or full sib matings may not be required, making experimental design more straightforward and saving considerable time. The second, and potentially greater, advantage offered by LD mapping is that traits (including QTLs) may be mapped to very small regions (particularly in outbreeders) thus enabling discovery of the underlying gene(s) and/or application for selection across a wider range of germplasm. Another advantage of LD mapping in self-pollinated crops with low diversity is that much more polymorphism is detected than in a biparental population. Different metrics can be used to assess LD, but some – in particular, r^2, have more utility for association genetics. Note here the paradox of high LD vs low LD. In the presence of high LD, lower marker density is required in a target region with greater potential for detecting markers strongly associated with the target gene polymorphism, even if distant physically. In the presence of low LD, many markers are required but the resolution of diagnostic markers is much higher, potentially to the level of the gene or QTN. LD relies on segregating variation in natural populations. As a result, LD mapping samples contain many more informative

meioses (i.e., all those that have occurred in the history of the sample) than traditional mapping population.

The extent of LD can vary across genomes and between species. Also, there is some indication of variation in genome-wide patterns of LD within a species, with some high LD regions interspersed with regions of low LD. Studies in model organisms such as *Drosophila*, maize and *Arabidopsis*, as well as humans, have informed plant geneticists of the potential complexity of phenomena that give rise to observed patterns of disequilibrium, including the heterogeneity of LD that can occur within species. Implications for association genetics are that global, or genome-wide averages may not adequately reflect patterns in specific regions, therefore patterns within regions of interest will need adequate elucidation for successful application of association genetics approaches. Population structure and biological behavior in particular, have pronounced effects on patterns of LD. Such differences have been observed in model plant species such as *Arabidopsis* and maize, where very different patterns have been observed. However, the current state of knowledge is such that even in model species, there is much to learn about the nature of LD and its underlying causes.

2.7 REFERENCES

Abecasis, G.R., Cookson, W.C., 2000, GOLD – graphical overview of linkage disequilibrium. Bioinformatics 16: 182–183.

Abott, R.J., Gomes, M.F., 1989, Population genetic structure and outcrossing rate of *Arabidopsis thaliana* (L.) Heynh. Heredity 62: 411–418.

Aguadé, M., 2001, Nucleotide sequence variation at two genes of the phenylpropanoid pathway, the FAH1 and F3H genes in *Arabidopsis thaliana*. Mol Biol Evol 18(1): 1–9.

Altshuler, D., Hirschhorn, J.N., Klannemark, M., Lingren, C.M., Vohl, M.-C., Nemesh, J., Lne, C.R., Schaffner, S.F., Bolk, S., Brewer, C., Tuomi, T., Gaudet, D., Hudson, T.J., Daly, M., Groop, L., Lander, E.S., 2000, The common PPARγ Pro12Ala polymorphisms is associated with decreased risk of type 2 diabetes. Nat Genet 26: 76–80.

Ardlie, K., Liu-Cordero, S.N., Eberle, M.A., Daly, M., Barret, J., Winchester, E., Lander, E.S., Kruglyak, L., 2001, Lower-than-expected linkage disequilibrium between tightly linked markers in humans suggests a role for gene conversion. Am J Hum Genet 69(3): 582–589.

Ball, R.D., 2005, Experimental designs for reliable detection of linkage disequilibrium in unstructured random population association studies. Genetics 170: 859–873.

Beer, S.C., Siripoonwiwat, W., O'Donoghue, L.S., Souza, E., Mathews, D., Sorrells, M.E., 1997, Associations between molecular markers and quantitative traits in an oat germplasm pool: can we infer linkages? J Agric Genom 3. [online] URL: http://www.ncgr.org/research/jag/papers97/paper197/indexp197.html.

Begun, D.J., Aquadro, C.F., 1992, Levels of naturally occurring DNA polymorphisms correlate with recombination rates in *D. Melanogaster*. Nature 356(6369): 519–520.

Begun, D.J., Aquadro, C.F., 1994, Evolutionary inferences from DNA variation at the 6-phosphogluconate dehydrogenase locus in natural populations of *Drosophila*: selection and geographic differentiation. Genetics 136(1): 155–171.

Bergelson, J., Stahl, E., Dudek, S., Kreitman, M., 1998, Genetic variation within and among populations of *Arabidopsis thaliana*. Genetics 148: 1311–1323.

Borevitz, J.O., Nordborg, M., 2003, The impact of genomics on the study of natural variation in *Arabidopsis*. Plant Physiol 132: 718–725.

Brown, G.R., Gill, G.P., Kaunz, R.K., Langley, C.H., Neale, D.B., 2004, Nucleotide diversity and linkage disequilibrium in loblolly pine. Proc Natl Acad Sci 42: 15255–15260.

Cannon, G.B., 1963, The effects of natural selection on linkage disequilibrium and relative fitness in experimental population of *Drosophila melanogaster*. Genetics 48: 1201–1216.

Cardon, L.R., Abecasis, G.R., 2003, Using haplotype blocks to map human complex trait loci. Trends Genet. 19: 135–140.

Chovnick, A., Schalet, A., Kernaghan, R.P., Kraus, M., 1964, The *rosy* cistron in *Drosophila melanogaster*: genetic fine structure analysis. Genetics 50: 1245–1259.

Clark, A.G., 1990, Inference of haplotypes from PCR-amplified samples of diploid populations. Mol Biol Evol 7: 111–122.

Corder, E.H., Saunders, A.M., Risch, N.J., Strittmatter, W.J., Schmechel, D.E., Gaskell Jr., P.C., Rimmler, J.B., Locke, P.A., Conneally, P.M., Schmader, K.E., Small, G.W., Roses, A.D., Haines, J.L., Pericak-Vance, M.A., 1994, Protective effect of apoliprotein E-type 2 allele for late on-s*et al*zheimer-disease. Nat Genet 7: 180–184.

Daly, M.J., Rioux, J.D., Schaffner, S.F., Hudson, T.J., Lander, E.S., 2001, High resolution haplotype structure in human genome. Nat Genet 29: 229–232.

David, J.R., Capy, P., 1988, Genetic variation of *Drosophila melanogaster* natural populations. Trends Genet 4: 106–111.

Devlin, B., Risch, N., 1995, A comparison of linkage disequilibrium measures for fine scale mapping. Genomics 29: 311–322.

Dvornyk, V., Sirviö, A., Mikkonen, M., Savolainen, O., 2002, Low nucleotide diversity at the *pal1* locus in the widely distributed *Pinus sylvestris*. Mol Biol Evol 19: 179–188.

Excoffier, L., Slatkin, M., 1995, Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12: 921–927.

Falconer, D.S., Mackay, T.F.C., 1996, Introduction to quantitative genetics (4th edition). Longman Group Limited, London.

Flint-Garcia, S., Thornsberry, J.M., Buckler IV, E.S., 2003, Structure of linkage disequilibrium in plants. Annu Rev Plant Biol 54: 357–374.

Frisse, L.R., Hudson, R.R., Bartoszewics, A., Wall, J.D., Donfack, J., Rienzo, D., 2001, Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. Am J Hum Genet 69: 831–843.

Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFliece, M., Lochner, A., Faggart, M., Liu-Cordero, S.N., Rotimi, C., Adebayo, A., Cooper, R., Ward, R., Lander, E.S., Daly, M.J., Altshuler, D., 2002, The structure of haplotype blocks in the human genome. Science 296: 2225–2229.

Gaut, B.S., Long, A.D., 2003 The lowdown on linkage disequilibrium. Plant Cell 15(7): 1502–1506.

Hagenblad, J., Nordborg, M., 2002, Sequence variation and haplotye structure surrounding the flowering time locus *FRI* in *Arabidopsis thaliana*. Genetics 161: 289–298.

Hansfstingl, U., Berry, A., Kellog, E.A., Costa III, J.T., Rüdiger, W., Ausubel, F.M., 1994, Haplotypic divergence coupled with lack of diversity at the *Arabidopsis thaliana* alcohol dehydrogenase locus: roles for both balancing and directional selection? Genetics 138: 811–828.

Häsbacka, J., de la Chapelle, A., Kaitila, I., Sistonen, P., Weaver, A., Lander, E., 1992, Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. Nat Genet 2: 204–211.

Hendricks, P., 1987, Gametic disequilibrium measures: proceed with caution. Genetics 117: 331–341.

Hill, W.G., Weir, B.S., 1994, Maximum-likelihood estimation of gene location by linkage disequilibrium. Am J Hum Genet 54: 705–714.

Innan, H., Tajima, F., Terauchi, R., Miyashita, N., 1996, Intragenic recombination in the *Adh* locus of the wild plant *Arabidopsis thaliana*. Genetics 143: 1761–1770.

Inoue, I., Nakajima, T., Williams, C.S., Quackenbush, J., Puryear, R., Powers, M., Cheng, T., Ludwig, E.H., Sharma, A.M., Hata, A., Jeunemaitre, X., Lalouel, J.M., 1997, A nucleotide substitution in the promoter of human angiotensinogen is associated with essential hypertension and affects basal transcription in vitro. J Clin Invest 99: 1786–1797.

Iso, H., Harada, S., Shinamoto, T., Sato, S., Kitamura, A., Sankai, T., Tanigawa, T., Iida, M., Komachi, Y., 2000, Angiotensinogen T174M and M235T variants, sodium intake and hypertension in non-drinking, lean Japanese men and women. J Hypertens 18: 1197–1206.

Jeffreys, A.J., Kauppi, L., Neumann, R., 2001, Intensely punctuate meiotic recombination in the class II region of the major histocompatibility complex. Nat Genet 29: 217–222.

Jennings, H.S., 1917, The numerical results of diverse systems of breeding, with respect to two pairs of characters, linked or independent, with special relations to the effect of linkage. Genetics 2: 97–154.

Jeunemaitre, X., Inoue, I., Williams, C., Charru, A., Tichet, J., Powers, M., Sharma, A.M., Gimenez-Roqueplo, A.P., Hata, A.S., Corvol, P., Lalouel, J.M., 1997, Haplotypes of angiotensinogen in essential hypertension. Am J Hum Genet 60: 1448–1460.

Jorde, L.B., 1995, Linkage disequilibrium as a gene mapping tool. Am J Hum Genet 56: 11–14.

Jung, M., Ching, A., Bhattramakki, D., Dolan, M., Tingey, S., Morgante, M., Rafalski, A., 2004, Linkage disequilibrium and sequence diversity in a 500-kbp region around the *adh1* locus in elite maize germplasm. Theor Appl Genet 109(4): 681–689.

Kaplan, N., Weir, B.S., 1992, Expected behavior of conditional linkage disequilibrium. American Journal of Genetics 51:333-343.

Kato, N., Sugiyama, T., Morita, H., Kurihara, H., Yamori, Y., Yazaki, Y., 1999, Angiotensinogen gene and essential hypertension in the Japanese: extensive association study and meta-analysis on six reported studies. J. Hypertens 17: 757–763.

Kerem, B., Rommens, J.M., Buchanan, J.A., Markiewicz, D., Cox, T.K., Chakravarti, A., Buchwald, M., Tsui, L.C., 1989, Identification of the cystic fibrosis gene: genetic analysis. Science 245: 1073–1080.

Knowler, W.C., Williams, R.C., Pettitt, D.J., Steinberg, A.G., 1988, $Gm^{3-5,13,14}$ and type 2 diabetes mellitus; an association in American Indians with genetic admixture. Am J Hum Genet 43: 520–526.

Kruglyak, L., 1999, Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat Genet 22: 139–144.

Krutovsky, K.V., Neale, D.B., 2005, Nucleotide diversity and linkage disequilibrium in cold-hardiness- and wood quality-related candidate genes in Douglas-fir. Genetics 171: 2029-2041.

Kunz, R., Kreutz, R., Beige, J., Distler, A., Sharma, A.M., 1997, Association between the angiotensinogen 235T-variant and essential hypertension in whites: a systematic review and methodological appraisal. Hypertension 30: 1331–1337.

Langley, C.H., Lazzaro, B.P., Philips, W., Heikkinen, E., Braverman, J.M., 2000, Linkage disequilibria and the site frequency spectra in su(s) and su(3(a)) regions of the *Drosophila melanogaster* X chromosome. Genetics 156(4): 1837–1852.

Levin, M.L., Bertell, R., 1978, Re: "Simple estimation of population attributable risk from case-control studies." Am J Epidemiol 108: 78–79.

Lewontin, R.C., 1964, The interaction of selection and linkage. I. General considerations; heterotic models. Genetics 49: 49–67.

Matsuoka, Y., Vigouroux, I., Goodman, M.M., Sanchez, G.J., Buckler, E., Doebley, J., 2002, Proc Natl Acad Sci USA 99: 6080-6084.

McVean, G.A., Myers, S.R., Hunt, S., Deloukas, P., Bentley, D.R., Donnelly, P., 2004, The fine scale structure of recombination rate variation in the human genome. Science 304: 581–584.

Nachman, M.W., 2002, Variation in recombination rates across the genome: evidence and implications. Curr Opin Genet Dev 12(6): 657–663.

Nordborg, M., 2000, Linkage disequilibrium, gene trees, and selfing: an ancestral recombination graph with partial self-fertilization. Genetics 154: 923–929.

Nordborg, M., Donnelly, P., 1997, The coalescent process with selfing. Genetics 146: 1185–1195.

Nordborg, M., Tavare, S., 2002, Linkage disequilibrium: what history has to tell us. Trends Genet 18: 83–90.

Nordborg, M., Borevitz, J.O., Bergelson, J., Berry, C.C., Chory, J., Hagenblad, J., Kreitman, M., Maloof, J.N., Noyes, T., Oefner, P.J., Stahl, E.A., Weigel, D., 2002, The extent of linkage disequilibrium in *Arabidopsis thaliana*. Nat Genet 30: 190–193.

Palaisa, K., Morgante, M., Williams, M., Rafalski, A., 2003, Contrasting effects of selection on sequence diversity and linkage disequilibrium at two phytoene synthase loci. Plant Cell 15: 1795–1806.

Pan, W.H., Chen, J.W., Fann, C., Jou, Y.S., Wu, S.Y., 2000, Linkage analysis with candidate genes: the Taiwan young-onset hypertension genetic study. Hum Genet 107: 210–215.

Parsch, J., Meiklejohn, C.D., Hartl, D.L., 2001, Patterns of DNA sequence variation suggest the recent action of positive selection in the *janus-ocnus* region of *Drosophila simulans*. Genetics 159(2): 647–657.

Pettersson, F., Oskar, J., Cardon, L.R., 2004, GoldSurfer: three dimensional display of linkage disequilibrium.

Phillips, M.S., Lawrence, R., Sachidanandam, R., Morris, A.P., Balding, J.D., Donaldson, M.A., Studebaker, J.F., Ankener, W.M., Alfisi, S.V., Kuo, F.-S., Camisa, A.L., Pazorov, V., Scott, K.E., Carey, B.J., Faith, J., Katari, G., Bhatti, H.A., Cyr, J.M., Derohannessian, V., Elosua, C., Forman, A.M., Grecco, N.M., Hock, C.R., Kuebler, J.M., Lathrop, J.A., Mockler, M.A., Natchtman, E.P., Restine, S.L., Varde, S.A., Hozza, M.J., Gelfand, C.A., Broxholme, J., Abecasis, G.R., Boyce-Jacino, M.T., Cardon, L.R., 2003, Chromosome-wide distribution of haplotype blocks and the role of recombination hotspots. Nat Genet 33: 382–387.

Pritchard, J.K., Przeworksi, M., 2001, Linkage disequilibrium in humans: models and data. Am J Hum Genet 69: 1–14.

Przeworski, M., 2002, The signature of positive selection at randomly chosen loci. Genetics 160: 1179–1189.

Przeworski, M., Wall, J.D., 2001, Why is there so little intragenic LD in humans? Genet Res 77: 143–151.

Ptak, S.E., Voelpel, S., Przeworski, M., 2004, Insights into recombination from patterns of linkage disequilibrium in humans. Genetics 167: 387–397.

Rafalski, A., Morgante, M., 2004, Corn and Humans: recombination and linkage disequilibrium in two genomes of similar size. Trends Genet 20(2): 103–111.

Rankinen, T., Gagnon, J., Perusse, L., Chagnon, Y.C., Rice, T., Leon, A.S., Skinner, J.S., Wilmore, J.H., Rao DC, Bouchard, C., 2000, AGT M235T and ACE ID polymorphisms and exercise blood pressure in the HERITAGE family study. Am J Physiol Heart Circ Physiol 279: H368–H374.

Reich, D.E., Cargill, M., Block, S., Ireland, J., Sabieti, P.C., Ritcher, D.J., Lavery, T., Kouyoumjian, R., Farhadian, S.F., Ward, R., Lander, E.S., 2001, Linkage disequilibrium in the human genome. Nature 41: 199–204.

Remington, D.L., Thornsberry, J.M., Matsuoka, Y., Wilson, L.M., Whitt, S.R., Doebley, J., Kresovich, S., Goodman, M.M., Buckler, E.S., 2001, Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc Natl Acad Sci USA 98: 11479–11484.

Rice, T., Rankinen, T., Province, M.A., Chagnon, Y.C., Perusse, L., Borecki, I.B., Bouchard, C., Rao, D.C., 2000, Genome-wide linkage analysis of systolic and diastolic blood pressure: the Quebec family study. Circulation 102: 1956–1963.

SAS Institute, Inc., 2004, SAS/Genetics 9.1 Users's Guide, Cary, NC: SAS Institute, Inc.

Sato, N., Katsuya, T., Nakagawa, T., Ishikawa, K., Fu, Y., Asai, T., Fukuda, M., Suzuki, F., Nakamura, Y., Higaki, J., Ogihara, T., 2000. Nine polymorphisms of angiotensinogen gene in the susceptibility to essential hypertension. Life Sci 68: 259–272.

Schaeffer, S.W., Walthour, C.S., Toleno, D.M., Olek, A.T., Miller, E.L., 2001, Protein variation in *Adh* and *Adh*-related in *Drosophila pseudoobscura*. Linkage disequilibrium between single nucleotide polymorphisms and protein alleles. Genetics 159(2): 673–687.

Shepard, K.A., Purugganan, M.D., 2003, Molecular population genetics of the *Arabidopsis CLAVATA2* region: The genomic scale of variation and selection in selfing species. Genetics 263: 1083–1095.

Souza, E., Sorrells, M.E., 1991, Relationships among 70 North American oat germplasms: I. Cluster analysis using quantitative characters. Crop Sci 31: 599–605.

Staessen, J.A., Kuznetsova, T., Wang, J.G., Emelianov, D., Vlietinck, R., Fagard, R., 1999, M235T angiotensinogen gene polymorphism and cardiovascular renal risk. J Hypertens 17: 9–17.

Tajima, F., 1989, Statistical method for testing the neutral mutation hypothesis by DNS polymorphism. Genetics 123: 585–595.

Tenaillon, M., Sawkins, M.C., Long, A.D., Gaut, R.L., Doebley, J.F., Gaut, B.S., 2001, Patterns of DNA sequence polymorphisms along chromosome 1 of maize (*Zea mays* ssp. Mays L.). Proc Natl Acad Sci USA 98: 9161–9166.

Terwilliger, J.D., 1995, A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic loci. Am J Hum Genet 56: 777–787.

The International HapMap consortium, 2005, A haplotype map of the human genome. Nature 437: 1299-1320.

Thornsberry, J.M., Goodman, M.M., Doebley, J., Kresovich, S., Nielsen, D., Buckler IV, E.S., 2001, *Dwarf8* polymorphisms associate with variation in flowering time. Nat Genet 28: 286–289.

Thumma, B.R., Nolan, M.F., Evans, R., Morgan, G.F., 2005, Polymorphisms in Cinnamoyl CoA reductase (CCR) are associated with variation in microfibril angle in *Eucalyptus* spp. Genetics 171:1257-1265.

Tian, D., Araki, H., Stahl, E.A., Bergelson, J., Kreitman, M., 2002, Signature of balancing selection in *Arabidopsis*. Proc Natl Acad Sci USA 99: 11525–11530.

Tishkoff, S.A., Dietzsch, E., Speed, W., Pakstis, A.J., Kidd, J.R. Cheung, K., Bonné-Tamir, B., Santachiara-Benerecetti, A.S., Moral, P., Krings, M., Pääbo, S., Watson, E., Risch, N., Jenkins, T., Kidd, K.K., 1996, Global patterns of linkage disequilibrium at the *CD4* locus and modern human origins. Science 271: 1380–1387.

Todokoro, S., Terauchi, R., Kawano, S., 1995, Microsatellite polymorphisms in natural populations of *Arabidopsis thaliana* in Japan. Jpn J Genet 70: 543–554.

Verrelli, B.C., Eanes, W.F., 2001, Clinal variation for amino acid polymorphisms at the *Pgm* locus in *Drosophila melanogaster*. Genetics 157(4): 1649–1663.

Wang, W., Thornton, K., Berry, A., Long, M., 2002, Nucleotide variation along the *Drosophila melanogaster* fourth chromosome. Science 295(5552): 134–137.

Weir, B.S., 1996, Genetic Data Analysis II, Sinaeur Sunderland, MA, USA.

Whitt, S.R., Wilson, L.M., Tenaillon, M.I,, Gaut, B.S., Buckler, E.S., 2002, Genetic diversity and selection in the maize starch pathway. Proc Natl Acad Sci USA 99: 12959–12962.

Yin, T.-M., DiFazio, S.P., Gunter, L.E., Jawdy, S.S., Boerjan, W., Tuskan, G.A., 2004, Genetic and physical mapping of *Melampsora* rust resistance genes in *Populus* and characterization of linkage disequilibrium and flanking genomic sequence. New Phytol 164: 95–105.

Zapata, C., Alvarez, C., 1983, On the detection of non-random association in natural populations of *Drosophila*. Mol Biol Evol 10: 823–841.

Chapter 3
WHAT ARE SNPs?

David Edwards[1], John W. Forster[1], David Chagné[2], and Jacqueline Batley[1]

3.1 SNP DEFINITION

In the simplest form, a single nucleotide polymorphism (SNP) is an individual nucleotide base difference between two DNA sequences. SNPs can be categorised according to nucleotide substitution as either transitions (C/T or G/A) or transversions (C/G, A/T, C/A, or T/G). As a nucleotide base is the smallest unit of inheritance, SNPs provide the ultimate form of molecular genetic marker. They also represent the most frequent type of genetic polymorphism, and the potential number of such markers is enormous in comparison with any but the most closely related genotypes within a given species (Rafalski 2002a,b). Sequence variation can have a major impact on how the organism develops and responds to the environment. Furthermore they are evolutionarily stable, not changing significantly from generation to generation (Lopez *et al*. 2005). SNPs provide an important source of molecular markers that are useful in genetic mapping, map-based positional cloning, detection of marker-trait gene associations through linkage and linkage disequilibrium (LD) mapping and the assessment of genetic relationships between individuals. The low mutation rate of SNPs makes them excellent markers for studying complex genetic traits and as a tool for the understanding of genome evolution (Syvanen 2001).

SNPs, at any particular site, could in principle involve four different nucleotide variants, but in practice they are generally biallelic. However, this disadvantage, when compared to multiallelic markers such as SSRs, is compensated by the relative abundance of SNPs. In humans, for a variation to be considered a true SNP, it must occur in at least 1% of the population. SNPs are suitable for automated discovery and detection, and can be applied to a wide range of purposes, including rapid identification of crop cultivars,

[1] Primary Industries Research Victoria, Victorian AgriBiosciences Centre, La Trobe R&D Park, Bundoora, Victoria 3083, Australia
[2] HortResearch, Plant Gene Mapping group, Private Bag 11030, Palmerston North, New Zealand

construction of ultra-high density genetic maps and association with genetic disorders (in humans and livestock) and agronomic traits (in livestock and crop plants).

SNPs provide an abundant source of DNA polymorphism in a number of eukaryote species. Information on the frequency, nature and distribution of SNPs in the majority of plant genomes is limited. However, the level of development and application of SNPs in higher plants, including some crop and tree species, is increasing, and consequently they provide an attractive marker system to plant breeders and geneticists. With the increasing availability of public sequence data and the rapid discovery of SNPs in plants, the development and application of SNP markers will continue to accelerate.

3.2 SNP FREQUENCY

SNPs can differentiate between related sequences both within an individual and between individuals in a population. In diploid species, in which an individual is heterozygous at a genetic locus, there are two homologous gene copies that may be differentiated by SNPs. The inheritance of each variant may be directly measured in the progeny. Detection of SNPs in individuals becomes complicated in the presence of gene or genome duplication. In these instances, it is often difficult to differentiate between homoeologous (between genome) and paralogous (within genome) duplication of genetic loci without detailed genetic inheritance studies. Because the majority of DNA in individuals within a related population is the same, genetic differences between individuals can be defined by SNPs. The frequency of SNPs (nucleotide diversity) and the haplotypic diversity (heterozygosity) between two individuals or within a population are direct measures of genetic diversity. Under conditions of forced inbreeding, such as recurrent backcrossing to parental individuals, sib-mating or mating between individuals with lower-degree relatedness, reduced genetic diversity and SNP frequency is observed. Such conditions may have arisen due to population reduction or isolation in natural populations (the so-called 'founder effect'). For domesticated crop plants, narrow genetic bases have contributed to corresponding reduced genetic diversity at the nucleotide level.

The frequency and nature of SNPs in plants is beginning to receive considerable attention. A number of reports in *Arabidopsis thaliana*, rice and maize have provided estimates of sequence diversity in these species. In many species, the analysis of DNA sequence variation has been confined to single genes or DNA fragments with the goal of defining gene structure, function or evolutionary relationships. It is known that SNPs are widely distributed throughout genomes, although various studies show that the occurrence and distribution of SNPs differs between species, in particular between inbreeding and outbreeding species, or in those species with a narrow genetic base. It is generally well accepted that some species, for example maize, are highly polymorphic, whilst others, such as soybean and melon, are less polymorphic. Detailed studies of sequence diversity have now been performed at selected loci for a range of plant species and in plants, the typical frequencies are in the range of 1 SNP every 100–300 bp.

The most advanced SNP studies in plants have been performed on model species where a large quantity of genomic or EST sequence is available. SNPs have been detected using high-throughput analysis in *A. thaliana* (Cho *et al*. 1999). ESTs are a good resource for SNP discovery and they have been used for SNP discovery in sugarbeet (Schneider *et al.* 2001), maize (Ching *et al.* 2002; Batley *et al.* 2003), rice (Nasu *et al.* 2002), soybean (Zhu *et al*. 2003) and sugarcane (Grivet *et al.* 2003). In soybean 280

SNPs from 143 amplicons (76.3 kb) have been identified (Zhu *et al.* 2003). In maize, one non-coding SNP/31 bp and one coding SNP/124 bp has been reported for 18 maize genes in 36 inbred lines.

A genome-wide polymorphism database of rice has been constructed defining polymorphisms between the cultivars Nipponbare (from sub-species *japonica*) and 93-11 (from sub-species *indica*) (Shen *et al.* 2004). The database contains 1,703,176 SNPs and 479,406 insertions/deletions (indels) (see Section 3.6 for further discussion on indels). This equates to approximately 1 SNP/268 bp in the rice genome. A similar study was also performed by Feltus *et al.* (2004). After aligning drafts of rice *indica* and *japonica* sequence and filtering to remove multiple copy and low-quality sequence, 384,341 candidate interspecific SNPs were identified, at a frequency of approximately 1.7 SNPs/kb. Due to the stringent filtering process, this is probably an underestimate of the real SNP frequency in rice. This work was performed again in 2005 (Yu *et al*. 2005) using alignments of the improved whole-genome shotgun sequences for *japonica* and *indica* rice. SNP frequencies varied from 3 SNPs/kb in coding sequence to 27.6 SNPs/kb in the transposable elements, with a genome wide measure of 15.13 SNPs/kb, or 1 SNP per 66 bp.

Further studies in rice have involved SNP discovery and characterisation in the *Piz* and *Piz-t* regions (Hayashi *et al.* 2004). The frequency was found to be similar to the previous studies, with an SNP found every 248 bp (Hayashi *et al.* 2004). The SNP frequency varied slightly depending on the cultivars being assessed. On average, 1 SNP was detected every 390 bp between cultivars Nipponbare and Zenith and 1 SNP per 173 bp between cultivars Nipponbare and Toride 1. The SNP frequency was higher between Zenith and Toride 1, with an SNP on average every 140 bp. In earlier studies, Yu *et al*. (2002) compared sequences from *japonica* and *indica* cultivars and found an average of 1 SNP every 170 bp, while Nasu (2002) reported a similar frequency for rice SNPs.

Extensive research has been performed on SNP frequency in barley. Russell *et al.* (2004) examined the frequency and distribution of SNPs within 23 genes associated with grain germination in barley in a range of accessions including European cultivars, landraces and wild barley. The frequency of SNPs was found to be 1 SNP every 78 bp. In a further study, the *Isa* (inhibitor of α-amylase) gene was sequenced in 16 barley genotypes to detect sequence polymorphisms (Bundock and Henry 2004). A total of 80 SNPs were identified in the 2,164 bp sequence, containing the *Isa* promoter, transcript and 3′-untranslated region (UTR), giving a high frequency of 1 SNP/27 bp. Kota *et al.* (2001) identified 72 polymorphic SNPs in seven genotypes of barley. The frequency of SNPs was estimated to be 1 every 240 bp. This was calculated from 52,140 bp of sequence from each genotype analysed. Similar studies have been performed in other crop species such as *Beta vulgaris* and *Zea mays*, for which the relevant frequencies were 1 every 60–130 bp and 104 bp, respectively (Schneider *et al.* 2001; Ching *et al.* 2002; Tenaillon *et al.* 2001). As expected, the frequency of SNPs in inbreeding species such as barley is lower than that observed in outbreeding species. This is further demonstrated in poplar, an out-breeding tree species, which exhibits a high level of variation. Cronk (2005) determined the presence of an SNP every 100 bp in poplar, increasing to 1 every 50 bp when geographically diverse species were included in the study.

In a study of 25 diverse genotypes of soybean (Zhu *et al.* 2003), a total of 280 SNPs were identified in 143 amplicons, totalling 76.3 kb sequence, providing 1 SNP per 273 bp. It was found that nucleotide diversity was lower in soybean than maize or *A. thaliana*, and this may be due to inbreeding. However, as *A. thaliana* is also self-pollinating, this does not explain all the findings. These results may also be due to

the narrow genetic base of soybean. SNP discovery has also been performed in lesser-known crops. Based on EST sequence information, fragments of 34 genes were amplified from five diverse quinoa (*Chenopodium quinoa* Willd.) accessions and the related weed species *C. berlandieri* and sequenced (Coles *et al.* 2005). Analysis of the quinoa EST sequences revealed a total of 51 polymorphisms in 20 EST sequences, including 38 SNPs and 13 indels. This was an average of 1 SNP every 462 bp, which increased to 1 SNP every 179 bp when *C. berlandieri* was included in the analysis. This SNP frequency is lower than that observed in barley (1/189 bp), maize (1/104 bp) and sugarbeet (1/130 bp), but similar to levels observed in soybean (1/503 bp) and *A. thaliana* (1/336 bp). Although the sample size was small, the SNP frequency reflects the narrow genetic base for cultivated quinoa.

Lopez *et al.* (2005) exploited a recently developed EST collection to identify SNPs in five cultivars of cassava (*Manihot esculenta* Crantz). One SNP per 905 bp was detected in intra-cultivar comparisons and 1 SNP per 1,032 bp was detected in inter-cultivar comparisons, based on data from 111 contigs, with an overall value of 1 SNP every 509 bp. This study also obtained further information on SNP frequency in six cultivars from 33 amplicons from 3′-EST and BAC end sequences. A total of 11 kb of sequence was obtained for each cultivar, with 186 SNPs being identified. Of these, 146 were observed within cultivars and 80 were observed between cultivars. The total frequency of SNPs was found to be one per 62 bp, a value similar to that observed for other crops. The intra-cultivar variation may be due to the presence of background heterozygosity and inbreeding depression within the lines. Cassava is also an ancient polyploid and predicted SNPs may be due to the presence of paralogous comparisons between members of multi-gene families.

In potato, 277 SNPs were identified between two alleles of the urease gene, with an average of 2.5 SNPs per 100 bp (Wittle *et al.* 2005). This average frequency of 1 SNP per 40 bp is relatively high for comparison between two alleles of a single copy gene. This is also reflected by studies of SNP variation in resistance gene analogues (RGAs) of cultivated potato, as described in Chapter 4.

3.3 SNP DISTRIBUTION

DNA is inherited in long stretches or blocks that are only separated by recombination events at meiosis. Because of this, groups of SNPs that are located in physical proximity to each other on the same stretch of DNA tend to be inherited together as a single linked group. A haplotype can be defined as a contiguous DNA sequence of an organism and may extend over physical distances characteristic of genes, gene clusters, chromosome segments, whole chromosomes or, in the case of asexual lineages, whole genomes. SNPs may be considered to define a haplotype, in that they are a series of DNA polymorphisms that differentiate between DNA sequences. As groups of SNPs that are in physical proximity tend to be inherited together (due to reduced capacity for genetic recombination and defining the extent of LD), haplotypes segregating in populations may be identified through the interrogation of one or a small number of diagnostic SNP loci.

The frequency of SNPs varies within each genome. Currently available data shows that the distribution of polymorphic sites is not random across the nuclear genome, or within a gene. SNPs can occur in coding and non-coding regions of the genome and at

different frequencies in different genomic regions. This uneven distribution may be due to differences in recombination rate, gene density, transmission pattern, selection strength and compositional pressure. Genomic regions with low recombination rates generally have reduced levels of polymorphisms (Rafalski and Morgante 2004). Regions subject to strong balancing selection (i.e. two or more alleles or haplotypes are maintained), such as those containing disease resistance genes, show the greatest diversity (Kuang *et al.* 2004).

The local abundance of SNPs within the genome varies due to a combination of the mutation rate that generates new polymorphisms and any positive or negative selection for regions linked to these mutations. SNP generation *de novo* may be more frequent outside of transcribed genic regions as these regions tend to exhibit greater levels of 5-methylcytosine (^{5me}C) abundance, an important factor in the generation of the most abundant C to T mutation due to deamination of ^{5me}C (which is aminothymidine) to T over evolutionary time. The majority of SNPs would be expected to be evolutionary neutral, that is, they would be neither selected for nor against, and their abundance in a population would vary due to random genetic drift. Rare deleterious mutations are counter-selected at a rate characteristic of the specific fitness penalty. For example, SNPs or Indels in transcribed sequences that lead to the production of altered proteins are relatively infrequent in populations when compared to similar polymorphisms within intron or untranscribed sequence. Selection, either natural or through breeding would lead to the removal of deleterious sequences from the population and increase the abundance of beneficial sequences. Selective pressure would apply to sequences in proximity to the selected sequence (the so-called 'hitch-hiking' phenomenon) unless they are separated by recombination during meiosis. Thus, strong selective pressure is likely to lead to genomic regions with reduced genetic diversity and fewer SNPs. This hypothesis is supported by the observation that in most organisms studied to date, SNPs are more prevalent in the non-coding regions of the genome. These mutations should theoretically only affect the phenotype if they cause a change in the regulation of gene expression, changing the expression pattern of surrounding transcribed regions. Within the coding regions, an SNP is either non-synonymous and results in an amino acid change, or is synonymous and does not alter the amino acid sequence and therefore is neutral. Non-synonymous SNPs may also be radical or conservative in nature, depending on transitions between positively charged, uncharged and negatively charged amino acid side-groups. Synonymous change may, however, potentially modify an RNA splice processing site resulting in phenotypic changes. SNPs have become popular tools for identifying genetic loci that contribute to phenotypic variation based on LD (see Chapters 2 and 7 for further discussion on the principles of LD).

The distribution of SNPs across the genome has been studied in a variety of plant species. Perhaps the most comprehensive study is in *A. thaliana*, where over 37,000 SNPs were identified by comparing partial genome sequence from the *Ler* accession with the near complete sequence of Col-0 (Schmid *et al.* 2003). The distribution of SNPs was found to be even across the five chromosomes, with the exception of centromeric regions, which contain few transcribed genes. In the ESTs studied, a total of 4,327 SNPs were identified. Analysis of amplicons derived from sequence tagged sites (STSs), corresponding to 4,955 consensus sequences revealed 3,773 SNPs. Of these, 2,922 (77%) were in non-coding regions of the genome. In the EST-derived SNPs, there was an average of 1 SNP per 336 bp. There was a higher ratio of synonymous to non-synonymous polymorphisms in EST compared to STS data, supporting the concept that expressed genes are more constrained by sequence evolution than randomly selected genomic loci.

A decreased frequency of SNPs in coding regions was also observed in quinoa (Coles *et al.* 2005). One SNP per 2,614 bp was observed in coding sequence, which increased to 1 SNP per 697 bp if the closely related weed species *C. berlandieri* was included in the analysis. The frequency of SNPs was much higher in the non-coding sequence, with an SNP per 385 bp, increasing to 1 every 144 bp in comparison to *C. berlandieri*. Of the SNPs in coding sequence, one was synonymous and three were non-synonymous. A detailed sequence analysis of four SSCP-SNP loci, over a panel of eight inbred pearl millet genotypes, revealed one SNP every 59 bp in introns, but considerably fewer in exons (Bertin *et al.* 2005). An elevated SNP frequency in non-coding sequence was also observed in maize, with 1 per 31 bp in non-coding regions and 1 per 124 bp in coding sequence (Ching *et al.* 2002). Five of the 18 SNPs in coding sequence were non-synonymous. In a study of SNP distribution in melon, 75% of the polymorphisms were located in introns and 3′-UTRs (Morales *et al.* 2004). Eleven SNPs (32%) were found in coding regions and the remaining 23 (68%) were found in 3′-UTR or intronic sequence. Seven of the eleven SNPs in coding sequence gave rise to synonymous changes. The proportion of synonymous compared to non-synonymous SNPs was also comparable with observations in maize, where 72% of SNPs in coding regions were synonymous (Ching *et al.* 2002). The higher presence of SNPs in non-coding regions has also been demonstrated in soybean (Zhu *et al.* 2003). These results suggest that UTRs and introns should be preferentially targeted for SNP discovery in candidate genes.

In a study of 25 diverse genotypes of soybean, 51 SNPs were identified in 28.7 kb coding sequence. Of these, 25 were synonymous and 26 were non-synonymous (Zhu *et al.* 2003). The rate of synonymous to non-synonymous base changes was lower in soybean than in maize, although similar to that seen in *A. thaliana*. Low diversity at non-synonymous sites is the result of selection against deleterious mutations. Out-crossing species are generally more effective at removing deleterious mutations as a consequence of large effective population size. Soybean and *A. thaliana* both exhibit low ratios of synonymous to non-synonymous mutation, suggesting the presence of a relatively high level of slightly deleterious mutations. The SNP distribution also varies in rice. Yu *et al*. (2005) aligned the improved whole-genome shotgun sequences for *japonica* and *indica* and found that SNP rates varied from 3 SNPs per kb in coding sequence to 27.6 SNPs per kb in the transposable elements. Furthermore, there were 4.72 SNPs per kb in the 5′UTR, 6.07 SNPs per kb in introns and 4.5 per kb in the 3′-UTR.

The *Isa* gene, which is significant for control of α-amylase activity during germination, was sequenced in 16 barley genotypes to detect sequence polymorphisms (Bundock and Henry 2004). A total of 80 SNPs and 23 indels were identified in 2,164 bp of sequence containing the *Isa* promoter, transcript and 3′-UTR. The frequency of SNPs was greatest in the 3′-non-translated region, downstream of the gene (1 SNP/16 bp), due to the contribution of comparison with sequences derived from wild barley (*Hordeum spontaneum*). One SNP per 75 bp was observed in the transcribed region, with 10 SNPs in the coding sequence, none in the 5′-UTR and 1 in the 3′-UTR. The region flanking the SSR in the promoter was highly polymorphic, with twice the number of SNPs expected given the overall frequency observed. This high frequency of SNPs surrounding SSRs has also been observed in maize (Mogg *et al.* 2002).

Two cultivars of soybean were distinguished by a non-synonymous transversion within the *GmNARK* (*Glycine max* nodule autoregulation receptor kinase) gene. Further sequence variants, including an indel and 5 SNPs, were detected in the intron and 5′-UTR respectively. There were a further 6 SNPs in the exons, all of which were synonymous (Kim *et al.* 2005). In the *Piz* and *Piz-t* regions of rice associated with rice blast resistance (Hayashi *et al.* 2004), SNPs were found every 248 bp (Hayashi *et al*. 2004).

Nucleotide polymorphism in the gene encoding phenylalanine ammonia-lyase (*Pal1*) of Scots pine (*Pinus sylvestris*) was studied by Dvornyk *et al.* (2002). A 2,045 bp exonic fragment of *Pal1* was sequenced in five megagametophytes from different individuals belonging to four populations, from Finland, Russia and Spain. Twelve polymorphisms were identified, and two alleles from a further 11 loci were studied (4,606 bp). Nine of the polymorphisms were synonymous and there were no introns in the sequence studied.

3.4 TRANSITIONS OR TRANSVERSIONS

SNPs are produced by mutations. The mutation frequency between any two nucleotides is not random but is dependent on the nucleotide base, the base sequence in its immediate proximity and the methylation status of the DNA. A major mechanism of spontaneous mutation is due to errors in DNA replication. Nucleotide bases in DNA can exist in two different structural forms (tautomers) called KETO and ENOL forms, but are predominantly found in the KETO form. Shifts to the ENOL form (tautomerisation) can alter pairing preferences, such that A may pair with C rather than T. Reversion of the tautomeric shift following DNA replication leads to fixation of a base mutation. The predicted average frequency of such processes is c. 1 per 10^4 bp copied, but the influence of fidelity maintenance systems such as polymerase proof-reading and post-replication mismatch repair results in observed frequencies of c. 1 in 10^{10} bp copied, corresponding to c. 1 in 10^6 per gene across a broad range of organisms.

Transitions are the most common form of SNP (Garg *et al.* 1999; Picoult-Newberg *et al*. 1999; Deutsch *et al.* 2001; Batley *et al.* 2003) reflecting the high frequency of the C to T mutation following deamination of methylated cytosine residues (Coulondre *et al.* 1978). C/T transitions constitute 67% of the SNPs observed in humans. Other variations in base substitution abundance are observed, but the underlying mechanisms for these differences remain to be explained (Batley *et al*. 2003).

Lopez *et al.* (2005) observed a significantly higher number of transitions than transversions in intra-cultivar (64% transitions) and inter-cultivar (65% transitions) comparisons in cassava. However, Coles *et al.* (2005) found an approximate 1:1 transition:transversion ratio in quinoa. A total of 20 transitions and 18 transversions were identified, increasing to 61 and 45, respectively, if the closely related weed species, *C. berlandieri,* was included in the analysis. This ratio was similar to those observed in maize, soybean (Zhu *et al.* 2003) and *A. thaliana*, but lower than the 2:1 ratios observed in sugarbeet, melon (Morales *et al.* 2004) and barley (Soleimani *et al.* 2003). The higher-than-expected C/T transition rate is likely to be due to the methylation effects described previously. Hayashi *et al.* (2004) found that 72–75% of SNPs between *indica* and *japonica* rice cultivars were transitions. This finding was supported by Feltus *et al.* (2004) who aligned drafts of the rice subspecies *japonica* and *indica* sequence and found that 65.8% SNPs were transitions and 34.2% were transversions. The high frequency of

transitions in this study is also compatible with the consequences of epigenetic modification of CG nucleotide motifs by DNA methylation in rice.

3.5 SNPs ASSOCIATED WITH ECONOMICALLY IMPORTANT GENES

As more genomes are being completely sequenced, interest is re-focusing on the discovery and analysis of intra-specific differences. SNPs can be used as simple genetic markers which may be identified in the vicinity of virtually every gene. There is potential for the use of SNPs to detect associations between the allelic forms of a gene and phenotypes, especially for common diseases in humans (see Chapter 2). SNPs have been identified in a number of plant genes of economic value. For example, SNPs were identified discriminating allelic variants of the potato urease gene in cultivar Desiree (Wittle *et al.* 2005).

SNPs associated with functional genes are candidate qualitative or quantitative trait nucleotides (QTNs) that are causally associated with the phenotypic effects of different alleles. However, the determination of QTNs is an intensive process which involves the use of data from methods such as induced mutagenesis, protein modelling and *in vitro* RNA and protein synthesis studies, as well as genetic analysis. For species without extensive LD, association studies can potentially be used to obtain very high map resolution, to the level of the QTN. Tree species such as Norway spruce (*Picea abies*), for which LD declines to minimal levels over very short distances (c. 50–100 bp) within genes (M. Morgante, pers. commun.), or the pine species *Pinus taeda*, for which the equivalent value is c. 1,500 bp, may be amenable to this form of analysis (Neale and Savolainen 2004). Species with moderate or high LD do not offer these advantages, and hence SNP haplotypes over the length of genes or gene clusters must be used to provide diagnostic tests for superior allele content.

3.6 INDELS

Small insertion or deletion events (indel for insertion/deletion) are another common form of genetic mutation. These mutations may be detected as SNPs as the insertion or deletion of nucleotides changes the sequence read. Indels may be produced by errors in DNA synthesis, repair or recombination, or may be due to the insertion and excision of transposable elements that often leave a characteristic DNA footprint of several nucleotide bases. For example, the relative abundance of eight base indels observed in maize by Bhattramakki *et al.* (2002) may be due to sequence duplication during insertion and excision of *Ac/Ds* transposable elements (Sutton *et al.* 1984).

Tenaillon *et al.* (2002) studied SNPs and indels located in previously published sequences from 21 loci on maize chromosome 1. Small indels (1–5 bp) were frequent, 56% of the indels being 1–2 bp in length and 92% were less than 20 bp in length. Furthermore, 5 of the 21 indels longer than 20 bp were found to be previously characterised Miniature Inverted-repeat Transposable Elements (MITEs). A total of 263 indels were observed in 17/21 loci. Indel size ranged from 1 to 640 bp, and the number per locus ranged from 2 to 59. This frequency of small indels was also observed in the *Piz* and *Piz-t* regions of rice. Of the 52 indels identified, 42 (81%) were 1–5 bp in length and only 4 were longer than 40 bp (Hayashi *et al.* 2004).

In a study of the urease gene in potato, 40 indels were observed within non-coding regions, of which 70% were 1–4 bp in length, 20% 5–10 bp and 10% (4 indels) were greater than 10 bp. The instances of these long indels may be explained by the relevant sequence features. One insertion was found to be due to a retrotransposon. A 30 bp indel is found in an array of 30 bp repeats within an intron and may have been caused by unequal cross-over, while a 34 bp indel is present in an SSR-containing region, which are known to undergo expansion and contraction (Wittle *et al.* 2005).

Morales *et al.* (2004) searched for indels in 34 ESTs between two distantly related melon genotypes. On average 1 indel was found per 1,666 bp. No indel was found inside the coding region. The indel length ranged from 1 to 13 bp, with single bp indels being the most frequent. This indel frequency was higher than in the total *A. thaliana* genome, in which one indel per 6.6 kb was observed (Jander *et al.* 2002). However, these data are not directly comparable to the melon study as both coding and non-coding regions were used in the *A. thaliana* study. Ching *et al.* (2002) examined the frequency and distributions of polymorphisms at 18 maize genes in 36 maize inbreds. Indels were found to be frequent in non-coding regions (1/85 bp) but rare in coding sequences.

In the genome wide polymorphism database of rice, using cultivars Nipponbare (*japonica*) and 93-11 (*indica*) (Shen *et al.* 2004), 479,406 indels were detected. This corresponds to approximately 1 indel per 953 bp in the rice genome. This indel frequency is higher than that observed in a similar study of the rice subspecies *indica* and *japonica* sequence by Feltus *et al.* (2004), who found approx. 0.11 indels/kb. However, due to the stringent sequence filtering performed in this later study, the result probably underestimates indel frequency in rice.

A total of 23 indels were identified between 16 barley genotypes in the 2,164 bp of *Isa* gene sequence (Bundock and Henry 2004), a measure of 1 indel per 94 bp. Four of these indels were within a microsatellite region and were excluded. Of the remaining 19 indels, 9 were 1 bp in length and the others ranged from 4 to 306 bp, giving an average frequency of 1 indel per 114 bp.

3.7 CONCLUDING REMARKS

SNPs are individual nucleotide base differences between DNA sequences and can represent differences between individuals or within populations. The specific base difference is determined by the cause of mutation and is non-random, with C to T transitions being the most frequent form. Insertion/deletion events (indels) are a special form of SNP caused by the addition or removal of DNA sequence, resulting in both length and sequence polymorphisms. The frequency of SNPs is dependent on both their generation and selection in populations. SNPs are generally evolutionally neutral, with frequencies varying due to random genetic drift. Some SNPs, particularly those associated with expressed genes, may be under positive or negative evolutionary selection pressure and will be maintained or rapidly removed from populations (Przeworski 2002; Bamshad and Wooding 2003). SNPs not separated by recombination at meiosis and thus in LD with other SNPs will be inherited as a linkage block and thus maintained at a frequency determined by the cumulative selection pressure of the haplotypic group. SNPs and indels are valuable molecular genetic markers due to both their abundance and relative stability in the genome, and can be applied as perfect molecular markers when identified within genes underlying observed traits.

3.8 REFERENCES

Bamshad, M., Wooding, S.P., 2003, Signatures of natural selection in the human genome. Nat. Rev. Genet. 4: 99–111.

Batley, J., Barker, G., O'Sullivan, H., Edwards, K.J., Edwards, D., 2003, Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. Plant Physiol. 132: 84–91.

Bertin, I., Zhu, J.H., Gale, M.D., 2005, SSCP-SNP in pearl millet – a new marker system for comparative genetics. Theor. Appl. Genet. 110: 1467–1472.

Bhattramakki, D., Dolan, M., Hanafey, M., Wineland, R., Vaske, D., Register, J.C. III, Tingey, S.V., Rafalski, A., 2002, Insertion–deletion polymorphisms in 3' regions of maize genes occur frequently and can be used as highly informative genetic markers. Plant Mol. Biol. 48: 539–547.

Bundock, P.C., Henry, R.J., 2004, Single nucleotide polymorphism, haplotype diversity and recombination in the *Isa* gene of barley. Theor. Appl. Genet. 109: 543–551.

Ching, A., Caldwell, K.S., Jung, M., Dolan, M., Smith, O.S., Tingey, S., Morgante, M., Rafalski, A.J., 2002, SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. BMC Genet. 3: 1–14.

Cho, R.J., Mindrinos, M., Richards, D.R., Sapolsky, R.J., Anderson, M., Drenkard, E., Dewdney, J., Reuber, T.L., Stammers, M., Federspiel, N., Theologis, A., Yang, W.H., Hubbell, E., Au, M., Chung, E.Y., Lashkari, D., Lemieux, B., Dean, C., Lipshutz, R.J., Ausubel, F.M., Davis, R.W., Oefner, P.J., 1999, Genome wide mapping with biallelic markers in *Arabidopsis thaliana*. Nat. Genet. 23: 203–207.

Coles, N.D., Coleman, C.E., Christensen, S.A., Jellen, E.N., Stevens, M.R., Bonifacio, A., Rojas-Beltran, J.A., Fairbanks, D.J., Maughan, P.J., 2005, Development and use of an expressed sequenced tag library in quinoa (*Chenopodium quinoa* Willd.) for the discovery of single nucleotide polymorphisms. Plant Sci. 168: 439–447.

Coulondre, C., Miller, J.H., Farabaugh, P.J., Gilbert, W., 1978, Molecular basis of base substitution hot spots in *Escherichia coli*. Nature 274**:** 775–780.

Cronk, Q.C.B., 2005, Plant eco-devo: the potential of poplar as a model organism. New Phytol. 166: 39–48.

Deutsch, S., Iseli, C., Bucher, P., Antonarakis, S.E., Scott, H.S., 2001, A cSNP map and database for human chromosome 21. Genome Res. 11**:** 300–307.

Dvornyk, V., Sirviö, A., Mikkonen, M., Savolainen, O., 2002, Low nucleotide diversity at the *pal*1 locus in the widely distributed *Pinus sylvestris*. Mol. Biol. Evol. 19: 179–188.

Feltus, F.A., Wan, J., Schulze, S.R., Estill, J.C., Jiang, N., Paterson, A.H., 2004, An SNP resource for rice genetics and breeding based on subspecies *Indica* and *Japonica* genome alignments. Genome Res. 14: 1812–1819.

Garg, K., Green, P., Nickerson, D.A., 1999, Identification of candidate coding region single nucleotide polymorphisms in 165 human genes using assembled expressed sequence tags. Genome Res. 9**:** 1087–1092.

Grivet, L., Glaszmann, J.-C., Vincentz, M., da Silva, F., Arruda, P., 2003, ESTs as a source for sequence polymorphism discovery in sugarcane: example of *Adh* genes. Theor. Appl. Genet. 106: 190–197.

Hayashi, K., Hashimoto, N., Daigen, M., Ashikawa, I., 2004, Development of PCR-based SNP markers for rice blast resistance genes at the *Piz* locus. Theor. Appl. Genet. 108: 1212–1220.

Jander, G., Norris, S.R., Rounsley, S.D., Bush, D.F., Levin, I.M., Last, R.L., 2002, *Arabidopsis* map based cloning in the post genome era. Plant Physiol. 129: 440–450.

Kim, M.Y., Van, K., Lestari, P., Moon, J.-K., Lee, S.-H., 2005, SNP identification and SNAP marker development for a GmNARK gene controlling supernodulation in soybean. Theor. Appl. Genet. 110: 1003–1010.

Kota, R., Varshney, R.K., Thiel, T., Dehmer, K.J., Graner, A., 2001, Generation and compairson of EST derived SSRs and SNPs in barley (*Hordeum vulgare* L.). Hereditas 135: 145–151.

Kuang, H., Woo, S.-S., Meyers, B.C., Nevo, E., Michelmore, R.W., 2004, Multiple genetic processes result in heterogeneous rates of evolution within the major cluster disease resistance genes in lettuce. Plant Cell 16: 2870–2894.

Lopez, C., Piegu, B., Cooke, R., Delseny, M., Tohme, J., Verdier, V., 2005, Using cDNA and genomic sequences as tools to develop SNP strategies in cassava (*Manihot esculenta* Crantz). Theor. Appl. Genet. 110: 425–431.

Mogg, R., Batley, J., Hanley, S., Edwards, D., O'Sullivan, H., Edwards, K.J., 2002, Characterisation of the flanking regions of *Zea Mays* microsatellites reveals a large number of useful sequence polymorphisms. Theor. Appl. Genet. 105: 532–543.

Morales, M., Roig, E., Monforte, A.J., Arús, P., Garcia-Mas, J., 2004, Single-nucleotide polymorphisms detected in expressed sequence tags of melon (*Cucumis melo* L.). Genome 47: 352–360.

Nasu, S., Suzuki, J., Ohta, R., Hasegawa, K., Yui, R., Kitazawa, N., Monna, L., Minobe, Y., 2002, Search for and analysis of single nucleotide polymorphisms (SNPs) in rice (*Oryza sativa, Oryza rufipogon*) and establishment of SNP markers. DNA Research 9: 163–171.

Neale, D.B., Savolainen, O., 2004, Association genetics of complex traits in conifers. Trends Plant Sci. 9: 325–330.

Picoult-Newberg, L., Ideker, T.E., Pohl, M.G., Taylor, S.L., Donaldson, M.A., Nickerson, D.A., Boyce-Jacino, M., 1999, Mining SNPs from EST databases. Genome Res. 9: 167–174.

Przeworski, M., 2002, The signature of positive selection at randomly chosen loci. Genetics 160: 1179–1189.

Rafalski, J.A., 2002a, Novel genetic mapping tools in plants: SNPs and LD-based approaches. Plant Sci. 162: 329–333.

Rafalski, J.A., 2002b, Applications of single nucleotide polymorphisms in crop genetics. Current Opin. Plant Biol. 5: 94–100.

Rafalski, A., Morgante, M., 2004, Corn and humans: recombination and linkage disequilibrium in two genomes of similar size. Trends Genet. 20: 103–111.

Russell, J., Booth, A., Fuller, J., Harrower, B., Hedley, P., Machray, G., Powell, W., 2004, A comparison of sequence-based polymorphism and haplotype content in transcribed and anonymous regions of the barley genome. Genome 47: 389–398.

Schmid, K.J., Rosleff Sörensen, T., Stracke, R., Törjék, O., Altmann, T., Mitchell-Olds, T., Weisshaar, B., 2003, Large-scale identification and analysis of genome wide single nucleotide polymorphisms for mapping in *Arabidopsis thaliana*. Genome Res. 13: 1250–1257.

Schneider, K., Weisshaar, B., Borchardt, D.C., Salamini, F., 2001, SNP frequency and allelic haplotype structure of *Beta vulgaris* expressed genes. Mol. Breed. 8: 63–74.

Shen, Y.-J., Jiang, H., Jin, J.-P., Zhang, Z.-B., Xi, B., He, Y.-Y., Wang, G., Wang, C., Qian, L., Li, X., Yu, Q.-B., Liu, H.-J., Chen, D.-H., Gao, J.-H., Huang, H., Shi, T.-L., Yang, Z.-N., 2004, Development of genome-wide DNA polymorphism database for map-based cloning of rice genes. Plant Physiol. 135: 1198–1205.

Soleimani, V.D., Baum, B.R., Johnson, D.A., 2003, Efficient validation of single nucleotide polymorphisms in plants by allele specific PCR, with an example from barley. Plant Mol. Biol. Rep. 21: 281–288.

Sutton, W.D., Gerlach, W.L., Schwartz, D., Peacock, W.J., 1984, Molecular analysis of *Ds* controlling element mutations at the Adh1 locus of maize. Science 223: 1265–1268.

Syvanen, A.C., 2001, Genotyping single nucleotide polymorphisms. Nat. Rev. Genet. 2: 930–942.

Tenaillon, M.I., Sawkins, M.C., Anderson, L.K., Stack, S.M., Doebley, J., Gaut, B.S., 2002, Patterns of diversity and recombination along Chromosome 1 of maize (*Zea mays* ssp. *mays* L.). Genetics 162: 1401–1413.

Wittle, C.-P., Tiller, S., Isidore, E., Davies, H.V., Taylor, M.A., 2005, Analysis of two alleles of the urease gene from potato: polymorphisms, expression and extensive alternative splicing of the corresponding mRNA. J. Exp. Botany 56: 91–99.

Yu, J., Hu, S., Wang, J., Wong, G.K., Songgang Li, S., Liu, B., Deng, Y., Dai, L., Zhou,Y., Zhang, X., Cao, M., Liu, J., Sun, J., Tang, J., Chen, Y., Huang, X., Lin, W., Ye, C., Tong, W., Cong, L., Geng, J., Han, Y., Li, L., Li, W., Hu, G., Xiangang Huang, X., Li, W., Li, J., Liu, Z., Li, L., Liu, J., Qi, Q., Liu, J., Li, L., Li, T., Wang, X., Lu, H., Wu, T., Zhu, M., Ni, P., Han, H., Dong, W., Ren, X., Feng, X., Cui, P., Li, X., Wang, H., Xu, X., Zhai, W., Xu, Z., Zhang, J., He, S., Zhang, J., Xu, J., Zhang, K., Zheng, X., Dong, J., Zeng, W., Tao, L. Ye, J., Tan, J., Ren, X., Chen, X., He, J., Liu, D., Wei Tian, W., Tian, C., Xia, H., Bao, Q., Li, G., Gao, H., Cao, T., Wang, J., Zhao, W., Li, P., Chen, W., Wang, X., Zhang, Y., Hu, J., Wang, J., Liu, S., Yang, J., Zhang, G., Xiong, Y., Li, Z., Mao, L., Zhou, C., Zhu, Z., Chen, R., Hao, B., Zheng, W., Chen, S., Guo, W., Li, G., Liu, S., Tao, M., Wang, J., Zhu, L., Yuan, L., Yang, H., 2002, A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). Science 296: 79–92.

Yu, J., Wang, J., Lin, W., Li, S., Li, H., Zhou, J., Ni, P., Dong, W., Hu, S., Zeng, C., Zhang, J., Zhang, Y., Li, R., Xu, Z., Li, S., Li, X., Zheng, H., Cong, L., Lin, L., Yin, J., Geng, J., Li, G., Shi, J., Liu, J., Lv, H., Li, J., Wang, J., Deng, Y., Ran, L., Shi, X., Wang, X., Wu, Q., Li, C., Ren, X., Wang, J., Wang, X., Li, D., Liu, D., Zhang, X., Ji, Z., Zhao, W., Sun, Y., Zhang, Z., Bao, J., Han, Y., Dong, L., Ji, J., Chen, P., Wu, S., Liu, J., Xiao, Y., Bu, D., Tan, J., Yang, L., Ye, C., Zhang, J., Xu, J., Zhou, Y., Yu, Y., Zhang, B., Zhuang, S., Wei, H., Liu, B., Lei, M., Yu, H., Li, Y., Xu, H., Wei, S., He, X., Fang, L., Zhang, Z., Zhang, Y., Huang, X., Su, Z., Tong, W., Li, J., Tong, Z., Li, S., Ye, J., Wang, L., Fang, L., Lei, T., Chen, C., Chen, H., Xu, Z., Li, H., Huang, H., Zhang, F., Xu, H., Li, N., Zhao, C., Li, S., Dong, L., Huang, Y., Li, L., Xi, Y., Qi, Q., Li, W., Zhang, B., Hu, W., Zhang, Y., Tian, X., Jiao, Y., Liang, X., Jin, J., Gao, L., Zheng, W., Hao, B., Liu, S., Wang, W., Yuan, L., Cao, M., McDermott, J., Samudrala, R., Wang, J., Wong, G.K.-S., Yang, H., 2005, The genomes of *Oryza sativa*: A history of duplications. PLoS Biology 3: 0266-0281.

Zhu, Y.L., Song, Q.J., Hyten, D.L., van Tassell, C.P., Matukumalli, L.K., Grimm, D.R., Hyatt, S.M., Fickus, E.W., Young, N.D., Cregan, P.B., 2003, Single nucleotide polymorphisms in soybean. Genetics 163: 1123–1134.

Chapter 4
SINGLE NUCLEOTIDE POLYMORPHISM DISCOVERY

David Edwards[1], John W. Forster[1], Noel O.I. Cogan[1], Jacqueline Batley[1], and David Chagné[2]

4.1 INTRODUCTION

As with the majority of molecular markers, one of the limitations of single nucleotide polymorphism (SNP) markers is the initial cost associated with their development. A variety of approaches have been adopted for the discovery of novel SNP markers in a wide range of organisms, including plants. These fall into three general categories, *in vitro* discovery, where new sequence data is generated, *in silico* methods that rely on the analysis of available sequence data and indirect discovery, where the base sequence of the polymorphism remains unknown. Methods for *in vitro* SNP discovery have been extended in recent years with the development of novel techniques for high-throughput resequencing. Furthermore, the reduced cost and increased throughput of SNP detection methods has enabled their extension for use in SNP discovery and validation. The value of the resulting data continues to drive technological developments, therefore it is difficult to predict the methods that are likely to be employed for SNP discovery in the future. Several current methods are detailed below and in Chapter 5. Methods for the discovery of SNPs from available sequence data are increasingly applied to a wide range of species, with some gene and genome-sequencing programs carefully selecting varieties considering their value for subsequent SNP discovery. Where there is a substantial quantity of sequence data available, *in silico* SNP discovery remains the cheapest and most efficient method for the identification of novel SNPs. When combined with validation using high-throughput resequencing methods, large numbers of SNPs may be identified and validated at a minimal cost. Where sequence data is limited or SNPs are required between specific lines or within certain genes, *in vitro* methods are generally more appropriate.

[1] Primary Industries Research Victoria, Victorian AgriBiosciences Centre, La Trobe R&D Park, Bundoora, VIC 3083, Australia

[2] HortResearch, Plant Gene Mapping group, Private Bag 11030, Palmerston North, New Zealand

4.2 *IN VITRO* APPROACHES FOR SNP DISCOVERY

4.2.1 "Nonsequencing" methods

The first techniques that will be presented are techniques that do not require the production of a large number of sequences to discover novel SNPs. These techniques are very popular because they are inexpensive and present the advantage of being applicable by any molecular biology laboratory since they use common reagents and equipments.

4.2.2 Restriction-based techniques: RFLP, CAPS, and dCAPS

Chronologically, the first method to be used for DNA polymorphism detection was restriction fragment length polymorphism (RFLP) (Botstein *et al.* 1980). This method was used successfully to detect point mutations occurring at restriction sites and was employed for mapping in a number of plant species (Keim *et al.* 1990). Nevertheless, this method is now rarely applied due to its technical limitations, i.e., labor-intensive and requiring large quantities of DNA. The next generation of molecular markers was based on the use of the polymerase chain reaction (PCR) technique. The first PCR-based marker, cleaved amplified polymorphic sequence (CAPS) (Konieczny and Ausubel 1993), is comparable to RFLP since it is based on the PCR-amplified fragments digestion using restriction endonuclease. The main drawback of CAPS is that, as with RFLPs, the SNP must occur within a restriction site. This restricts its use to a small minority of polymorphisms. To circumvent this problem, Neff *et al.* (1998) developed the dCAPS method ("derived" CAPS) where a restriction site can be created through the addition of a mismatch in a PCR primer located close to the SNP. In addition, the authors created a simple software system called *dCAPS Finder* (http://helix.wustl.edu/dcaps/dcaps.html), which facilitates the design of dCAPS markers. Although CAPS and dCAPS methods can be applied for genotyping SNPs in a relatively inexpensive way, these methods remain of low efficiency for SNP discovery. Indeed, a large number of restriction enzymes must be tested to find polymorphisms.

4.2.3 DNA conformation techniques: D/TGGE, SSCP, and heteroduplex analysis

Denaturing/temperature gradient gel electrophoresis (D/TGGE), single-stranded conformational polymorphism (SSCP), and heteroduplex-based methods (Figure 4.1) are based on the ability to distinguish the different conformations of short PCR-amplified DNA fragments. The DGGE (Myers *et al.* 1988) technique is based on the decreased electrophoretic mobility of partially melted double-stranded DNA molecules in polyacrylamide gels which contain a linear gradient of DNA denaturants, usually a combination of urea and formamide. TGGE is similar to DGGE but differs in the use of a temperature gradient to denature the DNA rather than chemical denaturing. TGGE and DGGE allow polymorphisms between DNA sequence strands to be detected to the resolution of a single base, making the methods applicable for SNP detection. As an example for plants, DGGE was successfully used for SNP discovery from pine expressed sequence tags (ESTs) and their subsequent genetic mapping (Brown *et al.* 2001; Gill *et al.* 2003).

SSCP (Orita *et al.* 1989) is a method for distinguishing between similar-sized DNA fragments according to the mobility of single-stranded DNA under polyacrylamide

gel electrophoresis. This method has been employed predominantly in human genetics, though SSCP has also been applied to detect SNPs in several plants like cereals (Martins-Lopes *et al.* 2001; Sato and Nishio 2003), forest trees (Plomion *et al.* 1999), horticultural trees (Etienne *et al.* 2002) and other crops (Hongtrakul *et al.* 1998; McCallum *et al.* 2001).

SNP detection can be based on resolving heteroduplex (i.e., mismatched hybridization between complementary DNA strands) from homoduplex (i.e., perfect hybridization) DNA fragments. Heteroduplexes can be formed during a heating/slow cooling procedure (Figure 4.1) with their subsequent differentiation from homoduplex sequences separated by polyacrylamide gel electrophoresis (Hauser *et al.* 1998). Heteroduplexes usually migrate slower than homoduplexes during electrophoresis due to the presence of mismatched base pairing. No sequence knowledge is needed prior to using this technique, which makes it suitable for SNP discovery in heterozygous individuals or pooled DNA.

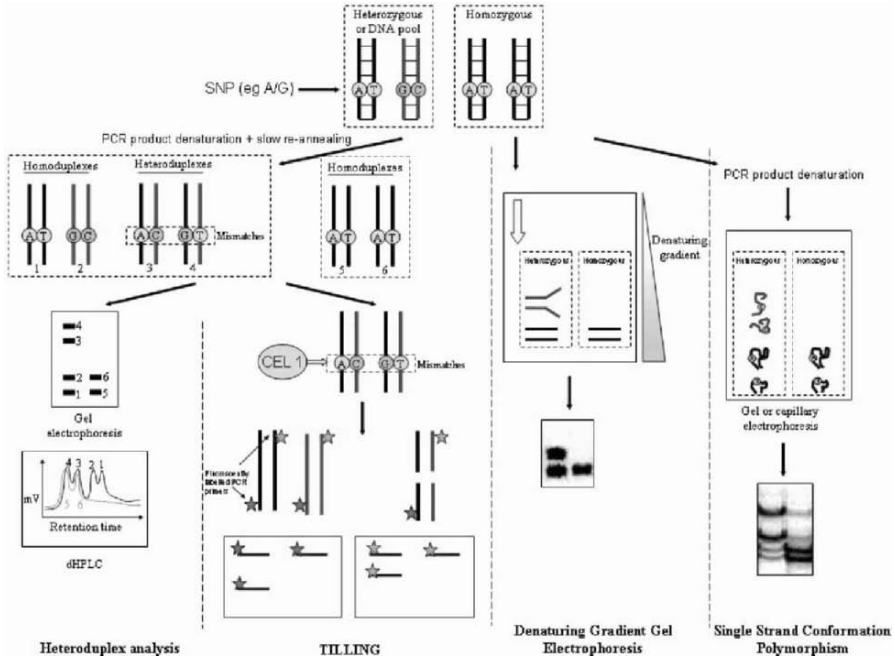


Figure 4.1. Nonsequencing SNP discovery methods: heteroduplex analysis, TILLING, DGGE, and SSCP. (see color plate)

D/TGGE, SSCP, and heteroduplex analysis are readily applicable in any molecular biology laboratory as they do not require sophisticated equipment to be used. The major drawbacks of these techniques, despite their relatively high efficiency in detecting SNPs, is their low-medium throughput, due to the use of polyacrylamide slab gels which require long migration times, and the use of ethidium bromide or silver staining methods. Indeed, these staining methods do not permit multiplexing and require the use of potentially hazardous chemicals. The speed and efficiency of SSCP and TGGE can be increased by the use of capillary electrophoresis systems (i.e., automated sequencers) and can also be

multiplexed by the application of fluorescent dye labels (Hebenbrock *et al.* 1995; Inazuka *et al.* 1997). As an example, Hsia *et al.* (2005) reported the use of temperature gradient capillary electrophoresis (TGCE), the capillary electrophoresis equivalent of TGGE, to detect SNPs in maize without prior knowledge of the polymorphic sequences. In addition, Kuhn *et al.* (2005) demonstrated the application of capillary electrophoresis-based SSCP in cocoa. However, the fact that the position and type of polymorphism are unknown when using D/TGGE, SSCP, or their capillary electrophoresis equivalents makes them less attractive to researchers who want to survey the position and nature of the polymorphism that may be associated with a trait variation.

As it has been reported in humans (Giordano *et al.* 1999), denaturing high-performance liquid chromatography (dHPLC) can also be used for detecting SNPs by heteroduplex analysis. dHPLC does not require gel-based genotyping procedures and is considered more accurate than polyacrylamide gel-based methods. DNA fragments are amplified by PCR, denaturated by heating, slowly cooled, and run through chromatographic columns using different temperatures. Because dHPLC assays can be performed in a relatively short time (5–15 min), are compatible with automation and do not require DNA resequencing, this method can provide an efficient means for relatively high-throughput SNP discovery and genotyping in plants (Kota *et al.* 2001).

4.2.4 TILLING

The targeting-induced local lesion in genomes (TILLING) method (Oleykowski *et al.* 1998; Till *et al.* 2003) is based on the use of a mismatch-specific endonuclease from the CEL1 family (Till *et al.* 2004). The CEL1 enzyme cleaves double-stranded DNA fragments at mismatch sites. These mismatch sites can be created during a denaturing/cooling procedure, with DNA pools or DNA derived from heterozygous lines. Unlike CAPS and dCAPS, TILLING does not require a site-specific restriction enzyme, which means that this method is portable to any type of SNP or INDEL without prior knowledge of the mutation position. In addition, this method may be automated and can be performed on automated sequencers such as the LI-COR (LI-COR, Lincoln, NE, USA), ABI3700 (Applied Biosystems, Foster City, CA, USA), or Megabace 1000 (GE Healthcare, Little Chalfont, UK), with restricted fragments visualized by fluorescent label detection. With the use of accurate DNA-sequencing gels, SNP positions can be determined through estimating the restriction fragment length.

Originally, TILLING was applied to identify mutations in specific genes of interest with subsequent analysis to determine the role of the genes (Greene *et al.* 2003), an approach often termed "reverse genetics." More recently, TILLING has been applied to study SNPs and functional mutations in natural populations, hence becoming "ecoTILLING" (Comai *et al.* 2004). This method is currently used for germplasm collection SNP screening in several plant species such as *Lotus japonicus* (Perry *et al.* 2003), wheat (Slade *et al.* 2005), and poplar (Gilchrist and Haughn 2005).

4.2.5 Chip-based method for SNP discovery

One of the major drawbacks for SNP discovery (and scoring) is the requirement for the PCR technique to reduce genome complexity. This is particularly problematic knowing that plants often have complex genomes. An ideal method for SNP discovery would be to scan the complete genome in a single reaction. However, only a few methods

that do not rely on PCR have been described to date. One particular method applies DNA chip technology to identify sequence polymorphisms. Borevitz *et al.* (2003) used a microarray, originally developed for gene expression studies, to identify new polymorphisms between *Arabidopsis thaliana* accessions (Col and L*er*). The difference in intensity between hybridization experiments was compared using similar statistical analysis as used for expression data. The authors demonstrated that the method could efficiently detect known polymorphisms and that detection is more efficient where the variation is close to the oligonucleotide features central base. Array-based discovery methods may represent the future for SNP discovery in particular cases where sequence information is great enough for gene array development but where there is not enough sequence information from different individuals to predict polymorphisms. This is particularly the case for species in which large EST databases have been generated from a small number of genotypes.

4.3 RESEQUENCING METHODS FOR SNP DISCOVERY

4.3.1 Pyrosequencing and the 454 technology

Pyrosequencing (Ahmadian *et al.* 2000) is a sequencing-by-synthesis method catalyzed by four kinetically well-balanced enzymes: DNA polymerase, ATP sulfurylase, luciferase, and apyrase (Figure 4.2). Pyrosequencing differs significantly from conventional Sanger sequencing. Rather than the extension of sequence fragments with the incorporation of labeled dideoxy nucleotide terminators, followed by detection of the labeled fragments, pyrosequencing incorporates the four different nucleotides in a defined order (e.g., CGAT) with detection concomitant with extension. Each addition of a nucleotide provokes an emission of light, which is detected as a peak on the *pyrogram*. The peak height is proportional to the number of nucleotides incorporated during a single step. This method of sequencing is significantly faster than Sanger DNA sequencing with 96 samples being sequenced within 5 min. This throughput makes it suitable for high-throughput SNP discovery and genotyping. Moreover, this method can be used for resequencing short fragments (up to 100 bp). In plants, pyrosequencing has been successfully applied in tetraploid potato (Rickert *et al.* 2002) and in loblolly pine (Brown *et al.* 2004). The recent implementation of the Genome Sequencer 20 System (Roche, Basel, Switzerland) developed by 454 Life Science (454 Life Science, Branford, CT, USA) allows large-scale resequencing (i.e., over 200,000 sequences produced simultaneously) of DNA fragments up to 150 bp, which is suitable for SNP detection in PCR amplicons. The method is based on the optimization of the pyrosequencing technique using fiber-optic slides and picoliter-scale volumes (Margulies *et al.* 2005).

Pyrosequencing fundamentally differs from Sanger's sequencing method in the order of nucleotide incorporation. Each nucleotide incorporation is accompanied by release of pyrophosphate (PPi) proportionally to the amount of nucleotide incorporated. ATP sulfurylase quantitatively converts PPi to ATP in the presence of adenosine 5′ phosphosulfate and this ATP permits the luciferase-mediated conversion of luciferin to oxyluciferin that generates visible light in amounts that are proportional to the amount of ATP. The light is detected by a charge-coupled device (CCD) camera and displayed

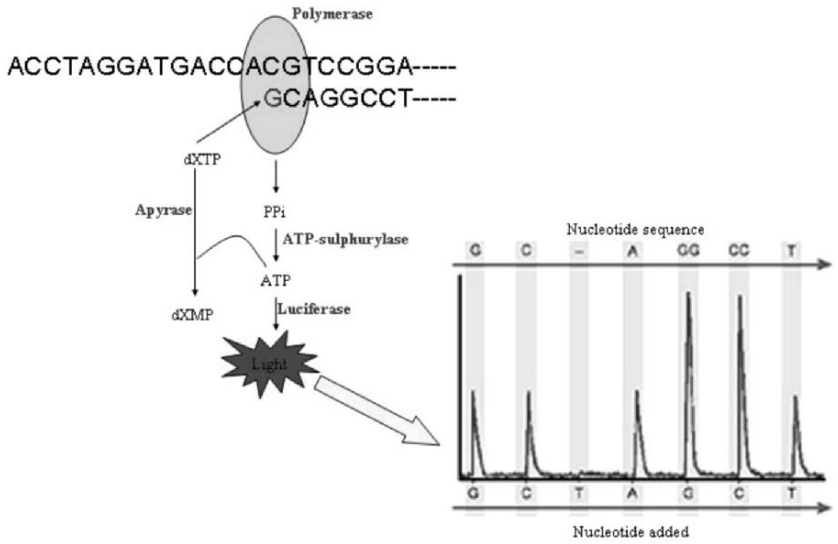


Figure 4.2. Pyrosequencing: principle.

as a peak in a "pyrogram." Peak height is proportional to the number of nucleotides incorporated. Apyrase continuously degrades unincorporated dNTP and excess ATP. After the degradation is completed, the next dNTP is added and a new pyrosequencing cycle is started. As the process continues, the complementary DNA strand is built up. To pyrosequence an unknown DNA sequence, a cyclic nucleotide dispensation order is generally used. As a result of each cycle of dATP, dGTP, dCTP, and dTTP dispensation, one of the four dNTPs is incorporated into the DNA template while the other dNTPs are degraded by Apyrase. Nucleotide sequence is determined from the order of nucleotide dispensation and peak height in the pyrogram. SNPs can be detected by aligning the sequences obtained by pyrosequencing or by a pattern recognition software.

4.3.2 MassArray

MassArray technology (Lau *et al.* 2000; Rodi *et al.* 2002) is based on the utilization of matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS). MALDI-TOF MS can detect differences between DNA fragments based on their molecular weight. As the molecular weight of the four nucleotides that make up DNA is different, this system is able to detect a single base variation in a PCR-amplified DNA fragment. The homogenous MassCleave (hMC) assay (Mattocks *et al.* 2004) is part of the MassArray platform developed by Sequenom (Sequenom, San Diego, CA, USA) and is suitable for SNP discovery. The principle of hMC is the following (Figure 4.3): PCR fragments between 300 and 700 bp in length are cleaved using an enzyme cutting at specific bases. Products are then run on a MALDI-TOF MS. The mass spectra obtained for the four cleavage reactions are compared to the theoretical spectra that were inferred

using reference sequences (i.e., the ones used for PCR design), or compared between different DNA pools or individuals. Differences between spectra can be due to sequence variations, the introduction or removal of a cleavage site, or mass shift due to the presence of an INDEL. The hMC method can be automated and multiplexed, which makes it a suitable method for high-throughput SNP discovery. However, to date there are no current examples of the use of the hMC technique for SNP discovery in plants, though the application has been applied successfully for studies in human genetics (Lau *et al.* 2000; O'Donnell *et al.* 1997).

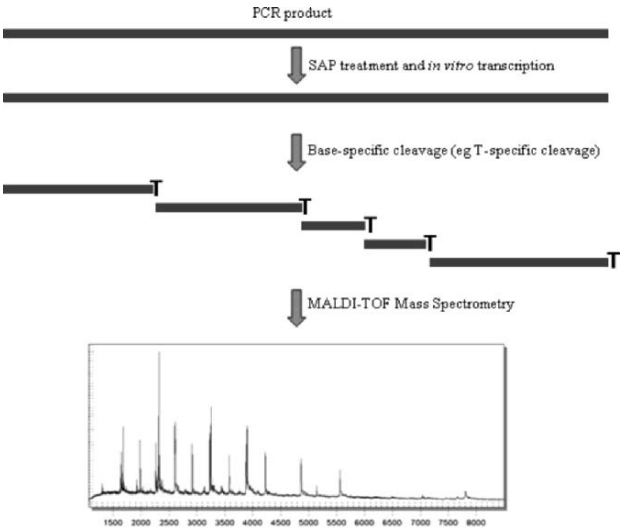


Figure 4.3. Homogenous MassCleave: principle. A PCR product is amplified, treated with SAP, and then *in vitro* transcribed. The transcription of the PCR product in RNA permits the base-specific cleavage using RNAse A. The resulting cleavage products are run on a MALDI-TOF MS, which generates a signal based on the fragment masses.

4.4 SNP DISCOVERY BASED ON SEQUENCING OF PCR AMPLICONS

4.4.1 General Principles

The most direct method for the discovery of SNP variation is DNA sequence analysis of genomic regions obtained by PCR amplification. The targeted DNA sequence is used to derive locus amplification primers (LAPs), which produce amplicons of a suitable size for analysis. These amplicons will generally be 500–700 bp in length, although the longer sequence reads that are characteristic of recent capillary electrophoresis platforms such as the ABI3730xl DNA Sequencer (Applied Biosystems, Foster City, CA, USA) may permit analysis of PCR products in the range of c. 1 kb. For gene-associated SNPs, each component of the transcriptional unit may be targeted, including 5′-untranslated regions (UTRs), coding sequence (CDS), and 3′-UTRs. "Tiling" of the gene provides a

particularly efficient method for gene-length SNP discovery. Access to full genomic sequences, such as those derived from large insert DNA libraries, permits direct primer design to upstream and downstream gene control elements and intragenic introns. Intron sequences may also be accessed by primer design to flanking exonic regions in cDNA sequences. However, without prior knowledge of intron–exon structure, recovery of intron sequences may be incomplete, due either to inadvertent primer design across intron–exon splice junctions, or to inefficient amplification of large introns. Certain genic regions are anticipated to show higher levels of SNP variation and may be preferentially targeted. These include the UTRs and introns, compared to exonic regions. As 3′-UTRs are frequently more extensive than their 5′-located counterparts, a number of studies have targeted these regions specifically.

Generation of PCR amplicons may be followed either by direct sequencing using one of the LAPs, or by cloning into a plasmid vector followed by clone-specific sequencing using a universal primer. The choice between direct sequencing and cloned amplicon sequencing is governed by a number of technical, statistical, and logistical considerations, and is highly influenced by the breeding system of the organism in question, as well as (in specialized cases such as conifer megagametophytes) by the ploidy level of the tissue used for SNP discovery.

Technical considerations include the efficiency and accuracy with which heterozygous SNPs may be identified by direct sequencing of an amplicon mixture derived from two (for diploid outbred genotypes) or more (for autopolyploid outbred genotypes) distinct haplotypes; the confounding effects of heterozygous indels, which produce overlapping phase shifts under conditions of direct sequencing; and similar effects arising from inadvertent amplification from multiple paralogous sequences.

Statistical considerations apply to the optimum number of cloned sequences selected for sequence analysis prior to alignment, based on expectations of allelic proportions, as well as the potential biasing effects of allele-specific PCR competition and paralogous sequence structure. The potential error rate associated with *in vitro* base substitution by thermostable polymerases must also be considered, as potential spurious SNPs may be generated in individual cloned sequences. This rate has been estimated at c. 1 in 10^3 bases replicated (Palumbi and Baker 1994), sufficiently high to require multiple clone sequencing for each allelic variant. In addition, cloned amplicon sequencing is prohibitive for large numbers of distinct genotypes, requiring appropriate experimental design in order to provide data of value across the broader germplasm pool of the target species. These issues are related to the logistical considerations, as amplicon cloning and sequencing is costly, laborious and time-consuming, especially during the process of manual sequence alignment.

4.4.2 Direct amplicon sequencing studies

SNP discovery by direct allele resequencing was originally performed in human genetics (Wang *et al.* 1998). For plants, the method has been used most effectively with either obligate or facultative inbreeding species. A number of representative studies have been performed in taxa such as maize (*Zea mays* L.) (Bhattramakki *et al.* 2002; Bhattramakki and Rafalski 2001; Ching *et al.* 2002; Mogg *et al.* 2002; Shattuck-Eidens *et al.* 1990), soybean (*Glycine max* L. Merr.) (Coryell *et al.* 1999; Zhu *et al.* 2003), *A. thaliana* (Jander *et al.* 2002; Olsen *et al.* 2004), wheat (*Triticum aestivum* L.) (Caldwell *et al.* 2004; Zhang *et al.* 2003), barley (*Hordeum vulgare* L.) (Bundock *et al.* 2003; Bundock

and Henry 2004; Russell *et al.* 2004), pearl millet (*Pennisetum glaucum* L.) (Gaut and Clegg 1993), and rice (*Oryza sativa* L.) (Bradbury *et al.* 2005; Hayashi *et al.* 2004; Jin *et al.* 2003).

For maize, the issue of paralogous sequences was minimized by predominant (but not exclusive) targeting of the 3′-ends of ESTs (Bhattramakki *et al.* 2002; Bhattramakki and Rafalski 2001; Ching *et al.* 2002). A similar approach was taken to discriminate between members of the cytochrome P450 gene family (Bundock *et al.* 2003), and for genes associated with grain germination in barley (Russell *et al.* 2004). For wheat, genome-specific primers were designed using pre-existing information on substitutions and indels in genes encoding ADP-glucose pyrophosphorylase and granule-bound starch synthase, and the specificity of amplification was determined through testing on nullisomic–tetrasomic (NT) substitution lines which permit discrimination between homoeologous gene sequences (Caldwell *et al.* 2004). In soybean, PCR products derived from a single standard genotype were pre-screened by gel electrophoresis in order to identify those primer sets that appeared to produce a single product, while those producing no or weak amplification, or multiple products, were discarded. In addition, sequencing from both ends with each amplification primer was used as necessary for additional quality control. Nonetheless, c. 20% of amplicons sequenced produced data attributable to heterogeneous template, demonstrating the importance of the paralogy problem (Zhu *et al.* 2003).

SNP variation between sequences from different homozygous genotypes was assessed visually (Ching *et al.* 2002), using the Phred/Phrap suite (Bhattramakki and Rafalski 2001; Ewing and Green 1998; Ewing *et al.* 1998), using Sequencher™ (Gene Codes, Ann Arbor, MI, USA; Bundock and Henry 2004; Caldwell *et al.* 2004; Mogg *et al.* 2002) or the PolyBayes SNP detection software (Zhu *et al.* 2003).

These and other studies have permitted estimates of SNP incidence over different germplasm samples. A comparison of genome sequences between two different accessions of *A. thaliana* predicted an SNP frequency of 1 per 6.6 kb (Jander *et al.* 2002). For the *A. thaliana CRY2* gene, comparison of a 3.2-kb region containing the entire transcriptional unit as well as over 1 kb of upstream and downstream sequences across 32 ecotypes revealed 90 SNPs and 12 indels, corresponding to frequencies of 1 per 36 bp and 1 per 267 bp, respectively. In soybean, comparison was based on 25 genotypes, of which 14 were estimated to have contributed 80.5% of allelic diversity present in North American varietal material. Resequencing was performed for 143 amplicons including coding and noncoding genic sequences selected from a total of 90 full-length genes and 88 cDNAs, as well as intergenic genomic sequences. A total of 280 SNPs were identified over 76.3 kb of genomic sequence, at a frequency of 1 per 272.5 bp (Zhu *et al.* 2003). In barley, a total of 2.7 kb from 23 grain germination-associated genes was resequenced across a panel of 24 cultivated barley accessions, eight landraces, and eight lines of the progenitor species *H. spontaneum*, identifying 1 SNP per 78 bp and 1 indel per 680 bp (Russell *et al.* 2004). Although the selection and range of germplasm clearly influences such estimates, the obligate inbreeding habit and narrow genetic bases typical of such species generally contributes to low SNP frequency. By contrast, higher values have been reported for facultative allogamous species such as maize. The study of 3′-UTR targeted amplicons in 22 amplicons from 18 genes was performed using 36 diverse maize genotypes, representing the major US-derived heterotic germplasm groups (Ching *et al.* 2002). Across a total of 6.9 kb of genomic sequence, the SNP frequency was 1 per 61 bp and the indel frequency was 1 per 126 bp. SNP frequency in coding sequence was 1 per 130.5 bp and in noncoding sequence was 47.7 bp, while the distribution of indels showed a similar

pattern. Further studies based on analysis of several hundred loci across eight inbred maize lines (Bhattramakki *et al.* 2002; Bhattramakki and Rafalski 2001) revealed comparable frequencies (1 SNP per 83 bp, 1 indel per 250 bp).

Direct sequencing has also been applied to obligate outbreeding (allogamous) species such as potato (*Solanum tuberosum* L.) (Rickert *et al.* 2003). BAC library clones containing sequences similar to nucleotide-binding site and leucine-rich repeat (NBS–LRR) type pathogen resistance genes were selected for analysis, and PCR amplicons were designed in candidate genomic regions. Comparative sequence analysis was performed using a panel of 17 autotetraploid and 11 diploid potato genotypes. A total of 78 amplicons with a total sequence length of 31 kb were reanalyzed across the germplasm panel. Predicted heterozygous indels were confirmed by sequencing from the opposite end of the amplicon with the second amplification primer, and SNP dosage in heterozygous autotetraploid combinations (i.e., ABBB, AABB, or AAAB) was estimated from overlapping sequence peak heights. A total of 1,498 SNPs and 127 indels were identified visually, corresponding to frequencies of 1 per 21 bp and 1 per 243 bp, respectively.

In conifers, SNP discovery by resequencing PCR products can be facilitated by the use of megagametophyte. Megagametophyte is a haploid endosperm developing from the maternal gamete, with nutritive functions for the surrounded zygote. The advantage of using megagametophyte for SNP discovery in conifers is that time-consuming and costly cloning procedures become unnecessary, given that sequencing reactions can be performed directly from PCR-amplified fragments to have access to haplotype sequences. The sequencing of several PCR products using the same endosperm gives the haplotype structure of the mother plant. This information can further be compared with sequenced PCR products from the diploid embryo to infer the paternal haplotype. The use of megagametophyte was very popular in the 1990s for genetic mapping in gymnosperms. This approach was recently employed for linkage disequilibrium studies using SNPs, in Japanese sugi (Kado *et al.* 2003), loblolly pine (Brown *et al.* 2004; Gill *et al.* 2003), and maritime and Monterrey pine (Pot *et al.* 2005).

4.4.3 Cloned amplicon sequencing studies

Despite the labor-intensive nature of amplicon cloning and sequencing (Zhang and Hewitt 2003), and the possibility of artifactual results due to *in vitro* recombination of cloned heteroduplexes (Tang and Unnasch 1995), the method provides a number of significant advantages. Linkage phase between contiguous heterozygous SNPs may be unambiguously determined in primary analysis, allowing SNP haplotype structure in the target region to be determined. In addition, as noted previously, heterozygous indels of variable length and paralogous sequences may be unambiguously identified.

In animal systems, amplicon cloning and sequencing has been used for species such as humpback whales (Palumbi and Baker 1994), black tiger prawn (Duda and Palumbi 1999), and turnip moth (LaForest *et al.* 1999). The results of several analyses in plant taxa have been published, and numerous studies are currently being performed in forestry, horticultural and forage species. In maize, amplicons corresponding to a c. 600 bp region of the *b* anthocyanin biosynthesis-regulatory gene were obtained from 18 different genotypes, including 18 inbred lines and 7 ancestral lines. Cloned amplicons were sequenced and aligned using CLUSTAL W (Thompson *et al.* 1994) to identify SNPs and indels (Selinger and Chandler 1999). The *teosinte-branched1* (*tb1*) domestication locus was targeted in cultivated maize and two species of the ancestral grass teosinte (*Z. mays*

ssp. *parviglumis* and *Z. mays* spp. *mexicana*) by amplicon cloning in the TOPO TA-cloning system (Invitrogen, Carlsbad, CA, USA) revealing limited coding sequence variation, but substantial promoter region divergence (Wang *et al.* 1999). Interspecific comparisons have also been made for the alcohol dehydrogenase (*Adh*) locus of *A. lyrata*, which is allogamous, and *A. thaliana*, which is autogamous. A combination of amplicon cloning and sequencing and direct sequencing strategies were used for *A. lyrata*, followed by alignment using CLUSTAL W (Savolainen *et al.* 2000).

The genomic complexity of hexaploid bread wheat has been addressed by sequence analysis of cloned amplicons derived from RFLP probes previously used for genetic map construction (Bryan *et al.* 1999). Low levels of SNP were detected between homologous sequences, at c. 1 per 1,000 bp. The majority of amplicons designed against template wheat sequences amplified at least two distinct products, reflecting potential homoeolocus and paralocus detection. Differences in the length of PCR products obtained with specific primers could be exploited to design genome-specific amplicons. Amplicon cloning has also been used to detect allelic variation in high molecular weight glutenin subunits from the wheat D-genome progenitor species *Aegilops tauschii* (Lu *et al.* 2005).

Amplicons from the nuclear ribosomal DNA internal transcribed spacer (ITS) region were obtained from individuals of different species of spruce (genus *Picea*) and cloned in pGEM vectors in order to identify SNPs capable of distinguishing black spruce (*Picea mariana*) and red spruce (*Picea rubens*) (Germano and Klein 1999). Nucleotide diversity has also been studied in the European aspen (*Populus tremula* L.) through analysis of 24 different trees from four different geographical sites (Ingvarsson 2005). Five gene loci (*Adh1*, *CI-1*, *GA20ox1*, *TI-3*, and *Gapdh*) were used for primer design to generate amplicons from each genotype that were directly cloned into the TA-cloning vector pCR2.1 and subsequently individually sequenced. Sequence alignments were performed using Sequencher, revealing an SNP frequency (across 6.2 kb) of 1 per 60 bp. Similar studies have been performed for other long-lived woody perennial species such as the silver birch, *Betula pendula* (Järvinen *et al.* 2003). Amplicons from the *PISTILLATA* (*PI*) homologue *BpMADS2* gene, spanning c. 2.4 kb, were derived from ten individuals from each of two Finnish populations. The amplicons were cloned into the pUC18 vector, sequenced and aligned to reveal limited haplotype diversity, with two common types in each population.

Amplicon cloning and sequencing has also been used for SNP discovery in potato, as a complement to the direct sequencing activities described above. A study of the LRR-encoding *StVe1* resistance gene in potato (Simko *et al.* 2004) was performed using a sample set of 30 North American tetraploid cultivars. PCR products (c. 839 bp in length) were directly cloned using the TOPO TA-cloning system, and a total of 600 cloned fragments (20 per cultivar) were sequenced. The average SNP incidence was 1 per 15 bp, but the nucleotide diversity was organized into a number of highly distinct haplotypes, of which three were detected in 97% of the analyzed cultivars. A paralogous sequence of 851 bp in length was also cloned and discriminated from the *StVe1* amplicon by PolyBayes analysis, through the presence of 2–6 bp indels.

4.5 CASE HISTORY: SNP DISCOVERY IN PERENNIAL PASTURE PLANT SPECIES

The Poaceae species perennial ryegrass (*Lolium perenne* L.) and the Fabaceae species white clover (*Trifolium repens* L.) are the most important components of temperate

pastoral agriculture systems, supporting grazing industries for dairy, meat, and wool production. The majority of research to date on molecular marker development and validation in out-crossing pasture species has been based on anonymous genetic markers, such as genomic DNA-derived simple sequence repeats (SSRs) and amplified fragment length polymorphisms (AFLPs) (Jones *et al.* 2002a, b, 2003). The paradigm for marker-assisted selection (MAS) that was established in autogamous plant species such as tomato, rice, and wheat involves the use of such markers to construct linkage maps, genetic trait dissection through QTL analysis, and selection of linked markers in selection schemes such as donor–recipient recurrent selection. The obligate outbreeding nature of pasture grasses and legumes clearly presents major limitations to the ready implementation of the inbreeding paradigm.

The most obvious solution to such problems is to develop candidate gene-based markers that show a functional association with the target trait region (Andersen and Lübberstedt 2003). Based on the population biology of perennial ryegrass and white clover (outbreeding with relatively large effective population sizes, at least for ecotypic populations), linkage disequilibrium (LD) is expected to extend over relatively short molecular distances. In this instance, it should be possible to identify diagnostic variants for the selection of individual parental genotypes on the basis of superior allele content. This will allow more efficient use of germplasm collections for parental selection. In addition, such "perfect" markers will allow highly effective progeny selection (Forster *et al.* 2004).

Large-scale gene sequence collections which have been generated by both incremental and EST discovery in perennial ryegrass and white clover provide the resource for functionally associated marker development, with c. 15,000 unigenes currently defined for each species (Sawbridge *et al.* 2003a, b). Selected genes have already been mapped as gene-associated RFLP and SSR loci (Barrett *et al.* 2004; Faville *et al.* 2005). RFLP markers are not readily implemented in molecular breeding, and SSRs are only present in a subset (generally less than 10%) of target genes. However, genic SNP markers can in principle be developed for any gene, and show the benefits of locus-specificity, high data fidelity, and high-throughput analysis. The experimental method for SNP discovery is based on cloning and sequencing of gene-specific amplicons from the heterozygous parents of two-way pseudo-testcross mapping families. The putative SNPs are then validated in the progeny set, and cross-validated in other sibships and diverse germplasm.

In perennial ryegrass, which is a diploid species ($2n = 2x = 14$), *in vitro* gene-associated SNP discovery process has been based on a three-part strategy. The "fast-track" component involves short ESTs, providing single SNP loci for structured map enhancement; "medium-track" involves full-length cDNAs, providing several SNP loci and partial haplotypic data; and "slow-track" is based on full-length genes with intron–exon structure, providing multiple SNP loci and determination of complete haplotype structures. Such data may be used to determine the extent of linkage disequilibrium and stability of gene-length SNP haplotypes, and to test for causal correlation between genotypic diversity and corresponding variation for related agronomic traits (Cogan *et al.* 2006).

"Proof-of-concept" for the *in vitro* discovery process was obtained with the perennial ryegrass *Lp*ASRa2 gene. The *Asr* gene family encodes a group of proteins that are transcriptionally induced by ABA treatment and water stress, and during fruit ripening. Osmotic and saline stress leads to up-regulation of the rice gene (Vaidyanathan *et al.* 1999), and the maize *Zm-Asr1* gene co-locates with QTLs for traits responsive to mild water stress (Jeanneau *et al.* 2002). *Lp*ASRa2 consequently provides an excellent candidate for

the assessment of correlation between genic sequence polymorphism and phenotypic variation. In addition, Southern hybridization analysis suggested that *Lp*ASRa2 is present as a single copy gene in the perennial ryegrass genome, reducing the potential complicating effects of sequence paralogy (Cogan *et al.* 2006).

The full-length *Lp*ASRa2 cDNA (890 bp) was tiled with four amplicons, covering 716 bp of the 5′-UTR, CDS, and 3′-UTR, and including a single 100 bp intron. A total of nine SNPs were detected within and between the parents of the $F_1(NA_6 \times AU_6)$ perennial ryegrass mapping family (Faville *et al.* 2004) by alignment of cloned amplicon sequences using Sequencher™. Of these, seven SNP loci showing segregation in the progeny were validated by single nucleotide primer extension (SNuPe, GE Healthcare, Little Chalfont, UK) assays on a MegaBACE 1000 automated capillary electrophoresis (CE) platform. The segregating markers were used to genotype the full $F_1(NA_6 \times AU_6)$ sibship, and were assigned to coincident locations on linkage group (LG) 4 of the NA_6 parental genetic map, directly adjacent to the corresponding RFLP locus (Faville *et al.* 2004; Figure 4.4A). Partial SNP haplotypic data for *Lp*ASRa2 revealed the maximum variant number of four, three of which are closely related, while the fourth is more divergent, defining two putative haplogroups (Olsen *et al.* 2004). Although the majority of the exon-located changes define synonymous amino acid changes (Figure 4.4B), two SNPs defined amino acid substitutions, one of which (glutamate to glutamine at coordinate 136) produces a radical change, and may be functionally significant. Alternatively, the characterized SNPs may be in LD with functionally significant changes in the transcriptional control regions (Paran and Zamir 2003), given haplotype stability over gene-length distances. Preliminary data from LD analysis of *Lp*ASRa2 and several full-length herbage quality genes using both $|D'|$and r^2 metrics suggests that LD blocks in perennial ryegrass genes are unlikely to extend further than 1–2 kb (Cogan *et al.* 2006).

Over 150 genes have been introduced into the *in vitro* SNP discovery for perennial ryegrass, of which 100 have been sequenced and aligned, with a total of 1,592 putative SNPs across 82 genes. SNuPe-validated SNPs were detected for 66 genes. Over a total of 87 kb of resequenced DNA, a relatively high SNP frequency (for four haplotypes) of 1 per 55 bp was observed, with higher incidence in intron compared to exon sequences, as anticipated. The validation rate from putative SNPs (predicted by alignment) to SnuPe-validated SNPs was c. 60%. Of the 40% of SNPs that were not validated, c. 15% were attributable to failed reactions. Reaction failures were probably due to LAP site mutations or reduced binding efficiency of the SnuPe interrogation primer, due to secondary structure within the primer, or primer site mutations due to SNP clustering. A second category (c. 25% of the total) arose from failure to detect segregation in the $F_1(NA_6 \times AU_6)$ family, despite the apparent presence of putative SNPs from the alignment process. The most likely explanations for this category are sequencing errors, although these are minimized by multiple cloned sequence determination, and identification of nonallelic SNPs through clustering of paralogous sequences from multigene families (Cogan *et al.* 2006).

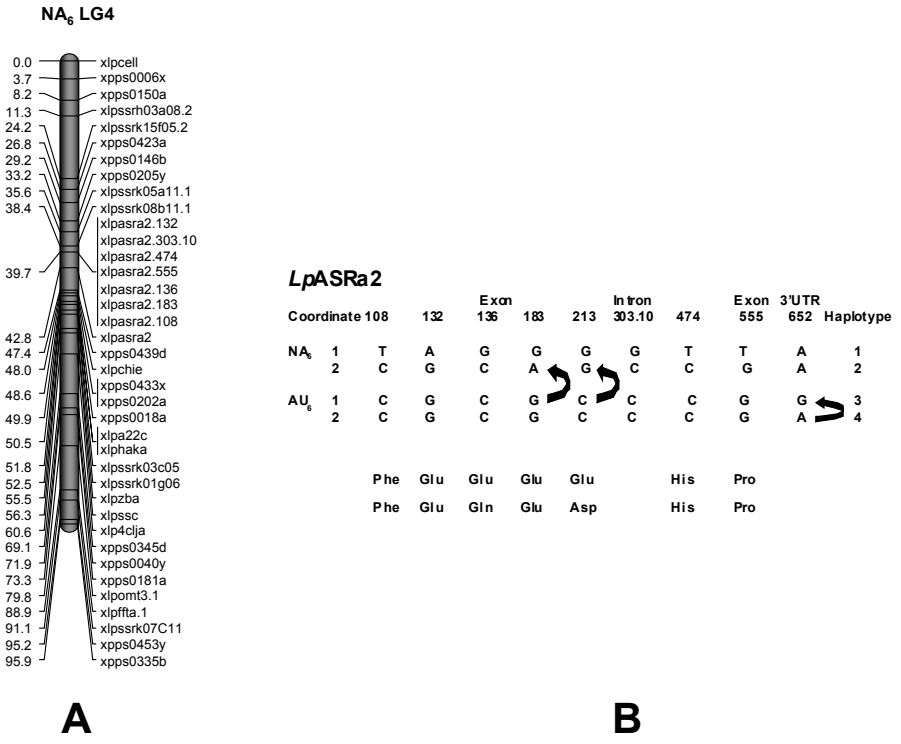


Figure 4.4. (A) Genetic linkage map of LG4 of the NA_6 parental map, showing the SNP loci (indicated as xlpasra2.coordinate number) in close linkage with the corresponding RFLP locus (xlpasra2). (B) *Lp*ASRa2 haplotype structures within and between the NA_6 and AU_6 parental genotypes. Arrows show putative mutational changes between members of the second haplogroup (haplotypes 2–4), and predicted translation products of exon-located SNP loci are indicated.

Although paralogous sequences may contribute to overestimation of SNP levels in perennial ryegrass when included in single alignments, a large number of template sequences clearly generated multiple amplicons that could be assembled into separate contigs for SNP identification. Validated allelic SNPs for different paralogous provide the means to distinguish and compare genomic locations between members of multigene families. Direct evidence for such effects has been provided by comparison of gene-derived SNP and RFLP loci. For instance, although the zinc transporter full-length cDNA *Lp*ZTa detected an RFLP locus on LG3 of the NA_6 parental map (Faville *et al.* 2004), a SNP locus derived from analysis of this gene was assigned to NA_6 LG1. The cDNA produced a complex Southern hybridization pattern consistent with a multigene family of 5–6 members, and multiple contigs were obtained following sequence alignment.

The complementary *in vitro* SNP discovery process for white clover is expected to be influenced by the allopolypoid genetic constitution of this species ($2n = 4x = 32$). The evolutionary origins of white clover are not fully understood, although two diploid species (*T. occidentale* D. Coombe and *T. pallescens* Schreber) are considered to be potential

progenitors (Badr *et al.* 2002; Chen and Gibson 1970a, b, 1971; Ellison *et al.* 2006). The structure of the EST–SSR based genetic map (Barrett *et al.* 2004) reveals homoeologous relationships between eight pairs of linkage groups. *In silico* alignment of SSR-containing EST sequences with whole genome sequence from model legume species has permitted comparisons between the subgenome structures of white clover and chromosome structure in barrel medic (*Medicago truncatula* Gaertn.).

Close to 50 white clover cDNAs have been selected from public databases, including the cyanogenesis-associated linimarase gene, *Tr*LIN, and from an unigene resource (Sawbridge *et al.* 2003b), including genes for flavonoid biosynthesis (relevant to bloat safety) and organic acid biosynthesis (relevant to aluminium tolerance and phosphorus acquisition). Two second generation two-way pseudo-testcross reference mapping families, designated $F_1(\text{Haifa}_2 \times \text{LCL}_2)$ and $F_1(\text{S184}_6 \times \text{LCL}_6)$, have provided parental DNA templates for *in vitro* SNP discovery. A large proportion of the analyzed genes produced multiple amplicons which could be assembled into separate contigs by application of high stringency alignment criteria in Sequencher™. A smaller proportion (< 25%) produced non overlapping contigs under moderately stringent conditions. Although intragenomic paralogy could account for this effect, in many instances homoeolocus amplification is probably responsible. For example, two distinct haplogroups obtained using the anthocyanidin reductase (banyuls) cDNA (*Tr*BANa) are differentiated by a large indel within an intron, in addition to multiple coding sequence differences (Figure 4.5). The assignment of allelic SNPs in each of putative homoeologues to the second-generation reference maps will permit clarification of these relationships.

4.5.1 *In silico* discovery of single nucleotide polymorphisms

Of the methods applied for the discovery of SNPs, the mining of sequence datasets should provide the cheapest source of abundant SNPs (Buetow *et al.* 1999; Gu *et al.* 1998; Picoult-Newberg *et al.* 1999; Taillon-Miller *et al.* 1998). Gene discovery and genome sequencing projects are increasingly considering SNP discovery in the selection of the starting material for nucleic acid extraction (Jander *et al.* 2002). With the development of high-throughput sequencing technology, large amounts of data have been submitted to the various DNA databases that may be suitable for data mining and SNP discovery. In particular, EST sequencing programs have provided a wealth of information, identifying novel genes from a broad range of organisms and providing an indication of gene expression level in particular tissues (Adams *et al.* 1995). EST sequence data may provide the richest source of biologically useful SNPs due to the relatively high redundancy of gene sequence, the diversity of genotypes represented within databases, and the fact that each SNP would be associated with an expressed gene (Picoult-Newberg *et al.* 1999). Candidate SNPs have been identified and validated from EST collections from a number of plant species including *Arabidopsis* (Schmid *et al.* 2003), barley (Kota *et al.* 2003), cassava (Lopez *et al.* 2005), melon (Morales *et al.* 2004), pine (Le Dantec *et al.* 2004), quinoa (Coles *et al.* 2005), tomato (Yang *et al.* 2004), and wheat (Somers *et al.* 2003). The continuing decrease in the cost of DNA sequencing is leading to a growing number of whole genome sequencing projects. This data increasingly enables the identification of SNPs in overlapping genomic sequence and also through comparison of genomic sequences with EST sequence data (Dawson *et al.* 2001; Jander *et al.* 2002; Taillon-Miller *et al.* 1998). Sequencing technologies continue

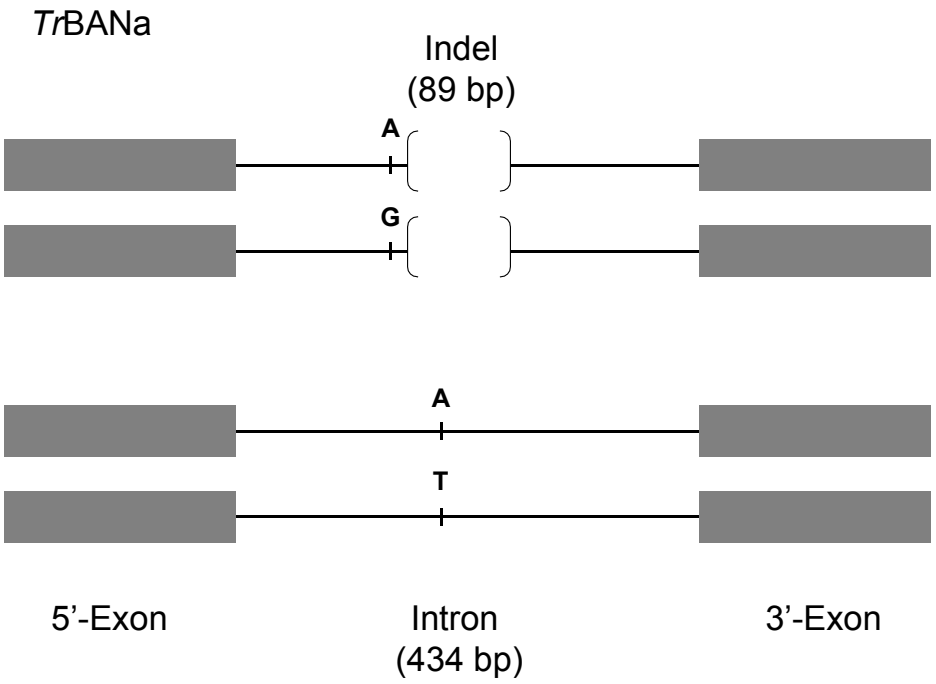


Figure 4.5. Schematic representation of putative homoeolocus structure in the white clover *Tr*BANa gene, including predicted allelic variation detected by *in vitro* SNP discovery.

to advance. However, high-throughput sequencing remains prone to inaccuracies as frequent as one base in every hundred. The false calling of these bases thereby hampers the electronic filtering of sequence data to identify potentially biologically relevant polymorphisms. The challenge of *in silico* SNP discovery is thus not the identification of polymorphic bases, but the differentiation of true SNP polymorphisms from the often more abundant sequence errors.

Several different sources of error need to be considered when differentiating between the sequence errors and true polymorphisms. Due to the automated nature of modern sequencing, human error is now rarely a factor. The principle source of sequence error is found in the automated reading of the raw chromatogram data. Here a balance exists between the desire to read as much sequence as possible and the confidence that bases are called correctly. Of the several algorithms available to call bases from chromatogram data, Phred is the most widely adopted standard (Ewing and Green 1998; Ewing *et al.* 1998). One benefit of this algorithm is that it provides a statistical estimate of the accuracy of calling each base and therefore provides a primary level of confidence that a sequence difference represents true genetic variation. There are several software packages which take advantage of this feature to estimate the confidence of sequence polymorphisms within alignments. PolyPhred integrates Phred base calling and peak information, within Phrap-generated sequence alignments (Green 1994), with alignments

viewed and marked for inspection using Consed (Gordon *et al.* 1998). More recently, this approach has been extended to include a Bayesian statistical method. PolyBayes (Marth *et al.* 1999) is a fully probabilistic SNP detection algorithm that calculates the probability that discrepancies at a given location of a multiple alignment represent true sequence variations as opposed to sequencing errors. The calculation takes into account the alignment depth, the base calls in each of the sequences, the associated base quality values (such as generated by the Phred trace analysis program or the Phrap fragment assembler), the base composition in the region, and the expected *a priori* polymorphism rate.

Where sequence trace files are available for the comparison of sequence trace files to filter out polymorphisms in traces of dubious quality, software such as PolyBayes and PolyPhred are the most efficient means to differentiate between true SNPs and sequence error. Unfortunately, complete sequence trace file archives are rarely available for large sequence datasets collated from a variety of sources. Furthermore, sequence quality-based SNP discovery does not identify errors in sequences which were incorporated prior to the base calling process. The principal cause of these prior errors is the inherently high error rate of the reverse transcription process required for the generation of cDNA libraries for EST sequencing. Similar errors are also inherent, though to a lesser extent, in any PCR amplification process that may be part of a sequencing protocol. In cases where trace files are unavailable, the identification of sequence errors can be based on two further methods to determine SNP confidence; redundancy of the polymorphism in an alignment, and co-segregation of SNPs with haplotype.

EST sequence datasets are most suited to redundancy-based SNP discovery. The highly redundant nature of EST datasets permits the selection of polymorphisms that occur multiple times within a set of aligned sequences. The frequency of occurrence of a polymorphism at a particular locus provides a measure of confidence in the SNP representing a true polymorphism and is referred to as the SNP redundancy score. By examining SNPs that have a redundancy score of two or greater, i.e., two or more of the aligned sequences represent the polymorphism, the vast majority of sequencing errors are removed. Although some true genetic variation is also ignored due to its presence only once within an alignment, the high degree of redundancy within the data permits the rapid identification of large numbers of SNPs without the requirement for sequence trace files.

While redundancy-based methods for SNP discovery are highly efficient, the nonrandom nature of sequence error may lead to certain sequence errors being repeated between runs due to conserved, complex DNA structures. Therefore, errors at these loci would have a relatively high SNP redundancy score and appear as confident SNPs. This source of error requires an additional method to differentiate them from true polymorphisms. A further measure of SNP confidence is based on haplotype co-segregation. While sequencing errors may occur at nonrandom positions within a sequencing read due to conserved sequence complexity, the probability of these errors being repeated between sequence reads remains random. True SNPs that represent divergence between homologous genes co-segregate to define a conserved haplotype, whereas nonrandom sequence errors do not co-segregate with haplotype. A co-segregation score based on the frequency of an SNP pattern occurring at multiple loci in an alignment allows ready identification of SNPs that do not co-segregate to define a haplotype. The SNP score and co-segregation score together provide a valuable means for estimating confidence in the validity of SNPs within aligned sequences independent of sequence trace files or the source of the sequence error.

Two methods currently apply a combination of redundancy and haplotype co-segregation; autoSNP (Barker *et al.* 2003; Batley *et al.* 2003), and SNPServer (Savage *et al.* 2005). AutoSNP combines sequence assembly such as D2 cluster (Burke et al. 1999), CAP3 (Huang and Madan 1999) or TGICL (Pertea *et al.* 2003) with redundancy-based SNP discovery and haplotype co-segregation scoring. A more recent application, SNPServer provides a real-time internet-based SNP discovery service. Sequences may be submitted for assembly with CAP3 or be submitted pre-assembled in ACE format. Alternatively, a single sequence may be submitted for BLAST comparison with a sequence database. Identified sequences are then processed for assembly with CAP3 and subsequent redundancy-based SNP discovery. SNPServer has an advantage in being the only real-time web-based software, which allows users to rapidly identify SNPs in sequences of interest using public data. Of the three methods for *in silico* SNP discovery, trace quality, redundancy, and haplotype co-segregation, none have yet been combined into a single software tool.

One approach that has yet to be applied for *in silico* SNP discovery is the use of comparative species SNP identification. Since the generation of SNPs is not completely random and the retention of SNPs within populations is subject to evolutionary pressure, SNPs are more likely to be observed at some positions within genes than others. An understanding of SNP position frequency within one species may therefore assist in the prediction of SNPs within similar genes from other species.

4.6 CONCLUSION

There are several approaches that may be undertaken for the discovery of SNPs in plant species. The method applied would be dependent on several factors, including the expected application of the discovered SNPs, the availability of gene or genome sequence and the availability of computational tools or laboratory facilities. Where only limited DNA sequence is available or large numbers of validated SNPs are required within a limited number of specific genes, an *in vitro* approach would be favored. Where large numbers of SNPs are required across a genome and a significant quantity of sequence data was available, an *in silico* approach may be more appropriate. Two factors are likely to influence SNP discovery in the future. These are the increasing ability to produce gene and genome sequence data at an ever reducing cost and the development of massive throughput genotyping systems for the assessment of tens of thousands of SNPs across thousands of genotypes. These applications can be applied for SNP discovery and validation within specific genes as well as defining global SNP frequencies across populations.

4.7 REFERENCES

Adams, M.D., Kerlavage, A.R., Fleischmann, R.D., Fuldner, R.A., Bult, C.J., Lee, N.H., Kirkness, E.F., Weinstock, K.G., Gocayne, J.D., White, O., Sutton, G., Blake, J.A., Brandon, R.C., Chiu, M.-W., Clayton, R.A., Cline, R.T., Cotton, M.D., Earle-Hughes, J., Fine, L.D., FitzGerald, L.M., FitzHugh, W.M., Fritchman, J.L., Geoghagen, N.S.M., Glodek, A., Gnehm, C.L., Hanna, M.C., Hedblom, E., Hinkle Jr., P.S., Kelley, J.M., Klimek, K.M., Kelley, J.C., Liu, L.-I., Marmaros, S.M., Merrick, J.M., Moreno-Palanques, R.F., McDonald, L.A., Nguyen, D.T., Pellegrino, S.M., Phillips, C.A., Ryder, S.E., Scott, J.L.,

Saudek, D.M., Shirley, R., Small, K.V., Spriggs, T.A., Utterback, T.R., Weidman, J.F., Li, Y., Barthlow, R., Bednarik, D.P., Cao, L., Cepeda, M.A., Coleman, T.A., Collins, E.-J., Dimke, D., Feng, P., Ferrie, A., Fischer, C., Hastings, G.A., He, W.-W., Hu, J.-S., Huddleston, K.A., Greene, J.M., Gruber, J., Hudson, P., Kim, A., Kozak, D.L., Kunsch, C., Ji, H., Li, H., Meissner, P.S., Olsen, H., Raymond, L., Wei, Y.-F., Wing, J., Xu, C., Yu, G.-L., Ruben, S.M., Dillon, P.J., Fannon, M.R., Rosen, C.A., Haseltine, W.A., Fields, C., Fraser, C.M., Venter, J.C., 1995, Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. Nature 377:3–17.

Ahmadian, A., Gharizadeh, B., Gustafsson, A.C., Sterky, F., Nyren, P., Uhlen, M., Lundeberg, J., 2000, Single-nucleotide polymorphism analysis by pyrosequencing. Analytical Biochemistry 280:103–110.

Andersen, J.R., Lübberstedt, T., 2003, Functional markers in plants. Trends in Plant Science 8:554–560.

Badr, A., Sayed-Ahmed, H., El-Shanshouri, A., Watson, L.E., 2002, Ancestors of white clover (*Trifolium repens* L.), as revealed by isozyme polymorphisms. Theoretical and Applied Genetics 106:143–148.

Barker, G., Batley, J., O'Sullivan, H., Edwards, K.J., Edwards, D., 2003, Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP. Bioinformatics 19:421–422.

Barrett, B., Griffiths, A., Schreiber, M., Ellison, N., Mercer, C., Bouton, J., Ong, B., Forster, J., Sawbridge, T., Spangenberg, G., Bryan, G., Woodfield, D., 2004, A microsatellite map of white clover. Theoretical and Applied Genetics 109:596–608.

Batley, J., Barker, G., O'Sullivan, H., Edwards, K.J., Edwards, D., 2003, Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. Plant Physiology 132:84–91.

Bhattramakki, D., Rafalski, A., 2001, Discovery and application of single nucleotide polymorphism markers in plants. In: R.J. Henry (ed) Plant Genotyping: The DNA Fingerprinting of Plants. CABI Publishing, Wallingford, Oxon, UK, pp 179–193.

Bhattramakki, D., Dolan, M., Hanafey, M., Wineland, R., Vaske, D., Register III, J.C., Tingey, S.V., Rafalski, A., 2002, Insertion–deletion polymorphisms in 3′ regions of maize genes occur frequently and can be used as highly informative genetic markers. Plant Molecular Biology 48:539–547.

Borevitz, J.O., Liang, D., Plouffe, D., Chang, H.S., Zhu, T., Weigel, D., Berry, C.C., Winzeler, E., Chory, J., 2003, Large-scale identification of single-feature polymorphisms in complex genomes. Genome Research 13:513–523.

Botstein, D., White, R.L., Skolnick, M., Davis, R.W., 1980, Construction of a genetic linkage map in man using restriction fragment length polymorphisms. American Journal of Human Genetics 32:314–331.

Bradbury, L.M.T., Fitzgerald, T.L., Henry, R.J., Jin, Q., Walters, D.E., 2005, The gene for fragrance in rice. Plant Biotechnology Journal 3:363–370.

Brown, G.R., Kadel, E.E., Bassoni, D.L., Kiehne, K.L., Temesgen, B., van Buijtenen, J.P., Sewell, M.M., Marshall, K.A., Neale, D.B., 2001, Anchored reference loci in loblolly pine (*Pinus taeda* L.) for integrating pine genomics. Genetics 159:799–809.

Brown, G.R., Gill, G.P., Kuntz, R.J., Langley, C.H., Neale, D.B., 2004, Nucleotide diversity and linkage disequilibrium in loblolly pine. Proceedings of the National Academy of Sciences of the United States of America 101:15255–15260.

Bryan, G.J., Stephenson, P., Collins, A., Kirby, J., Smith, J.B., Gale, M.D., 1999, Low levels of DNA sequence variation among adapted genotypes of hexaploid wheat. Theoretical and Applied Genetics 99:192–198.

Buetow, K.H., Edmonson, M.N., Cassidy, A.B., 1999, Reliable identification of large numbers of candidate SNPs from public EST data. Nature Genetics 21:323–325.

Bundock, P.C., Henry, R.J., 2004, Single nucleotide polymorphism, haplotype diversity and recombination in the Isa gene of barley. Theoretical and Applied Genetics 109:543–551.

Bundock, P.C., Christopher, J.T., Eggler, P., Ablett, G., Henry, R.J., Holton, T.A., 2003, Single nucleotide polymorphisms in cytochrome P450 genes from barley. Theoretical and Applied Genetics 106:676–682.

Burke, J., Davison, D., Hide, W., 1999, d2-cluster: a validated method for clustering EST and full-length cDNA sequences. Genome Research 9:1135–1142.

Caldwell, K.S., Dvorak, J., Lagudah, E.S., Akhunov, E., Luo, M.-C., Wolters, P., Powell, W., 2004, Sequence polymorphism in polyploid wheat and their D-genome diploid ancestor. Genetics 167:941–947.

Chen, C., Gibson, P.B., 1970a, Chromosome pairing in two interspecific hybrids of *Trifolium*. Canadian Journal of Genetics and Cytology 12:790–794.

Chen, C., Gibson, P.B., 1970b, Meiosis in two species of *Trifolium* and their hybrids. Crop Science 10:188–189.

Chen, C., Gibson, P.B., 1971, Karyotypes of fifteen *Trifolium* species in section Amoria. Crop Science 11:441–445.

Ching, A., Caldwell, K.S., Jung, M., Dolan, M., Smith, O.S., Tingey, S., Morgante, M., Rafalski, A.J., 2002, SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. BMC Genetics [electronic resource] 3:19.

Cogan, N.O.I., Ponting, R.C., Vecchies, A.C., Drayton, M.C., George, J., Dobrowolski, M.P., Sawbridge, T.I., Spangenberg, G.C., Smith, K.F., Forster, J.W., 2006, Gene-associated single nucleotide polymorphism (SNP) discovery in perennial ryegrass (*Lolium perenne L.*) Mol Genet Genomics 276:101-12.

Coles, N.D., Coleman, C.E., Christensen, S.A., Jellen, E.N., Stevens, M.R., Bonifacio, A., Rojas-Beltran, J.A., Fairbanks, D.J., Maughan, P.J., 2005, Development and use of an expressed sequenced tag library in quinoa (*Chenopodium quinoa* Willd.) for the discovery of single nucleotide polymorphisms. Plant Science 168:439–447.

Comai, L., Young, K., Till, B.J., Reynolds, S.H., Greene, E.A., Codomo, C.A., Enns, L.C., Johnson, J.E., Burtner, C., Odden, A.R., Henikoff, S., 2004, Efficient discovery of DNA polymorphisms in natural populations by Ecotilling. Plant Journal 37:778–786.

Coryell, V.H., Jessen, H., Schupp, J.M., Webb, D., Keim, P., 1999, Allele-specific hybridization markers for soybean. Theoretical and Applied Genetics 98:690–696.

Dawson, E., Chen, Y., Hunt, S., Smink, L.J., Hunt, A., Rice, K., Livingston, S., Bumpstead, S., Bruskiewich, R., Sham, P., Ganske, R., Adams, M., Kawasaki, K., Shimizu, N., Minoshima, S., Roe, B., Bentley, D., Dunham, I., 2001, A SNP resource for human chromosome 22: extracting dense clusters of SNPs from the genomic sequence. Genome Research 11:170–178.

Duda Jr., T.F., Palumbi, S.R., 1999, Population structure of the black tiger prawn, *Penaeus monodon*, among western Indian Ocean and western Pacific populations. Marine Biology 134:705–710.

Ellison, N.W., Liston, A., Szeimer, J.J., Williams, W.M., Taylor, W.L., 2006, Molecular phylogenetics of the clover genus (*Trifolium*-Leguminosae). Molecular Phylogenetics and Evolution 39: 688-705.

Etienne, C., Rotham, C., Moing, A., Plomion, C., Bodenes, C., Svanella-Dumas, L., Cosson, P., Pronier, V., Monet, R., Dirlewanger, E., 2002, Candidate genes and QTLs for sugar and organic acid content in peach (*Prunus persica* L. Batsch). Theoretical and Applied Genetics 105(1):145–159.

Ewing, B., Green, P., 1998, Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Research 8:186–194.

Ewing, B., Hillier, L., Wendl, M.C., Green, P., 1998, Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Research 8:175–185.

Faville, M.J., Vecchies, A.C., Schreiber, M., Drayton, M.C., Hughes, L.J., Jones, E.S., Guthridge, K.M., Smith, K.F., Sawbridge, T., Spangenberg, G.C., Bryan, G.T., Forster, J.W., 2004, Functionally associated molecular genetic marker map construction in perennial ryegrass (*Lolium perenne* L.). Theoretical and Applied Genetics 110:12–32.

Faville, M., Vecchies, A.C., Schreiber, M., Drayton, M.C., Hughes, L.J., Jones, E.S., Guthridge, K.M., Smith, K.F., Sawbridge, T., Spangenberg, G.C., Bryan, G.T., Forster, J.W., 2005, Candidate gene-based molecular marker map construction in perennial ryegrass (*Lolium perenne* L.). Theoretical and Applied Genetics 110:12–32.

Forster, J.W., Jones, E.S., Batley, J., Smith, K.F., 2004, Molecular marker-based genetic analysis of pasture and turf grasses. In: A. Hopkins, Z.-Y. Wang, M. Sledge, R.E. Barker (eds) Molecular Breeding of Forage and Turf. Kluwer, Dordecht, pp 197–239.

Gaut, B.S., Clegg, M.T., 1993, Nucleotide polymorphism in the Adh1 locus of pearl millet (*Pennisetum glaucum*) (Poaceae). Genetics 135:1091–1097.

Germano, J., Klein, A.S., 1999, Species-specific nuclear and chloroplast single nucleotide polymorphisms to distinguish *Picea glauca*, *P. mariana* and *P. rubens*. Theoretical and Applied Genetics 99:37–49.

Gilchrist, E.J., Haughn, G.W., 2005, TILLING without a plough: a new method with applications for reverse genetics. Current Opinion in Plant Biology 8:1–5.

Gill, G.P., Brown, G.R., Neale, D.B., 2003, A sequence mutation in the cinnamyl alcohol dehydrogenase gene associated with altered lignification in loblolly pine. Plant Biotechnology Journal 1:253–258.

Giordano, M., Oefner, P.J., Underhill, P.A.., Sforza, L.L.C., Tosi, R., Richiardi, P.M., 1999, Identification by denaturing high-performance liquid chromatography of numerous polymorphisms in a candidate region for multiple sclerosis susceptibility. Genomics 56:247–253.

Gordon, D., Abajian, C., Green, P., 1998, Consed: a graphical tool for sequence finishing. Genome Research 8:195–202.

Green, P., 1994, Phrap. Unpublished data (www.phrap.org).

Greene, E.A., Codomo, C.A., Taylor, N.E., Henikoff, J.G., Till, B.J., Reynolds, S.H., Enns, L.C., Burtner, C., Johnson, J.E., Odden, A.R., Comai, L., Henikoff, S., 2003, Spectrum of chemically induced mutations from a large-scale reverse-genetic screen in *Arabidopsis*. Genetics 164:731–740.

Gu, Z., Hillier, L., Kwok, P.-Y., 1998, Single nucleotide polymorphism hunting in cyberspace. Human Mutation 12:221–225.

Hauser, M.T., Adhami, F., Dorner, M., Fuchs, E., Glossl, J., 1998, Generation of co-dominant PCR-based markers by duplex analysis on high resolution gels. Plant Journal 16:117–125.

Hayashi, K., Hashimoto, N., Daigen, M., Ashikawa, I., 2004, Development of PCR-based SNP markers for rice blast resistance genes at the Piz locus. Theoretical and Applied Genetics 108:1212–1220.

Hebenbrock, K., Williams, P.M., Karger, B.L., 1995, Single strand conformational polymorphism using capillary electrophoresis with two-dye laser-induced fluorescence detection. Electrophoresis 16:1429–1436.

Hongtrakul, V., Slabaugh, M.B., Knapp, S.J., 1998, DFLP, SSCP, and SSR markers for △9-stearoyl-acyl carrier protein desaturases strongly expressed in developing seeds of sunflower: intron lengths are polymorphic among elite inbred lines. Molecular Breeding 4:195–203.

Hsia, A.-P., Wen, T.-J., Chen, H.D., Liu, Z., Yandeau-Nelson, M.D., Wei, Y., Guo, L., Schnable, P.S., 2005, Temperature gradient capillary electrophoresis (TGCE) – a tool for the high-throughput discovery and mapping of SNPs and IDPs. Theoretical and Applied Genetics 111:218–225.

Huang, X., Madan, A., 1999, CAP3: a DNA sequence assembly program. Genome Research 9:868–877.

Inazuka, M., Wenz, H.M., Sakabe, M., Tahira, T., Hayashi, K., 1997, A streamlined mutation detection system: multicolor post-PCR fluorescence labeling and single-strand conformational polymorphism analysis by capillary electrophoresis. Genome Research 7:1094–1103.

Ingvarsson, P.K., 2005, Nucleotide polymorphism and linkage disequilibrium within and among natural populations of European aspen (*Populus tremula* L., salicaceae). Genetics 169:945–953.

Jander, G., Norris, S.R., Rounsley, S.D., Bush, D.F., Levin, I.M., Last, R.L., 2002, *Arabidopsis* map-based cloning in the post-genome era. Plant Physiology 129:440–450.

Järvinen, P., Lemmetyinen, J., Savolainen, O., Sopanen, T., 2003, DNA sequence variation in *BpMADS2* gene in two populations of *Betula pendula*. Molecular Ecology 12:369–384.

Jeanneau, M., Gerentes, D., Foueillassar, X., Zivy, M., Vidal, J., Toppan, A., Perez, P., 2002, Improvement of drought tolerance in maize: towards the functional validation of the *Zm-Asr1* gene and increase of water use efficiency by over-expressing C4-PEPC. Biochimie 84:1127–1135.

Jin, Q., Waters, D., Cordeiro, G.M., Henry, R.J., Reinke, R.F., 2003, A single nucleotide polymorphism (SNP) marker linked to the fragrance gene in rice (*Oryza sativa* L.). Plant Science 165:359–364.

Jones, E.S., Dupal, M.P., Dumsday, J.L., Hughes, L.J., Forster, J.W., 2002a, An SSR-based genetic linkage map for perennial ryegrass (*Lolium perenne* L.). Theoretical and Applied Genetics 105:577–584.

Jones, E.S., Mahoney, N.L., Hayward, M.D., Armstead, I.P., Jones, J.G., Humphreys, M.O., King, I.P., Kishida, T., Yamada, T., Balfourier, F., Charmet, G., Forster, J.W., 2002b, An enhanced molecular marker based genetic map of perennial ryegrass (*Lolium perenne*) reveals comparative relationships with other Poaceae genomes. Genome 45:282–295.

Jones, E.S., Hughes, L.J., Drayton, M.C., Abberton, M.T., Michaelson-Yeates, T.P.T., Bowen, C., Forster, J.W., 2003, An SSR and AFLP molecular marker-based genetic map of white clover (*Trifolium repens* L.). Plant Science 165:531–539.

Kado, T., Yoshimaru, H., Tsumura, Y., Tachida, H., 2003, DNA variation in a conifer, *Cryptomeria japonica* (Cupressaceae sensu lato). Genetics 164:1547–1559.

Keim, P., Diers, B.W., Olson, T.C., Shoemaker, R.C., 1990, RFLP mapping in soybean: association between marker loci and variation in quantitative traits. Genetics 126:735–742.

Konieczny, A., Ausubel, F.M., 1993, A procedure for mapping *Arabidopsis* mutations using co-dominant ecotype-specific PCR-based markers. Plant Journal 4:403–410.

Kota, R., Wolf, M., Michalek, W., Graner, A., 2001, Application of denaturing high-performance liquid chromatography for mapping of single nucleotide polymorphisms in barley (*Hordeum vulgare* L.). Genome 44:523–528.

Kota, R., Rudd, S., Facius, A., Kolesov, G., Thiel, T., Zhang, H., Stein, N., Mayer, K., Graner, A., 2003, Snipping polymorphisms from large EST collections in barley (*Hordeum vulgare* L.). Molecular Genetics and Genomics 270:24–33.

Kuhn, D.N., Borrone, J., Meerow, A.W., Motamayor, J.C., Brown, J.S., Schnell, R.J., 2005, Single-strand conformation polymorphism analysis of candidate genes for reliable identification of alleles by capillary array electrophoresis. Electrophoresis 26:112–125.

LaForest, S.M., Prestwich, G.D., Löfstedt, C., 1999, Intraspecific nucleotide variation at the pheromone binding protein locus in the turnip moth, *Agrotis segetum*. Insect Molecular Biology 8:481–490.

Lau, E., Leushner, J., Patnaik, M., 2000, Automated detection of the factor V Leiden mutation using MALDI-TOF mass spectrometry on the MassARRAY system. Clinical Chemistry 46:1880.

Le Dantec, L.L., Chagné, D., Pot, D., Cantin, O., Garnier-Géré, P., Bedon, F., Frigerio, J.-M., Chaumeil, P., Léger, P., Garcia, V., Laigret, F., De Daruvar, A., Plomion, C., 2004, Automated SNP detection in expressed sequence tags: statistical considerations and application to maritime pine sequences. Plant Molecular Biology 54:461–470.

Lopez, C., Piégu, B., Cooke, R., Delseny, M., Tohme, J., Verdier, V., 2005, Using cDNA and genomic sequences as tools to develop SNP strategies in cassava (*Manihot esculenta* Crantz). Theoretical and Applied Genetics 110:425–431.

Lu, C.M., Yang, W.Y., Zhang, W.J., Lu, B.-R., 2005, Identification of SNPs and development of allelic specific PCR markers for high molecular weight glutenin subunit Dt×1.5 from *Aegilops tauschii* through sequence characterization. Journal of Cereal Science 41:13–18.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.-J., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L.I., Jarvie, T.P., Jirage, K.B., Kim, J.-B., Knight, J.R., Lanza, J.R., Leamon, J.H., Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A., Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F., Rothberg, J.M., 2005, Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380.

Marth, G.T., Korf, I., Yandell, M.D., Yeh, R.T., Gu, Z., Zakeri, H., Stitziel, N.O., Hillier, L., Kwok, P.-Y., Gish, W.R., 1999, A general approach to single-nucleotide polymorphism discovery. Nature Genetics 23:452–456.

Martins-Lopes, P., Zhang, H., Koebner, R., 2001, Detection of single nucleotide mutations in wheat using single strand conformation polymorphism gels. Plant Molecular Biology Reporter 19:159–162.

Mattocks, C., White, H.E., Owen, N., Durston, V.J., Harvey, J.F., Cross, N.C.P., 2004, An evaluation of the MassCLEAVE(tm) biochemistry for diagnostic screening. Journal of Medical Genetics 41:S75.

McCallum, J., Leite, D., Pither-Joyce, M., Havey, M.J., 2001, Expressed sequence markers for genetic analysis of bulb onion (*Allium cepa* L.). Theoretical and Applied Genetics 103:979–991.

Mogg, R., Batley, J., Hanley, S., Edwards, D., O'Sullivan, H., Edwards, K.J., 2002, Characterization of the flanking regions of *Zea mays* microsatellites reveals a large number of useful sequence polymorphisms. Theoretical and Applied Genetics 105:532–543.

Morales, M., Roig, E., Monforte, A.J., Arús, P., Garcia-Mas, J., 2004, Single-nucleotide polymorphisms detected in expressed sequence tags of melon (*Cucumis melo* L.). Genome 47:352–360.

Myers, R.M., Sheffield, V.C., Cox, D.R., 1988, Detection of single base changes in DNA: ribonuclease cleavage and denaturing gradient gel electrophoresis. In: K.E. Davies (ed) Genome Analysis: A Practical Approach. IRL, Oxford, pp 95–139.

Neff, M.M., Neff, J.D., Chory, J., Pepper, A.E., 1998, dCAPS, a simple technique for the genetic analysis of single nucleotide polymorphisms: experimental applications in *Arabidopsis thaliana* genetics. Plant Journal 14:387–392.

O'Donnell, M.J., Little, D.P., Braun, A., 1997, MassArray as an enabling technology for the industrial-scale analysis of DNA. Genetic Engineering News 17:39.

Oleykowski, C.A., Mullins, C.R.B., Godwin, A.K., Yeung, A.T., 1998, Mutation detection using a novel plant endonuclease. Nucleic Acids Research 26:4597–4602.

Olsen, K.M., Halldorsdottir, S.S., Stinchcombe, J.R., Weinig, C., Schmittt, J., Purugganan, M.D., 2004, Linkage disequilibrium mapping of *Arabidopsis CRY2* flowering time alleles. Genetics 167:1361–1369.

Orita, M., Suzuki, Y., Sekiya, T., Hayashi, K., 1989, Rapid and sensitive detection of point mutations and SNA polymorphisms using the polymerase chain reaction. Genomics 5:874–879.

Palumbi, S.R., Baker, C.S., 1994, Contrasting population structure from nuclear intron sequences and mtDNA of humpback whales. Molecular Biology and Evolution 11:426–435.

Paran, I., Zamir, D., 2003, Quantitative traits in plants: beyond the QTL. Trends in Genetics 19:303–306.

Perry, J.A., Wang, T.L., Welham, T.J., Gardner, S., Pike, J.M., Yoshida, S., Parniske, M., 2003, A TILLING reverse genetics tool and a web-accessible collection of mutants of the legume *Lotus japonicus*. Plant Physiology 131:866–871.

Pertea, G., Huang, X., Liang, F., Antonescu, V., Sultana, R., Karamycheva, S., Lee, Y., White, J., Cheung, F., Parvizi, B., Tsai, J., Quackenbush, J., 2003, TIGR gene indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics 19:651–652.

Picoult-Newberg, L., Ideker, T.E., Pohl, M.G., Taylor, S.L., Donaldson, M.A., Nickerson, D.A., Boyce-Jacino, M., 1999, Mining SNPs from EST databases. Genome Research 9:167–174.

Plomion, C., Frigerio, J.-M., Ridolfi, M., Pot, D., Pionneau, C., Bodénes, C., Kremer, A., Hurme, P., Savolainen, O., Avila, C., Gallardo, F., Canovas, F.M., David, H., Neutelings, G., Campbell, M., 1999, Developing SSCP markers in two *Pinus* species. Molecular Breeding 5:21–31.

Pot, D., McMillan, L., Echt, C., Le Provost, G., Garnier-Géré, P., Cato, S., Plomion, C., 2005, Nucleotide variation in genes involved in wood formation in two pine species. New Phytologist 167:101–112.

Rickert, A.M., Premstaller, A., Gebhardt, C., Oefner, P.J., 2002, Genotyping of SNPs in a polyploid genome by pyrosequencing (TM). Biotechniques 32(3):592–593.

Rickert, A.M., Jeong, J.H., Meyer, S., Nagel, A., Ballvora, A., Oefner, P.J., Gebhardt, C., 2003, First generation SNP/InDel markers tagging loci for pathogen resistance in the potato genome. Plant Biotechnology Journal 1:399–410.

Rodi, C.P., Storm, N., Darnhofer-Patel, B., Hartmer, R., Leppin, L., Bocker, S., Denissenko, M., van den Boom, D., 2002, MassARRAY (TM) analysis of fragmented nucleic acids: applications in typing, sequence validation, and targeted SNP discovery. European Journal of Human Genetics 10:299.

Russell, J., Booth, A., Fuller, J., Harrower, B., Hedley, P., Machray, G., Powell, W., 2004, A comparison of sequence-based polymorphism and haplotype content in transcribed and anonymous regions of the barley genome. Genome 47:389–398.

Sato, Y., Nishio, T., 2003, Mutation detection in rice waxy mutants by PCR–RF–SSCP. Theoretical and Applied Genetics 107:560–567.

Savage, D., Batley, J., Erwin, T., Logan, E., Love, C.G., Lim, G.A., Mongin, E., Barker, G., Spangenberg, G.C., Edwards, D., 2005, SNPServer: a real-time SNP discovery tool. Nucleic Acids Research 33:W493–W495.

Savolainen, O., Langley, C.H., Lazzaro, B.P., Fréville, H., 2000, Contrasting patterns of nucleotide polymorphism at the alcohol dehydrogenase locus in the outcrossing *Arabidopsis lyrata* and the selfing *Arabidopsis thaliana*. Molecular Biology and Evolution 17:645–655.

Sawbridge, T., Ong, E.-K., Binnion, C., Emmerling, M., McInnes, R., Meath, K., Nguyen, N., Nunan, K., O'Neill, M., O'Toole, F., Rhodes, C., Simmonds, J., Tian, P., Wearne, K., Webster, T., Winkworth, A., Spangenberg, G., 2003a, Generation and analysis of expressed sequence tags in perennial ryegrass (*Lolium perenne* L.). Plant Science 165:1089–1100.

Sawbridge, T., Ong, E.-K., Binnion, C., Emmerling, M., Meath, K., Nunan, K., O'Neill, M., O'Toole, F., Simmonds, J., Wearne, K., Winkworth, A., Spangenberg, G., 2003b, Generation and analysis of expressed sequence tags in white clover (*Trifolium repens* L.). Plant Science 165:1077–1087.

Schmid, K.J., Sörensen, T.R., Stracke, R., Törjek, O., Altmann, T., Mitchell-Olds, T., Weisshaar, B., 2003, Large-scale identification and analysis of genome-wide single-nucleotide polymorphisms for mapping in *Arabidopsis thaliana*. Genome Research 13:1250–1257.

Selinger, D.A., Chandler, V.L., 1999, Major recent and independent changes in levels and patterns of expression have occurred at the b gene, a regulatory locus in maize. Proceedings of the National Academy of Sciences of the United States of America 96:15007–15012.

Shattuck-Eidens, D.M., Bell, R.N., Neuhausen, S.L., Helentjaris, T., 1990, DNA sequence variation within maize and melon: observations from polymerase chain reaction amplification and direct sequencing. Genetics 126:207–217.

Simko, I., Haynes, K.G., Ewing, E.E., Costanzo, S., Christ, B.J., Jones, R.W., 2004, Mapping genes for resistance to Verticillium albo-atrum in tetraploid and diploid potato populations using haplotype association tests and genetic linkage analysis. Molecular Genetics and Genomics 271:522–531.

Slade, A.J., Fuerstenberg, S.I., Loeffler, D., Steine, M.N., Facciotti, D., 2005, A reverse genetic, nontransgenic approach to wheat crop improvement by TILLING. Nature Biotechnology 23:75–81.

Somers, D.J., Kirkpatrick, R., Moniwa, M., Walsh, A., 2003, Mining single-nucleotide polymorphisms from hexaploid wheat ESTs. Genome 46:431–437.

Taillon-Miller, P., Gu, Z., Li, Q., Hillier, L., Kwok, P.-Y., 1998, Overlapping genomic sequences: a treasure trove of single-nucleotide polymorphisms. Genome Research 8:748–754.

Tang, J., Unnasch, T.R., 1995, Discriminating PCR artifacts using directed heteroduplex analysis (DHDA). Biotechniques 19:902–905.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994, CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research 22:4673–4680.

Till, B.J., Reynolds, S.H., Greene, E.A., Codomo, C.A., Enns, L.C., Johnson, J.E., Burtner, C., Odden, A.R., Young, K., Taylor, N.E., Henikoff, J.G., Comai, L., Henikoff, S., 2003, Large-scale discovery of induced point mutations with high-throughput TILLING. Genome Research 13:524–530.

Till, B.J., Burtner, C., Comai, L., Henikoff, S., 2004, Mismatch cleavage by single-strand specific nucleases. Nucleic Acids Research 32:2632–2641.

Vaidyanathan, R., Kuruvilla, S., Thomas, G., 1999, Characterisation and expression pattern of an abscissic acid and osmotic stress responsive gene from rice. Plant Science 140:25–36.

Wang, D.G., Fan, J.-B., Siao, C.-J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., Kruglyak, L., Stein, L., Hsie, L., Topaloglou, T., Hubbell, E., Robinson, E., Mittmann, M., Morris, M.S., Shen, N., Kilburn, D., Rioux, J., Nusbaum, C,, Rozen, S., Hudson, T.J., Lipshutz, R., Chee, M., Lander, E.S., 1998, Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 280:1077–1082.

Wang, R.-L., Stec, A., Hey, J., Lukens, L., Doebley, J., 1999, The limits of selection during maize domestication. Nature 398:236–239.

Yang, W., Bai, X., Kabelka, E., Eaton, C., Kamoun, S., Van Der Knaap, E., Francis, D., 2004, Discovery of single nucleotide polymorphisms in *Lycopersicon esculentum* by computer aided analysis of expressed sequence tags. Molecular Breeding 14:21–34.

Zhang, D.X., Hewitt, G.M., 2003, Erratum: nuclear DNA analyses in genetic studies of populations: practice, problems and prospects (Molecular Ecology (2003) 12:563–584). Molecular Ecology 12:1687.

Zhang, W., Gianibelli, M.C., Ma, W., Rampling, L., Gale, K.R., 2003, Identification of SNPs and development of allele-specific PCR markers for gliadin alleles in *Triticum aestivum*. Theoretical and Applied Genetics 107:130–138.

Zhu, Y.L., Song, Q.J., Hyten, D.L., Van Tassell, C.P., Matukumalli, L.K., Grimm, D.R., Hyatt, S.M., Fickus, E.W., Young, N.D., Cregan, P.B., 2003, Single-nucleotide polymorphisms in soybean. Genetics 163:1123–1134.

Chapter 5

SINGLE NUCLEOTIDE POLYMORPHISMS GENOTYPING IN PLANTS

David Chagné[1], Jacqueline Batley[2], David Edwards[2], and John W. Forster[2]

5.1 INTRODUCTION

Single nucleotide polymorphism (SNP) markers are highly abundant in the genomes of the majority of organisms, including plants. They provide valuable markers for the study of agronomic or adaptive traits in plant species, using strategies based on genetic mapping or association genetics studies. The development of SNP markers usually follow a three-part progression consisting, chronologically, of SNP discovery based on analysis of a small set of individuals, validation in a larger set (i.e., to remove false positives due to sequencing errors or due to the presence of homeologous/paralogous sequences) and then genotyping in a large population. The present chapter will focus on methods that are applicable to large-scale SNP genotyping studies.

Syvänen (2001) and Kwok (2001) are among the most recent authors to publish complete reviews on SNP-genotyping techniques. In addition, a number of methods that use advanced technology (Invader assay, Pyrosequencing, Illumina fiber-optic array technology linked to bead-immobilized GoldenGate™ PCR technology, and Sequenom MALDI-TOF MS MassExtend™ technology) are described in detail in a recent book chapter (Kahl *et al.* 2005). Those reviews cover the different methods that can be employed for scoring SNPs, such as allele-specific oligonucleotide (ASO) hybridization, oligonucleotide ligation, single nucleotide primer extension, and enzymatic cleavage. Those methods are commonly used in combination with SNP detection technology platforms such as gel electrophoresis systems, fluorescent plate readers, flow cytometry, mass spectrometry, or oligonucleotide-based microarrays. Even with the latest technical advances that have occurred since these reviews were published, these methods still provide the core methodologies for SNP genotyping, in particular for plant association studies. The following chapter will not attempt to provide another fully comprehensive review, but will aim to describe the key features of the major technologies and attempt to analyze the requirements of the SNP scoring methods that can be used in plants. A range

[1] HortResearch, Plant Gene Mapping group, Private Bag 11030, Palmerston North, New Zealand
[2] Primary Industries Research Victoria, Victorian AgriBiosciences Centre, La Trobe R&D Park, Bundoora, VIC 3083, Australia

of high-throughput methods are currently being developed, in particular for model species such as humans (Syvänen 2005), but no example of rapid methods for very high-density whole-genome genotyping (WGG; i.e., at the level of millions of SNP data points) has been reported so far, providing one of the greatest challenges for the present generation of geneticists (and their collaborators in technology development) who wish to associate genotype and phenotype in their species of interest. Despite the ecological and economical significance of plant species, plant genetics can be considered as less advanced than human genetics. Therefore, association studies remain an even greater challenge for plant biologists, and the adoption of an adequate SNP-genotyping strategy and associated method provides one of the crucial obstacles that must be overcome.

One major concern with plant species is the complexity of their genomes compared with those for which standard methods have been developed. The SNP-genotyping systems developed for simple genomes like those of yeast are even difficult to translate to applications in human genetics (Syvänen 2005). Plant species, except for a few examples such as the model system *Arabidopsis thaliana*, have complex genomes with a large range of genome size (for instance, the genome of *Pinus pinaster* is up to 25.7 Gb in size, which is 150 times larger than that of *A. thaliana*), ploidy level (for instance, among the Poaceae family of cereals and grasses, rice is a diploid, maize is a paleotetraploid, and wheat is an allohexaploid), and genome structure (i.e., different species may contain dramatically different contents of repetitive DNA sequences). Different strategies can be used to circumvent the problem of genome complexity for association studies. The strategy employed will influence the choice of the appropriate techniques for SNP discovery (see previous chapter) and marker genotyping. As established earlier in Chapter 4, hundreds of thousands of SNPs have been recently identified in a number of plant species, using *in vitro*, *in silico*, or indirect discovery methods. The two main strategies which can be followed for SNP genotyping to obtain genetic correlation data are whole-genome scans and candidate gene-based approaches. The different scales of analysis required for these two approaches influence the choice and scale characteristics of the detection technology.

5.2 COMPARISON OF WHOLE-GENOME SCAN AND CANDIDATE GENE-BASED APPROACHES

5.2.1 Towards Whole-Genome Scans in Plants

The first strategy consists in scanning the whole-genome with a very large number of genetic loci (in the region of 10,000–100,000 or higher). This objective is difficult to achieve as it requires an extremely detailed knowledge of the genome under consideration, the availability of a large number of independent SNP markers, and a high-throughput detection method that can ideally be multiplexed on a very large-scale. For plant species, in which genomes can be relatively complex, for which linkage disequilibrium may only extend over short molecular distances because of the influence of reproductive systems, and for which SNP frequencies may be low (Rafalski and Morgante 2004), this approach can be difficult to apply. For instance, for well-characterized crop species such as maize or wheat, despite the availability of large EST data sets suitable for *in silico* SNP discovery and partial or complete physical maps, implementation of whole-genome scan-based association genetics methods would be a

major undertaking. The number of SNPs required for such analysis would substantially exceed any current technical capacity for genotyping. The same problem arises for other model systems such as rice (the model genome for grasses and cereals), tomato (the model species for Solanaceous plants), or poplar (the model species for trees), and is even more acute for the broad range of little-studied, "genomic-orphan" plant species, none of which possess sufficient SNP resources to consider a whole-genome scan with a sufficiently high marker density. An attempt to perform a whole-genome scan in *Arabidopsis* was recently reported (Törjek *et al.* 2003), but this study was based on a very low number of markers (i.e., 100 SNPs) compared with the larger number of markers ideally required. Although SNP markers provide the most effective current marker system for association genetics analysis, those inbreeding plant species that are descended from narrow domestication bottlenecks may show LD extending over map distances measured in centimorgans rather than physical distances in the range from Kb to Mb. In this case, other marker systems may be amenable to implementation for whole-genome scans, including restriction fragment length polymorphisms (RFLPs), amplified fragment length polymorphisms (AFLPs), and diversity array technology (DArT) (Jaccoud *et al.* 2001; Wenzl *et al.* 2004), because of the highly multiplex nature of the relevant assays and simple sequence repeats (SSRs), because of the highly polymorphic and multiallelic nature of the physical locus. In addition, SSR-based genetic maps have been developed for a number of the most important crop species and may consequently be directly used for this purpose. Studies of this nature have been performed in rice with SSRs (Semon *et al.* 2005), sugarcane (Jannoo *et al.* 1999), and sorghum (Deu *et al.* 2005) using RFLPs, as well as durum wheat, soybean, and other species. Although such studies are likely to be augmented and eventually supplanted by SNP-based surveys, the current data have been highly valuable for assessment of genome-wide patterns of LD.

For large nuclear plant genomes, the feasibility of whole-genome scan-based LD analysis is highly enhanced by methods for reduction of genome complexity. For the last 20 years, the global trend to reduce genome complexity in experimental DNA samples has been to use the polymerase chain reaction (PCR) technique. However, the scoring of millions of SNP loci spanning the entire genome over large numbers of test individuals cannot be realistically achieved by using PCR amplification, even with a high level of multiplexing, which is in any case often difficult to achieve. For that reason, more contemporary methods use whole-genome amplification (WGA) techniques (Telenius *et al.* 1992), which consist of amplifying total genomic DNA without a requirement for locus-specific oligonucleotide primers. It should be noted that although WGA is most commonly performed using PCR in combination with random oligomers (usually between 6 and 10 nucleotides in length), the recently developed multiple displacement amplification (MDA) technique (Dean *et al.* 2002) which is commercialized by GE HealthCare (GE Healthcare, Little Chalfont, UK) as GenomiPhi™, relies instead on isothermal rolling-circle replication catalyzed by bacteriophage 29 polymerase, and has been reported to provide superior genome coverage to methods such as degenerate oligonucleotide primed-polymerase chain reaction (DOP-PCR). The idea behind the use of WGA is to reduce the number of reactions to be performed for one individual to one single tube reaction. An example of genome complexity reduction using WGA was reported by Jordan *et al.* (2002) in which DOP-PCR was used to amplify *Arabidopsis* genomic DNA from different ecotypes. The authors suggested that the DOP-PCR amplified DNA could be used for SNP genotyping by direct sequencing or by ASO hybridization-based methods. Subsequently, several groups attempted to combine WGA

methods with microarray technologies, based on the potential for microarray to obtain high-density analysis. Following this idea, Matsuzaki *et al.* (2004) used a complexity-reduction assay to genotype more than 10,000 human SNPs based on oligonucleotide array-mediated detection. The complete genotyping assay, from the DNA template to the genotypic score, consists of the following steps: restriction enzyme digestion and universal adaptor ligation, amplification using sequences complementary to the adaptors as PCR primers, fragmentation and labeling, hybridization to the microarray, image scanning and data acquisition. This study presented one of the first examples of relatively high-resolution genotyping of a complex genome (i.e., average density of one SNP every 0.1 cM in the human genome), and represents one of the most promising techniques for high-throughput and accurate SNP genotyping. Similar high-throughput methods were recently developed such as the fiber-optic array-linked GoldenGate® assay (Illumina, Inc., San Diego, USA), the molecular inversion probe assay (Hardenbol *et al.* 2005), which utilizes an oligonucleotide ligation method (OLA; Iannone *et al.* 2000) and the Infinium™ WGG method (Gunderson *et al.* 2005), which combines WGA, allele-specific primer extension (ASPE; Taylor *et al.* 2001) and oligonucleotide array hybridization using the BeadArray™ technology (Illumina, Inc., San Diego, USA; Shen *et al.* 2005). Those latest assays make it possible to consider the completion of ambitious initiatives in plant species equivalent to the HapMap project, which requires high-resolution SNP haplotype definition across the genomes of members of multiple human populations (HapMap 2003). Even if such techniques are, in theory, transferable for application in any target organism, the application of such techniques for a plant species has not yet been demonstrated. The sole existing example of the use of microarray systems for SNP detection was performed through detection of single feature polymorphisms (SFPs) on the *Arabidopsis* Affymetrix GeneChip® (Borevitz *et al.* 2003). Interestingly, the use of a gene-expression-orientated array ensured that the analysis did not require any prior specific SNP development, but rather inferred SNP structure retrospectively through comparison of differential features. In this case, the characterization of such a number of chip-based SNP loci could be sufficient to permit complete coverage of the *Arabidopsis* genome, but this number would still not be sufficient for performing a whole-genome scan in species characterized by a less extensive LD, such as forest trees or out-breeding forage species (see Chapters 9 and 10). Moreover, and potentially of even higher significance, the assayed SNPs are only located in transcribed sequences. Since the Affymetrix chip-based experimental system is based on transcribed regions only it is unable (at present) to detect DNA variations arising only in noncoding regulatory regions, which have been proposed to account for the majority of quantitative trait variation in animal systems such as *Drosophila melanogaster* (Robin *et al.* 2002), and have been directly implicated in a high proportion of such variation in plant species (Paran and Zamir 2003). Nevertheless, availability of genome sequence for the two commonly used *Arabidopsis* ecotypes (i.e., Columbia [Col] and Landsberg *Erecta* [L*er*]) has provided access to a large quantity of ready-characterized SNPs located in both coding and noncoding regions (Jander *et al.* 2002). Furthermore, as plant genomes can now be sequenced in a relatively short time, as demonstrated through the completion of the genome sequence of poplar in less than two years (Brunner *et al.* 2004), it seems likely that the methods developed for SNP detection in model systems will soon be available for use by other crop biologists.

5.2.2 The Candidate Gene Approach

The second strategy which can be used for association studies in plants to reduce the complexity of the genomic regions that are targeted is the candidate gene approach. This approach consists of the characterization of SNPs present in a subset of specific genes identified using various strategies such as bioinformatics-based data mining, QTL analysis and linkage mapping, expression studies, transgenic modification by antisense RNA expression or RNA interference (RNAi), or positional cloning and physical mapping. The idea is to find the single base polymorphism that is directly causal of functional variation in the trait of interest (which is often termed the qualitative or quantitative trait nucleotide, QTN), or at least to find a SNP located within the functional gene or at a small physical distance from the gene. This strategy provides a good solution to the problems raised by the rapid decline of linkage disequilibrium observed in plant genomes (Rafalski and Morgante 2004), as the chances that linkage disequilibrium may be dissipated by a recombination event are extremely low in generational time (c. 10^{-6} per meiosis) when assaying a SNP located in a candidate gene, compared with much higher probabilities when using a more distant marker in a low-resolution genome scan. This numerically discrete strategy may consequently be applied to a large number of individuals (such as those present within germplasm collections).

5.3 SNP GENOTYPE SCORING METHODS

5.3.1 Sequencing Methods

A technique that may be widely used for SNP genotyping in candidate genes is the direct sequencing of PCR products. As described earlier in Chapter 4, sequencing is accurate and may also be used for SNP discovery and validation. Indeed, sequencing-derived data are often taken as a benchmark standard in studies for the evaluation of novel SNP-genotyping methods. Given that a large number of laboratories now possess automated sequencers or have access to facilities offering low-cost sequencing services, this method may also be highly effective in terms of cost and throughput. Current capillary electrophoresis technology permits sequencing of fragments of up to 1 Kb in length, which makes it possible to genotype several SNPs within the same sequence, and determine the haplotype structure within the sequenced fragment. Several examples of the application of SNP detection using sequencing in plants have been published. As an example, in forest trees, several authors (Brown *et al.* 2004; Gill *et al.* 2003; Pot *et al.* 2005) studied nucleotide diversity within candidate genes associated with wood formation and adaptive traits in three pine species. Sets of loci were sequenced across a range of natural populations, revealing heterogeneous patterns of diversity in the evaluated genes. The loci subjected to genotyping were chosen because they corresponded to genes of known function in wood formation, as well as co-locating with QTLs for wood quality that were previously identified by genetic mapping (Brown *et al.* 2003; Chagné *et al.* 2003).

Pyrosequencing (Ahmadian *et al.* 2000) may also be used for SNP genotyping through generation of short-read sequences, although the technique slightly differs from that used in standard Sanger–Coulsen sequencing chemistry (see Chapter 4). An example of the use of pyrosequencing was reported for genotyping of SNPs associated with grain

quality in barley (Polakova *et al.* 2003). Pyrosequencing is a very rapid and accurate method, and the cost per sample is relatively amenable to high-throughput analysis (i.e., 5,000 samples a day), even if the price of the requisite equipment platform (PSQ 96™, Pyrosequencing AB, Uppsala, Sweden) and associated analysis software is relatively high. The new large-scale sequencing 454 technology (Margulies *et al.* 2005), which exploits a pyrosequencing system in combination with solid-phase reaction support and picoliter-scale reaction volumes, offers a potential mechanism for dramatic increase in the scale of such sequencing efforts, and has already proven effective for resequencing of specific small-scale genomic regions.

5.3.2 DNA Conformation Methods

In vitro nonsequencing-based methods used for SNP discovery as described in Chapter 4 may also be used for SNP genotyping on a low- to medium-throughput scale. As an example, the SSCP method (Orita *et al.* 1989) was applied to SNP detection in pearl millet (Bertin *et al.* 2005). The authors showed that this method could detect SNPs located in introns, following a careful primer design procedure using various bioinformatics tools. According to this study, SSCP was sufficiently accurate to differentiate haplotypes. Although the position and type of the SNPs detected remain unknown, this method was suggested to be useful for association studies. Other DNA conformation methods like DGGE (Myers *et al.* 1987) and heteroduplex migration using dHPLC (Kota *et al.* 2001) may be used for SNP genotyping (Baumler *et al.* 2003; Schwarz *et al.* 2003). These methods are highly scalable, as they can be automated and run on higher-resolution and higher-throughput capillary electrophoresis systems (Hsia *et al.* 2005; Jander *et al.* 2004; Kourkine *et al.* 2002; Kuhn and Schnell 2005).

5.3.3 Allele-Specific PCR Amplification

Another method frequently used for genotyping SNPs is based on allele-specific PCR amplification (Figure 5.1; Newton *et al.* 1989). One of the PCR primers defining a sequence-tagged site (STS) is designed to preferentially amplify one of the SNP alleles, and the PCR fragments are subsequently separated on agarose-based electrophoresis gels. A mismatched base may be added close to the SNP site, three or four nucleotides upstream from the 3′-terminus of the primer, in order to enhance the preferential amplification of one of the alleles (Rust *et al.* 1993). This extremely simple method can be applied by any molecular biology laboratory, but remains a very low-throughput method which may only be applied as part of the candidate gene approach. The method is also vulnerable to false negative effects, as PCR failures cannot be reliably distinguished from genuine primer-binding discrimination in the absence of a reciprocal test. As an example, Délye *et al.* (2002) used this approach to identify herbicide resistant black-grass genotypes by designing allele-specific PCR markers within the chloroplastic *ACCase* gene. They demonstrated the efficacy of the method over a large number of samples (more than 1,000). Similarly, allele-specific amplification was used for genotyping of a supernodulation-related mutation located in the soybean *GmNARK* gene (Kim *et al.* 2005). The same method was employed successfully for mapping the *GmNARK* gene in a

F_2 population and the association between the SNP markers and the contrasting nodulation production trait was confirmed in different genetic backgrounds.

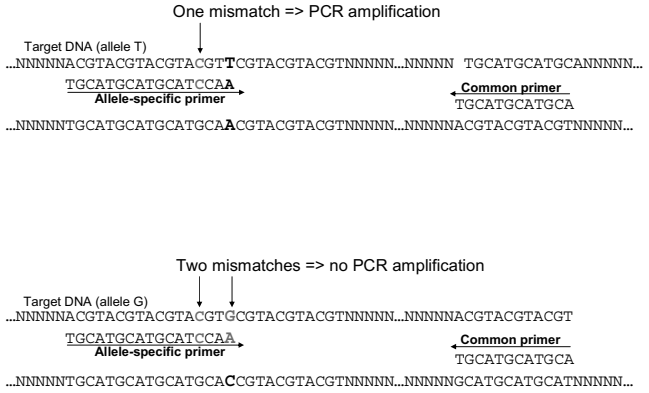


Figure 5.1. Allele-specific PCR amplification. In the example presented a G/T SNP is targeted by two PCR primers: one allele-specific primer is designed to anneal in its 3′ end to one of the SNP variants, but not to the other. In addition to the SNP site mismatch, a second mismatch is included on the third or fourth nucleotide situated upstream of the SNP site, in order that the PCR fails for the nonspecific allele, because of the low complementarity of the primer in the 3′ end. No fluorescent labeling is required and regular PCR conditions can be carried out.

5.3.4 Enzymatic Cleavage Methods

The derived cleaved amplified polymorphic sequence (dCAPS) method (Neff *et al.* 1998) represents a cost-effective system to convert SNPs into dominant PCR markers through incorporation of a mismatch into one of the PCR amplification primers in order to create a target restriction endonuclease site. Unlike the standard CAPS method, which is dependent on the presence of an SNP within a restriction site, dCAPS can be developed relatively easily for any SNP locus. The dCAPS method was successfully used to map SNPs linked to vernalization requirement in wheat (Iwaki *et al.* 2002) and to develop molecular markers linked to self-compatibility in sweet cherry (Ikeda *et al.* 2004).

The TILLING method (McCallum *et al.* 2000) differs slightly from CAPS or dCAPS in that most commonly used versions use a nuclease, such as *Cel*I, which cleaves mismatch-containing heteroduplex DNA. Although TILLING was originally developed for detection of mutations induced by chemical agents such as ethylmethanesulphonate (EMS), it may also be used in "ecoTILLING" applications to genotype SNPs in natural populations, and is hence suitable for association studies in plants (Comai *et al.* 2004; Gilchrist and Haughn 2005). One example is the demonstration of high levels of diversity in poplar, a long-lived woody perennial species which is outcrossing in nature and shows a wide natural range of distribution (Cronk 2005). However, the routine use of ecoTILLING in outbreeding plant species is likely to be highly exacting technically.

The Invader™ assay (Figure 5.2) is a relatively new technique designed specifically for genotyping SNPs (Mein *et al.* 2000; Olivier 2005). The technique uses two target

specific oligonucleotide probes (invader probe and SNP specific probe prolonged by a flap sequence) that anneal to the SNP and form a three-dimensional complex. The flap sequences are not complementary to the SNP site, but in the presence of the complex, an endonuclease (FEN) cleaves the flap, which is released and induces a fluorescent emission. The first generation of the Invader™ assay, although being a highly accurate method, does require the PCR amplification of the target DNA and the design of a specific secondary probe for each of the SNP alleles. This increases the cost of the method, which makes it unsuitable for high-throughput genotyping. The second generation of Invader™ assay, namely the Biplex Invader™ assay (Olivier *et al.* 2002), uses a serial invasive reaction, where two unlabeled allele-specific probes are designed. Each of them, if they anneal to the target SNP allele, releases a flap sequence which is complementary to a fluorescence resonance energy transfer (FRET) molecule, which fuels another cleavage reaction and then emits fluorescence. This method is characterized by a very high accuracy and a low failure rate, which makes it very attractive for plant biologists who want to genotype a small number of SNPs over large populations, as is characteristic of the candidate gene approach.

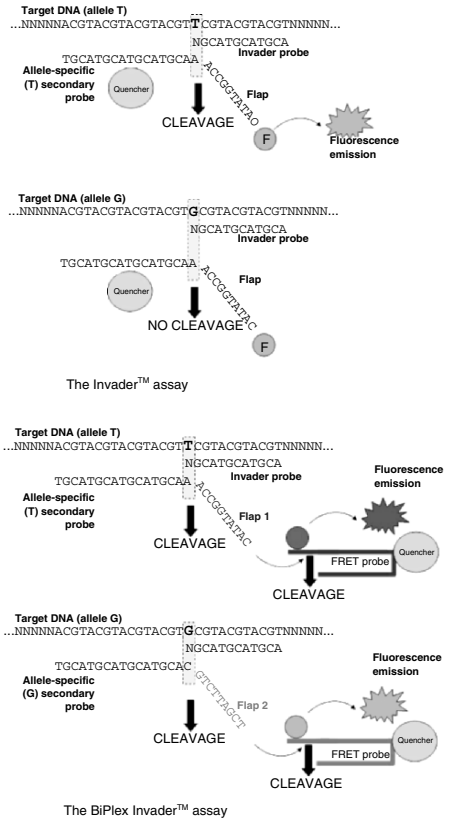


Figure 5.2. The Invader™ assay. (see color plate)

The principle is as follows: an oligonucleotide invader probe is designed to anneal immediately next to the variable site, in the opposite direction to an allele-specific probe (secondary probe) prolonged by a flap in 5′-orientation. (A) In the first generation of Invader™ assay, a quencher molecule and a fluorophore were attached to the secondary probe. If the secondary probe is complementary to the SNP allele, in the presence of the invader probe a three-dimensional complex is formed, which induces cleavage by a flap structure-specific endonuclease (FEN). The cleaved 5′-flap fragment then triggers a reaction between the quencher molecule, the fluorophore, and the cleaved fragment, which results in a fluorescent emission. In the case where no invasive complex is created (i.e., the secondary probe is not complementary to the SNP allele), the FEN does not perform cleavage and no fluorescence is emitted. (B) A second generation of Invader™ assay, called the Biplex Invader™ assay (Olivier *et al.* 2002) was recently developed and allows the detection of both alleles in the same reaction tube as well as the use of nonlabeled allele-specific probes. The technique uses the same principle of cleavage using the FEN. However, the flap sequence prolonging the secondary probe is complementary to a fluorescent resonance energy transfer (FRET) molecule which is labeled with a fluorophore and a quencher. Two secondary probes specific to both SNP alleles can be added into the same tube and each of the flaps are complementary to a FRET labeled with a different fluorescent dye. The flap released by the FEN cleavage anneals to the complementary FRET, which creates another complex that is targeted by the FEN, inducing a fluorescent emission.

5.3.5 Hybridization With Allele-Specific Oligonucleotide Probes

The ASO probe hybridization method is based on the interaction of solid-phase immobilized oligonucleotides with labeled template DNA obtained by standard PCR or WGA. The variant SNP site is usually located in the central position of a 25-mer oligonucleotide feature (Figure 5.3), although the oligonucleotide length and SNP position can vary according to the detection system which is employed. An example of the use of such oligonucleotide probes for detecting SNP variants formatted in high-density multiplex arrays was discussed earlier in the chapter (Borevitz *et al.* 2003). Similarly, ASO was used for genotyping SNPs located in microsatellite flanking sequences in maize (Mogg *et al.* 2002).

One popular method based on ASO hybridization is the Taqman™ assay (Livak 1999; Livak *et al.* 1995). The principle of this method is as follows: an oligonucleotide probe labeled with a fluorescent dye and specific to a SNP allele is combined with PCR primers that are capable of amplifying the SNP-containing region. The 5′-endonuclease activity of the *Taq* polymerase releases the fluorescent molecule, which can be detected by a real-time PCR (Heid *et al.* 1996) instrument (e.g., ABI 7700; Applied Biosystems, Foster City, CA, USA). The Taqman™ assay is highly rapid and accurate and the equipment may also be used for other genomics applications such as expression studies by reverse transcriptase polymerase chain reaction (RT-PCR), which contributes to a flexible investment. In terms of cost, the need for labeled probe increases the cost per sample, and so the technique is only suitable for small-to-medium-throughput projects. An example of the use of the Taqman™ technology in plants was reported in potato (De Jong *et al.* 2003). Mutations in the dihydroflavonol 4-reductase genes that co-segregate with the *R* gene (controlling production of red anthocyanin pigments in the skin) were

genotyped over a range of potato clones spanning different genotypes of the *R* locus. Interestingly, the Taqman™ assay could distinguish between different allele dosages (as potato has an autopolyploid genetic constitution) which is an attractive and often critical feature for analysis of those plant species that show complex higher ploidy levels (such as wheat or kiwifruit) or are derived paleopolyploids (such as apple and maize).

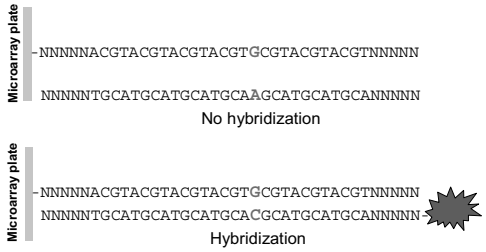


Figure 5.3. Allele-specific oligonucleotide hybridization. A oligonucleotides feature with the SNP site in its central position is bound to a microarray glass plate. Under stringent hybridization conditions, the complementary allele will anneal to the fixed oligonucleotide and a fluorescent signal attached to the probe will be detected. (see color plate)

5.3.6 Oligonucleotide Ligation Assay

The oligonucleotide ligation assay (OLA; Landegren *et al.* 1988) is based on the properties of a long characterized enzymatic reaction in which two adjacent oligonucleotides may be covalently joined by a DNA ligase when annealed to a complementary DNA target (Figure 5.4). Both primers must have perfect base pair complementarity at the ligation site, which makes it possible to discriminate two alleles at a SNP site. There are several applications which have been developed to detect SNP variation using OLA, including colorimetric assays in ELISA plates (Tobe *et al.* 1996), separation of the ligated oligonucleotide that have been labeled with a fluorescent dye on an automated sequencer, or rolling-circle amplification (RCA) with one of the ligation probes bound to a microarray surface (Faruqi *et al.* 2001; Lizardi *et al.* 1998). The RCA method can be used directly on genomic DNA, which makes it suitable for genome scan approaches (Alsmadi *et al.* 2003) as well as the candidate gene approach. No example of the use of the OLA technique has been reported in plants for association studies yet, but the technique has been used in cattle, which suggests that it may also be applied to crop species (Dunner *et al.* 2003).

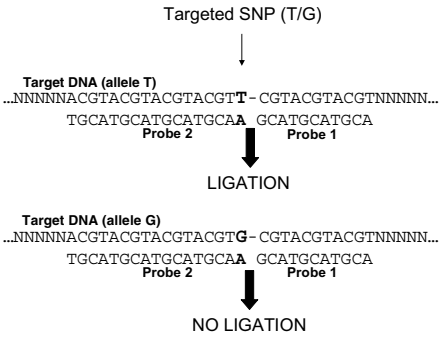


Figure 5.4. Oligonucleotide ligation assay (OLA). The OLA is based on the ligation of two probes hybridizing next to the SNP site. The joining of the two probes using a DNA ligase depends on the probes hybridization, their juxtaposition on the target sequence and the perfect complementarity at the joining site. If the allele-specific probe is not specific to the SNP variant, the ligation does not occur.

5.3.7 Minisequencing/Primer Extension

A popular method which was designed specifically for genotyping SNPs is the minisequencing technique (Syvänen 1999; Syvänen *et al.* 1990), also called the primer extension technique. The principle of this method is as follows: a detection primer is designed to target a sequence immediately upstream of the SNP. Then, the 3′-terminus of the oligonucleotide is extended by a DNA polymerase using labeled ddNTPs (Figure 5.5). Therefore, one terminating fluorescent dye corresponds to each individual base, which makes it possible to detect up to four allelic variants for a variable site and discriminate heterozygous from homozygous genotypes. Different detection platforms such as microarrays (Pastinen *et al.* 1997), capillary electrophoresis systems (Pastinen *et al.* 1996), pyrosequencing (Ekstroem *et al.* 2000), flow cytometry (Chen *et al.* 2000), mass spectrometry (Buetow *et al.* 2001; Haff and Smirnov 1997; Li *et al.* 1999; Tang *et al.* 1999) or fluorescence plate readers (Chen *et al.* 1999; Hsu *et al.* 2001; Lopez-Crapez *et al.* 2005) can be employed with the minisequencing method, demonstrating its flexibility of adaptation to different analytical technologies.

As an example in plants, Törjek *et al.* (2003) used minisequencing to develop a set of 112 SNP markers in *A. thaliana* using the SNaPshot™ assay combined with the use of an ABI 3700 automated sequencer (Applied Biosystems, Foster City, CA, USA) and a matrix-assisted laser desorption/ionization time of flight (MALDI-TOF) mass spectrometer. Both platforms allowed the set of markers to be multiplexed (such that 5,376 data points were collected in this study), which suggested that the method can be used as a medium- to high-throughput genotyping system. In crop plants, the primer extension technique was employed for studying the association between variations in the β-amylase gene and the fermentation properties of barley (Paris *et al.* 2002). The authors

used the SNuPe technique (GE Healthcare, Little Chalfont, UK) to genotype their SNPs over a range of barley breeding lines. Similarly, the SNuPe method was employed to genotype SNPs linked to microsatellite loci in maize inbred lines (Batley *et al.* 2003), using a Megabace capillary sequencer (GE Healthcare, Little Chalfont, UK). A set of SNPs linked to a leaf rust resistance gene in wheat (Tyrka *et al.* 2004) was also genotyped by the SNuPe technique. Interestingly, Lee *et al.* (2004) compared the primer extension technique with three other methods (i.e., ASPE, OLA, and direct hybridization), using a flow cytometry instrument as a detection system (Luminex, Austin, TX, USA). Results of the four methods were compared with SNP genotype scores obtained with the SNaPshot kit as a positive control. Overall, minisequencing and ASPE using flow cytometric detection methods were shown to be effective methods for the provision of codominant markers, as both SNP alleles can be discriminated and are represented in the same reaction tube. However, minisequencing methods do show some demerits in terms of cost and time, as the reactions need to be treated before SNP detection using exonuclease I and shrimp alkaline phosphatase (SAP) to degrade excess PCR primers and dNTPs prior to DNA polymerization.

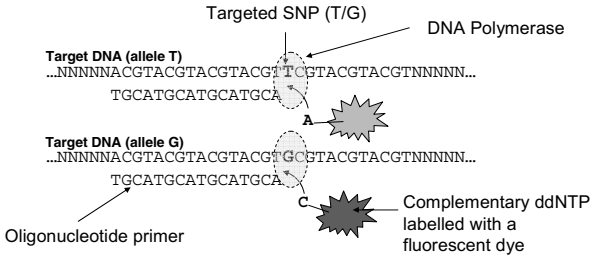


Figure 5.5. Minisequencing or primer extension. An oligonucleotide primer immediately flanking the SNP is extended using a DNA polymerase. Fluorescently labeled terminating nucleotides are incorporated, with a different dye color for every nucleotide. The oligonucleotide can be attached to a solid-phase array, separated in a capillary electrophoresis system, by a flow cytometry instrument, by mass spectrometry, or revealed by a fluorescent plate reader. (see color plate)

5.4 FUTURE TRENDS FOR THE GENOTYPIC ANALYSIS OF SNPs FOR ASSOCIATION STUDIES IN PLANTS

As shown in this chapter, a large diversity of SNP-genotyping techniques is available for plant biologists to conduct association studies. All these techniques differ in accuracy, cost per sample, and the number of data points (i.e., number of SNPs $\times$ number of individuals) that can be created (Table 5.1). Overall, the choice of the techniques is

Table 5.1. Relative costs of different SNP-genotyping methods. Note that we have used relative rankings in terms of cost as exact values will change with time

Technique	Scale	Equipment and platform cost	Reagents cost
Sequencing			
Resequencing (Sanger)	Low- to medium-throughput	High	High
Pyrosequencing	Medium-throughput	High	Medium
454 sequencing	High-throughput	High	Low
DNA conformation			
SSCP	Low- to medium-throughput	Low	Low
DGGE	Low- to medium-throughput	Low	Low
dHPLC	Low- to medium-throughput	Medium	Low
Allele-specific amplification	Low-throughput	Low	Low
Enzymatic cleavage methods			
CAPS and dCAPS	Low-throughput	Low	Low
TILLING	Medium-throughput	Medium	Low
Invader assay	Medium- to high-throughput	Medium	Medium–High
Allele-specific oligonucleotides			
Microarray-based	High-throughput	High	High
Taqman	Medium-throughput	Medium	High
Oligonucleotide ligation assay			
ELISA colorimetric assay	Medium-throughput	Low	Low
Rolling-circle amplification	High-throughput	Medium	Medium
Minisequencing			
Allele-specific primer extension	Medium- to high-throughput	High	Medium

specific to the project and the crop considered. The HapMap project (HapMap 2003) provides a striking example of an ambitious international consortium dedicated to genotyping a very large number of SNPs over a range of individuals derived from multiple human groups. Plant geneticists may well regard such a project with envy, not least because of the high level of financial investment required, and speculate on the feasibility of performing such studies in plant genomes. Plant genetics laboratories, whether located in academia, the public service sector or the private sector, have typically been geared to low- to medium-throughput genetic analysis and are often multidisciplinary in nature, with expertise ranging from classical quantitative and population genetics to recent molecular biology disciplines, studying gene function and expression or genome structure and organization. For this reason, the method chosen for SNP genotyping would have to fit with other technical requirements. In the particular case of genetic analysis, the method chosen would need to be amenable to association studies, as well as linkage mapping or marker-assisted breeding. All these applications require different numbers of loci to be considered and different scales of plant samples to be characterized. Genetic trait dissection, generally based at present on linkage mapping and QTL analysis, is characterized by relatively small numbers of closely related genotypes (150–300) and large numbers of genetic markers (200–400). By contrast, implementation of validated genetic marker-trait gene associations in molecular plant breeding is characterized by relatively large numbers of individuals (typically 1,000–10,000) and small numbers of markers (5–25). Association genetic analysis, as an aspect of DNA profiling, shows heterogeneous scale requirements from one to thousands of markers, and from tens to thousands of individuals. The flexible scales of genotyping analysis imply a necessity for equally flexible genotyping platforms, ideally modular in nature, to service the different requirements.

The model plant species *A. thaliana* may provide a good model for association studies through a whole-genome scan strategy. However, unlike the majority of crop plant species, *A. thaliana* does not possess a particular complex genome. Furthermore, its reproductive system is not shared by the majority of plant species and *A. thaliana* has not experienced a strong domestication bottleneck as has occurred for many major crop species, implying some differences in the structure and distribution of LD compared to that seen in other plants. On the other hand, the self-pollinating (autogamous) breeding system of *A. thaliana* is shared by some crop species such as wheat, barley, rice, tomato, sorghum, pearl millet, and others, and on this basis, LD information from the model species may prove useful for other species. It is clear, however, that intensive association genetics studies must be performed in each target species or species group (as described in Chapters 9–11), and the candidate gene-based approach seems to offer the most feasible current option for such analysis. As previously described, plant genomics tends to follow trends established in its human counterpart, such as complete sequencing of several plant genomes and the establishment of large EST data sets. If a highly multiplexed, high-throughput, accurate, low-cost technique is developed for human genetics, such a system will be rapidly assimilated into plant genetics. The GoldenGate™ and Infinium™ assays commercialized by Illumina (Illumina, San Diego, CA, USA) present a number of highly attractive features, especially the capacity to process large numbers of SNP loci over multiple DNA samples using a microtiter plate format, and the capability to produce modular rearrangements of array elements to address different scales of analysis, as described earlier. The identification of large numbers of validated SNP loci is an issue in the generation of such systems, but the methodologies described in Chapter 4 will provide suitable sets for all major crop species in the near future. A prototype barley SNP-based Illumina system for assay of 1,536 gene-associated SNPs has recently been developed in collaboration with the Scottish Crop Research Institute (SCRI), Dundee, UK (R. Waugh, personal communication). The performance of this prototype will provide important information on general applicability to other crop species. If the 454 DNA sequencing technology (Margulies *et al.* 2005) can be successfully adapted to large-scale genotyping applications, this may offer another attractive route for plant geneticists who want to identify SNPs and apply LD mapping.

5.5 REFERENCES

Ahmadian, A., Gharizadeh, B., Gustafsson, A.C., Sterky, F., Nyren, P., Uhlen, M., Lundeberg, J., 2000, Single-nucleotide polymorphism analysis by pyrosequencing. Analytical Biochemistry 280:103–110.

Alsmadi, O.A., Bornarth, C.J., Song, W., Wisniewski, M., Du, J., Brockman, J.P., Faruqi, A.F., Sun, Z., Du, Y., Wu, X., Egholm, M., Abarzúa, P., Lasken, R.S., Driscoll, M.D., 2003, High accuracy genotyping directly from genomic DNA using a rolling circle amplification based assay. BMC Genomics 4:21.

Batley, J., Mogg, R., Edwards, D., O'Sullivan, H., Edwards, K.J., 2003, A high-throughput SNuPE assay for genotyping SNPs in the flanking regions of *Zea mays* sequence tagged simple sequence repeats. Molecular Breeding 11:111–120.

Baumler, S., Felsenstein, F.G., Schwarz, G., 2003, CAPS and DHPLC analysis of a single nucleotide polymorphism in the cytochrome b gene conferring resistance to strobilurins in field isolates of *Blumeria graminis* f. sp *hordei*. Journal of Phytopathology-Phytopathologische Zeitschrift 151:149–152.

Bertin, I., Zhu, J.H., Gale, M.D., 2005, SSCP-SNP in pearl millet – A new marker system for comparative genetics. Theoretical and Applied Genetics 110:1467–1472.

Borevitz, J.O., Liang, D., Plouffe, D., Chang, H.S., Zhu, T., Weigel, D., Berry, C.C., Winzeler, E., Chory, J., 2003, Large-scale identification of single-feature polymorphisms in complex genomes. Genome Research 13:513–523.

Brown, G.R., Bassoni, D.L., Gill, G.P., Fontana, J.R., Wheeler, N.C., Megraw, R.A., Davis, M.F., Sewell, M.M., Tuskan, G.A., Neale, D.B., 2003, Identification of quantitative trait loci influencing wood property traits in Loblolly pine (*Pinus taeda* L.). III. QTL verification and candidate gene mapping. Genetics 164:1537–1546.

Brown, G.R., Gill, G.P., Kuntz, R.J., Langley, C.H., Neale, D.B., 2004, Nucleotide diversity and linkage disequilibrium in loblolly pine. Proceedings of the National Academy of Sciences of the United States of America 101(42):15255–15260.

Brunner, A.M., Busov, V.B., Strauss, S.H., 2004, Poplar genome sequence: functional genomics in an ecologically dominant plant species. Trends in Plant Science 9:49–56.

Buetow, K.H., Edmonson, M., MacDonald, R., Clifford, R., Yip, P., Kelley, J., Little, D.P., Strausberg, R., Koester, H., Cantor, C.R., Braun, A., 2001, High-throughput development and characterization of a genomewide collection of gene-based single nucleotide polymorphism markers by chip-based matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. Proceedings of the National Academy of Sciences of the United States of America 98:581–584.

Chagné, D., Brown, G., Lalanne, C., Madur, D., Pot, D., Neale, D., Plomion, C., 2003, Comparative genome and QTL mapping between maritime and loblolly pines. Molecular Breeding 12:185–195.

Chen, X., Kwok, P.-Y., Levine, L., 1999, Fluorescence polarization in homogeneous nucleic acid analysis. Genome Research 9:492–498.

Chen, J., Iannone, M.A., Li, M.-S., Taylor, J.D., Rivers, P., Nelsen, A.J., Slentz-Kesler, K.A., Roses, A., Weiner, M.P., 2000, A microsphere-based assay for multiplexed single nucleotide polymorphism analysis using single base chain extension. Genome Research 10:549–557.

Comai, L., Young, K., Till, B.J., Reynolds, S.H., Greene, E.A., Codomo, C.A., Enns, L.C., Johnson, J.E., Burtner, C., Odden, A.R., Henikoff, S., 2004, Efficient discovery of DNA polymorphisms in natural populations by Ecotilling. Plant Journal 37:778–786.

Cronk, Q.C.B., 2005, Plant eco-devo: the potential of poplar as a model organism. New Phytologist 166:39–48.

Dean, F.B., Hosono, S., Fang, L.X.W., Fawad Faruqi, A., Bray-Ward, P., Sun, Z., Zong, Q., Du, Y., Du, J., Driscoll, M., Song, W., Kingsmore, S.F., Egholm, M., Lasken, R.S., 2002, Comprehensive human genome amplification using multiple displacement amplification. Proceedings of the National Academy of Sciences of the United States of America 99:5261–5266.

De Jong, W.S., De Jong, D.M., Bodis, M., 2003, A fluorogenic 5′ nuclease (TaqMan) assay to assess dosage of a marker tightly linked to red skin color in autotetraploid potato. Theoretical and Applied Genetics 107:1384–1390.

Délye, C., Matéjicek, A., Gasquez, J., 2002, PCR-based detection of resistance to acetyl–CoA carboxylase-inhibiting herbicides in black-grass (*Alopecurus myosuroides* huds) and ryegrass (*Lolium rigidum* gaud). Pest Management Science 58:474–478.

Deu, M., Ratmadass, A., Hamada, M.A., Noyer, J.L., Diabate, M., Chantereau, J., 2005, Quantitative tratit loci for head-bug resistance in sorghum. African Journal of Biotechnology 4:247–250.

Dunner, S., Miranda, M.E., Amigues, Y., Cañon, J., Georges, M., Hanset, R., Williams, J., Ménissier, F., 2003, Haplotype diversity of the myostatin gene among beef cattle breeds. Genetics Selection Evolution 35:103–118.

Ekstroem, B., Alderborn, A., Hammerling, U., 2000, Pyrosequencing for SNPs. Proceedings of SPIE – The International Society for Optical Engineering 3926:134–139.

Faruqi, F.A., Hosono, S., Driscoll, M.D., Dean, F.B., Alsmadi, O., Bandaru, R., Kumar, G., Grimwade, B., Zong, Q., Sun, Z., Du, Y., Kingsmore, S., Knott, T., Lasken, R.S., 2001, High-throughput genotyping of single nucleotide polymorphisms with rolling circle amplification. BMC Genomics 2:4.

Gilchrist, E.J., Haughn, G.W., 2005, TILLING without a plough: a new method with applications for reverse genetics. Current Opinion in Plant Biology 8:15.

Gill, G.P., Brown, G.R., Neale, D.B., 2003, A sequence mutation in the cinnamyl alcohol dehydrogenase gene associated with altered lignification in loblolly pine. Plant Biotechnology Journal 1:253–258.

Gunderson, K.L., Steemers, F.J., Lee, G., Mendoza, L.G., Chee, M.S., 2005, A genome-wide scalable SNP genotyping assay using microarray technology. Nature Genetics 37:549–554.

Haff, L.A., Smirnov, I.P., 1997, Single-nucleotide polymorphism identification assays using a thermostable DNA polymerase and delayed extraction MALDI-TOF mass spectrometry. Genome Research 7:378–388.

HapMap, 2003, The International HapMap Project: The International HapMap Consortium. Nature 426:789–796.

Hardenbol, P., Yu, F., Belmont, J., MacKenzie, J., Bruckner, C., Brundage, T., Boudreau, A., Chow, S., Eberle, J., Erbilgin, A., Falkowski, M., Fitzgerald, R., Ghose, S., Iartchouk, O., Jain, M., Karlin-Neumann, G., Lu, X., Miao, X., Moore, B., Moorhead, M., Namsaraev, E., Pasternak, S., Prakash, E., Tran, K., Wang, Z., Jones, H.B., Davis, R.W., Willis, T.D., Gibbs, R.A., 2005, Highly multiplexed molecular inversion

probe genotyping: over 10,000 targeted SNPs genotyped in a single tube assay. Genome Research 15:269–275.

Heid, C.A., Stevens, J., Livak, K.J., Williams, P.M., 1996, Real time quantitative PCR. Genome Research 6:986–994.

Hsia, A.-P., Wen, T.-J., Chen, H.D., Liu, Z., Yandeau-Nelson, M.D., Wei, Y., Guo, L., Schnable, P.S., 2005, Temperature gradient capillary electrophoresis (TGCE) – a tool for the high-throughput discovery and mapping of SNPs and IDPs. Theoretical and Applied Genetics 111:218–225.

Hsu, T.M., Chen, X., Duan, S., Miller, R.D., Kwok, P.-Y., 2001, Universal SNP genotyping assay with fluorescence polarization detection. BioTechniques 31:560–570.

Iannone, M.A., Taylor, J.D., Chen, J., Li, M.-S., Rivers, P., Slentz-Kesler, K.A., Weiner, M.P., 2000, Multiplexed single nucleotide polymorphism genotyping by oligonucleotide ligation and flow cytometry. Cytometry 39:131–140.

Ikeda, K., Watari, A., Ushijima, K., Yamane, H., Hauck, N.R., Iezzoni, A.F., Tao, R., 2004, Molecular markers for the self-compatible S4? – Haplotype, a pollen-part Mutant in sweet cherry (*Prunus avium* L.). Journal of the American Society for Horticultural Science 129:724–728.

Iwaki, K., Nishida, J., Yanagisawa, T., Yoshida, H., Kato, K., 2002, Genetic analysis of *Vrn-B1* for vernalization requirement by using linked dCAPS markers in bread wheat (*Triticum aestivum* L.). Theoretical and Applied Genetics 104:571–576.

Jaccoud, D., Peng, K., Feinstein, D., Kilian, A., 2001, Diversity arrays: a solid state technology for sequence information independent genotyping. Nucleic Acids Research 29(4):e25.

Jander, G., Norris, S.R., Rounsley, S.D., Bush, D.F., Levin, I.M., Last, R.L., 2002, *Arabidopsis* map-based cloning in the post-genome era. Plant Physiology 129:440–450.

Jander, G., Norris, S.R., Joshi, V., Fraga, M., Rugg, A., Yu, S., Li, L., Last, R.L., 2004, Application of a high-throughput HPLC-MS/MS assay to *Arabidopsis* mutant screening; evidence that threonine aldolase plays a role in seed nutritional quality. Plant Journal 39:465–475.

Jannoo, N., Grivet, L., Seguin, M., Paulet, F., Domaingue, R., Rao, P.S., Dookun, A., D'Hont, A., Glaszmann, J.C., 1999, Molecular investigation of the genetic base of sugarcane cultivars. Theoretical and Applied Genetics 99:171–184.

Jordan, B., Charest, A., Dowd, J.F., Blumenstiel, J.P., Yeh, R.-F., Osman, A., Housman, D.E., Landers, J.E., 2002, Genome complexity reduction for SNP genotyping analysis. Proceedings of the National Academy of Sciences of the United States of America 99:2942–2947.

Kahl, G., Mast, A., Tooke, N., Shen, R., van den Boom, D., 2005, Single nucleotide polymorphisms: detection techniques and their potential for genotyping and genome mapping. In: Meksem, K., Kahl, G. (eds). The Handbook of Plant Genome Mapping: Genetic and Physical Mapping. Wilecy-VCH Verlag GmbH & Co., KGaA, Weinheim, pp. 75–104.

Kim, M.Y., Van, K., Lestari, P., Moon, J.-K., Lee, S.-H., 2005, SNP identification and SNAP marker development for a *GmNARK* gene controlling supernodulation in soybean. Theoretical and Applied Genetics 110:1003–1010.

Kota, R., Wolf, M., Michalek, W., Graner, A., 2001, Application of denaturing high-performance liquid chromatography for mapping of single nucleotide polymorphisms in barley (*Hordeum vulgare* L.). Genome 44(4):523–528.

Kourkine, I.V., Hestekin, C.N., Buchholz, B.A., Barron, A.E., 2002, High-throughput, high-sensitivity genetic mutation detection by tandem single-strand conformation polymorphism/heteroduplex analysis capillary array electrophoresis. Analytical Chemistry 74:2565–2572.

Kuhn, D.N., Schnell, R.J., 2005, Use of capillary array electrophoresis single-strand conformational polymorphism analysis to estimate genetic diversity of candidate genes in germplasm collections. Methods in Enzymology 395:238–258.

Kwok, P.-Y., 2001, Methods for genotyping single nucleotide polymorphisms. Annual Review of Genomics and Human Genetics 2:235–258.

Landegren, U., Kaiser, R., Sanders, J., Hood, L., 1988, A ligase-mediated gene detection technique. Science 241:1077–1080.

Lee, S.-H., Walker, D.R., Cregan, P.B., Boerma, H.R., 2004, Comparison of four flow cytometric SNP detection assays and their use in plant improvement. Theoretical and Applied Genetics 110:167–174.

Li, J., Butler, J.M., Tan, Y., Lin, H., Royer, S., Ohler, L., Shaler, T.A., Hunter, J.M., Pollart, D.J., Monforte, J.A., Becker, C.H., 1999, Single nucleotide polymorphism determination using primer extension and time-of-flight mass spectrometry. Electrophoresis 20:1258–1265.

Livak, K.J., 1999, Allelic discrimination using fluorogenic probes and the 5′ nuclease assay. Genetic Analysis – Biomolecular Engineering 14:143–149.

Livak, K.J., Flood, S.J., Marmaro, J., Giusti, W., Deetz, K., 1995, Oligonucleotides with fluorescent dyes at opposite ends provide a quenched probe system useful for detecting PCR product and nucleic acid hybridization. Genome Research 4:357–362.

Lizardi, P.M., Huang, X., Zhu, Z., Bray-Ward, P., Thomas, D.C., Ward, D.C., 1998, Mutation detection and single-molecule counting using isothermal rolling-circle amplification. Nature Genetics 19:225–232.

Lopez-Crapez, E., Bazin, H., Chevalier, J., Trinquet, E., Grenier, J., Mathis, G., 2005, A separation-free assay for the detection of mutations: combination of homogeneous time-resolved fluorescence and minisequencing. Human Mutation 25:468–475.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.-J., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L.I., Jarvie, T.P., Jirage, K.B., Kim, J.-B., Knight, J.R., Lanza, J.R., Leamon, J.H., Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A., Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F., Rothberg, J.M., 2005, Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380.

Matsuzaki, H., Loi, H., Dong, S., Tsai, Y.-Y., Fang, J., Law, J., Di, X., Liu, W.-M., Yang, G., Liu, G., Huang, J., Kennedy, G.C., Ryder, T.B., Marcus, G.A., Walsh, P.S., Shriver, M.D., Puck, J.M., Jones, K.W., Mei, R., 2004, Parallel genotyping of over 10,000 SNPs using a one-primer assay on a high-density oligonucleotide array. Genome Research 14:414–425.

McCallum, C.M., Comai, L., Greene, E.A., Henikoff, S., 2000, Targeting induced local lesions in genomes (TILLING) for plant functional genomics. Plant Physiology 123:439–442.

Mein, C.A., Barratt, B.J., Dunn, M.G., Siegmund, T., Smith, A.N., Esposito, L., Nutland, S., Stevens, H.E., Wilson, A.J., Phillips, M.S., Jarvis, N., Law, S., De Arruda, M., Todd, J.A., 2000, Evaluation of single nucleotide polymorphism typing with invader on PCR amplicons and its automation. Genome Research 10:330–343.

Mogg, R., Batley, J., Hanley, S., Edwards, D., O'Sullivan, H., Edwards, K.J., 2002, Characterization of the flanking regions of *Zea mays* microsatellites reveals a large number of useful sequence polymorphisms. Theoretical and Applied Genetics 105:532–543.

Myers, R.M., Maniatis, T., Lerman, L.S., 1987, Detection and localization of single base changes by denaturing gradient gel electrophoresis. Methods in Enzymology 155:501–527.

Neff, M.M., Neff, J.D., Chory, J., Pepper, A.E., 1998, dCAPS, a simple technique for the genetic analysis of single nucleotide polymorphisms: experimental applications in *Arabidopsis thaliana* genetics. Plant Journal 14:387–392.

Newton, C.R., Graham, A., Heptinstall, L.E., Powell, S.J., Summers, C., Kalsheker, N., Smith, J.C., Markham, A.F., 1989, Analysis of any point mutation in DNA. The Amplification refractory mutation system (ARMS). Nucleic Acids Research 17:2503–2516.

Olivier, M., 2005, The Invader® assay for SNP genotyping. Mutation Research – Fundamental and Molecular Mechanisms of Mutagenesis 573:103–110.

Olivier, M., Chuang, L.M., Chang, M.S., Chen, Y.T., Pei, D., Ranade, K., de Witte, A., Allen, J., Tran, N., Curb, D., Pratt, R., Neefs, H., de Arruda Indig, M., Law, S., Neri, B., Wang, L., Cox, D.R., 2002, High-throughput genotyping of single nucleotide polymorphisms using new biplex invader technology. Nucleic Acids Research 30:e53.

Orita, M., Suzuki, Y., Sekiya, T., Hayashi, K., 1989, Rapid and sensitive detection of point mutations and SNA polymorphisms using the polymerase chain reaction. Genomics 5:874–879.

Paran, I., Zamir, D., 2003, Quantitative traits in plants: beyond the QTL. Trends in Genetics 19:303–306.

Paris, M., Jones, M.G.K., Eglinton, J.K., 2002, Genotyping single nucleotide polymorphisms for selection of barley beta-amylase alleles. Plant Molecular Biology Reporter 20:149–159.

Pastinen, T., Syvänen, A.-C., Partanen, J., 1996, Multiplex, fluorescent, solid-phase minisequencing for efficient screening of DNA sequence variation. Clinical Chemistry 42:1391–1397.

Pastinen, T., Kurg, A., Metspalu, A., Peltonen, L., Syvänen, A.-C., 1997, Minisequencing: a specific tool for DNA analysis and diagnostics on oligonucleotide arrays. Genome Research 7:606–614

Polakova, K., Laurie, D., Vaculova, K., Ovesna, J., 2003, Characterization of beta-amylase alleles in 79 barley varieties with pyrosequencing. Plant Molecular Biology Reporter 21:439–447.

Pot, D., McMillan, L., Echt, C., Le Provost, G., Garnier-Géré, P., Cato, S., Plomion, C., 2005, Nucleotide variation in genes involved in wood formation in two pine species. New Phytologist 167:101–112.

Rafalski, A., Morgante, M., 2004, Corn and humans: recombination and linkage disequilibrium in two genomes of similar size. Trends in Genetics 20:103–111.

Robin, C., Lyman, R.F., Long, A.D., Langley, C.H., Mackay, T.F.C., 2002, hairy: a quantitative trait locus for drosophila sensory bristle number. Genetics 162:155–164.

Rust, S., Funke, H., Assmann, G., 1993, Mutagenically separated PCR (MS-PCR): a highly specific one step procedure for easy mutation detection. Nucleic Acids Research 21:3623–3629.

Schwarz, G., Sift, A., Wenzel, G., Mohler, V., 2003, DHPLC scoring of a SNP between promoter sequences of HMW glutenin x-type alleles at the Glu-D1 locus in wheat. Journal of Agricultural and Food Chemistry 51:4263–4267.

Semon, M., Nielsen, R., Jones, M.P., McCouch, S.R., 2005, The population structure of African cultivated rice *Oryza glaberrima* (Steud.): evidence for elevated levels of linkage disequilibrium caused by admixture with *O-sativa* and ecological adaptation. Genetics 169:1639–1647.

Shen, R., Fan, J.-B., Campbell, D., Chang, W., Chen, J., Doucet, D., Yeakley, J., Bibikova, M., Garcia, E.W., McBride, C., Steemers, F., Garcia, F., Kermani, B.G., Gunderson, K., Oliphant, A., 2005, High-throughput SNP genotyping on universal bead arrays. Mutation Research – Fundamental and Molecular Mechanisms of Mutagenesis 573:70–82.

Syvänen, A.-C., 1999, From gels to chips: 'Minisequencing' primer extension for analysis of point mutations and single nucleotide polymorphisms. Human Mutation 13:1–10.

Syvänen, A.-C., 2001, Accessing genetic variation: genotyping single nucleotide polymorphisms. Nature Reviews Genetics 2:930–942.

Syvänen, A.-C., 2005, Toward genome-wide SNP genotyping. Nature Genetics 37:S5–S10.

Syvänen, A.-C., Aalto-Setala, K., Harju, L., Kontula, K., Soderlund, H., 1990, A primer-guided nucleotide incorporation assay in the genotyping of apolipoprotein E. Genomics 8:684–692.

Tang, K., Fu, D.-J., Julien, D., Braun, A., Cantor, C.R., Köstek, H., 1999, Chip-based genotyping by mass spectrometry. Proceedings of the National Academy of Sciences of the United States of America 96:10016–10020.

Taylor, J.D., Briley, D., Nguyen, Q., Long, K., Iannone, M.A., Li, M.-S., Ye, F., Afshari, A., Lai, E., Wagner, M., Chen, J., Weiner, M.P., 2001, Flow cytometric platform for high-throughput single nucleotide polymorphism analysis. BioTechniques 30:661–669.

Telenius, H., Carter, N.P., Bebb, C.E., Nordenskjold, M., Ponder, B.A.J., Tunnacliffe, A., 1992, Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer. Genomics 13:718–725.

Tobe, V.O., Taylor, S.L., Nickerson, D.A., 1996, Single-well genotyping of diallelic sequence variations by a two-color ELISA-based oligonucleotide ligation assay. Nucleic Acids Research 24:3728–3732.

Törjek, O., Berger, D., Meyer, R.C., Müssig, C., Schmid, K.J., Sörensen, T.R., Weisshaar, B., Mitchell-Olds, T., Altmann, T., 2003, Establishment of a high-efficiency SNP-based framework marker set for *Arabidopsis*. Plant Journal 36:122–140.

Tyrka, M., Blaszczyk, L., Chelkowski, J., Wisniewska, H., Lind, V., Kramer, I., Weilepp, M., Ordon, F., 2004, Development of the single nucleotide polymorphism marker of the wheat *Lr1* leaf rust resistance gene. Cellular and Molecular Biology Letters 9:879–889.

Wenzl, P., Carling, J., Kudrna, D., Jaccoud, D., Huttner, E., Kleinhofs, A., Kilian, A., 2004, Diversity Arrays Technology (DArT) for whole-genome profiling of barley. Proceedings of the National Academy of Sciences of the United States of America 101:9915–9920.

Chapter 6
SNP APPLICATIONS IN PLANTS

Jacqueline Batley[1] and David Edwards[1]

6.1 INTRODUCTION

The development of high-throughput methods for the detection of single nucleotide polymorphisms (SNPs) and small indels (insertion/deletions) has led to a revolution in their use as molecular markers. SNPs are increasingly becoming the marker of choice in genetic analysis and are used routinely as markers in agricultural breeding programs (Gupta *et al.* 2001). SNPs have uses in plants for many molecular genetic marker applications. These applications include high-resolution genetic map construction, linkage disequlibrium-based association mapping, genetic diagnostics, genetic diversity analysis, cultivar identification, phylogenetic analysis, and characterization of genetic resources (Rafalski 2002a). The applications of SNPs in crop genetics have been extensively reviewed by Rafalski (2002a, b) and Gupta *et al.* (2001). However, these reviews highlight that for several years SNPs will coexist with other marker systems. The use of SNPs will become more widespread with the increasing availability of crop genome sequence, the reduction in cost, and the increased throughput of SNP assays.

DNA sequence differences are the basic requirement for the study of molecular genetics. The assay utilized to genotype polymorphisms is dependent on the technology and sequence information available. The hybridization method of restriction fragment length polymorphism (RFLP) has largely been superseded by amplification-based technologies following the advent of polymerase chain reaction (PCR). Random amplified polymorphic DNAs (RAPDs) and amplified fragment length polymorphisms (AFLPs) are frequently used to produce many markers from a single reaction without prior knowledge of the sequence of genomes of interest. Furthermore, diversity arrays technology (DArT) can detect and type DNA variation at several hundred genomic loci in parallel without relying on sequence information (Wenzl *et al.* 2004).

[1] Primary Industries Research Victoria, Victorian AgriBiosciences Centre, La Trobe R&D Park, Bundoora, Victoria 3083, Australia

These techniques may be powerful for amplifying single loci within a single reaction and for assessing genetic diversity or genetic mapping in a species where limited sequence information is available. However the markers are anonymous. Furthermore, unlike RAPDs and RFLPs, direct SNP assays provide the exact nature of the allelic variants. The genetic analysis of SNPs is gaining interest, due to the ever-increasing availability of sequence data, revealing their abundance. This abundance allows the construction of very high-density genetic maps, offering the potential to detect associations between allelic forms of a gene and observed phenotypes. SNPs are far more prevalent than microsatellites, and therefore, may provide a high-density of markers at a locus of interest. The abundance of SNPs offsets the disadvantage of bi-allelism, compared to the multi-allelic nature of microsatellites. The low mutation rate of SNPs also makes them excellent markers for studying complex genetic traits and as a tool for the understanding of genome evolution (Syvanen 2001). SNPs may be used to interrogate haplotype structure and can be applied for linkage disequilibrium (LD) studies. This application is described in detail in Chapter 9.

Table 6.1. A comparison of features and applications for AFLP, RFLP, SSR, and SNP molecular genetic markers. Y = Yes, N = No. Scores are based on the authors' experience with these markers

Applications	**RFLP**	**SSR**	**SNP**	**AFLP**
Genetic mapping	Y	Y	Y	Y
Comparative mapping	Y	Y	Y	
Framework mapping	Y	Y	Y	Y
Region-specific marker saturation			Y	Y
Map-based gene cloning	Y	Y	Y	
Bulk segregant analysis	Y	Y	Y	Y
Marker-assisted selection	Y	Y	Y	
Varietal/line identification	Y	Y	Y	Y
Genetic diversity studies		Y	Y	Y
Novel allele detection	Y	Y	Y	
Features (1 = poor, 10 = good)				
Loci density	7	3	10	8
Speed of assay	1	7	8	8
Capacity for automation	1	7	7	7
Robustness	7	7	8	8
Polymorphism level	6	8	6	9
Capacity for multiplexing	1	9	10	3
Quantity/quality DNA required	1	10	10	4
Cross-species amplification	8	5	5	2
Sequence information required	5	1	2	10
Multi-allelic	8	7	2	8
Cost-effective per assay	1	7	5	9
Technique is patented	N	N	N	Y
Codominant	Y	Y	Y	N

In the following paragraphs, we will outline the various applications of the different marker systems in genetic studies while a comparison of their features and utility is presented in Table 6.1.

6.2 GENETIC DIVERSITY

Information on genetic diversity and relationships among lines and varieties is of importance to plant breeders for the improvement of crop plants. This knowledge is valuable for germplasm conservation, inbred line identification and assignment to heterotic groups, and planning crosses in line and hybrid experiments. A knowledge of genetic diversity is also valuable for the identification of novel alleles which may then be introgressed into elite backgrounds within breeding programs.

Molecular markers are powerful tools to assess genetic variation within and between populations of plants. Previously, assessment of diversity on a genome wide scale was based on marker systems such as AFLPs, SSRs, or isozymes (Vigouroux *et al.* 2005). SNPs have been applied to assess diversity within specific genes or genomic regions for a number of years, and the results have been extrapolated to infer phylogenetic relationships between species. However, the advent of high-throughput SNP technology enables SNP-based genetic diversity assessment on a genome wide scale. Osman *et al.* (2003) used allele-specific oligonucleotide hybridization to detect polymorphisms among different accessions of Tongkat Ali (*Eurycoma longifolia*), a plant used in herbal remedies and health supplements. Forty seven plants from six geographic regions of Malaysia were studied. An average of 64% loci were polymorphic, and the populations were found to exhibit a high degree of diversity. Thus these SNPs will prove useful in preserving diversity in domesticated populations. In maize, genetic diversity was studied using SNPs at 21 loci along chromosome 1 (Tenaillon *et al.* 2002). This study facilitated an understanding of the forces contributing to genetic diversity in maize. SNPs have also been used for cultivar identification in malting barley (Dusabenyagasani *et al.* 2003) and wheat cultivars (Kirkpatrick *et al.* 2002). These assays could also be applied to distinctness, uniformity and stability testing and assessment of plant breeder's rights (Chiapparino *et al.* 2004).

6.3 PHYLOGENETIC AND EVOLUTIONARY ANALYSIS

Plant phylogenetic and evolutionary studies have traditionally relied on sequence diversity, and therefore SNPs, in genes of interest. Nuclear and chloroplast genes are a rich source of phylogenetic information for evolutionary analysis in plants, where the diversity of the sequence and genotyping of these SNPs can be used to infer phylogenetic and evolutionary relationships in a wide variety of species. Traditionally, genes or genomic fragments have been PCR amplified and resequenced in a wide variety of lines. Through analysis of SNP diversity and conservation between sequences from individuals, inheritance may be inferred. By considering rates of mutation, a molecular clock may also be applied to estimate the timing of species divergence. Molecular phylogenetics had recently also been applied in a study of maize genome evolution. Through a comparison of the terminal inverted repeats of transposons in regions of the maize genome, the order and timing of waves of historical transposon activity in this species has been elucidated (SanMiguel *et al.* 1998). Increasing quantities of sequence and SNP data for genes in a

wide variety of species is slowly uncovering the molecular mechanisms of evolution within genomes and between species. It is possible to utilize other molecular markers for phylogenetic analysis, however, without the knowledge of the sequence variation, degrees of similarity only can be assessed and homoplasy cannot be ruled out.

6.4 GENETIC MAPPING

Genetic studies involving linkage mapping, map-based positional cloning, and QTL mapping require data from large sets of genetic markers. The abundance of SNPs, combined with methods for their high-throughput discovery and detection, makes them suitable markers for these applications. SNPs identified within ESTs or large genomic fragments maintained within bacterial artificial chromosomes (BACs) can be applied for genetic mapping of complex traits. This enables the genetic mapping of specific genes of interest and assists in the identification of linked or perfect markers for traits, as well as increasing density of markers on genetic maps (Rafalski 2002b). BAC SNP markers also allow the integration of genetic and physical maps.

SNPs can be used to develop haplotyping systems for genes or regions of interest (Rafalski 2002a). The information provided by SNPs is useful when several SNPs define haplotypes in the region of interest. Only a small subset is then required to define the haplotype, and therefore need to be assayed. The use of SNPs for identifying haplotype structure, and subsequent uses for LD studies, will be covered in Chapter 9.

SNPs can be applied for genetic mapping, positional cloning, QTL mapping, and association mapping. When SNPs are applied for high-resolution genetic mapping, they can enable the development of saturated genetic maps. This has been demonstrated both on the large- and small-scale, in both model and less widely grown crop species. If a whole genome scan is to be undertaken, trait mapping by allele association requires high marker density which can readily be provided by SNPs. A genome wide set of SNP markers in *Arabidopsis thaliana* has been identified for these purposes (Schmid *et al.* 2003). Alternatively, a targeted approach may be undertaken for the mapping of candidate genes or the fine mapping of specific genomic regions which may have previously been identified through QTL mapping.

The use of SNPs to genetically map genes has also been demonstrated by Ching and Rafalski (2002). This research showed that abundance of SNPs makes them useful for placing ESTs or candidate genes onto a genetic map, which has been previously constructed with other markers. Previously, mapping ESTs predominantly involved using RFLPs or by CAPS, both of which require the presence of restriction enzyme polymorphisms. The use of SNPs for gene mapping has a further advantage in that this approach can be gene-specific, whereas RFLPs frequently assay multiple loci. Zhu *et al.* (2003) characterized SNPs and studied LD in soybean. A further objective of the research was to develop a strategy for SNP discovery, for the development of a SNP-based soybean linkage map. This would create a transcript map for soybean with candidate genes to associate with quantitative trait loci. A high-density transcript map of barley is being produced (Kota *et al.* 2001), and this will facilitate alignment of existing linkage maps in barley and permit identification of ESTs associated with traits of interest. Moreover, the SNPs can be used for syntenic studies with other related species. As an example of this approach in a minor crop, five SNPs were genetically mapped in melon, using three different genotyping assays (Morales *et al.* 2004). Genetic mapping of genes

and BAC end sequences using SNPs has also been performed in cassava (Lopez *et al.* 2005). SNPs are being applied in maize for the generation of a high-resolution genetic map, which will act as a framework to anchor BAC contigs. This data is being managed in a database of the maize community (Sanchez-Villeda *et al.* 2003).

6.5 MARKER-ASSISTED SELECTION

One of the most often cited benefits of genetic markers for plant breeding has been their use in marker-assisted selection (MAS), exploiting the markers as selection tools in crop breeding programs (Koebner and Summers 2002). This allows the breeder to achieve early selection of a trait, or a combination of traits. This is particularly useful when the trait concerned is under complex genetic control, or when field trials are unreliable or expensive. By increasing favorable allele frequency early in the breeding process, a larger number of small populations can be carried forward in the breeding process, each of which has been prescreened to remove or reduce the frequency of unfavorable alleles.

Molecular markers are 100% heritable, therefore using these markers to select for a low heritable trait is more effective and less expensive than phenotypic selection for that trait. Molecular markers are essential for the mapping of candidate genes, marker-assisted breeding, and the map-based cloning of genes underlying traits. Marker-assisted breeding has previously utilized molecular markers such as RAPDs, RFLPs, AFLPs, CAPS, and microsatellites (SSRs). However, these marker systems are frequently labor intensive and time-consuming and the associated costs constrain the ability to perform high-throughput genotyping on breeding populations or germplasm. RFLPs are particularly unsuitable for large-scale MAS due to the high cost implications of their implementation to screen large numbers of individual plants. PCR-based markers are preferable to RFLPs due to their potential for high-throughput and reduced costs. PCR-based methods only require small quantities of DNA and are therefore suitable for the screening and selection of plants at early seedling stages. However, the application of each PCR-based marker technology may have limitations. Many PCR-based technologies are impractical for use as MAS tools, as they are either too complex for automation (AFLPs), or demonstrate poor reproducibility (RAPDs). Genotyping with CAPS requires the use of a restriction endonuclease and is therefore dependent on a polymorphism in the restriction site. AFLPs are anonymous markers. SSRs are a useful tool for MAS, however the markers are often only loosely linked to the polymorphism responsible for the trait, rather than being 100% diagnostic. Markers loosely linked to a trait may suffer from recombination between the marker and the gene. Linked markers are also not usually transferable between populations originating from different parents, due to lack of polymorphism. Markers within the gene responsible for the trait are considered perfect markers. These are highly valuable for breeding as the possibility of recombination between the marker and gene is essentially eliminated and they are frequently transferable between populations. SSRs suffer from homoplasy (alleles which are identical by size, but not by descent) making them less suitable than SNPs for MAS studies. The abundance of SNPs in plant genomes makes them attractive tools for MAS and map-based cloning and SNPs and indel molecular markers can be applied for MAS.

SNPs are highly stable markers which may contribute directly to phenotype and they can serve as a powerful tool for MAS. Once SNP markers are found to be associated

with a target trait, they can be applied by plant breeders for MAS to identify individual plants containing a combination of alleles of interest from large segregating populations. SNPs can be identified within or in close proximity to genes underlying agronomic traits. Although the SNP may not be responsible for the mutant phenotype, they may be applied for MAS and for the positional cloning of the gene in question (Gupta *et al.* 2001). Association of SNPs with genes of economic value has already been demonstrated. SNP markers for supernodulation in soybean have been identified (Kim *et al.* 2005). The identified SNP in the *GmNARK* gene indicates the presence of the hypernodulating mutation. The SNP was converted to a single nucleotide amplified polymorphism (SNAP) marker to allow direct MAS for supernodulation at an early growth stage without the need to inoculate and phenotype roots.

ESTs have been utilized in sugarcane for the identification of SNP markers associated with the *Adh* genes (Grivet *et al.* 2003). The *Adh* gene family encodes a key enzyme, alcohol dehydrogenase, in the glycolytic pathway and is well characterized in a number of plant species, providing an ideal model for SNP discovery and analysis. These demonstrate the principles of the application in sugarcane and can be used for genetic mapping and QTL analysis as well as for MAS.

A high-throughput SNP genotyping system has been developed and used to select barley alleles carrying superior alleles of β-amylase, a key enzyme involved in the degradation of starch during the malting process (Paris *et al.* 2002). The four allelic forms of the enzyme were unambiguously identified by genotyping two SNPs using the SnuPE system. A CAPS marker has also been developed enabling the transfer of the marker to other laboratories which do not have SnuPE assay capabilities. These assays provide a rapid and inexpensive method for screening large numbers of individual plants, allowing the introgression of the desirable allele into breeding programs. Further work on MAS using SNPs in barley include identification of SNPs in the *Isa* gene, which has a likely role in defense against pathogens. This gene was sequenced and screened for SNPs across 16 genotypes (Bundock and Henry 2004). This study showed there is little diversity in cultivated barley and that SNPs could be a useful tool for the introduction of novel alleles from wild barley. Furthermore, SNPs associated with grain germination have been characterized across 23 varieties (Russell *et al.* 2004) for their suitability for implementation in MAS.

A SNP marker has been developed for the waxy gene controlling amylose content in rice. Amylose is the main component controlling the cooking and nutritional properties of cereals. Low amylose varieties are considered desirable, and in rice, it has been shown that the high and low amylose types can be differentiated based on a SNP near the waxy gene. This marker will be applied for MAS for the low amylose trait in seedlings (Gupta *et al.* 2001). Further SNPs associated with important genes in rice include a SNP marker for the dwarfing gene. The SNP was identified within an SSR flanking sequence and is used for selection in a wide range of crosses. SNP-based markers for rice-blast resistance genes have also been developed (Hayashi *et al.* 2004). These markers enabled the mapping of the *Piz* and *Piz-t* genes, demonstrating that the SNPs are a valuable tool for gene mapping, map-based cloning and MAS in rice.

In wheat, the SNP found to alter the protein structure of adenine phosphoribosyl transferase has been identified (Xing *et al.* 2005). This gene encodes the key enzyme which converts adenine to adenosine monophosphate in the purine salvage pathway. In wheat, further SNPs in genes of interest have been identified, including the *Lr1* leaf rust resistance gene (Tyrka *et al.* 2004). Infections can lead to severe yield losses and

therefore the desire is to grow resistant cultivars. The development of the SNP marker in the *Lr1* gene has been a dramatic improvement on the STS marker previously used, which was not specific in 50% of cultivars tested. The growing number of wheat SNP markers available will open the possibility of introducing multiplexed assays, targeting loci to pyramid trait selection during wheat breeding.

Work has also been performed on MAS in less developed crop species. One hundred and thirty-two SNPs in quinoa have been identified from ESTs (Coles *et al.* 2005). It was found that the SNP development from ESTs was a practical method for developing species-specific markers and may provide the molecular differentiation required to monitor gene flow between cultivated quinoa and weedy species. Furthermore, these will prove valuable in MAS projects aimed at improving quinoa via exotic gene introgression. Further potential applications in plants include the results of a study of nucleotide diversity in the *pal1* locus of Scots pine (Dvornyk *et al.* 2002). This gene is predicted to be associated with ozone tolerance, pathogen defense, and metabolism of exogenous compounds, and SNPs within it could prove valuable for MAS in this species.

6.6 CONCLUDING REMARKS

SNPs are increasingly becoming the marker of choice for a wide range of applications including genetic mapping, MAS, and diversity analysis. As the availability of SNPs increases, they are displacing other forms of molecular markers for these applications. As costs associated with SNP discovery and detection continue to fall, SNPs will increasingly be associated with agronomic traits and will be applied for crop improvement through parental selection and MAS. Of the marker systems available, each has their own benefits and limitations. AFLPs are anonymous markers and do not provide sequence information, RFLPs are time-consuming and laborious, SSRs have the benefit that they are transferable between related organisms. However, for LD studies SNPs have significant advantages over SSRs, due to their greater frequency and specificity in the genome.

6.7 REFERENCES

Bundock, P.C., Henry, R.J., 2004, Single nucleotide polymorphism, haplotype diversity and recombination in the *Isa* gene of barley. Theor. Appl. Genet. 109:543–551.

Chiapparino, E., Lee, D., Donini, P., 2004, Genotyping single nucleotide polymorphisms in barley by tetra-primer ARMS-PCR. Genome 47:414–420.

Ching, A., Rafalski, A., 2002, Rapid genetic mapping of ESTs using SNP pyrosequencing and indel analysis. Cell. Mol. Biol. Lett. 7:803–810.

Coles, N.D., Coleman, C.E., Christensen, S.A., Jellen, E.N., Stevens, M.R., Bonifacio, A., Rojas-Beltran, J.A., Fairbanks, D.J., Maughan, P.J., 2005, Development and use of an expressed sequenced tag library in quinoa (*Chenopodium quinoa* Willd.) for the discovery of single nucleotide polymorphisms. Plant Sci. 168:439–447.

Dusabenyagasani, M., Perry, D., Lee S.-J., Demeke, T., 2003, Genotyping malting barley varieties registered in Canada with SNP markers. In: XI Plant and Animal Genome Meeting, San Diego, CA.

Dvornyk, V., Sirviö, A., Mikkonen, M., Savolainen, O., 2002, Low nucleotide diversity at the *pal*1 locus in the widely distributed *Pinus sylvestris*. Mol. Biol. Evol. 19:179–188.

Grivet, L., Glaszmann, J.-C., Vincentz, M., da Silva, F., Arruda, P., 2003, ESTs as a source for sequence polymorphism discovery in sugarcane: example of *Adh* genes. Theor. Appl. Genet. 106:190–197.

Gupta, P.K., Roy, J.K., Prasad, M., 2001, Single nucleotide polymorphisms: a new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plants. Curr. Sci. 80:524–535.

Hayashi, K., Hashimoto, N., Daigen, M., Ashikawa, I., 2004, Development of PCR-based SNP markers for rice blast resistance genes at the *Piz* locus. Theor. Appl. Genet. 108:1212–1220.

Kim, M.Y., Van, K., Lestari, P., Moon, J.-K., Lee S.-H., 2005, SNP identification and SNAP marker development for a GmNARK gene controlling supernodulation in soybean. Theor. Appl. Genet. 110:1003–1010.

Kirkpatrick, R., Somers, D.J., Moniwa, M., Walsh, A., Riemer, E., 2002, Variety identification using single nucleotide polymorphisms in hexaploid wheat. In: X Plant and Animal Genome Meeting, San Diego, CA.

Koebner, R., Summers, R., 2002, The impact of molecular markers on the wheat breeding paradigm. Cell. Mol. Biol. Lett. 7:695–702.

Kota, R., Varshney, R.K., Thiel, T., Dehmer, K.J., Graner, A., 2001, Generation and comparison of EST derived SSRs and SNPs in barley (*Hordeum vulgare* L.). Hereditas 135:145–151.

Lopez, C., Piegu, B., Cooke, R., Delseny, M., Tohme, J., Verdier, V., 2005, Using cDNA and genomic sequences as tools to develop SNP strategies in cassava (*Manihot esculenta* Crantz). Theor. Appl. Genet. 110:425–431.

Morales, M., Roig, E., Monforte, A.J., Arús, P., Garcia-Mas, J., 2004, Single-nucleotide polymorphisms detected in expressed sequence tags of melon (*Cucumis melo* L.). Genome 47:352–360.

Osman, A., Jordan, B., Lessard, P.A., Muhammad, N., Haron, M.R., Riffin, N.M., Sinskey, A.J., Rha, C., Housman, D.E., 2003, Genetic diversity of *Eurycoma longifolia* inferred from single nucleotide polymorphisms. Plant Physiol. 131:1294–1301.

Paris, M., Jones, M.G.K., Eglinton, J.K., 2002, Genotyping single nucleotide polymorphisms for selection of barley β-amylase alleles. Plant Mol. Biol. Rep. 20:149–159.

Rafalski, J.A., 2002a, Novel genetic mapping tools in plants: SNPs and LD-based approaches. Plant Sci. 162:329–333.

Rafalski, J.A., 2002b, Applications of single nucleotide polymorphisms in crop genetics. Curr. Opin. Plant Biol. 5:94–100.

Russell, J., Booth, A., Fuller, J., Harrower, B., Hedley, P., Machray, G., Powell, W., 2004, A comparison of sequence-based polymorphism and haplotype content in transcribed and anonymous regions of the barley genome. Genome 47:389–398.

Sanchez-Villeda, H., Schroeder, S., Polacco, M., McMullen, M., Havermann, S., Davis, G., Vroh-bi, I., Cone, K., Shrapova, N., Yim, Y., Scultz, L., Duru, N., Musket, T., Houchins, K., Fang, Z., Gardiner, J., Coe, E., 2003, Development of an integrated laboratory information management system for the maize mapping project. Bioinformatics 19:2022–2030.

SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y., Bennetzen, J.L., 1998, The paleontology of intergene retrotransposons of maize. Nat. Genet. 20:43–45.

Schmid, K.J., Rosleff Sörensen, T., Stracke, R., Törjék, O., Altmann, T., Mitchell-Olds, T., Weisshaar, B., 2003, Large-scale identification and analysis of genome wide single nucleotide polymorphisms for mapping in *Arabidopsis thaliana*. Genome Res. 13:1250–1257.

Syvanen, A.C., 2001, Accessing genetic variation: genotyping single nucleotide polymorphisms. Nat. Rev. Genet. 2:930–942.

Tenaillon, M.I., Sawkins, M.C., Anderson, L.K., Stack, S.M., Doebley, J., Gaut, B.S., 2002, Patterns of diversity and recombination along Chromosome 1 of maize (*Zea mays* ssp. *mays* L.). Genetics 162:1401–1413.

Tyrka, M., Blaszczyk, L., Chelkowski, J., Lind, V., Kramer, I., Weilepp, M., Wisniewska, H., Ordon, F., 2004, Development of the single nucleotide polymorphism of the wheat *Lr1* leaf rust resistance gene. Cell. Mol. Biol. Lett. 9:879–889.

Vigouroux, Y., Mitchell, S., Matsuoka, Y., Hamblin, M., Kresovich, S., Smith, S.C., Jaqueth, J., Smith, O.S., Doebley, J., 2005, An analysis of genetic diversity across the maize genome using microsatellites. Genetics 169:1617–1630.

Wenzl, P., Carling, J., Kudrna, D., Jaccoud, J., Huttner, E., Kleinhofs, A., Killian, A., 2004, Diversity arrays technology (DArT) for whole genome profiling of barley. Proc. Natl Acad. Sci. USA 101:9915–9920.

Xing, Q., Ru, Z., Li, J., Zhou, C., Jin, D., Sun, Y., Wang, B. 2005, Cloning a second form of adenine phosphoribosyl transferase gene (TaAPT2) from wheat and analysis of its association with thermosensitive genic male sterility (TGMS). Plant Science, 169: 37−45.

Zhu, Y.L., Song, Q.J., Hyten, D.L., van Tassell, C.P., Matukumalli, L.K., Grimm, D.R., Hyatt, S.M., Fickus, E.W., Young, N.D., Cregan, P.B., 2003, Single nucleotide polymorphisms in soybean. Genetics 163:1123–1134.

Chapter 7
LINKAGE DISEQUILIBRIUM MAPPING CONCEPTS

H. Nihal De Silva[1] and Roderick D. Ball[2]

7.1 INTRODUCTION

In this section we introduce the basic statistical concepts needed for association mapping.

If the data is good enough statistical analysis is hardly needed. Hence the quote:

> "If your experiment needs statistics you ought to have done a better experiment." (Rutherford)

The physicist Rutherford did not have to contend with biological variation. If your experiment is not quite good enough for Rutherford, any simple summary statistics would still suffice. Any effects would be large compared to their standard errors. This is not the case for most biological experiments.

Mathematically we can view the genome as a disjoint set of lines, one for each chromosome, with a distance measured along each line, such that the distance between any two loci is related to the per meiosis probability of recombination in the interval between the loci. The implications of this structure for gene mapping are:

- That *linkage maps* can be constructed with a set of markers and distances between markers estimated; location in the genome can be specified in terms of position in an interval between markers on the linkage map.
- Within a family, if there is a causal locus affecting the trait, then linked markers will also be associated with the trait, with effect size reduced by $(1 - 2r)$, where r is the recombination rate between the two loci.
- Within a population, historical linkage disequilibrium between loci will be reduced by $(1 - 2r)$ per generation of random mating.

Using the genome map and samples from populations or families we can obtain information on marker locations and on locations of QTL, i.e. locations on the genome contributing to variation in a trait, where the genotypes are not directly observed.

[1] The Horticulture & Food Research Institute Limited (HortResearch), Mt Albert Research Research Centre, 120 Mt Albert Road, P.B. 92169, Auckland, New Zealand.
[2] Ensis (New Zealand Forest Research Institute Limited), 49 Sala Street, P. B. 3020, Rotorua, New Zealand.

Statistical methods need to take into account this genome structure. For example, marker–trait associations give information on trait loci, but statistical estimates and tests for marker–trait associations at nearby markers will be correlated, and hence there is no independent evidence for an association.

In association mapping, the effect being tested is only a small proportion of the total variation in a trait, and to make matters more difficult the loci affecting a given trait are only several among thousands or even hundreds of thousands of possible candidates found from the set of all possible loci, markers or genes in the genome. Use of common statistical methods has lead to many published spurious associations (Chapter 8). Large costly experiments are needed. Therefore we have to be careful with the use of statistics, and make the best possible use of available data.

There are two main schools of statistics, frequentist and Bayesian. The frequentist considers the sampling properties of real-valued random variables or "statistics", while the Bayesian approach uses probability theory to obtain probability distributions for unknown parameters. A critical comparison of Bayesian and frequentist approaches to statistical estimation and inference for association mapping, will be given throughout this chapter and Chapter 8, and will show that there are substantial differences between methods that can lead to spurious associations. In the rest of this section we introduce the concepts of statistical estimation and inference illustrated by results and interpretation in simple examples, for linkage disequilibrium estimation and association mapping. A comprehensive range of methods is given in more detail in Chapter 8.

7.2 QTL AND LD MAPPING COMPARED

Similarities and differences between QTL and LD mapping are summarised in Table 7.1. The main differences are that LD mapping detects historical linkage disequilibrium generated in a population while QTL mapping detects linkage disequilibrium generated within a family or pedigree.

In a population, recombinations affecting the association between a gene and a marker may occur over many generations. This potentially gives a much finer resolution for mapping QTL than pedigrees used for traditional QTL mapping (linkage analysis), where recombinations occur over at most several generations. The extent of LD in the population varies between species and populations (cf. Chapter 8, Section 8.3.1), but may be as small as 4 kb. However, achieving the potential resolution may require many markers to cover the genome and large sample sizes especially where the extent of LD is small. (cf. the power calculations in Chapter 8, Section 8.3.2.)

LD mapping is based on a random population sample, which is *observational data*. In any observational study associations found may not be causal – associations may be due to correlation with unmeasured causal factors or population structure.

A QTL mapping family is equivalent to a designed experiment. In a classical designed experiment, treatments are randomly assigned to experimental units, randomising the effects of extraneous factors, giving unbiased, and with sufficient sample sizes, accurate, estimates of treatment effects. A QTL mapping family can be thought of as a random sample from the set of possible progeny. In each meiosis, recombinations occur randomly, randomising the effects of unlinked markers, and generating associations between pairs of loci

Table 7.1. QTL mapping and LD mapping compared

QTL	LD
Equivalent to a designed experiment.[1] • Every individual genotyped and phenotyped. • Potentially affected by undetected or un-modelled loci. • Randomisation generated by experimental crosses. • Association between a marker and QTL depends only on recombination distance between the marker and QTL.	Observational study. • Information on recombinations generated in un-genotyped ancestral genealogy inferred from final generation. • Potentially affected by undetected or un-modelled loci or population structure. • Population-level LD depends on age of mutations, population history and recombination distance between markers.
Detects within family LD generated in pedigree. • Marker and trait loci probably in linkage equilibrium in the population. • Linkage phase of marker–QTL associations may vary between families, and needs to be verified for each family if used for marker-aided selection (MAS). Otherwise selecting on marker genotypes may choose the unfavourable allele. Necessity to verify reduces the potential benefit of early selection for long lived species (e.g. forest trees). • Any given QTL may not be segregating in some families. Many QTL allele-trait associations will be missed in single family QTL mapping studies.	Detects LD generated historically in population. • Marker and trait loci in linkage disequilibrium. Marker–trait association should persist across families. • Reduced necessity to verify associations – verification for MAS could be limited to several families. • With sufficient sample size, depending on marker spacings, allele frequencies and effect sizes, any QTL allele can be detected.
Lower resolution (typically 1–50 cM, depending on marker density, sample size and QTL heritability). • Too many base pairs in QTL interval for brute force sequencing. • Typically of the order of 100 markers to cover the genome. • Prior odds of the order of 1/10 per marker.	Potentially finer resolution (down to 4 kb depending on extent of LD in the population, marker density, sample size and QTL heritability). • If the extent of LD is not too high, the LD region can be sequenced to locate and clone genes. • Potentially hundreds of thousands of SNP markers to cover the genome. • Prior odds of the order of 1/50,000 per marker or candidate gene.

depending only on the recombination distance between the loci. Effects of unlinked markers are randomised. The set of progeny is also in random order according to our model for meiosis. Hence, other environmental factors including maternal environment are also randomised with respect to progeny genotypes. Hence marker genotypes (for any marker or set of markers) are equivalent to treatments in a designed experiment. Hence estimates of marker–trait associations are unbiased, and with sufficient sample size, accurate.

If existing QTL mapping populations are available, information from QTL and LD mapping studies can be combined. Use of combined QTL and LD mapping populations is discussed in Chapter 8, where it is shown that a combined approach can be more efficient than LD mapping alone.

7.3 STATISTICAL ESTIMATION

Statistical estimation is the estimation of unknown parameters. The simplest example is a population mean, estimated by a sample mean. More generally, parameters may be any unknown parameters included in a statistical model for the process generating the data. Estimated quantities are indicated with a "hat," e.g. the estimate of D is denoted by $\hat{D}$. If the quantities are given by Greek letters the sample estimate may also be denoted using the corresponding Roman letter, e.g. the estimate of σ^2 may be denoted by s^2. In addition to the estimate, we usually require some measure of variability, e.g. a standard deviation of the estimate.

In Example 7.1 we demonstrate estimating the linkage disequilibrium coefficient D. Estimates are obtained, first by a simple plug-in approach, then using maximum likelihood estimation. Standard errors and 95% confidence intervals are obtained from the maximum likelihood method.

Example 7.1. Estimating D.

Suppose the observed counts from a sample of size $n = 100$ individuals from a population were as in Table 7.2.

The linkage equilibrium coefficient D (cf. Chapter 2) is defined as:

$$D = \Pr(A, T) - \Pr(A)\Pr(T)\,. \tag{7.1}$$

We can estimate $\Pr(A)$, $\Pr(T)$ and $\Pr(A, T)$ from the sample proportions $n_A/100$, $n_T/100$, $n_{AT}/100$, respectively, and plug these values into Equation (7.1) to solve for D:

$$\hat{p}_A = n_A/n = 0.1,\ \hat{p}_T = n_T/n = 0.3,\ \hat{D} = n_{AT}/n - \hat{p}_A\hat{p}_T = 0.07\,. \tag{7.2}$$

Table 7.2. Observed counts for two loci from a sample of 100

	A	*G*	Total
T	$n_{AT} = 10$	$n_{GT} = 20$	$n_T = 30$
C	$n_{AC} = 0$	$n_{GC} = 70$	$n_C = 70$
Total	$n_A = 10$	$n_G = 90$	$n = 100$

We refer to this as the "plug-in" estimate. From Equation (7.1), using the delta-method (cf Weir 1996, Chapter 2), an approximate variance and standard error for $\hat{D}$ are calculated as:

$$\begin{aligned}
\text{var}(\hat{D}) &\approx \text{var}(\hat{p}_{AT}) + p_T^2\text{var}(\hat{p}_A) + p_A^2\text{var}(\hat{p}_T) - 2p_A\text{cov}(\hat{p}_{AT}, \hat{p}_T) \\
&\quad -2p_T\text{cov}(\hat{p}_{AT}, \hat{p}_A) + 2p_Ap_T\text{cov}(\hat{p}_A, \hat{p}_T) \\
&= \frac{1}{n}p_{AT}(1-p_{AT}) + \frac{1}{n}p_T^2p_A(1-p_A) + \frac{1}{n}p_A^2p_T(1-p_T) \\
&\quad -\frac{2}{n}p_Ap_{AT}(1-p_T) - \frac{2}{n}p_Tp_{AT}(1-p_A) + \frac{2}{n}p_Ap_T(p_{AT}-p_Ap_T) \\
&= 0.00036 \\
\text{se}(\hat{D}) &= \sqrt{\text{var}(\hat{D})} = 0.019\,, \qquad (7.3)
\end{aligned}$$

where we have used: $\text{var}(\hat{p}) = \frac{1}{n}p(1-p)$, for $p = p_A, p_T, p_{AT}$, $\text{cov}(\hat{p}_A, \hat{p}_T) = \frac{1}{n}(p_{AT} - p_Ap_T)$, $\text{cov}(\hat{p}_{AT}, \hat{p}_A) = \frac{1}{n}p_{AT}(1-p_A)$, and $\text{cov}(\hat{p}_{AT}, \hat{p}_T) = \frac{1}{n}p_{AT}(1-p_T)$, and replaced p_A, p_T, p_{AT} by their estimates.

The covariances in (7.3) can be derived using *indicator variables* (cf Weir 1996, Chapter 2). Let x_i be indicator variables for the allele A at the first locus, in the ith sampled individual: $x_i = 1$ if the allele is A, and 0 otherwise. Similarly let y_i be the indicator variables for the allele T at the second locus. Then

$$\hat{p}_A = \frac{1}{n}\sum x_i, \quad \hat{p}_T = \frac{1}{n}\sum y_i, \quad \text{and} \quad \hat{p}_{AT} = \frac{1}{n}\sum x_iy_i \qquad (7.4)$$

The covariance between $\hat{p}_A$ and $\hat{p}_T$ is calculated as:

$$\begin{aligned}
\text{cov}(\hat{p}_A, \hat{p}_T) &= \frac{1}{n^2}\text{cov}\left(\sum x_i, \sum y_j\right) \\
&= \frac{1}{n^2}\sum \text{cov}(x_i, y_i) \\
&= \frac{1}{n}(p_{AT} - p_Ap_T) \qquad (7.5)
\end{aligned}$$

since $\text{cov}(x_i, y_i) = E(x_iy_i) - E(x_i)E(y_i) = p_{AT} - p_Ap_T$, and $\text{cov}(x_i, y_j) = 0$ for $i \neq j$, since alleles for different individuals are independent.

The covariance between $\hat{p}_{AT}$ and $\hat{p}_A$ is calculated as:

$$\begin{aligned}
\text{cov}(\hat{p}_{AT}, \hat{p}_A) &= \frac{1}{n^2}\text{cov}\left(\sum x_iy_i, \sum x_j\right) \\
&= \frac{1}{n}p_{AT}(1-p_A) \qquad (7.6)
\end{aligned}$$

since $\text{cov}(x_iy_i, x_i) = E(x_iy_ix_i) - E(x_iy_i)E(x_i) = p_{AT} - p_{AT}p_A$, where we have used $x_i^2 = x_i$, and since $\text{cov}(x_iy_i, x_j) = 0$ for $i \neq j$. The covariance between $\hat{p}_{AT}$ and $\hat{p}_T$ is calculated similarly.

The mean and standard error are a good estimate summary of a distribution, provided the distribution is symmetric and approximately normal. However, for $p_A = 0.1$, $p_T = 0.3$ the disequilibrium coefficient D must lie between minimum and maximum limits of -0.03

and 0.07. In this example, the standard error, 0.032, is a substantial fraction of the length of the parameter space, and $\hat{D}$ is on the boundary of the parameter space, so the sampling distribution is likely to be skewed. A mean and standard error may not be a good summary. Therefore, to calculate a confidence interval we use the method of maximum likelihood.

Maximum likelihood estimation of D. The likelihood function is

$$f(x \mid p_A, p_T, D) = \frac{100!}{10!20!0!70!} p_{AT}^{n_{AT}} p_{GT}^{n_{GT}} p_{AC}^{n_{AC}} p_{GC}^{n_{GC}} , \tag{7.7}$$

where $x = (n_{AT}, n_{GT}, n_{AC}, n_{GC}) = (10, 20, 0, 70)$ denotes the observed counts. Taking logs and dropping the initial constant term in Equation (7.7) gives the log-likelihood:

$$L = n_{AT} \log p_{AT} + n_{GT} \log p_{GT} + n_{AC} \log p_{AC} . \tag{7.8}$$

Solving for the cell probabilities $p_{AT}, p_{GT}, p_{AC}, p_{GC}$ in terms of p_A, p_T, D we obtain

$$\begin{aligned} L(p_A, p_T, D) = {} & n_{AT} \log(p_A p_T + D) + n_{GT} \log(p_T(1 - p_A) - D) + \\ & n_{AC} \log(p_A(1 - p_T) - D) + n_{GC} \log((1 - p_A)(1 - p_T) + D) . \end{aligned} \tag{7.9}$$

R calculations for maximum likelihood estimation of D are shown in Figure 7.1. We do not describe the R language in detail, only essential aspects of our code, referring the reader to the manual for further information. The key steps are to define the likelihood function `ld.loglik()` and then to maximise the likelihood. Maximisation uses the generic optimisation function `optim()`. Valid values for the parameters p_A, p_T are in unit interval, while D is constrained to lie within an interval depending on p_A, p_T. The optimisation uses an unconstrained parameterisation in terms of $\theta = (\text{logit}(p_A), \text{logit}(p_B), \text{logit}(D'))$. In the R code, D' is denoted by `D1`. The R function `ld.loglik.x()` calculates the log-likelihood in terms of the transformed parameters θ. Note that the maximum likelihood estimates agree almost exactly with the "plug-in" estimates above.

Confidence intervals and standard errors for $\hat{D}$. The log-likelihood is plotted as a function of D, with p_A, p_T set to their estimates, in Figure 7.2. Note that the ML estimate of D is on the boundary of the interval. Hence we have a *non-regular maximum likelihood estimation problem*. If the maximum of the likelihood occurred in the interior of the region a confidence interval for $\hat{D}$ would be obtained by referring $\Lambda = -2(L - L_{\max})$, to the chi-squared ($\chi^2$) distribution with 1 degree of freedom, where $L_{\max}$ is the value of the maximised likelihood. However, since the estimate is on the boundary, there is heuristically speaking only 1/2 a degree of freedom. Since we only observe one tail of the distribution, p-values for a given χ^2 value are doubled (Stram and Lee 1994). Therefore, for a 95% confidence interval we find the value of D giving a value of Λ corresponding to the 97.5% points of the χ_1^2 distribution. Confidence intervals are shown in Table 7.3. For comparison, confidence intervals were calculated in three ways (1) using a naïve χ_1^2 approximation to the likelihood, ignoring the fact that $\hat{D}$ is on the boundary of the parameter space; (2) using the adjusted χ_1^2 approximation, (Stram and Lee 1994) and (3) by numerical integration of the likelihood function. If $\hat{D}$ was close to but not on the boundary the need for adjustment would be indicated by part of the confidence interval extending beyond the parameter space. The correct limit would be somewhere between the naïve and adjusted values.

```
> # the log-likelihood function in terms of pa, pt, D.
> logit <- function(p,tol=1.0e-4){p1 <- ifelse(p<tol,tol,
+             ifelse(p>1-tol,1-tol,p)); log(p1/(1-p1))}
> anti.logit <- function(q){u <- exp(q);u/(1+u)}
> ld.loglik <- function(pa,pt,D){
+   10*log(pa*pt+D) + 20*log(pt*(1-pa) - D) +
+   70*log((1-pa)*(1-pt)+ D) }
> # the log-likelihood function re-parameterised in terms of
> # logit(pa), logit(pt), logit(D1).
> ld.loglik.x <- function(theta){
+   pa <- anti.logit(theta[1])
+   pb <- anti.logit(theta[2])
+   Dx <- anti.logit(theta[3])
+   Dmax <- min(pa*(1-pb),(1-pa)*pb)
+   Dmin <- max(-pa*pb,-(1-pa)*(1-pb))
+   D <- Dmin + (Dmax-Dmin)*Dx
+   ld.loglik(pa=pa,pt=pb,D=D) }
> pAhat <- 0.1; pThat <- 0.3;  D0 <- 0.05
> theta0 <- c(logit(pAhat),logit(pThat),
+             logit((D0 - Dmin)/(Dmax-Dmin)))
> res <- optim(theta0,function(theta){-ld.loglik.x(theta)},
+              method="Nelder-Mead")
> res$par
[1] -2.1971691 -0.8472953 14.5655742
> # back transform the parameters
> pAhat.res <- anti.logit(res$par[1])
> pThat.res <- anti.logit(res$par[2])
> Dmax.res <- min(pAhat.res*(1-pThat.res),(1-pAhat.res)*pThat.res)
> Dmin.res <- max(-pAhat.res*pThat.res,-(1-pAhat.res)*(1-pThat.res))
> Dhat <- Dmin.res + (Dmax.res - Dmin.res)*anti.logit(res$par[3])
> c(pAhat=pAhat.res,pThat=pThat.res,Dhat=Dhat)
    pAhat     pThat      Dhat
    0.100     0.300     0.070
> # calculate lower limit for 95% c.i. for D
> xp <- seq(Dmin.res,Dmax.res,length=1000)
> yp <- ld.loglik(pa=0.1,pt=0.3,D=xp)
> # avoid singular value at lower limit
> xp1 <- xp[-1]; yp1 <- yp[-1]
> # naive chi-squared on 1 d.f. 0.0532
> approx(yp1,xp1, xout=max(yp1) - qchisq(0.95,1)/2)$y
[1] 0.05324738
> # adjusted, P = 0.5 Pr(X^2 >d) where X^2 ~ chi-squared on 1 d.f.
> approx(yp1,xp1, xout=max(yp1) - qchisq(0.975,1)/2)$y
[1] 0.04886265
> # integral method, giving lower 5% point of approx. posterior
> I <- sum(exp(yp1))
> approx(cumsum(exp(yp1))/I,xp1,xout=0.05)$y
[1] 0.04766237
```

Figure 7.1. *R* output for maximum likelihood estimation of D in Example 7.1.

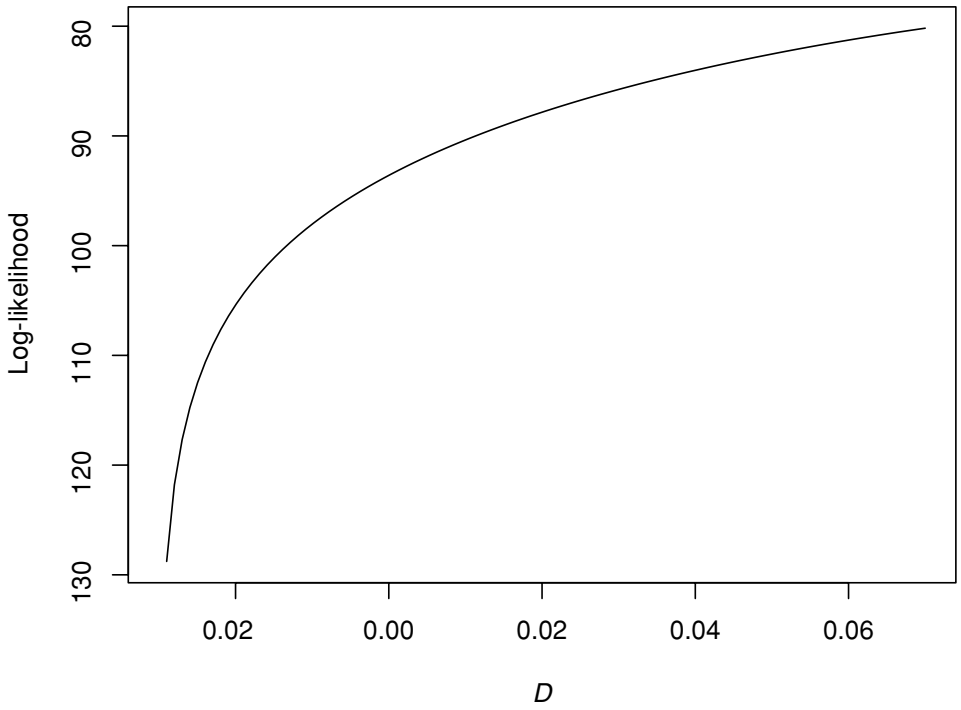


Figure 7.2. Log-likelihood versus D for Example 7.1.

Foreshadowing discussion of Bayesian methods in Chapter 8, we note that the likelihood integral method, (method 3 in Table 7.3), is equivalent to a Bayesian approach, if a uniform prior for D is assumed, and we assume the values of p_A, p_T are known. In this case the adjusted χ^2_1 and integral methods give similar answers, and the naïve approach overestimated the lower limit of the confidence interval for D by about 0.005 or 10% of its value. In general, Bayesian and likelihood methods give similar answers for estimation with large sample sizes or uniform priors. This example, perhaps unfortunate in its complexity, serves to illustrate the pitfalls and complexities inherent in the non-Bayesian approach, except in the simplest cases.

7.4 STATISTICAL INFERENCE

Statistical inference refers to quantifying the evidence for an effect. Statistical inference is important to association mapping, because we need to quantify evidence for the existence of a genetic effect at a locus in a genomic region.

The ultimate goal is to determine and locate causal factors, i.e. genes affecting variability in traits of interest. Gene mapping, either association mapping or QTL mapping, looks for statistical associations between genetic markers and traits. However, association or correlation does not imply causality. Gene mapping exploits and also must contend with the fact that physical proximity between markers and/or genes generates a correlation between

Table 7.3. Estimates and confidence intervals for D

	Method	Estimate	95% c.i.
1.	χ^2_1, naïve	0.07	0.0532–0.07
2.	χ^2_1, adjusted	0.07	0.0488–0.07
3.	Likelihood integral	0.07	0.0476–0.07

genotypes at different loci. To succeed we must first find good evidence for an association, and then use that information in further experiments or applications, e.g. sequencing a region of the genome to find a gene, or using a marker in MAS to select for superior genotypes. Since there are many possible loci, e.g. hundreds of thousands of SNP markers, and only a moderate number of genes expected to substantially influence a given trait, any given marker has an *a priori* low probability of being closest to the true gene.

The statistical problem is to identify which subsets of markers are likely to be close to the genes and/or have good predictive value. This evidence should enable us to make effective or optimal decisions. Hence, statistical methods for quantifying evidence for associations and comparing different possible subsets (or the corresponding models) are critical. The large number of published spurious associations, discussed in Chapter 8, illustrates the need for more rigorous statistical evidence. In this subsection we review methods of statistical inference, which are used to quantify evidence for associations.

7.4.1 Frequentist Inference

Frequentist measures of evidence include the p-value from hypothesis tests and the type I error rate, usually denoted by α. For testing scientific hypotheses, generally two models are compared: the *null hypothesis*, e.g. $H_0 : \theta = 0$; and the *alternative hypothesis*, e.g. $H_1 : \theta \neq 0$. A test statistic measuring departures from the null hypothesis is chosen. The p-value is defined as the probability of observing more extreme values of the test statistic under H_0 than the observed value: e.g. if T is a test statistic taking positive values

$$p = \Pr(T \geq T_{\text{obs}} \mid H_0)\,. \tag{7.10}$$

The two measures p and α were used in different ways by Fisher and Pearson. Fisher used the p-value as a measure of evidence in its own right and did not agree with the Neyman–Pearson school's use of error rates from hypothetical repeated samples (Fisher 1959). To be a valid error rate, the type I error rate α for a test should be determined pre-experimentally. This means that α does not depend on the data, and hence cannot be an efficient summary of the evidence. The p-value, on the other hand is efficient, if based on a sufficient statistic, but is not a valid error rate. Modern treatments generally integrate the p-value and type I error rates in an apparently seamless treatment, resulting in widespread confusion in the literature, a common misconception being that

> "The p-value is the probability of being wrong if the null hypothesis is true."

What is true is that if α is pre-experimentally determined and we reject H_0 whenever $p \leq \alpha$ then

> "α is the probability of being wrong if the null hypothesis is true."

This confusion results from the close association and similar sounding wording used around p and α. The common mistake and temptation is to assume α is the observed p-value. The observed p-value is the lowest, i.e. most optimistic, value to which we could set α and still have rejected H_0. But it is an error to set $\alpha = p$.

If the probability of being wrong when H_0 is true is not p when a p-value of p is observed, then what is it? In other words, given p, what is α? Sellke *et al.* (2001) give an approximate answer, which they call the *conditional* α, which we denote by α_c. A lower bound for the conditional α is given by

$$\alpha_{\mathrm{c}} \geq (1 + [-ep\log(p)]^{-1})^{-1}\,. \tag{7.11}$$

For example if $p = 0.05$, $\alpha_{\mathrm{c}} \geq 0.289$. Suddenly, $p = 0.05$ does not look like very good evidence.

The main problem with p-values is how to use and interpret the p-value, when is it good evidence and how should we make a decision? Problems with the interpretation of p-values have been pointed out by, e.g. Edwards *et al.* (1963), Berger and Sellke (1987), Berger and Berry (1988), and demonstrated in a genetics context by Ball (2001, 2005). In Chapter 8 we show that the strength of evidence implied by a given p-value depends on sample size (Chapter 8, Table 8.1). Interpretation of statistical evidence, including p-values and other measures, and how to make descisions is considered below.

Multiple comparisons. Many comparisons are made, e.g. for each marker in a genome scan, in a genomics experiment. Multiple comparisons procedures control the type I error rate for a set of tests. If n independent tests are made under the null hypothesis the probability that one or more type I errors are made is given by the Bonferroni correction:

$$\mathrm{FWER} = 1 - (1 - \alpha)^n\,. \tag{7.12}$$

In association mapping, the Bonferroni correction is overly conservative for two reasons (1) because tests are highly correlated between adjacent markers when markers are closely spaced and (2) because there is no reason that we need to make the overall probability of even a single type I error low when selecting putative loci from a whole genome scan. We can afford several errors provided most of the "detected" loci are real, i.e. the proportion of false discoveries is not too high.

7.4.2 The False Discovery Rate

The *false discovery rate* (FDR) has been proposed as an alternative to controlling type I errors (Benjamini and Hochberg 1995, 2000; Benjamini and Yekutieli 2001). The FDR is discussed here, because it has become a popular alternative to classical frequentist inference for microarray experiments, where there are many gene effects being tested in each experiment. The FDR is an improvement over type I error rates, but not a necessary concept in addition to Bayesian posterior probabilities – in fact the FDR is the average posterior probability of H_0 for a set of effects. We note below some potential difficulties with use of the FDR in association studies.

The motivation of Benjamini and Hochberg for introducing the FDR was that experience in genomics studies showed that comparison-wise thresholds (e.g. $\alpha = 0.05, 0.01$) gave too many false positives, whereas using genome-wise thresholds certainly cuts down

on the false positives, but it does so at the expense of also eliminating most of the true positives. Perhaps there is an optimum somewhere in between.

Classical frequentist inference controls "error rates" because of an inability to calculate a probability that H_1 is true. However, the FDR is a more useful quantity than the p-value or type I error rate. Given a set of putative gene effects, end users are interested only in the proportion of the given gene effects which are real, estimated by $1 - \text{FDR}$. End users are not interested in the number of other markers or genes that were or might have been tested and rejected (as given by the α threshold) to obtain the significant effects.

The false discovery rate is defined as the expected proportion of false discoveries among the rejected null hypotheses. We give a variant here, the positive false discovery rate (pFDR, Storey 2003), given by:

$$\text{pFDR} = E(V/R \mid R > 0)\,, \tag{7.13}$$

where V is the number of false positives when H_0 is rejected and R is the number of rejected null hypotheses. The positive false discovery rate avoids a technical problem with the denominator in Equation (7.13) if the number of rejected null hypotheses is zero. Storey (2003) shows that the positive false discovery rate has a Bayesian interpretation:

$$\text{pFDR} = \Pr(H_0 \mid p < \alpha)\,. \tag{7.14}$$

The FDR is the average probability that H_0 is true for the set of "detected" or "significant" effects from a testing procedure. The false discovery rate can be calculated in the frequentist paradigm when there are many *exchangeable* tests, meaning that the effects being tested are *a priori* indistinguishable. In effect, this is a Bayesian approach with prior probabilities per gene estimated from the data.

FDR computations. If a large number, m, of multiple exchangeable effects are being tested, the FDR is controlled at level α by the "step-up procedure" of Benjamini and Hochberg as follows:

- Sort the p values in increasing order and let $p_{(i)}$ denote the ith ordered p-value.
- Optionally, plot the ordered p-values versus i (Figure 7.3).
- Find the last p-value $p_{(k)}$ which lies below the line $p(i) = \frac{\alpha}{\pi_0 m} \times i$, i.e.

$$k = \max\left\{i : p_{(i)} \leq \frac{i\alpha}{m\pi_0}\right\} \tag{7.15}$$

- H_0 is rejected for $p_{(i)} \leq p_{(k)}$.

Conversely, for given k we can estimate α by applying the procedure for various α and interpolating in the sequence of k-values found, i.e. for given k we can estimate the FDR. In microarray or genomic studies the proportion of true effects is low, so $\pi_0 \approx 1$, and $\pi_0 = 1$ can be used in Equation (7.15). This is conservative, still providing FDR $\leq \alpha$, and is a good approximation when the proportion of true effects is low, as is often the case in gene mapping. Additionally, if tests are positively correlated, as when testing multiple linked markers, the FDR is conservative (Benjamini and Yekutieli 2001). However, in

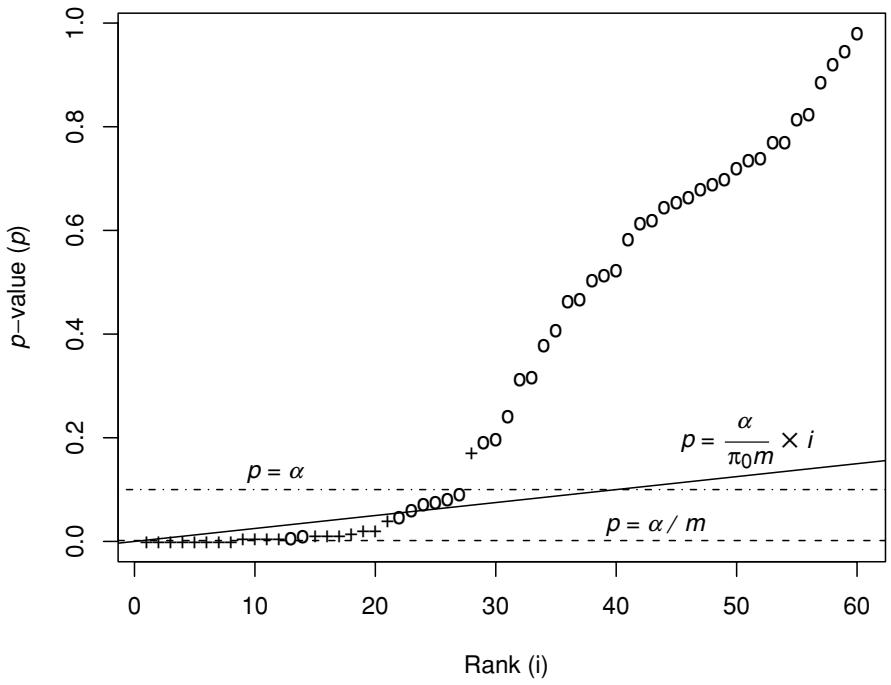


Figure 7.3. FDR computation. Data are ranked p-values from 60 simulations of which 40 were under the null hypothesis (plotted as "o") and 20 under the alternative (plotted as "+"). Parameters $\alpha = 0.1$ and $\pi_0 = 2/3$ are being used. Lines corresponding to comparison-wise ($p = \alpha$), FWER ($p = \alpha/m$) and FDR ($p = \alpha/(\pi_0 m) \times i$) thresholds are shown. The largest p-value below the line $p(i) = \alpha/(\pi_0 m) \times i$ is at rank 22. This and all smaller p-values are selected by the Benjamini and Hochberg (1995) "step-up procedure," controlling the FDR at level α.

association mapping, where there are many closely spaced markers along the genome it seems likely that the FDR, estimated this way, will be overly conservative. Hence the FDR is probably more suited to microarray data than multi-locus association studies.

For the data shown in Figure 7.3, $k = 22$ effects are selected. The false discovery rate was controlled at $\alpha = 0.1$, so the expected number of false discoveries was 2.2. The actual number of false discoveries is binomial with $n = 22$ and $p = 0.1$. The actual number of false discoveries was 3, which is slightly higher but equivalent to within sampling error. There was one false negative at rank 28.

Benjamini and Hochberg's estimation of the FDR above requires a large number of multiple exchangeable effects. The large number of effects makes it possible to calculate the FDR without explicitly using a prior. In the Bayesian paradigm described next, the FDR is calculated as the average posterior probability of H_0 for the "detected" effects. Bayesian calculation of posterior probabilities, and hence the FDR, do not require a large number of effects. Probabilities can be calculated even for a single effect.

7.4.3 Bayesian Inference

Bayesian statistical inference uses Bayes factors and posterior probabilities. Bayesian statistics calculates the posterior probability distribution for parameters $g(\theta \mid x)$, where x is the data. This is a probability distribution for the unknown parameters which combines information from prior knowledge and the data. Where there are two or more possible models posterior probabilities for models can also be calculated. The Bayes factor for comparing two models or hypotheses is the ratio of the posterior probability of the data under the two models. The Bayes factor is given by

$$B = \frac{\Pr(\text{data} \mid H_1)}{\Pr(\text{data} \mid H_0)}. \tag{7.16}$$

The Bayes factor measures the strength of evidence in the data.

Thus, a Bayes factor of 10 means the data are 10 times more likely to have occurred if H_1 is the true model for the process generating the data than if H_0 is the true model. Bayesian statistics requires *prior probability distributions*, which represent our knowledge before observing data. Priors are required in order to be able to compute posterior probabilities.

Probabilities can be exact or nearly so, as in probabilities for outcomes in roulette, throwing a dice, tossing a coin or of getting a certain hand in cards, where the process generating the data is carefully constructed or thought to be well understood. Frequentists consider only such probabilities. However, probabilistic models can be more flexible allowing for uncertainty in probabilities themselves. This is vital for real world applications, where the processes are not usually so cut and dried. Fisher thought that Bayesian methods apply only with exact prior probabilities, which he described as "probability of the mathematicians" but not when probabilities are not exactly known:

> "While as Bayes perceived, the concept of mathematical probability affords a means, in some cases, of expressing inferences from observational data, involving a degree of uncertainty, and of expressing them rigorously, in that the nature and degree of the uncertainty is specified with exactitude, yet it is by no means axiomatic that the appropriate inferences, though in all cases involving uncertainty, should always be rigorously expressible in terms of this same concept." (Fisher 1959, p. 37)

Exact probabilities apply, for example to roulette or coin tossing, but not, for example to betting on horses, where the process generating the data is imperfectly understood.

In genetics, allele frequencies provide examples where probabilities are not exactly known. Fisher (1959, pp. 18–20) considered the probability of observing heterozygous or homozygous mice in a cross. Suppose there are two alleles A, a, and suppose the AA genotypes have black coats while Aa genotypes have brown coats. The aa genotypes were not considered. Given the parental genotypes AA, Aa the probability of each offspring to be heterozygous is exactly known, i.e. 0.5. However, if the parental genotypes are unknown we have to use our prior knowledge of the population from which the parents are drawn. We may have prior knowledge from previous experiments. But what if there is no prior information? In practice this is unlikely, e.g. we would have observed a certain number of black and brown mice, but suppose for argument's sake there is no prior

information. Simply assuming equal probabilities of 0.5 for heterozygous and homozygous parents is not a good strategy, as Fisher rightly pointed out. This was perhaps not so rightly attributed as following from Laplace's suggestion of assigning equal probabilities to the various alternatives in the absence of prior information; the mistake being to effectively specify an exact value for the allele frequency p below, giving equal probabilities of heterozygous and homozygous parents, which he knew would probably be wrong, and which is wrong in principle, because it ignores the knowledge that the parental mice come from a population. Instead, Laplace's principle should be applied to values of p. The modern Bayesian approach is to say the population allele frequency for A is p, and let p have a Beta$(0.5, 0.5)$ prior distribution. This gives a *hierarchical model*

$$p \longrightarrow (g_{\mathrm{p}}, g_{\mathrm{m}}) \longrightarrow g_{x,i}, \tag{7.17}$$

where $g_{\mathrm{p}}, g_{\mathrm{m}}$ denote the paternal and maternal genotypes and $g_{x,i}$ denotes the ith offspring genotype. In the hierarchical model, probabilities for each parameter depend on the values of its ancestors. For example g_{p} depends on p. If Hardy–Weinberg equilibrium applies, $\Pr(g_{\mathrm{p}} = AA) = p^2$, $\Pr(g_{\mathrm{p}} = Aa) = 2p(1-p)$, $\Pr(g_{\mathrm{p}} = aa) = (1-p)^2$. The progeny genotype $g_{x,i}$ depends on g_{p} and on g_{m}. We have $\Pr(g_{x,i} = Aa \mid g_{\mathrm{p}} = Aa \text{ and } g_{\mathrm{m}} = aa) = 0.5$.

Note: A hierarchical model is similar to a family tree, where the probabilities for the genotypes of an individual depend on the genotypes of its ancestors. The Beta family of distributions Beta(a, b) gives a range of shapes, useful as prior distributions for proportions, with any given mean value $a/(a+b)$ and variance $ab/[(a+b)^2(a+b+1)]$, ranging from uninformative Beta(1/2,1/2) or Beta(1,1) to highly informative distributions when $a+b$ is large. Chapter 8, Example 8.3 gives a Bayesian analysis for a case–control test using Beta distributions.

Use and interpretation of statistical evidence. Using p-values, one strategy would be to select loci with $p \leq \alpha$, for some α. The "detected" loci would then be further tested, or regions around these loci genotyped, etc. However, the problem is: what is the best value of α to use? The strength of evidence implied by a given p-value increases with decreasing p-value, for a given experimental design and test setup. However, there is no interpretation of the p-value as evidence independent of sample size. This can be seen from Chapter 8, Table 8.1, where correspondences between p-values and Bayes factors for association tests are given. Fisher himself said that scientists should not make decisions (Fisher 1959, p. 101) based on p-values or error rates.

Use and interpretation of p-values is problematic, particularly in gene mapping, where effects may be small, sample sizes large and each effect has an *a priori* low chance of being real. Equation (7.11) gives an indication of a valid error rate and hence a better indication of the strength of evidence when a given p-value is obtained.

Interpretation of Bayesian posterior probabilities and/or Bayes factors is, in principle, straightforward. The reader can make the decision which maximises their expected utility. The expected utility is obtained by summing or integrating over the set of possible values for unknown parameters, and averaging over possible models if a unique true model is not known or unequivocally determined from the data.

The Bayes factor gives a direct measure of the strength of evidence favouring one hypothesis or model over another. Posterior probabilities for each hypothesis, assuming

one is true, can be obtained from the Bayes factor and the reader's own prior probabilities for each hypothesis to be true. This is useful in gene mapping where prior probabilities are low.

To find the optimal set of effects to choose for further investigation or use in applications, we simply choose all effects where the expected benefit outweighs the increased cost, i.e. the expected utility (or marginal profit) is positive. Evaluating the expected utility for the ith effect requires the posterior probability that the effect is real and the posterior distribution for the estimated effect. If the effect is β_i and the utility (benefit − cost) is $U(\beta_i)$, the posterior probability that the ith gene effect is real or not real is $\Pr(H_{1,i} \mid y), \Pr(H_{0,i} \mid y)$, respectively, where $H_{0,i}, H_{1,i}$ are, respectively, the null and alternative hypotheses for testing the ith effect, and the posterior distribution for β_i assuming $H_{1,i}$ is true is $g(\beta_i \mid H_{1,i}, y)$ then the expected utility from using gene i is

$$\int U(\beta_i) g(\beta_i \mid H_{1,i}, y) \mathrm{d}\beta_i \times \Pr(H_{1,i} \mid y) + C \times \Pr(H_{0,i}) \,, \tag{7.18}$$

where C is the utility if $H_{0,i}$ is true. If sample sizes are sufficiently large, the posterior distribution $g(\beta_i \mid H_{1,i}, y)$ in Equation (7.18) can be approximated by a normal distribution with mean $\hat{\beta}_i$ and standard deviation $\mathrm{se}(\hat{\beta}_i)$ obtained from maximum likelihood.

FDR and posterior probabilities. We have noted that the FDR is an average of posterior probabilities for a set of selected gene effects. Posterior probabilities can be recovered as successive differences from a sequence of FDRs as follows. First, sort the effects in order of increasing p-values. Then estimate the false discovery rate FDR_i, for effects $1, \ldots, i$, and similarly estimate FDR_{i+1}, when the effect $i+1$ is added. Since FDR_i is the average posterior probability of H_0 for the first i effects and FDR_{i+1} is the average for the first $i+1$ effects, these two rates are related by:

$$\mathrm{FDR}_{i+1} = \frac{i\mathrm{FDR}_i + \Pr(H_{0,i+1} \mid y)}{i+1} \,. \tag{7.19}$$

Then, solve for the (approximate) posterior probabilities $\Pr(H_{0,i+1} \mid y)$ in Equation (7.19) by equating the false discovery rates to their estimates.

7.4.4 Multi-Locus Methods

The single marker approach to QTL or association mapping is to apply a test successively to each marker or locus along the genome. Evidence is obtained for an association between each given marker and the trait. For example, interval mapping (Lander and Botstein 1989) does a likelihood ratio test for an association at each genomic locus (including between markers).

However, as noted previously, different marker genotypes are not independent, and nearby markers may be quite highly correlated. Thus, for example, a "significant" association between a marker and the trait may be a result of a causal association with a gene in the vicinity of a different marker. Or, there may be two or more linked QTL affecting the trait. Single marker methods, while giving some indication of QTL location, e.g. where marker–trait associations are most "significant", or where there is a peak in the interval mapping likelihood ratio, cannot effectively determine the genetic architecture, i.e. the

number and location of QTL. Additionally, the most "significant" markers tend to be those whose effects have been overestimated, a phenomenon known as *selection bias* (Miller 1990). To avoid selection bias, effects should be re-estimated in an independent population (Miller 1990) or a Bayesian model selection method can be used (e.g. Ball 2001).

Multi-locus methods give a more direct link between statistical inference and the genetic architecture. Stepwise regression simply chooses the model which best fits the data, possibly with some adjustment for the number of parameters. A "model" consists of a subset of selected markers, on which the trait is regressed. Stepwise regression is naïve in this context because the best model is generally not unequivocally identified; on the contrary there may be many models consistent with the data. Inferences or optimal decisions cannot be made simply by assuming the "best" model is the true model, particularly when the quantities of interest, e.g., the genetic architecture are strongly related to the model. The best approach is to use Bayesian model selection introduced by George and McCulloch (1993). The Bayesian model selection approach considers all possible models according to their probabilities. Estimates are averaged over models and inference is based on the total probability of models where a proposition is satisfied.

Since no model is selected, there is no selection bias. This approach is applied to QTL mapping in Ball (2001), with approximate posterior probabilities for models estimated using the BIC criterion.

By considering all models according to their posterior probabilities (cf. Raftery *et al.* 1997), it is possible to obtain unbiased estimates of effects, and to make inferences about the genetic architecture (Ball 2001, discussed in Sillanpää and Corander 2002; see also Yandell *et al.* 2002; Bogdan *et al.* 2004).

The Bayesian model selection analysis from Ball (2001) was applied to a linkage group with five markers in a QTL mapping family. Markers were in pseudo-backcross configuration, so only a single additive effect per marker was fitted. The prior probability per marker was 0.1, approximately equivalent to Poisson distribution with an average rate of 10 QTL over the whole genome. Statistics for the ten most probable models are shown in Table 7.4. Each row of Table 7.4 corresponds to a model, except for the final row which shows the total probabilities for markers. A "T" in the column for a marker indicates that marker is selected in the model, e.g. model 1 has marker M2 only selected. This model had an $R^2 = 18.6$ and a posterior probability of 50.5%. The marginal probability for M2 is the total probability for models with marker M2 selected, which is 68.5%, and for M3 the marginal probability is 39.5%. Other markers have probability less than 6%.

The null model, model 6, had posterior probability 1.1%. Thus, the posterior probability for one or more QTLs to be present is $100 - 1.1 = 98.9\%$. The probability for model size 1 is obtained by summing the probabilities for models with $k = 1$, i.e. $50.5 + 28.0 = 78.5\%$. The probability for model size 2 is obtained similarly as $9.0 + 4.9 + 2.7 + 1.0 + 0.7 + 0.5 + 0.4 = 19.3\%$. Thus, there is a 1.1%, 78.5%, 19.3% posterior probability that there is 0, 1 or 2 QTLs, respectively, present in the linkage group.

For case–control studies we have previously used an *indirect* method, where the number of cases and controls is fixed, and the marker allele frequencies become random. We call this an indirect method because the putative explanatory variables (here allele frequencies) appear in the model as responses, while the response (here disease status, case or control) appears as an explanatory variable (or treatment factor). This approach

Table 7.4. Top ten models for a linkage group with five markers

	Markers								
Model	M1	M2	M3	M4	M5	k	R^2	Prob	Cum.p
1	F	T	F	F	F	1	18.6	50.5	50.5
2	F	F	T	F	F	1	17.4	28.0	78.4
3	F	T	T	F	F	2	23.8	9.0	87.4
4	F	T	F	T	F	2	22.7	4.9	92.3
5	F	T	F	F	T	2	21.5	2.7	95.0
6	F	F	F	F	F	0	0.0	1.1	96.1
7	T	F	T	F	F	2	19.6	1.0	97.1
8	T	T	F	F	F	2	18.9	0.7	97.8
9	F	F	T	F	T	2	18.3	0.5	98.4
10	F	F	T	T	F	2	17.6	0.4	98.8
Total	2.1	68.5	39.5	5.9	3.7				100.0

is convenient when a single marker is being studied. For multi-locus methods, we need to revert to the *direct* method of analysis where the putative explanatory variables such as marker allele frequencies or genotypes appear as explanatory variables in the model and the response, disease status (case or control), appears as the response in our model. This is a binary response and is modelled as a generalised linear model (McCullagh and Nelder 1989). A generalised linear model has two parts. The first part is a linear model related to expectations of the observed data by a non-linear *link function* $g(\cdot)$:

$$g(E(y)) = X\beta, \tag{7.20}$$

X is a matrix of explanatory variables and β is a vector of regression coefficients. The second part is a possibly non-normal *error distribution* $f(y \mid \theta)$, where θ are model parameters. For binary data, $E(y) = p$, the error distribution is the Bernoulli distribution and the link function is generally taken to be the logit function $g(p) = \log(p/(1-p))$ for $0 \leq p \leq 1$. The model is

$$f(y \mid p) = p^y(1-p)^{(1-y)} \qquad \text{for } y = 0, 1, \tag{7.21}$$

where

$$\log(p/(1-p)) = X\beta \quad \text{or} \quad p = \exp(X\beta)/(1+\exp(X\beta))\,. \tag{7.22}$$

In principle, the Bayesian model selection approach described above can be applied where X is taken to be the model matrix based on a set of marker loci. However, caution is advised with binary generalised linear models because the usual model comparison statistic, the *deviance*, is approximately distributed as a χ^2 for binomial data, only when np and $n(1-p)$ are greater than 5. This is not the case for binary data, which is binomial with $n = 1$. For the same reason, the BIC criterion may not provide a good approximation to posterior probabilities for models, and *Markov chain Monte Carlo* (MCMC) sampling methods may be required. See Chapter 8, Example 8.7 for an application of MCMC.

7.4.5 Experimental Design and Power

The *power* of an experiment is defined as the probability of detecting an effect of given size, usually the smallest size we wish to detect.

The traditional (frequentist) approach is to design experiments with power to detect an effect of a certain specified size, where effects are deemed to be "detected" if $p < \alpha$, for given pre-specified type I error rate α. The power of an experiment can be improved by getting better data, e.g. more accurate measurements, or by getting more data of the same quality, i.e. increasing the sample size.

It is important to realise that the power refers to the probability of detecting the true size of an effect, not the estimated size of effect. If the power to detect the true size of effect is good there is usually no problem with estimates, whether Bayesian or frequentist. The "significant" effects, will be affected by selection bias if the power to detect the true effect is not high. For effects near the borderline of significance, i.e. $p \approx \alpha$, the power will be low. Post-experimentally, we may estimate an effect but this effect may well be overestimated. For example suppose $\alpha = 0.05$, and we observe a borderline p-value of 0.05, corresponding to an estimated effect ($\hat{d}$) of approximately twice its standard error ($\sigma(\hat{d})$) in absolute value. The true effect could quite probably be one standard error less, or quite possibly even two standard errors less than the estimate. If the estimated effect is only twice its standard error, then the true effect could be very small or even zero. So, the power could be very low and selection bias arbitrarily high in percentage terms. If $\alpha = 0.05$, the power to detect an effect of two standard errors is 0.5, since the estimated effect is equally likely to be greater or less than the true effect. If, however, we observe $p = 6.3 \times 10^{-5}$, then the estimated effect size is approximately 4 times its standard error, and the true effect is quite likely (approximately 97.5% probability) to be at least two standard errors. We can fairly safely say the power to detect the effect is at least 0.5. An approximate calculation (Ball 2005, Appendix B) shows that if $d \geq 4\sigma(\hat{d})$, and H_0, H_1 are *a priori* equally likely, then the posterior probability of H_0 for a detected effect is less than 0.1.

An alternative approach, which avoids the problem of how to choose α, is discussed in Chapter 8. The approach, from Ball (2005) is to design experiments with given power to detect effects of interest with given Bayes factor.

7.4.6 Summary

- Single marker hypothesis tests have been used in genome scans because they were a known and easily computed method.
- There are problems with the interpretation of p-values.
  - Commonly used $p = 0.05$ can be very weak evidence.
  - A p-value is not a valid error rate. The observed p-value is *not* the probability of being wrong if H_0 is true.
- We have no way to determine the optimum value of α to use.
  - Commonly used $\alpha = 0.05, 0.01$ gives too many false positives.

  - Multiple positive tests are likely for linked markers in the neighbourhood, of a causal locus, complicating the interpretation of error rates.
  - Multiple comparisons do not help, we still have to determine α.
  - Tests on multiple linked markers in association studies are not independent, complicating the interpretation of multiple tests.

- FDR is better, in principle than p-values.
- The FDR is approximately equivalent to average posterior probabilities of H_0 for a set of selected effects, but:
  - The FDR requires data from many exchangeable tests.
  - The FDR may be overly conservative when applied to many closely linked markers in association studies.
- Using the Bayesian approach avoids these problems, because we can calculate Bayesian posterior probabilities for propositions of interest, e.g. the probability that an effect is real, or that a QTL is in a region.
- The posterior probability for H_1 gives us a probability that the effect being tested is real.
- The statistical problem in association mapping is to select loci associated with variation in a trait. This is a model selection problem, not a hypothesis testing problem.
- Bayesian model selection considers multiple models according to their probabilities, and can be used to infer the genetic architecture of a trait.
- The Bayes factor gives the strength of evidence in the data.
- In general, experiments should be designed with good power to detect effects with a reasonable Bayes factor, e.g. at least $B = 20$.
- In genomics, we are trying to select a small number of effects from the whole genome. The prior probabilities will be low. Bayes factors as high as $1,000,000$ may be required.
- Required sample sizes will be substantially larger, than for experiments designed to detect effects with $\alpha = 0.05$.

7.5 EXPERIMENTAL DESIGNS FOR ASSOCIATION STUDIES

Common experimental designs for association studies include:

- A random population sample of unrelated individuals.
- A case–control test.

- A transmission disequilibrium test (TDT), based on transmission of alleles in many small families.

- A study of related individuals in a pedigree (e.g. in a breeding population).

A random population sample assumes there is no population structure, otherwise population structure must be allowed for in the analysis.

If there is population structure, such as might be generated by recent admixture of populations, then spurious associations can be generated. For example, after a population admixture long-range linkage disequilibrium can be formed as a result of allele frequencies differences between the source populations (Chapter 8, Example 8.8). These associations represent genuine linkage disequilibrium, but are spurious from the point of view of locating genes, because they do not imply close proximity of marker and trait loci.

Pritchard *et al.* 2000a, 2000b give a Bayesian method (STRAT) for testing for population structure (Chapter 8, Section 8.3.5). These methods require a set of unlinked markers, usually one per chromosome that can differentiate between the subpopulations. The true population structure and subpopulation membership cannot be determined exactly, but this uncertainty can be taken into account.

Case–control and the discrete TDT designs are primarily used for diseases. The case–control study design selects approximately equal-sized samples of cases and controls, but otherwise samples randomly from the population. The TDT test is based on transmission of marker alleles from parents to offspring in many small families.

Individuals of known relatedness can be sampled from existing pedigrees, and analysed using a mixed model (Chapter 8, Section 8.3.6). This approach also combines linkage and linkage disequilibrium analysis, and also controls for population structure, similar to the TDT. The effective sample size of the LD part of a pedigree analysis is reduced to number of founders of the pedigree. Thus, a pedigree where all individuals were descended from four grandparents would only have an effective sample size of 4 for the LD part, in which case the study would be essentially a QTL study.

An association study may be combined with a QTL mapping study (Chapter 8, Section 8.3.7), using the association mapping study to refine location information from the QTL mapping study. Marker genotyping can be reduced by genotyping markers only within QTL regions. This may be more efficient than an association study alone, particularly if a substantial QTL mapping trial already exists. The QTL mapping trial when combined with a random population study or a pedigree can also perform a role similar to the TDT of reducing spurious associations.

7.5.1 Case–control studies

For many diseases, the prevalence of cases in the population is much lower than the prevalence of controls. A random sample from the population would typically have few cases, and therefore low power to detect effects on disease incidence. The case–control study design remedies this by selecting separate samples of affected individuals (cases) and unaffected individuals (controls). The sample sizes for cases and controls are generally approximately equal.

Table 7.5. Contingency table for a genotypic case–control test

Disease status	Marker genotype M_1M_1	M_1M_2	M_2M_2	Total
Case	u	v	w	$n_D = u + v + w$
Control	x	y	z	$n_H = x + y + z$
Total	$u + x$	$v + y$	$w + z$	$n = n_D + n_H$

Table 7.6. Contingency table for an allelic case–control test

Disease status	Marker allele M_1	M_2	Total
Case	a	b	$2n_D = a + b$
Control	c	d	$2n_H = c + d$
Total	$a + c$	$b + d$	$2n = a + b + c + d$

Consider a case–control study with n_D cases and n_H healthy controls. The data for genotypic and allelic case–control tests is summarised in contingency tables (Tables 7.5 and 7.6).

The Pearson χ^2-test or Fisher's exact test can be used to test for an association between disease status and marker genotype or allele frequencies. The χ^2 approximation to the test statistic is approximately valid only if the expected cell counts are five or more or at least 80% of cells in the table. Also, the Yates continuity correction is recommended when sample sizes are small (Sokal and Rohlf 1969). Fisher's exact test does not depend on any asymptotic approximation, so should be used if cell counts are low. Fisher's exact test conditions on the marginal totals, i.e. considers the distribution of the test statistics only for tables with the given marginal totals, while the χ^2-test considers all possible tables with the given total under the null hypothesis, H_0, of no association. According to Fisher (1959), the Pearson χ^2 is not appropriate because the marginal totals are known, and therefore the appropriate reference set is only the set of tables with the given marginal totals. Other tables are irrelevant. Both approaches have the drawback that they produce a p-value whose distribution is uniform under H_0 but the distribution under H_1 is not considered. Low p-values may be unlikely under H_0 but may be equally unlikely under H_1, and hence are not necessarily good evidence for H_1. This problem is addressed by using Bayes factors in Chapter 8.

Recall that the χ^2 test statistic for a contingency table is

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}\,, \tag{7.23}$$

where O_i, E_i are the observed and expected cell counts for the ith cell, respectively.

For 2×2 tables such as Table 7.6, the χ^2 test statistic is conveniently calculated as:

$$\chi^2 = \frac{(a+b+c+d)(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)} \sim \chi^2_1 \qquad \text{under } H_0\,. \tag{7.24}$$

Table 7.7. Contingency table for genotypic case–control test for association between the sickle cell locus and malaria (Example 7.2)

Disease	Marker genotype			
status	AA	AS	SS	Total
Case	309	5	1	315
Control	485	95	3	583
Total	794	100	4	898

Table 7.8. Contingency table for allelic case–control test for association between the sickle cell locus and malaria (Example 7.2)

	Marker allele	
Status	A	S
Case	623	7
Control	1065	101

Example 7.2. A case–control study for malaria.

Genotypic and allelic contingency tables for an association with malaria are shown in Tables 7.7 and 7.8. Data are from Ackerman *et al.* (2005). Marker alleles are A (normal), and S (HbS sickle cell mutation).

For the allele-based test we calculate

$$\chi^2 = \frac{(a+b+c+d)(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)} = 41.3\,. \tag{7.25}$$

The p-value is $\Pr(\chi^2_1 \geq 41.3) = 1.3 \times 10^{-10}$, which is very low, suggesting strong evidence for an effect. This is confirmed by the Bayes factor of 1.0×10^{10} calculated in Chapter 8.

This and other examples discussed (e.g. the Alzheimer's – APOE association in Chapter 8) are examples where the association was already known. We should bear in mind that it is more difficult to find associations from a whole genome scan, than testing a single locus where the marker locus associated with the trait is already known. The statistical evidence has to overcome the low prior probability for any given marker locus to be within a small genomic interval of the given trait locus. Additionally, the statistical evidence must discriminate between the causal locus and nearby loci which may also be in linkage disequilibrium with the trait.

7.5.2 Transmission Disequilibrium Tests

The transmission disequilibrium test (TDT) (Spielman *et al.* 1993) is based on transmission of marker alleles from parents to offspring in many small families. We discuss two forms of TDT test here: the discrete TDT test based on many small families (trios) with a single affected offspring; and the S-TDT test based on *discordant sib pairs*, where one affected and one unaffected offspring are sampled from each family. The TDT test

Table 7.9. Contingency table for transmission of alleles in a TDT test based on parent–offspring trios

Transmitted alleles	Non-transmitted alleles	
	1	2
1	n_{11}	n_{12}
2	n_{21}	n_{22}

requires both parental and offspring genotypes. The S-TDT (Spielman and Ewens 1998) requires only genotypes from each discordant sib pair, so can be used where parental DNA is not available. The SDT test (Horvath and Laird 1998) can be used when data from more than one affected and unaffected sib are available.

The TDT test is based on the fact that each parental allele is transmitted randomly, with 50% probability, to each offspring. If an allele is associated with a disease phenotype, then there will be a higher or lower proportion of the marker amongst the cases.

The TDT test statistic for a bi-allelic marker is given by

$$T = \frac{(n_{12} - n_{21})^2}{(n_{12} + n_{21})} \sim \chi^2_1 \qquad \text{under } H_0\,, \tag{7.26}$$

where n_{ij} are as in Table 7.9. For a multi-allelic marker with m alleles, the TDT test statistic is given by

$$T = \frac{m-1}{m} \sum_{i=1}^{m} \frac{(n_{i\cdot} - n_{\cdot i})^2}{(n_{i\cdot} + n_{\cdot i} - 2n_{ii})} \sim \chi^2_{(m-1)}\,, \tag{7.27}$$

where n_{ij} is the number of times allele i is transmitted and allele j is not transmitted, $n_{i\cdot} = \sum_j n_{ij}$ and $n_{\cdot i} = \sum_j n_{ji}$ (Weir 2001).

Note: The Spielman *et al.* TDT test was based on an earlier idea (Terwilliger and Ott 1992) of presenting data related to transmission and non-transmission of alleles, as in Table 7.9.

Allele transmission status is known if one parent is heterozygous and the other homozygous (e.g. $Aa \times aa$) or both parents are heterozygous with different alleles.

The TDT test simultaneously tests for linkage and linkage disequilibrium:

- If there is no linkage disequilibrium, that means that the marker and trait loci are independent in the population. Hence any parental trait QTL allele will be independent of the parental marker alleles. Hence there will be no association between QTL allele and the transmitted marker allele.

- If there is linkage disequilibrium between marker and QTL alleles, but no linked QTL this means that there is an association between parental marker and QTL alleles, but because there is no linkage ($r = 0.5$), the QTL allele transmitted will be independent of the marker allele transmitted. Again, there will be no association between the transmitted QTL allele and the transmitted marker allele.

- If there is linkage disequilibrium and linkage, linked QTL effects with $r < 0.5$ will be reduced by a factor $(1 - 2r)$.

This reduces spurious associations from population structure. The disadvantage of the TDT design is the need to genotype trios – each transmission requires genotyping of markers from three individuals – two parents and one offspring. One parent should be heterozygous and the other homozygous for the marker. Power may be lower than for a random population sample with the equivalent amount of genotyping.

Most spurious associations between markers unlinked to the trait locus will be eliminated by the TDT test. However, some spurious associations between markers linked to the trait locus may still occur. These associations are "spurious" in the sense that although the marker is linked to the trait locus it may still be very far from the trait locus compared to the potential resolution of the association study. This should be considered in the analysis and experimental design.

Example 7.3. A TDT test.

Frequencies of parent–offspring genotype combinations for 120 nuclear families, each with one affected sibling, are shown in Table 7.10. Frequencies of transmitted and non-transmitted alleles are shown in Table 7.11.

Note: When both parents are homozygous (parental genotypes 1,6) the transmitted and non-transmitted alleles are the same, affecting n_{11}, n_{22} only, hence these crosses do not contribute to the test statistic. When one parent is homozygous (parental genotypes 2,5) and the other parent heterozygous only the allele transmitted from the heterozygous parent contributes to the test statistic.

From Table 7.11 we have $n_{12} = 39$ and $n_{21} = 86$. The TDT test statistic is

$$T = \frac{(39 - 86)^2}{(86 + 39)} = 17.7\,. \tag{7.28}$$

The p-value is $\Pr(\chi^2_1 \geq 17.7) = 2.6 \times 10^{-5}$. The Bayes factor, calculated in Chapter 8 is 610, representing strong evidence, but not strong enough to overcome low prior odds in genomic studies.

Sib-based TDT tests. Sib-based TDT tests (Curtis 1997; Spielman and Ewens 1998; Boehnke and Langefeld 1998; Horvath and Laird 1998; see Monks *et al.* 1998 for a comparative review) compare marker allele frequencies or summary statistics between affected and unaffected sibs. These tests are useful when parental genotypes are unavailable, e.g. for late onset diseases.

Table 7.10. Parental and offspring genotypes for the TDT test (Example 7.3)

Parental genotype	Offspring (case) genotype			Total
	M_1M_1	M_1M_2	M_2M_2	
1. $M_1M_1 \times M_1M_1$	18	0	0	18
2. $M_1M_1 \times M_1M_2$	21	30	0	51
3. $M_1M_1 \times M_2M_2$	0	8	0	8
4. $M_1M_2 \times M_1M_2$	2	15	19	36
5. $M_1M_2 \times M_2M_2$	0	3	3	6
6. $M_2M_2 \times M_2M_2$	0	0	1	1

Table 7.11. Transmitted and non-transmitted alleles for TDT test (Example 7.3)

Transmitted allele	Non-transmitted allele		Total
	M_1	M_2	
M_1	99	39	138
M_2	86	16	102
Total	185	55	240

We consider two forms of sib-based TDT tests, the S-TDT test (Spielman and Ewens 1998) and the SDT test (Horvath and Laird 1998). The S-TDT test compares marker allele frequencies between affected and unaffected sibs from a large number of nuclear families. The S-TDT gives a test for both linkage and linkage disequilibrium if only one affected and one unaffected sib per family are used. If data from more than one affected and/or unaffected sib is used the S-TDT gives a test for linkage (Monks *et al.* 1998; Weir 2001). The SDT test bases inference on a summary statistic calculated for each family, based on whether or not the average allele frequency for affected sibs differs from the family mean. The SDT tests for both linkage and linkage disequilibrium provided there is at least one affected and unaffected sib per family.

The S-TDT test. Let y_i denote the number of M_1 alleles in affected sibs, a_i, u_i the number of affected and unaffected sibs, and r_i, s_i the number of M_1M_1 and M_1M_2 genotypes in the ith nuclear family, respectively. The test statistic is

$$T = \frac{Y - A}{\sqrt{V}} \sim N(0,1) \qquad \text{asymptotically, under } H_0, \tag{7.29}$$

where

$$Y = \sum_{i=1}^{n} y_i, \tag{7.30}$$

$$A = \sum_{i=1}^{n} (2r_i + s_i)a_i/t_i, \tag{7.31}$$

$$V = \sum_{i=1}^{n} \frac{a_i u_i [4r_i(t_i - r_i - s_i) + s_i(t_i - s_i)]}{t_i^2(t_i - 1)}\,. \tag{7.32}$$

Y is the total number of M_1 alleles in affected sibs, A is an estimate of the total expected value of Y under H_0 (assuming the allele frequency for affected sibs is the same as for all sibs) and V is an estimate of the variance of $Y - A$.

The TDT and S-TDT tests can be combined, with the test statistic, Z, given by:

$$Z = \frac{(Y + n_{12}) - (A + (n_{12} + n_{21})/2)}{\sqrt{V + (n_{12} + n_{21})/4}} \sim N(0,1)\,, \tag{7.33}$$

where n_{12}, n_{21} are as in Table 7.9, for the parent–offspring trios, and Y, A, V are calculated as above for the discordant sib-pair data.

Figure 7.4. Families for the S-TDT test (Example 7.4). Affected offspring are shown in black.

Table 7.12. Values of t, a, u, r, s and y for families in Example 7.4

Family	t	a	u	r	s	y
1	6	2	4	2	1	4
2	3	2	1	1	1	3
3	4	2	2	0	3	1

Example 7.4. An S-TDT test.

Three families are shown in Figure 7.4. Affected sibs are shown as solid black boxes or circles; unaffected sibs are shown with blank boxes or circles. Marker alleles for each individual are indicated as, e.g. $(1, 2)$. The summary statistics t, a, u, r, s and y for marker allele 2 are summarised for each family in Table 7.12.

Applying Equations (7.30–7.32) we obtain $Y = 8, A = 5.17, V = 2.21$, and hence

$$T = \frac{8 - 5.17}{\sqrt{2.21}} = 1.90\,. \tag{7.34}$$

The p-value is $\Pr(|Z| \geq 1.9) = 2(1 - \Phi(1.9)) = 0.057$, where $\Phi(\cdot)$ is the standard normal c.d.f., which is not significant at the 5% level. The Bayes factor, calculated in Chapter 8, is 1.94, which is less than one, indicating weak evidence for an association. This is not surprising, since the example was chosen to be small to illustrate the calculations. Much larger sample sizes will be needed in genomic association studies.

The SDT test. As noted above, the SDT test (Horvath and Laird 1998) tests for both linkage and linkage disequilibrium if multiple affected and unaffected offspring per family are used. The SDT test compares the average rate of occurrences of a given allele between affected and unaffected sibs scoring $1, 0, -1$ for a family if rate of occurrences for affected sibs is greater, equal or less (respectively) than the rate for unaffected sibs.

The total score, S, is

$$S = \sum_{i=1}^{n} \text{sign}(d_i)\,, \tag{7.35}$$

where $d_i = y_i/a_i - [(2r_i + s_i) - y_i]/u_i$ is the difference in rates of occurrence of the allele, and $\text{sign}(d_i)$ is $1, 0, -1$ if d_i is positive, zero or negative, respectively.

The number of non-zero differences is

$$W = \sum_{i=1}^{n} \text{sign}(d_i)^2\,, \tag{7.36}$$

and the test statistic is given by

$$T = \frac{S^2}{W} \sim \chi_1^2 \,. \tag{7.37}$$

Exercise. Calculate the SDT statistic and p-value for the three families of Example 7.4.

7.5.3 Choice of Experimental Designs for Association Studies in Plants

Any of the above designs can be used in plants. However, there are few published examples of association studies in plants. Hence we are guided by the human genetics literature. Statistical considerations and the need for evidence are similar.

Considerations of which design to use will vary among plant species, depending considerations such as the relative cost and time required for establishing and growing experimental crosses, and genotyping.

Long-lived perennials such as forest trees will have many similar considerations to human trials. It may take a number of years from establishment of a trial before some traits can be evaluated. Hence, current studies are often constrained by the availability of existing trials. Phenotypes or genotypes of parent trees may be missing.

If new trials are being established for association mapping in plants it is possible to include multiple replicates of each genotype in a field trial. This may be more efficient, as the field trial can use blocking to control within-site variability, and additionally the use of replicated genotypes increases the effective heritability of the trait, which can substantially reduce error for low heritability traits. Only one ramet of each clone would need to be fully genotyped, although some markers from each ramet will probably need to be genotyped to verify identity. Such alternatives would need to be evaluated for each species.

The ability to clone plant genotypes also makes the TDT design easier to implement and more efficient, since parents and offspring trios can be simultaneously cloned and grown in the same environment.

7.5.4 Summary

- A range of experimental design types is available, including a random population sample, a case–control study, a TDT test and a pedigree with mixed model analysis.
- Any of these design types may be combined with a QTL mapping study, giving reduced genotyping.
- There is scope to develop new experimental designs for association mapping in plants, using clonal replication of genotypes in field trials, with potential to simultaneously increase the effective heritability, control environmental variation and reduce residual errors.

7.6 SUMMARY

- Single marker hypothesis tests have been used in genome scans because they were a known and easily computed method. But results of single marker tests are not directly related to the genetic architecture.

- There are problems with the interpretation of p-values.
  - A p-value of 0.05 can be a very weak evidence.
  - A p-value is not a valid error rate. The observed p-value is *not* the probability of being wrong if H_0 is true.
- We have no way to determine the optimum value of α to use.
  - Commonly used $\alpha = 0.05, 0.01$ gives too many false positives.
  - Multiple positive tests are likely for linked markers in the neighbourhood, of a causal locus, complicating the interpretation of error rates.
  - Multiple comparisons do not help, we still have to determine α.
  - Tests on multiple linked markers in association studies are not independent, complicating the interpretation of multiple tests.
- FDR is better, in principle, than p-values but requires data from many exchangeable tests, and may be overly conservative when applied to many closely linked markers in association studies.
- Bayesian measures of evidence are readily interpretable, independent of the data or experimental designs used.
- Using Bayesian methods it is possible to calculate posterior probabilities for a marker trait association to be real, or for a causal effect to lie within a given region.
- The statistical problem in association mapping is to select loci associated with variation in a trait. This is a model selection problem, not a hypothesis testing problem.
- Multi-locus methods can be used to infer the genetic architecture, by giving probability distributions for number and locations of QTL.
- The Bayes factor gives the strength of evidence in the data.
- Experiments should be designed with good power to detect effects with a reasonable Bayes factor, e.g. at least $B = 20$. In genomics we are trying to select a small number of effects from the whole genome. Prior probabilities will be low. Higher Bayes factors will be required.
- A range of experimental design types is available, including a random population sample, a case–control study, a TDT test and a pedigree with mixed model analysis.
- Any of these design types may be combined with a QTL mapping study, giving reduced genotyping.
- There is scope to develop new experimental designs for association mapping in plants, using clonal replication of genotypes in field trials, with potential to simultaneously increase the effective heritability, control environmental variation and reduce residual errors.

7.7 REFERENCES

Ackerman, H., Usen, S., Jallow, M., Sisay-Joof, F., Pinder, M., Kwiatkowski, D.P. 2005, A comparison of case–control and family-based association methods: the example of sickle cell and malaria. Ann. Human Genet. 69:559–565.

Ball, R.D. 2001, Bayesian methods for quantitative trait loci mapping based on model selection: approximate analysis using the Bayesian Information Criterion. Genetics 159:1351–1364.

Ball, R.D. 2005, Experimental designs for reliable detection of linkage disequilibrium in unstructured random population association studies. Genetics 170:859–873. `http://www.genetics.org/cgi/content/abstract/170/2/859`

Benjamini, Y., Hochberg, Y. 1995, Controlling the false discovery rate a practical and powerful approach to multiple testing. J. R. Stat. Soc. B 57:289–300.

Benjamini, Y., Hochberg, Y. 2000, On the adaptive control of the false discovery rate in multiple testing with independent statistics. J. Educ. Behav. Stat. 25:60–83.

Benjamini, Y., Yekutieli, D. 2001, The control of the false discovery rate in multiple testing under dependency. Ann. Stat. 29:1165–1187.

Berger, J., Berry, D. 1988, Statistical analysis and the illusion of objectivity. Am. Scientist 76:159–165.

Berger, J.O., Sellke, T. 1987, Testing a point null hypothesis: The irreconcilability of P values and evidence (with discussion). J. Am. Stat. Assoc. 82:112–139.

Boehnke, N., Langefeld, C.D. 1998, Genetic association mapping based on discordant sib pairs: the discordant-alleles test. Am. J. Hum. Genet. 62:950–961.

Bogdan, M., Ghosh, J.K., Doerge, R.W. 2004, Modifying the Schwarz Bayesian information criterion to locate multiple interacting quantitative trait loci. Genetics 167:989–999.

Curtis, D. 1997, Use of siblings as controls in case–control association studies. Ann. Hum. Genet. 61:319–333.

Edwards, W., Lindman, H., Savage, L.J. 1963, Bayesian statistical inference for psychological research. Psychol. Rev. 70:193–242.

Fisher, R.A. 1959, *Statistical methods and scientific inference*, 2nd ed., T. and A. Constable, Edinburgh.

George, E.I., McCulloch, R.E. 1993, Variable selection via Gibbs sampling. J. Am. Stat. Assoc. 88:881–889.

Horvath, S., Laird N.M. 1998, A discordant-sibship test for disequilibrium and linkage: no need for parental data. Am. J. Hum. Genet. 63:1886–1897.

Lander, E.S., Botstein, D. 1989, Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185–199.

McCullagh, P., Nelder, J.A. 1989. *Generalized linear models*, 2nd ed., Chapman & Hall/CRC, London.

Miller, A.J. 1990: *Subset selection in regression*, Monographs on Statistics and Applied Probability 40, Chapman & Hall, London.

Monks, A.A., Kaplan, N.L., Weir, B.S. 1998, A comparative study of sibship tests of linkage and/or association. Am. J. Hum. Genet. 63:1507–1516.

Pritchard, J.K., Stephens, M., Donnelly, P. 2000a, Inference of population structure using multilocus genotype data. Genetics 155:945–959.

Pritchard, J.K., Stephens, M., Rosenberg, N.A., Donnelly, P. 2000b, Association mapping in structured populations. Am. J. Hum. Genet. 67:170–181.

Raftery A.E., Madigan D., Hoeting J.A. 1997, Bayesian model averaging for linear regression models. J. Am. Stat. Assoc. 92:179–191.

Sellke, T., Bayarri, M.J., Berger, J.O. 2001, Calibration of p values for testing precise null hypotheses. Am. Statistician 55:62–71.

Sillanpää, M.J., Corander, J. 2002, Model choice in gene mapping: what and why. Trends Genet. 18:301–307.

Sokal, R.R., Rohlf, F.J. 1969, *Biometry: The principles and practice of statistics in biological research.* W. H. Freeman, San Francisco, p. 776

Spielman, R.S. Ewens, W.J. 1998, A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. Am. J. Hum. Genet. 62:450–458.

Spielman, R.S., McGinnis, R.E., Ewens, W.J. 1993, Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am. J. Hum. Genet. 52:506–516.

Storey, J.D. 2003, The positive false discovery rate: A Bayesian interpretation and the q-value. Ann. Statist. 31:2013–2035.

Stram, D.O., Lee, J.W. 1994, Variance components testing in the longitudinal mixed effects model. Biometrics 50:1171–1177.

Terwilliger, J.D., Ott, J. 1992, A haplotype based haplotype relative risk approach to detecting allelic associations. Hum. Hered. 42:337–346.

Yandell B.S., Jin C., Satagopan, J.M., Gaffney, P.J. 2002, In: Discussion of: Model selection approach for the identification of quantitative trait loci in experimental crosses, by Broman and Speed. J. Roy. Stat. Soc. B. 64:731–775.

Weir, B.S. 2001, *Population Genetic Data Analysis 2001*, Southern Summer Institute of Statistical Genetics, North Carolina State University.

Chapter 8
STATISTICAL ANALYSIS AND EXPERIMENTAL DESIGN

Roderick D. Ball[1]

8.1 INTRODUCTION

The goal of association mapping is to locate genes and/or predict genetic effects, to allow selection of favourable genotypes.

Association mapping, also known as 'linkage disequilibrium mapping' or 'LD mapping' aims to detect and locate genes relative to a map of existing genetic markers. Location information is obtained because the distance between the gene and a marker on a chromosome is one factor influencing the strength of association between the gene and marker. In a population, recombinations affecting the association between a gene and marker may occur over many generations. This potentially gives a much finer resolution for mapping QTL than pedigrees used for linkage analysis.

Early attempts to find associations for complex diseases or quantitative traits have led to many published associations which are likely to be spurious (Terwilliger and Weiss 1998; Altshuler *et al.* 2000; Emahazion *et al.* 2001; Neale and Savolainen 2004).

Altshuler *et al.* (2000) (discussed in Gura 2000) retested 13 published associations of SNPs with type II diabetes in an independent population. Only one was significant. These results are summed up by Altshuler (Hampton 2000):

> The lack of replication of the others points to the need for larger samples, controls for population differences, *and stronger statistical evidence prior to claiming an association*. (emphasis added)

Terwilliger and Weiss (1998, Figure 4) show the distribution of around 260 reported p-values from association studies in two journals, and note that there is no evidence of departure from the uniform distribution (i.e. no evidence of any real effect):

> ... investigators are too frequently gambling on and publishing results in situations where the evidence is not at all compelling.

Neale and Savolainen (2004) note that candidate gene associations have been criticised as being unreliable with insufficient sample size cited as a contributing factor.

[1]Ensis (New Zealand Forest Research Institute Limited), 49 Sala Street, P. B. 3020, Rotorua, New Zealand

Emahazion *et al.* (2001) retested published associations for a number of markers with 13 genes putatively associated with Alzheimer's disease, found from case–control studies, noting

> ...limited ability of typical association studies based on candidate genes to discern the true medium sized signals from false positives...

Except for the APOE ϵ_4 allele (with $p \approx 0.003\%$), which was used as a 'positive control' in their study, only 2.8% were 'verified' with $p < 0.05$ and only one had $p < 0.01$, which rose to $p = 0.33$ after allowing for 60 comparisons.

False positives, publication bias, population structure, heterogeneity (i.e. variability resulting from other genetic and environmental risk factors) and conservative multiple corrections procedures were cited as causes of problems, with the first three factors contributing to spurious associations and the latter two factors contributing to failure to detect true effects. These comments point to problems with the use of statistics, particularly the interpretation of p-values.

The rest of this chapter consists of two main sections: the statistical analysis section (Section 8.2) and the experimental design section (Section 8.3).

The statistical analysis section (Section 8.2) covers general approaches for testing scientific hypotheses, including comparison of frequentist and Bayesian approaches, and comparison of model-based and empirical approaches for single marker or multiple marker (haplotype) analyses. To understand and rectify the problems with spurious associations, we revisit the fundamentals of statistical inference with respect to the problem of testing scientific hypotheses, comparing frequentist and Bayesian methods in Section 8.2.1. Problems are noted with the use of frequentist p-values as commonly used, and Bayesian alternatives given.

The ability to detect LD is determined by factors including the extent of LD, size of QTL effects, i.e. trait genetic architecture. These factors are important considerations for experimental design. The experimental design section (Section 8.3) covers how experiments can be designed with power to detect effects with a given strength of evidence, and considers the main types of experimental design in separate subsections with examples and statistical analyses appropriate to each.

8.2 STATISTICAL ANALYSIS

Testing for an association between a marker and a trait is an example of testing a scientific hypothesis. We first revisit the fundamentals of statistical inference with respect to testing scientific hypotheses, including the commonly used frequentist hypothesis testing, with p-values as a measure of evidence, and the Bayesian approach with Bayes factors and posterior probabilities as evidence. It is shown that the two approaches are substantially different, for testing scientific hypotheses, and very small p-values are needed to obtain even modest evidence according to the Bayesian framework, when sample sizes are sufficiently large to detect the many small effects underlying quantitative traits and complex diseases.

We then consider various statistical approaches ranging from simple single marker analysis of variance (ANOVA) or t-tests to more complex statistical models such as coalescent-based approaches to analysis of haplotype data in a gene region.

Statistical methods specific to each experimental design are discussed in the experimental design section (Section 8.3).

8.2.1 Testing scientific hypotheses

Testing of scientific hypotheses, such as Einsteins' theory of relativity versus Newton's theory or that a new diet or drug is effective, or that a certain marker is associated with variation of a trait has three important characteristics:

1. There is a signficant cost to being wrong. We would not want to use Einstein's much more difficult theory if Newton's was the true model. Patients might die if we give them a drug which is not effective. Similarly if a marker–trait association was spurious we would not want to select that marker for further testing or applications.

2. Testing scientific hypotheses often involve comparing a simpler model to a more general one. The null hypothesis of no effect is a subset of the hypothesis where there is an effect, i.e. the hypothesis is of the form (8.1) below. Berger and Berry (1988) refer to such hypotheses as 'precise hypotheses', to distinguish them from 'one-sided' hypotheses such as $\theta < a$ versus $\theta > a$. When testing 'one-sided' hypotheses Bayesian and frequentist approaches consider the probability of the same events. For testing precise hypotheses Bayesian and frequentist approaches consider the probabilities of different events.

3. It is important to take into account prior probabilities for the hypothesis to be true, because the prior probability for a real effect may be small. A random SNP marker will have a low probability of being closest to the true locus.

Cases when Bayesian and frequentist inference are similar. A common use of hypothesis testing is selecting a model for a given dataset, where there is no cost, except for a negligible amount of computer time, resulting from including one or more spurious effects in the model. In this case the cost of wrongly selecting H_1 when the estimated effect is small, and the utility is mainly governed by predictive accuracy of the model. In this situation, frequentist inference gives similar results with large sample sizes as Bayesian inference with a *reference prior*, and using a p-value of 0.05 may be quite reasonable (Bernardo 1999; Ball 2005, Appendix B). Moreover, for many datasets commonly encountered in statistical practice, the variables are only measured because it is believed that there is a likely effect, hence prior probability would not usually be very low. For testing 'one-sided' hypotheses such as $H_{\mathrm{l}} : \theta < a$ versus $H_{\mathrm{u}} : \theta > a$, Bayesian and frequentist inference also gives similar results for large samples (Casella and Berger 1987).

In this chapter we show that for testing scientific hypotheses frequentist and Bayesian methods are not similar – respectable-sized Bayes factors correspond to very small p-values, therefore when only p-values are given the evidence for an effect is exaggerated, which can lead to spurious associations.

Non-Bayesian or 'frequentist' statistics

Non-Bayesian statistics, otherwise known as 'classical' or 'frequentist' statistics bases inference on values of a test statistic. Frequentist hypothesis tests compare the *null*

hypothesis (H_0) to the *alternative hypothesis* (H_1). For testing scientific hypotheses, such as a non-zero marker–trait association, H_1 typically has one or more unknown parameters than H_0, e.g.

$$H_0 : \theta = 0 \quad \text{versus} \quad H_1 : \theta \neq 0\,. \tag{8.1}$$

The observed value of a test statistic, T, chosen to measure departures from H_0 is compared to its sampling distribution under H_0. The observed value, $T_{\rm obs}$, of T is compared to its value under repeated sampling.

As noted in Chapter 7, making a decision based on p-values is problematic. A low p-value means there is evidence that H_0 may not be the perfect model for the data. A low p-value means the probability $\Pr(T \geq T_{\rm obs} \mid H_0)$ is small (cf. Equation (7.7)). However, the corresponding probability under H_1 may be equally small. Some threshold has to be chosen, but there is no method for choosing the optimal threshold. Choosing $p = 0.05$ as a threshold, i.e. 'rejecting' H_0 and choosing H_1, when $p \leq 0.05$ may or may not be a good strategy, whether p is the comparison-wise, genome-wise or experiment-wise p-value.

Bayesian statistics

Bayesian statistics is statistics soundly based on probability theory. Probability theory is used to represent one's knowledge about a system. Prior to observing data this is known as the *prior distribution*. Bayes' theorem is used to update the prior distribution to incorporate information in the data (Bayes 1763).

Bayesian and frequentist statistics give broadly similar answers for parameter estimation when there is sufficient data, relative to the complexity of the model, so that the prior has little effect. For testing scientific hypotheses such as (8.1) however the results are not similar (Berger and Sellke 1987), and for larger sample sizes the difference is greater (Ball 2005; Table 8.1).

Table 8.1. p-Values corresponding to various Bayes factors, for testing for linkage disequilibrium between a bi-allelic marker and QTL.

	Bayes factor (B)						
n	$1/20$	$1/10$	$1/5$	1	5	10	20
300	0.270	0.136	0.069	0.0139	2.83×10^{-3}	1.42×10^{-3}	7.18×10^{-4}
432	0.188	0.094	0.047	0.0096	1.94×10^{-3}	9.73×10^{-4}	4.89×10^{-4}
600	0.135	0.068	0.034	0.0068	1.38×10^{-3}	6.92×10^{-4}	3.47×10^{-4}
864	0.093	0.047	0.023	0.0047	9.49×10^{-4}	4.76×10^{-4}	2.39×10^{-4}
1,200	0.067	0.034	0.017	0.0034	6.84×10^{-4}	3.40×10^{-4}	1.71×10^{-4}
1,728	0.047	0.023	0.012	0.0023	4.69×10^{-4}	2.35×10^{-4}	1.18×10^{-4}
2,400	0.033	0.017	0.008	0.0017	3.37×10^{-4}	1.69×10^{-4}	8.44×10^{-5}
3,756	0.021	0.011	0.005	0.0010	2.15×10^{-4}	1.07×10^{-4}	5.37×10^{-5}
4,800	0.017	0.008	0.004	0.0008	1.68×10^{-4}	8.38×10^{-5}	4.19×10^{-5}

Reprinted from Ball (2005).

Bayesian updating takes the form

$$g(\theta \mid x) = \frac{f(x \mid \theta)\pi(\theta)}{\int f(x \mid \theta)\pi(\theta)\mathrm{d}\theta}\,, \tag{8.2}$$

where $g(\theta \mid x)$ is the *posterior distribution* of the unknown parameters θ given the data x, $\pi(\theta)$ is the *prior distribution* of the parameters and $f(x \mid \theta)$ is the *likelihood*, i.e. probability of observing the data for a given value of the parameters.

Note how information about x given θ in $f(x \mid \theta)$ has been turned into information about θ given x in $g(\theta \mid x)$.

Note: the technical difficulty in implementing Bayesian computations lies in evaluating the integral in (8.2), which is often analytically intractable. Nowadays, calculating the integral is generally avoided by using computationally intensive Markov chain Monte Carlo (MCMC) sampling methods. Gibbs sampling, Metropolis sampling and variants can be used to obtain a sample from $g(\cdot)$, and quantities of interest easily calculated from this sample (see, e.g. Gelfand *et al.* 1990; Gelman *et al.* 1995; Gilks *et al.* 1996). This methodology gives great modelling flexibility, and avoids the need for asymptotic (requiring large samples) and distributional (e.g. requiring normal, independent identically distributed data) assumptions.

The *Bayes factor* is defined as the ratio of the probability of observing the observed data under H_1 to that under H_0:

$$B = \frac{\Pr(\text{data} \mid H_1)}{\Pr(\text{data} \mid H_0)}\,. \tag{8.3}$$

The Bayes factor measures how much more likely the data are under H_1 than under H_0. If $B = 1$ the data are equally likely under H_0 as under H_1, i.e. there is no evidence either way. Values close to 1 are weak evidence. High values (greater than 1) are evidence for H_1, low values (less than 1) are evidence against H_1, or for H_0.

Given prior probabilities $\pi(H_0)$, and $\pi(H_1)$, for each hypothesis the corresponding posterior probabilities $\Pr(H_0 \mid \text{data})$ and $\Pr(H_1 \mid \text{data})$ are determined from the Bayes factor by

$$\frac{\Pr(H_1 \mid \text{data})}{\Pr(H_0 \mid \text{data})} = \frac{\Pr(\text{data} \mid H_1)}{\Pr(\text{data} \mid H_0)} \times \frac{\pi(H_1)}{\pi(H_0)}\,. \tag{8.4}$$

In other words,

$$\text{posterior odds} = \text{Bayes factor} \times \text{prior odds}\,. \tag{8.5}$$

Equation (8.5) is a consequence of Bayes' theorem, and states that the Bayes factor is the factor by which prior odds have increased to give posterior odds as a result of observing the data. The posterior odds are how much more likely we believe H_1 to be true than H_0 *after* observing the data. If H_1 and H_0 are the only two possibilities, then $\Pr(H_1 \mid \text{data}) + \Pr(H_0 \mid \text{data}) = 1$, i.e. the evidence can be equivalently specified by giving any one of the three quantities $\Pr(H_1 \mid \text{data})$, $\Pr(H_0 \mid \text{data})$ or the posterior odds, whichever is convenient.

Clearly, both the Bayes factor and prior odds are important factors contributing to the posterior odds. If the prior odds are low, a higher Bayes factor, i.e. stronger evidence from the data are required, to convince us of the likelihood of an effect.

Note:

1. The Bayes factor does not depend on prior odds. It does, however, depend on the prior distribution for parameters under H_1, especially the parameter(s) being tested, e.g. θ in (8.1).

2. The Bayes factor compares the probability of the data under both hypotheses, whereas the p-value considers only the probability of an event under H_0.

3. The Bayes factor or posterior probability considers only the observed data, unlike the p-value which considers the probability of unobserved values of the test statistic, under unobserved repeated sampling.

4. For a given experimental design and test setup, the smaller the p-value, the larger the Bayes factor will be. However the p-value needed to obtain a given Bayes factor gets smaller with increasing sample size (Table 8.1; Ball 2005). A p-value of 0.05 can correspond to evidence *against* H_1, e.g. with $n = 1,728$, $p = 0.047$ corresponds to a Bayes factor $B = 1/20$.

5. The Bayes factor has a natural interpretation as the strength of evidence from (8.5). The p-value is the probability of an unobserved event, and has no such interpretation independent of experimental design and test setup.

6. The p-value tends to exaggerate the evidence for H_1. A p-value of much less than 0.05 is needed to correspond to $B = 20$, i.e. 20-fold increase from prior to posterior odds in association studies. For example if $n = 300$ we need $p = 7.18 \times 10^{-4}$ to obtain a Bayes factor $B = 20$ in Table 8.1.

Various techniques are available for calculating Bayes factors, or the marginal probabilities $\Pr(H_i \mid \text{data})$, $i = 0, 1$, forming the numerator and denominator in the equation for the Bayes factor. The method of Spiegelhalter and Smith (1982) gives Bayes factors based on non-informative priors for one-way ANOVA models (Equation (8.8)). This form of the Bayes factor is used to test for differences between marker classes in independent population sample sizes in Section 8.3.2, and for designing experiments with power to detect effects with a given Bayes factor. Direct integration is used for the case–control studies in Section 8.3.3. The Savage–Dickey density ratio (Equation (8.32); Dickey 1971) gives the Bayes factors for nested hypotheses if the marginal posterior for the variable being tested can be evaluated. The Savage–Dickey density ratio is applied to calculate equivalent Bayes factors for the TDT, S-TDT transmission disequilibrium tests from Chapter 7, and for the TDT-Q1 test for continuous traits, in Section 8.3.4. General methods of estimating $\Pr(H_i \mid \text{data})$ from MCMC samplers are given by Raftery (1996).

Summary

Scientific hypotheses, such as whether a new drug is effective or whether a genomic region is associated with a trait, are tested statistically by Bayesian or frequentist methods. Scientific hypotheses often correspond to a 'precise hypothesis' of the form (8.1), where H_0 is a subset of H_1 obtained by setting one variable to zero. Frequentist inference uses

the p-value which measures probabilities of more extreme values of a test statistic than that observed if H_0 is the true model. In Bayesian inference the Bayes factor measures the strength of evidence, while posterior probabilities combine the evidence with prior probabilities for effects.

Bayesian and frequentist inference give similar results for testing one-sided hypotheses, or where there is negligible cost to making the wrong decision. For testing scientific hypotheses this is not the case: there is generally a substantial cost to making the wrong decision, and we are testing precise hypotheses. Bayesian and frequentist inference are not similar. For a given experiment, smaller p-values correspond to stronger evidence, but there is no general interpretation of the p-value as strength of evidence for H_1. Very small p-values are needed to correspond to a respectable Bayes factor with the kind of sample sizes needed for association studies. Therefore, we do not recommend p-values for testing scientific hypotheses. Bayes factors and/or posterior probabilities should be used instead.

8.2.2 Statistical approaches

There are a range of possible approaches to statistical analysis of association studies.

> There are numerous different approaches to significance testing of LD, ranging from simple contingency table chi-square tests through to complex likelihood-based procedures. If strong enough LD exists, any of the methods should give similar results. A more important issue than how to do the analyses is how to interpret the results. (Terwilliger and Weiss 1998)

Statistical approaches vary in model complexity and assumptions from simple ANOVA or t-tests for single markers, to more complex multi-locus models involving multiple markers or haplotypes. Multi-locus approaches can be model based using the coalescent, or mixed models based on IBD probabilities, to take account of correlation between similar haplotypes or more empirical approaches (Table 8.2).

The general approach is to compare statistical models with and without the association being tested, allowing for other relevant information, e.g. pedigree or marker locations, etc.

Single marker association studies versus haplotype-based analysis

The latter models make assumptions about population history, e.g. using the coalescent to simulate possible ancestral genealogies, and base inference on the simulated genealogies. This has the effect of allowing for similar effects of similar haplotypes. This should theoretically be more efficient, however the literature is divided.

Liu *et al.* (2001) claim, but do not prove, that haplotype analysis is more efficient:

> ... simply looking at the marginal dependency between each marker and disease status in a case/control sample of chromosomes is clearly inefficient. For an LD mapping strategy to be optimal in fine mapping, it is essential to consider the information observed in a set of contiguous markers (i.e., haplotypes).

In a review paper, Nielsen and Zaykin (2001) noted the literature was divided. Akey *et al.* (2000) suggest 'significant improvement in power and robustness of association

Table 8.2. Types of statistical approaches to association modelling.

	Single marker or one haplotype at a time	Multi-locus (haplotypes) (allow for correlation between haplotypes)	
		Model based	Empirical
Frequentist	Simple t-test or ANOVA or linear mixed models	Linear mixed models(1)	
Bayesian	t-test model or ANOVA model or linear-mixed models	Linear mixed models(1)	
		Coalescent(2,3)	Empirical(4,5)
		Uniform shrinkage prior(6)	

1. Correlations between haplotypes estimated from IBD probabilities (Section 8.3.6; Meuwissen and Goddard 2000, 2001).
2. Liu *et al.* (2001).
3. Zöllner and Pritchard (2005).
4. Molitor *et al.* (2003).
5. Product partition model or Bayesian model selection, this section.
6. Uniform shrinkage prior, combining coalescent and empirical methods, this section.

tests' while Long and Langley (1999) and Kaplan and Morris (2001) conclude 'single marker tests are at least as powerful as haplotype-based tests.' This was without considering the loss of information, when estimating haplotypes. Haplotype data are available where chromosome segments have been sequenced, or can be estimated where individuals have sufficiently many progeny.

In practice, haplotypes may need to be estimated from genotypic data, further reducing the power of haplotype-based methods. For haplotype estimation see, e.g. Stephens *et al.* (2001). Although higher LD may be found with the 'right' haplotypes, or group of haplotypes, there are many such possibilities, each with lower prior probability, hence requiring stronger evidence to reliably detect.

A pragmatic recommendation is to consider the haplotype-based approach where haplotype data from closely spaced loci is available for one or a small number of gene regions.

Coalescent-based models for haplotypes

> ...the coalescent is a stochastic process that provides good approximations to the distribution of ancestral histories that arise from classical forward-time models such as Fisher-Wright (Fisher 1930; Wright 1931) and Moran population models. (Drummond *et al.* 2002)

Coalescent-based approaches consider the relationships between haplotypes in the context of possible ancestral genealogies. This effectively generates a covariance structure on haplotypes. Similar haplotypes are likely to have a more recent common ancestor, and therefore are more likely have similar effects.

In essence, the coalescent simulates the evolutionary process backwards in time, considering recombinations and mutations. A *coalescence* occurs when a single segment is the common ancestor of two later segments. Figure 8.1 (Figure 1 from Nordborg and Tavaré 2002) shows a possible genealogy of a short chromosomal segment. The blue, green and red chromosome segments at the bottom of the figure are traced backwards in time, i.e. upwards in the figure, to their most recent common ancestor. Four events are labelled: Event 1 is a recombination and Events 2–4 are coalescences. The colour coding shows which parts of a chromosome are ancestral to which parts of chromosomes lower in the tree. Above a coalescence, multiple colours indicate that a segment is ancestral to multiple segments, e.g. at Event 4 the left-hand chromosome is ancestral to the red, green and blue chromosomes, while at Event 2 the left-hand chromosome is partly ancestral to both red and blue, and partly ancestral to red alone.

There is no uniquely determined ancestral genealogy, rather inference needs to consider possible genealogies according to their probabilities. Liu *et al*. (2001) give a fully Bayesian approach using MCMC to simulate from genealogies according to their probabilities in a coalescent model, illustrated with applications to cystic fibrosis and Friedrich's ataxia disease haplotypes.

The coalescent approach requires knowledge of several parameters, e.g. recombination and mutation rates, and embodies assumptions about the evolutionary process which may or may not accurately reflect the population history for the species or gene being considered. Estimates of the evolutionary parameters can be obtained from temporal sequence data for some species (Drummond *et al.* 2002). Fearnhead and Donnelly (2001) estimate recombination rates from population genetic data.

For further information on the coalescent see Kingman (1982), Hudson (1983, 1990), Nordborg (2001), Griffiths and Marjoram (1997), Stephens (2001), Nordborg and Tavaré (2002) and Stephens and Donnelly (2003) (with discussion by Bahlo *et al.* 2003; Wilson 2003).

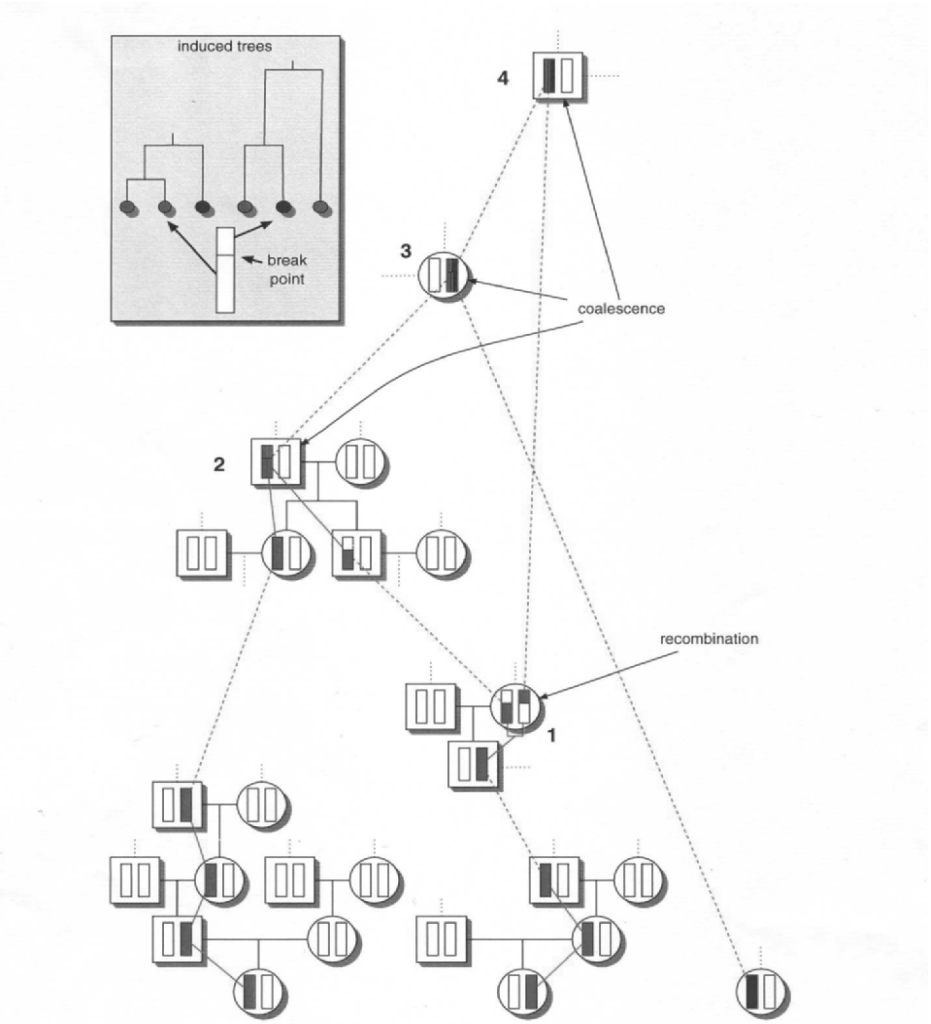
Mixed models for haplotypes

An alternative to the coalescent-based approach is to use a linear mixed model with haplotype effects as random effects. The correlations between haplotype random effects are given from IBD probabilities, which may be estimated from coalescent-based simulations, pedigree data or an analytical formula (Meuwissen and Goddard 2000, 2001).

When individuals are sampled from a known pedigree, the mixed model can also incorporate random effects representing polygenic variation, with covariance structure given by the additive relationship matrix for a pedigree. This controls for population structure resulting from non-random mating in the pedigree. Mixed models for haplotypes in a pedigree are discussed in Section 8.3.6.

Empirical multi-marker approaches

An alternative to the use of the coalescent in multi-marker or haplotype-based analyses is a purely empirical approach. A simple empirical approach would be *ad hoc* testing of each haplotype of interest versus the rest.



An example of a genealogy for three copies of a short chromosomal segment. Tracing the segmental lineages back in time, the following events occur: **1**, the "green" lineage undergoes recombination and splits into two lineages, which are then traced separately; **2**, one of the resulting green lineages coalesces with the "magenta" lineage, creating a segment, part of which is ancestral to both green and magenta, part of which is ancestral to magenta only; **3**, the "blue" lineage coalesces with the lineage created by event 2, creating a segment that is partially ancestral to blue and magenta, partially ancestral to all three colours; **4**, the "other" part of the green lineage coalesces with the lineage created by event 3, creating a segment that is ancestral to all three colours in its entirety. The recombination event induces different genealogical trees on either side of the break: these are shown in the inserted figure.

Figure 8.1. Example genealogy illustrating the coalescent (Nordborg and Tavaré 2002). (see color plate)

A more sophisticated approach, allowing for groupings of haplotypes with similar effects could be based on the 'product partition model' (Hartigan 1990; Barry and Hartigan 1992) estimated using computationally intensive Bayesian methods. With this approach, results from each possible grouping of haplotypes are combined according to the posterior probability for the grouping.

A third empirical approach is to use Bayesian model selection, where a 'model' consists of a selected set of markers regressed on the trait (Chapter 7, Section 7. 4. 4). This is less sophisticated, but by reducing the set of possible models, more computationally efficient than the product partition model. All possible models within the class of linear models with subsets of markers as explanatory variables are considered according to their probabilities. Ball (2001), reviewed in Sillanpää and Corander (2002), illustrate Bayesian model selection for QTL mapping with approximate posterior probabilities for models obtained using the Bayesian Information Criterion (BIC; Schwarz 1978). Unconditional estimates of effects, not subject to selection bias, are obtained by Bayesian model averaging (cf. Raftery *et al.* 1997). The same approach can be applied to association mapping.

In the first instance, additive terms for each marker would be included as possible variables, but epistasis can also be included, essentially by adding epistatic terms with the appropriate prior probability. Bogdan *et al.* (2004) adapt the BIC criterion to achieve the same effect. This approach is limited by the number of variables which can simultaneously be considered (about 30). This is not a problem for additive models, if a single linkage group in the QTL mapping context, or haplotypes in a single small chromosome region in the association mapping context, are being studied. It may not be possible to simultaneously consider all possible epistatic interactions between loci, because the number of possible models may be too large. One approach is to limit the interactions to loci already detected in additive models.

When the space of all models is too large, an alternative to considering all possible models is to search through the space of all possible models with an MCMC sampler. Since the MCMC sampler samples from models with probability proportional to their posterior probability in the long run, mainly models with reasonably high probability would be sampled. Sillanpää and Bhattacharjee (2005) is a recent MCMC approach using indicator variables to give a similar modelling framework, although they do not specifically consider interactions. This has the advantage that it is implemented in BUGS (Spiegelhalter *et al.* 1995). BUGS is a programming language and system for specifying Bayesian hierarchical models, and generating a Gibbs sampler. Implementing a model in BUGS is much quicker than developing an MCMC sampler from scratch in a conventional programming language, e.g. C, and additionally it is much easier to check BUGS code, and have confidence in the sampler. An analysis using BUGS is given in Section 8.3.4.

Note: an MCMC approach to Bayesian model selection was first given in George and McCulloch (1993). Variables not selected were given a prior concentrated around 0. This sampler was best for uncorrelated predictors, and could have poor convergence otherwise. Other Bayesian multi-locus methods for LD mapping include Kilpikari and Sillanpää (2003) and Meuwissen *et al.* (2001).

The empirical- and model-based coalescent approaches could be combined using a 'uniform shrinkage prior'. Two models represented by $f_1(x \mid \theta_1)$ (e.g. representing the model-based coalescent approach), and $f_2(x \mid \theta_2)$ (e.g. representing the empirical

approach) can be combined with a uniform shrinkage prior (Natarajan and Kass 2000) given by:

$$f(x \mid \theta_1, \theta_2, \lambda) = \lambda f_1(x \mid \theta_1) + (1 - \lambda) f_2(x \mid \theta_2)\,. \tag{8.6}$$

The uniform shrinkage prior is so named because the shrinkage parameter λ varying from 0 to 1 controls the relative influence of each model, and has a uniform prior distribution. Allowing for $\lambda < 1$ relaxes the strong model assumptions, allowing the data to say how much of the stronger model assumptions apply.

Summary

There are a range of approaches to the analysis of association study data. For inference, the general approach is to compare models with and without the effect being tested. Single marker analyses comparing one marker allele or haplotype versus the rest can easily be carried out using standard methods.

If haplotype data are available in a small genomic region, such as the vicinity of a functional locus, it may be more efficient to use haplotype-based methods: the fully Bayesian BLADE method based on sampling from the set of possible ancestral genealogies according to their posterior probabilities; or a mixed model, with random effects for haplotypes, and a covariance structure estimated from the coalescent or a deterministic formula. The assumptions inherent in the coalescent-based models can be avoided by using empirical models. Reduced dependence on assumptions comes at a possible cost of reduced accuracy or power. Further experience is needed to tell which of these approaches is most effective, and when.

8.2.3 Sources of 'spurious' associations or bias

In addition to problems with use of p-values there are a number of other potential causes of 'spurious' associations listed along with suggested possible solutions in Table 8.3.

Table 8.3. Problems and suggested solutions.

Problem	Solutions
p-values	Use Bayesian methods, Bayes factors and posterior probabilities.
Population substructure	Test for substructure. If present use STRAT type method (8.3.5) or TDT design (8.3.4), or allow for relatedness in a pedigree design (8.3.6).
Epistasis	When major additive genes or markers have been found allow for possible epistasis using a Bayesian model selection approach.
Non-genetic factors	Allow for factors as fixed or random effects in a mixed model.
(G×E) interactions	Allow for interactions as fixed or random effects in a mixed model.

8.3 EXPERIMENTAL DESIGNS

The main choices available to the experimenter are the number of individuals to sample, the number of markers to genotype per individual, and which traits to study. Power calculations allow choice of sample size so that the experiment has power to detect a QTL with given effect size and LD. Factors affecting the sample size required are summarised in Figure 8.2.

In this section we consider the various possible experimental designs – independent samples from a population without substructure (Section 8.3.2), case–control tests (Section 8.3.3), many small families for TDT type tests (Section 8.3.4), populations with substructure (Section 8.3.5) and samples from related individuals (i.e. pedigrees, Section 8.3.6). A strategy combining QTL and LD mapping is considered in subsection 8.3.7. Statistical methods specific to each type of design are discussed in the relevant subsections.

Design of experiments with power to detect effects with given Bayes factor for independent population samples is discussed in Section 8.3.2, and the independent population sample methods applied to results from candidate gene studies in *Eucalyptus* and maize in Examples 8.1 and 8.2. Frequentist and Bayesian case–control analyses are compared in Example 8.3 (the malaria data from Chapter 7, with variants). The power of case–control studies to detect genomic associations is assessed in Example 8.4 (APOE linkage disequilibrium data). A full Bayesian analysis using BUGS to implement a Gibbs sampler for simulated TDT-Q1 data are given in Example 8.7. Example 8.8 shows how LD are generated following admixture between sub-populations with differing allele frequencies.

8.3.1 Extent of linkage disequilibrium

The extent of LD is a major determinant of the resolution and cost of association studies. Information on the extent of LD is available for several species (Table 2.5).

If the extent of LD is short range, e.g. 4 kb, there is potentially very high resolution, but to exploit this requires genotyping many markers each with lower prior probability, hence stronger evidence is needed for each putative association, hence higher sample sizes are also needed. At the other extreme, if the extent of LD is long range, e.g. 10 cM the resolution is no more than available from modest QTL mapping pedigrees.

As noted in Chapter 2, The extent of LD found varies widely depending on mating system, species, population history, genomic region and sub-population sampled.

The extent of LD can vary within a species, e.g. if there is a sub-population with smaller effective population size (e.g. Europeans in Chapter 2, Table 2). This could have resulted from a small number of humans colonising Europe, as per the 'out of Africa' theory (Cavalli-Sforza and Cavalli-Sforza 1993; Foley 1995; Stringer and McKie 1996; Crow 2002; Sykes 2001; Wells 2003; Oppenheimer 2003).

The population history may include population bottlenecks, subdivisions, expansions or admixtures. In a population bottleneck, allele frequencies and LD values are subject to random genetic drift at a rate inversely proportional to population size, with effects proportional to the length of time a population is at that size. The alleles in each generation are a sample from the previous generation. For an allele at population proportion p the sample proportion $\hat{p}$, in the next generation is binomial with parameters n, p, where n is the population size. The variance of small binomial samples is proportional to p in absolute

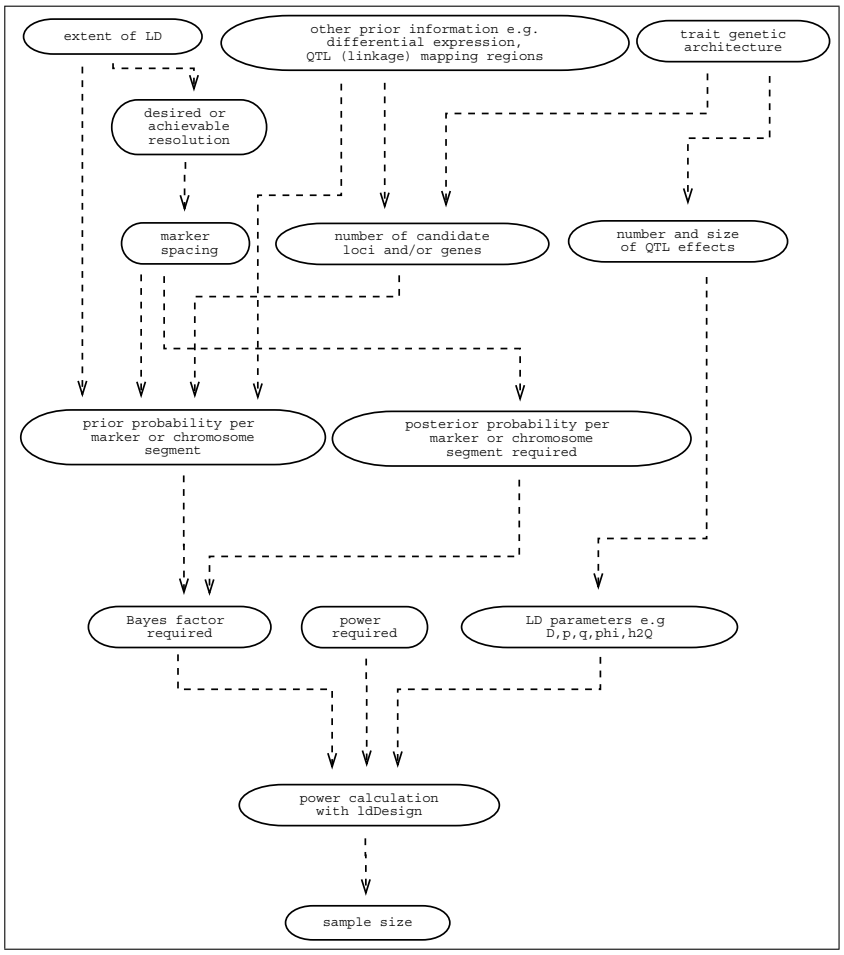


Figure 8.2. Sample size determination for detecting linkage disequilibrium.

terms, but in relative terms is inversely proportional to p:

$$\text{var}(\hat{p}/p) = (1-p)/(np)\,. \tag{8.7}$$

This shows that low-frequency alleles have greater relative variation in frequency from generation to generation. For a given population size the frequency of low-frequency alleles is more susceptible to random drift.

These factors contribute to the considerable variation in observed LD between loci in a population. Coalescent simulations have been used to study the resulting effects (Nordborg and Tavaré 2002).

LD patterns reflect the population history, including recent man-made influences, with short range LD patterns reflecting more ancient history, and long range patterns reflecting recent history, admixtures and inbreeding. So, paradoxically, there may be simultaneously short range of LD observed when single gene regions are examined, as well as long range recently introduced LD between more distant markers from a sample. How best to untangle this information is a challenging statistical problem.

Summary

The extent of LD is an important consideration in the design of association studies. There is currently limited information on extent of LD, and LD patterns in many species, however it is clear that the extent of LD varies widely between species, populations and genomic loci.

8.3.2 Independent sample from a population without substructure

Frequentist analysis

Suppose we have a random sample of n individuals from the population in consideration. For each individual suppose we observe the trait y and a bi-allelic marker with genotypes MM, Mm, mm. The ANOVA table is shown in Table 8.4. The p-value for testing for a difference between marker classes is obtained by referring the F-value to its distribution under H_0, i.e. the F-distribution.

Bayesian analysis

A Bayes factor for ANOVA models corresponding to a nearly non-informative prior is given by Spiegelhalter and Smith (1982) who obtain, for a one-way ANOVA model

$$B = \left[\frac{1}{2}\frac{(m+1)}{n}\prod_{i=1}^{m} n_i\right]^{-1/2}\left[1+\frac{(m-1)}{(n-m)}F\right]^{n/2}, \tag{8.8}$$

where m is the number of groups, n_i the number in each group, n the total sample size and F the classical F-value as in Table 8.4.

Table 8.4. ANOVA table for single marker analysis.

	df	SS	MS	F
Between marker classes	$\nu_1 = 2$	$\mathrm{SS_b}$	$\mathrm{MS_b} = \mathrm{SS_b}/\nu_1$	$F = \mathrm{MS_b}/\mathrm{MS_w}$
Within marker classes	$\nu_2 = n - 3$	$\mathrm{SS_w}$	$\mathrm{MS_w} = \mathrm{SS_w}/\nu_2$	

Reprinted from Ball (2005). Genetics 170:859–873.

Frequentist power calculations

Power is the probability of 'detecting' an effect, when a test statistic exceeds a predetermined threshold. The classical power calculation gives designs with power to 'detect' an effect with a given p-value. Luo (1998) corrected in Ball (2005) gives a deterministic power calculation for detecting a given level of LD between a bi-allelic marker and a bi-allelic QTL.

Bayesian power calculations

As noted above, p-values of, e.g. 0.05 or even much smaller may correspond only to weak evidence for an association. This is particularly a problem when trying to detect small effects associated with quantitative traits or complex diseases (Ball 2005). To better quantify the evidence for an effect, and avoid spurious associations resulting where p-values correspond to weak evidence, Ball (2005) adapts the method of Luo (1998), to give designs with given power to detect an effect with a given Bayes factor. More generally, this approach can be used to adapt any deterministic power calculation.

For power calculations, we replace n_i in (8.8) by their expected values

$$n_1 = np^2, \quad n_2 = 2np(1-p), \quad n_3 = n(1-p)^2. \tag{8.9}$$

where p is the population allele frequency of M giving

$$B \approx \left[4n^2p^3(1-p)^3\right]^{-1/2} \left[1 + \frac{2}{(n-3)}F\right]^{n/2}, \tag{8.10}$$

and F is the value of the classical F-statistic (Ball 2005), which corresponds to the p-value via the F-distribution.

This is implemented in the `ldDesign` R package (Ball 2004). Results for the examples from Luo (1998) are shown in Table 8.5. Additional columns are the Bayes factor, B, for the design and the sample size $n_{B_{20}}$ needed for a Bayes factor of 20 with power 0.9. Note that none of the original designs has $B > 1$ when $p = 0.05$.

Whole genome scans

Two approaches to finding genes are whole genome scans and candidate gene-based approaches. Whole genome scans may be applied in species such as maize, rice and poplar where the genome has been sequenced and substantial resources can be invested. Table 8.6 shows the sample sizes which are needed to obtain various posterior odds for

Table 8.5. Comparison with results from Luo (1998). Results are shown for the 12 example populations (cf. Luo Tables 2, 3) with sample size n, marker and QTL allele frequencies p, and q, linkage disequilibrium D, QTL heritability h^2_{Q} and dominance ratio ϕ. $\mathcal{P}_{0.05}$ is the power to detect an effect with $\alpha = 0.05$, B is the corresponding Bayes factor and $n_{B_{20}}$ is the sample size required to achieve a Bayes factor of 20 with power 0.9.

Populations	n	p	q	D	h^2_{Q}	ϕ	$\mathcal{P}_{0.05}$	B	$n_{B_{20}}$
1	100	0.5	0.5	0.1	0.1	0.0	0.18	0.88	1,837
2	200	0.5	0.5	0.1	0.1	0.0	0.34	0.42	1,837
3	200	0.5	0.5	0.2	0.1	0.0	0.91	0.42	381
4	200	0.5	0.5	0.1	0.2	0.0	0.62	0.42	849
5	200	0.5	0.5	0.1	0.1	0.5	0.31	0.42	2,047
6	200	0.5	0.5	0.1	0.1	1.0	0.25	0.42	2,640
7	200	0.3	0.3	0.1	0.1	0.0	0.46	0.54	1,211
8	200	0.7	0.7	0.1	0.1	0.0	0.46	0.54	1,211
9	200	0.3	0.5	0.1	0.1	0.0	0.39	0.54	1,476
10	200	0.5	0.3	0.1	0.1	0.0	0.39	0.42	1,513
11	200	0.4	0.6	0.1	0.2	1.0	0.45	0.45	1,259
12	200	0.6	0.4	0.1	0.2	1.0	0.54	0.45	995

Reprinted from Ball (2005).

Table 8.6. Sample sizes required for power of 0.9 of detection of linkage disequilibrium between a bi-allelic QTL and a bi-allelic marker with given posterior odds for linkage disequilibrium with $D = 0.1$, $p = 0.5$ and $q = 0.5$ in a genome scan with 500,000 SNP markers. Prior probability per marker is assumed to be 1/50,000.

		Sample size	
Posterior odds	Bayes factor	$h^2_{\mathrm{Q}} = 0.05$	$h^2_{\mathrm{Q}} = 0.01$
1/20	2,500	5,572	30,640
1/5	10,000	6,008	32,792
1	50,000	6,524	35,397
5	250,000	7,031	37,949
20	1,000,000	7,465	40,089

Reprinted from Ball (2005).

associations with small effect QTL in a genome scan. These values were calculated using the `ldDesign` R package (Ball 2004).

Prior probabilities per marker in Table 8.6 are based on the expected number of QTL affecting the trait and number of markers. This was based on the assumption that QTL are equally likely to occur anywhere on the genome and assuming an expected number of 10 QTL. This corresponds to a Poisson prior probability distribution for the number of QTL in the genome with rate $\lambda_{\mathrm{G}} = 10$. The prior distribution for locations of QTL, assuming they exist is generally assumed to be uniform: the probability that a QTL is within a small genomic interval of width Δx is

$$\Pr[\text{QTL in } (x, x + \Delta x)] = \lambda_{\mathrm{G}} \frac{\Delta x}{L_{\mathrm{G}}}, \tag{8.11}$$

where L_{G} is the genome length in base pairs.

With 500,000 SNP markers this equates to an average probability of 1/50,000 per marker. For unequally spaced markers, the prior probability for a marker at position x_i with flanking markers at x_{i-1}, x_{i+1} we take $\Delta x = 1/2(x_{i+1} - x_{i-1})$, which is the width of the sub-interval of points closer to x_i than to the flanking markers, in (8.11). The prior probability π_i for the ith marker is then given by:

$$\pi_i = \Pr[\text{QTL in } (1/2(x_i - x_{i-1}), 1/2(x_{i+1} - x_i))] = \lambda_{\mathrm{G}} \frac{x_{i+1} - x_{i-1}}{2L_{\mathrm{G}}} \,. \tag{8.12}$$

The Poisson distribution with rate 10 QTL per genome has mean 10, and standard deviation about 3, and has 95% of its probability in the range from 4 to 17. Hence there is some flexibility in the prior – we are not assuming the number of QTL is exactly 10. If this prior is too precise, we can allow for more uncertainty in the number of QTL by using a mixture of Poisson distributions, e.g. a mixture of Poisson distributions with means 3,5,10,20, with mixing probability 0.25 for each rate, which has a mean and standard deviation approximately 9.5, and 7.2, respectively, and 95% of its probability in the range 1–26. Power calculations for this composite prior can be obtained by noting that for a given n the power to obtain a given Bayes factor B is the same, regardless of the prior, but the posterior probabilities are different. In general, for a mixture prior π, with prior probabilities per marker π_i, and mixing proportions c_i for $1 \leq i \leq k$

$$\pi = c_1\pi_1 + \cdots + c_k\pi_k \,, \tag{8.13}$$

the posterior probability for H_1 is given by:

$$\Pr(H_1 \mid \pi) = \sum_{i=1}^{k} c_i \frac{B\pi_i}{1 - \pi_i + B\pi_i} \,. \tag{8.14}$$

Where there is no prior information for a given locus, we may be guided by the number of QTL found at other loci, or by information on similar traits in other species. The trait genetic variance gives an upper bound for variance explained by each individual QTL. Results from QTL mapping studies also contain useful prior information, e.g. undetected QTL are likely to be small enough to a reasonable chance of escaping detection. An upper bound on QTL magnitude for undetected QTL together with the amount of unexplained genetic variance gives a lower bound for the number of undetected QTL. For example, if the QTL detection experiment was sufficiently powerful so that each undetected QTL explains 5% or less of the total variance, and there are two detected QTL explaining, in total 20% of the variance of a trait with heritability 50%, that leaves 30% of the variance, which is genetic, unexplained. Therefore, there should be at least six loci explaining the remaining 30%. A prior rate of $\lambda_{\mathrm{G}} = 8$ loci per genome would be reasonable. QTL mapping studies also contain prior information on the locations of detected QTL (cf. Section 8.3.7).

Candidate gene mapping

For more resource-limited species or where the genome has not been sequenced a candidate gene-based approach may be preferred. Candidate genes may be pre-selected based on prior information from one or more of the following:

Table 8.7. Sample sizes required for power of 0.9 of detection of linkage disequilibrium between a bi-allelic QTL and a bi-allelic marker with given posterior odds for linkage disequilibrium with $D = 0.1$, $p = 0.5$ and $q = 0.5$ in a set of 50,000 markers representing candidate genes. Prior probability per marker is assumed to be 1/5,000.

		Sample size	
Posterior odds	Bayes factor	$h^2_{\mathrm{Q}} = 0.05$	$h^2_{\mathrm{Q}} = 0.01$
1/20	250	4,826	26,808
1/5	1,000	5,288	29,093
1	5,000	5,808	31,658
5	25,000	6,322	34,223
20	1,00,000	6,762	36,406

Reprinted from Ball (2005).

– In a QTL mapping region, or,

– Associated with the trait in a model species, or,

– Associated with processes likely to affect the trait in a model species, or,

– Differentially expressed genes.

Table 8.7 shows the sample sizes, which are needed to obtain various posterior odds for associations with small effect QTL in a genome scan. These values were calculated using the `ldDesign` R package (Ball 2004).

Prior probabilities per marker in Table 8.7 again assume a Poisson distribution with rate 10 QTL for the number of QTL. If there are 50,000 candidate genes, this equates to a prior probability of 1/5,000 per candidate.

The prior probability is clearly an important factor in our ability to find genes. Candidate genes with a lower prior probability need a higher Bayes factor to obtain the same posterior probability.

Example 8.1. A candidate gene-based association study in Eucalyptus.

Thumma *et al.* (2005) study associations between SNPs and haplotypes in a candidate gene *Cinnamoyl CoA reductase* (*CCR*), 'a key lignin gene, which has been shown to affect physical properties of the secondary cell wall in *Arabidopsis*', and microfibril angle (MFA). This is the (mean) angle at which microfibrils making up the S2 layer of the secondary cell wall are oriented relative to the longitudinal axis of wood fibres. High MFA is associated with lower longitudinal stiffness, and poor stability in juvenile wood, hence there is interest in breeding to reduce MFA or selection of trees to avoid high MFA.

Note: this example, and the next, was chosen because there are as yet few examples of published associations in plants with sufficient information for us to calculate Bayes factors. Many papers use frequentist measures of evidence and similar sample sizes, so problems with statistical evidence identified here are likely to be widespread, as is the case in the human genetics literature (cf. Terwilliger and Weiss 1998; Altshuler *et al.* 2000; Emahazion *et al.* 2001).

Within the candidate gene, 25 SNP markers were tested for associations with MFA in an independent sample of $n = 290$ *E. nitens* trees. A putative association with SNP21 explained an estimated 4.6% of the variation, and had a reported experiment-wise p-value of 0.0002 (Thumma *et al.*, Table 2). The comparison-wise p-value corresponding to an effect explaining 4.6% of the variation, with the given allele frequencies was calculated by us as 0.00023 .

The apparently strongest associations (SNP20, SNP21) were not segregating in the validation families. However, associations were 'validated' ($p < 0.05$), for nearby markers, in two full-sib families of *E. nitens* ($n = 287, p = 0.02$) and *E. globulus* ($n = 148, p = 0.04$). In the validation samples the effect sizes were smaller, and with less significant p-values, than in the association population. At this point readers should ask themselves: how good is the evidence? Should we consider the associations validated?

The results for the most 'significant' associations from Thumma *et al.*, Tables 3, 5 are shown in Table 8.8. To better assess the evidence from the population and validation samples separately and combined, we converted all p-values to individual comparison-wise p-values, calculated the corresponding F-values and then calculated the Spiegelhalter and Smith Bayes factors (Equation (8.8); R function `SS.oneway.bf()` from `ldDesign`). For the association population, we calculated the comparison-wise p-value based on the reported percent variation explained (4.6%), and the allele frequency for the SNP. The p-values for the validation populations were already comparison-wise.

Frequentist interpretation. The p-values show a 'highly significant' association in the population sample, supported by significant associations in the two QTL mapping families.

Bayesian interpretation. The Bayes factors show strong evidence in the data ($B = 98$) for an effect in the association population, but very weak evidence in the validation families ($B = 1.5, 1.1$). A Bayes factor of 98 normally represents strong evidence, however if the prior odds are low as in Tables 8.6 and 8.7, the posterior probabilities for an association will be low.

Note that the 'validation' of this association in the QTL mapping families, even if the evidence was good, would be supporting evidence for, but would not validate an association with SNP21. An association in the QTL mapping families could result from QTL at some distance (e.g. 20 cM) from the SNP locus, in either the Bayesian or frequentist paradigms. A better approach to combining QTL and association mapping inference is to use the QTL posterior probability distribution to improve the prior odds for the association mapping analysis. This approach is studied in Section 8.3.7.

Prior and posterior probabilities for various priors are shown in Table 8.9. Priors 1, 2 and 3 are given for a random SNP, a random candidate gene and a candidate gene with some fairly strong prior information, respectively.

Table 8.8. Statistics for markers with 'significant' associations with MFA.

Population	n	Marker	Freq	%Var	P	B
E. nitens association pop.	290	SNP21	0.31	4.6	0.00023	98.4
E. nitens family	287	SNP18	0.5	0.45	0.02	1.5
E. globulus family	148	SNP120	0.5	0.69	0.04	1.1

Table 8.9. Prior and posterior probabilities for an association with SNP21, for various priors.

Prior information	Prior probability	Posterior odds	Posterior probability
1. Random SNP	1/50,000	1/508	0.002
2. Random SNP within candidate gene	$1/1,00,000^{\dagger}$	1/1,016	0.001
3. Differentially expressed candidate gene	$1/40^{\ddagger}$	2.5/1	0.72

†Based on assuming 40,000 genes in the genome and 10 genes expected to affect the trait.

‡A value that might occur in similar experiments, e.g. assuming 200 differentially expressed genes, 10 genes expected to significantly affect the trait and 50% of genes significantly affecting the trait expected to be differentially expressed.

For example, for a random SNP selected from the genome the prior probability per SNP might be 1/50,000, posterior odds are 1/508 and the posterior probability is 0.002.

Clearly posterior probabilities for a real effect are low except in case 3, where the candidate gene is *a priori*, not unlikely. The authors did not give prior probabilities for an association. Their candidate gene was selected from a set of differentially expressed genes, and was also associated with stiffness in *Arabidopsis*. However, the associations in *Arabidopsis* are not with the trait in consideration, i.e. MFA. In the absence of other evidence, since we have no reason to expect a lignin gene to causally affect MFA, we would, use prior 1 or 2. For respectable posterior odds of 20:1 or more, with the Bayes factor obtained, the prior odds should be at least 1:5. If prior odds of 1:5 (around 800 times better than for a random candidate gene, representing stronger evidence than the data) are used, these need careful justification.

Selection bias The reduction in estimated magnitude of the effects, in the validation population compared to the association population, could be due to validation with different markers. This phenomenon is also typical of *selection bias*. Significant effects, originally estimated from the same population used to test for significance tend to be biased upwardly, a phenomenon known as selection bias. Estimates free of selection bias should be given. These can be obtained, either by using an independent population, or, in a Bayesian context by considering multiple models, not just the models where the marker is selected (cf. Ball 2001, for application in a multi-marker-QTL mapping context), and averaging over models according to their posterior probabilities.

As a special case, this applies when a single marker or haplotype is being tested. In this case there are two possible models. These correspond to H_1, the *alternative hypothesis*, where the marker is selected and H_0, the null hypothesis where there is no effect, i.e. the effect is zero, respectively. Allowing for selection bias means allowing (with non-zero probability) for the possibility that H_0 is true, in which case the effect is zero. Otherwise selection bias occurs if $\Pr(H_0 \mid \text{data})$ is not small. The unconditional estimate of marker effects is obtained by averaging effects in each model according to the posterior probabilities. With priors 1 and 2 in Table 8.9, the resulting estimates would be very small since

the posterior probabilities for H_1 are small. With prior 3, the posterior probability of 0.72 would mean the estimates reduce by a factor of 0.72 and the percentage variation explained reduces by the square of this factor, or 0.52.

Power Frequentist methods give approximately valid results, approximately free of selection bias and without the need to use an independent sample, if the power to detect the *true effect* is good, e.g. 0.9 . The difficulty is that we do not know the true effect, we only have the estimated effect. Often, even if the power to detect the estimated effect is reasonable, the true effect may be smaller, hence suffer from selection bias. We could be reasonably sure of good power if the lower limit of a 95% confidence interval for the estimated effect was larger than the value for which the power is 0.9. Often, experiments are designed with power of 0.9 to detect an effect with $p = 0.05$, i.e. two standard deviations greater than zero. The 95% c.i. for the estimated effect would then be greater than this value if the effect was at least four standard deviations greater than zero, or a p-value of around 0.0001.

Finally, we examine the power of the experiment using `ldDesign` (Ball 2004, 2005). The power, calculated using `ldDesign`, of the experiment to detect LD with $D = 0.1, 0.2$, with the given allele frequencies is shown in Figure 8.3. Power to detect LD with $D = 0.1$, with a Bayes factor of 20 is very low (0.04), but nearly 0.5 to detect LD with $D = 0.2$ (nearly its maximum for the given allele frequencies). The indicated sample size for a power of 0.9 is 575, or nearly twice the size of the experiment. To detect a QTL explaining 5% of the variation with $D = 0.1$ and a Bayes factor of 20 or more requires a population of around 2,730, or almost 10 times the size.

Figure 8.4 shows power to detect LD between a bi-allelic marker and QTL with a given Bayes factor, as a function of sample size. Allele frequencies are assumed to be 0.31 (the same as for SNP21) for both marker and QTL. Each panel corresponds to a combination of $D = 0.1$ or 0.2, and QTL heritability $h^2_{\mathrm{Q}} = 0.01$ or 0.05, i.e. explaining 1% or 5% of trait variation. Within each panel power curves are given for power to detect associations with Bayes factors of 20, 1,000 or 1,000,000.

```
> ld.power(n=290,p=0.31,q=0.31,h2=0.05,phi=0,Bf=20,D=0.1)
       n  power
[1,] 290  0.038
> ld.power(n=290,p=0.31,q=0.31,h2=0.05,phi=0,Bf=20,D=0.2)
       n  power
[1,] 290  0.495
> ld.design(p=0.31,q=0.31,D=0.1,h2=0.05,Bf=20,phi=0,power=0.9)
[1] 2727.228
> ld.design(p=0.31,q=0.31,D=0.2,h2=0.05,Bf=20,phi=0,power=0.9)
[1] 575.1845
```

Figure 8.3. Power calculations with `ldDesign`.

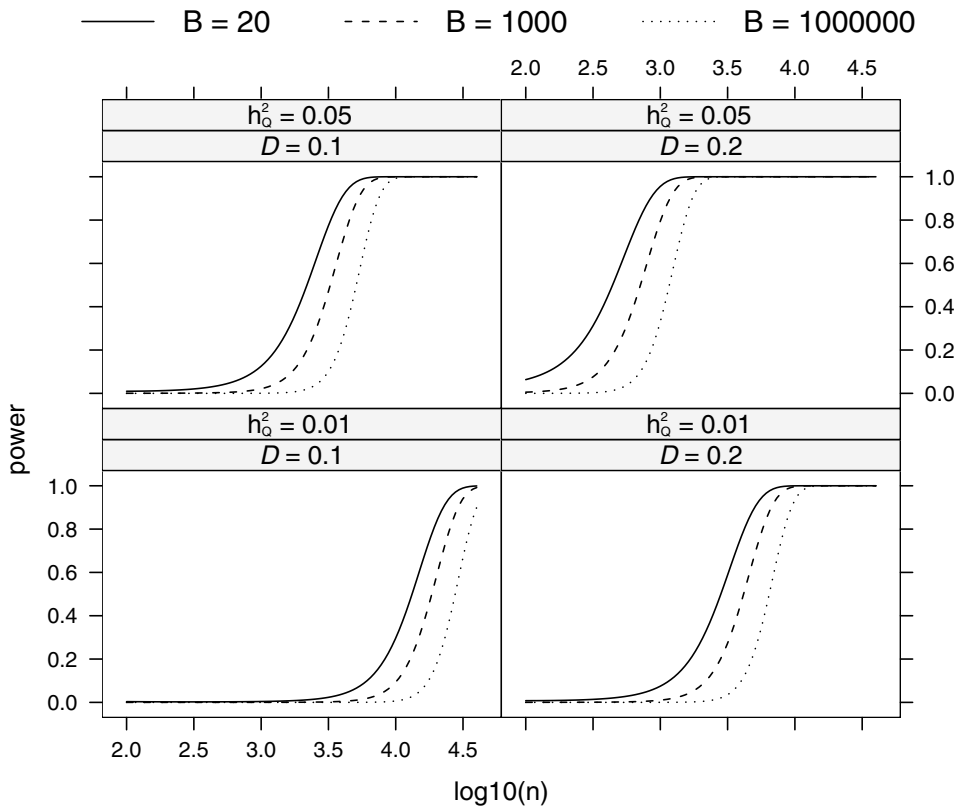


Figure 8.4. Power versus sample size for various levels of disequilibrium.

Example 8.2. A candidate gene-based association study in maize.

Thornsberry *et al.* (2001) study associations between polymorphisms of the *Dwarf8* gene and time to flowering in maize. Table 8.10 shows test statistics for time to silking in five-fields. Values for the whole gene are based on the log-likelihood ratio statistics $\ln(\Lambda_{\max})$ which are the maximum values over 41 haplotypes for each field. The comparison-wise p-value for the log-likelihood ratio statistic was calculated by reference to the χ^2 distribution on one degree of freedom

$$2 \times \ln(\Lambda) \sim \chi^2_1, \tag{8.15}$$

and comparison-wise p-values were calculated for the deletion flanking *sh2* by reference to the t-distribution. Bayes factors were based on assuming allele frequencies of 0.5. Somewhat higher values were obtained for Bayes factors based on allele frequencies of 0.1 and 0.9 (not shown).

As with the previous example fairly high Bayes factors were obtained, but strong prior information (prior odds of no more than 20:1 against an effect, compared to 1/4,000 for

Table 8.10. Likelihood ratios, comparison-wise p-values and Bayes factors for time to silking in five fields from Thornsberry *et al.* (2001). Bayes factors are based on allele frequencies of 0.5.

	Whole gene			Deletion flanking *sh2*		
Field	$\ln(\Lambda_{\max})$	P	B	Effect $\pm$s.e. (days)	P	B
S1999/A	9.00	2.2×10^{-5}	1,754.2	-10 ± 3	0.001	36.0
S1999/B	8.11	5.6×10^{-5}	704.1	-9 ± 3	0.003	13.7
S1999/C	7.08	1.7×10^{-4}	244.7	-10 ± 3	0.001	36.0
W1999	7.67	9.0×10^{-5}	448.3	-7 ± 2	0.0007	60.3
W2000	8.56	3.5×10^{-5}	1,117.1	-8 ± 3	0.009	5.6

Reprinted from Thornsberry *et al.* (2001).

a random candidate gene assuming 10 genes affecting the trait), or lower for a random haplotype within the gene is needed to obtain respectable posterior odds.

Note: readers may notice the variability in the Bayes factor between fields. The log Bayes factors from the final column in Table 8.10 had a sample standard deviation of 0.95. This level of variability (if applying independently across haplotypes) would contribute a 3.6–20.3-fold increase (95% c.i. for the maxima of the log Bayes factor) in the Bayes factor by chance as a result of maximising over 41 haplotypes. With a 20-fold reduction, the whole-gene Bayes factors become comparable to the *sh2* values.

Summary

Standard statistical methods, such as the frequentist ANOVA method, can be used to analyse associations in independent samples. However, due to problems with the interpretation of p-values, Bayes factors and posterior probabilities for H_1 are the recommended measures of evidence.

Using a correspondence between p-values for one-way ANOVA models and the Spiegelhalter and Smith Bayes factors enables us to use existing power calculations to find the sample sizes required to detect effects with a given Bayes factor. The same technique is useful for estimating a Bayes factor based on results where only p-values are published. Results consistently showed that the p-value is not a reliable measure of evidence. The p-values corresponding to a respectable Bayes factor were very low, and varied considerably.

The power of an experiment to detect an association between a bi-allelic marker and QTL with given sample size, allele frequencies and LD coefficient can be calculated using the `ldDesign` R library. Or, the sample size required for a given power can be calculated.

Prior probabilities and the Bayes factor combine to give posterior probabilities. Prior distributions are a mathematical representation of prior knowledge. Priors are subjective, there is no 'right' or 'wrong' prior; different observers will have different priors. With experience, priors can be chosen which are a reasonable representation of the available prior knowledge. Priors based on the Poisson distribution can be used for the number of QTL present in the genome, and this information used to obtain probabilities that a QTL is present in a given genomic region, e.g. the vicinity of a marker. Good prior information may substantially increase prior odds, hence reducing the sample size needed. But it is important not to overstate prior information. If necessary, mixtures of Poisson distributions can be used to obtain less informative priors.

The key to obtaining a high posterior probability for detected QTL is to design the experiment with good power to detect QTL with a given Bayes factor, where the Bayes factor is chosen sufficiently large to overcome the low prior odds.

The methods were applied to candidate gene studies in *Eucalyptus* and maize (Examples 8.1 and 8.2). Respectable Bayes factors of around 100 were obtained in both examples, but these were not high enough to overcome the low prior odds for candidate genes. Lessons learnt from these examples include:

- Approximate Bayes factors can be found from experiments where comparison-wise p-values are reported if the sample sizes in each marker class are given.
- Large Bayes factors are needed to overcome the low prior odds in genome scans or candidate gene studies.
- Low prior probabilities for a genome scan or candidate gene region apply even if testing a single gene region, unless there is independent evidence for the region to contain loci affecting the trait in consideration.
- When power is not good estimates of effects for the 'detected' markers will be inflated by selection bias.
- QTL mapping results can support but do not validate LD mapping associations.

8.3.3 Case–control studies

The observed counts and expected proportions in a case–control test with two marker classes are shown in Table 8.11. The proportions p_{ij} are by definition proportions of the row totals. They are not independent since proportions add up to 1 across rows. Under H_1, proportions p_{12} and p_{22} of the allele S under cases and controls are not necessarily the same, so the model can be parameterised by p_{12}, p_{22}. Under H_0 the proportions p_{12} and p_{22} are the same by hypothesis, so the model can be parameterised by p_{12} alone, setting $p_{22} = p_{12}$.

Example 8.3. A case–control test for malaria.

The number of occurrences of each allele for cases and controls is shown in Table 8.12.
Frequentist analysis (cf. Chapter 7).

$$\chi^2 = 41.3 \sim \chi^2_1, \quad P = 1.33 \times 10^{-10}\,. \tag{8.16}$$

Table 8.11. Observed counts and expected proportions in a case–control test with two marker classes.

	Observed counts		Expected proportions	
	A	S	A	S
Case	n_{11}	n_{12}	p_{11}	p_{12}
Control	n_{21}	n_{22}	p_{21}	p_{22}

Table 8.12. Frequencies of alleles in the case–control test for the malaria data.

	A	S
Case	623	7
Control	1,065	101

Bayesian analysis. Under H_1, let p_{12}, p_{22} be the expected proportions of allele S for cases and controls, respectively, with Beta$(1/2, 1/2)$ prior distributions. Under H_0 we assume $p_{12} = p_{22}$ and let p_{12} have a Beta$(1/2, 1/2)$ prior.

Note: the Beta prior is a conjugate prior for binomial sampling, meaning that if the prior is a Beta distribution, and a binomial sample is observed, the posterior distribution is also a Beta distribution. If the prior for p is Beta(a, b), and k successes are observed in n Bernoulli trials, then the posterior is Beta$(a + k, b + n - k)$. A Beta$(1/2, 1/2)$ distribution is the standard Jeffreys prior (Jeffreys 1961) with mean 0.5, and information equivalent to one Bernoulli trial. The density for a Beta(a, b) distribution is

$$f(p \mid a, b) = \frac{1}{B(a, b)} p^a (1 - p)^b \,, \tag{8.17}$$

where $B(a, b)$ is the value of the Beta function given by

$$B(a, b) = \int_0^1 p^a (1 - p)^b \mathrm{d}p \,, \tag{8.18}$$

i.e. the factor needed to make $f(p \mid a, b)$ in Equation (8.17) a probability density. Values of $B(a, b)$ can be calculated with the standard R function `beta()`. When the values of $B(a, b)$ are very small it is best to work the logarithm of the values calculated directly with the R function `lbeta()`.

We now calculate the Bayes factor, by explicitly integrating out p_{12}, p_{22} for H_1 and p_{12} for H_0.

$$\begin{aligned} \Pr(\text{data} \mid H_1) &= \int\int \binom{630}{7} p_{12}^7 (1 - p_{12})^{623} \times p_{12}^{0.5} (1 - p_{12})^{0.5} / B(0.5, 0.5) \times \\ &\quad \binom{1,166}{101} p_{22}^{101} (1 - p_{22}^{1,065}) \times p_{22}^{0.5} (1 - p_{22})^{0.5} / B(0.5, 0.5) \mathrm{d}p_{12} \mathrm{d}p_{22} \\ &= \binom{630}{7} \binom{1,166}{101} \frac{B(623.5, 7.5) B(1, 065.5, 101.5)}{B(1/2, 1/2) B(1/2, 1/2)}. \end{aligned} \tag{8.19}$$

Similarly

$$\Pr(\text{data} \mid H_0) = \binom{630}{7} \binom{1,166}{101} \frac{B(1, 688.5, 108.5)}{B(1/2, 1/2)} \tag{8.20}$$

so the Bayes factor is

$$\begin{aligned} B &= \frac{\Pr(\text{data} \mid H_1)}{\Pr(\text{data} \mid H_0)} = \frac{B(623.5, 7.5) B(1, 065.5, 101.5)}{B(1/2, 1/2) B(1, 688.5, 108.5)} \\ &= 1.0 \times 10^{10} \,. \end{aligned} \tag{8.21}$$

Table 8.13. χ^2 and Fisher's exact test statistics and Bayes factors for three case–control datasets.

Dataset	Data	Statistics
I	$\begin{pmatrix} 623 & 7 \\ 1{,}065 & 101 \end{pmatrix}$	$X^2 = 41.3$ $P_{\chi^2} = 1.3 \times 10^{-10}$ OR $= 8.31\,(3.91, 21.6)$ $P_{\text{Fisher}} = 9.9 \times 10^{-13}$ $B = 1.0 \times 10^{10}$
II	$\begin{pmatrix} 602 & 28 \\ 1{,}065 & 101 \end{pmatrix}$	$X^2 = 10.0$ $P_{\chi^2} = 9.5 \times 10^{-4}$ OR $= 2.04\,(1.31, 3.26)$ $P_{\text{Fisher}} = 7.7 \times 10^{-4}$ $B = 13.9$
III	$\begin{pmatrix} 594 & 36 \\ 1{,}065 & 101 \end{pmatrix}$	$X^2 = 5.0$ $P_{\chi^2} = 0.02$ OR $= 1.56\,(1.04, 2.39)$ $P_{\text{Fisher}} = 1.4 \times 10^{-2}$ $B = 0.5$

More generally if n_{ij} are as in Table 8.11, the Bayes factor is

$$B = \frac{B(n_{11} + 0.5, n_{12} + 0.5)B(n_{21} + 0.5, n_{22} + 0.5)}{B(0.5, 0.5)B(n_{11} + n_{21} + 0.5, n_{12} + n_{22} + 0.5)} \tag{8.22}$$

Frequentist and Bayesian analyses are compared for 3 possible case – control datasets in Table 8.13. Fisher's exact test was computed using the R function `fisher.test()`, with a two-sided alternative. In Dataset I (the malaria data) the chi-squared test p-value P_{χ^2} is of the order of 10^{-10}, the Fisher exact test even smaller at $P_{\text{Fisher}} = 9.9 \times 10^{-13}$, and the corresponding Bayes factor of the order of 10^{10}, representing very strong evidence. Even after allowing for prior odds of 1/500,000 the posterior probability for an association will be high. Datasets II and III had higher values of the S allele for the cases and lower values of the A allele than Dataset I but had the same row totals. Dataset II had $P_{\chi^2} \leq 0.001$, and P_{Fisher} slightly smaller. Both of these values are commonly considered 'highly significant' in frequentist analyses. The corresponding Bayes factor was 13.9, representing only moderate evidence and not enough to overcome the low prior odds for most associations. Dataset III had $P_{\chi^2} = 0.02$, and P_{Fisher} similar. These values are normally considered 'significant' in frequentist analysis. The corresponding Bayes factor was only 0.5, representing weak evidence *against* H_1.

Odds ratios and relative risks. The odds ratio is given by

$$\text{OR} = \frac{p_{12}p_{21}}{p_{11}p_{22}} = \frac{p_{12}(1 - p_{22})}{(1 - p_{12})p_{22}}\,. \tag{8.23}$$

The model for H_1 could have been parameterised in terms of odds ratios, rather than p_{12}, p_{22}. Nevertheless we can compute the posterior distribution for odds ratios from the

```
> p12.sim <- rbeta(3000,7.5,623.5)
> p22.sim <- rbeta(3000,101.5,1065.5)
> OR.sim <- p12.sim*(1-p22.sim)/((1-p12.sim)*p22.sim)
> # mean =0.13,median=0.12, 95% ci = (0.05,0.24)
> stats(OR.sim)
     mean   stdev     sem    2.5%     25%     50%     75%   97.5%
    0.128   0.048 0.00088   0.053   0.092   0.121   0.156   0.239
```

Figure 8.5. R calculations for simulation from the posterior distribution of the odds ratio for Dataset I.

Table 8.14. Posterior statistics for log-odds ratios. Conditional estimates are made assuming H_1 is true. Unconditional estimates average estimates under H_0 and H_1, according to their posterior probabilities. The posterior probabilities $p_{H_1} = \Pr(H_1 \mid \text{data})$, were estimated assuming prior odds of 1/4,000 appropriate for a candidate gene if 10 genes out of 40,000 are expected to contribute to the disease.

Dataset	Conditional on H_1			Unconditional			
	Mean	Standard deviation	95% c.i.	Mean	Standard deviation	p_{H_1}	Selection bias (%)[1]
I	−2.13	0.39	(−2.94,−1.41)	−2.13	0.39	1.0000	0
II	−0.71	0.22	(−1.15,−0.29)	−0.0074	0.0443	0.0030	9,495
III	−0.44	0.20	(−0.84,−0.05)	−0.00027	0.00555	0.0001	162,900

[1]Selection bias is estimated as the bias from assuming H_1 is true as a percentage of the unconditional estimate.

posteriors for p_{12}, p_{22} which are:

$$p_{12} \sim \text{Beta}(n_{12} + 1/2, n_{11} + 1/2)\,, \quad p_{22} \sim \text{Beta}(n_{22} + 1/2, n_{21} + 1/2)\,, \tag{8.24}$$

under H_1.

R code for the odds ratio simulation for Dataset I is shown in Figure 8.5. We use the fact that p_{12} and p_{22} are independent (since p_{12} depends only on the case data and p_{22} depends only on the control data). Posterior statistics for the log-odds ratio for each dataset are shown in Table 8.14. Both conditional and unconditional estimates are shown. Unconditional estimates are obtained from the mixture distribution

$$f(\theta) \sim (1 - p_{H_1}) f(\theta \mid H_0) + p_{H_1} f(\theta \mid H_1), \tag{8.25}$$

where p_{H_1} is the posterior probability of H_1, and $f(\theta \mid H_0), f(\theta \mid H_1)$ are the posteriors for θ (here the log-odds ratio) under H_0, H_1, respectively.

Note: the odds ratios in Figure 8.5, are the reciprocals of those calculated for the Fisher exact test in Table 8.13, due to the use of a different convention.

Under H_0 the log-odds ratios are all zero, hence non-zero posterior probability for H_0 leads to the unconditional estimates being smaller. The conditional estimates are affected by selection bias, because effects over-estimated in absolute value tend to be selected, by whatever criteria is used, whereas the unconditional estimates are not (cf. Ball 2001). Selection bias is small in Dataset I since the posterior probability is close to 1, and large in Datasets II and III where the posterior probability is close to 0.

Example 8.4. APOE gene and Alzheimer's disease.

The APOE gene has three alleles ϵ_2, ϵ_3 and ϵ_4 affecting susceptibility to Alzheimer's disease. Nielsen and Weir (N&W 2001) simulate power for allele-based case–control tests and the TDT to detect associations between two SNP markers (SNP1 and SNP2) located near the APOE gene locus and the disease. Of interest is whether associations between two SNP markers and the disease could be detected by association mapping. Power was reported as around 57% for the allele-based case–control test and 50% for the TDT test with 50 cases and 50 controls, to detect an association with SNP2 (marker 'N' in N&W, Table II), at significance level $\alpha = 0.05$ (N&W, Fig. 1, p. 259).

To make the probability calculations required for simulations, we make the following statistical assumption: that conditional on the APOE genotypes the disease status and marker genotypes are independent. This *conditional independence* assumption is equivalent to the biological assumption that the APOE locus is the only locus affecting the disease that is in linkage disequilibrium with the marker. The conditional independence model is represented as a graphical model in Figure 8.6. This is the same type of model used to represent probabilistic models for Bayesian analysis using BUGS in Section 8.3.4 (cf. Figure 8.13).

Allele frequencies, LD values and disease penetrances from N&W are shown in Tables 8.15, 8.16 and 8.17. Using these values and Bayes' theorem we calculate probabilities for APOE genotypes conditional on case or control status (Equation (8.26) and (8.27)).

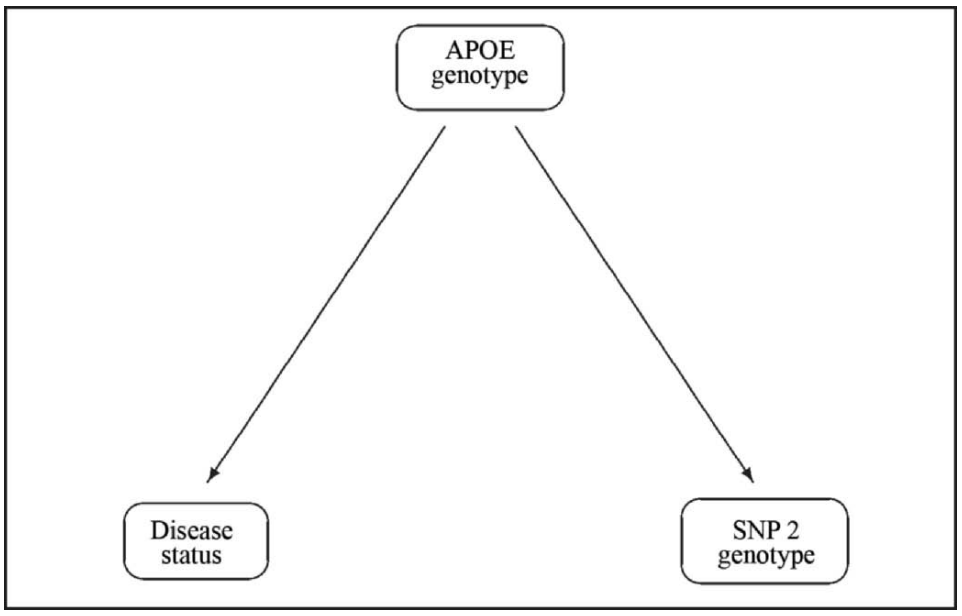


Figure 8.6. Graphical representation of probabilistic model relating APOE genotypes, SNP marker genotypes and Alzheimers' disease status.

Table 8.15. Allele frequencies for SNP2 and APOE.

SNP2		APOE		
N_1	N_2	ϵ_2	ϵ_3	ϵ_4
0.15	0.85	0.085	0.779	0.137

Table 8.16. Disequilibrium and recombination rates between SNP markers and APOE.

	Disequilibria			Recombination
Marker	$D_{\epsilon 2,1}$	$D_{\epsilon 3,1}$	$D_{\epsilon 4,1}$	$c\,(\%)$
M (SNP1)	0.07149	−0.1169	0.04545	0.05
N (SNP2)	0.04545	−0.1169	0.07140	0.5

Table 8.17. Disease penetrances for APOE genotypes.

Genotype (g)	$\epsilon_2\epsilon_2$	$\epsilon_2\epsilon_3$	$\epsilon_2\epsilon_4$	$\epsilon_3\epsilon_3$	$\epsilon_3\epsilon_4$	$\epsilon_4\epsilon_4$
Pr(case $\mid g$)	0.0432	0.0288	0.0576	0.0480	0.130	0.600

$$\begin{aligned} \Pr(\epsilon_k\epsilon_l \mid \text{case}) &= \Pr(\text{case} \mid \epsilon_k\epsilon_l)\Pr(\epsilon_k\epsilon_l)/\Pr(\text{case}), && (8.26)\\ \Pr(\epsilon_k\epsilon_l \mid \text{control}) &= \Pr(\text{control} \mid \epsilon_k\epsilon_l)\Pr(\epsilon_k\epsilon_l)/\Pr(\text{control}). && (8.27) \end{aligned}$$

The probabilities Pr(case $\mid \epsilon_k\epsilon_l$) are the penetrances, and $\Pr(\epsilon_k\epsilon_l) = \Pr(\epsilon_k)\Pr(\epsilon_l)$, where $\Pr(\epsilon_k)$ and $\Pr(\epsilon_l)$ are the allele frequencies. Pr(case) is the disease prevalence (7.3%), which can be obtained by summation over k, l of the numerator of the right-hand side of Equation (8.26). Probabilities and conditional probabilities for the controls in Equation (8.27) are obtained from the corresponding probabilities for cases by subtraction from 1.

Next we calculate probabilities for marker genotypes given the APOE genotypes and disease status

$$\begin{aligned} \Pr(N_iN_j \mid \epsilon_k\epsilon_l, \text{case}) &= \Pr(N_iN_j \mid \epsilon_k\epsilon_l)\\ &= \Pr(N_i \mid \epsilon_k)\Pr(N_j \mid \epsilon_l)\\ &= (\Pr(N_i)+D_{\epsilon_k,i}/\Pr(\epsilon_k))(\Pr(N_j)+D_{\epsilon_l,j}/\Pr(\epsilon_l))\,, && (8.28) \end{aligned}$$

where the first equality follows by conditional independence of the disease and marker genotypes. The disequilibrium coefficients $D_{\epsilon_k,1}$ are given in Table 8.16, and the coefficients $D_{\epsilon_k,2}$ given by $D_{\epsilon_k,2} = -D_{\epsilon_k,1}$ because N is bi-allelic. A derivation is given in Figure 8.7. The same equations apply if disease status is set to control, i.e.

$$\Pr(N_iN_j \mid \epsilon_k\epsilon_l, \text{control}) = \Pr(N_iN_j \mid \epsilon_k\epsilon_l, \text{case})\,. \quad (8.29)$$

The above probabilities were used to simulate populations for assessing the power of allele-based case–control studies to detect associations with Alzheimer's disease and SNP2, following the procedure in Figure 8.8.

$$\Pr(N_2 \mid \epsilon_k) = \Pr(N_2) + D_{\epsilon_k,2}/\Pr(\epsilon_k) \tag{8.30}$$

and

$$\begin{aligned} \Pr(N_2 \mid \epsilon_k) &= 1 - \Pr(N_1 \mid \epsilon_k) \\ &= 1 - (\Pr(N_1) + D_{\epsilon_k,1}/\Pr(\epsilon_k)) \\ &= \Pr(N_2) - D_{\epsilon_k,1}/\Pr(\epsilon_k) \end{aligned} \tag{8.31}$$

Comparing (8.30) and (8.31) we see that $D_{\epsilon_k,2} = -D_{\epsilon_k,1}$.

Figure 8.7. Derivation of $D_{\epsilon_k,2}$ in terms of $D_{\epsilon_k,1}$.

1. Simulate APOE genotypes for cases and controls using the probabilities $\Pr(\epsilon_k\epsilon_l \mid \text{case})$, $\Pr(\epsilon_k\epsilon_l \mid \text{control})$.
2. Simulate marker genotypes using the probabilities $\Pr(N_iN_j \mid \epsilon_k, \epsilon_l, \text{case})$, $\Pr(N_iN_j \mid \epsilon_k, \epsilon_l, \text{control})$.
3. Form the 2×2 table of disease status and marker values.
4. Calculate Bayes factors using Equation (8.22).
5. Estimate power as proportion of Bayes factors greater than the threshold(s) of interest.

Figure 8.8. Simulations for power of case–control studies to detect associations with Alzheimer's disease.

Power to detect the association between marker SNP2 and the disease with various Bayes factors, estimated from 3,000 simulated populations for each sample size, is shown in Table 8.18. We see that a sample of size $n = 50$ cases and $n = 50$ controls with 50% power to obtain a p-value less than 0.05 has 53% power to obtain a Bayes factor of 1 (similar to the power to obtain a p-value of 0.05 from N&W, hence the $p = 0.05$ is approximately equivalent to a Bayes factor $B = 1$ here), but low power to obtain a Bayes factor of 20. A sample size of $n = 200$ is sufficient to obtain a Bayes factor of 20 with 80% power, useful if we already have strong prior information on the location of the gene, while a sample size of $n = 600$, sufficient to obtain a Bayes factor of 1,000,000 with 95% power, would suffice for a genome scan, with prior odds of 1/50,000 per marker.

Summary

Data for single marker tests in case–control studies can be summarised as a contingency table, and associations tested using the χ^2 or Fisher exact tests, or Bayesian methods.

In Example 8.3, the frequentist χ^2 and Fisher exact tests were compared with Bayesian inference for several 2×2 contingency tables. Bayesian inference for Example 8.3 illustrates calculating the Bayes factor by explicit integration, made possible because of the

Table 8.18. Power of case–control test with n cases and n controls to detect the association between marker SNP2 and Alzheimer's disease with given Bayes factors. Power was estimated from 3,000 simulated populations for each sample size. Bayes factors were calculated using Equation (8.22).

Bayes	n				
factor	50	200	400	600	800
1	0.532	0.966	1.000	1.000	1.000
20	0.153	0.809	0.994	1.000	1.000
100	0.063	0.666	0.981	1.000	1.000
1,000	0.016	0.449	0.940	0.998	1.000
1,000,000	0.000	0.062	0.593	0.952	0.997

use of a conjugate Beta prior. As with other examples, there were 'significant' p-values, corresponding to only weak evidence according to the Bayes factor. Again, selection bias in estimated effects occurred in the datasets where posterior probabilities for H_1 were not high.

Example 8.4 illustrates probability calculations for the multiple LD coefficients which occur when there are more than two alleles, and simulations to obtain the power to detect LD between a marker and trait locus with a given Bayes factor. The sample sizes considered by Nielsen and Weir, of 50 cases and 50 controls had about 50% power to detect the marker with $p = 0.05$. A sample size of 200 cases and 200 controls is required for power 80% to detect the association with Bayes factor 20. To detect the associations in a genome scan requires a Bayes factor of around 1,000,000, and a sample size of 600 cases and 600 controls. As with other examples, to reliably detect the associations with the Alzheimer's locus (assuming its position was not already known), in a genome scan would require substantially larger sample sizes than those indicated by traditional frequentist power calculations.

8.3.4 Transmission disequilibrium (TDT) tests

The transmission disequilibrium test (TDT; Spielman *et al.* 1993) tests for an association between transmission of an allele and a trait. The TDT tests for both linkage and linkage disequilibrium, hence eliminating problems with spurious associations due to population structure, between unlinked markers. There may still be problems with spurious associations between markers that are linked but not tightly linked, compared to the resolution of LD.

The TDT test requires many small families, with a single progeny where one parent is heterozygous and the other parent is homozygous for a marker. Figure 8.10 shows sample families from a TDT test with transmission status indicated by $T = 1$ (marker allele A transmitted from the heterozygous parent) or $T = 0$ (A not transmitted).

TDT for discrete traits. The TDT and S-TDT tests were introduced in Chapter 7, Section 7.5.2. We calculate the Bayes factors for the TDT test from Example 7.3, and for the S-TDT test from Example 7.4, using the Savage–Dickey Bayes factor estimate for nested models (Dickey 1971).

We first introduce the Savage–Dickey Bayes factor. Suppose H_0 is a subset of H_1 with $\theta = 0$. For such nested models, the Bayes is given by the Savage–Dickey ratio, the ratio of

prior to posterior densities at 0

$$B = \frac{\pi(\theta = 0)}{g(\theta = 0 \mid y)}, \tag{8.32}$$

where θ denotes the parameter being tested (here a), and $g(\theta = 0 \mid y)$ is the marginal posterior density for θ. If there are additional parameters ψ, common to H_0, H_1, these are integrated over to obtain the marginal posterior $g(\theta = 0 \mid y)$.

Example 8.5. Bayes factor calculation for the TDT.

Recall n_{12} and n_{21} were the numbers of times allele 1 but not allele 2 was transmitted, and the number of times allele 2 but not allele 1 was transmitted (Table 7.9).

We condition on $n_{12} + n_{21}$, the number of times exactly one allele was transmitted. Under the null hypothesis, alleles 1 and 2 are equally likely to be transmitted, so n_{12} has a binomial distribution with $n = n_{12} + n_{21}$, and $p = 0.5$. Under the alternative hypothesis n_{12} has a binomial distribution with $p = p_1$. We use a non-informative Beta$(0.5, 0.5)$ prior for p_1.

$$n_{12} \sim \text{Binomial}(n_{12} + n_{21}, 0.5) \qquad \text{under } H_0\,, \tag{8.33}$$

$$n_{12} \sim \text{Binomial}(n_{12} + n_{21}, p_1) \qquad \text{under } H_1\,. \tag{8.34}$$

Recall that the Beta distribution is the conjugate prior for binomial sampling – if the prior is Beta(a, b), and k successes are observed in n Bernoulli trials the posterior is Beta$(a + k, b + n - k)$. Hence, under H_1 the posterior for p_1 under H_1 is Beta$(n_{12} + 0.5, n_{21} + 0.5)$. In Example 7.3, we have $n_{12} = 39$ and $n_{21} = 86$, so the posterior is Beta$(39.5, 86.5)$. The prior and posterior densities for p_1 under H_1 are

$$\pi(p) = \frac{1}{B(0.5, 0.5)} p^{0.5}(1-p)^{0.5}, \tag{8.35}$$

$$g(p \mid n_{12}, n_{21}) = \frac{1}{B(0.5, 0.5)} p^{39.5}(1-p)^{86.5}, \tag{8.36}$$

where $B(a, b)$ is the beta function (cf. Example 8.3).

H_0 and H_1 are nested models with H_0 corresponding to $p_1 = 0.5$. Therefore the Bayes factor is given by the Savage–Dickey density ratio:

$$B = \frac{\pi(p_1 = 0.5)}{g(p_1 = 0.5 \mid n_{12}, n_{21})} = \frac{0.5}{B(0.5, 0.5)} \times \frac{B(39.5, 86.5)}{0.5^{126}} = 610.8. \tag{8.37}$$

Recall the p-value from Chapter 7, Example 7.3 was 2.6×10^{-5}.

The density-ratio calculation is conveniently done using the R function `dbeta()` (Figure 8.9).

□

```
> # Savage-Dickey Bayes factor calculation for the TDT test.
> dbeta(0.5,0.5,0.5)/dbeta(0.5,86.5,39.5)
[1] 610.809
```

Figure 8.9. R calculation of the Savage–Dickey density ratio for the Bayes factor.

Example 8.6. Bayes factor calculation for the S-TDT.

Similar to the previous example, the Bayes factor calculation uses the Savage–Dickey density ratio. Whereas, the TDT was based on a proportion, the S-TDT is based on a normalised test statistic based on averages for affected sibs and all sibs within a family.

To apply the Savage–Dickey density ratio we rescale the test statistic, as a function of the number, n of families in the test, to a statistic Z_n which is estimating the same quantity for all n. We then use a nearly non-informative prior for the test statistic, equivalent to the sampling distribution of Z_1 under H_0.

Write the test statistic as

$$T_n = \frac{(Y_n - A_n)/n}{\sqrt{V_n}} \sim N(0, 1/n), \tag{8.38}$$

where n is the number of families in the test.

Recall that

$$V_n = \sum_{i=1}^{n} V_i, \tag{8.39}$$

where

$$V_i = \frac{a_i u_i [4r_i(t_i - r_i - s_i) + s_i(t_i - s_i)]}{t_i^2(t_i - 1)}\,. \tag{8.40}$$

Now let Z_n be given by

$$\begin{aligned} Z_n &= T_n/\sqrt{n} && (8.41)\\ &= \frac{(Y_n - A_n)/n}{\sqrt{\overline{V}_n}} \sim N(0, 1/n) \quad \text{under } H_0\,, && (8.42) \end{aligned}$$

where $\overline{V}_n = V_n/n$. Notice that the quantities in the numerator and denominator for Z_n are stable, i.e. estimating the same quantity, independent of n.

Under H_1 the sampling variance for Z_n is $1/n$, and its estimate is the value of the statistic. We take a prior for Z, the quantity that Z_n is estimating under H_1 to be the same as the sampling distribution for Z_1, i.e. $N(0, 1)$. By construction, this is a nearly non-informative prior with the same information as a single experimental unit. The posterior distribution for Z under H_1 is then given by

$$z \mid Z_{\text{obs}} \quad \sim \quad N\left(\frac{n}{n+1} Z_{\text{obs}}, \frac{1}{n+1}\right). \tag{8.43}$$

The prior and posterior densities are

$$\pi(z) = \phi(z) = \frac{1}{\sqrt{2\pi}} \exp(-z^2/2), \tag{8.44}$$

$$g(z \mid Z_{\rm obs}) = \phi\left(\frac{n}{\sqrt{n+1}} Z_{\rm obs}\right) \sqrt{n+1}. \tag{8.45}$$

From Example 7.4 we have $T = 1.90$, and $n = 3$, so $Z_{\rm obs} = T/\sqrt{3} = 1.10$. The Savage–Dickey density Bayes factor estimate is

$$B = \frac{\pi(z=0)}{g(z=0 \mid Z_{\rm obs})} = \frac{\frac{1}{\sqrt{2\pi}}}{\frac{1}{\sqrt{2\pi}} \exp(-\frac{n^2}{n+1} Z_{\rm obs}^2/2)\sqrt{n+1}} \tag{8.46}$$

$$= \frac{\exp(\frac{n^2}{n+1} Z_{\rm obs}^2/2)}{\sqrt{n+1}} = 1.94\,, \tag{8.47}$$

representing very weak evidence for H_1. Recall that the p-value was 0.057.

TDT for quantitative traits. Allison (1997) considers five variants of TDT tests for quantitative traits, called TDT-Q1-Q5. TDT-Q1 assumes random sampling. TDT-Q3,Q4 assume 'extreme sampling', a form of selective genotyping where trios are pre-selected so the offspring lie in the tails of the phenotypic distribution. TDT-Q1–Q4 use only families with one heterozygous parent, while TDT-Q5 attempts to use information from all possible matings.

Why it works. Consider the alleles transmitted from the heterozygous parent. If a QTL is linked to the marker, and the QTL allele Q is associated with the marker allele A in the population, then Q will be transmitted along with A more (or less) often than not, generating an association. If the QTL alleles are Q, q, and the marker alleles are A, a, the probability the QTL allele is Q, conditional on transmission status ($T = 0$ or $T = 1$, Figure 8.10) if the marker allele is A is:

$$\Pr(Q \mid T=1) = \Pr(Q \mid A)(1-r) + \Pr(Q \mid a)r\,, \tag{8.48}$$

$$\Pr(Q \mid T=0) = \Pr(Q \mid A)r + \Pr(Q \mid a)(1-r)\,. \tag{8.49}$$

If Q and A are not linked but are 'spuriously' associated due to allele frequency differences between sub-populations, the recombination process will ensure the transmission of Q and

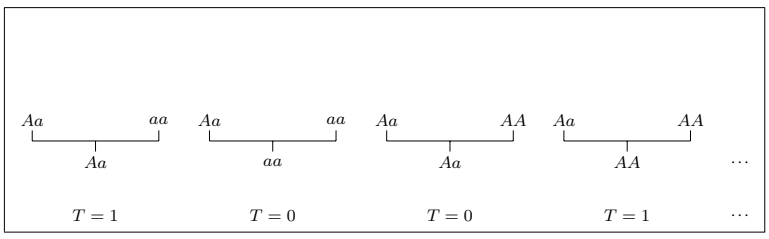


Figure 8.10. Sample TDT test families.

A from the heterozygous parent are independent, hence there will be no association: ($r = 0.5$ in Equations (8.48), (8.49) and hence $\Pr(Q \mid T = 1) = \Pr(Q \mid T = 0) = \Pr(Q)$, i.e. T and Q are independent). This eliminates completely spurious associations due to population structure; however some partially spurious associations between linked loci may remain. These are associations where the recombination distance between marker and QTL is less than 0.5 but still large compared to the resolution of the association mapping experiment. Partially spurious associations, where r is less than 0.5 in Equation (8.48), will reduce by a factor $(1 - 2r)$ in magnitude, so could still be substantial for small to moderate values of r, e.g. with $r = 0.1, 0.2$, the association is reduced by only 20 or 40%, respectively. These 'small' values of r correspond to genomic intervals which are nevertheless large compared to the resolution of the association mapping experiment.

The QTL allele transmitted from the homozygous parent will be random, reflecting population allele frequencies, whether or not A is transmitted from the heterozygous parent, hence will not contribute to the expected trait value conditional on transmission of A, but will contribute to variability.

Note: In practice a number of markers will be tested. The heterozygosity condition can be obtained for each marker by selecting a subset of families where the condition applies. For families with more than one offspring, only one progeny can be selected at random.

Frequentist analysis

The TDT-Q1 is analysed with a standard t-test, the TDT-Q2 with a χ^2-test (Allison 1997, p. 678), and the TDT-Q3, with a modified t-test (Allison 1997, p. 679). The χ^2-test tests for independence between transmission status and phenotype class (L for $y < Z_{\mathrm{L}}$ and U for $y > Z_{\mathrm{U}}$ where Z_{L} and Z_{U} are the thresholds used for selective genotyping).

Sample power calculations from Allison (1997) are shown in (Table 8.19). These are based on a comparison-wise α value of 0.0001, noting the need

> ... to maintain a genome-wide α at the desired level ... further development in this regard may be needed.

Note also that these values are based on the assumption that the marker locus is the trait locus, so these values are upper bounds to the power.

The sample sizes shown are the number of families genotyped. Not surprisingly, the TDT-Q2–Q5 designs using selective genotyping have more power per family genotyped. Which design is more efficient for the end user depends on the relative costs of obtaining and phenotyping the families. Additionally if multiple traits are being considered, the advantage of selective genotyping reduces, as different subsets are needed for each trait. Not surprisingly, the t-test is more powerful than the χ^2-test (TDT-Q2 versus TDT-Q3), since the t-test requires more assumptions, while the phenotypic data are grouped into categories for the χ^2-test.

Bayesian analysis

A full Bayesian model can be fitted to the data. Alternatively, where the t-test analysis has been done, equivalent Bayes factors can be calculated using Equation (8.8), noting that $F = t^2$, has an F distribution on 1, $n - 1$ d.f.

Table 8.19. Sample sizes required for 80% power of the TDT–Q1-Q5 tests from Allison (1997), for an additive QTL, assuming a type I error rate of $\alpha = 0.0001$, with QTL explaining 5, 10% of the phenotypic variance, and allele frequencies $p = 0.1, 0.3, 0.5$.

	TDT-Q1	TDT-Q2	TDT-Q3	TDT-Q4	TDT-Q5
$h^2_{\mathrm{Q}} = 0.05$					
$p = 0.1$	308	120	93	94	115
$p = 0.3$	727	247	214	224	212
$p = 0.5$	873	294	259	272	237
$h^2_{\mathrm{Q}} = 0.10$					
$p = 0.1$	147	68	43	41	60
$p = 0.3$	351	125	98	102	105
$p = 0.5$	426	149	121	127	118

Reprinted from Allison (1999), American Journal of Human Genetics 60:676–690.

Table 8.20. Equivalent Bayes factors for the TDT-Q1 tests from Table 8.19.

	$h^2_{\mathrm{Q}} = 0.05$			$h^2_{\mathrm{Q}} = 0.1$		
	n	α	B	n	α	B
$p = 0.1$	308	0.0001	192	147	0.0001	296
$p = 0.3$	727	0.0001	120	351	0.0001	146
$p = 0.5$	873	0.0001	109	426	0.0001	129

Equivalent Bayes factors for the TDT-Q1 with α as in Table 8.19 are shown in Table 8.20.

The choice of $\alpha = 0.0001$ has given some more respectable Bayes factors, but these vary from 109 to 296, nearly a three-fold range depending on QTL heritability and allele frequencies. These Bayes factors are still too low to use for genome scans or candidate genes without strong prior information (cf. Tables 8.6, 8.7, Datasets 1, 2).

We next illustrate the implementation and analysis of the full Bayesian model including an MCMC sampler using the BUGS language (Spiegelhalter *et al.* 1995), with assumptions as in TDT-Q1 for simulated data.

Example 8.7. Bayesian analysis of simulated TDT-Q1 data.

Data simulation. R functions for simulating TDT data and calculating Bayes factors are shown in Figure 8.11. Calculations for three replicate simulations with additive QTL explaining 5% of the variation and sample size N = 873 trios, and allele frequencies p = 0.5 (corresponding to power 0.8 in Table 8.19) are shown in Figure 8.12. Data from the first simulation is also analysed using a full Bayesian model below.

Assume families are sampled at random, and allele frequencies for the marker A, a are $p, (1 - p)$, with genotype frequencies, assuming Hardy–Weinberg equilibrium,[2] of AA, Aa, aa of $p^2, 2p(1-p), (1-p)^2$, and with genotype-expected values of $\mu_{AA} = \mu + 2a$,

[2]This assumption is not necessary, but we would otherwise need estimates of genotypic frequencies.

```
library(ldDesign)
sim.TDT <- function(h2q,N,p=0.5,mu=0,phi=0,Vp=1,Vq=h2q*Vp,
                    Ve=Vp-Vq){
# simulate data from a TDT for biallelic marker at trait locus
# Cf Allison 1997
# h2q:  QTL variance as a proportion of total
# N: number of family trios
# p: allele frequency for the allele 'A', (1-p) for allele 'a'
# mu: population mean
# phi: dominance proportion d=phi*a
# phi=0 for additive, phi=1 for complete dominance
# Vp: total phenotypic variation
# Vq: QTL variance
# Ve: residual variance
# set initial values for a and d
# calculate QTL variance and scale to give required variance
a0 <- sqrt(2*Vq)
d0 <- phi*a0
muq0 <- p^2*2*a0 + 2*p*(1-p)*(a0+d0)
Vq0 <- p^2*(2*a0 - muq0)^2 + 2*p*(1-p)*(a0+d0 - muq0)^2 +
       (1-p)^2*(0 - muq0)^2
sqrt.ratio <- sqrt(Vq/Vq0)
a <- a0*sqrt.ratio
d <- d0*sqrt.ratio
family.type.levels <- c("Aa x aa", "Aa x AA")
genotype.levels <- c("aa","Aa","AA")
genotype.means <- c(mu,mu+a+d,mu+2*a)
family.type <- sample(size=N, c(1,2),prob=c((1-p)^2,p^2),
                      replace=TRUE)
transmissions <- rbinom(n=N,size=1,prob=0.5)
progeny.genotypes <- ifelse(family.type=="Aa x aa", ifelse(
    transmissions==1,"Aa","aa"), ifelse(transmissions==1,"AA","Aa"))
progeny.phenotypes <- genotype.means[match(progeny.genotypes,
                          genotype.levels)] +  sqrt(Ve)*rnorm(N)
list(progeny.phenotypes=progeny.phenotypes,
     transmissions=transmissions,
     progeny.genotypes=progeny.genotypes)
}
calc.bf.TDT <- function(data){
# data: dataframe as generated by sim.TDT()
summ1 <- summary(aov(progeny.phenotypes ~ transmissions,
                     data=data))
ns <- table(data$transmissions)
N <- sum(ns)
F.value <- summ1[[1]]$"F value"[1]
list(N=N,ns=ns,F.value=F.value,B=SS.oneway.bf(group.sizes=ns,
     Fstat=F.value))
}
```

Figure 8.11. R functions for simulating TDT data, and calculating the Spiegelhalter and Smith Bayes factors using the R function `SS.oneway.bf()` from `ldDesign`.

```
> sim.df2a <- sim.TDT(h2q=0.05,p=0.5,N=873)
> sim.df2b <- sim.TDT(h2q=0.05,p=0.5,N=873)
> sim.df2c <- sim.TDT(h2q=0.05,p=0.5,N=873)
> # columns are
> # N:       total sample size,
> # n0, n1:  number of non-transmissions, transmissions
> # F.value: value of F statistic from ANOVA
> # B:       corresponding Bayes factor
> rbind(unlist(calc.bf.TDT(sim.df2a)),
+       unlist(calc.bf.TDT(sim.df2b)),
+       unlist(calc.bf.TDT(sim.df2c)))
       N   n0   n1 F.value   B
[1,] 873  451  422    14.9  93
[2,] 873  409  464    16.0 154
[3,] 873  443  430    18.6 555
```

Figure 8.12. R calculations for three simulated datasets.

$\mu_{Aa} = \mu + a + d$, $\mu_{aa} = \mu$. Let C_i denote the event that a family meets the selection criteria, $T_i = 1$, $T_i = 0$ denote the event that the allele A is transmitted (resp., not transmitted), from the heterozygous parent, and y_i denote phenotype of the ith offspring.

Step 1. Write down the model. The key is to note that the TDT ignores the family genotypes, and looks at transmission only, and that transmission occurs with probability 0.5 and is independent of family type. Conditional on the heterozygous parent being Aa, the legal family types $Aa \times aa$ and $Aa \times AA$ occur with probability $(1-p)^2$, p^2, respectively.

The likelihood is

$$f(y \mid \mu, a, d, \sigma_{\mathrm{e}}^2) = \prod_{i=1}^{n} f(y_i \mid T_i), \tag{8.50}$$

where

$$f(y_i \mid T_i = 1) \sim N(\mu_1, \sigma_1^2), \tag{8.51}$$
$$f(y_i \mid T_i = 0) \sim N(\mu_0, \sigma_0^2), \tag{8.52}$$

where

$$\mu_0 = \frac{(1-p)^2\mu_{aa} + p^2\mu_{Aa}}{p^2 + (1-p)^2}, \tag{8.53}$$
$$\mu_1 = \frac{(1-p)^2\mu_{Aa} + p^2\mu_{AA}}{p^2 + (1-p)^2}, \tag{8.54}$$
$$\sigma_0^2 = \sigma_{\mathrm{e}}^2 + \frac{(1-p)^2(\mu_{aa} - \mu_0)^2 + p^2(\mu_{Aa} - \mu_0)^2}{p^2 + (1-p)^2}, \tag{8.55}$$
$$\sigma_1^2 = \sigma_{\mathrm{e}}^2 + \frac{(1-p)^2(\mu_{Aa} - \mu_1)^2 + p^2(\mu_{AA} - \mu_1)^2}{p^2 + (1-p)^2}. \tag{8.56}$$

Note: Equations for $\mu_0, \mu_1, \sigma_0^2, \sigma_1^2$ are given in Allison (1997, p. 677). Our equations may differ due to differing conventions and a possible error in the equation for μ_{Y_1} (here denoted by μ_1), there.

Step 2. Represent the hierarchical model as a graphical model. The graphical model is shown in Figure 8.13. Points to note are:

1. The graph is a *directed graph*, with the convention that the arcs are directed downwards, i.e. parent nodes are located above their descendants. The probability distribution of the variable(s) at a node needs to be given as a function of its parent nodes. Parameter values, for two nodes which are not direct descendants of each other, are *conditionally independent*, given the values of their common 'ancestors' in the graph.

2. Parameters specific to individual observations (here the transmission status and phenotypic value for the offspring of a family trio) are located within the lower box. The nested boxes signify multiple pages, with one page for each datum.

3. To simplify the diagram, parameters $\mu_{AA}, \mu_{Aa}, \mu_{aa}$ have been grouped together in a single node, as have parameters $\mu_0, \mu_1, \sigma_0^2, \sigma_1^2$.

4. BUGS uses precisions (reciprocal of variances) as parameters instead of variances.

5. Parameters η_i, τ_i are the means and precisions for the ith observation, given by

$$\eta_i = \begin{cases} \mu_0 & \text{if } T_i = 0 \\ \mu_1 & \text{if } T_i = 1 \end{cases} \quad \text{and} \quad \tau_i = \begin{cases} 1/\sigma_0^2 & \text{if } T_i = 0 \\ 1/\sigma_1^2 & \text{if } T_i = 1 \end{cases} . \tag{8.57}$$

Step 3. Write down the distributions of nodes in the graph in terms of their 'parents'. For example the distribution of y_i conditional on its parents is normal with mean η_i and precision τ_i.

Top level nodes, i.e. those with no parents are given prior distributions, obviously not involving any other variables.

Step 4. Implement the Gibbs sampler in BUGS code. BUGS code is shown in Figure 8.14. We do not describe the BUGS language in detail, only essential aspects of our code, referring the reader to the BUGS manual for further information.

The BUGS code consists of initial declarations of variables and constants, specifying the data file, and optional file of initial values, followed by the main body of the program, where the distribution of each variable in the graphical model is specified in terms of the values of its parents. Distributions are specified viz

```
tau.e ~ dgamma(1.0,1.0)I(0.7,1.3);
```

meaning that `tau.e` has a gamma distribution with shape and rate parameters 1, 1, respectively. The optional `I(0.7,1.3)` notation restricts the distribution to the interval `(0.7,1.3)`, allowing BUGS to use Metropolis sampling. This was required for `tau.e` because BUGS could not otherwise choose an update method.

Our default priors for μ and a were normal with mean 0 and precision 1, i.e. similar to the precision of the phenotypic distribution.[3] However, even with Metropolis sampling,

[3]With actual data we may have more informative priors.

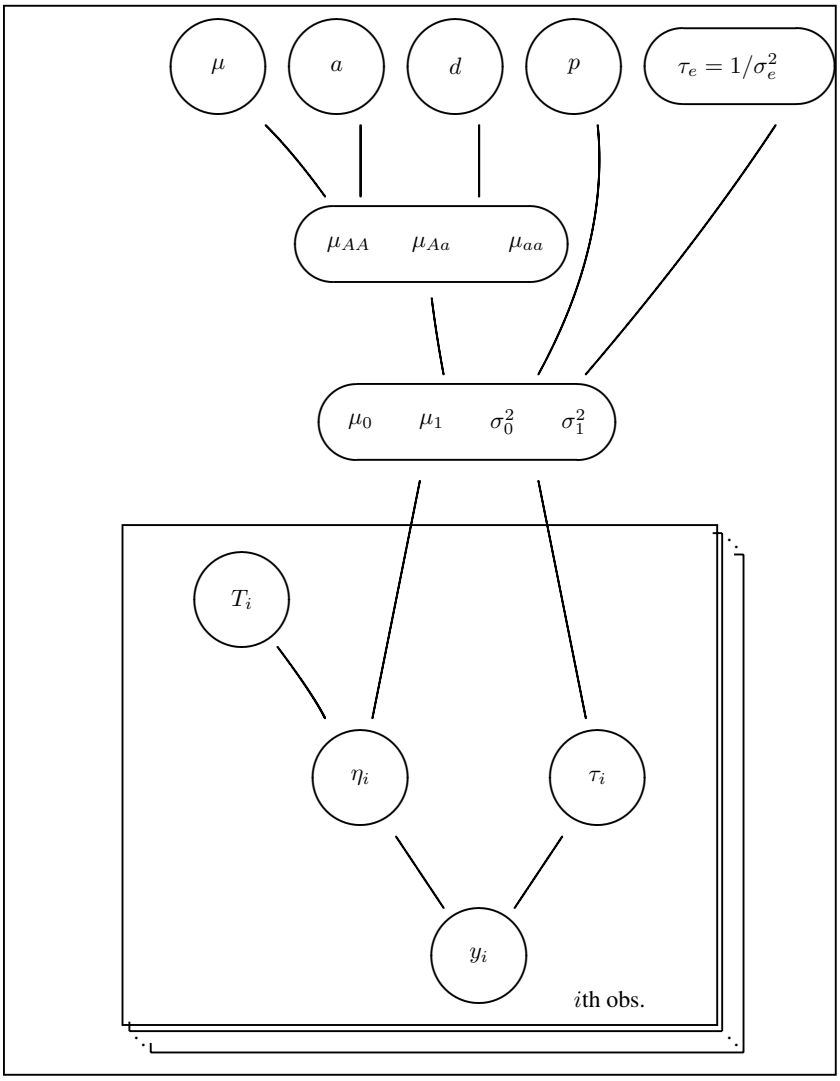


Figure 8.13. Graphical representation of a Bayesian hierarchical model for a TDT model (TDT-Q1) for quantitative traits.

```
model tdtq1; const
  N=873;
var
  mu,a,d,p,tau.e,mu.AA,mu.Aa,mu.aa,mu0,mu1,sigma2.0,sigma2.1,
  sigma2.e,mu.pop,sigma2.q,h2.q,eta[N],tau[N],y[N],T[N],
  mu.values[21],imu,pmu[21],a.values[21],ia,pa[21];
data in "tdtq1n873.dat"; inits in "tdtq1.ini"; {
  # Metropolis sampling for tau.e, indicated by I(0.7,1.3)
  tau.e ~ dgamma(1.0,1.0)I(0.7,1.3);
  # unable to choose update for  mu, a
  # so use categorical priors,
  # with normal prior, even with Metropolis sampling
  # values in the data file
  #  mu ~ dnorm(0.0, 1.0)I(-0.5,0.5);
  imu ~ dcat(pmu[]);  mu <- mu.values[imu];
  #  a ~ dnorm(0.0,1.0);
  ia ~ dcat(pa[]);  a <- a.values[ia];
  # set d to 0 for an additive model, or use a prior similar to a
  d <- 0.0;  p <- 0.5;
  sigma2.e <- 1/tau.e;
  mu.AA <- mu + 2*a;  mu.Aa <- mu + a + d;  mu.aa <- mu;
  mu.pop <- mu.AA*p*p + mu.Aa*2*p*(1-p) + mu.aa *(1-p)*(1-p);
  sigma2.q <- p*p*(mu.AA - mu.pop)*(mu.AA - mu.pop) +
              2*p*(1-p)*(mu.Aa - mu.pop)*(mu.Aa - mu.pop) +
              (1-p)*(1-p)*(mu.aa - mu.pop)*(mu.aa - mu.pop);
  h2.q <- sigma2.q/(sigma2.q + sigma2.e);
  mu0 <- ((1-p)*(1-p)*mu.aa + p*p*mu.Aa)/(p*p + (1-p)*(1-p));
  mu1 <- ((1-p)*(1-p)*mu.Aa + p*p*mu.AA)/(p*p + (1-p)*(1-p));
  sigma2.0 <- sigma2.e + ((1-p)*(1-p)*(mu.aa - mu0)*(mu.aa - mu0) +
              p*p*(mu.Aa - mu0)*(mu.Aa - mu0))/(p*p + (1-p)*(1-p));
  sigma2.1 <- sigma2.e + ((1-p)*(1-p)*(mu.Aa - mu1)*(mu.Aa - mu1) +
              p*p*(mu.AA - mu1)*(mu.AA - mu1))/(p*p + (1-p)*(1-p));
  for(ii in 1:N){
    eta[ii] <- mu0*step(0.5-T[ii]) + mu1*step(T[ii]-0.5);
    tau[ii] <- 1.0/(sigma2.0*step(0.5-T[ii]) +
               sigma2.1*step(T[ii]-0.5));
    y[ii] ~ dnorm(eta[ii], tau[ii]);
  }
}
```

Figure 8.14. BUGS code for the TDT-Q1.

BUGS could not choose an updating method for μ and a with these priors so these variables were discretised into 21 steps and, for convenience, given a uniform prior on the discrete values. In order to make best use of all 21 steps in the discretisations, the discretisations were adapted so as to extend slightly beyond the ranges of the posterior distributions from an initial run. The discretisation for μ is implemented with the BUGS parameters `imu` (categorical index), and `pmu` (probabilities for each index value) and `mu.values` (corresponding values for μ), and similarly for a, with parameters `ia, pa` and `a.values`, with values defined in the data file.

The intermediate parameters $\mu_{AA}, \mu_{Aa}, \mu_{aa}, \mu_0, \mu_1, \sigma_0^2, \sigma_1^2, \eta_i, \tau_i$ (present mainly for convenience and readability), assigned with '<−', are *deterministic nodes*, since they are completely determined by their parents.

Step 5. Run the sampler, and examine the output.

The Gibbs sampler was run for 60,000 iterations. Each iteration took approximately 1 s on a Linux machine (Kernel 2.4, with a 2.4 GHz Pentium processor). The output was examined using `coda` (Spiegelhalter *et al.* 1995; Plummer *et al.* 2005), an R (or Splus) package for BUGS output diagnosis and analysis.

Graphs of BUGS output, for parameters μ, a, τ_e, and derived parameters h_{Q}^2 are shown in Figure 8.16.

The left-hand column of figures shows the trace of sampler estimates for iterations 1,001–60,000. The solid traces indicate frequent visits of the sampler to high and low values of the variables indicating good mixing. The right-hand column of figures shows density estimates for the marginal posterior distribution of each parameter. The small bumps in the density estimate for μ and a are an artefact of the discretisations of these variables. Graphs were produced using `coda`. The density estimates were obtained by `coda` using kernel smoothing. A number of diagnostics are provided in `coda`. The Raftery and Lewis diagnostics (Raftery and Lewis 1992, 1995) are shown in Figure 8.15, calculated using the R function `raftery.diag()` in `coda`.

A run of at least 3,746 is indicated if it is desired to estimate the $q = 2.5\%$ quantiles of the distributions with an accuracy of $r = \pm\, 0.5\%$, with probability 0.95. A run of at least 1,377 is needed to estimate the 50% quantiles with an accuracy of $\pm\, 5\%$, with probability 0.95.

Note: In Figures 8.15 and 8.16 only the variables $\mu, a, \tau_e, h_{\mathrm{Q}}^2$ are shown. In general, it is important to examine all variables for convergence. The mean parameter μ is not of particular interest here, however in our experience problems with convergence are often apparent from values of μ, since μ enters the likelihood for every observation, particularly if prior distributions or initial values are poorly specified.

Note: No diagnostic can guarantee convergence of a MCMC sampler. Apparent convergence can persist for a large number of iterations in pathological cases or complex problems. MCMC samplers are best used by statisticians with a good intuitive grasp of the models and parameterisations being used. In most cases convergence problems can be overcome by various techniques, e.g. re-parameterising or block updates, which are beyond our scope. Under general conditions the Gibbs sampler can be shown to converge geometrically (see, e.g. Tierney 1994), and some authors recommend formally proving convergence for each sampler. However, the geometric convergence can still be extremely slow, and most well-constructed MCMC samplers converge orders of magnitude faster than

```
> raftery.diag(bugs1[,c("mu","a","tau.e","h2.q")])

Quantile (q) = 0.025
Accuracy (r) = +/- 0.005
Probability (s) = 0.95

You need a sample size of at least 3746 with these values of q, r,
and s
> raftery.diag(bugs1[,c("mu","a","tau.e","h2.q")],q=0.5,r=0.05,
               s=0.95)

Quantile (q) = 0.5
Accuracy (r) = +/- 0.05
Probability (s) = 0.95

       Burn-in  Total Lower bound  Dependence
       (M)      (N)   (Nmin)       factor (I)
 mu    11       1148  385          2.98
 a     11       1377  385          3.58
 tau.e 2        381   385          0.99
 h2.q  11       1363  385          3.54
```

Figure 8.15. Raftery and Lewis diagnostics for the TDT-Q1 BUGS output from and initial run of 1,000 iterations.

the theoretical bounds. Where these techniques fail is well beyond the realm of traditional asymptotic statistical methods.

Summary statistics for the marginal posterior distributions of parameters are shown in Figure 8.17. The `sem` column shows standard errors of the estimated posterior means, calculated naively based on variance of the sampler output. These may overstate the precision because successive samples are auto-correlated. The effects of auto-correlation can be reduced by calculating standard errors based on batch means, where a batch is a group of successive iterations (Roberts 1996). The standard errors, re-calculated based on batch means for batches of size 100, are calculated using the `batchSE()` function in Figure 8.17.

Finally, from the MCMC output we estimate the Bayes factor for comparing the models H_1 (with a) and H_0 (with $a = 0$), as a sub-model. Recall that the Savage–Dickey Bayes factor estimate (Dickey 1971) is given by the ratio of prior to posterior densities at 0

$$B = \frac{\pi(\theta = 0)}{f(\theta = 0 \mid y)}, \tag{8.58}$$

where θ denotes the parameter being tested (here a), and $f(\theta = 0 \mid y)$ is the marginal posterior density for θ.

The density $f(\theta = 0 \mid y)$ can be estimated from BUGS output since

$$f(\theta = 0 \mid y) \approx \Pr(0 \leq \theta \leq \epsilon)/\epsilon, \tag{8.59}$$

by definition of $f(\cdot)$ as a probability density. The choice of ϵ should be small enough to give

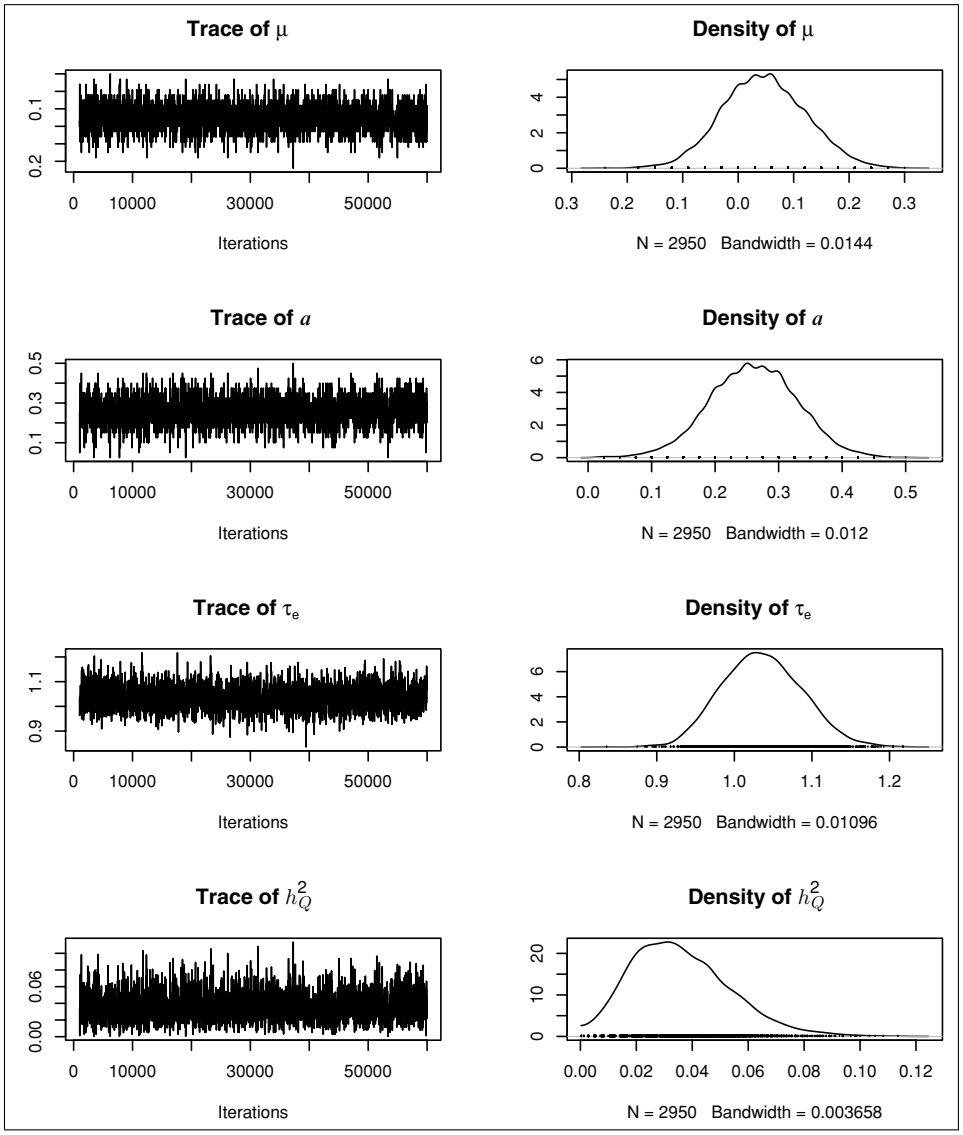


Figure 8.16. BUGS output for simulated TDT-Q1 data. Data were simulated with $N = 873$ families with one parent heterozygous for a bi-allelic marker locus coincident with an additive QTL locus with effect $a = 0.31$ corresponding to 5% of the variation, and allele frequency 0.5. This sample size had power 0.8 to detect an effect with $\alpha = 0.0001$ in Table 8.19, and $B = 109$ in Table 8.20. The sampler was run for 60,000 iterations. The first 1,000 iterations were removed as 'burn in' and, to reduce the amount of data for plotting, only every 20th iteration is plotted.

```
> stats <- function (x, na.rm = T, quants = c(0.025, 0.25, 0.5,
+                    0.75, 0.975)){
+    if (na.rm)  x <- x[!is.na(x)]
+    if (length(x > 0)) {
+        c(mean = mean(x), stdev = stdev(x), sem = sem(x),
+          quantile(x, quants))
+    }else NA
+ }
# 1,000 iterations
>  t(apply(as.matrix(run1.1k[,c("mu","a","tau.e","h2.q")]),2,
+          stats))
        mean  stdev      sem     2.5%    25%    50%    75%  97.5%
mu    0.0516 0.0760 0.002404 -0.09000 0.0000 0.0600 0.1200 0.1807
a     0.2571 0.0703 0.002222  0.12500 0.2000 0.2500 0.3000 0.4000
tau.e 1.0356 0.0525 0.001661  0.93952 0.9993 1.0347 1.0730 1.1377
h2.q  0.0353 0.0179 0.000567  0.00756 0.0207 0.0325 0.0461 0.0755

> batchSE(run1.1k[,c("mu","a","tau.e","h2.q")])
     mu       a   tau.e    h2.q
0.00579 0.00552 0.00140 0.00135

# 10,000 iterations
> t(apply(as.matrix(run1.10k[,c("mu","a","tau.e","h2.q")]),2,
+         stats))
        mean  stdev      sem     2.5%    25%    50%    75%  97.5%
mu    0.0465 0.0756 0.000756 -0.09000 0.0000 0.0600 0.0900 0.1800
a     0.2611 0.0686 0.000686  0.12500 0.2250 0.2500 0.3000 0.4000
tau.e 1.0357 0.0518 0.000518  0.93659 0.9996 1.0341 1.0705 1.1383
h2.q  0.0362 0.0177 0.000177  0.00784 0.0239 0.0338 0.0469 0.0767

> batchSE(run1.10k[,c("mu","a","tau.e","h2.q")])
      mu        a    tau.e     h2.q
0.002048 0.001883 0.000583 0.000480

# 60,000 iterations
> t(apply(as.matrix(run1.60k[,c("mu","a","tau.e","h2.q")]),2,
+         stats))
        mean  stdev      sem     2.5%    25%    50%    75%  97.5%
mu    0.0473 0.0743 3.06e-04 -0.09000 0.0000 0.0600 0.0900 0.1800
a     0.2604 0.0676 2.78e-04  0.12500 0.2250 0.2500 0.3000 0.4000
tau.e 1.0360 0.0513 2.11e-04  0.93832 1.0007 1.0347 1.0699 1.1397
h2.q  0.0360 0.0175 7.19e-05  0.00792 0.0239 0.0333 0.0462 0.0758

> batchSE(run1.60k[,c("mu","a","tau.e","h2.q")])
      mu        a    tau.e     h2.q
0.000860 0.000778 0.000247 0.000199
```

Figure 8.17. Calculation of summary statistics for parameters from TDT-Q1 BUGS output.

a good approximation to $f(\theta = 0 \mid y)$ in Equation (8.59) but large enough so that there are a reasonable number of posterior samples less than ϵ with which to estimate the probability $\Pr(0 \leq \theta \leq \epsilon)$ on the right-hand side of the equation. If necessary more iterations of the sampler can be run to enable this. The calculation in R, giving a Bayes factor of $B = 89.7$, is shown in Figure 8.18.

□

```
> # 10,000 iterations
> tbl <- table(run1.10k[,"a"] < 0.0625)
> tbl
FALSE  TRUE
 8986    14
> # marginal probability for -0.0125 <= a <= 0.0625
> mpprob0 <- (tbl/sum(tbl))[2]
> mpprob0
   TRUE
0.00156
> prior.prob0 <- 0.0750/0.525
> prior.prob0
[1] 0.143
> # ratio of prior to posterior densities, B=91.8
> # B=91.8
> prior.prob0/mpprob0
TRUE
91.8
> # 60,000 iterations
> tbl <- table(run1.100k[,"a"] < 0.0625)
> tbl
FALSE  TRUE
58906    94
> mpprob0 <- (tbl/sum(tbl))[2]
> mpprob0
   TRUE
0.00159
> prior.prob0 <- 0.0750/0.525
> prior.prob0
[1] 0.143
> # B=89.7
> prior.prob0/mpprob0
TRUE
89.7
```

Figure 8.18. Savage–Dickey Bayes factor estimation for the TDT-Q1.

Summary

This section considers the TDT test for continuous data.

Bayes factors were estimated from the F-values ($F = t^2$) from the t-test for an effect of transmission, using the Spiegelhalter and Smith method.

Bayes factors corresponding to $\alpha = 0.0001$, for the designs considered by Allison (1997) varied from 109 to 296. These were much more respectable Bayes factors than those corresponding to $\alpha = 0.05$, but not large enough to give high posterior probabilities for markers from a genome scan.

Example 8.7 illustrates a full Bayesian analysis for the TDT test for simulated data with $n = 873$ trios.

A Bayesian hierarchical model was fitted using Gibbs sampling – a MCMC method, which generates a sample approximately from the posterior distribution. A Bayesian graphical model was constructed (Figure 8.13). Distributions for parameters in terms of their parents were coded in BUGS language (Figure 8.14).

Posterior estimates, of interest, e.g. posterior means and standard deviations, are easily obtained from the Gibbs sampler output. Marginal distributions for a set of one or more parameters are obtained by simply ignoring the other parameters.

Diagnostics and posterior summary statistics were obtained from the Gibbs sampler output using the R CODA package. The Bayes factor was estimated from the Gibbs sampler output using the Savage–Dickey density ratio.

The Bayes factor calculated using the Savage–Dickey density ratio (89.7) from the computationally intensive MCMC sampler output was similar to that obtained by the easy to compute Spiegelhalter and Smith method (93.1). This is consistent with our experience in other problems where the amount of information in the prior for the parameter being tested is comparable to the information in one data point. Of course, the MCMC output can be used to compute other useful information such as distributions of parameters, and predictions of genetic gain.

The Bayes factor of 89.7 represents strong evidence for an effect, but not strong enough to overcome low prior odds in a genome scan. Readers interested in fitting Bayesian models using MCMC are advised to study this example, and the examples provided with BUGS, in detail; most of the methods also apply to other designs and models considered in this chapter.

8.3.5 Populations with substructure

If there are two or more sub-populations with differing allele frequencies, linkage disequilibrium can be generated between loci (markers and/or QTL) without regard to genomic location. The resulting associations are considered spurious, since they are clearly not useful for gene discovery. This happens for example if there has been a recent admixture (Example 8.8).

Example 8.8. A population admixture.

Suppose sub-population 1 has Qq, Aa at 10:90 ratios, and that sub-population 2 has Qq, Aa at 90:10, and that there is no association between Q and A in either sub-population. After mixing the populations in the ratio 50:50 we have, within the combined population:

$$\Pr(Q) = 0.1 \times 0.5 + 0.9 \times 0.5 = 0.5, \tag{8.60}$$

$$\Pr(A) = 0.1 \times 0.5 + 0.9 \times 0.5 = 0.5, \tag{8.61}$$

$$\Pr(A, \text{pop1}) = \Pr(\text{pop1} \mid A)\Pr(A) = \Pr(A \mid \text{pop1})\Pr(\text{pop1}). \tag{8.62}$$

Therefore

$$\begin{aligned} \Pr(\text{pop1} \mid A) &= \Pr(A \mid \text{pop1})\Pr(\text{pop1})/\Pr(A) \\ &= 0.1 \times 0.5/0.50 \\ &= 0.1. \end{aligned} \tag{8.63}$$

Similarly

$$\Pr(\text{pop2} \mid A) = 0.9 \times 0.5/0.5 = 0.9, \tag{8.64}$$

$$\begin{aligned} \Pr(Q \mid A) &= \Pr(Q \mid A, \text{pop1})\Pr(\text{pop1} \mid A) + \Pr(Q \mid A, \text{pop2})\Pr(\text{pop2} \mid A) \\ &= \Pr(Q \mid \text{pop1})\Pr(\text{pop1} \mid A) + \Pr(Q \mid \text{pop2})\Pr(\text{pop2} \mid A) \\ &= 0.1 \times 0.1 + 0.9 \times 0.9 \\ &= 0.82, \end{aligned} \tag{8.65}$$

where the second equality uses the assumed within population independence of Q and A.

$$\Pr(Q, A) = \Pr(Q \mid A)\Pr(A) = 0.82 \times 0.50 = 0.41. \tag{8.66}$$

By definition, the LD coefficient, D, is given by

$$D = \Pr(Q, A) - \Pr(Q)\Pr(A) = 0.41 - 0.5 \times 0.5 = 0.16\,, \tag{8.67}$$

which is a substantial level of LD. However minor allele frequency differences lead to only small amounts of LD. □

Pritchard *et al.* (2000a, b), give Bayesian methods for testing and allowing for population structure, where the population may be stratified into several sub-populations. The number of sub-populations and the assignment of individuals to sub-populations are unknown. Information on population structure is obtained from a set of unlinked auxilliary markers.

The Bayesian approach is to simulate from the probability distribution of possible sub-populations. Each individual in the sample is assigned a set of unknown parameters representing the proportions of the individual's alleles coming from each population. The MCMC sampler is generated by sampling from the conditional distributions of each of these parameters in turn. These conditional distributions are related to the probability of belonging to a sub-population given the values of the auxilliary markers. The number of sub-populations is also allowed to vary using a 'reversible jump' MCMC technique (Green 1995).

This is the approach taken in the Structure method (Pritchard *et al.* 2000b):

1. If the population structure were known, the population can be divided into k sets S_i each without structure:
$$S = \bigcup_{i=1}^{k} S_i \,. \tag{8.68}$$
In this case the analysis can take the population substructure into account, e.g. by allowing for different allele frequencies among populations.

2. In the case where the population structure is unknown, i.e. the S_i above are unknown, but k is known in Equation (8.68), a Bayesian approach is used where additional indicator parameters indicate which of the subsets S_i each individual belongs. The distribution of parameters is obtained using MCMC.

3. The general case, where k is also unknown, is modelled using reversible jump Markov Chain Monte Carlo (RJMCMC; Green 1995). For each value of k there is a different model, as per case 2, and the model dimension varies with k. RJMCMC constructs 'jumps' between models, and assuming the sampler converges, gives a sample from the joint distribution of all models (sampled according to their posterior probabilities), and of parameters within models.

In the 'STRAT' test (Pritchard *et al.* 2000b, for case–control data, generalised by Thornsberry *et al.* 2001 for quantitative traits) a likelihood ratio statistic is constructed from the MCMC output. This stops short of a fully Bayesian approach.

More generally, in a fully Bayesian approach, for each possible population structure from the MCMC output, the within sub-population disequilibrium estimates can be obtained and the results averaged over possible population substructures according to their posterior probabilities. An important point to note is that the population structure and membership may not be determined uniquely. A fully Bayesian approach would take this uncertainty into account by giving probabilities for membership in each sub-population, e.g. an individual may be in sub-populations S_1, S_2, S_3 with probabilities 0.3, 0.2, 0.5, respectively.

Example 8.9. Structure analysis of a population admixture.

A population of 200 individuals, similar to example 6, was simulated with values of the markers M1, M2, M3, M4, M5 with 100 individuals from each of the sub-populations. Allele frequencies for the common allele varied from 90% down to 60% (Table 8.21). Simulated populations were analysed using `structure version 2.1`. The program was run for 2,000 iterations burn in and 10,000 further iterations, although further iterations are recommended for estimating the marginal probabilities $\Pr(D \mid K)$), where K specifies the number of sub-populations assumed. Recall that the marginal probabilities, $\Pr(D \mid K)$, are the values used in the calculation of Bayes factors for comparing models with different values of K (cf. Equation (8.3)). When all five markers were used the correct number (K=2) of sub-populations was identified with high probability by `structure`. Most individuals were predicted to belong to a single population with probability around 90%.

Table 8.21. Simulated marker frequencies.

	Marker										
	M1		M2		M3		M4		M5		
Allele	a1	a2	a1	a2	a1	a2	a1	a2	a1	a2	a3
pop1	0.9	0.1	0.8	0.2	0.7	0.3	0.6	0.4	0.16	0.20	0.64
pop2	0.1	0.9	0.2	0.8	0.3	0.7	0.4	0.6	0.20	0.00	0.80

Table 8.22. Posterior statistics for linkage disequilibrium coefficients. A population of size 200 was simulated with independent values for markers and QTLs within each sub-population. Allele frequencies within each sub-population were simulated as in Table 8.21. Linkage disequilibrium between the markers and corresponding QTL is shown for the combined population (pop1 $\cup$ pop2), for each sub-population separately (pop1, pop2), and for the sub-populations estimated by structure (pop1 (est.), and pop2 (est.)).

	$\hat{D}$ (95% c.i.)		
	M1–Q1	M2–Q2	M3–Q3
pop1 $\cup$ pop2	0.160 (0.140,0.18)	0.090 (0.070,0.11)	0.040 (0.02,0.07)
pop1	-0.001 (-0.010,0.01)	-0.0004 (-0.020,0.02)	-0.001 (-0.03,0.03)
pop2	-0.007 (-0.014,0.002)	0.030 (0.005,0.06)	-0.004 (-0.03,0.02)
pop1 (est.)	0.004 (-0.007,0.02)	0.030 (0.006,0.06)	-0.005 (-0.03,0.02)
pop2 (est.)	0.006 (-0.005,0.02)	0.003 (-0.020,0.03)	0.002 (-0.03,0.03)

For the purposes of estimating within sub-population disequilibrium, individuals were assigned to the sub-population with the highest probability of membership. Linkage disequilibrium coefficients and 95% credible intervals are shown in Table 8.22. Disequilibrium coefficients between the markers and QTLs (Q1–Q5) simulated at the same frequencies within sub-populations as the corresponding markers are shown for the combined populations, for the individual sub-populations, and for the estimated sub-populations.

Note that the disequilibrium is approximately 0 within sub-populations, and is zero or nearly zero to within experimental error within estimated sub-populations. We conclude that the structure analysis has successfully reduced the admixture disequilibrium to negligible levels in this example. This is not surprising since all but four individuals were assigned to the correct sub-population. □

Summary

We have seen that LD can be generated by population structure, and that the population structure analysis methods (Pritchard *et al.* 2000b; Thornsberry *et al.* 2001) can be effective at removing effects of population structure by estimating sub-population membership probabilities using a set of preferably unlinked control markers.

There are some caveats to the population structure analysis. A full Bayesian analysis is not yet available – the current methods give a likelihood ratio test, which gives a p-value, which as we have seen is not a reliable measure of evidence for an association.

To overcome this limitation, we suggest calculating approximate Bayes factors using the methods of previous sections, based on the likelihood ratio p-value as if it was from the standard ANOVA method. Strictly speaking, these methods do not apply to the more complex model, however they should give a good general indication. Similarly, we expect the power calculations for independent samples should apply approximately, with a moderate increase in sample size needed to allow for the more complex model, though this has yet to be fully tested.

A further caveat is that the population structure analysis can be affected by deviations from HWE or by null alleles. These factors, if present, may increase the number of subpopulations estimated by the program `structure`.

8.3.6 Samples from related individuals (pedigrees)

Many plant breeders will have access to populations where pedigree information is available. This material cannot be regarded as an independent population sample because individuals are related. However, these populations may still contain LD useful for association studies.

The methods in this subsection take into account relatedness between individuals in the analysis of marker–trait associations from a known pedigree. The methodology uses mixed models to allow for correlation between haplotype effects, with correlation structure based on IBD probabilities, and also to allow for polygenic effects, with covariance structure given by the additive relationship matrix from quantitative genetics. In so doing, the methods combine linkage and linkage disequilibrium information. The linkage or 'QTL mapping' information is generated by recombinations within the pedigree, detected by marker genotypes of parents and their offspring. The LD or association mapping, information, is generated by ancestral recombinations, and detected by population level associations between individuals.

The effectiveness of a pedigree population for LD mapping depends on the effective population size of the pedigree. The pedigree will probably be recorded for relatively few generations, and if it has formed from only a few founders the effective sample size for detecting population level LD is no larger than the number of founders. For example, a large single family provides no significant population level LD information, since there are effectively only two parents sampled from the population, and offspring will replicate most of the parental chromosomes in large blocks.

Incorporating polygenic random effects in the model via the additive relationship matrix effectively controls for population structure within the pedigree (Sillanpää and Bhattacharjee 2005). The pedigree analysis may, however, still be affected by spurious associations from population structure present when the founders were obtained. Relatedness between the founders would probably be unknown, and still needs to be checked and/or controlled by methods of Section 8.3.5. This might happen if a breeding population was obtained from material taken from several native provenances, as is the case for *P. radiata*. If individuals' ancestry cannot be traced back to the provenances, the genomes of currently growing trees may be a mixture of provenances, with unknown mixing probabilities, which can be estimated by the program `structure`.

Frequentist analysis

Meuwissen *et al.* (2002) use combined linkage disequilibrium and linkage information to fine map a QTL in cattle in a known pedigree. They fit a mixed model

$$y = \mu + Zh + u + e, \tag{8.69}$$

$$h \sim N(0, G\sigma^2_{\mathrm{h}}), \tag{8.70}$$

$$u \sim N(0, A\sigma^2_{\mathrm{u}}), \tag{8.71}$$

$$e \sim N(0, \sigma^2_{\mathrm{e}}), \tag{8.72}$$

where μ is the overall mean, h are random haplotype effects, u are random polygenic effects and e are residual errors. The haplotypes are based on markers close to the QTL locus. An 'infinite alleles model' is assumed so that each haplotype potentially has a different effect.

Note that each individual has two haplotypes. The number of haplotypes is greater than the number of individuals so haplotype effects cannot be estimated individually, however haplotype effects are correlated. Haplotype effects are identical if the corresponding QTL alleles are IBD. It follows that haplotype effects are correlated, with correlation matrix, G, given by the IBD probabilities for the QTL. The correlation between polygenic effects is given by the 'additive relationship matrix' A (Falconer and Mackay 1996).

The IBD probability calculation is based on Meuwissen and Goddard (2001):

- The calculation is different for each putative QTL locus. Haplotypes can be based on up to around 15 closely spaced markers around the putative QTL locus.
- For base haplotypes (first generation genotyped) Meuwissen and Goddard (M&G) use a modified coalescent to estimate IBD probabilities. Briefly, assume an effective population size of N_{e}, and T generations of random mating. Either simulate the coalescent (M&G 2000), or use the analytical formulae from M&G (2001). In a given simulated coalescent, haplotypes are considered IBD if they have coalesced within the T generations. IBD probabilities for a pair of haplotypes are estimated as the proportion of simulated coalescents where the haplotypes coalesced.
- For subsequent generations estimate IBD, using parental and marker information.

Similar to interval mapping QTL approaches (Lander and Botstein 1989), the analysis is repeated for each putative QTL position and likelihood ratios calculated at each position. A p-value is obtained by referring the likelihood ratio statistic to its sampling distribution under the null hypothesis of no effect. As demonstrated in previous sections, there are problems with the interpretation of p-values.

Note:

1. Meuwissen *et al.* fitted their model using ASREML (Gilmour *et al.* 2000). Analysis using the publicly available `nlme` R package is also possible, by forming the Choleski decomposition of the matrices G and A, and incorporating the Choleski factor into the Z-matrices, effectively transforming the sets of random effects to independent random effects, enabling the model to be fitted using the standard `nlme` covariance matrix classes as in Figure 8.19. This technique is used in the `lmeSplines` R package (Ball 2003).

```
library(nlme)
# QTL analysis
# given: trait y, G, A matrices
# calculate Z matrices for paternal and maternal haplotypes
# individual: a factor coding individual animals or plants
Zhp <- model.matrix(~ individual -1)
Zhm <- model.matrix(~ individual -1)
# Choleski matrix for G
Rg <- chol(G,pivot=FALSE)
Zh <- cbind(Zhp, Zhm) %*% t(Rg)
# Choleski matrix for G
Ra <- chol(A, pivot=FALSE)
Za <- Ra
# model with polygenic effects only
fit0 <- lme(y ~ 1, random=list(all=pdIdent(~Za -1)))
# model with QTL plus polygenic effects
fit1 <- lme(y ~ 1, random=list(all=pdBlocked(list(
            pdIdent(~Zh -1), pdIdent(~Za -1)))))
# compare models, LR test etc.
anova(fit0,fit1)
```

Figure 8.19. R code for mixed model QTL analysis (8.69)–(8.72) combining linkage and linkage disequilibrium.

2. The major computational difficulty in fitting the mixed models is evaluating the inverse of the matrix A, for large pedigrees.

Bayesian analysis

The mixed model of Equations (8.69)–(8.72) is almost Bayesian in that random effects have probability distributions. To make a full Bayesian model requires only specifying priors on the variance components $\sigma^2_{\rm h}, \sigma^2_{\rm a}, \sigma^2_{\rm e}$. As in previous sections Bayes factors and posterior probabilities are used for inference. An MCMC sampler can be generated and Bayes factors for comparing models and posterior probabilities can be calculated (cf. Section 8.3.4).

Note: There are some similarities between this approach and the 'BLADE' method (Liu *et al.* 2001 discussed in Section 8.3.2 above). Meuwissen and Goddard (2000, 2001) simulate or calculate IBD probabilities based on possible ancestral genealogies and use the IBD probabilities in a mixed model analysis, while Liu *et al.* simulate possible ancestral genealogies from a coalescent process with inference based on analysis of each of the simulated genealogies. The mixed model approach has the advantage of being able to incorporate pedigree information, and control for population structure, but the disadvantage of using fixed estimates of IBD probabilities, in the mixed model. This means one is effectively conditioning on the IBD probabilities being the true values in the mixed model analysis. This is the price paid for the convenience of using a more standard mixed model, with easier implementation in R or BUGS. The full Bayesian coalescent-based model con-

ditions on the population assumptions inherent in the coalescent, as does the Meuwissen and Goddard IBD estimation, but not on possible values of IBD probabilities that might be consistent with these assumptions.

□

Summary

A sample from a known pedigree combines QTL and LD mapping information in a single dataset. Mixed model analysis for a pedigree combines haplotype effects (at or around a single locus), with a correlation structure based on IBD probabilities, and polygenic effects. The effectiveness of the pedigree sample for LD mapping depends on the breadth and sample size of individuals from which the pedigree was founded.

Incorporating polygenic effects via the additive relationship matrix controls for population structure generated within the pedigree, but not for population structure when the founders were chosen, since relatedness between the founders is probably unknown. Population structure analysis on the founders is recommended.

For further information on models combining pedigree and LD information, see Wu and Zeng (2001), Wu *et al.* (2002), Farnir *et al.* (2002), Perez-Enciso (2003), Fan and Jun (2003), Lund *et al.* (2003), Meuwissen and Goddard (2004) and Lee and Van der Werf (2005). Other approaches to calculation of IBD probabilities include Heath (1997, 2002), (a stochastic MCMC method for use in large pedigrees, available as a software package Loki), Pong-Wong *et al.* (2001) and Gao and Hoeschele (2005) (deterministic methods). The deterministic methods are faster but are approximate, and/or ignore uncertainty in haplotypes.

8.3.7 QTL and LD mapping combined

In this subsection we consider combining information from QTL and LD mapping. Unlike Section 8.3.6, where population and pedigree information are combined in a single dataset, this subsection considers QTL and LD analysis on distinct datasets, where the results of QTL analysis are used as prior information for the LD analysis.

Brute force genotyping of, e.g. 500,000 SNP markers for large numbers of individuals would be prohibitive. In this section we consider a strategy for reducing the amount of genotyping by combining QTL (linkage) mapping and LD (association) mapping. A QTL mapping family is used to narrow down the range of the genomic region to search for associations, reducing the amount of genotyping required, and in the process increasing the prior odds per marker genotyped.

In the Bayesian paradigm, successive datasets can be analysed sequentially with the posterior distribution from each analysis being the prior for the next analysis. This makes sense logically since the posterior distribution represents our knowledge after the ith analysis which is the same as our knowledge prior to the $i + 1$st analysis. The same posterior distributions are obtained as if the datasets are analysed jointly in a single model. Utilising this fact, a natural way to approach combined analysis for separate QTL and LD mapping datasets is to use the posterior distribution from the QTL analysis as the prior distribution for the LD analysis. As before, if QTL mapping data is available, but the LD experiment

Table 8.23. Sample sizes and amount of genotyping required to locate a QTL when searching the genome using QTL and LD mapping combined. Assume there are 10 QTL explaining 5% of the variation, $D = 0.1$ or $D = 0.2$, allele frequencies 0.5 for QTL and marker for closest marker to the trait locus, a genome of 3×10^9 bases, extent of LD 6 kb, 500,000 SNP markers available at a spacing of 6 kb, giving prior probabilities per marker of 1/50,000. The QTL mapping results assume there are 12 chromosomes and 20 markers per chromosome at a spacing of 10 cM. Results are given for an overall posterior probability of 0.9 for an association.

	Number of QTL progeny genotyped			
	$n_{\mathrm{QTL}} = 100$	$n_{\mathrm{QTL}} = 400$	$n_{\mathrm{QTL}} = 1,000$	$n_{\mathrm{QTL}} = 3,000$
QTL: PProb(H_0)	0.5	0.2	< 0.001	< 0.001
se$(\hat{x})$	12.2 cM	6.5 cM	4.1 cM	2.4 cM
Number of SNPs in QTL interval	10,167	5,417	3,417	2,000
Average prior odds per SNP	1/20,333	1/6,770	1/3,417	1/2,000
Bayes factor required from LD	183,000	60,938	30,750	18,000
n_{LD}				
$D = 0.2$	1,589	1,508	1,451	1,407
$D = 0.1$	6,909	6,554	6,345	5,713
QTL-marker genotyping	24,000	96,000	240,000	720,000
LD-marker genotyping				
$D = 0.2$	16.2×10^6	8.2×10^6	5.0×10^6	2.8×10^6
$D = 0.1$	70.2×10^6	35.5×10^6	21.7×10^6	11.4×10^6
Total genotyping				
$D = 0.2$	16.2×10^6	8.3×10^6	5.2×10^6	3.5×10^5
$D = 0.1$	70.3×10^6	35.6×10^6	21.9×10^6	12.1×10^6

has not yet been done we can design the LD experiment with given power to obtain a sufficiently high Bayes factor to obtain a reasonably high posterior probability after the LD analysis. Next, we apply this approach to locating small effect QTL.

Table 8.23 shows results for sample sizes and amount of genotyping required to detect a QTL explaining 5% of the variation of a trait. Results are given for various sizes ($n_{\mathrm{QTL}} =$ 100, 400, 1,000, 3,000) of the QTL mapping family. For each family size the average standard error (se$(\hat{x})$) of the estimate of QTL location was calculated by simulation of an additive QTL. The QTL interval was assumed to be two standard deviations either side of the estimate, although smaller values could be considered and may be more cost-effective, at the risk of loosing some QTL. The number of SNPs within the QTL region was calculated and average prior odds per SNP were determined from this. Then, the Bayes factor required to obtain the required posterior probability of 0.9 calculated and the sample sizes (n_{LD}) for this calculated using the R function `ld.design()` from the `ldDesign` package (Ball 2004, 2005).

There are a number of factors which could be varied in searching for an optimal design – we have considered only two special cases here. Nevertheless, the results suggest, with

the extent of LD considered, that a significant efficiency gain can be achieved by combining QTL and LD mapping, and that the optimal QTL mapping population size will often be quite large. There are still quite a large number of SNPs to genotype per individual within the QTL region. Hence, the LD genotyping dominated the QTL genotyping except for the largest QTL sample size, and the maximum disequilibrium. Except for QTL population size 100, which had posterior probability of only 0.5, the width of the QTL region decreased gradually in inverse proportion to the square root of the QTL mapping population size. The least amount of total genotyping was for the largest QTL population of size $n_{\mathrm{QTL}} = 3,000$, with a fivefold reduction in genotyping compared to $n_{\mathrm{QTL}} = 100$. In this case, depending on phenotyping costs, larger QTL mapping populations should be considered before embarking on LD mapping. Values are given for both $D = 0.2$ and $D = 0.1$, with the latter being the minimum disequilibrium expected within the marker interval, by assumption. The total genotyping was still decreasing between $n_{\mathrm{QTL}} = 1,000$ and $n_{\mathrm{QTL}} = 3,000$, for both $D = 0.2$ and $D = 0.1$, so the optimum may be even higher.

Similar results are shown in Table 8.24 where the extent of LD is assumed to be $60\,\mathrm{kb}$. In this case the prior odds per SNP have increased tenfold compared to the previous case.

Table 8.24. Sample sizes and amount of genotyping required to locate a QTL when searching the genome using QTL and LD mapping combined. Assume there are 10 QTL explaining 5% of the variation, $D = 0.1$ or $D = 0.2$, allele frequencies 0.5 for QTL and marker for closest marker to the trait locus, a genome of 3×10^9 bases, extent of LD 60 kb, 50,000 SNP markers available at a spacing of 60 kb, giving prior probabilities per marker of 1/5,000. The QTL mapping results assume there are 12 chromosomes and 20 markers per chromosome at a spacing of 10 cM. Results are given for a posterior probability of 0.9 for an association.

	Number of QTL progeny genotyped			
	$n_{\mathrm{QTL}} = 100$	$n_{\mathrm{QTL}} = 400$	$n_{\mathrm{QTL}} = 1,000$	$n_{\mathrm{QTL}} = 3,000$
QTL: PProb(H_0)	0.5	0.2	< 0.001	< 0.001
se$(\hat{x})$	12.2 cM	6.5 cM	4.1 cM	2.4 cM
Number of SNPs in QTL interval	1,017	541	342	200
Average prior odds per SNP	1/2,033	1/677	1/342	1/200
Bayes factor required from LD	18,300	6,094	3,075	1,800
n_{LD}				
$D = 0.2$	1,408	1,323	1,268	1,222
$D = 0.1$	6,193	5,826	5,603	5,435
QTL-marker genotyping	24,000	96,000	240,000	720,000
LD-marker genotyping				
$D = 0.2$	1.4×10^6	7.2×10^5	4.3×10^5	2.4×10^5
$D = 0.1$	6.3×10^6	3.2×10^6	1.9×10^6	1.1×10^6
Total genotyping				
$D = 0.2$	1.5×10^6	8.1×10^5	6.7×10^5	9.6×10^5
$D = 0.1$	6.3×10^6	3.3×10^6	2.1×10^6	1.8×10^6

The optimal design appears to be approximately when $n_{\mathrm{QTL}} \approx 1,000$ when $D = 0.2$, and $n_{\mathrm{QTL}} \approx 3,000$, with up to approximately a threefold reduction in genotyping compared to $n_{\mathrm{QTL}} = 100$ when $D = 0.1$.

Summary

This subsection shows that QTL mapping and LD mapping analysis and experimental design can be profitably combined.

The posterior distributions from QTL analysis can be used as prior distributions for the LD analysis. The Bayes factor required from the LD mapping population for a given posterior probability is reduced for loci within a QTL mapping region.

Brute force genotyping of all markers in a genome scan for a sufficiently large population is very costly due to the very large amount of total genotyping. One possible strategy is to restrict genotyping of the LD mapping population to QTL regions. We have seen that this can result in reduced overall genotyping compared to a stand-alone LD mapping approach. Considering a single trait, the examples suggest that the optimal strategy is to use even larger QTL mapping populations than those currently used, prior to LD mapping, in order to find small effect genes.

A by-product of this approach is that, most spurious associations due to population structure will be eliminated by the QTL mapping study. If the QTL mapping intervals are small, e.g. with a sufficiently large QTL mapping family size this can be more effective than the TDT.

8.4 SUMMARY

From the point of view of statistical testing, association mapping for quantitative traits or complex diseases is characterised by :

1. Small effects requiring large sample sizes to detect, and,
2. Low prior odds for the effects, requiring additional evidence and/or stronger evidence from the data.

The frequentist hypothesis testing framework is not suited to testing scientific hypotheses: problems with p-values are accentuated with the large sample sizes required, and the frequentist approach does not consider prior odds. Bayesian statistics is ideally suited to the problem, since the Bayes factors or posterior probabilities do not depend on sample size for their interpretation. Bayesian prior probabilities represent the low prior odds, or prior information from alternative sources such as QTL mapping studies, differential expression microarray experiments or results from other species. Frequentists see priors as introducing an undesirable element of subjectivity, but here they are an essential part of the problem.

Thumma *et al.* (2005) conclude that:

> Careful selection of candidate genes through different approaches such as micro array analysis, EST database searches and QTL mapping is very important as a large amount of effort is needed for LD mapping. Success of LD mapping in out-crossing plants therefore depends upon careful selection of candidate genes...

Section 8.3.7 shows that combining QTL and LD mapping is an effective strategy for reducing the total amount of genotyping. The number of markers to genotype per individual is reduced by restricting to QTL regions, and the prior odds per marker is increased, resulting in fewer individuals needing to be genotyped in the association population for a given Bayes factor. Where the extent of LD is low, quite large QTL mapping populations can be profitably used. Of course, the gain from this strategy reduces as more traits are studied. In addition, other prior information can further improve prior odds.

In addition, it is important for the experiment to be designed with good power, and to use a reliable measure of evidence, such as the Bayes factor, and to give estimates free of selection bias. It is essential to consider and justify prior probabilities for markers. Otherwise many misleading and spurious associations will be generated. Large sample sizes are needed, but these are not out of reach of major companies or international cooperative efforts for economically important species.

We have given Bayesian power calculations (Ball 2005) for the case of independent samples only, as these are not yet available for other designs. The existing power calculations can, however, give a good indication of the sample size required in other cases. In principle, extending the methods to designs for TDT tests should not be difficult. Where an existing non-Bayesian power calculation exists this can be used in conjunction with the R function `SS.oneway.bf()` from `ldDesign`, or Equation (8.8). Using this function, calculate the Bayes factor equivalent to the p-value used for the design. This may be repeated, decreasing the p-value used in the power calculation until a sufficiently large Bayes factor is obtained.

For Bayesian haplotype analysis or analysis allowing for population substructure, simulations can be carried out to assess the additional sample size required for the more complex models. We conjecture that the analysis allowing for population substructure is equivalent to reducing degrees of freedom by an amount comparable to the number of unlinked markers used for the population structure analysis, which makes only a minor difference if the total sample size is much larger.

There are as yet relatively few published association studies in plants. One reason may be concerned about spurious associations resulting from population structure. We have considered four approaches which can be used to control and/or test for effects of population structure:

1. Where population structure is unknown, the population structure analysis methods of Section 8.3.5 can be used to reduce spurious effects in random population samples.

2. Where pedigree information is available, incorporating relationship information via the additive relationship matrix in a mixed model with polygenic effects (Section 8.3.6) controls for population structure generated within the pedigree.

3. Utilising many small families, the TDT design eliminates spurious associations between unlinked loci, but some 'partly spurious' associations between linked loci may remain at recombination distances less than 0.5.

4. Combining QTL and LD mapping as discussed in Section 8.3.7 eliminates spurious associations that are not within QTL mapping intervals.

A combination of two or more of these approaches can be applied for greater effectiveness.

Problems with epistasis are best addressed after most of the additive effects (genes or loci) contributing to a trait are identified. Then possible interactions between detected loci, and possibly between detected major loci and other loci can be examined.

Statistical methods cannot definitively establish causality. The best we can do is rule out likely causes of non-causal or spurious associations, and give putative associations with reasonably high posterior probability. Then, putative effects can be verified by functional testing.

8.5 REFERENCES

Akey, J., Jin, L., Xiong, M. 2000, Haplotypes vs single marker linkage disequilibrium tests: what do we gain? Eur. J. Hum. Genet. 9:291–300.

Allison, D.B. 1997, Transmission disequilibrium tests for quantitative traits. Am. J. Hum. Genet. 60:676–690.

Altshuler, D., Hirschhorn, J.N., Klannemark, M., Lindgren, C.M., Vohl, M.-C., Nemesh, J., Lane, C.R., Schaffner, S.F., Bolk, S., Brewer, C., Tuomis, T., Gaudet, D., Hudson, T.J., Daly, M., Groop, L., Lander, E.S. 2000, The common PPARγ Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. Nat. Genet. 26:76–80.

Bahlo, M., Thomson, R., Speed, T. 2003, Discussion of: "Ancestral inference in population genetics models with selection" by M. Stephens and P. Donnelly. Aust. NZ J. Stat. 45:427–428.

Ball, R.D. 2001, Bayesian methods for quantitative trait loci mapping based on model selection: approximate analysis using the Bayesian Information Criterion. Genetics 159:1351–1364.

Ball, R.D. 2003, lmeSplines – an R package for fitting smoothing spline terms in LME models. R News 3/3 p24–28.
`http://cran.r-project.org/src/contrib/Descriptions/lmeSplines.html`

Ball, R.D. 2004, ldDesign – an R package for design of experiments for detection of linkage disequilibrium.
`http://cran.r-project.org/src/contrib/Descriptions/ldDesign.html`

Ball, R.D. 2005: Experimental designs for reliable detection of linkage disequilibrium in unstructured random population association studies. Genetics 170:859–873.
`http://www.genetics.org/cgi/content/abstract/170/2/859`

Barry, D., Hartigan, J.A. 1992, Product partition models for change point problems. Ann. Stat. 20:260–279.

Bayes, T. 1763, An essay towards solving a problem in the doctrine of chances. Philos. Trans. R. Soc. 53:370–418.

Berger, J., Berry, D. 1988, Statistical analysis and the illusion of objectivity. Am. Sci. 76:159–165.

Berger, J.O., Sellke, T. 1987, Testing a point null hypothesis: the irreconcilability of P values and evidence (with discussion). J. Am. Stat. Assoc. 82:112–139.

Bernardo, J.M. 1999, Nested hypothesis testing: the Bayesian reference criterion. In: *Bayesian Statistics 6*. J.M. Bernardo, J.O. Berger, A.P. Dawid, A.F.M. Smith (Eds.) Oxford University Press, Oxford, pp. 101–130 (with discussion).

Bogdan, M., Ghosh, J.K., Doerge, R.W. 2004, Modifying the Schwarz Bayesian information criterion to locate multiple interacting quantitative trait loci. Genetics 167:989–999.

Brown, G.R., Gill, G.P., Kuntz, R.K., Langley, C.H., Neale, D.B. 2004, Nucleotide diversity and linkage disequilibrium in loblolly pine. Proc. Natl Acad. Sci. USA 42:15255–15260.

Casella, G., Berger, R.L. 1987, Reconciling Bayesian and frequentist evidence in the one-sided testing problem. J. Am. Stat. Assoc. 82:106–111.

Cavalli-Sforza, L.L., Cavalli-Sforza, F. 1993, The great human diasporas – the history of diversity and evolution (Italian original Chi Siamo: La Storia della Diversit'a Umana). ISBN 0-201-44231-0 (paperback), 1993.

Crow, T.J. (Ed.) 2002, The speciation of modern *Homo Sapiens*. ISBN 0-19-726311-9 (paperback) 2002.

Dickey, J.M. 1971, The weighted likelihood ratio, linear hypothesis on normal location parameters. Ann. Math. Stat. 42:204–223.

Drummond, A.J., Nicholls, G.K., Rodrigo, A.G., Solomon, W. 2002, Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics 161:1307–1320.

Dunner, S., Charlier, C., Farnir, F., Brouwers, B., Canon, J., *et al.* 1997, Towards interbreed IBD fine mapping of the *mh* locus: double-muscling in the *Asturiana de los Valles* breed involves the same locus as in the *Belgian Blue* cattle breed. Mamm. Genome 8:430–435.

Emahazion, T., Feuk, L., Jobs, M., Sawyer, S.L., Fredman, D., *et al.* 2001, SNP association studies in Alzheimer's disease highlight problems for complex disease analysis. Trends Genet. 17:407–413.

Falconer, D.S., Mackay, T.F.C. 1996, *Introduction to Quantitative Genetics*. Addison-Wesley Longman, Harlow, England.

Fan, R., Jung, J. 2003, High-resolution joint linkage disequilibrium and linkage mapping of quantitative trait loci based on Sibship data. Human Heredity 56:166–187.

Farnir, F., Grisart, B., Coppieters, W., Riquet, J., Berzi, P., *et al.* 2002, Simultaneous mining of linkage and linkage disequilibrium to fine map quantitative trait loci in outbred half-sib pedigrees: revisiting the location of a quantitative trait locus with major effect on milk production on bovine chromosome 14. Genetics 161:275–287.

Fearnhead, P., Donnelly 2001, Estimating recombination rates from population genetic data. Genetics 159:1299–1318.

Fisher, R.A. 1930, *The Genetical Theory of Natural Selection.* Clarendon Press, Oxford.

Foley, R. 1995, Humans before humanity. ISBN 0-631-20528-4 (paperback).

Gao, G., Hoeschele, I. 2005, Approximating identity-by-descent matrices using multiple haplotype configurations on pedigrees. Genetics 171:365–376.

Gelfand, A.E., Hills, S.E., Racine-Poon A., Smith, A.F.M. 1990, Illustration of Bayesian inference in normal data models using Gibbs sampling. J. Am. M. Stat. Assoc. 85:972–985.

Gelman, A., Carlin, B., Stern H.S., Rubin D.B. 1995, *Bayesian Data Analysis*. Chapman and Hall, London.

George, E.I., McCulloch, R.E. 1993, Variable selection via Gibbs sampling. J. Am. Stat. Assoc. 88(423):881–889.

Gilks, W.R., Spiegelhalter, D.J., Richardson, S. (Eds.) 1996, *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London.

Gilmour, A.R., Cullis, B.R., Welham, S.J. 2000, ASREML Reference Manual. NSW Agriculture, Orange, Australia.

Green, P.J. 1995, Reversible jump Markov Chain Monte Carlo computation and Bayesian model determination. Biometrika 82:711–732.

Griffiths, R.C., Marjoram, P. 1997, An ancestral recombination graph. pp. 257–270. In: *Progress in Population Genetics and Human Evolution.* P. Donnelly and S. Tavaré (Eds.), Springer, Berlin Heidelberg New York.

Gura, T. 2000: Can SNPs deliver on susceptibility genes? Science 293:593–595.

Hampton, T. 2000, Research Brief, Focus, 29 September 2000. Harvard University.

Hartigan, J.A. 1990, Partition models. Commun. Stat. Theory Meth. 19:2745–2756.

Heath, S.C. 1997, Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. Am. J. Hum. Genet. 61:748–760.

Heath, S. 2002, *Loki 2.4.5 – A package for multipoint linkage analysis on large pedigrees using reversible jump Markov chain Monte Carlo.* Centre National de Genotypage, Evry Cedex, France.
`http://bioweb.pasteur.fr/docs/doc-gensoft/loki/loki_doc.ps`

Hudson, R.R. 1983, Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. 23:183–201.

Hudson, R.R. 1990, Gene genealogies and the coalescent process. Oxf. Surv. Evol. Biol. 7: 1–43.

Jeffreys, H. 1961, *Theory of probability*, 3rd ed. Oxford University Press, London.

Kingman, J.F.C. 1982, On the genealogy of large populations. J. Appl. Prob. 19A:27–43.

Kaplan, N., Morris, R. 2001, Issues concerning association studies for fine mapping a susceptibility gene for a complex disease. Genet. Epidemiol. 20:432–457.

Kilpikari, R., Sillanpää, M.J. 2003, Bayesian analysis of multilocus association in quantitative and qualitative traits. Genet. Epidemiol. 25:122–135.

Lander, E.S., Botstein, D. 1989, Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185–199.

Lee, S.H., Van der Werf, J.H.J. 2005, The role of pedigree information in combined linkage disequilibrium and linkage mapping of quantitative trait loci in a general complex pedigree. Genetics 169:455–466.

Liu, J.S., Sabatti, C., Teng, J., Keats, B.J.B., Risch, N. 2001, Bayesian analysis of haplotypes for linkage disequilibrium mapping. Genome Res. 11:1716–1724.

Long, A.D., Langley, C.H. 1999: The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. Genome Res. 9:720–731.

Lund, M.S., Sorensen, P., Guldbrandtsen, P., Sorensen, D.A. 2003, Multitrait fine mapping of quantitative trait loci using combined linkage disequilibria and linkage analysis. Genetics 163:405–410.

Luo, Z.W. 1998, Detecting linkage disequilibrium between a polymorphic marker locus and a trait locus in natural populations. Heredity 80:198–208.

Meuwissen, T.H.E., Goddard, M.E. 2000, Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. Genetics 155:421–430.

Meuwissen, T.H.E., Goddard, M.E. 2001, Prediction of identity-by-descent probabilities from marker haplotypes. Genet. Sel. Evol. 33:605–634.

Meuwissen, T.H.E., Hayes, B.J., Goddard, M.E. 2001, Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829.

Meuwissen, T.H.E., Karlsen, A., Lien, S., Oldsaker, I., Goddard, M. 2002, Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. Genetics 161:373–379.

Meuwissen, T.H.E., Goddard, M.E. 2004, Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. Genet. Sel. Evol. 36:261–279.

Molitor, J., Majoram, P., Thomas, D. 2003, Fine scale mapping of disease genes with multiple mutations via spatial clustering techniques. Am. J. Hum. Genet. 73:1368–1384.

Natarajan, R., Kass, R.E. 2000, Reference Bayesian methods for generalized linear mixed models. J. Am. Stat. Assoc. 95:227–237.

Neale, D.B., Savolainen, O. 2004, Association genetics of complex traits in conifers. Trends Plant Sci. 9:325–330.

Nielsen, D.M., Weir, B.S. 2001, Association studies under general disease models. Theor. Popul. Biol. 60:253–263.

Nielsen, D.M., Zaykin, D. 2001, Association mapping: where we've been, where we're going. Expert Rev. Mol. Diagn. 1(3):89–97.

Nordborg, M. 2001, Coalescent theory. pp 179–212. In: *Handbook of Statistical Genetics.* D. J. Balding, *et al.* (Eds.), Wiley, New York.

Nordborg, M., Tavaré, S. 2002, Linkage disequilibrium: what history has to tell us. Trends Genet. 18:83–90.

Oppenheimer, S. 2003, The Real Eve: Modern Man's Journey Out of Africa. ISBN 0-786-71192-2 (Hardcover).

Perez-Enciso, M. 2003, Fine mapping of complex trait genes combining pedigree and linkage disequilibrium information: a Bayesian unified framework. Genetics 163:1497–1510.

Plummer, M., Best, N., Cowles, K., Vines, K. 2005, Coda Version 0.9-5: Output analysis and diagnostics for MCMC. `http://www-fis.iarc.fr/coda/`

Pong-Wong, R., George, A.W., Woolliams, J.A., Haley, C.S. 2001, A simple and rapid method for calculating identity-by-descent matrices using multiple markers. Genet. Sel. Evol. 33:453–471.

Pot, D., McMillan, L., Echt, C., Le Provost, G., Garnier-Géré, P., Cato, S., Plomion, C. 2005, Nucleotide variation in genes involved in wood formation in two pine species. New Phytol. 167:101–112.

Pritchard, J.K., Stephens, M., Donnelly, P. 2000a, Inference of population structure using multilocus genotype data. Genetics 155:945–959.

Pritchard, J.K., Stephens, M., Rosenberg, N.A., Donnelly, P. 2000b, Association mapping in structured populations. Am. J. Hum. Genet. 67:170–181.

Pritchard, J.K., Rosenberg, N.A. 1999, Use of unlinked markers to detect population stratification in association studies. Am. J. Hum. Genet. 65:220–228.

Raftery, A.E., Lewis, S.M. 1992, One long run with diagnostics: implementation strategies for Markov chain Monte Carlo. Stat. Sci. 7:493–497.

Raftery, A.E., Lewis, S.M. 1995, The number of iterations, convergence diagnostics and generic Metropolis algorithms. In: *Practical Markov Chain Monte Carlo.* W.R. Gilks, D.J. Spiegelhalter, and S. Richardson, (Eds.), Chapman, Hall, London.

Raftery, A.E. 1996, Hypothesis testing and model selection. Chapter 10, pp. 163–188. In: *Markov chain Monte Carlo in practice.* W.R. Gilks, S. Richardson, and D.J. Spiegelhalter (Eds.).

Raftery, A.E., Madigan, D., Hoeting, J.A. 1997, Bayesian model averaging for linear regression models. J. Am. Stat. Assoc. 92:179–191.

Roberts, G.O. 1996, Markov chain concepts related to sampling algorithms. pp 45–58. In: *Markov Chain Monte Carlo in Practice*. W.R. Gilks, S. Richardson, D.J. Spiegelhalter (Eds.), Chapman, Hall, London.

Schwarz, G. 1978, Estimating the dimension of a model. Ann. Stat. 6:461–464.

Sillanpää, M.J., Bhattacharjee, M. 2005, Bayesian association-based fine mapping in small chromosomal segments. Genetics 169:427–439.

Sillanpää, M.J., Corander, J. 2002, Model choice in gene mapping: what and why. Trends Genet. 18:301–307.

Spiegelhalter, D., Smith, A.F.M. 1982, Bayes factors for linear and log-linear models with vague prior information. J.R. Stat. Soc. B 44:377–387.

Spiegelhalter, D., Thomas, A., Best, N., Gilks, W. 1995, *BUGS – Bayesian inference using Gibbs sampling Version 0.50.* MRC Biostatistics Unit, Cambridge.
`http://www.mrc-bsu.cam.ac.uk/bugs`

Spielman, R.S., McGinnis, R.E., Ewens, W.J. 1993, Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am. J. Hum. Genet. 52:506–516.

Stephens, M. 2001, Inference under the coalescent. pp 213–238 In: *Handbook of Statistical Genetics.* D.J. Balding, *et al.* (Eds.), Wiley, New York.

Stephens, M., Smith, M.J., Donnelly, P.A. 2001, A new statistical method for haplotype reconstruction from population data. Am. J. Hum. Genet. 66:978–989.

Stephens, M., Donnelly, P. 2003, Ancestral inference in population genetics models with selection (with discussion). Aust. NZ J. Stat. 45:395–430.

Stringer, C., McKie, R. 1996, *African Exodus*. Owl Books, London.

Sykes, B. 2001, *The Seven Daughters of Eve: The Science That Reveals Our Genetic Ancestry*. W.W. Norton, New York.

Terwilliger, J.D., Weiss, K.M. 1998, Linkage disequilibrium mapping of complex disease: fantasy or reality? Curr. Opin. Biotechnol. 9:578–594.

Thornsberry, J.M., Goodman, M.M., Doebley, J., Kresovich, S., Nielsen, D., *et al.* 2001, Dwarf8 polymorphisms associate with variation in flowering time. Nat. Genet. 28:286–289.

Thumma, B.R., Nolan, M.F., Evans, R., Moran, G.F. 2005, Polymorphisms in *Cinnamoyl CoA Reductase* (*CCR*) are associated with variation in microfibril angle in *Eucalyptus* spp. Genetics 10.1534/genetics.105.042028.

Tierney, L. 1994, Markov chains for exploring posterior distributions. In: *Proceedings to the 1991 Interface Symposium*, also available as Technical report #560 (revised), School of Statistics, University of Minnesota.

Wells, S. 2003, *The Journey of Man: A Genetic Odyssey*. Princeton University Press, Princeton.

Wilson, S. 2003, Discussion of: "Ancestral inference in population genetics models with selection" by M. Stephens and P. Donnelly. Aust. NZ J. Stat. 45:423–426.

Wright, S. 1931, Evolution in Mendelian populations. Genetics 16:97–159.

Wu, R., Zeng, Z.-B. 2001, Joint linkage and linkage disequilibrium mapping in natural populations. Genetics 157:899–909.

Wu, R., Ma, C.X., Casella, G. 2002, Joint linkage and linkage disequilibrium mapping in natural populations. Genetics 160:779–792.

Zöllner, S., Pritchard, J.K. 2005, Coalescent-based association mapping and fine mapping of complex trait loci. Genetics 169:1071–1092.

Chapter 9

LINKAGE DISEQUILIBRIUM-BASED ASSOCIATION MAPPING IN FORAGE SPECIES

Mark P. Dobrowolski[1] and John W. Forster[2]

9.1 INTRODUCTION

Forage species provide herbage for grazing, hay and silage production servicing livestock production industries in both tropical and temperate regions of the world. The grazing industries are responsible for dairy, meat and fibre production. Temperate forage grasses include perennial ryegrass (*Lolium perenne* L.), Italian ryegrass (*Lolium multiflorum* Lam.), meadow fescue (*Festuca pratensis* Huds.), tall fescue (*Festuca arundinacea* Schreb.), cocksfoot (*Dactylis glomerata* L.), Kentucky bluegrass (*Poa pratensis* L.), smooth bromegrass (*Bromus inermis* L.) and harding grass (*Phalaris aquatica* L.). Temperate forage legumes include white clover (*Trifolium repens* L.), red clover (*Trifolium pratense* L.), subterranean clover (*T. subterraneum* L.), bird's foot trefoil (*Lotus corniculatus* L.) and lucerne/alfalfa (*Medicago sativa* L.). Tropical forage species include grasses such as buffelgrass (*Pennisetum ciliare* L.), and members of the genera *Brachiaria* and *Paspalum*, as well as legumes such as round-leafed cassia (*Chamaecrista rotundifolia*), siratro (*Macroptilium atropurpuerum*) and members of the genera *Stylosanthes*, *Centrosema* and *Desmodium*. A range of temperate and warm-season grasses are also important for non-forage applications such as turf and amenity cultivation. The temperate turf grasses include *Lolium*, *Festuca*, *Poa* and *Agrostis* (bentgrass) species, while warm-season and tropical turf grasses include switchgrass (*Panicum virgatum* L.), seashore paspalum and bahiagrass (*Paspalum vaginatum* Swartz *and Paspalum notatum* Flugge) and members of the *Zoysia* (zoysiagrass) and *Cynodon* (bermuda grass) genera (Forster *et al*. 2001a).

This chapter will focus mainly on the potential application of linkage disequilibrium (LD)-based association mapping to perennial ryegrass and white clover. These two

[1] Primary Industries Research Victoria, Plant Genetics and Genomics Research Platform, Hamilton Centre, Mt. Napier Road, Hamilton, Victoria 3300, Australia
Molecular Plant Breeding Cooperative Research Centre, Australia

[2] Primary Industries Research Victoria, Plant Genetics and Genomics Research Platform, Victorian AgriBiosciences Centre, La Trobe R&D Park, Bundoora, Victoria 3083, Australia

species are cultivated, generally in combination, in grassland-producing regions of Northern Europe, the Pacific North–West of the United States, Japan, South-Eastern Australia, and New Zealand, and provide high-quality forage with superior palatability and nutrient content (Forster *et al*. 2001a). The aspects of the biology, life history, and genomic and genetic resources of these two predominantly outbreeding species, discussed in the following sections, provide both challenges and opportunities for association mapping studies. Progress in association studies for silage maize (*Zea mays* L.) is also described. This cereal species has been developed for silage production, particularly in Europe, and has been the subject of more intensive "proof-of-concept" studies for association mapping in recent times (Thornsberry *et al*. 2001; Rafalski and Morgante 2004).

9.2 FACTORS INFLUENCING ASSOCIATION MAPPING STRATEGIES FOR PERENNIAL RYEGRASS AND WHITE CLOVER

9.2.1 Taxonomy

Perennial ryegrass is a member of the Poaeae tribe of the Poodae super-tribe in the Pooideae sub-family of the grass and cereal family Poaceae (Soreng and Davis 1998). The *Lolium* and *Festuca* genera are closely allied, and the most nearly related major cereal species is cultivated oats (*Avena sativa* L.) within the Aveneae tribe of the Poodae. The Triticeae cereal tribe (wheat, barley and rye) is located within the Triticodae super-tribe of the Pooidae. Rice (*Oryza sativa* L.), by contrast, is located in the Poaceae sub-family Bambusoideae. Translation genomics from rice to perennial ryegrass based on whole genome DNA sequence data consequently traverses a significant phylogenetic distance, but the use of partial genomic sequence and expressed sequence tag (EST) data from wheat (*Triticum aestivum* L.: Powell and Langridge 2004) and barley (*Hordeum vulgare* L.) exploits closer taxonomic affinities.

White clover is a member of the Trifolieae tribe of the cool-season Galegoid clade in the Papilinoideae sub-family of the legume family Fabaceae (Doyle and Luckow 2003). The most closely related genus is *Melilotus* (sweet clovers) and the genus *Medicago*, including alfalfa, is also part of the Trifolieae. As a consequence, the model legume species barrel medic (*Medicago truncatula* Gaertn.) shares a common ancestor relatively recently in evolutionary time with white clover. Translational genomics based on whole genome sequencing of *M. truncatula* (Young *et al*. 2005; Zhu *et al*. 2005) is consequently anticipated to be highly efficient for members of the *Trifolium* genus. The other model legume species, *Lotus japonicus* Gifu, is also a Galegoid legume located in a separate tribe, Loteae.

9.2.2 Genome Structure

Members of the *Lolium* genus are diploids with a fundamental chromosome number of 7 ($2n = 2x = 14$). The genome size of perennial ryegrass has been estimated through measurements of nuclear DNA content by microdensitometry (Hutchinson *et al*. 1979; Seal and Rees 1982). A 2C value of 4.16 pg corresponds to a haploid genome size of c. 1.6×10^9 bp. The individual genome sizes of other *Lolium* species vary, with the inbreeding taxa such as *Lolium temulentum* exhibiting nuclear DNA contents c. 50% larger than those of the outbreeding species. In common with other Poaceae family members, the genomes of *Lolium* species contain large numbers of dispersed repetitive sequences, frequently belonging to major retroelement families (Jenkins *et al*. 2000).

White clover is an allotetraploid species with a fundamental chromosome number of 8 ($2n = 4x = 16$). The evolutionary origin of white clover is not fully understood, although two diploid species, *T. occidentale* D. Coombe and *T. nigrescens* Viv and another allotetraploid species, *T. uniflorum* L. have been considered to be potential progenitors (Chen and Gibson 1970; 1971; Badr *et al.* 2002). More recently, studies of choloroplast DNA and nuclear ribosomal DNA variation have implicated the diploid species *T. occidentale and T. pallescens* Schreber as the progenitor taxa (Ellison *et al.* 2006). Measurements of nuclear DNA content have been performed using Feulgen microdensitometry (Grime and Mowforth 1982; Campbell *et al*. 1999) and flow cytometry (Aramuganathan and Earle 1991). A range of values between 2.07 and 3.0 pg were reported. The lower value corresponds to a haploid genome size of c. 8×10^8 bp, implying an average sub-genome size close to 400 Mb and comparable to that of *M. truncatula*.

9.2.3 Reproductive Behaviour

Perennial ryegrass is an obligate outbreeding species, with a gametophytic self-incompatibility (SI) system controlled by two loci (*S* and *Z*). Incompatible matings occur when the alleles at both loci in the male gametophyte (pollen grain) match one of the two alleles at each locus in the female sporophyte (Cornish *et al*. 1979). This mechanism ensures a very low level of self-fertilisation, and also limits the level of fertility in closely related individuals such as full-sibs. The proportions of compatible matings between related individuals depend on the degree of allelic complexity at the SI loci. Genetic studies of both natural and synthetic populations have revealed a large number of different alleles (c. 20) of each of the *S* and *Z* loci (Devey *et al*. 1994; Fearon *et al*. 1994). The *S*/*Z*-based SI system is present in a wide range of outbreeding Poaceae species. Genetic studies of SI have been performed in the reed canary grass species *Phalaris coerulescens* and cereal rye (*Secale cereale* L.), among others (Baumann *et al*. 2000).

White clover is also an obligate outbreeding species, with a gametophytic SI system controlled by a series of alleles at a single locus (*S*) (Attwood 1940, 1941, 1942a). Rare instances of self-compatibility have been reported (Attwood 1942b; Yamada *et al*. 1989), presumably due to the presence of self-fertile (*Sf*) alleles at the SI locus.

The three well-recognised outbreeding species within the *Lolium* genus (perennial ryegrass, Italian ryegrass and annual ryegrass [*Lolium rigidum* Gaud.] chiefly differ in reproductive development and growth habit, with a continuous range of variation from short-lived annual ecotypes of annual ryegrass at one extreme to long-lived perennial ecotypes of perennial ryegrass at the other extreme. The different species show a high degree of interfertility, and are reproductively isolated from one another in nature through separate geographical locations and mean flowering time. The annual-type ryegrasses are more characteristic of Mediterranean environments and are early flowering, while the perennial-type ryegrasses are characteristic of cooler temperate environments and are generally later flowering (Forster *et al*. 2005a).

White clover also exhibits perennial growth behaviour, with known genetic variation for floral development traits (Connolly 1990; Williams *et al*. 1998). The stoloniferous growth habit of white clover strongly contributes to vegetative persistence. Morphogenetic traits are generally negatively correlated with flowering date (Cogan *et al*. 2006), and such relationships may be related to the observed decline of vegetative growth following onset of flowering, due to inhibition of stolon elongation (Kawanabe *et al*. 1963).

9.2.4 Varietal Development

Due to the obligate outbreeding natures of both perennial ryegrass and white clover, both natural and synthetic populations are highly genetically heterogeneous. Varietal development is typically based on the following process: evaluation of base populations containing 2,000–5,000 individuals; selection of c. 200 potential parental clones (Vogel and Pedersen 1993); and polycrossing to generate a synthetic 1 (Syn1) population. The number of foundation individuals may vary from as low as four for perennial ryegrass to 50–100 for polyploid species such as tall wheat grass and alfalfa (Bray and Irwin 1999).

9.2.5 Genomic Resources

High-throughput gene discovery by EST sequencing has generated significant genomic resources for the temperate pasture species. A collection of 44,534 perennial ryegrass ESTs was generated from single pass sequencing of randomly selected clones from 29 cDNA libraries that represent a range of plant organs (leaf, root, seed, etc.), developmental stages (vegetative, reproductive, etc.) and environmental conditions (Sawbridge *et al*. 2003a). EST redundancy was resolved through assembly with the CAP3 application, leading to identification of 12,170 unigenes. Similarly, a collection of 42,017 white clover ESTs was generated from 16 cDNA libraries obtained from a broad range of plant organs, developmental stages and environmental conditions (Sawbridge *et al*. 2003b). Each of the sequences was annotated by comparison to GenBank and SwissProt public sequence databases and automated intermediate Gene Ontology (GO) annotation was obtained (Spangenberg *et al*. 2005).

Complementing the EST resources, large insert DNA libraries have been generated using bacterial artificial chromosome (BAC) vectors. The perennial ryegrass BAC library was constructed using *Hin*dIII-generated partial DNA fragments in the pBeloBACII vector, and consists of 50,304 clones with an average genome size of 113 kb, corresponding to 3.4 genome equivalents. The white clover BAC library was constructed in a similar fashion, and consists of 50,302 clones with an average genome size of 101 kb, corresponding to 6.3 genome equivalents (Spangenberg *et al*. 2005). BAC library screening has been performed for both species using both macroarray hybridisation and PCR analysis of microtitre plate pools.

9.2.6 Genetic Resources

The development of molecular genetic markers and associated genetic maps for perennial ryegrass has been comprehensively reviewed by Forster *et al*. (2001a, 2004), while the application of genetic marker analysis to trait dissection has been reviewed by Yamada and Forster (2005).

A comprehensive set (c. 400) of unique perennial ryegrass genomic DNA-derived SSR (LPSSR) markers has been developed using enrichment library technology (Jones *et al*. 2001). This resource has been augmented by the results of similar studies on a smaller scale by Kubik *et al*. (2001) and Lauvergeat *et al*. (2005). The perennial ryegrass EST collection has also been used for the development of a set of 310 EST-SSR primer pairs (Faville *et al*. 2004). More recently, gene-associated single nucleotide polymorphism (SNP) markers have been developed through both *in vitro* and *in silico*

discovery in perennial ryegrass and white clover (Spangenberg *et al*. 2005; Shinozuka *et al*. 2005; Cogan *et al*. 2006b; Chapter 4 in this volume).

Development of the first generation reference genetic map of perennial ryegrass was performed through the use of public domain genetic markers, including restriction fragment length polymorphisms (RFLP) and amplified fragment length polymorphisms (AFLPs). This was achieved through coordination of the International *Lolium* Genome Initiative (ILGI), using the p150/112 one-way pseudo-testcross population of 183 F_1 genotypes. The ILGI reference map, constructed through collaboration between Victorian DPI, Australia; Institute of Grassland and Environmental Research (IGER), UK; Yamanashi Prefectural Dairy Experiment Station (YPDES) and the National Agricultural Research Centre for Hokkaido Region (NARCH), Japan and the Institut National de la Recherche Agronomique (INRA), France contained c. 200 AFLP loci and 109 heterologous RFLP loci (detected by wheat, barley, oat and rice cDNA probes), allowing the inference of comparative relationships between perennial ryegrass and other Poaceae species (Jones *et al*. 2002a). The ILGI map was enhanced through the addition of 93 LPSSR loci, providing the basis of framework genetic mapping in other populations (Jones *et al*. 2002b).

A second generation reference genetic mapping family was developed based on the $F_1(NA_6 \times AU_6)$ two-way pseudo-testcross family of 157 F_1 genotypes, generating two parental genetic maps. The consolidated genetic maps included 43 LPSSR loci, 88 EST-RFLP loci and 71 EST-SSR loci on the NA_6 parental map, with a total length of 963 cM; and 49 LPSSR loci, 67 EST-RFLP loci and 58 EST-SSR loci on the AU_6 parental map, with a total length of 779 cM (Faville *et al*. 2004).

Trait dissection for perennial ryegrass has been performed in multiple populations to allow quantitative trait locus (QTL) analysis. The p150/112 population has been analysed for traits such as vegetative and reproductive morphogenesis, reproductive development and winter-hardiness, and herbage quality (Yamada *et al*. 2004; Cogan *et al*. 2005), while the $F_1(NA_6 \times AU_6)$ population has been studied for a range of root and shoot morphogenesis, photosynthetic efficiency, pseudostem water soluble carbohydrate (WSC) content and crown rust resistance characters (Forster *et al*. 2004). Other perennial ryegrass populations have been analysed to detect genetic control of crown rust resistance (Dumsday *et al*. 2003; Muylle *et al*. 2005a,b), vernalisation response (Jensen *et al*., 2005) and flowering time variation (Armstead *et al*. 2004; Warnke *et al*. 2004).

The development of molecular genetic markers for white clover has been reviewed by Forster *et al*. (2001a). A comprehensive set (c. 400) of unique white clover genomic DNA-derived SSR (TRSSR) markers was developed using enrichment library construction technology (Kölliker *et al*. 2001a). The white clover EST library was also been used to develop 792 EST-SSR primer pairs (Barrett *et al*. 2004).

Genetic map development in white clover was performed using a combination of TRSSR and AFLP markers. The reference mapping population was the $F_2(I.4R \times I.5J)$ family that was developed at IGER, Aberystwyth, UK, with parental genotypes from fourth and fifth generation inbred lines descended from plants containing the rare self-fertile (S_f) allele. A single F_1 plant was self-pollinated to generate an F_2 population of 150 individuals (Michaelson-Yeates *et al*. 1997). The level of genetic polymorphism between the inbred parents, as assessed with TRSSR markers, was 48% of those markers showing efficient amplification. The $F_2(I.4R \times I.5J)$ map contained 135 loci (78 TRSSR and 57 AFLP) on 18 linkage groups (two more than the karyotypic number), with a total map

length of 825 cM. The extent of map construction was limited by high levels of segregation distortion, affecting 39% of the TRSSR loci, with the majority distorted towards the heterozygous genotypic class (Jones *et al*. 2003). A higher-resolution genetic map largely based on EST-SSR markers was constructed using the F_1(Sustain 6525-2 × NRS 364-7) mapping family (Barrett *et al*. 2004). The EST-SSR markers detected homoeologous locations between the ancestral genomes at high frequency, and provided the basis for standard chromosome nomenclature development.

The F_2(I.4R × I.5J) genetic map has been exploited for QTL analysis of a number of vegetative morphogenesis, reproductive morphogenesis and reproductive development traits (Cogan *et al*. 2006a). Target traits were measured across a number of years of clonal replication, and geographical sites in Wales and Scotland, United Kingdom. Individual environment analyses detected a large number of QTLs for each trait, with QTL clustering for correlated traits. Multi-environment combined analysis revealed genomic locations that are relatively insensitive to genotype × environment effects. The F_1(Sustain 6525-2 × NRS 364-7) population has also been used for QTL analysis, specifically targeting seed production traits such as inflorescence density, yield per inflorescence and thousand-seed weight (Barrett *et al*. 2005). Stability of QTL effects was observed across temporal replication, along with co-location of QTLs for correlated traits.

9.2.7 Population Structure

Genetic diversity analysis has been performed for perennial ryegrass using AFLP and SSR-based marker systems (Roldán-Ruiz *et al*. 2000; Guthridge *et al*. 2001; Forster *et al*. 2001b, 2005a) and has revealed larger levels of genetic variation within than between populations. Varieties based on small numbers of parental genotypes (restricted-base varieties) were found to show lower levels of intrapopulation diversity and to be more readily discriminated than those based on larger numbers of parental individuals (non-restricted base varieties). AFLP profiling has also been used to determine levels of genetic variability within and between white clover populations (Kölliker *et al*. 2001b). As for perennial ryegrass, the majority of genetic variation was detected within rather than between populations, and divergent varieties were largely discriminated on the basis of AFLP profile. Bulking at the genotypic level followed by AFLP analysis was used to determine the level of congruence between morphophysiological and genotypic variation in white clover.

9.3 IMPLICATIONS OF OUTBREEDING REPRODUCTIVE BEHAVIOUR FOR ASSOCIATION MAPPING STRATEGIES

For effective implementation of marker-assisted selection (MAS), use of genetic markers that are in linkage of varying strength, rather than directly associated with the gene of interest, is a problem for both inbreeding and outbreeding species. For this reason, closely linked markers, ideally flanking the target region, are preferred. However, given fixation of the target region in an inbred background to generate a homogeneous variety, the problem of potential reversal of linkage between favourable gene variant and selected marker allele is eliminated. In the context of a genetically heterogeneous synthetic population, complete fixation of a target genomic region will be difficult and

slow to achieve, and consequently, the probability of recombination to decouple the favourable marker–trait allele combination will be high. This logic implies that diagnostic genetic markers are of even higher potential value for outbreeding than inbreeding crops.

A further problem for the implementation of MAS for pasture species is the large number of parental genotypes that are generally used in the polycross design for synthetic population development. The number of foundation individuals varies between restricted base varieties (4–6 parents) and non-restricted base varieties (6–100+ parents) (Guthridge *et al*. 2001; Forster *et al*. 2001b). Even for restricted base varieties, the process of tagging each gene variant in the parental genotypes with linked markers would imply multiple cycles of genetic trait-dissection in pair cross-derived mapping families. This contrasts with the situation for inbreeding plant species, in which the trait-dissection process in a sib-ship derived from crossing the future donor and recipient lines provides all relevant information for subsequent recurrent selection. One way to overcome this multiplicity of marker-trait allele associations would be to pre-introgress the desired combination into each of the selected parents. However, this implies a prior round of MAS, and does not adequately address the logistical complexity problem for molecular breeding of outcrossing forages, even for single gene traits. All of these considerations suggest that diagnostic genetic markers provide the ideal system for molecular breeding of forage species.

Intensive breeding of the outcrossing pasture species dates from the early years of the twentieth century, and was based on selection from adapted ecotypes. The domestication process is consequently relatively recent, implying that many contemporary varieties are derived from landraces with large effective population sizes. The reproductive habit and presumptive population structure of outbreeding forage species would be expected to dispose towards limited LD, extending over relatively short molecular distances (Mackay 2001; Forster *et al*. 2004). This is especially true for long-established populations derived from a large number of founding parents, as expected for ecotypes and long-established varieties, in which many rounds of recombination have occurred. These factors would tend to favour association studies based on candidate genes (Andersen and Lübberstedt 2003) rather than whole genome scans (see Chapter 5), although newly synthesised populations with small numbers of parents may prove suitable for limited genome-wide marker-based analysis.

9.4 CURRENT RESULTS FROM ASSOCIATION MAPPING STUDIES

To date, LD-based association mapping studies in forages have been restricted to perennial ryegrass and to silage maize, which as a facultative inbreeding species may be expected to show different patterns of LD from the grasses and clovers. The trait focus of the relevant studies has been on cold tolerance, flowering time, and forage quality characters. The approaches taken have moved from whole genome scans using anonymous markers to the more focused approach of selecting particular genes as candidates for testing more specific association with putatively correlated phenotypic traits.

In perennial ryegrass, Skøt *et al*. (2002, 2005b) adopted a whole genome scan approach using AFLP profiling to generate a large number of markers. Natural populations, with pre-existing genecological data, were chosen to represent the geographical range of perennial ryegrass within Europe (Sackville Hamilton *et al.* 2002).

AFLP marker frequencies within populations were tested for association with the phenotype of populations as a whole, rather than with the individual plant phenotype. For the cold tolerance trait, marker frequency-trait associations were regarded as spurious if the marker was not also correlated with average winter temperature and altitude clines of the population origins, independent of average summer temperature clines, in addition to these correlations being consistent across geographical origin (Skøt *et al.* 2002). One AFLP marker locus was identified as associated with cold tolerance using this method, and was prioritised for further investigation to assess its predictive value for this trait.

A similar approach using previously analysed phenotypes of natural populations was employed to analyse AFLP marker association with flowering time variation in perennial ryegrass (Skøt *et al.* 2005b). Five marker loci showed association with flowering time variation, based on linear regression analysis, following exclusion of distinct populations from the analysis due to concerns regarding population structure. Twenty nine of the 590 AFLP bands tested could be mapped in a full-sib genetic mapping family (F_2[Aurora × Perma]: Armstead *et al*. 2002), based on polymorphism in the mapping population and the assumption that co-migrating bands represent homologous loci. Three of the five markers that were associated with flowering time were closely linked in a region of linkage group (LG) 7, which contains a large quantitative trait locus (QTL) for heading date variation accounting for 70% of the phenotypic variance (Armstead *et al*. 2004). These three markers also revealed significant LD in pair-wise comparisons. However, the effects of residual population structure were still evident in the data, as many unlinked marker pairs also showed significant LD. Studies using AFLPs for whole genome scans are vulnerable to various problems, including: the potential confounding effects of AFLP band size homoplasy; population structure, which was clearly evident in the latter study and was not measured in the former, leading to spurious associations; and unanticipated ecological parameters that define other differentiated traits within and between comparator groups. Another potential problem was the use of population-level phenotypic data rather than individual plant performance, and comparison of this data to population-based allele frequencies. Use of population-based data is especially problematic considering the high levels of intrapopulation genetic diversity, often greater than 80%, that is frequently observed in perennial ryegrass populations (Guthridge *et al.* 2001; Kubik *et al.* 2001; Dobrowolski *et al.* 2005). As previously stated, LD is also predicted to decay rapidly in outbreeding species such as perennial ryegrass (Flint-Garcia *et al.* 2003) and for this reason, and the various problems with use of whole genome scans, LD-based association mapping in forages has now shifted to the use of genotype-specific phenotypic data and a candidate gene-based approach.

In silage maize, a candidate gene-based approach was used to test the association between digestibility and sequence haplotypes of the maize peroxidase gene *ZmPox3* (Guillet-Claude *et al.* 2004b) and the three *O*-methyltransferase genes, *CCoAOMT2*, *CCoAOMT1*, and *AldOMT* (Guillet-Claude *et al.* 2004a). These genes were chosen as candidates for association studies on the basis of presumptive role in lignin biosynthesis and co-location with QTLs for lignin content and cell wall digestibility based on analysis of genetic mapping populations. Genotypic data was obtained by direct sequencing of the haplotypes of these genes from various silage maize lines. LD decayed rapidly, reaching values of $r^2 = 0.2$ within 200–1,200 bp, as seen in other studies of diverse maize populations (Remington *et al.* 2001). Associations were found between the digestibility phenotype and distinct haplotypes containing insertions in both *ZmPox3* and *CCoAOMT2*

(Guillet-Claude *et al.* 2004a, b). However, the investigators recognised that population structure was not accounted for in testing for the associations, and these effects may have given rise to spurious haplotype–phenotype correlations.

Candidate gene-based association mapping studies in perennial ryegrass have so far targeted forage quality and flowering traits. Skøt *et al.* (2005a) selected 100 genotypes from each of nine European populations showing wide variation in heading date, and measured this trait from replicated plants grown in pots following vernalisation. In parallel, herbage quality traits were measured in replicate, including WSC content, nitrogen content and dry matter digestibility, using near infrared reflectance spectroscopy (NIRS).

Two candidate genes were targeted for analysis by Skøt *et al.* (2005a): the *LpHd1* gene, which is the putative perennial ryegrass ortholocus of the rice *Hd1* photoperiodic control gene; and *LpAlkInv*, an alkaline invertase gene which has been mapped to LG 6 of the perennial ryegrass genetic map, coincident with a WSC QTL. SNP discovery in *LpAlkInv* was based on analysis of 24 genotypes that span the phenotypic range variation in the larger sample set. The 6,328 bp *LpAlkInv* genomic clone, composed of six exons and five introns, was tiled with overlapping PCR primers and the resulting amplicons from each genotype were directly sequenced. Minimal problems due to paralogous sequence amplification and sequence frameshifts due to heterozygous indels were reported, and heterozygous SNPs were identified based on overlapping peaks in sequence traces. Across the 48 possible haplotypes, an average of one SNP was identified per 28 bp. LD between SNP loci decayed to an r^2 value of 0.1 over 2–3 kb distances. Association analysis based on individual SNP genotype rather than SNP haplotypes revealed no significant correlations with WSC content variation, but possible correlations with heading date variation. Equivalent analysis is being performed for the 7.3 kb *LpHd1* genomic sequence, including an assessment of the impact of population structure on the association analysis. Other candidate gene-based studies in perennial ryegrass (Ponting *et al*. 2005) have observed similar values for decay of LD, with r^2 values dropping below 0.1 over 2 kb distances between SNP loci covering 6,137 bp of the forage quality candidate gene *LpFT1* (putative sucrose:fructose 6-fructosyltransferase) (Lidgett *et al*. 2002). However in the *Lp1-SST* (1-sucrose:sucrose fructosyltransferase) gene (Chalmers *et al*. 2003), little LD was evident between SNP loci covering 4,269 bp and the equivalent study to detect LD decay rate in the *LpASRa2* (abscisic acid, stress, ripening) gene (Forster *et al*. 2005b; Cogan *et al.* 2006b) was limited by the short distance (447 bp) between the most distal SNPs. These analyses were based on a set of 81 diverse perennial ryegrass genotypes. With the addition of genotypes from the closely related ryegrass taxa *L. hybridum*, *L. multiflorum*, *L. rigidum*, *L. temulentum* higher levels of LD were observed, presumably due to the confounding effects of population structure, and specific *LpASRa2* haplotypes were evidently conserved across species.

9.5 CONCLUSIONS

The demonstrated rapid decay of LD over short physical distances in the genomes of outbreeding forage species provides support for the view that whole genome scans are unlikely to identify regions of the genome that are causally responsible for agronomically-important phenotypic variation, given the constraints of current technology. By contrast, the candidate gene-based approach should be highly suitable, allowing the

identification of individual SNPs or gene-length SNP haplotypes associated with superior allele content (Forster *et al*. 2004). The success of this approach, however, depends on the effective selection of candidates that are genuinely functionally correlated with the traits of interest. The validation of such candidates requires a multidisciplinary approach with supporting data obtained through QTL co-location, microarray-based transcriptome analysis, transgenic modification, induced mutagenesis and other approaches. For perennial ryegrass and white clover, large-scale EST resources, BAC libraries, trait-specific mapping populations, spotted cDNA arrays and efficient transformation systems are currently available (Spangenberg *et al*. 2005). Candidate gene-based methods for LD analysis of these species are consequently highly feasible and are anticipated to have important impacts on breeding practices.

9.6 REFERENCES

Andersen, J.R., Lübberstedt, T., 2003, Functional markers in plants. Trends in Plant Science 8: 554–560.

Armstead, I.P., Turner, L.B., King, I.P., Cairns, A.J., Humphreys, M.O., 2002, Comparison and integration of genetic maps generated from F_2 and BC_1-type mapping populations in perennial ryegrass. Plant Breeding 121: 501–507.

Armstead, I.P., Turner, L.B., Farrell, M., Skøt, L., Gomez, P., Montoya, T., Donnison, I.S., King, I.P., Humphreys, M.O., 2004, Synteny between a major heading-date QTL in perennial ryegrass (*Lolium perenne* L.) and the *Hd3* heading-date locus in rice. Theoretical and Applied Genetics 108: 822–828.

Arumuganathan, K., Earle, E.D., 1991, Nuclear DNA content of some important plant species. Plant Molecular Biology Reporter 9: 208–218.

Attwood, S.S., 1940, Genetics of cross-incompatibility among self-incompatible plants *of Trifolium repens*. Journal of the American Society of Agronomy 32: 955–968.

Attwood, S.S., 1941, Controlled self- and cross-pollination of *Trifolium repens*. Journal of the American Society of Agronomy 33: 538–545.

Attwood, S.S., 1942a, Oppositional alleles causing self-incompatibility in *Trifolium repens*. Genetics 27, 333–338.

Attwood, S.S., 1942b, Genetics of pseudo-self-incompatibility and its relation to cross-incompatibility in *Trifolium repens* L. Journal of Agricultural Research 64: 699–709.

Badr, A., Sayed-Ahmed, H., El-Shanshouri, A., Watson, L.E., 2002, Ancestors of white clover (*Trifolium repens* L.), as revealed by isozyme polymorphisms. Theoretical and Applied Genetics 106: 143–148.

Barrett, B., Griffiths, A., Schreiber, M., Ellison, N., Mercer, C., Bouton, J., Ong, B., Forster, J., Sawbridge, T., Spangenberg, G., Bryan, G., Woodfield, D., 2004, A microsatellite map of white clover (*Trifolium repens* L.). Theoretical and Applied Genetics 109: 596–608.

Barrett, B.A., Baird, I.J., Woodfield, D.R., 2005, A QTL analysis of white clover seed production. Crop Science 45: 1844–1850.

Baumann, U., Juttner, J., Bian, X.-Y., Langridge, P., 2000, Self-incompatibility in the grasses. Annals of Botany 85 (Supplement A): 203–209.

Bray, R.A., Irwin, J.A.G., 1999, *Medicago sativa* L. (lucerne) cv. Hallmark. Australian Journal of Experimental Agriculture 39: 643–644.

Campbell, B.D., Caradus, J.R., Hunt, C.L., 1999, Temperature responses and nuclear DNA amounts of seven white clover populations which differ in early spring growth rates. New Zealand Journal of Agricultural Research 42: 9–17.

Chalmers, J., Johnson, X., Lidgett, A., Spangenberg, G., 2003, Isolation and characterisation of a sucrose:sucrose 1-fructosyltransferase gene from perennial ryegrass (*Lolium perenne* L.). Journal of Plant Physiology 160: 1385–1391.

Chen, C.C., Gibson, P.B., 1970, Chromosome pairing in two interspecific hybrids of *Trifolium*. Canadian Journal of Genetics and Cytology 12: 790–794.

Chen, C.C., Gibson, P.B., 1971, Karyotypes of fifteen *Trifolium* species in section Amoria. Crop Science 11: 441–445.

Cogan, N.O.I., Smith, K.F., Yamada, T., Francki, M.G., Vecchies, A.C., Jones, E.S., Spangenberg, G.C., Forster, J.W., 2005, QTL analysis and comparative genomics of herbage quality traits in perennial ryegrass (*Lolium perenne* L.). Theoretical and Applied Genetics 110: 364–380.

Cogan, N.O.I., Abberton, M.T., Smith, K.F., Kearney, G., Marshall, A.H., Williams, A., Michael-Yeates, T.P.T., Bowen, C., Jones, E.S., Vecchies, A.C., Forster, J.W., 2006a, Individual and multi-environment combined analyses identify QTLs for morphogenetic and reproductive development traits in white clover (*Trifolium repens* L.). Theoretical and applied Genetics 112: 1401–1415.

Cogan, N.O.I., Ponting, R.C., Vecchies, A.C., Drayton, M.C., George, J., Dobrowolski, M.P., Sawbridge, T.I., Spangenberg, G.C., Smith, K.F., Forster, J.W., 2006b, Gene-associated single nucleotide polymorphism (SNP) discovery in perennial ryegrass (*Lolium perenne* L.) Mol Genet Genomics 276: 101–12.

Connolly, V., 1990, Seed yield and yield components in ten white clover cultivars. Irish Journal of Agricultural Research 29: 41–48.

Cornish, M.A., Hayward, M.D., Lawrence, M.J., 1979, Self-incompatibility in ryegrass. I. Genetic control in diploid *Lolium perenne* L. Heredity 43: 95–106.

Devey, F., Fearon, C.H., Hayward, M.D., Lawrence, M.J., 1994, Self-incompatibility in ryegrass. 11. Number and frequency of alleles in a cultivar of *Lolium perenne* L. Heredity 73: 262–264.

Dobrowolski, M.P., Bannan, N.R., Ponting, R.C., Forster, J.W., Smith, K.F., 2005, Population genetics of perennial ryegrass (*Lolium perenne* L.): differentiation of pasture and turf cultivars. In: *Molecular breeding for the genetic improvement of forage crops and turf*. Humphreys M.O. (ed.). Wageningen Academic Publishers: The Netherlands. p. 273.

Doyle, J.J., Luckow, M.A., 2003, The rest of the iceberg. Legume diversity and evolution in a phylogenetic context. Plant Physiology 131: 900–910.

Dumsday, J.L., Smith, K.F., Forster, J.W., Jones, E.S., 2003, SSR-based genetic linkage analysis of resistance to crown rust (*Puccinia coronata* Corda f. sp. *lolii*) in perennial ryegrass (*Lolium perenne* L.). Plant Pathology 52: 628–637.

Ellison, N.W., Liston, A., Szeimer, J.J., Williams, W.M., Taylor, W.L., 2006, Molecular phylogenetics of the clover genus (*Trifolium*-Leguminosae). Molecular Phylogenetics and Evolution 39: 688−705.

Faville, M., Vecchies, A.C., Schreiber, M., Drayton, M.C., Hughes, L.J., Jones, E.S., Guthridge, K.M., Smith, K.F., Sawbridge, T., Spangenberg, G.C., Bryan, G.T., Forster, J.W., 2004, Functionally-associated molecular genetic marker map construction in perennial ryegrass (*Lolium perenne* L.). Theoretical and Applied Genetics 110: 12–32.

Fearon, C.H., Cornish, M.A., Hayward, M.D., Lawrence, M.J., 1994, Self-incompatibility in ryegrass. 10. Number and frequency of alleles in a natural-population *of Lolium perenne* L. Heredity 73: 254–261.

Flint-Garcia, S.A., Thornsberry, J.M., Buckler, E.S.I., 2003, Structure of linkage disequilibrium in plants. Annual Review of Plant Biology 54: 357–374.

Forster, J.W., Jones, E.S., Kölliker, R., Drayton, M.C., Dumsday, J., Dupal, M.P., Guthridge, K.M., Mahoney, N.L., van Zijll de Jong, E., Smith, K.F., 2001a, Development and Implementation of Molecular Markers for Forage Crop Improvement. In: *Molecular breeding of forage crops*. Spangenberg G. (ed.). Kluwer Academic Press: Dordecht. pp. 101–133.

Forster, J.W., Jones, E.S., Kölliker, R., Drayton, M.C., Dupal, M.P., Guthridge, K.M., Smith, K.F., 2001b, DNA profiling in outbreeding forage species. In: *Plant genotyping – the DNA fingerprinting of plants*. Henry R (ed.). CABI Press: New York. pp. 299–320.

Forster, J.W., Jones, E.S., Batley, J., Smith, K.F., 2004, Molecular marker-based genetic analysis of pasture and turf grasses. In: *Molecular breeding of forage and turf*. Hopkins A., Wang Z.-Y., Sledge M., Barker R.E. (eds.). Kluwer Academic Press: Dordecht. pp. 197–239.

Forster, J.W., Jones, E.S., Smith, K.F., Guthridge, K.M., Dupal, M.P., Howlett, S., Hughes, L.J., Garvie, S., Preston, C., 2005a, Molecular Marker Technology for the Study of Molecular Variation and Comparative Genetics in Pasture Grasses. In: *Plant Genome: Biodiversity and Evolution, Volume 1 Pt. B: Phanerogams*. Sharma A.K., Sharma A. (eds.). Science Publishers: Enfield, NH. pp. 119–155.

Forster, J.W., Cogan, N.O.I., Vecchies, A.C., Ponting, R.C., Drayton, M.D., George, J., Dumsday, J.L., Sawbridge, T.I., Spangenberg, G.C., 2005b, Gene-associated single nucleotide polymorphism (SNP) discovery in perennial ryegrass (*Lolium perenne* L.). In: *Molecular breeding for the genetic improvement of forage crops and turf*. Humphreys M.O. (ed.). Wageningen Academic Publishers: The Netherlands. p. 199.

Grime, J.P., Mowforth, M.A., 1982, Variation in genome size – an ecological interpretation. Nature 299: 151–153.

Guillet-Claude, C., Birolleau-Touchard, C., Manicacci, D., Fourmann, M., Barraud, S., Carret, V., Martinant, J.P., Barriere, Y., 2004a, Genetic diversity associated with variation in silage corn digestibility for three *O*-methyltransferase genes involved in lignin biosynthesis. Theoretical and Applied Genetics 110: 126–135.

Guillet-Claude, C., Birolleau-Touchard, C., Manicacci, D., Rogowsky, P.M., Rigau, J., Murigneux, A., Martinant, J.P., Barriere, Y., 2004b, Nucleotide diversity of the *ZmPox3* maize peroxidase gene:

relationships between a MITE insertion in exon 2 and variation in forage maize digestibility. BMC Genetics 5: 16.

Guthridge, K.M., Dupal, M.P., Kolliker, R., Jones, E.S., Smith, K.F., Forster, J.W., 2001, AFLP analysis of genetic diversity within and between populations of perennial ryegrass (*Lolium perenne* L.). Euphytica 122: 191–201.

Hutchinson, J., Rees, H., Seal, A.G., 1979, An assay of the activity of supplementary DNA in *Lolium*. Heredity 43: 411–421.

Jenkins, G., Head, J., Forster, J.W., 2000, Probing meiosis in hybrids of *Lolium* (Poaceae) with a discriminatory repetitive genomic sequence. Chromosoma 109: 280–286.

Jensen, L.B., Andersen, J.R., Frei, U., Xing, Y., Taylor, C., Holm, P.B., Lübberstedt, T., 2005, QTL mapping of vernalisation response in perennial ryegrass (*Lolium perenne* L.) reveals co-location with an orthologue of wheat VRN1. Theoretical and Applied Genetics 110: 527–536.

Jones, E.S., Dupal, M.P., Kölliker, R., Drayton, M.C., Forster, J.W., 2001, Development and characterisation of simple sequence repeat (SSR) markers for perennial ryegrass (*Lolium perenne* L.). Theoretical and Applied Genetics 102: 405–415.

Jones, E.S., Mahoney, N.L., Hayward, M.D., Armstead, I.P., Jones, J.G., Humphreys, M.O., King, I.P., Kishida, T., Yamada, T., Balfourier, F., Charmet, C., Forster, J.W., 2002a, An enhanced molecular marker-based map of perennial ryegrass (*Lolium perenne* L.) reveals comparative relationships with other Poaceae species. Genome 45: 282–295.

Jones, E.S., Dupal, M.D., Dumsday, J.L., Hughes, L.J., Forster, J.W., 2002b, An SSR-based genetic linkage map for perennial ryegrass (*Lolium perenne* L.). Theoretical and Applied Genetics 105: 577–584.

Jones, E.S., Hughes, L.J., Drayton, M.C., Abberton, M.T., Michaelson-Yeates, T.P.T., Forster, J.W., 2003, An SSR and AFLP molecular marker-based genetic map of white clover (*Trifolium repens* L.). Plant Science 165: 531–539.

Kawanabe, S., Yoshihara, K., Okada, T., Ueno, M., Hidaka, M., 1963, Studies on summer depression of pasture crops. 3. Influence of flower bud removal upon vegetative growth of Ladino clover. Journal of the Japanese Society for Grassland Science 9: 31–41.

Kölliker, R., Jones, E.S., Drayton, M.C., Dupal, M.P., Forster, J.W., 2001a, Development and characterisation of simple sequence repeat (SSR) markers for white clover (*Trifolium repens* L.). Theoretical and Applied Genetics 102: 416–424.

Kölliker, R., Jones, E.S., Jahufer, M.Z.Z., Forster, J.W., 2001b, Bulked AFLP analysis for the assessment of genetic diversity in white clover (*Trifolium repens* L.). Euphytica 121: 305–315.

Kubik, C., Sawkins, M., Meyer, W.A., Gaut, B.S., 2001, Genetic diversity in seven perennial ryegrass (*Lolium perenne* L.) cultivars based on SSR markers. Crop Science 41: 1565–1572.

Lauvergeat, V., Barre, P., Bonnet, M., Ghesquiére, M., 2005, Sixty simple sequence repeat markers for use in the *Festuca-Lolium* complex of grasses. Molecular Ecology Notes 5: 401–405.

Lidgett, A., Jennings, K., Johnson, X., Guthridge, K., Jones, E., Spangenberg, G., 2002, Isolation and characterisation of a fructosyltransferase gene from perennial ryegrass (*Lolium perenne*). Journal of Plant Physiology 159: 1037–1043.

Mackay, T.F.C., 2001, The genetic architecture of quantitative traits. Annual Review of Genetics 35: 303–309.

Michaelson-Yeates, T.P.T., Marshall, A., Abberton, M.T., Rhodes, I., 1997, Self-incompatibility and heterosis in white clover (*Trifolium repens* L.). Euphytica 94: 341–348.

Muylle, H., Baert, J., Van Bockstaele, E., Moerkerke, B., Goetghebeur, E., Roldán-Ruiz, I., 2005a, Identification of molecular markers linked with crown rust (*Puccinia coronata* f.sp. *lolii*) resistance in perennial ryegrass (*Lolium perenne*) using AFLP markers and a bulked segregant approach. Euphytica 143: 135–144.

Muylle, H., Baert, J., Van Bockstaele, E., Petijs, J., Roldán-Ruiz, I., 2005b, Four QTLs determine crown rust (*Puccinia coronata* f.sp. *lolii*) resistance in a perennial ryegrass (*Lolium perenne*) population. Heredity 95: 348–357.

Ponting, R.C., Drayton, M.D., Cogan, N.O.I., Dobrowolski, M.D., Spangenberg, G.C., Smith, K.F., Forster, J.W., 2005, SNP discovery and haplotypic variation in full-length herbage quality genes of perennial ryegrass (*Lolium perenne* L. In: *Molecular breeding for the genetic improvement of forage crops and turf*. Humphreys MO (ed.). Wageningen Academic Publishers: The Netherlands. p. 196.

Powell, W., Langridge, P., 2004, Unfashionable crop species flourish in the 21st century. Genome Biology 5: 233.

Rafalski, A., Morgante, M., 2004, Corn and humans: recombination and linkage disequilibrium in two genomes of similar size. Trends in Genetics 20: 103–111.

Remington, D.L., Thornsberry, J.M., Matsuoka, Y., Wilson, L.M., Whitt, S.R., Doebley, J., Kresovich, S., Goodman, M.M., Buckler, E.S.I., 2001, Structure of linkage disequilibrium and phenotypic associations

in the maize genome. Proceedings of the National Academy of Sciences of the United States of America 98: 11479–11484.

Roldán-Ruiz I., Dendauw J., Van Bockstaele J., Depicker, E., De Loose, M., 2000, AFLP markers reveal high polymorphic rates in ryegrasses (*Lolium* spp.). Molecular Breeding 6: 125–134.

Sackville Hamilton, N.R., Skøt, L., Chorlton, K.H., Thomas, I.D., Mizen, S., 2002, Molecular genecology of temperature response in *Lolium perenne*: 1. Preliminary analysis to reduce false positives. Molecular Ecology 11: 1855–1863.

Sawbridge, T., Ong, E.-K., Binnion, C., Emmerling, M., McInnes, R., Meath, K., Nguyen, N., Nunan, K., O'Neill, M., O'Toole, F., Rhodes, C., Simmonds, J., Tian, P., Wearne, K., Webster, T., Winkworth, A., Spangenberg, G., 2003a, Generation and analysis of expressed sequence tags in perennial ryegrass (*Lolium perenne* L.). Plant Science 165: 1089-1100.

Sawbridge, T., Ong, E.-K., Binnion, C., Emmerling, M., Meath, K., Nunan, K., O'Neill, O., O'Toole, F., Simmonds, J., Wearne, K., Winkworth, A., Spangenberg, G., 2003b, Generation and analysis of expressed sequence tags in white clover (*Trifolium repens* L.). Plant Science 165: 1077–1089.

Seal, A.G., Rees, H., 1982, The distribution of quantitative DNA changes associated with the evolution of the diploid Festuceae. Heredity 49: 179–190.

Shinozuka, H., Hisano, H., Ponting, R.C., Jones, E.S., Cogan, N.O.I., Forster, J.W., Yamada, T., 2005, Molecular cloning and genetic mapping of perennial ryegrass protein kinase CK2α-subunit genes. Theoretical and Applied Genetics 112: 167–177.

Skøt, L., Sackville Hamilton, N.R., Mizen, S., Chorlton, K.H., Thomas, I.D., 2002, Molecular genecology of temperature response in *Lolium perenne*: 2. Association of AFLP markers with ecogeography. Molecular Ecology 11: 1865–1876.

Skøt, L., Humphreys, J., Armstead, I.P., Humphreys, M.O., Gallagher, J.A., Thomas, I.D., 2005a, Approaches for associating molecular polymorphisms with phenotypic traits based on linkage disequilibrium in natural populations of *Lolium perenne*. In: *Molecular breeding for the genetic improvement of forage crops and turf*. Humphreys M.O. (ed.). Wageningen Academic Publishers: The Netherlands. p. 157.

Skøt, L., Humphreys, M.O., Armstead, I., Heywood, S., Skot, K.P., Sanderson, R., Thomas, I.D., Chorlton, K.H., Hamilton, N.R.S., 2005b, An association mapping approach to identify flowering time genes in natural populations of *Lolium perenne* (L.). Molecular Breeding 15: 233–245.

Soreng, R.J., Davis, J.I., 1998, Phylogenetics and character evolution in the grass family (Poaceae): simultaneous analysis of morphological and chloroplast DNA restriction site character sets. Botanical Reviews 64: 1–85.

Spangenberg, G., Forster, J.W., Edwards, D., John, U., Mouradov, A., Emmerling, M., Batley, J., Felitti, S., Cogan, N.O.I., Smith, K.F., Dobrowolski, M.P., 2005, Future directions in the molecular breeding of forage and turf. In: *Molecular breeding for the genetic improvement of forage crops and turf*. Humphreys M.O. (ed.). Wageningen Academic Publishers: The Netherlands. pp. 83–97.

Thornsberry, J.M., Goodman, M.M., Doebley, J., Kresovich, S., Nielsen, D., Buckler IV, E.S., 2001, *Dwarf8* polymorphisms associate with variation in flowering time. Nature Genetics 28: 286–289.

Vogel, K.P., Pedersen, J.F., 1993, Breeding systems for cross-pollinated forage grasses. Plant Breeding Reviews 11: 251–274.

Warnke, S.E., Barker, R.E., Jung, G., Rouf Mian, M.A., Saha, M.C., Brilman, L.A., Dupal, M.D., Forster, J.W., 2004, Genetic linkage mapping of an annual × perennial ryegrass population. Theoretical and Applied Genetics 109: 294–304.

Williams, T.A., Abberton, M.T., Thornley, W.J., Evans, D.R., Rhodes, I., 1998, Evaluation of seed production potential in white clover (*Trifolium repens* L.) varietal improvement programs. Grass Forage Science 53: 197–207.

Xing, Q., Ru, Z., Li, J., Zhou, C., Jin, D., Sun Y., Wang, B. Cloning a second form of adenine phosphoribosyl transferase gene (TaAPT2) from wheat and analysis of its association with thermo-sensitive genic male sterility (TGMS) (2005) Plant Science, 169 (1), pp. 37–45.

Yamada, T., Forster, J.W., 2005, QTL analysis and trait dissection in ryegrasses (*Lolium* spp.). In: *Molecular breeding for the genetic improvement of forage crops and turf*. Humphreys M.O. (ed.). Wageningen Academic Publishers: The Netherlands. pp. 43–53.

Yamada, T., Higuchi, A., Fukuoka, A., 1989, Recurrent selection of white clover (*Trifolium repens* L.) using self-compatible plants. I. Selection of self-compatible plants and inheritance of a self-compatibility factor. Euphytica 44: 167–172.

Yamada, T., Jones, E.S., Cogan, N.O.I., Vecchies, A.C., Nomura, T., Hisano, H., Shimamoto, Y., Smith, K.F., Forster, J.W., 2004, QTL analysis of morphological, developmental and winter hardiness-associated traits in perennial ryegrass (*Lolium perenne* L.). Crop Science 44: 925–935.

Young, N.D., Cannon, S.B., Sato, S., Kim, D., Cook, D.R., Town, C.D., Roe, B.A., Tabata, S., 2005, Sequencing the genespaces of *Medicago truncatula* and *Lotus japonicus*. Plant Physiology 137: 1174–1181.

Zhu, H., Choi, H.-K., Cook, D.R., Shoemaker, R.C., 2005, Bridging model and crop legumes through comparative genomics. Plant Physiology 137: 1189–1196.

Chapter 10

GENE-ASSISTED SELECTION: APPLICATIONS OF ASSOCIATION GENETICS FOR FOREST TREE BREEDING

Phillip L. Wilcox[1], Craig E. Echt[2], and Rowland D. Burdon[3]

SUMMARY

This chapter describes application of association genetics in forest tree species for the purposes of selection. We use the term gene-assisted selection (GAS) to denote application of marker–trait associations determined via association genetics, which we anticipate will be based on polymorphisms associated with expressed genes. The salient features of forest trees are reviewed, including existing and somewhat limited knowledge of linkage disequilibrium (LD), as well as genomic information for both conifers and hardwoods. The relatively short span of LD in largely undomesticated and outbred forest tree species offer good prospects for precisely locating quantitative trait nucleotide (QTN), but necessitates wise candidate gene selection and generation of nongenic sequences, which could be limiting, particularly for conifers. Prerequisites for successful application are discussed, and include suitable populations for detecting LD; powerful quantitative genetic and bioinformatic capabilities; large EST libraries, if not whole genomic sequences, to identify candidate genes; and other capabilities for studying functional genomics; as well as a mix of quantitative genetics, tree breeding, and molecular biology skills. Experimental designs for tree improvement applications are also described, as well as analytical methods. For existing tree improvement practice, GAS should be applicable in virtually all population strata, although careful evaluation on a case-by-case basis will be needed to determine the appropriate implementation pathway(s). Such evaluation will likely include numerical simulation. GAS also fits well with other biotechnologies used for tree improvement. A number of impediments to

[1] Cellwall Biotechnology Centre, Scion (New Zealand Forest Research Institute), Private Bag 3020, Rotorua, New Zealand

[2] USDA Forest Service, Southern Institute of Forest Genetics 23332 MS Highway 67, Saucier, MS 39574, USA

[3] Ensis Genetics, Scion (New Zealand Forest Research Institute), Private Bag 3020, Rotorua, New Zealand

application are also discussed, including institutional barriers; implementation costs; certain molecular mechanisms underpinning variation; and modes of gene action such as epistasis and genotype-environment interaction.

10.1 INTRODUCTION

Many of the generic applications of association genetics described in this book apply to forest trees as well as other plant species. However, in implementation of association genetics plantation forest tree species differ from most other plant species, because of the unique combination of physical and genetic characteristics of forest trees, as well as the state of existing genomic information. This is especially so for breeding applications, as tree breeding is often very different from breeding of other plant species, particularly annual crop plants. In this chapter we focus on association genetics specifically in the context of tree breeding applications, in part because the most frequent use of association genetics may well be in the areas of selection and breeding. We describe where association genetics can be used in existing tree breeding programs, as well as new technologies under development, and discuss experimental components necessary for demonstration of concept and, ultimately, operational implementation. We also identify some potential limitations and challenges for successful implementation of association genetics in a tree improvement context.

In this section we provide relevant background to the chapter by reviewing the salient biological features of forest tree species, as well as the current state of knowledge of genomics in forest trees, and what is known about patterns of LD in tree species. We also introduce the term "gene-assisted selection" (GAS) used to denote the application of information from association genetics in a selection context, and compare and contrast this with marker-assisted selection (MAS), which uses information from marker-trait associations in pedigreed mapping populations for within-family selection.

10.2 DISTINGUISHING FEATURES OF FOREST TREES

Key features of most forest tree species include their large size and long lifespan; predominantly outbreeding behavior; slowness to express their phenotype as well as to reach reproductive maturity; and high levels of synteny within genera, and among conifers, within orders. The size and longevity of trees has both benefits and drawbacks. In terms of the latter, size can create major complications for both conventional breeding and the application of DNA polymorphism for selection. The complications involve both delayed expression of traits, and high costs of producing and managing the genetic material. For phenotypic selection, the delayed expression of traits may preclude effective selection for a number of years. It similarly affects any cross-referencing of phenotype with either genomic markers or QTN. The size of trees, along with the lifespan, means that field-testing trees is very expensive, either for a selection population in itself or for establishing relationships between phenotypic values and DNA polymorphisms. Unless the cost is accepted, which is a problem in itself, this in turn will tend to restrict both the potential selection intensity and the quality of information available on the relationships in question. In contrast, however, a key benefit of the long lifespan is the lasting presence of genotypes across years, even decades or centuries, almost "immortalizing" populations. Such a benefit can allow for repeated measurements over time on the same populations, further leveraging genotypic data, and/or allow for repeated DNA collections and therefore continued generation of genotypic data.

Trees are predominantly outbreeding, meaning that population-wide LD between quantitative trait loci (QTL) and neutral-marker alleles will generally be lacking, unless the base population(s) is/are quite strongly structured in one or more of certain ways. Such structuring will tend to be limited in wind-pollinated species, unless breeding populations are composed of recently admixed populations from distinct progenitor provenances or species. Interspecific hybrid breeding populations represent an example of this. Within species we would expect an absence of population-wide LD between QTL- and neutral-marker alleles; therefore LD will be confined to individual families, such that detection and quantification of QTL need to be undertaken independently for each family. Given the large population sizes needed for each family, unless there are extremely large QTL effects, this creates a very powerful incentive to develop GAS, based on establishing the effects of QTN. A key point here is that for among-family selection, which is common in tree improvement, marker–trait relationships ascertained via QTL mapping may have little or no predictive value for among-family selection.

Forest tree species are also frequently slow to reproduce, resulting in breeding generations within tree improvement programs that typically exceed a decade for conifer species, or much more for some angiosperms such as certain oaks (*Quercus* spp.). This contrasts with annual crop species such as corn, where two generations per year are possible in commercial breeding programs. A compensating feature of many tree species is that once reproductive maturity is attained, the numbers of seed produced can be very large, and seed production can last for decades, albeit seasonally, therefore facilitating generation of potentially large populations for experimental purposes. Furthermore, many tree species can be clonally reproduced, allowing for more precise estimation of genotypic value as well as allowing longer-term storage of specific genotypes.

An associated feature of forest trees is the high level of genetic load with deleterious effects of inbreeding (Williams and Savolainen 1996). Related matings for the most part greatly reduce fitness and frequently lead to phenomena such as embryo lethality that result in segregation distortion (Kuang *et al.* 1999), reduced rates of growth, and abnormal phenotypes (Williams and Savolainen 1996). Such effects, combined with the slow onset of reproduction, effectively eliminate the opportunity to develop homozygous lines, therefore populations used for association genetics and QTL mapping alike are typically heterozygous, and show strong variation both phenotypically and genetically.

A further, but mitigating, characteristic of forest trees is the high level of synteny among species, and even among genera, especially in conifers. The potential advantage of this is the leveraging of sequence information across species, as well as information regarding the functional role(s) of specific genes in trait variation. Furthermore, because different species frequently produce structures that are phenotypically very similar e.g., woody tissues), opportunities are enhanced for cross-referencing genomic information among species.

10.3 STATUS OF GENOMIC INFORMATION IN TREE SPECIES

Successful application of association genetics in forest trees, like all other species, requires considerable genomic information, either in the species of interest or in some highly syntenic species. Currently, forest tree species straddle the pre- and postgenome divide, with the majority (especially conifers) in the former. Recently, the full genome sequence of a poplar (*Populus*) has been determined, a first for a forest tree species (http://www.jgi.doe.gov/poplar). A further effort is currently underway in *Eucalyptus*

(www.ieugc.up.ac.za). Extensive EST databases have been developed for a number of species within these genera, although some questions have been raised about the level of EST representation; based on gene predictions, it is estimated that as much as 75% of genes are not represented in EST databases (unpublished results cited in Plomion *et al.* 2005). In conifers, most DNA sequence information is restricted to EST databases, which exist for a number of commercially important *Pinus* and *Picea* species, in addition to Douglas-fir (*Pseudostuga menziesii*) and a number of other conifer species. Most of these resources are publicly available (e.g., http://fungen.botany.uga.edu/Projects/Pine/Pine.htm, http://dendrome.ucdavis.edu/Gen_res.htm, http://web.ahc.umn.edu/biodata/nsfpine/), although some are proprietary. EST sequences have been determined using cDNA libraries constructed from a wide range of tissues, including developing xylem and cambium, roots, floral structures, and needles/leaves. For conifers, only a limited amount of nongenic sequence has been generated, and is usually associated with genic sequences (e.g., regulatory elements). While an increased amount of gDNA sequence data is likely, the large size of conifer genomes means it is unlikely that full sequence will be available within a short timeframe. Some technologies, such as sequencing Cot-based libraries and bacterial artificial chromosomes, may facilitate generation of a limited amount of genomic sequence data.

Linkage maps have been constructed for a wide range of tree species, primarily for the purposes of QTL studies (see Sewell and Neale 2000 and references therein), based upon a wide array of commonly used marker systems, including ESTs. A number of comparative mapping studies have also been undertaken (Devey *et al.* 1999; Echt *et al.* 1999; Chagné *et al.* 2003), elucidating the synteny referred to above, particularly among conifers. Linkage maps have been constructed for the most, if not all, commercially important forest plantation species, although applications in breeding programs have not been as widespread. However, relatively few studies have been undertaken evaluating synteny of QTL across species. One such study – comparing traits of adaptive significance in *Quercus robur* and *Castanea sativa* – found conservation of QTL for timing of bud burst but not for height or carbon-isotope discrimination (Casasoli *et al.* 2006). Telfer *et al.* (2006) reported nonrandom coincidence of QTL for wood density between *Pinus radiata* and *Pinus taeda*.

Linkage maps have been used extensively for QTL detection studies, mostly in full- or half-sib families. With the notable exception of disease resistance (e.g., Kinloch *et al.* 1970; Wilcox *et al.* 1996), the vast majority of QTL for commercially relevant traits appear to be of small effect only (Wilcox *et al.* 1997; Sewell and Neale 2000; Brown *et al.* 2003; Devey *et al.* 2004), indicating a large number of genes involved in variation of a particular trait. Implications for association genetics in an applied breeding context are that large population sizes will be needed to detect such QTL in sufficient quantity. This is discussed in more detail later in this chapter.

More recently, a range of gene expression technologies have also been applied in forest tree species, in particular microarrays (Kirst *et al.* 2004; Paux *et al.* 2004), and more recently reverse transcriptase polymerase chain reaction (RT-PCR), elucidating the level of gene expression in specific tissue types. This, coupled with EST databases and a suite of bioinformatics tools available, has generated much knowledge about the relative levels of gene expression, including both temporal and spatial variation in tissue of interest for a suite of genes. Such expression studies will be useful for selecting candidate genes for association genetics studies.

As with many other plant and animal species, however, the roles of genes in trait variation are largely unknown. To date, there are no reports of QTL having been cloned from forest tree species, partly due to the large number of candidates within QTL confidence intervals, but also because of the length of time required for trait expression of transformants arising from complementation studies, as well as the largely subtle effects expected for most QTL, together requiring considerable experimental resources to confirm complementation.

10.4 LD AND NUCLEOTIDE DIVERSITY IN FOREST TREE SPECIES

LD and nucleotide diversity, insofar as the latter governs functional variation, are the two key parameters for evaluating the efficacy of association genetics. To date, there have been relatively few extensive studies of LD in forest trees (see Gupta *et al.* 2005 for a review of LD in higher plants). Studies conducted in the 1980s with relatively limited numbers of polymorphic isozyme loci indicated limited or no LD, as would be expected in outbred species with relatively large effective population sizes. Mitton *et al.* (1980) found higher-than-expected digenic LD (6 out of 30 locus pairs) in *Pinus ponderosa*. Similarly, Roberds and Brotschol (1985) found evidence for age-related differences in the incidence of LD in *Liriodendron tulipifera*. Muona and Szmidt (1985) reported no evidence of LD in either pollen or megagametophytes in *Pinus sylvestris*. A study in *Pinus contorta* by Epperson and Allard (1987) showed higher-than-expected LD, but was limited to certain locus combinations, with some closely linked loci not in LD. Geburek (1998) also reported higher-than-expected digenic LD in *Picea abies*, although most were restricted to two or less subpopulations. In most of the aforementioned isozyme-based studies, nonrandom mating and/or selection on a limited number of loci were the most frequent explanations offered for higher-than-expected observed LD.

Studies with DNA-based markers have tended to reveal similar results. Bucci and Menozzi (1995) reported no LD in a small sample of *P. abies* using RAPD markers. A later study in *P. radiata*, involving microsatellite marker loci from a range of linkage groups, also indicated very little genome-wide LD (Kumar *et al.* 2004). More recently, a number of results from DNA sequence have been reported for conifers as well as *Eucalyptus* (Thumma *et al.* 2005) and *Populus* (Yin *et al.* 2004), surveying LD patterns in relatively small regions in and around expressed genes. Results to date generally indicate very short regions of LD, particularly in conifers where r^2 values tend to decrease to zero within a few hundreds to low thousands of base pairs (Table 10.1, and associated references), although there is considerable variability even within genes. Some exceptions have been noted in the average length of LD within genera; Yin *et al.* (2004) reported significant LD in regions around the *MXC3* resistance gene in *Populus trichocarpa* in the order of 16–34 kb. These results indicate that while on average the amount of LD is confined to relatively short spans in forest tree species, variations need to be taken into account, which can only be characterized via empirical data on genes of interest.

Table 10.1. Estimates of linkage disequilibrium and nucleotide diversity in plantation forest tree species based on DNA markers and candidate genes

Genus and species	Extent of LD	Metric(s)	No. of genes	Nucleotide diversity (Synonymous or not)		References
				yes	no	
Pinus radiata	No evidence between unlinked SSR markers	r^2	N/A	N/A	N/A	Kumar *et al.* (2004)
P. radiata	Not estimated	N/A	1	0.0300	0.0043	Cato *et al.* (2006)
P. radiata	Not estimated	N/A	8	0.0008	0.00005	Pot *et al.* (2005)
P. pinaster	Not estimated	N/A	8	0.0003	0.00015	Pot *et al.* (2005)
P. sylvestris	None observed within approx. 2 kb[a]	r^2	11	0.0056	0.0022	Dvornyk *et al.* (2002)
P. taeda	2,000 bp	$r^2 \sim 0.2$	19	0.0064	0.0011	Brown *et al.* (2004b)
Pseudostuga menziesii	1,000 bp	$r^2 \sim 0.1$	18	0.0105	0.0021	Krutovsky and Neale (2005)
Picea abies	100 bp 200 bp	$r^2 \sim 0.2$	?	Not provided	Not provided	Unpublished results cited in Rafalski and Morgante (2004)
Eucalyptus nitens	"Similar results to maize and *Pinus*"	r^2	1	Not estimated	Not estimated	Thumma *et al.* (2005)
Populus trichocarpa	Up to 34 kb	Not provided	1	Not estimated	Not estimated	Yin *et al.* (2004)
Populus tremula	<500 bp	$r^2 < 0.05$	5	0.0220	0.0059	Ingvarsson (2005)

[a]Analyses based on one gene only.

Nucleotide diversity in forest tree species appears to be variable both among and within species. In most conifers, typical reported values range between ca. 10^{-2} and 10^{-4}, with some variation within species (Krutovsky and Neale 2005). Overall, forest trees appear to show more such diversity than humans, but slightly less than that observed in species such as maize (Brown *et al.* 2004b). Diversity appears to be lower in coding sequences, with nonsynonymous substitutions being less frequent than synonymous substitutions, although rarely are such differences reported as being statistically significant – for example, Brown *et al.* (2004b) found no evidence for selection in 19 genes in *P. taeda,* while Krutovsky and Neale (2005) reported evidence for selection in *P. menziesii* in three of 18 expressed genes. Cato *et al.* (2006) reported evidence for selection in a putative dehydrin gene in *P. radiata*, and found weak associations with the same gene and wood density and growth rate.

The moderate nucleotide diversity, coupled with the typically low LD per base pair, indicates a relatively high number of haplotypes per genic region. For example, Krutovsky and Neale (2005) found that there were approximately 2–3 haploblocks per gene, thus on average, 4–5 single nucleotide polymorphisms (SNPs) would be needed to adequately cover most single genes for association genetics applications.

What is the significance of these results for association genetics in conifers? Firstly, the observed levels of nucleotide diversity indicate there is sufficient polymorphism for association genetics studies. Secondly, the relatively small regions of LD give some cause for optimism regarding functional assignment, as the small regions of LD observed within most genes indicate the possibility of implicating genes (or even small regions within, or associated with, genes) in trait variation. The disadvantage is that relatively

detailed studies will be needed, typically assaying many polymorphisms in regions of interest, necessitating judicious targeting of candidate regions to limit the number of genes to be screened. Such detailed studies are costly and time-consuming, particularly if applied breeding is the key objective. However, short stretches of LD mean there is some potential for using association genetics to assign putative function to genes, and will be of use to those seeking to determine molecular mechanisms underpinning phenotypic variation.

To date, relatively few results have been reported from association genetics experiments, although this should change. Kumar *et al.* (2004) found only weak evidence for association between polymorphic SSR markers and a number of traits in a small female-tester mating design in *P. radiata*. Since then, Brown *et al.* (2004a) reported a putative association between an SNP within an α-tubulin, and earlywood microfibril angle, a key component influencing performance of structural-grade timber in conifers. More recently, Thumma *et al.* (2005) reported an association between polymorphism encoding a putative splice-site variant in a Cinnamoyl CoA Reductase (*CCR*) gene in *Eucalytus nitens* and microfibril angle. A mutation in the putative functional homologue of this gene in *Arabidopsis thaliana* proved to cause the *IRX4* phenotype (Jones *et al.* 2001). Cato *et al.* (2006) reported an association in *P. radiata* between polymorphisms in a putative stress-response gene, with both wood density and growth rate in large association population.

10.5 GENE-ASSISTED SELECTION VERSUS MARKER-ASSISTED SELECTION

One of the key features of outcrossing species such as forest trees is the expectation of widespread linkage equilibrium within unstructured populations, and conversely, the expectation of strong LD within specific pedigrees. The latter has been extensively utilized to date in the field of QTL mapping based on pedigreed populations (usually full-sib families), leading to the development of linkage maps for a wide range of species and demonstration of the potential for within-family MAS. This approach, however, has various limitations, including the restriction of selection to within specific families for which the marker allele–trait associations have been previously established (Strauss *et al.* 1992; Johnson *et al.* 2000; Wilcox *et al.* 2001).

From a tree breeding perspective, the key feature of association genetics is the opportunity to select both among and within families, by establishing relationships between polymorphisms and heritable trait variation outside of any family structure. However, because LD is restricted to relatively small chromosomal regions in forest tree species, we consider that the most likely polymorphisms to be associated with trait variation are those within, or associated with, expressed genes. For this reason we use the term “gene-assisted selection” to denote the application of within- and/or among-family selection based on polymorphisms shown to be associated with trait variation in unstructured populations, i.e., association genetics.

The idea of selecting genotypes based on DNA sequence variation is not new – the concept of MAS is indeed based on the same principles, i.e., selecting on the phenotypic-, and/or discrete isozyme-, and/or DNA-sequence variants that are correlated, through linkage, with phenotypic variation in commercially relevant traits. There are key differences between MAS and GAS, however (Table 10.2), from perspectives of both research and operational implementation. Here, the terms GAS and MAS are used primarily to define differences relevant to typical forest tree breeding; we refer to MAS

as a technology for within-family selection only, in contrast to GAS, where selection can in theory be applied at the family level, in addition to individual genotypes within families, without prior pedigree information. These differences are not trivial with respect to the objectives and design of the underlying experiments needed to detect and quantify marker–trait associations. For example, for MAS, marker–trait associations are generally detected using pedigreed mapping populations, thereby maximizing linkage disequilibria between neutral markers and QTL that control detectable proportions of the phenotypic variation. For GAS, researchers basically accept and work with the existing levels of (dis)equilibria, however incomplete, that prevail in populations within which there are no recognized patterns of interrelatedness. Marker systems are likely to differ also, although in limited cases there may be some overlap. For MAS, selectively neutral marker systems adequate for development of moderate-density linkage maps and high-throughput (HTP) genotyping are considered satisfactory (e.g., RAPDs, AFLPs, microsatellites). For GAS, however, we consider it is more likely that polymorphisms associated with candidate gene sequences, i.e., SNPs, and insertions/deletions (indels), would be the marker systems of choice.

Table 10.2. Comparisons of requirements for MAS based on QTL detection and GAS based on association genetics in a tree breeding context

Attribute	MAS	GAS
Detection goal	Quantitative trait *locus* – i.e., chromosomal regions within specific pedigrees within which a QTL is located	Quantitative trait *nucleotide* – i.e., maximize causative sequence(s)
Genomic resolution	Low – moderate density linkage maps only required	High disequilibria within small physical regions usually needed (<2 kb) Linkage disequilibrium experiments: unrelated
Experimental design for detection	Defined pedigrees, e.g., three and two generation pedigrees/families, half-sib families	individuals (association tests), or large numbers of small unrelated families (transmission disequilibrium tests, TDTs)
Applicable to	Within-family forwards[a] selection only, within specific families where associations detected	Plus-tree selection, among- *and* within-family forwards selection, within reference population[b]
Marker neutrality	Neutral	Non neutral
Marker specificity	Non-trait-specific	Trait-specific[c]
Marker discovery costs	Moderate	Moderate for few traits, high for many traits
Prescreening[d] for functional association required?	No	Yes[e]
Opportunity to identify co-adapted gene complexes	Moderate	Good
Number of markers required	200–300 codominant markers per genome on average	>5 prescreened markers per gene on average, likely >5 genes per trait

[a]Selection among latest generations of breeding-population off spring.
[b]As defined by populations used for detection experiments.
[c]Except when polymorphism is in disequilibria with gene(s) controlling more than one trait.
[d]Prescreening defined as the need to select candidate sequences based on some *a priori* expectation of association or causation (e.g., candidate genes).
[e]Assuming lack of genome-wide, ultra-high density marker maps.

10.6 GENERIC BENEFITS OF GAS

Stromberg *et al.* (1994) classified generic benefits relating to the use of DNA markers for selection into three areas: earlier selection; cheaper, more cost-effective selection; and increased selection intensity. In the context of GAS in a tree improvement program these also apply, but for the sake of completeness, can be expanded. The following, partly overlapping areas are where we consider most of the potential benefits will be:

(1) *Earlier selection*. Perhaps the single most important limiting factor in plantation forest tree improvement has been selection age. The vast majority of characteristics do not adequately express their genotypic value until one-quarter to one-half of rotation age, which is a key factor influencing the long generation intervals typical of most tree breeding programs. GAS, like MAS, offers the tantalizing prospect of selecting at an emergent seedling stage, rather than waiting for up to many years for adequate trait expression. Such early selection can be used as a substitute for direct phenotypic selection, or as a complement in a multistage selection procedure, or simultaneously with information on phenotype. The net effect will be to increase selection intensity (see (3)). A further benefit, particularly in the cases of plus-tree and among-family selection, is the prospect of screening individuals without need to generate and evaluate offspring, which will further reduce generation interval by directly evaluating genotype.

(2) *Cheaper, more cost-effective selection*. Knowledge of the sequence variants and their effects on phenotype offers opportunity to select based on sequence only which could reduce or perhaps ultimately eliminate need for field screening. Field testing is one of the most expensive components of tree breeding programs, and sequence-based selection is likely to be cheaper, particularly for multitrait breeding objectives where expensive-to-measure traits such as wood properties are involved. Furthermore, advances in DNA technologies offer further reductions in costs in the medium term, whereas phenotypic measurements are likely to remain relatively expensive. One factor to consider, however, is the reasonably high cost of establishing marker–trait associations, which means that a large-scale breeding operation may be needed to justify use of GAS. Nonetheless, these costs can be reduced through various means such as pooling DNA samples (e.g., Germer *et al.* 2000). Moreover, the associations are expected to hold across a number of generations, so costs can be spread accordingly provided generation intervals are short. However, sample sizes necessary for detection of marker–trait association in LD populations may require at least several thousand genotypes for small-effect QTL for even modest levels of power and ability to infer association (Ball 2005; Chapter 8).

(3) *Increased selection intensity*. This can result partly from the low cost of producing young propagules that can be screened by GAS and partly from the higher-throughput evaluation capacity. Even with current moderate- to high-throughput genotyping technologies, there is capacity to screen far more genotypes than can be field-tested, at potentially much lower cost. Thus, genetic gains are likely to increase, particularly with multitrait breeding objectives that will tend to require larger numbers of selection candidates. In fixed-resource phenotypic-screening programs the addition of another trait into a breeding objective will typically incur costs in gain for any single specific trait unless the “new” and “existing” traits are strongly

and favorably correlated. Such a cost can be reduced with increased selection intensities, but in contemporary breeding programs this usually means more phenotypic evaluations (often on progenies) and possibly introduction of new genotypes into breeding populations. GAS could be used as a surrogate selection tool in these situations, although there may be the challenge of establishing the requisite associations simultaneously in several traits.

(4) *Reduced need for phenotypic selection*. The combined result of selection that is cheaper and/or earlier and/or more intensive may mean, in theory, at least, that GAS could ultimately replace phenotypic selection. This is based on the intriguing possibility that concomitant advances in genomics, proteomics, and metabolomics could eventually lead to development of predictive models that integrate information on gene sequences with information on environmental influences to predict phenotype, thus reducing reliance on phenotypic selection, and basing genetic selection entirely upon DNA sequence. The reduced reliance on field testing has several distinct advantages, including a reduction in costs and/or a concomitant increase in effectiveness of a tree improvement program via reallocating financial resources to other components of the operational program. Field testing is one of the most costly items in a tree improvement program, not just in terms of data collection, but also trial establishment and maintenance, and to a lesser extent, analyzing data and maintaining records. While the need for various forms of field experiments will likely persist even once all genes are sufficiently well characterized with respect to effects on trait variation (e.g., genetic gain trials), significant cost reductions should become possible.

(5) *Increased flexibility for operational evaluation and selection of genotypes*. Knowledge of phenotypic value associated with specific DNA sequences that can be applied across unrelated genotypes expands the scope of potential application. GAS can be applied to plus-tree selection, as well as among- and within-family selection, in contrast to MAS, where associations between marker alleles and trait variability are family-specific, and are thus applicable to within-family selection only. Therefore, in theory a genetic value can be placed on any specific individual based on DNA sequence information, where sequence has some nonzero association with trait value. While implementation of GAS would lead to more field trialling initially because of the need to find sufficient associations between markers and traits, ultimately, GAS could reduce need for "common-garden" testing, and allows new introductions to be evaluated without progeny evaluations (as is typically practiced).

(6) *Complementary/synergistic fit with both existing and new genetics technologies to enhance genetic gains*. Because various genetic technologies are available for forest tree improvement, in addition to an array of new technologies currently being developed, there are typically alternative routes to delivery of genetic gain. GAS potentially offers an additional technological route, in that it can either complement – or possibly supplant – phenotypic selection, but in addition, fits well with newer technologies. We describe in more detail in Section 10.9 the fit with new biotechnologies.

(7) *Prediction of genotypic value and enhanced opportunities for optimizing combinations of genotypic, site, and silvicultural characteristics*. Eventually, the knowledge of DNA sequences underpinning heritable variation could be combined with knowledge of key environmental and silvicultural influences to predict phenotypic characteristics. While this is a far-reaching goal, it is a tantalizing

possibility that knowledge of the causative nucleotides in combination with the extent to which environment affects the roles in particular characteristics could be combined to design combinations of genotypes and silviculture that optimize returns to forest growers. Such a capability would be extremely beneficial for designing genotypes with particular characteristics in mind, and would also aid silviculturists in designing genotype-specific regimes to maximize value, as well as a more optimal matching of genotypes to sites.

(8) *Provision of experiments that could ultimately lead to identification of actual QTN*. Because GAS typically develops initially from correlation rather than causation, identification of causative QTN may not be necessary for selection. However, the candidate genes and experiments necessary for identifying which polymorphisms are associated with trait variation (described below) are also necessary components for identifying the actual QTN. This knowledge is a key step in elucidating molecular mechanisms underpinning quantitative variation, and this information could be used to design new strategies for creating and utilizing variation, by identifying genomic regions that, when further altered, could lead to creation of additional useful variation.

(9) *Provision of experiments to answer questions about the genetic structures of forest tree populations and provide key information that could assist in management of breeding populations*. A benefit of the experimental infrastructure established for association genetics is the opportunity to generate genome-wide information that could be used to elucidate genetic phenomena such as presence/absence of trait variation, population structure and history, and evidence of selection. The genetic architecture of trait variation can be defined as the frequencies, location, magnitude, and mode(s) of action of QTL/N effects underpinning quantitative traits. While QTL mapping has been very informative in this regard, the results are relevant only to the pedigree(s) used, rather than to whole populations. Association genetics may therefore be more relevant for understanding the genetic landscape of trait variation in forest trees. While large, essentially panmictic populations cannot be expected to have appreciable across-family linkage disequilibrium, cryptic structuring may exist which generates significant disequilibrium. For example, localized population bottlenecks, followed by coalescences, could easily cause this. Such LD could provide valuable clues to "metapopulation" history. Despite wind pollination, various factors can generate population structure in conifers (Mitton 1992). Interesting possibilities of structure exist in populations derived from recent admixture. In *P. radiata*, the exotic, domesticated "land races" still have large elements of the wild state. Interestingly, they evidently represent a genetically recent fusion of two of the native populations, Año Nuevo and Monterey (Burdon 1992; Burdon *et al.* 1998), which may provide a basis for some admixture disequilibrium. A further benefit of association genetics is that DNA sequence data derived from both genic and nongenic regions can reveal much about genetic history of those regions. Departures from Hardy–Weinberg equilibrium could reveal presence of previously undetected genetic phenomena such as presence/absence of inbreeding. Indeed, genetic variance (and gain) estimates are based on assumptions regarding relatedness of parents used in genetic tests. Such data can be used to check these assumptions and provide empirical data for more accurate estimates. Similarly, sequence data from genic regions can reveal evidence of selection (see Section 1.3 for recent examples in forest trees). Such evidence – which can be generated on a

relatively small subset of genotypes – could be an effective prescreen for genes more likely to be associated with trait variations, although some caveats apply regarding power to detect effects of selection (Wright and Gaut 2005).

10.7 PREREQUISITES FOR FEASIBILITY

10.7.1 Basic Prerequisites for Operational Implementation

Successful application of association genetics for forest tree breeding must depend on the context of a well-structured breeding program. Genetic variation for economic traits is essential, and must be proven, while important genetic correlations between different economic traits need to be at least reasonably understood. Achieving this will entail major progress towards obtaining the populations needed for detecting associations between DNA polymorphisms and phenotypes. Efficient assays, which can be used on young trees, are important for this purpose, just as they are for conventional breeding. This will generally require new measurement technology, and/or easily measured juvenile traits that are good proxies for harvest-age economic traits. For wood quality, the SilviScan instrument (e.g., Evans 1994; Evans *et al.* 1999) has been developed to measure several detailed anatomical properties, and this has been complemented by an improved understanding of how such properties affect processing- and product-performance characteristics. Resistance to certain diseases can be assayed by inoculation trials of young seedlings (e.g., Powers *et al.* 1982). Very early evaluation for growth rate, however, can be very problematic: juvenile–mature correlations can be low, physiological variables can show highly nonlinear relationships with performance, and metabolite fluxes can be far more important than metabolite concentrations.

More specific requirements for applying GAS include quantitative capabilities, both in providing appropriate material to furnish phenotypic data and in managing, analyzing, and interpreting phenotypic and genomic data; access to HTP genotyping technologies; and good marker selection. This involves selection of candidate genes that could be associated with quantitative variation, and discovery and evaluation of important polymorphisms. We discuss each of these requirements.

Operational implementation will depend not only on meeting the various technical conditions listed above, but also on meeting organizational and even institutional requirements. Between the tree breeders and the genomic scientists there need to be close communication and considerable mutual education. Allocation of resources to the various parties will be a continuing challenge. A further challenge will lie in maintaining a strategic focus, whereby GAS and other new technologies can be used to best long-term advantage.

The total scale of undertakings for successful development and application of GAS will typically require collaboration between institutions, including industry, specialist research organizations, and universities. This will need to be achieved in the face of a climate of competitive bidding for research funding and the various pressures to appropriate Intellectual Property for individual organizations' own gain.

10.7.2 Quantitative genetic skills for experimental design and analyses

Effective application of association genetics for selection applications also requires both good experimental designs and analytical skills so that sufficient numbers of QTN

can be detected and utilized. These issues are discussed in more detail elsewhere in this book (Chapters 7 and 8). We cover components relevant to application of association genetics for tree breeding.

10.7.2.1 Experimental Design

A key prerequisite for GAS is the identification of DNA polymorphisms for selection. But what kind of experiments and what analytical methods are necessary? This has been covered to some extent in Chapter 8, so here we confine our discussion to issues relevant to tree breeding.

One of the few benefits of tree breeding is that unstructured (or loosely structured) populations already exist due to the nature of breeding programs, which usually consist of breeding populations with moderate numbers of heterozygous genotypes that show considerable genetic variation, despite being subject to phenotypic selection as a prerequisite to introduction in breeding populations. Moreover, such populations have usually been extensively progeny-tested, sometimes with clonally replicated progenies, for which phenotypic records have been generated for a range of commercially important traits. In addition, most programs maintain reasonable records of the geographic locations of the original first generation selections, as well as good knowledge of the range of genetic diversity represented in the naturally occurring populations – which may or may not contribute to breeding populations.

However it is necessary to bear in mind what information is needed from association tests that could be of use to breeders. Firstly, sufficient numbers of markers associated with QTL/N are required to obtain worthwhile genetic gains, implying the necessity for moderate–high power of detection of QTL/N. In addition, the genomic location of these polymorphisms and their magnitudes of effect, as well as modes of gene action and population allele frequencies, are also key pieces of information. Furthermore, it is necessary to account for population structure, as the impact of population structure can affect both the validity of any detected associations (Pritchard and Rosenberg 1999) as well as the estimates of gene-substitution effects (Deng 2001). Methods are available to do this (Pritchard *et al.* 2000; Thornsberry *et al.* 2001; Yu *et al.* 2006), and some experimental designs can account for such admixture (Allison 1997; Wu *et al.* 2002). In the relatively few studies undertaken to date there is very little evidence of population structuring in forest trees – no evidence was found in Douglas-fir (Krutovsky and Neale 2005) or loblolly pine (Brown *et al.* 2004b) which are both wind-pollinated conifer species, nor in *E. nitens* (Thumma *et al.* 2005). However, population structure has been indicated for other species. For example, Lagercrantz and Ryman (1990) reported presence of structure among populations of *P. abies* based on both allozyme and morphological (but not genecological) variability, a result at least in part, of population disruption during the most recent glaciation.

In order to determine appropriate experimental designs, it is germane to briefly review what is known about the genetic architecture of traits of commercial value in forest tree species. Numerous studies have been conducted using QTL mapping populations usually involving full- or half-sib families for most forest tree species of commercial value. For traits such as disease and insect resistance there are well-documented examples of major genes (Devey *et al.* 1995; Wilcox *et al.* 1996) although it is unlikely that resistance to pests and pathogens is solely conferred through major genes alone. For quantitatively inherited traits, which appear to be the norm for the majority of

commercially important traits, there has been some debate regarding the true nature of the underlying variation. Early studies involving relatively small populations indicated genetic variation was dominated by a few genes of moderate effect, however, these results were difficult to repeat, even in the same families (Wilcox *et al.* 1997; Sewell and Neale 2000). Interpretations of those early studies may therefore have been erroneous in that results are also consistent with genetic architecture involving genes of small effect only, similar to that described in corn (Beavis 1994), and subsequent verification, when done, have indicated this to be the case (Wilcox *et al.* 1997; Sewell and Neale 2000; Brown *et al.* 2003). Therefore for most traits, we contend that the underlying genetic architecture is most likely to be dominated by genes of relatively small effect contributing a few percent of the variation at most (e.g., Devey *et al.* 2004). An exception may be that interspecific hybrids could involve genes of moderate–large effect (e.g., Bradshaw and Stettler 1995), although small-effect genes may also have a role. Experimental designs for association genetics will therefore need to be cognizant of these architectures, particularly genes of small effect, if selection is going to be effective.

A number of different experimental designs could be used to detect associations between QTL/N and polymorphisms, such as an unstructured population consisting of putatively unrelated (or distantly related) genotypes; or combined with information on progeny (analogous to a TDT design, except using quantitative traits); or alternatively a hybrid QTL–LD population (see Chapter 8 and references therein). Some of these approaches have been evaluated in a manner more relevant to forest trees (e.g., Wu *et al.* 2002; Ball 2005). Furthermore, some of the genetic characteristics of forest trees parallel humans (e.g., high levels of heterozygosity, adverse effects of inbreeding, longevity), for which much has been written in regard to the theory and efficacies of specific experimental designs and analytical procedures, and are therefore relevant to tree species. We review some of this literature here, and refer the reader to Chapter 8 for a more extensive review.

A number of theoretical studies have been conducted, particularly in comparing designs with and without use of information from sibs. A somewhat unclear picture has emerged to date, however, partly because of differing assumptions and input values used for simulations. Long and Langley (1999) showed that for smaller-effect QTL (~5% of phenotypic variance), unstructured or random populations were more powerful than TDT-based designs, and that power increased more when greater numbers of individuals rather than markers were used. Moreover, they concluded that unstructured populations sample sizes ≥500 individuals would suffice to detect small-effect QTL assuming a Type-1 error rate of 0.05. A further and nontrivial finding was that equally large populations would be needed to verify any detected associations.

Wu *et al.* (2002) developed theory for combined linkage- and linkage-disequilibrium mapping, based on use of genotypic information from a single parent combined with genotypic and phenotypic information from offspring, analogous to multiple half-sib families, as in often used in breeding population testing. They compared different combinations of family numbers and sizes, and compared the power to detect a segregating QTL of large effect with an unstructured population without information from progenies. In contrast to Long and Langley (1999), they found that simulation results indicated that use of information from progenies was more powerful than unstructured populations only, particularly with low disequilibrium, assuming the same number of individuals genotyped. Results also indicated that few families with many offspring per family were more powerful than many families with few offspring. A key benefit of this

approach is that the use of progenies obviates the need to independently evaluate population structure. However, because these results were based upon a single QTL with a large effect (both additive and dominance terms equal to residual error), relevance of these results may well be limited, as individual QTL effects are typically much less than residual variance. Therefore these results would need more careful evaluation using a range of QTL effects more relevant to known genetic architectures.

Most of the above studies have involved estimating power with comparison-wise Type-1 error rates in the region of 0.01–0.05. However, such values may be problematic in reality because actual results in that range of *P*-value may not be equate to strong evidence for an association. Using a Bayesian approach based on theory originally developed by Luo (1998), Ball (2005) calculated that *P*-values in the range of 0.01–0.05 actually represented weak evidence against an association for sample sizes in the 432–1,200 individuals in an unstructured population. *P*-values in the range of 10^{-4} would be more indicative of evidence for an association, assuming high prior expectation for an association (see Chapter 8). This also implies that larger sample sizes than those generally reported above would be needed.

Ball (2005) also showed that very large sample sizes are necessary for high power (0.9) of detection of QTL with small effects (explaining 1–5% of total variance) when using either candidate genes or a genome scan in an unstructured population. To obtain high power with strong posterior odds (Bayes Factor >20) with moderate disequilibium ($D' = 0.1$), sample sizes ranging from 6,800 to 40,100 would be necessary to detect QTL of 5 and 1% effect, respectively. Such sample sizes are based in part on relatively low prior odds, which may be increased through generation of additional experimental and biological information on specific genes (e.g., expression profiles, evidence of selection), therefore sample sizes could be reduced. However, even with relatively high prior odds, sample size requirements will still be relatively high. Furthermore, Ball (2005) quantified the power to detect QTL when marker and QTN frequency differed. Even with very large sample sizes (19,200 and 38,400 genotypes), there is relatively low power to detect rare QTN with intermediate marker allele frequencies, even when in almost complete disequilibria. This is an important consideration, given that long-term genetic gains are driven by low-frequency QTN, along with mutations that arise during the selection period.

What can be concluded regarding optimal experimental designs based upon the work described above, and what are the implications for tree breeding programs? Firstly, moderate- to large-effect genes are likely to be easily detected using material from existing breeding populations, as long as there are sufficient numbers (200–1,000 putatively unrelated genotypes with phenotypic records available). For smaller-effect genes, which are likely to dominate the genetic architecture of quantitative traits in particular, much larger sample sizes are likely to be needed; therefore augmentation of existing breeding populations with genotypes from natural populations may be necessary. The implication here is that such augmentation will require common-garden experimentation, which is time-consuming, and could delay or militate against use of association genetics. Furthermore, maintenance of genetic diversity of nonbreeding population genotypes is also a necessity. Optimal designs with sufficient power for detection of small-effect QTL will therefore need to be ascertained in the context of tree improvement programs, most likely necessitating numerical simulation on a case-by-case basis.

Experimental designs could nonetheless be incorporated into tree improvement programs even if additional genotypes are necessary: such populations will be useful for other purposes (such as parameter estimation for new traits), particularly if progenies are incorporated. Indeed some redesign of breeding strategies may well be necessary if genetic tests are to take effective advantage of association genetics.

10.7.2.2 Analytical Methods

A further requirement for successful implementation of GAS is the use of appropriate methods for analyzing results from association tests. Some parameters such as population structure, linkage disequilibrium, and evidence for natural selection are estimable from sequence data generated on a small subset of genotypes, which could be used as a prescreen for a larger association test. For the latter, it would be necessary to use only those polymorphisms not in LD with other polymorphisms ("haplotype-tagged" polymorphisms), which would be determined in such a prescreen.

A number of analytical approaches could be used, depending on the experimental design. For most experimental designs, population structure will need to be tested for and, if present, taken into account. After examination of evidence for population structure, a number of parameters need to be simultaneously estimated for effective application. These include gene-substitution effects, population structure, frequency of both marker and QTN, mode(s) of gene action, and genotype × environment interaction (if present). Methods for estimating such parameters are discussed more fully in Chapter 8.

Preliminary analyses for detection of marker–trait associations can be undertaken using simple regression or ANOVA-based approaches, which can be undertaken in a variety of software packages. Specific software such as PowerMarker (www.powermarker.net) and TASSEL (www.maizegenetics.net) can also undertake limited analyses. While these may be useful for indicating a potential association, more detailed analyses are required for adequate statistical inference and gain estimates. While maximum-likelihood methods have been developed to estimate key parameters (e.g., Wu and Zeng. 2001), the estimates tend to be 'prone to selection bias' if the same data are used to estimate parameters as well as detect associations (Ball 2001), and thus may be unreliable. Overestimation of some gene-substitution effects has been reported for QTL mapping (Beavis 1994; Ball 2001). Bayesian methods may be more appropriate here (see Chapter 8). Methods to reduce or eliminate selection bias have been developed for QTL mapping in pedigreed populations (e.g., Ball 2001), and extension to commonly used experimental designs for association genetics may be useful.

A further consideration for the experimental design is the actual nature of the molecular data. Data can come from haplotypes (such as directly sequencing each copy of a gene in the diploid genotypes, or genotyping haploid tissue), or directly obtaining marker genotypes at each polymorphic site without surrounding sequence information. The key difference here is that with haplotypic data, the phase relationships between polymorphic sites are known for each copy of a polymorphic region in an individual. In contrast, for marker-genotype data, phase relationships are not known. Haplotypic data are considered, by some, to be more powerful for detection of marker–trait associations, as information from multiple polymorphisms can be condensed into discrete haplotypic classes (e.g., Lynch and Walsh 1997). Long and Langley (1999) found that marker-based methods were as powerful if not more powerful in some situations, than "simple"

haplotype-based methods, and simulations suggested lower Type-1 error rates. Genotypic data are sometimes cheaper to obtain, as direct sequencing is not necessary.

10.7.3 Access to Appropriate Genotyping Facilities

HTP facilities are necessary for sequencing and genotyping, for both detecting associations and operational selection. Extensive sequencing and resequencing are required, even if only a small subset of genotypes are used for initial scans of candidate gene regions. HTP genotyping is an obvious prerequisite, given the large amount of data generation necessary for adequately conducting powerful association tests. Whether or not specific breeding programs choose to develop "in-house" capacity or choose to outsource this component will be a choice made on a case-by-case basis.

10.7.4 Marker selection

Appropriate marker systems are an obvious prerequisite for detection of associations between marker and trait variation, along with HTP genotyping for selection purposes. But how many and what types of markers are needed for association genetics? Requirements for association genetics and subsequent selection applications differ substantially from those for QTL mapping (Table 10.2), primarily because disequilibrium per base pair is likely to be substantially less for apparently unstructured populations versus pedigreed QTL mapping populations. Forest trees present specific problems here. The outbreeding behavior, in particular, means that regions of LD tend to be very small, typically in the range of 0.3–2 kb (Table 10.1). In addition, gymnosperms in particular have typically large genomes (Murray 1998) adding further complications.

Several approaches could – at least in theory – be used to select polymorphisms to detect marker–trait associations. These include:

- *Use of the same markers as those developed for QTL mapping*, for example, SSR and EST markers. For most forest tree species, total number of markers used for linkage and QTL mapping is generally in the range of several hundred to low thousands, and therefore insufficient to achieve adequate resolution for association mapping given the typically small stretches of LD. Moreover, many of these loci are likely to amplify phenotypically neutral regions of the genome, or at least do not appear to be strongly correlated with trait variation even in specific pedigrees where disequilibrium is much greater, so such markers are unlikely to be adequate for association genetics. Nonetheless, these markers can be useful for revealing population structuring, which needs to be taken into account in association tests. It is also possible that a small number of loci could be in disequilibrium with QTN.
- *Whole-genome sequencing* (and resequencing), such as that undertaken in humans and a small number of important domesticated animal species. This involves complete (or near-complete) genome sequencing, followed by *in silico* polymorphism identification, after which a subset of polymorphisms are chosen for whole-genome scanning based on the patterns of observed disequilibrium. Such an approach is costly and technically challenging with existing sequencing technologies in highly repetitive and large genomes such as gymnosperms. For example, in *P. radiata*, assuming a 1C genome content of 22×10^9 bp (Murray 1998), with a 1,000 bp haplotype block size on average, we calculate that 22

million markers would be needed. To genotype a 1,000-tree population at a cost of US 5 cents per marker per genotype, would cost in excess of US $1 billion! Even for the smaller angiosperm genomes, assuming 1% of the size of the above example with a similar haploblock size, cost is still well beyond the reach of most tree breeding programs.

– *Partial genome sequencing* (and resequencing) of specific (rather than entire) genomic regions. This is a more limited approach than that described above. Genomics technologies targeting gene-rich areas such as Cot-based selection methods, which target low-copy-number regions, may be an alternative to whole-genome sequencing/resequencing. Such an approach may be more financially acceptable, particularly for hardwoods which have smaller genomes than conifers. Further research is needed, however, to determine if such methods could be effective at targeting QTN, as success of this approach is predicated on whether or not the QTN are located mainly in either low-copy-number regions or regions of low methylation. Genic regions within such "short" genomic stretches would need to be identified, which could be done using gene-searching algorithms, and/or alignment with relevant EST databases. QTN discovery via this method could still be expensive, however, as high polymorphism rate and low LD per base pair mean that SNP discovery would be expensive. Moreover, gene families could further complicate this approach in the more complex, larger genomes such as in conifers.
– *Preselection of candidate genes*, followed by polymorphism discovery, within these genes as well as the surrounding regions. We consider marker selection using this approach more promising than all of the above, primarily due to cost. With this approach, nucleotide variants within the transcribed sequence and the surrounding regulatory regions would then be assayed for association with trait variation. Such candidates could be selected using various approaches, which are described in more detail in the following section.

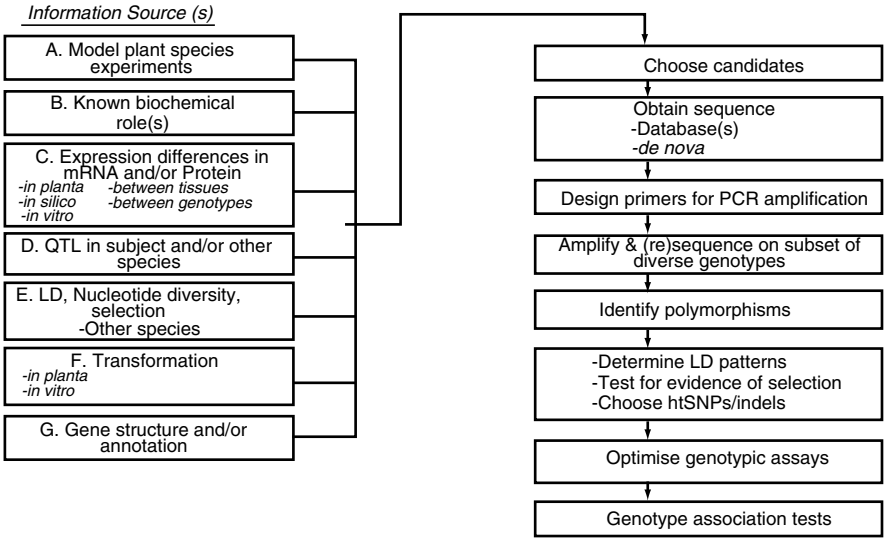


Figure 10.1. Generic process for selecting candidate genes and generating polymorphism information on association tests.

The overall process from gene selection to generation of genotypic data on association tests is described in Figure 10.1. It should be noted that this process assumes population structure has been already evaluated.

10.7.4.1 Candidate-gene selection

Generic methods for candidate gene selection are described in more detail elsewhere in this book. Here, we outline more specific approaches that could be considered, noting that except for *Populus* and *Eucalyptus*, there will be very little genome-wide data available for subject, although for most commercially important genera extensive EST sequence information is available, if not in the species of interest, then in a closely related species. Note, too, that selection of candidate genes can be based on more than a single criterion, although the relative efficacies of the various criteria are not yet known. Such criteria include:

- Choosing orthologous genes to those in model plant species that have been shown to have a role in traits of interest (Figure 10.1, Box A). For example, Thumma *et al.* (2005) found that polymorphism in an intronic region of a *CCR* gene was statistically associated with microfibril angle in *E. nitens* in a small association population. This gene was chosen because it is homologous to the *IRX4*-causing *CCR* in *A. thaliana*. However, it is not yet known to what extent and which plant model systems can predict roles of the homologous genes governing endogenous variation in forest tree species. If, in the more complex conifer genomes, there is a greater tendency for large gene families affording some degree of functional redundancy, information from short-lived angiosperms could be of limited value.
- Similar to the above, but using information on mutations and knowledge of gene sequences (and expression patterns of the sequences) from other forest tree species. For example, while an annual-plant model system could have limited applicability, a model system based on a woody perennial (e.g., *Populus*) could be more useful. In either case, the role of comparative genomics is crucial.
- Endogenous genes based on known or suspected role(s) in relevant biochemical pathways (Figure 10.1, Box B), e.g., genes involved in lignin biosynthesis as a preliminary choice to investigate natural variation in lignin chemistry. Much molecular information has been generated on this topic, and the key regulatory genes have been identified (e.g., Huntley *et al.* 2003). Such an approach has been used in mammalian systems, although with mixed success. For example, the Booroola gene in sheep (*FecB*), which causes elevated fecundity, was initially thought to be due to natural variation in FSH, a gene encoding a follicle-stimulating hormone. However, subsequent linkage analysis showed otherwise (Dodds *et al.* 1993), which was later verified by identifying the causative gene.
- Information from transcript profiling (Figure 10.1, Box C), identifying genes whose expression patterns are correlated with specific traits. A number of differential-expression technologies have been developed, including micro-arrays, cDNA–AFLP and similar approaches, and are now extensively used, although not as tools in breeding programs. Such technologies do reveal many candidates – possibly too many to be used as a screening tool alone. Moreover,

heritable variation may arise for reasons other than differential expression of allelic variants. In reviews of cloned plant QTL, only three of ten QTL whose mechanisms were determined were shown to be due to differential expression (Salvi and Tuberosa 2005). Nonetheless, combining expression-profiling technologies with QTL mapping shows considerable promise. A number of studies have shown this hybrid approach to be useful in identifying the genes potentially causing trait variation (Wayne and McIntyre 2002). For example, Kirst *et al.* (2003) reported a candidate gene underpinning a major-effect QTL in an interspecific *Eucalyptus* hybrid. Furthermore, Cato *et al.* (2006) reported a dehydrin gene associated with both wood density and growth rate in *P. radiata* that showed allelic differences in transcript abundance in different wood-forming tissues within the same genotype.

- A variant of the above, using proteomics rather than mRNA populations. The lack of complete correspondence between translation and transcription may be a useful means to eliminate those genes that are less likely to contribute to trait variation. Moreover, this approach has promise in that it may also identify gene products whose contribution to trait variation may be due to reasons other than differential expression (e.g., protein folding, etc.). Such an approach has not been extensively tried yet, at least not in forest trees.
- Expressed genes that consistently colocalize with QTL regions in multiple pedigreed QTL mapping populations, either within or across species (Figure 10.1, Box D). In practice, this could be of limited value, as confidence intervals around QTL are likely to cover much of a chromosome, particularly where sample size is limited (Dupuis and Siegmund 1999). Nonetheless, pedigreed mapping populations could be used as an additional screening step. However, caution is recommended: small–moderate size QTL mapping populations could be of limited value as they may not be sufficiently powerful to detect QTL, therefore the lack of association is not conclusive; or else the QTN may not be segregating in the particular pedigree(s) being used. If using information from another species to infer trait association in the subject species, then evidence for nonrandom colocation of QTL for traits of interest should be determined *a priori*, otherwise use of information from other species will be of little value.
- Genes that have been shown to be associated with variation in traits of interest via association genetics in other species (Figure 10.1, Box E). Caveats regarding utility of transferability of QTL across species mentioned above also apply. Nonetheless, marker–trait associations that occur in homologous sequences across species may also serve as independent validation of associations.
- Use of genetic transformation to determine potential role(s) of candidate genes (Figure 10.1, Box F). This approach involves modification of endogenous gene function in some manner, e.g., enhancer trapping, RNAi, over-expression, etc. However, for forest trees, such approaches have limited promise, particularly in species where trait expression takes years, and/or have low transformation efficiencies. Other technical problems could also be limiting, e.g., sense suppression in the case of over-expression. Regulatory issues could also impact, particularly where field trials are necessary. However, this approach may be useful in cases where *in vitro* or early-assay systems have been developed, particularly where transient expression can result in a discernable phenotype.

As we learn more about the function of specific genes alone, and in concert with other genes, other criteria are likely to be added to the above list. Moreover, as more information from each of these sources becomes available, it will be possible to evaluate the relative efficacy of each of these criteria. Suffice to say, the roles of structural and comparative genomics, proteomics, molecular biology, as well as knowledge of physiological roles of specific genes, are crucial. Very few of these skills are currently utilized by, or available within, current tree breeding programs.

Of interest too, are the identity and nature of regulatory regions associated with candidate genes (Morgante and Salamini 2003; Paran and Zamir 2003). Because trait variation could be a result of gene regulation, there is a need to ascertain – via *de novo* sequencing if necessary – regulatory sequences. This should be easily achievable for promoter sequences in close proximity to open reading frames, but may be more difficult for transacting enhancer elements, particularly if such sequences are not known *a priori*.

10.7.4.2 Polymorphism discovery and evaluation

Following selection of candidate genes, further evaluations are required (Figure 10.1). These involve resequencing on a subset of genotypes to identify specific polymorphisms, and to determine patterns of disequilibrium before choosing a subset for testing for associations with traits of interest. SNPs and indels are the most likely forms of polymorphism to be useful, although other forms (such as repeat sequences) could also be useful. Polymorphisms that need to be detected and evaluated include not only those nonsilent substitutions in coding regions, but also polymorphisms in noncoding regions such as introns, and 5′ regulatory regions, particularly if they are not in disequilibrium with polymorphisms in coding regions. Patterns of disequilibria will need to be determined on a gene-by-gene basis, unless some general patterns emerge that can be applied across all genes. The relatively short span of disequilibria observed in forest trees (Table 10.2) – at least by some statistics, such as r^2 – will necessitate extensive SNP discovery and evaluation throughout the relevant genic regions.

Detection of SNPs and evaluation of disequilibria require genomic sequence information, some of which can be obtained from EST databases, but regulatory and intronic regions will need to be sequenced from genomic DNA. This step – polymorphism detection – is likely to be very time-consuming and labor-intensive, particularly in species where little EST and/or gDNA sequence information is available, and may well limit the rate of implementation, as individual polymorphisms will need evaluation and assays for chosen SNPs will need to be optimized for large-scale genotyping. For most forest tree species, technologies are needed that expedite polymorphism detection and resolution without the need for extensive sequence information.

It may therefore be useful to implement further marker-selection criteria at this point, prior to extensive SNP optimization and/or resequencing. Possible criteria include whether the sequence data generated reveal any evidence indicating a possible role in trait variation – such as evidence of selection, which can be obtained from examining patterns of nucleotide substitution in coding and noncoding regions for example. For example, Cato *et al.* (2006) reported elevated levels of nonsynonymous substitution in a dehydrin gene in *P. radiata* that had previously been shown to colocate with wood density QTL, and was subsequently shown to be associated with both growth rate and

wood density in an association population consisting of 1,700+ genotypes. Whether or not such criteria will be broadly effective is yet to be determined, in part because some QTN may not be under natural selection, yet still of use for artificial selection.

Once polymorphisms are detected and optimized for genotyping, a subset of polymorphisms will need to be selected and screened across some form of association population for which phenotypic data are available. Numbers required per gene (and associated regulatory regions) will depend upon the number of statistically independent regions per base pair and the size of the region being evaluated. It may therefore be necessary to screen tens of polymorphisms per genic region, although Krutovsky and Neale (2005) estimated less than ten would suffice for all but large genes. Also, because the size of the populations is likely to be in the order of many hundreds to thousands (below), high-throughput SNP genotyping is likely to be necessary. A range of technologies are available for this, and technology developments in this area are ongoing. Access to such technologies is obviously required, at affordable cost.

10.8 HOW MIGHT *GAS* BE INTEGRATED INTO A TREE IMPROVEMENT PROGRAM

The generic advantages of using association genetics in tree breeding have already been stated (cf. Stromberg *et al.* 1994). For effective use there are many possibilities. Some of the issues will be common to both MAS (including marker-based and marker-assisted selection) and true GAS based on QTN, and some will be specific to one or the other. To be effective, use in tree breeding of nucleotide–trait associations derived from association genetics must be integrated with essentially the existing tree improvement practice. Such practice includes the arrangement and structuring of breeding populations, and the manner in which genetic gain is delivered into plantation forests. For the future, the practices can be modified as true GAS becomes possible.

Tree breeding differs from much traditional crop plant breeding because of various factors, including relatively little history of domestication, moderate–high levels of genetic load, and long generation intervals imposed by slowness to reach reproductive competence and/or late expression of trait values. Forest tree breeding tends therefore to take a population-based approach involving many genotypes, where populations are usually structured into a hierarchy (Burdon 1988):

- At the lowest level are unimproved *gene resources* (essentially undomesticated genotypes).
- From these, the next level, the *breeding population*, is or already has been chosen.
- From which in turn the best genotypes are chosen (usually progeny-tested) for the *production population*, from which planting stock is derived for forest plantations.

This hierarchy of populations is schematically like a pyramid with the widest genetic diversity at the base, and the narrowest genetic variation at the top level of genetic improvement. Within this scheme, there can be many variations and refinements. Movement of genetic material will tend to be very much up the hierarchy, in the nature of replenishing genetic diversity in the upper levels.

At the start of a breeding program, before any progeny testing, the production population and the breeding population are often one and the same. Thereafter, the

breeding population becomes the "engine room" for cumulative genetic advance, building up frequencies of favorable alleles through successive cycles of mating, genetic recombination, and selection. For clonal forestry, clonal selection will typically be done within crosses between top-ranked parents which may be common to both the breeding population and existing seed orchards.

To complicate matters, tree breeding typically involves multitrait breeding objectives, and some programs also develop specific breeds that focus on improving differing sets of traits (Jayawickrama and Carson 2000). Application of GAS in tree improvement programs needs to fit into this general framework in a cost-effective manner. We will now consider potential applications of GAS in the context of such population hierarchies.

10.8.1 Plus-Tree Selection Applications

In programs where new plus-tree selections are required, GAS may be useful as a prescreening tool either to increase selection intensity, or to cull candidates down to those of sufficient promise to warrant costs of testing, and of forwards selection among offspring. Here, GAS has, in theory, the advantage of favoring selection well before full phenotypic expression, therefore increasing the available number of selection candidates. However, this may be constrained by the cost of phenotyping relative to genotyping, plus the desideratum of ascertaining marker–trait associations for the multiple traits that comprise a breeding goal. Nonetheless, marker–trait associations could be accumulated over time from association tests, and utilized as they become available, thereby increasing scope for adding new material into breeding populations. Similarly, genotypes could be identified for immediate deployment, in addition to incorporating them into breeding populations – assuming propagation systems exist to cost-effectively multiply selected genotypes without detrimental effects of maturation. For instance, in response to a biotic crisis (e.g., outbreak of a new disease or pest) GAS could be directly applied to identify genotypes more likely to be resistant to the pathogen or pest, rather than undertake laborious phenotypic screening. Specific genes could then be integrated more quickly into the relevant populations. Prospects for widespread application of GAS for plus-tree selection may be limited in practice; however, as population sizes for detecting associations would most likely exceed those required for breeding population advancement. Moreover, knowledge of nucleotide–trait associations may come to hand too late for fresh plus-tree selection, especially with traits of late expression.

10.8.2 Breeding Population Applications

Breeding populations in forest trees tend to comprise many genotypes, sometimes exceeding 1,000 parents, most of which are putatively unrelated plus trees and/or their offspring. Co-ancestry is usually minimized, to avoid deleterious and sometimes unpredictable effects of inbreeding, often via use of sublines (Burdon and Namkoong 1983). Substructuring of breeding populations is often undertaken, utilizing "main" and "elite" populations, generally with more intensive data gathering and selection in the smaller elite populations, to secure genetic gains sooner than in the main populations. Phenotypic evaluation in breeding populations is usually done on offspring that are planted in common-garden genetic tests, which allow breeding population advancement

by forwards selection for the multitrait criteria. Backwards selection, from progeny-test results, is also used to rank parents, particularly for production populations.

For breeding population advancement, the same marker–trait associations as might be used for plus-tree selection described above could be used for selecting among and within families, to increase selection intensity, as an early selection tool, and/or to reduce costs. However, even within breeding populations, specific applications will be context-dependent. For example, in main populations, which are generally less intensively managed than elite populations, GAS could be used as a surrogate for more expensive-to-measure traits. Here, phenotypic data could be generated on cheap-to-assay traits (e.g., growth rate) and GAS used for more expensive or later-expressing traits (e.g., certain wood properties). However, for the time being, DNA polymorphisms are likely to characterize less additive genetic variation than phenotypic records, resulting in potentially less gain for traits selected just on marker information. Such a reduction could be offset by increasing selection intensity among, and particularly, within families. Trade-offs will need to be carefully evaluated, initially at least via simulation.

For any breeding, an ideal is saving rare or low-frequency QTN that have current or contingently favorable additive effects. Such alleles can be the key to longer-term genetic gain and/or coping with a biotic crisis. For detecting, preserving and increasing the frequencies of these QTN, instead of losing them to genetic drift, GAS may be crucial. However, such a pursuit may well be deemed too expensive for breeding programs that are dominated by shorter-term financial imperatives.

In elite populations, with the fewer families for intensive measurement and selection, opportunities may exist for more intensive selection and faster turnover of generations. For combined among- and within-family selection, there is more scope to increase selection intensity within families. Because association tests identify markers in strong disequilibria with QTN, it may be relatively easy to detect pedigrees within which the predominant linkage phase is reversed. Undetected reverse-phase linkages are likely to be serious within small elite populations, or any other small breeding groups within the breeding population; simulation would again be helpful in quantifying potential reductions in gain.

Reducing generation intervals through use of GAS would depend on the trees becoming reproductively competent before trait expression. However, if markers or actual QTN were used as a surrogate for trait expression, genotypes could be screened as soon as sufficient tissue can be spared for DNA assays, even in germinating seedlings. Some conifers, in particular, are typically reproductively competent before selection age for at least some commercially important traits, creating a real potential for use of GAS to shorten generation interval. However, this would require marker–trait associations that explain substantial additive genetic variance for at least some important breeding goal traits. While this could one day be achieved, it is currently more likely to have associations that explain only a proportion of additive variance for just subset of traits. Thus, trade-offs between expected gain per generation and rate of generation turnover will need to be carefully evaluated.

It is more likely that, in the shorter term at least, selection in elite breeding populations would be implemented in a multistage approach, using marker information as an early screening tool, followed by phenotypic records. Such an approach could either increase selection intensity (by screening more genotypes), or reduce costs of phenotypic evaluation by short-listing genotypes for field testing, to achieve the same gain. Alternatively, using GAS to select for later-onset traits – if the nucleotide–trait

associations are established – could reduce generation interval, by concomitantly using phenotypic records on the earlier-expressed traits. A simple example could be in breeding objectives that incorporate both growth rate (if it is only expressed well at an advanced age) and, say, resistance to a disease for which empirical phenotypic screening is possible in very young seedlings.

There are other generic breeding population applications for GAS, which apply alike to both main and elite populations. These include more powerful selection via correlation breakers,[4] reselection, and as a surrogate for later-onset and/or expensive-to-measure traits. Such applications, while generic in nature, seem appropriate for where the need is greatest – more likely in elite populations.

Selection for recombinants of known QTN that break adverse genetic correlations between breeding goal traits is especially attractive. Detection of such recombinants would not require field testing, and can involve many more genotypes than could be field-tested, thus raising the probability of encountering the desired correlation breakers. Such genotypes would then need field testing, as confirmation, which would be done anyway in breeding population advancement.

A challenge will exist in applying GAS to new breeding goal traits in breeding populations. Tree breeding not only usually involves multitrait breeding goals, but also new traits are sometimes added to breeding goals in response to changes in market perceptions and values. Information for establishing the requisite associations for using GAS may be already available, even if the trait was not originally part of the breeding goal; otherwise, the major effort of fresh association tests may be needed. Alternatively, existing association tests may be screened for those new traits, and any subsequent association used for backwards and/or forwards selection, rather than extensively screening multiple progeny tests over successive generations for the same traits. For selecting a new trait, the greater selection intensity allowed by forwards selection would be very attractive, but at the risk of a new generation's decay of LD. As usual, correct choices of candidate genes will be key to making this approach cost-effective, especially finding the polymorphisms in strong LD with the QTN if not the actual QTN.

Related to this, is the potential to use GAS as a surrogate for phenotypic evaluations that are either expensive or involve destructive sampling. While establishing associations between markers and traits would of course require expensive phenotypic evaluations as part of the operational development; it may well be cheaper to use this route than to continue "trawling" numerous genetic tests over several cycles of breeding. Where assessment is necessarily destructive, there may be limited opportunity to measure progeny tests because of their inherent value for assaying other traits; therefore, GAS could be used as a surrogate for destructively sampled traits – if the requisite associations have already been established. Clonal replication of individual offspring, however, would effectively avert loss of material to destructive sampling.

QTN conferring resistance or tolerance to specific pests or pathogens may be particularly amenable to GAS. Pathotype-specific resistance genes of large effect are known in forest tree pathosystems (Kinloch *et al.* 1970; Wilcox *et al.* 1996), and in some cases are of great commercial potential despite their specificity. Identifying the QTN underpinning such pathotype-specific resistance, or finding polymorphisms in strong

[4] The type of correlations that can in principle be attacked effectively in this way would be correlations resulting from important chromosomal linkages that are persisting following fusion of differentiated ancestral populations, rather than correlations stemming from pleiotropic effects

disequilibrium with these QTN, has the benefit of obviating the need for screening families with specific pathotypes, to determine which families carry which resistance genes. Combining or "pyramiding" different resistance genes, preferably within the same individuals, can promise resistance that is durable against mutations and genetic shifts in the pathogen (Burdon 2001). Thus, phenotyping costs can be much reduced, as well as time required for manipulation of frequencies. This may be a great advantage in the event of a biotic crisis where low-frequency resistance is required to quickly combat a new disease or pest. The advantage would be increased by the desirability of pyramiding different resistance factors. Genotypes carrying such QTN can be identified in the breeding population (including directly estimating QTN frequencies), enabling among- and within-family selection to be carried out over a large proportion of the breeding population. In such circumstances, it is likely that at least some of the resistant genotypes will be suboptimal for other traits, so GAS might be used to select for other properties to reduce the loss in genetic gain.

Despite the prevalence of inbreeding depression in forest trees, use of inbreeding as a breeding tool has attractions because it can theoretically amplify the expression of additive gene effects (e.g., Burdon and Russell 1999; Russell *et al.* 2003). In most species, however, the challenge will be to "purge" highly deleterious recessive alleles ("hard" genetic load) that threaten viability and/or often mask the expression of favorable additive gene effects in inbred lines (e.g., Williams and Savolainen 1996). MAS has promise for such purging, because QTL effects of hard load should be relatively easy to detect in individual pedigrees in order to purge such alleles even in the heterozygous state (cf. Kuang *et al.* 1999). Use of GAS in this way, however, may not really work, because such genetic load almost certainly represents alleles that are individually rare but occur at very many loci and are therefore very unlikely to be involved in any general LD.

10.8.3 Production and Deployment Populations

Production populations comprise the genotypes that either provide seed for deployment into plantation forests, or are used for large-scale vegetative propagation for clonal forestry. These populations usually have a few tens of genotypes at any one time, and actually represent subsets of the breeding populations and are subject to most of the same considerations as the breeding populations for the applications of GAS. As subsets, they represent a relatively narrow genetic base compared to the breeding- and gene-resource populations. Related matings are avoided as far as possible, to avert inbreeding depression. Various systems are used to deliver commercial planting stock. Some programs use open-pollinated seed orchards, to produce seedlings. Other programs use control-pollination technologies, where top genotypes are pollinated with pollens from either single or multiple parents. Seed from these either provides seedlings for planting stock, or is vegetatively multiplied as nursery cuttings or as plantlets raised from *in vitro* culture, but, despite the average level of genetic improvement, this still produces uncharacterized segregating offspring genotypes. For clonal forestry, genotypes produced by intercrossing top parents are subject to a further round of testing and selection, before identifying and mass-propagating top clones for deployment.

Production populations are of key importance, as it is these populations from which seed and plant producers obtain most of their revenues, thus additional costs associated with this form of selection can be offset in a shorter time period than the breeding

population applications, as few if any products are delivered to forest growers directly from breeding populations. Furthermore, there is continual pressure on breeding programs to deliver gains to commercial plantations faster and/or at greater rates. Production populations are therefore more likely to be target populations for applying GAS, at least in the shorter term.

GAS, along with its variants, has obvious possibilities for selecting individual offspring for clonal forestry and/or subsequent vegetative amplification of a narrow range of genotypes – such as in situations where "family forestry" is combined with vegetative amplification. The parents – while they may already have been selected with the aid of GAS – will almost certainly still be highly heterozygous, so the expected genetic variation within any sort of family will be considerable for most quantitatively inherited traits. Where GAS is based on markers in LD with the QTN rather than on the QTN itself, response to selection of a limited number of clones in a limited number of families could be very vulnerable to reversals of the prevailing linkage phase, especially as this material will represent one more generation for decay of LD to occur in. On the other hand, the small number of families should make it relatively easy to verify linkage phases in individual pedigrees. The results of Wilcox *et al.* (2001) indicate that this scenario could be cost-effective in the context of within-family selection (MAS) based on neutral DNA markers.

In selecting clones for clonal forestry the potential of GAS for selecting rare recombinants, especially involving QTN, looks particularly attractive, because such recombinants could not be produced reliably through sexual reproduction within any reasonable timeframe.

Where new traits must be addressed in the breeding goal, the emphasis in selection for production populations is likely to shift in favor of forwards selection over backwards selection, which is likely to favor use of GAS if the appropriate associations can be established.

For disease resistance (and possibly some cases of insect-pest resistance), the potential of GAS for advantageous pyramiding of resistance factors looks especially valuable. This could be all the more important where durability of resistance may depend on certain individual resistance alleles remaining at minority frequencies, in pyramiding at the level of the population rather than the individual genotype.

10.8.4 Summary: Selection Application in Forest Tree Species

This section has outlined generic applications of marker–trait associations obtained from association genetics for tree breeding programs. Overall, GAS can be applied at the various strata and substrata in the genetic hierarchy of a classical tree breeding program. Within each of these strata there are opportunities to increase genetic gains by increasing selection intensity, more accurate selection, reduced costs of field testing and phenotypic evaluation, and possibly to speed up responses to changes in breeding objectives. Specific applications would, however, need to be carefully and quantitatively evaluated on a case-by-case basis, particularly in light of the fact that results from association tests will most likely come from a limited range of traits where only a proportion of the extant variation is accounted for by assayable polymorphisms, at least in the short term. Furthermore, because of the additional costs of this form of selection compared to phenotypic selection alone, it is likely that the initial application will be in the production populations, where

investment in nearer-to-market selection applications are likely to have more immediate pay-back.

10.9 FIT WITH OTHER BIOTECHNOLOGIES USED IN TREE IMPROVEMENT

As already stated, a key feature of GAS is the complementary fit with other genetic technologies, including those currently under development. For these new technologies to be applicable, they need to be more cost-effective at delivering genetic gains than conventional technologies. A number of new technologies are under development, and are at various stages of readiness for implementation in tree breeding programs. Here, we consider examples of new technologies that can be used to complement GAS and greatly enhance its effectiveness.

10.9.1 Within-Family Selection Based on DNA Marker–QTL Associations (MAS)

Scope exists for integrating GAS strategy with that of MAS. Because most commercially important tree breeding programs are now well into advanced-generation selection, there is significant emphasis on within-family selection in order to maintain the breadth of genetic base and avoid undue build-up of co-ancestry. MAS could be used for within-family selection although some limitations have been noted (Strauss *et al.* 1992; Kerr and Goddard 1997; Johnson *et al.* 2000), including the need for large individual family sizes necessary for achieving genetic gain for most quantitatively inherited characteristics (Wilcox *et al.* 2001). Given the high cost of detection of marker–trait associations for MAS on a family-by-family basis, it is likely that in breeding programs using MAS, detection of marker–trait associations will have been undertaken in only a subset of families in their respective breeding programs. Here, GAS could be used both as an aid to among-family selection and to augment MAS for within-family selection where family-specific marker–trait associations for MAS are not available. There are two potential benefits in doing this: firstly, increased genetic gains for reasons outlined above, and secondly, alleviation of the accelerated build-up of co-ancestry that could occur with the operational dependence on MAS. With MAS, accelerated co-ancestry could arise through MAS being available only for a small proportion of pedigrees which could therefore contribute disproportionate numbers of selections. More broadly applicable marker–trait associations (i.e., GAS), by facilitating selection from all pedigrees, would not be conducive to the same build-up of co-ancestry. Given the large sample sizes per family that are needed to detect QTL so as to achieve moderate genetic gains from MAS (Wilcox *et al.* 2001), practicing MAS across large numbers of essentially unrelated families becomes prohibitive. In comparison, GAS requires much lower sample sizes when averaged across the number of parents in breeding populations (discussed below). However, this advantage could be offset to some extent by the need to identify and assay many more polymorphisms per candidate gene, although there is potential to reduce sampling costs due to techniques such as pooling DNA samples from phenotypic extremes (Michelmore *et al.* 1991). Moreover, in specific cases such as dominant major genes for disease and insect resistance (cf. Bus *et al.* 2000), which do not require large sample sizes for detection, MAS is likely to be an effective means of obtaining gain; when thus detected, such genes may then be amenable to use of GAS, with the help

of comparative genomics based on DNA sequences in other plants. Similarly, family-specific effects associated with inbreeding (e.g., lethals, loci contributing to reduced vigor and general fitness status) may be better dealt with via MAS on a pedigree-specific basis as co-ancestry builds up, with GAS used to select for other characteristics. This could be especially important for purging highly deleterious alleles if aggressive inbreeding were to be adopted as a breeding tool.

Combined use of experimental infrastructure for both MAS and GAS has potential benefits also. Pedigreed QTL detection populations (as would be used for MAS) with association genetics population (as used for GAS) have been evaluated as a means of fine-mapping QTL (see Chapter 8 and references therein). Such an approach could be used to reduce confidence intervals around QTL location, thereby narrowing the range of potential candidates and effectively increasing the probability of choosing the appropriate genes.

10.9.2 Genetic Engineering

For operational use of genetic engineering, it is always important to do transformations on carefully chosen recipient genotypes. This is partly because inherently poor recipients will remain poor even after transformation, and partly because transformation costs are still high because of both the inherent costs of the protocols and the low success rate resulting from the inexact nature of contemporary transformation technologies. Selection of recipient genotypes, however, may be constrained by the fact that transformation may need to be done on embryogenic material. This creates a special attraction for the sort of very early selection that GAS can afford, by using DNA data (along with prior family information) to identify top candidates for transformation.

In addition to the operational use of genetic engineering there is the role of genetic engineering to establish the roles of candidate genes, which may serve to inform conventional breeding via indicating which genes are likely to result in phenotypic effects. In practice, this could be limited because of the time and expense of genetic modification, although some genes could be identified in this manner (see Section 3.4.1).

10.9.3 *In Vitro* Propagation Technologies

With the various technologies for *in vitro* propagation (e.g., organogenesis and somatic embryogenesis), the opportunity for early identification of top genotypes has benefits when both amplifying limited quantities of top genetic material, as well as for development of material for clonal testing and deployment. This form of early selection not only increases selection intensity, but also could be used to increase the efficiency of tissue culture by identifying genotypes more likely to propagate well – although having to select for propagation behavior is liable to be at the expense of potential genetic gain in other directions. This also applies to *in vivo* vegetative propagation. However, with a number of propagation technologies in various commercially important forest tree species, further development of propagation technologies may be required to fully utilize the potential from GAS.

10.9.4 Accelerated Flowering and Rejuvenation Technologies

Accelerated flowering technologies may be crucial to realizing at least some of the benefits of GAS and MAS. Such technologies can make it possible to capitalize on the early selection afforded by GAS to dramatically reduce length of breeding cycles and the lead time for deployment of genetic gains, thereby achieving more effective utilization of endogenous variability. For example, breeding cycles in contemporary commercially important conifer species are still 14–20 years in length, with selection requiring 4–10 years, and flower induction and seed production requiring a further 5–8 years. Flowering-on-command, coupled with selection based on DNA sequence information, could reduce the time for identification of top genotypes dramatically – in theory to much less than a year. Indeed, reducing the time required for floral induction, fertilization, and seed production could increase rates of gain by as much as three times, depending upon the reduction of generation interval.

Rejuvenation technologies achieve the opposite to accelerated flowering in operational breeding. The prospective benefits of rejuvenation for realizing genetic gain are great (Burdon 1982; Bonga and von Aderkas 1993), but they generally interplay less specifically with GAS than do the benefits of accelerated flowering.

10.9.5 Technologies to Study Pathways of Gene Action

GAS experiments (LD populations) are also useful as screening populations for identifying potential causative QTN, allowing integration of molecular and selection technologies by sharing common experimental platforms. The potential offered by association genetics experiments to identify candidates offers molecular biologists the opportunity to use genetics to inform roles and functions, thereby elucidating the particular roles of specific genes and the manners in which they might interact at a whole-organism level, either informing or complementing *in vitro* or model plant studies. Benefits arising from identification of causal mechanisms and pathways, apart from improved understanding of the molecular basis for heritable variation, include identifying genes (and methods) to create and exploit variation based on understanding the causal mechanisms (including potential pleiotropic effects). In the shorter term, a further benefit includes the identification of which and what type of genes could be targeted to create new "mutations" (via transformation) of potentially larger effect (Section 10.9.2 and above).

10.10 LIMITATIONS AND CHALLENGES

While the potential for GAS in tree breeding looks positive, implementation in commercial breeding programs faces a number of key obstacles. These include the high cost of implementation, institutional barriers, and technical impediments due to certain molecular mechanisms underpinning trait variation. We briefly discuss each of these below.

A key impediment to uptake is the high up-front cost of implementation, which is particularly important given that most commercial breeding programs need to bear most or all of the entire costs, whereas the benefits of genetic gain tend to accrue further down

the forestry value chain, which can take decades to materialize. Reasons for high implementation costs include:

- *High cost of establishing marker–QTN associations*. In order to achieve adequate experimental power, large experiments are likely to be needed (above). Furthermore, such experiments are likely to be costly to measure, particularly as most breeding objectives involve multiple traits, and typically include expensive-to-assess wood-property traits.
- *Costs associated with polymorphism discovery and genotyping*. Polymorphism discovery consists of extensive amounts of resequencing, followed by elucidation of disequilibrium patterns after which subsets of SNPs are chosen for genotyping in association tests (Figure 10.1). Because a number of polymorphisms per gene will be needed as well as several genes per QTL interval (unless prior information indicates a clear choice), there is substantially more evaluation and genotyping required per QTN than compared with MAS using pedigreed QTL mapping populations, although with the latter marker–trait associations need to be ascertained on a pedigree-by-pedigree basis. Such costs are not trivial, and may only be offset by investment from public funding agencies or by collaborations with organizations undertaking association genetics studies for purposes other than selection. Associated with sequencing and genotyping costs is the necessity to access facilities to undertake such work, although access to technologies could be attained through service providers and existing laboratories.
- *Additional skills needed for operational implementation in breeding programs*. These include competency in marker technologies, genomics and cellular biochemistry (primarily for candidate gene selection), and quantitative genetics methods relevant to detecting and estimating linkage disequilibria. Such skills usually require teams rather than single individuals, which therefore requires additional investment to establish and maintain an infrastructure associated with such teams, unless such skills can be acquired via collaboration.
- *Occurrence of genotype × environment interaction*. This will increase the number of experimental populations that will need to be deployed, although deploying cloned experimental populations could minimize additional genotyping. Even if selection for specific environments is not needed, good coverage in terms of test environments may still be needed (cf. Johnson and Burdon 1990).
- *Intergenerational changes in relationship between QTN and phenotype*. These could arise for example with disease/pest resistance genes, where shifts in the pathogen/pest population could change predictive value of QTN(s). Similarly, changes over time in environments, or even silvicultural practices, could likewise change the nature and/or extent of the causative associations in a manner that may not be easy to predict. Such changes would be likely to make certain costs recurring.

High costs mean GAS is unlikely to be an attractive option for species and/or breeding objectives with low commercial value. Even for species with greater commercial value, the additional investment may not be considered affordable, particularly for existing operational programs that lack additional financial resources with

which to develop and implement the operational infrastructure necessary for GAS. Therefore careful evaluation of specific implementation strategies and including costs and benefits are most likely to be necessary.

Certain mechanisms underpinning trait variation could also prevent effective development of GAS. An example particularly relevant to species with limited commercial value and/or relatively limited availability of nongenic DNA sequence (particularly those with large genomes) is where causative QTN occur many kilobases distal to expressed genes. Such is the case for the *Vgt1* locus in corn, which has been shown via association genetics to map to a 2 kb region that is 70 kb away from the nearest open reading frame (Salvi *et al.* 2006). If such distal transacting regulatory factors dominate trait variability, then extensive amounts of gDNA resequencing will be required. This would significantly add to costs, as well as reduce efficacy, particularly for large-genome species, effectively precluding application in gymnosperms, as well as a number of hardwood species. Another example is where trait architecture is predominantly composed of clusters of small-effect QTN per QTL. Such architecture is theoretically possible, and further experimentation will reveal whether or not this is the case. Experiments of sufficient power will be necessary, increasing cost and time required to detect QTN. Furthermore, genotyping costs per unit of gain will be greater, potentially offsetting expected benefits.

Another technical limitation is the predictive value of associations in the light of potential modes of gene action, particularly epistasis. Nucleotide substitution effects would usually be estimated by averaging over allelic combinations sampled in association tests. However, the selected variants may not be well represented in association tests, so the predictive value of multilocus QTN could be limited in the presence of epistasis. Evidence from genetic tests in conifers indicates that large-effect epistasis is unlikely to be prevalent, but does not rule out smaller epistatic effects. Such interactions are plausible, given the nature of interdependent biosynthetic pathways that give rise to phenotype, but may not be observed (or even important) in large outbred deployment populations that are typically derived from open- and control-pollinated seed orchards. Conversely, for clonal forestry, where GAS could potentially be used to identify candidates for further testing, such interactions could be important, particularly if candidates available to be screened are unlikely to include optimal multitrait genotypes because of biological limitations on the numbers of seed that could be produced for screening.

A specific, potentially important class of epistasis, is co-adapted gene complexes. This phenomenon is possible in forest trees, although some surprising cases have been observed of essentially independent inheritance of traits that would seem to have common adaptive significance (Howe *et al.* 2003). If, however, such complexes do exist, they must be considered when generating and selecting new variants, necessitating the detection and if necessary, management of, haplotypic complexes. Fortunately, further experimentation to detect such complexes may be unnecessary, as existing technologies combined with association test populations may well be adequate. We envision that such research will be undertaken over the next few years. If present, means of managing co-adapted complexes in tree breeding programs will need to be implemented; although this may not be difficult in theory, it may present major logistical challenges.

GAS may have little or no utility for backwards selection and reselection within existing breeding and production populations, particularly where progeny tests are already established and measured for other traits. Such instances may not be rare, as

breeding objectives and strategies are frequently being revised, and new traits are often introduced into breeding programs in response to factors such as new biological pressures and/or market signals. In these cases, it may be more cost-effective to screen extant families for new properties. In breeding programs with limited resources, the short-term cost-effectiveness of such approaches may restrict or prevent investment in technologies such as GAS which are longer-term in delivery of improved germplasm, unless marker–trait relationships can be easily undertaken in association tests that result in a significant proportion of trait variation being explained by markers.

Institutional barriers to implementation also exist. In the case of breeding cooperatives and companies whose programs are based on phenotypic selection, barriers can exist to understanding the nature and complexities of molecular genetics applications as most programs have tended not to use such tools routinely, and when done, usually in some conceptually easy application such as verification of parentage or clonal identity. Convincing such organizations, which tend to be conservative, to implement this technology, may be difficult particularly in light of the few results to date that clearly demonstrate ease of detecting associations let alone actual genetic gains. Furthermore, fluctuations in the relative economics of plantation forestry and frequent ownership changes can prevent adequate investment from nongovernment sources to appropriately develop and implement the technology. This may be particularly important where plantation ownership is dominated by investors with short-term financial goals, therefore unwilling to participate in more longer-term activities such as association genetics.

For reasons described above, we foresee that GAS is most likely to be implemented in breeding programs where there are good operational links between molecular geneticists and tree breeders (as well as others), either moderate to high product values or sufficient scale to allow costs to be widely spread, and sufficient investment over the requisite period of time to enable discovery of suitable numbers of marker–QTN relationships.

10.11 CONCLUSIONS

Application of association genetics in plantation forest tree species has the potential to increase genetic gains from among- and/or within-family selection via a number of routes such as increased selection intensities and/or earlier selection. Such selection can be applied to virtually all strata of hierarchically structured populations used in tree improvement, although it is likely that the most immediate applications will be in populations used to provide seed for commercial plantations, owing to the relatively shorter timeframe to recover additional costs associated with detecting marker–trait associations. Other potential benefits include cheaper selection, reduced need for phenotypic selection, and complementary fit with other biotechnologies used either commercially or in research, as well as use of the same experimental infrastructure for purposes other than selection.

The few studies to date of LD in forest trees indicate relatively short spans of LD, implying that finding disequilibria between causative QTN will need to be undertaken via judiciously chosen candidate genes (hence use of the term “gene-assisted selection”), particularly in conifers where large genomes effectively preclude cost-effective whole genome resequencing.

There are a number of important prerequisites for GAS to be successful. These include effective integration of existing tree breeding skills with molecular genetics,

genomics, and bioinformatics, as well as relevant statistical skills. In addition, access to adequate populations with which to detect sufficient numbers of small-effect QTN are a key requirement. Access to genomics and genotyping facilities are also critical, as are accessed to technologies that will improve the ability to choose appropriate candidate genes.

There are, however, some potential impediments to implementation of association genetics in tree breeding. These include the high costs of detecting marker–trait associations relative to product value and long rotation lengths of forest trees; certain modes of gene action which may preclude effective detection of associations, particularly in conifers; and institutional barriers associated with understanding and investing in new technologies.

10.11 REFERENCES

Allison, D.B., 1997, Transmission-disequilibrium tests for quantitative traits. Genetics 60:676–690.

Ball, R.D., 2001, Bayesian methods for quantitative trait loci mapping based on model selection: approximate analysis using the Bayesian Information Criterion. Genetics 159:1351–1364.

Ball, R.D., 2005, Experimental designs for reliable detection of linkage disequilibrium in unstructured random population association studies. Genetics 170:859–873.

Beavis, W.D., 1994, The power and deceit of QTL experiments: lessons from comparative QTL studies. pp. 250–266. In: Proceedings of the 49th Annual Corn and Sorghum Industry Research Conference. American Seed Trade Association, Washington, DC.

Bonga, J.M., von Aderkas, P., 1993, Rejuvenation of tissues from mature conifersand its implications for propagation *in vitro*. In: Clonal Forestry (Eds. M.R. Ahuja, W.J. Libby) pp. 182–199. Springer-Verlag, Berlin Heidelberg.

Bradshaw, H.D., Stettler, R.F., 1995, Molecular genetics of growth and development in *Populus*. IV. Mapping QTLs with large effects on growth, form, and phenology traits in a forest tree. Genetics 139:963–973.

Brown, G.R., Bassoni, D.L., Gill, G.P., Fontana, J.R., Wheeler, N.C., Megraw, R.A., Davis, M.F., Sewell, M.M., Tuskan, G.A., Neale, D.B., 2003, Identification of quantitative trait loci influencing wood property traits in loblolly pine (*Pinus taeda* L.) III. QTL Verification and candidate gene mapping. Genetics 164:1537–1546.

Brown, G.R., Gill, G.P., Kuntz, R.J., Beal, J.A., Nelson, C.D., Wheeler, N.C., Penttila, B., Roers, J., Neale, D.B., 2004a, Associations of candidate gene single nucleotide polymorphism with wood property phenotypes in loblolly pine (Abstr.). Plant and Animal Genome XII, 10–14 January 2006, San Diego, CA.

Brown, G.R., Gill, G.P., Kuntz, R.J., Langley, C.H., Neale, D.B., 2004b, Nucleotide diversity and linkage disequilibrium in loblolly pine. Proceedings of the National Academy of Sciences of the United States of America 101:15255–15260.

Bucci, G., Menozzi, P., 1995, Genetic variation of RAPD markers in a *Picea abies* Karst. population. Heredity 75:188–197.

Burdon, R.D., 1982, The Roles and Optimal Place of Vegetative Propagation in Tree Breeding Strategies. In: Proceedings of IUFRO Meeting on Genetics and Breeding Strategies pp. 66–83. Sensenstein, Germany.

Burdon, R.D., 1988, Recruitment for breeding populations: objectives, genetics, and implementation. In: Proceedings of Second International Conference on Quantitative Genetics (Eds. B.S. Weir, E.J. Eisen, M.M. Goodman, G. Namkoong) pp. 555–572. Sinauer, Sunderland, MA.

Burdon, R.D., 1992, Genetic survey of *Pinus radiata*. 9: general discussion and implications for genetic management. New Zealand Journal of Forest Science 22:174–198.

Burdon, R.D., 2001, Genetic diversity and disease resistance: some considerations for research, breeding and deployment. Canadian Journal of Forest Research 32:596–606.

Burdon, R.D., Namkoong, G., 1983, Multiple populations and sublines. Silvae Genetica 32:221–222.

Burdon, R.D., Russell, J.H., 1999, Inbreeding depression in selfing experiments: statistical issues. Forest Genetics 5:179–189.

Burdon, R.D., Firth, A., Low, C.B., Miller, M.A., 1998, Multi-site provenance trials of *Pinus radiata* in New Zealand. Forest Genetic Resources No 26. pp. 3–8. FAO, Rome.

Bus, V.G., Gardiner, S.E., Bassett, H.C.M., Ranarunga, C., Rikkerink, E.H.A., 2000, Marker assisted selection for pest and disease resistance in the New Zealand apple breeding programme. Acta Horticulture 538:541–547.

Casasoli, M., Derory, J., Morera-Dutrey, C., Brendel, O., Porth, I., Guehl, J.M., Villani, F., Kremer, A., 2006, Comparison of quantitative trait loci for adaptive traits between oak and chestnut based on an expressed sequence tag consensus map. Genetics 172:533–546.

Cato, S.A., Pot, D., Kumar, S., Douglas, J., Gardner, R.C., Wilcox, P.L., 2006, Balancing selection in a dehydrin gene associated with increased wood density and decreased radial growth in *Pinus radiata* (Abstr.). Plant and Animal Genome XIV, 14–18 January 2006, San Diego, CA.

Chagné, D., Brown, G., Lalanne, C., Madur, D., Pot, D., Neale, D., Plomion, C., 2003, Comparative genome and QTL mapping between maritime and loblolly pines. Molecular Breeding 12:185–195.

Deng, H.-W., 2001, Population admixture may appear to mask, change or reverse genetic effects of genes underlying complex traits. Genetics 159:1319–1323.

Devey, M.E., Delfino-Mix, A., Kinloch, B.B., Neale, D.B., 1995, Random amplified polymorphic DNA markers tightly linked to a gene for resistance to white pine blister rust in sugar pine. Proceedings of the National Academy of Sciences of the United States of America 92:2066–2070.

Devey, M.E., Sewell, M.M., Uren, T.L., Neale, D.B., 1999, Comparative mapping in loblolly and radiata pine using RFLP and microsatellite markers. Theoretical and Applied Genetics 99:656–662.

Devey, M.E., Groom, K.A., Nolan, M.F., Bell, J.C., Dudzinski, M.J., Old, K.M., Matheson, A.C., Moran, G.F., 2004, Detection and verification of quantitative trait loci for resistance to Dothistroma needle blight in *Pinus radiata*. Theoretical and Applied Genetics 108:1056–1063.

Dodds, K.G., Montgomery, G.W., Tate, M.L., 1993, Testing for linkage between a marker locus and a major gene locus in half-sib families. Journal of Heredity 84:43–48.

Dupuis, J., Siegmund, D., 1999, Statistical methods for mapping quantitative trait loci from a dense set of markers. Genetics 151:373–386.

Dvornyk, V., Sirviö, A., Mikkonen, M., Savolainen, O., 2002, Low nucleotide diversity at the *pal1* locus in the widely distributed *Pinus sylvestris*. Molecular Biology and Evolution 19:179–188.

Echt, C.S., Vendramin, C.D., Nelson, C.D., Marquardt, P., 1999, Microsatellite DNA as shared genetic markers among conifer species. Canadian Journal of Forest Research 29:365–371.

Epperson, B.K., Allard, R.W., 1987, Linkage disequilibrium between allozymes in natural populations of lodgepole pine. Genetics 115:341–352.

Evans, R., 1994, Rapid measurement of transverse measurements of tracheids in radial wood specimens of *Pinus radiata*. Holzforschung 48:168–172.

Evans, R., Kibblewhite, R.P., Stringer, S., 1999, Variation of microfibril angle, density and fibre orientation in twenty-nine *Eucalyptus nitens* trees. Appita Journal 50:487–494.

Geburek, T., 1998, Genetic variation of Norway spruce (*Picea abies* [L.] Karst.) populations in Austria 1. Digenic disequilibrium and microspatial patterns derived from allozymes. Forest Genetics 5:221–230.

Germer, S., Holland, M.J., Higuchi, R., 2000, High-throughput SNP allele-frequency determination in pooled DNA samples by kinetic PCR. Genome Research 10:258–266.

Gupta, P.K., Rustgi, S., Kulwal, P.L., 2005, Linkage disequilibrium and association studies in higher plants: present status and future prospects. Plant Molecular Biology 57:461–485.

Howe, G.T., Aitken, S.N., Neale, D.B., Jermstad, K.D., Wheeler, N.C., Chen, T.H.H., 2003, From genotype to phenotype: unravelling the complexities of cold adaptation in forest trees. Canadian Journal of Forest Research 33:1247–1266.

Huntley, S.K., Ellis, D., Gilbert, M., Chapple, C., Mansfield, S.D., 2003, Significant increases in pulping efficiency in C4H–F5H transformed poplars: improved chemical savings and reduced environmental toxins. Journal of Agricultural Food Chemicals 51:6178–6183.

Ingvarsson, P.K., 2005, Nucleotide polymorphism and linkage disequilibrium within and among natural populations of European aspen (*Populus tremula* L., Salicaceae). Genetics 169:945–953.

Jayawickrama, K.J.S., Carson, M.J., 2000, A breeding strategy for the New Zealand Radiata Pine Breeding Co-operative. Silvae Genetica 49:82–90.

Johnson, G.R., Burdon, R.D., 1990, Family-site interaction in *Pinus radiata*: implications for progeny testing strategy and regionalised breeding in New Zealand. Silvae Genetica 39:55–62.

Johnson, G.R., Wheeler, N.C., Strauss, S.H., 2000, Financial feasibility of marker-aided selection in Douglas-fir. Canadian Journal of Forest Research 30:1942–1952.

Jones, L., Ennos, A.R., Turner, S.R., 2001, Cloning and characterization of irregular xylem4 (irx4): a severely lignin-deficient mutant of *Arabidopsis*. The Plant Journal 26:205–216.

Kerr, R.J., Goddard, M.E., 1997, Comparison between the use of MAS and clonal tests in tree breeding programmes. In: IUFRO '97 Genetics of Radiata Pine (Eds. R.D. Burdon, J.M. Moore) pp. 297–303. Proceedings of NZFRI/IUFRO Conference 1–4 December and Workshop 5 December, Rotorua, New Zealand FRI Bulletin No. 203.

Kinloch, B.B., Parks, G.K., Flower, C.W., 1970, White pine blister rust: simply inherited resistance in sugar pine. Science 167:193–195.

Kirst, M.E., Myburg, A.A., Sederoff, R.R., 2003, Genetical genomics of *Eucalytptus*: combining expression profiling and genetic segregation analysis (Abstr.). Plant and Animal Genome XI, 11–15 January 2003, San Diego, CA.

Kirst, M., Myers, R.M., De León, J.P.G., Kirst, M.E., Scott, J., Sederoff, R., 2004, Coordinated genetic regulation of growth and lignin revealed by quantitative trait locus analysis of cDNA microarray data in an interspecific backcross of eucalyptus. Plant Physiology 135:2368–2378.

Krutovsky, K.V., Neale, D.B., 2005, Nucleotide diversity and linkage disequilibrium in cold hardiness and wood quality related candidate genes in Douglas-fir. Genetics 171:2029–2041.

Kuang, H., Richardson, T.E., Carson, S.D., Bongarten, B., 1999, Genetic analysis of inbreeding depression in plus tree 850.55 of *Pinus radiata* D. Don. II. Genetics of viability genes. Theoretical and Applied Genetics 99:140–146.

Kumar, S., Echt, C.S., Wilcox, P.L., Richardson, T.E., 2004, Testing for linkage disequilibrium in the New Zealand radiata pine breeding population. Theoretical and Applied Genetics 108:292–298.

Lagercrantz, U., Ryman, N., 1990, Genetic structure of Norway spruce (*Picea abies*): concordance of morphological and allozymic variation. Evolution 44:38–53.

Long, A.D., Langley, C.H., 1999, The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. Genome Research 9:720–731.

Luo, Z.W., 1998, Detecting linkage disequilibrium between a polymorphic marker locus and a trait locus in natural populations. Heredity 80:198–208.

Lynch, M., Walsh, B., 1997, Genetics and analysis of quantitative traits. Sinauer Associates, Sunderland, MA.

Michelmore, R.W., Paran, I., Kesseli, R.V., 1991, Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. Proceedings of the National Academy of Sciences of the United States of America 88:9828–9832.

Mitton, J.B., 1992, The dynamic mating system of conifers. New Forests 6:187–216.

Mitton, J.B., Sturgeon, K.B., Davis, M.L., 1980, Genetic differentiation in ponderosa pine along a steep elevational transect. Silvae Genetica 29:100–103.

Morgante, M., Salamini, F., 2003, From plant genomics to breeding practice. Current Opinion in Biotechnology 14:214–219.

Muona, O., Szmidt, A.E., 1985, A multilocus study of natural populations of *Pinus sylvestris*. In: Lecture notes in Bioinformatics. (Ed H.-R. Gregorious) pp. 226–240. Springer Verlag, Berlin.

Murray, B.G., 1998, Nuclear DNA amounts in gymnosperms. Annals of Botany 82(Supplement A):3–15.

Paran, I., Zamir, D., 2003, Quantitative traits in plants: beyond the QTL. Trends in Genetics 19:303–306.

Paux, E., Tamasloukht, M.B., Ladouce, N., Sivadon, P., Grima-Pettenati, J., 2004, Identification of genes preferentially expressed during wood formation in *Eucalyptus*. Plant Molecular Biology 55:263–280.

Plomion, C., Richardson, T.E., MacKay, J., 2005, Advances in forest tree genomics: forest trees workshop, plant and animal genome XIII conference, San Diego, CA, January 2005. New Phytologist 166:713–717.

Pot, D., McMillan, L.K., Echt, C.S., Le Provost, G., Garnier-Gere, P., Cato, S.A., Plomion, C., 2005, Nucleotide variation in genes involved in wood formation in two pine species. New Phytologist 167:101–112.

Powers, H.R., Hubbard, S.D., Anderson, R.L., 1982, Resistance to diseases and pests in forest trees. In: Proceedings of Third International Workshop on Genetics of Host-Parasite Interactions in Forestry (Eds. H.M. Heybroek, B.R. Stephan, K. von Weissenberg). pp. 427–434. Pudoc, Wageningen, The Netherlands.

Pritchard, J.K., Rosenberg, N.A., 1999, Use of unlinked genetic markers to detect population stratification in association studies. American Journal of Human Genetics 65:220–228.

Pritchard, J.K., Stephens, M., Rosenberg, N.A., Donnelly, P., 2000, Association mapping in structured populations. Genetics 67:170–181.

Rafalski, J.A., Morgante, M., 2004, Corn and humans: recombination and linkage disequilibrium in two genomes of similar size. Trends in Genetics 20:103–111.

Roberds, J.H., Brotschol, J.V., 1985, Linkage disequilibrium among allozyme loci in natural populations of *Liriodendron tulipifera* L. Silvae Genetica 34:137–141.

Russell, J.H., Burdon, R.D., Yanchuk, A.D., 2003, Inbreeding depression and variance structures for height and adaptation in self- and outcross *Thuja plicata* families in varying environments. Forest Genetics 10:171–184.

Salvi, S., Tuberosa, R., 2005, To clone or not to clone plant QTLs: present and future challenges. Trends in Plant Sciences 10:1360–1385.

Salvi, S., Sponza, G., Morgante, M., Tomes, D., Tuberosa, R., 2006, Confirmation of the maize flowering time QTL *Vgt1* by association mapping (Abstr.). Plant and Animal Genome XIV, 14–18 January 2006, San Diego, CA.

Sewell, M.M., Neale, D.B., 2000, Mapping quantitative traits in forest trees. In: Molecular biology of woody plants, forestry Sciences (Eds. S.M. Jain, S.C. Minocha) pp. 407–433. Kluwer Academic Publishers, The Netherlands.

Strauss, S.H., Lande, R., Namkoong, G., 1992, Limitations of molecular marker-aided selection in forest tree breeding. Canadian Journal of Forest Research 22:1050–1061.

Stromberg, L.D., Dudley, J.D., Rufener, G.K., 1994, Comparing conventional early generation selection with molecular marker assisted selection in maize. Crop Science 34:1221–1225.

Telfer, E.J., Echt, C.S., Nelson, C.D., Wilcox, P.L., 2006, Comparative mapping in *Pinus radiata* and *P. taeda* reveals co-location of wood density-related QTL (Abstr.). Plant and Animal Genome XIV, 14–18 January 2006, San Diego, CA.

Thornsberry, J.M., Goodman, M.M., Doebley, J., Kresovich, S., Nielsen, D., Buckler, E.S., 2001, *Dwarf8* polymorphisms associate with variation in flowering time. Nature Genetics 28:286–289.

Thumma, B.R., Nolan, M.F., Evans, R., Moran, G.F., 2005, Polymorphisms in *Cinnamoyl CoA Reductase* (*CCR*) are associated with variation in microfibril angle in *Eucalyptus* spp. Genetics 171:1257–1265.

Wayne, M.L., McIntyre, L.M., 2002, Combining mapping and arraying: an approach to candidate gene identification. Proceedings of the National Academy of Sciences of the United States of America 99:14903–14906.

Wilcox, P.L., Amerson, H.V., Kuhlman, E.G., Liu, B.-H., O'Malley, D.M., Sederoff, R.R., 1996, Detection of a major gene for resistance to fusiform rust disease in loblolly pine by genomic mapping. Proceedings of the National Academy of Sciences of the United States of America 93:3859–3864.

Wilcox, P.L., Richardson, T.E., Carson, S.D., 1997, Nature of quantitative trait variation in *Pinus radiata*: insights from QTL detection experiments. In: IUFRO '97 Genetics of Radiata Pine (Eds. R.D. Burdon, J.M. Moore) pp. 304–312. Proceedings of NZFRI/IUFRO Conference 1–4 December and Workshop 5 December, Rotorua, New Zealand FRI Bulletin No. 203.

Wilcox, P.L., Carson, S.D., Richardson, T.E., Ball, R.D., Horgan, G.P., Carter, P., 2001, Benefit-cost analysis of DNA marker-based selection in progenies of *Pinus radiata* seed orchard parents. Canadian Journal of Forest Research 31:2213–2224.

Williams, C.G., Savolainen, O., 1996, Inbreeding depression in conifers: implications for breeding strategy. Forest Science 42:102–117.

Wright, S.I., Gaut, B.S., 2005, Molecular population genetics and the search for adaptive evolution in plants. Molecular Biology and Evolution 22:506–519.

Wu, R., Zeng, Z.-B., 2001, Joint linkage and linkage disequilibrium mapping in natural populations. Genetics 157:899–909.

Wu, R., Ma, C.-X., Casella, G., 2002, Joint linkage and linkage disequilibrium mapping of quantitative trait loci in natural populations. Genetics 160:779–792.

Yin, T.M., DiFrazio, S.P., Gunter, L.E., Jawdy, S.S., Boerjan, W., Tuskan, G.A., 2004, Genetic and physical mapping of *Melampsora* rust resistance genes in *Populus* and characterization of linkage disequilibrium and flanking genomic sequence. New Phytologist 164:95–105.

Yu, J., Pressoir, G., Briggs, W.H., Bi, I.V., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Kresovich, S., Buckler, E.S., 2006, A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature Genetics 38:203–208.

Chapter 11

PROSPECTS OF ASSOCIATION MAPPING IN PERENNIAL HORTICULTURAL CROPS

Erik H.A. Rikkerink[1], Nnadozie C. Oraguzie[2] and Susan E. Gardiner[3]

11.1 INTRODUCTION

Many horticultural crops share several characteristics that complicate genetic analysis including long generation intervals, protracted evaluation times, high costs of breeding inputs, slow maturation, and polyploidy. Partly as a result of the limited economic impact of individual species new technologies have been incorporated into breeding strategies in horticultural crops at a relatively slow pace. As outlined in previous sections, association mapping has only recently begun to be applied to plants and there is no published data yet on horticultural crops. We outline here the characteristics of perennial horticultural species that impinge on the application of association mapping, assess the potential impact of this technology and propose some guidelines for incorporating association mapping into conventional horticultural breeding programs. For the sake of simplicity, we will illustrate many of the points in the discussion with examples from the Rosaceae, a family of plants that has a diverse range of uses and ranks third in economic importance in the USA and temperate regions throughout the world. The most valuable fruit producing crops in this family include apple (*Malus*), pear (*Pyrus*), stonefruit (*Prunus*), and strawberry (*Fragaria*). The Rosaceae also contain a wide variety of ornamental plants including roses (*Rosa*), flowering cherry (*Prunus*), crabapple (*Malus*), and quince (*Cydonia*).

[1] The Horticulture and Food Research Institute of New Zealand Limited (HortResearch), Mount Albert Research Centre, Private Bag 92169, Auckland, New Zealand

[2] HortResearch, Hawke's Bay Research Centre, Private Bag 1401, Havelock North, New Zealand

[3] HortResearch, Palmerston North Research Centre, Private Bag, Palmerston North, New Zealand

11.2 CHARACTERISTICS OF PERENNIAL HORTICULTURAL CROPS

Horticultural crops share some characteristics with each of the crop groups covered in Chapters 9 and 10. Since some of these properties have already been discussed we simply outline the major areas of commonality and difference between the crop groups in Table 11.1. We present an overview of the combination of characteristics which are more or less peculiar to horticultural crops and then go on to outline in more detail where horticultural crops differ in their status and/or biological nature as it impinges on the application of association mapping.

Table 11.1. A comparison of the major characteristics of crop species

	Crop Group		
Characteristic	**Forage/Agronomy**	**Forestry**	**Perennial Horticulture**
Economic impact	Several major impact species	Several major impact species	Large number of moderate impact species
Breeding systems	Both in-breeding and out-breeding	Largely out-breeding	Largely out-breeding
Generation intervals	Annual and perennial, months	Perennial, years	Perennial, years
Maintenance and testing costs	Varied, commonly smaller plants and lower costs	Large trees, expensive	Shrubs to large trees, moderate to expensive
Ploidy	Diploid and polyploidy	Mostly diploid	Polyploidy common
Genome size	Small to large	Moderate to large	Small to moderate

The most valuable group of horticultural crops, on a gross margin per hectare basis, is the fruit crops. From a breeder's point of view, fruit crops differ from most agronomic or forest crops because of a peculiar combination of features including high heterozygosity, asexual propagation, their perennial nature, and the perishability of their products. At the same time these attributes make them attractive candidates for marker-assisted and/or gene-assisted selection. Most fruit crops maintain high levels of heterozygosity in individuals and an allelic richness in their primary germplasm pools. There are, however, some important exceptions to this generalization (such as peach within the Rosaceae). In nature diversity is maintained by various mechanisms that actively promote out-crossing. Such a high degree of diversity might be disadvantageous if horticulture relied on the sexual cycle to generate the individual plants that yield product. However fruit crops were amongst the first plants where techniques of asexual propagation were discovered and utilized. In most cases, production can rely on asexual propagation of individuals that enables the fruit breeder to exploit all the genetic effects, additive and non-additive as they are expressed in the phenotypes of superior individuals. These crops are mostly perennial with many featuring large plant size, long productive period, an extended juvenile phase for seedlings, and a marketable product that cannot be assessed until a seedling is physiologically mature. Added complications derive from

multiple biotic and abiotic factors that can affect both quality and quantity during both preharvest and postharvest periods.

11.2.1 Economic Impact

Horticultural crops are extremely versatile and exhibit a great diversity in terms of their applications and the habitats that they require for successful cultivation. Their range of uses include: fresh fruit, processed fruit, juices, ornamental crops, food extracts and additives, convenience foods, and specialty health foods. An unfortunate negative side effect of this versatility is that the low to moderate economic impact of any one single species is a factor that differentiates horticultural crops from the other groupings. While the economic impact has no direct bearing on the ability to apply a particular scientific approach (such as association mapping) it has a major influence on cost–benefit considerations. Although none of the horticultural crops would feature in the main tier of crops at the international level, several of them do feature as critically important crops for particular nation states, or large regions within the larger nation states. Perennial horticultural crops that constitute major exporting crops in major geographical regions include: apple in Washington State (USA) and New Zealand, citrus in Brazil and Florida (USA), stonefruit in Spain, kiwifruit (mainly *A. deliciosa* and *A. chinensis* cultivars) in New Zealand, Chile, Italy, and France, pear (European) in Australia, UK, and France, and nashi pear in Japan. Consequently there are active research programs in these crops that usually include the advanced molecular genetic techniques that are a significant component of the prerequisites for association mapping. In addition there are some species groupings within the horticultural crops (most notably those that belong to the Rosaceae family) that may, in some respects at least, be treated as a unitary genetic system since they demonstrate a degree of co-linearity (Dirlewanger *et al.* 2004). The progress made in the genome analysis of the grasses is partly aided by the fact that they can be treated as a single genetic system (Bennetzen and Freeling 1993) and a strategy of a multi-species unitary genetic system has been proposed and widely accepted within the Rosaceae research community as one of the ways around the economic impact factor. We do not yet know however how well (or if at all) association mapping or any of its design components can be “transported” across some of these plant families. Major biological differences between members of the same family or even a single species (such as self-fertilizing and out-breeding members in several important *Prunus* species including sweet cherry, almond, and Japanese apricot), may play an overriding role in determining the ability to apply association mapping principles developed in one subgroup to others.

11.2.2 Breeding Characteristics

Most fruit breeding programmes can be represented as two-stage selection programs (see Luby and Shaw 2001). In stage 1, large populations of non-replicated individuals are evaluated and a small proportion is selected for extensive asexual propagation for stage 2 testing in replicated trials. The large plant size and long life cycle, and especially a long juvenile period, have the greatest negative impact on cost and time efficiency of fruit breeding programs during stage 1. This requires large areas of land for plant maintenance, and fruit evaluation is also labour intensive. The land and labour may be required for many years and at the end of the process a high proportion of inferior seedlings are destined for culling. Stage 2, though usually more land- and labour-intensive

per genotype evaluated, is focused on a drastically reduced number of elite genotypes because of the intense selection in stage 1. This two-stage independent culling (tandem selection) constrains breeding opportunities more so than crops in which single stage selection is practical.

The fruit breeder requires selected individuals to exceed a certain level of performance (or culling level), for each of a host of traits desired by producers, processors, and consumers. The culling levels for most traits are usually independent of one another. A common breeding practice is to weight each trait relative to its importance for the commercial success of a cultivar. Most fruit breeders also take advantage of multiple stage selection by emphasizing a limited number of traits when first evaluating non-replicated seedlings, and then considering the full suite of traits in advanced testing of clonally replicated genotypes. The simultaneous selection for multiple oligogenic or polygenic traits ensures that only a small proportion of individuals will have favourable alleles at a large enough number of loci to be judged superior, but necessitates evaluation of large (stage 1) populations to increase the probability of obtaining and identifying these superior individuals.

In-breeding depression was first recognized as an evolutionary force by Darwin. Perhaps because of the importance of in-breeding depression, out-breeding has been a very common adaptation amongst plants in general and in perennial horticultural crops in particular. Although the majority of flowering plants are hermaphrodites and therefore potentially capable of selfing, they have evolved a number of different mechanisms to eliminate or control the degree of in-breeding (Dellaporta and Calderon-Urrea 1993), underlining the importance of controlling in-breeding. These mechanisms include sexual dimorphism (dioecy), separation of the male and female reproductive organs in space (monoecy and herkogamy) or time (dichogamy), and gametophytic or sporophytic self-incompatibility systems. These different behaviours will almost certainly have a significant influence on the nature of linkage disequilibrium (LD) and therefore on the strategies adopted for association mapping. Some correlation can be found between plants that largely disperse their seed with the aid of birds (mainly small nuts and berries) and dioecy. One hypothesis is that dispersal efficiency can be increased by only developing fruit in half the plants, and in fact these higher dispersal rates may be required for dioecy to be an advantage over other forms of sexual behaviour (Barot and Gignoux 2004). There may also be a link with the perennial nature of most fruits as their long-lived nature is another factor that can help overcome seed dispersal difficulties associated with the dimorphic state. The fruiting (female) plants can produce higher densities of fruit since they need not waste resources on producing pollen. This hypothesis could also be extended to other, less polar, forms of sexual dimorphism (e.g. gynodioecy). This could be why there are more examples of sexual dimorphism in fruit crops (strawberry, grape, papaya, kiwifruit) than might be expected by chance. There are likely to be different consequences on populations depending on the exact nature of the in-breeding control, e.g. whether it is partial or complete and depending on the type of control mechanism. Self-incompatibility systems can encode a large number of different alleles. In these cases there is an inherent advantage for individuals carrying rare alleles which leads to a type of balancing selection dependent on frequency (Charlesworth *et al.* 2005).

Breeding characteristics can vary even in closely related crops, such as the example of the self-compatible and self-incompatible *Prunus* members referred to above. This means that, in theory, quite different strategies of experimental design may be required for even closely related crops. The comparison between in-breeding and out-breeding

Prunus will be particularly interesting. Self-compatible behaviour in this genus appears to segregate as a single gene trait. The probable molecular basis of the difference between at least some of the selfing and non-selfing members amongst *Prunus* species has recently been determined by correlation with a defective pollen S component that is physically linked to the pistil S component (Ushijima *et al.* 2004). This suggests that self-compatibility may be a relatively recent adaptive event in *Prunus* and that their ancestors were self-incompatible. A comparison of LD between compatible and incompatible *Prunus* members could therefore be very interesting. Comparisons like this will require considerable effort and careful planning of sampling. This is illustrated by a comparison of haplotype structure between selfing *Arabidopsis thaliana* populations and the closely related self-incompatible *A. lyrata* that indicates that the effect of complex population history can make it very difficult to draw conclusions from the analysis of limited datasets (Wright *et al.* 2003).

To date, existing evidence appears to suggest that out-breeders generally have a lower degree of LD than in-breeders (see discussion comparing maize and *Arabidopsis* in Chapter 2). One of the most obvious effects of out-breeding behaviour is the rapid dissociation of multi-gene complexes (Dobzhanksy 1972) unless genetic linkage of the alleles in question is maintained. Therefore there might be a selective advantage for out-breeding crops to maintain higher rates of LD in areas of the genome where important complexes of alleles that rely on genetic linkage drag reside. There is now some evidence that loose clustering of genes from the same biosynthetic pathway does occur in plant genomes (Lee and Sonnhammer 2003) and the hypothesis is that these are maintained by selective pressure. If that is the case, this might result in a mosaic of regions of high and low LD across the genome that, in turn, could have important consequences on our ability to detect statistically significant differences in LD in these regions. It should be stressed however, that LD data exists for relatively few plants and it is dangerous to draw too many conclusions based on a limited number of examples. It is also dangerous to draw conclusions about LD in the entire genome if the data have been generated by analysis of LD at a small number of specific loci. In the case of out-breeding species it will be particularly important to develop some understanding of how LD varies across the genome of several species with different biological behaviours, before we may be able to extrapolate from behaviour to genome wide LD with any degree of confidence.

The nature of the in-breeding control mechanism may also have an unusual effect on LD around the regions of the genome encoding the specific biological trait that limits in-breeding. The most extreme examples of linkage drag are the sex determining chromosome of some organisms displaying sexual dimorphism. It is known that sex chromosomes in mammals show a high degree of degeneration and near perfect LD along long stretches of the chromosome (Sykes 2003). This has probably occurred as a result of eliminating crossing-over in these regions so that a set of linked genes that determine sex can be perfectly co-inherited to maintain the sexual dimorphism state. Forms of dioecy are utilized by at least five members of the horticultural perennials namely grape (Dalbo *et al.* 2000), kiwifruit (Harvey *et al.* 1997), some strawberry species (Ashman 1999), figs (Weiblen *et al.* 2001), and papaya (Liu *et al.* 2004). Although in plants the dimorphic behaviour is not as "pure" as in animals and often includes intermediate states between dioecy and hermaphroditism, there is now growing evidence that repression of recombination, and therefore possibly also degeneration events, occur in plant sex chromosomes (Guttman and Charlesworth 1998; Charlesworth and Guttman 1999; Liu *et al.* 2004). Whether comparable events are occurring near loci such as the self-incompatibility system is not known,

although there is some evidence to suggest that recombination may be unusually low in this region in some species at least (Wang *et al.* 2003). Since good candidates for the pollen determinant of self-incompatibility have recently been identified near the S-RNase (pistil determinant) self-incompatibility locus in a number of different crops (Lai *et al.* 2002; Ushijima *et al.* 2004), it should now be possible to look for repression of recombination in the region between or around these genes. Given that the degree of LD in low recombination regions might be expected to be unusually high compared with the rest of the genome, the overriding effect of the sex locus will need to be taken into account when it comes to analysing LD.

11.2.3 Generation Intervals, Maintenance, and Testing Costs

Many of the horticultural crops with the largest economic impact have comparatively long generation times, particularly the horticultural tree crops. While these generation times are a distinct disadvantage in terms of the speed at which scientific progress can be made, the flip side of this coin does imply that the impact of incorporating major time saving measures into breeding programs can have a relatively greater impact on progress in these crops. The possible impact of association mapping and LD analysis in horticultural crops is therefore significant. The generation time also has a major impact on the maintenance and testing costs. In particular the long juvenility periods of some of the crops in question require several seasons of growth before gathering of the phenotypic data that breeders will ultimately require to be able to predict or manipulate with molecular markers and techniques. The quality of the phenotypic data is one of the major limiting factors determining the potential for successful identification of real associations, because of the susceptibility of phenotype to a range of environmental influences. For some characteristics extra care will be required to collect phenotypic data over several seasons and several locations to enable researchers to differentiate strictly environmental influences from genetic influences.

For most perennial fruit crops, the cost of carrying seedlings in a breeding program is much greater than for annual crops. Thus selection during the early juvenile phase produces the maximum cost savings (Luby and Shaw 2001). Many of the costs associated with plant maintenance are connected with some of the above characteristics and with generation time in particular. These costs are not helped by the fact that many of the crops are large trees or shrubs during their productive lifespan and therefore require large amounts of land and constant human intervention in the form of pruning, fertilizing, and preventative treatment for pests and diseases, etc. The smaller berry crops, such as blueberry in the Ericaceae, grape in the Vitaceae, and strawberry and raspberry in the Rosaceae, do not utilize the same amount of orchard space but still require a high degree of intervention to ensure a high quality product. The testing of the fruit products is another significant expense. Fruit is highly perishable and is subject to numerous interactions of genetic effects with demanding consumer organoleptic and texture preferences.

11.2.4 Genome and Genomics Status

The ploidy, genome size, and status of genomics analysis in a particular crop will significantly influence how and when association mapping can be applied. Polyploidy is quite common in horticultural crops and many of the most important crops are either

predominantly polyploid (e.g. many of the *Prunus*, *Fragaria*, and *Actinidia* species) or can be polyploid (e.g. *Malus* species). Fortunately genome sizes in the horticultural crops are often of moderate size and can sometimes be very small. For example the small haploid genome size of diploid *Fragaria* species is comparable to model species such as *Arabidopsis* (Antonius and Ahokas 1996, Sargent *et al.* 2004). There is evidence that several of the species with larger genome sizes (such as the Maloideae) are probably cryptic polyploids – making their effective haploid genome size in terms of gene content smaller than the nuclear DNA content data indicate. However, it is likely that neither polyploidy or cryptic polyploidy will simplify any of the association mapping strategies, as both types of polyploids may well have taken advantage of genome duplication events to generate extra levels of specialization/adaptation of genes at homologous loci. For any genome scanning based approaches (see below) the absolute genome size will need to be taken into account (together with the structure of LD in that genome) when deciding on the number of markers that will be needed to give adequate coverage of the genome. In the true polyploid, in particular, the inability to readily identify linkage phase and thus chromosome haplotypes may well hamper subsequent analysis. It may also result in much more complex inheritance behaviour. The impact of phenomena such as polysomic inheritance and double reduction on association mapping is largely uncharted territory. Indeed the novel effects of polyploidy are only now beginning to be addressed in pedigree based linkage analysis (Luo *et al.* 2004) and fingerprinting studies (De Silva *et al.* 2005).

The feasibility of developing a sufficient number of markers to match the genome coverage required will depend on the resources that are already available in the species of interest (see Table 11.2). The existence of a large number of EST sequences, e.g. in crops such as grape and apple, offer a route by which large numbers of single nucleotide polymorphisms (SNPs) can be identified – particularly if the data are derived from a number of genotypes and/or from out-breeding species (as are most of these crops). Existing genetic maps with a number of genome anchoring markers such as microsatellites and/or RFLPs will enable researchers to integrate the genome position of these SNPs efficiently with any association data generated – thus enabling a rapid integration of LD mapping and more traditional genetic mapping approaches. The markers and map positions can also be used to assess the level of disequilibrium amongst alleles of unlinked markers and linked markers. In this way, a picture of the background level of disequilibrium across the genome that is not associated with genetic linkage can be developed. These markers can also be used to assess population structure, which can have a major influence on LD. The most significant horticultural perennials namely banana, apple, citrus, grape, and peach have already developed many of these genetic and genomics resources, but there are many other less valuable crops where most of these basic resources are still missing.

Table 11.2. Genomics status of resources for a selection of valuable perennial horticultural species

[*1] Genome data are mostly derived from the Kew database of genome sizes (http://www.rbgkew.org.uk/cval/homepage.html) Release 3.0, December 2004. Usually data from the most valuable member of the genus is given, except in the case of strawberry where the value is from a diploid species and a few cases where the closest relative that has data from the same genus has been used to provide an approximate value

[*2] Nucleotide sequences from the genus lodged in Genbank as at September 2005

[*3] Development of a physical map of the genome is in progress; [*4] Significant numbers of ESTs known to exist outside the public domain; [*5] Since limited information on sequence status or genetic mapping technology is available for the genera *Cydonia* (quince), *Eriobotrya* (loquat), *Asimina* (pawpaw), *Annona* (custard apple/cherimoya), *Mangifera* (mango), *Olea* (olive), *Feijoa* (feijoas), or *Ficus* (figs), their totals have been combined under "*Others*".

Family/ *Genus*	Common names	Genome (Mb)[*1]	Genbank sequences[*2]	Sequencing status	Other resources
Rosaceae/ *Malus*	Apple, Crabapples	743	203,829	Extensive EST libraries, BACs and cosmid libraries, partially completed genome[*3]	Micro-arrays, several genetic maps
Rosaceae/ *Prunus*	Peach, Apricot, Cherry, Plum, Nectarine	262	43,975	BAC libraries, partially completed genome[*3]	Multiple genetic maps
Rosaceae/ *Fragaria*	Strawberry	98-164	7,385[*4]		Limited BACs, micro-arrays, several genetic maps
Rosaceae/ *Pyrus*	Pear, Nashi pear	496	664[*4]		Several genetic maps and some markers
Rosaceae/ *Rubus*	Raspberry, Blackberry, Boysenberry, Loganberry	294	237	BAC library	Some molecular markers and maps
Vitaceae/ *Vitis*	Grape (table and wine)	417	207,428	Extensive EST libraries, BAC libraries	Some genetic maps

Rutaceae/ *Citrus*	Orange, Lemon, Lime, Tangerine, Grapefruit	368	130,991	BAC libraries	Some genetic maps
Lauraceae/ *Persea*	Avocado	907	8,859		Some molecular markers
Ericaceae/ *Vaccinium*	Blueberry, Cranberry, Huckleberry	588–3,528	4,832[*4]		Some molecular markers
Bromeliaceae/ *Ananas*	Pineapple	539	5,703		Some molecular markers and maps
Musaceae/ *Musa*	Banana, Plantain	613	3,199[*4]	BAC libraries BAC-end sequence	Limited genetic map information
Grossulariaceae/ *Ribes*	Black & Red currants, Gooseberry	534	2,540		Limited genetic map information
Caricaceae/ *Carica*	Papaya	368	1,657	BAC library	Limited genetic map information
Actinidiaceae/ *Actinidia*	Kiwifruit	760	444[*4]	Partial OVERGO based mapping, BACs	Microarray, some genetic maps & markers
Others		441–1,911	1,313[*5]	-	-

11.3 THE POTENTIAL IMPACT OF ASSOCIATION MAPPING ON HORTICULTURAL CROPS

It is likely that association mapping will have the most immediate and largest impact on the tier of crops with the greatest economic value. There probably are ways that this impact can be spread across lower value crops. We would therefore expect association mapping strategies to be applied first in banana, grape, citrus fruit, apple, pineapple, and stonefruit. There are some technology transfer strategies that could benefit the lower value crops during the interim period when full scale association mapping technology remains beyond the reach of such crops (see later). There are likely to be three main separate (but partially linked) approaches involving association mapping that can be used to benefit these crops. The approach at one end of the spectrum concentrates on the improved delivery of markers that can be used for marker aided selection (MAS). At the other end of the spectrum is the use of whole genome scans in order to identify the allele(s) of the gene(s) responsible for a particular phenotype of interest. In between these approaches lies a candidate gene-based approach. We discuss these three approaches in detail below.

11.3.1 The Marker-Assisted Selection (MAS) Approach

Improvements to marker technology will probably cover several areas, including (but not limited to) closer markers and therefore a more accurate "predictability" content, more widespread applicability of markers across populations, and a revolution in the range of markers available and our ability to screen them accurately (and relatively cheaply) across large breeding populations of thousands of plants. This approach does not necessarily require large numbers of markers at the outset (although its power of delivering useful markers would certainly benefit from this) and could probably be applied to both low and higher value crops.

There is huge potential for MAS to speed up genetic improvement in perennial horticultural crops particularly, through reduction of generation interval/breeding cycle following juvenile phase selection. Linkage studies have been very successful in identifying markers for simply inherited traits in a number of fruit crops including blueberry, strawberry, peach, pear and apple (Gardiner *et al.* 2005; Mnejja and Arus 2006; Yamamoto *et al.* 2006). However, these traits can equally be selected for easily by the breeder based on individual phenotype. It would seem that the most practical and potentially powerful use of MAS in fruit breeding will be in recurrent backcrossing schemes to facilitate introgression of simply inherited traits and also accelerate return to recipient parent genome (Hospital *et al.* 1992; Tanksley and Nelson 1996). Recurrent backcrossing has been traditionally used to introgress genes for resistance into commercial apple (*Malus x domestica* Borkh) cultivars. In the simplest case of a single major disease resistance gene (*Vf* for scab resistance in apple caused by *Venturia inaequalis*) introgression, it took up to seven generations to develop varieties with economic potential (Bus *et al.* 2001). It is anticipated that use of MAS for combined foreground selection (to determine presence of introgressed gene) and background selection (to accelerate return to 'recipient' parent genotype at the other loci) will reduce

the breeding cycle by two or more generations. However, most studies on marker-assisted introgression in fruit crops are still in their early stages and it is difficult to judge the efficacy of this approach.

How MAS for complex traits based on QTL mapping will benefit fruit crops will depend on a number of factors including the genetic architecture of traits in question, the accuracy of the phenotyping technique, the type and size of populations used, the density and functionality of the genetic markers and the repeatability of results. The prospects are not promising following initial prognosis by Luby and Shaw (2001) summarized below. The mostly out-breeding nature, selection schemes employed by breeders and other life history characteristics of these crops would seem to negate or complicate maintenance of any marker–gene associations detected in their diverse gene pools. Hence, there will always be the need to establish and evaluate marker associations for each cross or each recombination cycle. The reason is simple-in the diverse gene pool of an out-crossing fruit crops there is less chance that a particular allele will be linked to a particular phenotype of interest in the germplasm at large. In contrast, marker–gene associations established in self-pollinated crops where most initial MAS studies were carried out would be expected to break down very slowly because of their mostly in-breeding behaviour and intense selection. In fruit crops particularly, in the case of simultaneous selection for multiple polygenic traits, marker-QTL linkages will be more uncertain because of varying degrees of repulsion/coupling linkage phases (particularly where genetic correlations between traits are low and probably negative) making LD in specific crosses more difficult to evaluate. Under these circumstances, it is to be expected that the same loci may not have the main influence on inheritance of the same traits in different parents. Also, the same marker alleles may not be segregating in different crosses as progenies can only inherit alleles from their parents. Therefore, it will be necessary to conduct separate marker-QTL linkage analysis for each cross or population for which MAS is used. In the case of markers that are developed in the light of sequence information (e.g. SNPs, SSRs) this may not always be so serious since the marker may be adaptable to detect other alleles. However, dominant markers with little sequence information such as AFLPs and RAPDs would be severely affected by these limitations. There may also be a large number of loci controlling a quantitative trait and these may be in different repulsion phase linkage arrangement in the two parents. In such cases it is likely that a high number of recombinants will be recovered in the F_1 populations and this would lead to poor resolution in QTL mapping. This poor resolution means that the methodology identifies large DNA segments with potentially hundreds of candidate genes. It can then take several more years to produce the populations for fine scale mapping to narrow the distance between marker and gene. This increases the cost of field testing and phenotyping further, particularly if the trees need to be bearing fruit to measure the critical phenotype(s). Identification of markers by association mapping strategies may well be less prone to some of the above problems. Theoretically the power of detection of quantitative traits may be improved by employing association mapping. On the other hand other problems such as population structure may mitigate some of the gains in power of detection. These gains may come at the expense of having to generate a much greater number of genotypes – but then the high throughput capabilities required for this are rapidly becoming a reality (see Chapters 3–5). It is also important to be cautious about equating all reported associations with success, as the real strength of published evidence may be highly variable (Ball 2005; and see Chapter 8). The resolution of some maps can definitely be improved by association mapping strategies – as these

strategies naturally encompass a far greater number of meioses than the segregating family approach. This is a particular advantage for large trees as discussed above. An additional advantage is that the populations utilized can be designed to target different levels of detailed mapping – thus permitting an incremental approach to improving the closeness of the association between marker and trait.

11.3.2 Whole Genome Scanning Approaches

A second approach involves using association mapping to identify the alleles of genes responsible for major important phenotypic variations. This may, to some degree, displace the "traditional" map-based cloning route for particular alleles of genes. In the first instance these traits will tend to be simply inherited traits of major benefit. There are a number of potential advantages of association mapping over mapping progenies derived from deliberate crosses. The most important of these advantages is the much greater genetic resolution that is potentially deliverable – in some experimental designs-particularly in out-breeders. When sufficient background information about the species has been generated and appropriate experimental design has been applied (see Chapter 8) it may be possible to "land" on the gene responsible for the phenotype in question. The possibility of success in this type of approach will depend on the accuracy of not only the marker scores but also, most crucially, of the phenotypic measurements that are being compared with marker information. This strategy would probably require the capability to perform whole genome scans for LD at some point (or alternatively some sort of combination of candidate gene and LD strategies in the interim, see below). Inevitably this would eventually require a large number of SNPs across the genome and a very high throughput method for mapping these SNPs. The number of polymorphic sites required is probably somewhere between 10,000 and 1,000,000 depending on the nature of the crop. In crops where the region in LD is measured in the kb range the numbers would need to be very large (and would currently be beyond the economic scope of most crops). By scanning markers that cover the entire genome for statistically significant associations between the phenotype under investigation and the markers, it may be possible to identify polymorphisms very closely linked to (or inside) the coding region of the gene responsible for the phenotype. There are now also haplotype variations of the single marker scanning approach that may be even more powerful at detecting significant associations.

While initial recommendations focused on increasing the density of markers available for whole genome scanning as a strategy to improve the outcomes of association mapping, as a result of the human genome analysis (which has piloted many of the advances in LD) there has been a growing realization that, after a point, less can be more. As the density of markers increases the propensity for type 1 errors (false positive associations) also increases. The strategy to maximize the benefits of association mapping requires a balancing act between generating type 1 and type 2 errors and where the balance lies should be determined to some extent by the cost of verification analyses. Like QTL analysis, association mapping approaches will require strategies for verifying any potential associations and therefore it becomes important to minimize type 1 errors to reduce the cost of unnecessary verification efforts being "wasted" on chance associations or associations due to population structure. This then has to be counterbalanced with the desire not to miss true positives (type 2 error). Since it is likely to require significant

investment in generating the sequence information upon which the subsequent large scale marker analysis would have to be based, it is likely that the genome scanning strategy will initially be limited to the more valuable tier of crops namely banana, grape, apple, citrus, and stone-fruits.

11.3.3 The Candidate Gene Approach

This strategy offers an intermediate level of commitment between the marker-assisted selection and genome scan approaches discussed above by enabling a much more limited application of markers in candidate gene regions to yield almost equal statistical power of detection to the "large-scale" application of whole genome scanning. A candidate gene based LD strategy recently identified an association between mutations in Cinnamoyl CoA reductase and a wood quality trait affecting wood stiffness and strength in *Eucalyptus* (Thumma *et al.* 2005). Candidate gene studies are likely to be very versatile and widely applicable – and perhaps limited more by the biological similarity of the "donor" and "receptor" systems for the character targeted for transfer than the degree of sequence similarity. This versatility comes from the fact that several factors, such as the relationship between the species that are the sources for the candidate gene(s) used as comparisons and the species where the candidates are sought, the expected evolutionary forces that operate on the genes in question (i.e. purifying versus diversifying selection), and prior knowledge about the degree of conservation of some gene families, can all be taken into account when searching for candidate genes to be used in such a strategy. Candidate gene approaches are also greatly accelerated by the availability of large scale EST sequence resources such as those in apple, grape, and citrus. If available, whole genome sequence information can be used to identify candidate genes in particular parts of the genome where genes of interest are known to reside based on initial association mapping data. These candidate genes can even be derived from a close relative as demonstrated by recent comparative mapping approaches that rely on synteny (Perovic *et al.* 2004). We would expect that the whole genome sequence of the first Rosaceae species, perhaps peach (which might be available within a few years), may be able to act as a springboard for candidate gene association mapping strategies for any members of the Rosaceae where a sufficient degree of synteny has been identified to putatively connect the mapped genome region with the syntenic region in the sequenced species.

11.3.4 Technology Transfer to Lower Value Crops

One route of knowledge and technical transfer to the less valuable horticultural crops will be by the identification of genes in the more valuable species followed by direct transfer of the genes to the "lower tier" species. This would be no different from the way that *Arabidopsis* has acted as a model system for all plant species. This could include either the direct transfer of the gene itself (which is likely to be quite successful in closely related species), or the identification of the likely homologue in the species into which the technology is being transferred. *Arabidopsis*, while of great value as a model system to all plant species, has some severe limitations when it comes to technology transfer into most major horticultural crops. Probably most importantly, it does not have many of the fruit characteristics that are important in major horticultural crops, and a secondary consideration is that it is not closely related to many of these crops. This means that some of the biological characteristics that are particularly important, and

perhaps "peculiar" to a horticultural crop, will require the development of model systems that are more closely related to the crop and share that particular important characteristic. Sometimes these traits are shared across several unrelated species – such as the growing importance of dwarfing rootstocks for many horticultural crops. At other times even closely related crops may have important differences that may limit the application of discoveries. For example, within the Rosaceae several different fruit types (achenes, drupes, pomes) can be recognized, and their evolutionary relationship is unclear (Morgan *et al.* 1994) suggesting that there could be limited applicability of at least some important fruit characteristics across these species. In fact there is now also evidence supported by DNA based phylogeny that suggests different fruit types have repeatedly evolved in distinct lineages (Knapp 2002). Another major difference between members of the Rosaceae for example is that it includes both climacteric fruits (such as apple) and non-climacteric fruits (such as strawberry). Even within the genus *Pyrus*, there are both climacteric (European pears – *P. communis)* and non-climacteric (Asian pear – *P. pyrifolia*) species. The influence of such important differences on the success of direct routes of technology transfer is as yet largely unknown. Another method of technology transfer (already covered above) is the application of candidate gene approaches that make use of data from other species.

11.4 STRATEGIES FOR ACCELERATING THE ADOPTION OF NEW TECHNOLOGIES IN HORTICULTURAL CROPS

In this section we will consider if there are ways that the adoption of association mapping based technologies can be accelerated. These strategies will need to include both scientific and collaborative strategies. Perhaps the best way to ensure adoption is to develop strategies for integrating research across the often fragmented research sectors in these crops at the levels of both research discipline and nation states. Integration of a number of disciplines is required from breeding through to molecular biology, physiology, pathology, chemistry, bioinformatics, and biostatistics to name just a few. The breadth of disciplines that potentially could play a role highlights the difficulty of achieving such integration. International collaboration is made more difficult by the fact that the scientists concerned often focus on research for competing industries, and have different research priorities.

The potential benefit of better integration/collaboration far outweighs the difficulties and this is being recognized in some sectors already. Within the Rosaceae, for example, a group of international researchers has met several times to attempt to integrate the various species into a multi-species "model system", where it is planned that the strength of the different species, in terms of their rate of progress in different areas, will be used to maximize the benefit for all species in the family. This type of structure may be particularly appropriate for crop groupings which, in their own right, do not cross the value threshold required to attract the level of investment that would enable the full power of association mapping technologies (such as whole genome scanning) to be applied. However, as a group these species may well make a much more attractive prospect for investment. Smaller scale integration may be possible when it comes to solving particular problems that affect several different crops and that may be able to rely on common underlying mechanisms in these crops. In these cases a model system for the problem could

be a focus of collaboration building on the particular strengths of that crop that are peculiar to the problem at hand.

A major effort is also needed to make the different scientific disciplines both accessible and understandable to all researchers. This can help to counteract the growing degree of specialisation of researchers as they are required to focus on narrower fields of interest in order to cope with the information explosion. A major role in this respect is likely to be carried out by the development of database systems, and their bioinformatic interfaces, to display and summarize information. These systems will also allow the data to be shared almost instantaneously between groups that may well be on opposite sides of the globe and thus can result in reducing the amount of unnecessary duplication in labour intensive steps such as genome annotation. The ultimate measure of success of association mapping strategies, however, will be their integration with breeding practices. A major strategy should be to develop ways of incorporating association mapping-based design principles, particularly into the more traditionally based breeding programs.

11.5 GUIDELINES FOR INCORPORATING ASSOCIATION MAPPING STRATEGIES INTO CONVENTIONAL HORTICULTURAL TREE BREEDING

Fruit breeders are interested in new techniques that can improve their genetic efficiency in selection and reduce the risk of failing to identify superior individuals. Such new techniques must, however, have a certain level of cost–benefit advantage over the older techniques, if they are to be widely adopted. Given the highly variable nature of horticultural crop species it is likely that multiple guidelines will need to be developed for these crops. It is not practical to attempt to do this here as this will require a high degree of specialized expertise for each crop concerned. There will probably be some common themes on which we will elaborate below. There are several different strategies for applying association mapping that requires different levels of resource commitment and we will present these in order from lowest to highest below. Different levels will be appropriate for different crops at different times and we suggest that each crop builds gradually to a situation where the full power of association mapping can be applied.

In most horticultural perennials we know nothing about the level of LD in the crops concerned and relatively little about population structure. In this situation we can at best develop hypotheses by extending what we know about the biology of these crops. An obvious prerequisite for applying LD based analysis in a crop that requires a modest investment of resources is to generate base-line data on population structure and the extent of LD across the genome. LD should also be analysed in populations consisting of various levels of known and/or deduced inter-relatedness of plants. If it is anticipated that a modified form of MAS using association mapping is likely to be the only affordable route of incorporation into breeding strategies, then a certain approach and experimental design is required. In this instance the focus could be on utilizing the LD that exists in breeding populations and among commercial varieties to accelerate selection approaches. In that case moderate density marker scans (1,000–5,000) may well give sufficient power to firstly detect, and then follow many of the useful associations in subsequent deliberate crosses. These would only require moderate throughput capabilities and could conceivably even utilize a mixture of different marker types including existing informative markers such as microsatellites and perhaps even RFLPs in the case of

candidate genes. The interpretation of these results will need some caution as different marker systems can deliver different levels of LD (see Section 2.5.3.1). These systems could be used to initiate low density genome scans for LD amongst markers in relatively narrow populations, perhaps even with existing breeding populations. We would expect the maximum extent of LD to be present in such populations given that they have had little time to approach equilibrium. Since these are the populations where we might first want to utilize the technology to aid breeding processes, this would also seem a reasonable place to start. If we can learn something about LD from pilot studies with existing markers and populations, this can then inform our decision making when it comes to developing the marker technology required for high density genome scans and the platforms to score these markers (see Chapter 5).

The next level of commitment of resources involves using the candidate gene approach. This can be used to "enrich" for markers that are likely to be linked to a particular phenotype and therefore requires the researcher and breeder to prioritize which phenotypes to screen for in the first instance. This approach is partly limited by the set of phenotypes for which candidate gene approaches are feasible (i.e. based on information correlating particular gene sequences or gene families with particular traits in other plant systems). Good candidates for this type of approach are genes such as resistance gene candidates – which have already been successfully used to simplify the identification of resistance genes by mapping in segregating families (Paal *et al.* 2004). There are also large gene families likely to be involved in controlling a large number of different traits (e.g. transcription factors and protein kinases) which might be able to be utilized in a "blind approach" to attempt to correlate alleles of particular genes with phenotype. This type of "blind" association method could then be the forerunner of a comprehensive application of whole genome scans – involving the most significant level of resource commitment.

As mentioned above there are some common themes that could be integrated into a guideline for incorporating association mapping strategies into more traditionally focused breeding programs. One place to start is with the germplasm that is maintained and exploited for breeding gains. The strategy adopted for incorporating gains into new varieties will have an important influence on the nature of the germplasm that can be exploited by association mapping. One important consideration is whether introducing genes by artificial gene transfer techniques is likely to be a viable addition to more traditional selection based strategies for cultivar improvement. Given that gene transfer methods are much less susceptible to linkage drag (in theory there would be no linkage drag at all if single genes are transferred) – the introgression of useful characters from a much wider germplasm base could be anticipated with these methods. This is particularly the case for the long-lived out-breeders which are common amongst this group of plants. In outcrossing plants classical introgression cannot be performed and pseudo-backcrosses have to substitute. If an artificial gene transfer strategy is adopted then a much wider germplasm base might be utilized for association mapping, than if it is deemed such a strategy is not yet viable. The wider germplasm base would also lend itself better to identifying genes by "landing" on the gene since much smaller regions would be expected to be in LD with phenotype across a wider germplasm base. While the investment required for such a strategy would be considerable for each crop, major components such as DNA sequencing are rapidly decreasing in cost and the potential for incorporating highly novel characteristics that could provide a very valuable point of

cultivar differentiation in the market is likely to offset these costs in the medium to long-term.

The above strategies are not mutually exclusive. In fact the MAS strategy will probably be the first adopted in most crops and it can then be gradually extended to encompass the candidate gene and eventually the full genome scanning strategy. In some cases a crop may be able to go directly to a candidate gene strategy, depending on the nature of the biological question being tackled.

How can LD mapping address some of the major issues in fruit breeding? The likelihood of success from a cost–benefit point of view will depend on the traits targeted, their mechanism of inheritance (simple, oligogenic, or polygenic), gene action/effect, and mating system (self-pollinated, out-crossing, asexual). These factors will then need to balance any potential cost reductions and/or price advantage that might result from the technology. In the context of cost reduction, the LD mapping approach may not need large numbers of segregating populations initially for mapping quantitative traits, thus costs associated with making crosses and maintaining large tree populations in the orchard may reduce. However, these advantages could initially be offset by the substantial upfront cost associated with establishing large databases of gene/marker sequence information and developing high throughput genotyping assays to facilitate scans for marker trait associations. In the case of crops like apple, citrus and peach where some of the sequence resources required to initiate the development of such a system are already available, these costs will be less (but still substantial).

Some of the cost of running assays may be mitigated through DNA pooling by combining phenotype extremes. This approach has recently been used successfully in humans (Butcher *et al.* 2005; Sham *et al.* 2002; Zeng *et al.* 2005) but would only be feasible for finding associations for a particular predetermined phenotype used to devise the pools. Depending on the breeding approach adopted and if it includes following traits for several successive generations – it is likely that many marker–trait associations established could hold across a number of generations, so costs could be spread accordingly. Another cost driver is technological advances. Recent advances in DNA technologies have made large scale EST sequencing efforts viable, and even a significant number of whole genome sequencing projects viable, when these were considered virtually unaffordable just ten years ago. These advances are likely to continue to reduce the cost of developing the required sequence databases and individual genotype assays further.

Marker techniques could also be targeted at traits which are more expensive to measure by other approaches (labour intensive field assessment, physiological, biochemical, physical, chemical, or consumer preference measurements). In the traditional breeding process many of these assessments would tend to be carried out at the second stage to reduce their costs. Association methods might allow for their incorporation into the first stage culling process, thus making the selection process much more efficient and allowing breeders to increase the population size from which the superior individuals are selected. As for other molecular marker technologies it may be possible to screen with markers at an earlier stage of plant growth (prefruiting seedlings) as well and this could save considerable amounts of orchard space and costs.

The initial financial hurdles for developing genome scanning capability for any particular crop are considerable. However, once developed and over the initial cost hurdles, markers are likely to be cheaper particularly when combining the multiple traits required to meet breeding objectives. Since it is a relatively new technology, an unknown

factor is to what extent this technology will be able to reduce the cost of other measurements and assessments without compromising the outcome. The strategy of application will need to be thought out carefully for each crop. Given that the biology of some horticultural crops may be quite unique compared with other organisms where association mapping has been applied already, it would seem prudent to test the success of some of these strategies first, before applying them on a large scale.

Accurate estimations of costs between association mapping and "competing" technologies are difficult as they need to take into account a complex series of interdependent costs including maintenance of plant material, phenotyping, data collection, population sizes required, and genotyping of individuals. Amos and Page (2001) developed methods for comparing the costs of detecting genetic factors by linkage and association mapping. Because they were developed for comparing mapping approaches in the human system they did not take into account factors peculiar to plant breeding such as the extra costs of germplasm maintenance. They determined that the cost effectiveness of LD methods was greater for traits with lower single-locus heritability, whereas family based linkage analysis appeared to be more cost effective for traits with high single-locus heritability.

In terms of generating a price advantage it is likely that the application of these refined techniques would significantly improve the chances of delivering superior genotypes ahead of other breeding programs, assuming they are not using similar techniques. Conversely any programs not using these more efficient techniques would run the risk of falling behind their competitors.

11.6 PROBLEMS AND QUESTIONS IN THE APPLICATION OF ASSOCIATION MAPPING TO DOMESTICATED CROPS

There are a number of unusual properties or questions associated with agricultural crops that have been caused by the intervention of humans in the natural selection process. We do not yet fully understand how these interventions have affected LD in domesticated crops. One example is the effect of the different types of mating systems in plants covered above. There are other questions, for example what is the influence of bottlenecks caused by human breeding or selection? Deliberate breeding is a relatively recent event since human agriculture has operated on a relatively small geological timescale. How have the bottlenecks caused by these efforts affected LD in the "domesticated germplasm" of horticultural and other crops? The influence of deliberate breeding on LD is even less significant in crops with long generation times since the opportunity to go through a large number of generations has simply not been there. Does this mean that relatively little distortion of the LD caused by selecting the initial breeding parents from larger wild populations have been added by subsequent deliberate crossing and selection?

A further question revolves around how one deals with problems caused by admixture in breeding populations? How do you detect that there has been a recent case of admixture that potentially could give rise to many "spurious associations"? There are some methods for assessing admixture using unlinked markers (Chapters 7 and 8) and perhaps these should be routinely applied to populations used for association mapping strategies. In some crops the tendency to avoid in-breeding may have artificially increased the extent of admixture. Is this likely to be more extreme where breeding has

included utilizing distinct (but closely related) species (e.g. crossing European and Asian Pears) which is not an uncommon strategy in plants? Wild sister species are often used as a strategy for introgressing new pathogen and insect resistance factors into their domesticated relatives. These types of questions will continue to offer fertile ground for research with both a practical and intellectual interest.

11.7 ACKNOWLEDGMENTS

E.H.A.R., N.C.O., and S.E.G. were funded by the Horticulture and Food Research Institute of New Zealand Ltd (HortResearch). We thank reviewers for their comments and the staff of the HortResearch Publication Unit for their editorial assistance. Genome size data reproduced with the permission of the Trustees of the Royal Botanic Gardens, Kew http://www.rbgkew.org.uk/cval/homepage.html, Bennett MD, Leitch IJ. 2005. Plant DNA C-values database (release 4.0, Oct. 2005).

11.8 REFERENCES

Amos, C.I., Page, G., 2001, Cost of linkage versus association methods. In: Rao D.C., Province M.A. (eds) Genetic Dissection of Complex Traits, pp. 213–221. Academic Press, San Diego, CA, USA.

Antonius, K., Ahokas, H., 1996, Flow cytometric determination of polyploidy level in spontaneous clones of strawberries. Hereditas 124: 285.

Ashman, T.L., 1999, Quantitative genetics of floral traits in a gynodioecious wild strawberry *Fragaria virginiana*: implications for the independent evolution of female and hermaphrodite floral phenotypes. Heredity 83: 733–741.

Ball, R.D., 2005, Experimental designs for reliable detection of linkage disequilibrium in unstructured random population association studies. Genetics 170: 859–875.

Barot, S., Gignoux, J., 2004, How do sessile dioecious species cope with their males? Theoretical Population Biology 66: 163–173.

Bennetzen, J.L., Freeling M., 1993, Grasses as a single genetic system: genome composition, collinearity and compatibility. Trends in Genetics 9: 259–261.

Bus, V., Brooking, L., Davis, L., Norling, C., Ranatunga, C., Gardiner, S., 2001. Accelerated breeding for apple. In: Halligan, L. (ed) The New Zealand Controlled Environment Laboratory (NZCEL) Workshop Proceedings. Use of Controlled Environments in Containment Research, June 2001, pp. 25–27.

Butcher, L.M., Meaburn, E., Knight, J., Sham, P.C., Schalkwyk, L.C., Craig, I.W., Plomin, R., 2005, SNPs, microarrays and pooled DNA: identification of four loci associated with mild mental impairment in a sample of 6000 children. Human Molecular Genetics 14: 1315–1325.

Charlesworth, D., Guttman, D.S., 1999, The evolution of dioecy and plant sex chromosome systems. In: Ainsworth C.C. (ed) Sex Determination in Plants, pp. 25–49. BIOS Scientific Publishers Ltd, Oxford, UK.

Charlesworth, D., Vekemans, X., Castric, V., Glemin, S., 2005, Plant self-incompatibility systems: a molecular evolutionary perspective. New Phytologist 168: 61–69.

Dalbo, M.A., Ye, G.N., Weeden, N.F., Steinkellner, H., Sefc, K.M., Reisch, B.I., 2000, A gene controlling sex in grapevines placed on a molecular marker-based genetic map. Genome 43: 333–340.

De Silva, H.N., Hall, A.J., Rikkerink, E., McNeilage, M.A., Fraser, L., 2005, Estimation of allele frequencies in polyploids under certain patterns of inheritance. Heredity [Epub ahead of print].

Dellaporta, S.L., Calderon-Urrea, A., 1993, Sex determination in flowering plants. Plant Cell 5: 1241–1251.

Dirlewanger, E., Graziano, E., Joobeur, T., Garriga-Caldere, F., Cosson, P., Howad, W., Arus, P., 2004, Comparative mapping and marker-assisted selection in Rosaceae fruit crops. Proceedings of the National Academy of Sciences of the United States of America 101: 9891–9896.

Dobzhanksy, T., 1972, The genetics of the evolutionary process. Columbia University Press, New York, NY, USA.

Gardiner, S.E., Bus, V.G.M., Rusholme, R.L., Chagné, D., Rikkerink, E.H.A., 2005, Apple. In: Kole C (ed) The Genomes: A Series on Genome Mapping, Molecular Breeding & Genomics. Science Publishers, Inc., Enfield, NH, USA, Plymouth, UK.

Guttman, D.S., Charlesworth, D., 1998, An X-linked gene has a degenerate Y-linked homologue in the dioecious plant *Silene latifolia*. Nature 393: 263−266.

Harvey, C.F., Gill, G.P., Fraser, L.G., McNeilage, M.A. 1997, Sex determination in *Actinidia*. 1. Sex-linked markers and progeny sex ratio in diploid *A. chinensis*. Sexual Plant Reproduction 10: 149–154.

Hospital, F., Chevalet, C., Mulsant, P., 1992, Using markers in gene introgression breeding programs. Genetics 132: 1199–1210.

Knapp, S., 2002, Tobacco to tomatoes: a phylogenetic perspective on fruit diversity in the Solanaceae. Journal of Experimental Botany 53: 2001–2022.

Lai, Z., Ma, W., Han, B., Liang, L., Zhang, Y., Hong, G., Xue, Y., 2002, An F-box gene linked to the self-incompatibility (S) locus of Antirrhinum is expressed specifically in pollen and tapetum. Plant Molecular Biology 50: 29–42.

Lee, J.M., Sonnhammer, E.L.L., 2003, Genomic gene clustering analysis of pathways in eukaryotes. Genome Research 13: 875–882.

Liu, Z., Moore, P.H., Ma, H., Ackerman, C.M., Ragiba, M., Yu, Q., Pearl, H.M., Kim, M.S., Charlton, J.W., Stiles, J.I., Zee, F.T., Paterson, A.H., Ming, R., 2004, A primitive Y chromosome in papaya marks incipient sex chromosome evolution. Nature 427: 348–352.

Luby, J.J., Shaw, D.V., 2001, Does marker-assisted selection make dollars and sense in a fruit breeding programme? HortScience 36: 872–879.

Luo, Z.W., Zhang, R.M., Kearsey, M.J., 2004, Theoretical basis for genetic linkage analysis in autotetraploid species. Proceedings of the National Academy of Sciences of the United States of America 101: 7040–7045.

Mnejja, M., Arus, P., 2006, Microsatellite transportability across Rosaceae crops. In: 3rd International Rosaceae Genomics conference, p. 57, War Memorial Conference Center, Napier New Zealand, 19–22 March 2006.

Morgan, D.R., Soltis, D.E., Robertson, K.R., 1994, Systematic and Evolutionary implications of *rbcL* sequence variation in Rosaceae. American Journal of Botany 81: 890–903.

Paal, J., Henselewski, H., Muth, J., Meksem, K., Menendez, C.M., Salamini, F., Ballvora, A., Gebhardt, C., 2004, Molecular cloning of the potato Gro1-4 gene conferring resistance to pathotype Ro1 of the root cyst nematode Globodera rostochiensis, based on a candidate gene approach. Plant Journal 38: 285–297.

Perovic, D., Stein, N., Zhang, H., Drescher, A., Prasad, M., Kota, R., Kopahnke, D., Graner, A., 2004, An integrated approach for comparative mapping in rice and barley with special reference to the Rph16 resistance locus. Functional & Integrative Genomics 4: 74–83.

Sargent, D.J., Davis, T.M., Tobutt, K.R., Wilkinson, M.J., Battey, N.H., Simpson, D.W., 2004, A genetic linkage map of microsatellite, gene-specific and morphological markers in diploid Fragaria. Theoretical and Applied Genetics 109: 1385–1391.

Sham, P., Bader, J.S., Craig, I., O'Donovan, M., Owen, M., 2002, DNA Pooling: a tool for large-scale association studies. Nature Reviews Genetics 3: 862–871.

Sykes, B., 2003, Adam's Curse – a future without men. Bantam, London.

Tanksley, S.D., Nelson, J.C., 1996, Advanced backcross QTL analysis: a method for the simultaneous discovery and transfer of valuable QTLs from updated germplasm into elite breeding lines. Theoretical and Applied Genetics 92: 191–203.

Thumma, B.R., Nolan, M.F., Evans, R., Moran, G.F., 2005, Polymorphisms in Cinnamoyl CoA Reductase (CCR) are associated with Variation in Microfibril Angle in Eucalyptus spp. Genetics: genetics.105.042028.

Ushijima, K., Yamane, H., Watari, A., Kakehi, E., Ikeda, K., Hauck, N.R., Iezzoni, A.F., Tao, R., 2004, The S haplotype-specific F-box protein gene, SFB, is defective in self-compatible haplotypes of Prunus avium and P. mume. Plant Journal 39: 573–586.

Wang, Y., Wang, X., McCubbin, A.G., Kao, T.H., 2003, Genetic mapping and molecular characterization of the self-incompatibility (S) locus in Petunia inflata. Plant Molecular Biology 53: 565–580.

Weiblen, G.D., Yu, D.W., Wes, S.A., 2001, Pollination and parasitism in functionally dioecious figs. Proceedings of Biological Science 22: 651–659.

Wright, S.I., Lauga, B., Charlesworth, D., 2003, Subdivision and haplotype structure in natural populations of *Arabidopsis lyrata*. Molecular Ecology 12: 1247–1263.

Yamamoto, T., Terakami, S., Nishitani, C., Kimura, T., Sawamura, Y., Hirabashi, T., Hayashi T., 2006, Genome mapping in pear. In: 3rd International Rosaceae Genomics Conference, p. 55, War Memorial Conference Center, Napier New Zealand, 19–22 March 2006.

Zeng, D., Lin, D.Y., 2005, Estimating haplotype-disease associations with pooled genotype data. Genetic Epidemiology 28: 70–82.

Color Plates

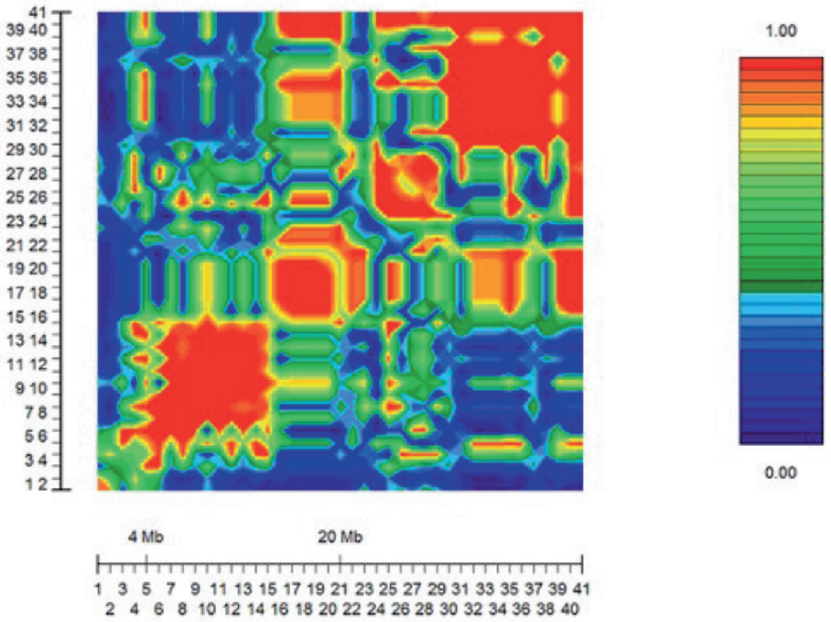


Figure 2.5. Pairwise $|D'|$ for 45 SNPs within a linked region (figure from GENESTAT, http://www.meb.ki.se/genestat/, courtesy of the Swedish National Biobanking program, Wallenberg consortium north).

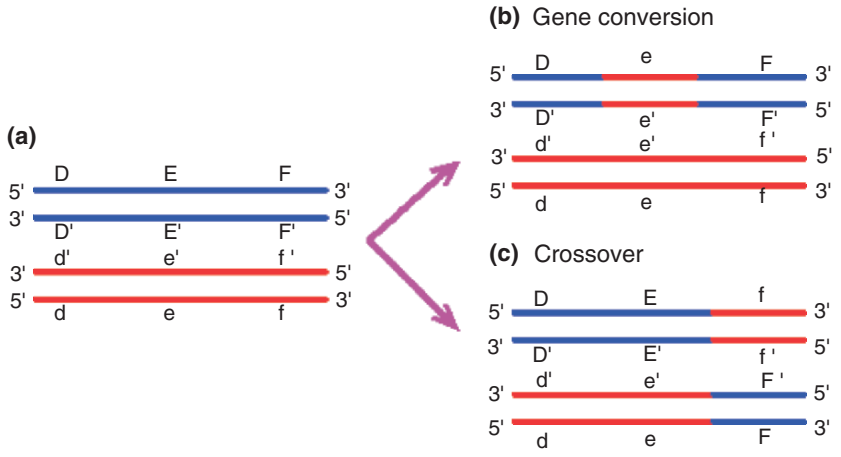


Figure 2.6. A simplistic diagram showing the major difference between gene conversion and crossover. (A) Two DNA molecules. (B) Gene conversion after mismatch correction – the red DNA donates part of its genetic information (e–e' region) to the blue DNA. (C) DNA crossover – the two DNAs exchange part of their genetic information (f–f' and F–F').

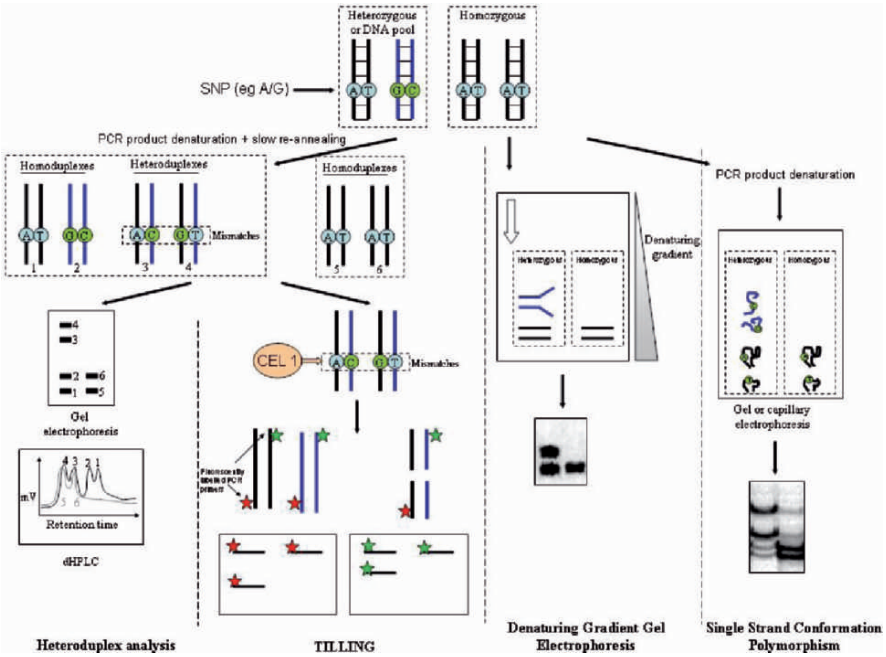


Figure 4.1. Nonsequencing SNP discovery methods: heteroduplex analysis, TILLING, DGGE, and SSCP.

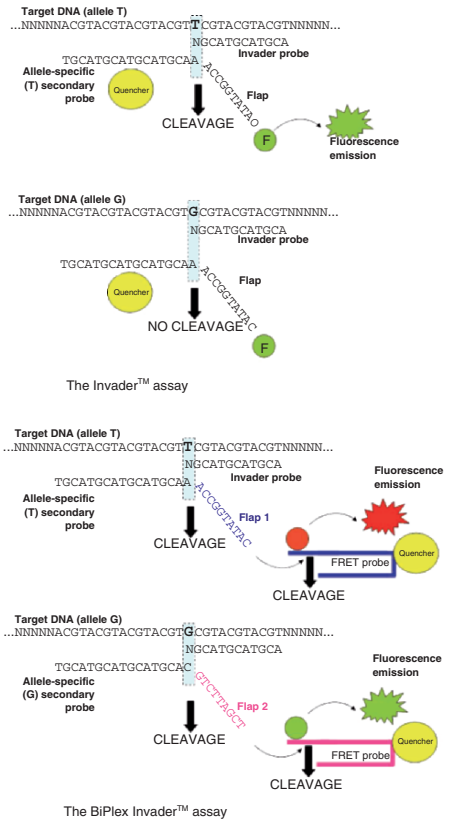


Figure 5.2. The Invader™ assay.

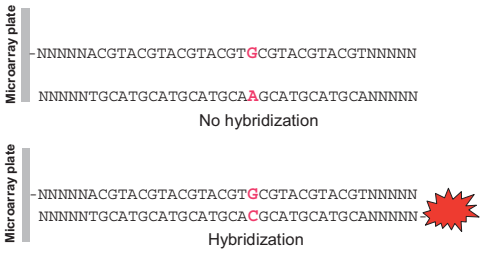


Figure 5.3. Allele-specific oligonucleotide hybridization. A oligonucleotides feature with the SNP site in its central position is bound to a microarray glass plate. Under stringent hybridization conditions, the complementary allele will anneal to the fixed oligonucleotide and a fluorescent signal attached to the probe will be detected.

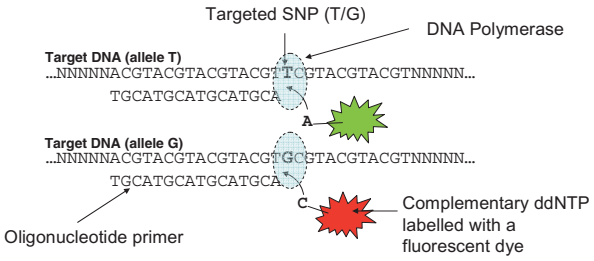
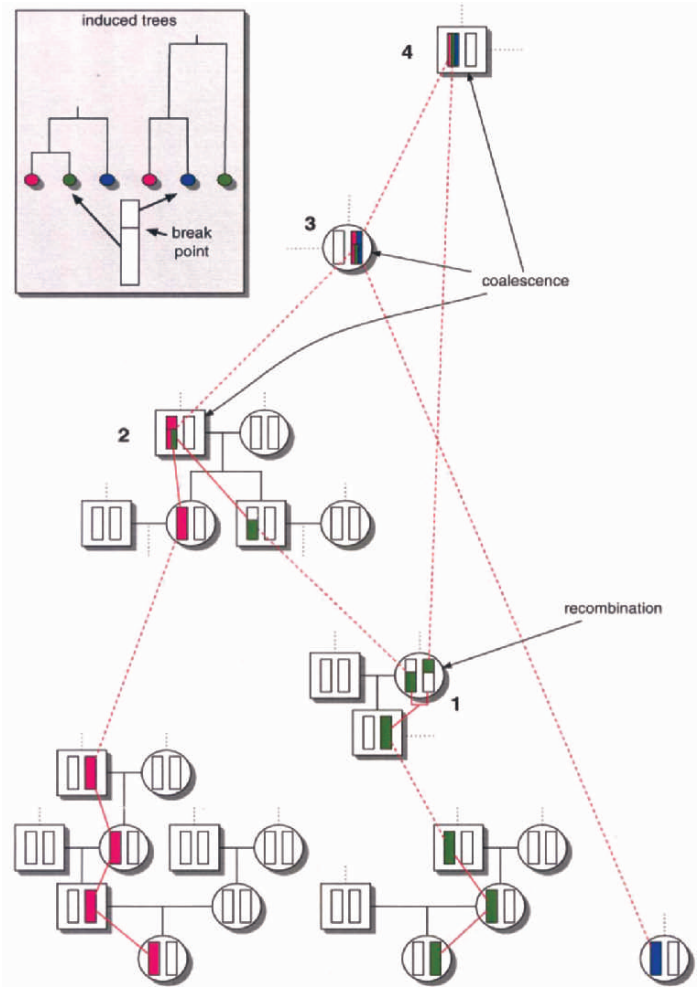


Figure 5.5. Minisequencing or primer extension. An oligonucleotide primer immediately flanking the SNP is extended using a DNA polymerase. Fluorescently labeled terminating nucleotides are incorporated, with a different dye color for every nucleotide. The oligonucleotide can be attached to a solid-phase array, separated in a capillary electrophoresis system, by a flow cytometry instrument, by mass spectrometry, or revealed by a fluorescent plate reader.



An example of a genealogy for three copies of a short chromosomal segment. Tracing the segmental lineages back in time, the following events occur: **1**, the "green" lineage undergoes recombination and splits into two lineages, which are then traced separately; **2**, one of the resulting green lineages coalesces with the "magenta" lineage, creating a segment, part of which is ancestral to both green and magenta, part of which is ancestral to magenta only; **3**, the "blue" lineage coalesces with the lineage created by event 2, creating a segment that is partially ancestral to blue and magenta, partially ancestral to all three colours; **4**, the "other" part of the green lineage coalesces with the lineage created by event 3, creating a segment that is ancestral to all three colours in its entirety. The recombination event induces different genealogical trees on either side of the break: these are shown in the inserted figure.

Figure 8.1. Example genealogy illustrating the coalescent (Nordborg and Tavaré 2002).

INDEX

G

H

I

L