Kirkhouse Trust

JOHN INNES CENTRE

MOLECULAR MARKER TECHNIQUES FOR CROP IMPROVEMENT Part II Genetic Mapping

COURSE MANUAL NOVEMBER 2005

UNIVERSITY OF AGRICULTURAL SCIENCES, BANGALORE, INDIA

CONTENTS

1. Introduction

1.1 Acknowledgements

The Kirkhouse Trust has been involved with the UAS Bangalore for three years. In this period, we have supported a number of activities in the university, but our primary project has centred on the training courses in marker assisted selection. The first of these, which was the first major activity of the Trust, took place in November 2003, a collaborative endeavour involving the John Innes Centre and the UAS. We are especially pleased that this course was successfully repeated by the UAS in 2004. These activities brought other benefits to the Trust. We have learned a great deal about working with colleagues in parts of the world where scientific research is still developing; about the problems they face and how we can best support them. The experience has shaped the way in which we are developing projects at other centres and in other countries. It has focussed our attention on the grain legumes, important crops, which are relatively neglected by most aid agencies.

The 2003 and 2004 courses focussed on characterising genome sequence diversity in Lablab cultivars. The 2005 course will explore how this can be used in mapping traits and incorporated into breeding programmes. As scientists, we must take a detached view as to the usefulness of this technology; there are circumstances in which traditional breeding methods are more economical. An important part of this course will be an objective evaluation of the potential benefits. This caution notwithstanding, we have committed support for a research project to develop more molecular markers for Lablab.

The 2005 course will take place in the laboratory refurbished for Trust supported activities. We are pleased to acknowledge the persistent efforts on the Trust's behalf of Professor T.K.S.Gowda and other members of the Department of Biotechnology in supervising this work and in other aspects of our activities. We are grateful for his generosity in making the facilities we provide available to other departments in the university. Many others have helped in the 2004 course and in preparing for the 2005 course; among these are Drs A. Mohan Rao, P.H. Ramanjini Gowda, P. Mahadevu, and Mr S.C. Venkatesh.

The principal organisers of the 2005 course are Dr Maggie Knox and Professor Noel Ellis, who developed and ran the successful 2003 course. A number of experts in the use of MAS in India have agreed to lecture on the course, including Drs Mohapatra from IARI, New Delhi, K.V. Bhat from NBPGR New Delhi, Girish Kumar, and C.T. Hash from ICRISAT. Others joining the teaching team from outside India are Drs Ian Mackay, D.J. Kim and Robert

Koebner; administrative support will be provided by the Department of Biotechnology and Drs Sonia Morgan and Janice Henderson. We are delighted that two international Lablab experts, Drs Brigitte Maass and Bruce Pengelly can also be present for part of the course.

Once more, Professor K. VijayRaghavan, the Director of the National Centre of Biological Sciences in Bangalore and his staff have been generous in making available the Centre's Computer Lab and other facilities for the course.

We are grateful to Dr T.K. Prabhakara Setty and Professor Sheelavantar the Research Director, and the Vice Chancellor of the University of Agricultural Science, Bangalore, for their generous support of the Trust's work in the University and to Professor Sharat Chandra for his unstinting help and advice.

Professor Sir Edwin Southern
Oxford, 13th October 2005

1.2 Challenges for plant breeding & the role of molecular genetics

Agriculture faces the problem of increasing demand from an expanding population, coupled to threats of reduced area for production as a consequence of climate change for example through water deficit, soil salinity or unpredictable weather at harvest. In essence within the next 50 years agricultural productivity will need to double. No doubt the major determinants of the success or failure of this endeavour will depend on factors remote from molecular genetics and in the end it will be farmers that produce food and inevitably most will have limited resources, what then is the role of resource rich molecular genetics?

While molecular genetics will not feed people, I think it can help plant breeders who in turn will supply seed that is the raw material for agriculture. A basic problem for the exploitation of molecular genetics in plant breeding is its cost. To obtain one data point from molecular marker analysis costs about 0.5$, and this does not include the costs associated with marker development, or determining the association between a marker allele and a relevant trait. Deploying markers with this cost in breeding programs presupposes a high commercial value of the crop, or a great willingness for public sector investment.

A major challenge is therefore to reduce the cost of molecular genetics in plant breeding. This may rely on technical developments, but a major means to reduce cost is to ensure that molecular marker methods are appropriate to the context in which they are used. In the first of this series of courses we focussed on molecular markers providing genotype information associated with germplasm because this is simply a means for attaching knowledge to breeding materials. Once this is done the information, especially for inbreeding crops, can be disseminated and used at very low additional cost.

The primary aim of the first course was to facilitate the use of molecular markers in the choice of breeding material notably the selection parental lines. The objective in this second course is to explore the exploitation of molecular markers in later aspects of breeding. Having selected parents that contrast for breeding traits that are also maximally distinct as judged by molecular markers we should be able to use this molecular diversity to obtain deeper knowledge of the genetics of traits that are the focus of breeding programmes, again we will focus on *Lablab purpureus* (*Dolichos lablab*), as a crop of specific interest in UAS Bangalore.

Professor Noel Ellis 19th July 2005

1.3 *Lablab purpureus* (*Dolichos*) – Characteristics & plant details

See: *Lablab Purpureus* (Dolichos) – Characteristics & plant details (p9) of Part I Course Manual.

1.4 Course Administration

1.5 Course programme
Course timetable:

The course aims to cover three component parts as outlined below:

(Part i) Data generation and collection - How do markers work? This includes SNPs, SSRs etc., and can include experimental work on Lablab populations.

(Part ii) Data analysis – all of the issues about segregation ratios and mapping, including QTL analysis. This could analyse the Lablab data.

(Part iii) Generation of markers and populations. This includes comparative genomics both for comparative marker positions and primer design. Here we also need to consider population structures and get some input from the breeders on the best way to include marker analysis in a breeding strategy. The issue of marker diversity in relation to populations also comes in here.

All of these components will be discussed with all course participants. However towards the end of the course part iii will be subdivided into two sections with the intention that participants could dig a little deeper into the subject as best meets their needs:

1) The means to incorporate marker analysis in breeding strategies.
2) Comparative genomics and the construction of comparative genetic maps.

NB all participants will cover these issues within the course as a whole, the intention is to tailor the further details to individual needs.

For the main part of the course the participants will be in two cohorts (A & B). This separation is purely for the purpose of timetabling and the efficient use of lab / computing room space. the separation into two groups for the specialisms a and b will probably require a re grouping, and participants may wish to change their mind on which avenue they choose during the progress f the course.

The proposed timetable is presented below. Note that the group as a whole will meet at UAS in the morning for debriefing and discussion of the day's planned events. Hopefully the two cohorts will be able to help each other by highlighting difficulties or specific areas hat are of particular interest to pursue.

Colour codes

Lecture / data analysis	NCBS split sessions also at UAS
Lab session	UAS
Group joins together	
Breather	

Sunday 6th Nov.	groups arrive
Monday 7th Nov.	Course planning : final details

		Morning	Afternoon
Tuesday	8	Use of KT funded facilities/applying for research grants/verbal presentations	
Tuesday	8	course set up	
Wednesday	9	Open seminars	Open seminars

Group A

day	date	start of day	morning	afternoon
Thursday	10	intro to first two days	1) part ii Introduction to genetic mapping	2) part ii Statistics in genetic mapping
Friday	11	round up and plans for today	3) part i checking DNA preps: FTA paper	4) part i checking DNA preps: FTA paper
Saturday	12	Late start	5) Lablab workshop (afternoon and evening)	
Sunday	13			
Monday	14	intro to second two days	6) part ii Construction of genetic maps	7) part ii examples and data sets
Tuesday	15	round up and plans for today	8) part i PCR markers	9) part i Microsatellite markers
Wednesday	16	intro to third two days	10) part iii Comparative genetic maps	11) part iii (2) Markers for comparative genetics
Thursday	17	round up and plans for today	12) part i (part iii) SNP / allele specific PCR	13) part i Marker analysis / Data collation
Friday	18	intro to fourth two days	14) part iii (1 or 2) Comp. genetics or markers & traits	15) part iii (1 or 2) Comp. genetics or markers & traits
Saturday	19	round up and plans for today	16) part iii (1) Workshop on MAS in breeding	
Sunday	20			
Monday	21	round up and plans for today	17) part ii Introduction to QTL	18) part iii (1) Statistics in MAS and Breeding
Tuesday	22	round up and plans for today	19a) part ii Interpreting map data	19b) part ii Interpreting map data
Wednesday	23		20) Preparing posters and scientific papers	

Group B

day	date	start of day	morning	afternoon
Thursday	10	intro to first two days	3) part i checking DNA preps: FTA paper	4) part i checking DNA preps: FTA paper
Friday	11	round up and plans for today	1) part ii Introduction to genetic mapping	2) part ii part iStatistics in genetic mapping
Saturday	12	Late start	5) Lablab workshop (afternoon and evening)	
Sunday	13			
Monday	14	intro to second two days	8) part i PCR markers	9) part i Microsatellite markers
Tuesday	15	round up and plans for today	6) part ii Construction of genetic maps	7) part ii examples and data sets
Wednesday	16	intro to third two days	12) part i (part iii) SNP / allele specific PCR	13) part i Marker analysis / Data collation
Thursday	17	round up and plans for today	10) part iii Comparative genetic maps	11) part iii (2) Markers for comparative genetics
Friday	18	intro to fourth two days	17) part ii Introduction to QTL	18) part iii (1) Statistics in MAS and Breeding
Saturday	19	round up and plans for today	16) part iii (a) Workshop on MAS in breeding	
Sunday	20			
Monday	21	round up and plans for today	14) part iii (1 or 2) Comp. genetics or markers & traits	15) part iii (1 or 2) Comp. genetics or markers & traits
Tuesday	22	round up and plans for today	19a) part ii Interpreting map data	19b) part ii Interpreting map data
Wednesday	23		20) Preparing posters and scientific papers	

item	title	person
1	Introduction to genetic mapping	Noel
2	Statistics in genetic mapping	Noel
3	checking DNA preps: FTA paper	Venkatesh \ Mohan
4	checking DNA preps: FTA paper	Venkatesh \ Mohan
5	5) Lablab workshop	Maggie Knox, Byre Gowda, Brigette Maas, Bruce Pengally,Ian Ian Mackay + nuritionist
6	Construction of genetic maps	Noel / Maggie
6	5) Lablab workshop	Maggie Knox, Byre Gowda, Brigette Maas, Bruce Pengally,Ian Ian Mackay + nuritionist
7	examples and data sets	Noel / Maggie
8	PCR markers	Venkatesh \ Mohan
9	Microsatellite markers	Venkatesh \ Mohan
10	Comparative genetic maps	Noel \ DJ Kim
11	Markers for comparative genetics	Noel \ DJ Kim
12	SNP / allele specific PCR	DJ / Noel / Maggie / Venkatesh / Mohan
13	Marker analysis / Data collation	DJ / Noel / Maggie / Venkatesh / Mohan
14	Marker design or QTL analysis	Noel / Ian
15	Marker design or QTL analysis	Noel / Ian
16	Workshop on MAS in breeding	Girish Kumar, Robert Koebner and Tom Hash (plus debate on MAS)
17	Introduction to QTL	Ian / Noel / Robert
17	Workshop on MAS in breeding	Girish Kumar, Robert Koebner and Tom Hash (plus debate on MAS)
18	Statistics in MAS and Breeding	Ian / Noel / Robert
19	Interpreting map data	Noel / Robert / Ian
20	Interpreting map data	Noel / Robert / Ian

1.6 Speakers and teachers

1.7 Course participants

2. An introduction to genetic mapping

Mendelian genetics is one of the great scientific theories: it has simplicity, predictive power and a wide scope. Mendelian genetics has three basic propositions: (1) Parents contribute equally to their offspring; (2) Characters have discrete determinants; and (3) These determinants behave independently both in the formation of gametes and in the association of gametes to form a zygote. Sometimes this is formulated as two 'laws'. The first is the law of segregation (essentially 1 and 2 above) and the second is the law of independent assortment (3).

For diploid organisms (peas, people, parrots …) there are at most two different types of determinant for any character and the usual Mendelian ratios follow. However there are also polyploid organisms (alfalfa, bananas, clover … and many crop species are polyploid) where there may be more than two different types of determinant for any character. In these cases segregation can differ from 'the usual Mendelian ratios' but follows the same basic laws.

Since the early 20th century a number of examples of 'non-Mendelian' genetics have been found; for example, of plastid genetics, bacterial genetics, transposable elements and especially pertinent to this course, genetic linkage and quantitative traits. These have not been taken to disprove Mendel's theory, but are seen as special cases where some additional information is needed to understand the way inheritance works. Mendel's basic ideas remain intact.

Genetic linkage was discovered very early in the history of genetics even before the chromosome theory was accepted. William Bateson (the first director of the John Innes Institute) knew about this but was, for most of his life, implacably opposed to the chromosomal theory of inheritance[1]. We now know that the independent segregation of types of determinants (alleles) is a consequence of the way that chromosomes behave at meiosis. There are many more genes (determinants) than chromosomes so this means that some alleles tend to follow each other into the same gametes (linkage in coupling) or tend to go into opposite gametes (linkage in repulsion). This means that genetic linkage is a modification of Mendel's ideas about independent segregation.

There was a long and somewhat bitter argument in the early days of genetics between 'biometricians' and 'Mendelian genetics', the source of this dispute is rather hard to understand now, but the nub of it is

[1] He could not accept that something identical in every cell could be responsible for the huge diversity of cell types.

clearly described by Lewontin[2]. This appears to have been a dispute about the discreteness of allelic states. Although some allelic differences seem clear (eg. round vs wrinkled seeds) others are less so (small vs large, or not-so-small seeds). The issue at stake was whether these different classes of traits (quantitative vs qualitative) are determined in the same way, by the same type of genes. Thomas Mather ascribed some of these properties to a distinct class of determinant that he called 'polygenes'. The current accepted view is that quantitative traits are determined by alleles of weak effect, or where the effect is comparable with environmentally determined variation. Often these weak alleles are distributed over many genes, so their segregation is difficult to follow.

A related issue was whether genes were indivisible units of inheritance, hence the name locus. If they were not (as was first shown in bacterial genetics) then could some discrete states be combinations of nearby differences in adjacent determinants or within individual extended genes? There are indeed some examples of haplotypes that determine individual character states. This complication could be regarded as modification to the Mendelian view of discrete determinants with different forms, but there are alternative ways of thinking about it.

From this discussion we can see that Mendelian genetics is a robust framework for understanding inheritance. However we now know a lot about the underlying molecular mechanism so we can choose whether to think in molecular terms or as Mendel did, in either case we must think carefully about the meanings of some of the entities under discussion and when that is done much that appears different between the two modes of thought becomes less so.

In this course we will address the two issues of genetic linkage and quantitative traits. The course will draw heavily on Dominique de Vienne's text book[3] (cited as de Vienne pp in the text below), and where appropriate individual papers will be cited. In the section below some words are highlighted and these are discussed in the glossary.

2.1 Linkage

Genetic mapping, like diversity analysis, depends on the scoring of marker data. Markers can be of a wide range of types, from flower colour to bands on gels. However, the main difference between genetic mapping and diversity analysis is that for genetic mapping, population structure is defined; while for diversity studies this is not known a

[2] Lewontin R.C. (1974) The Genetic Basis of Evolutionary Change, Columbia University Press [ISBN 0-231-08318-1]
[3] de Vienne D. (2003) Molecular Markers in Plant Genetics and Biotechnology. Science Publishers, Inc. Enfield (NH) USA, Plymouth UK.

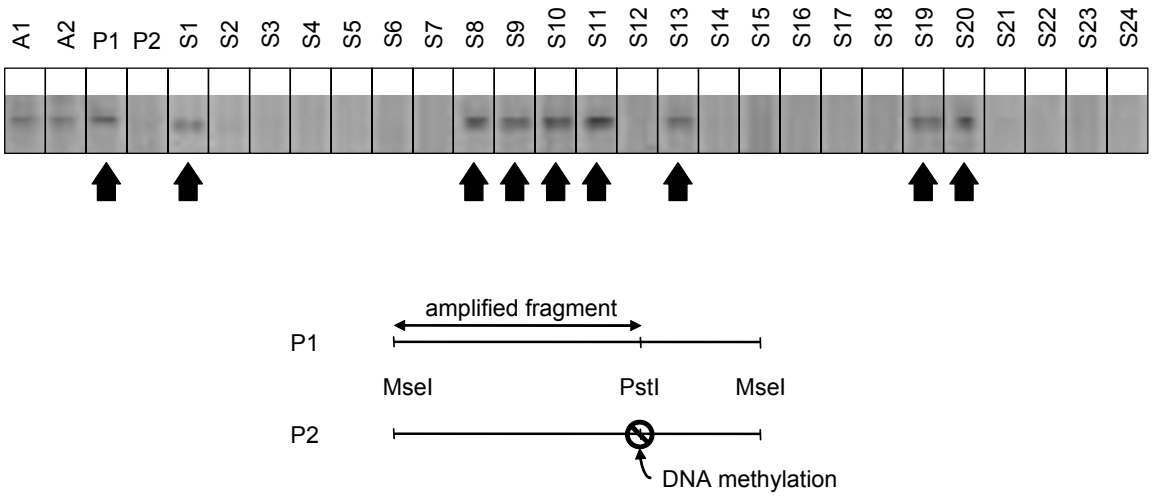
priori. When looking for associations in diversity analysis we ask which samples share allelic states, and so score data according to these alleles. In segregation analysis we want to know which parent the alleles are derived from, and we are searching for associations between alleles from the same parent in a segregating population. So for segregation analysis parent-of-origin is the key point rather than allele type (see below). The alleles from the two parents are expected to be found in non-parental combinations among the offspring in a segregating population, and these new patterns are recombinant. Those that tend to be found together in the same individuals (segregants) have experienced little recombination and are said to be associated or linked.

2.2 Parent-of-origin

It seems a simple matter to decide the parent of origin of a given marker, but this is not always the case. This section will consider some, thankfully rare cases where parent-of-origin may not be easy to determine. In classical genetics unstable mutant alleles may revert, so in a cross to wild type an allele may appear to be wild type but in fact be inherited from the mutant parent. An accurate assessment of the parent of origin of a marker or gene assumes that it has high expressivity and penetrance.

Figure 3.2.1 illustrates this in the context of a molecular marker. In this figure it appears that the arrowed segregants have inherited the allele revealed by the band arrowed in P1. However, the tracks labelled A1 and A2 are an alternative assay for this allelic difference and show a different result: both parents appear to carry the same DNA sequence. In this example the difference between P1 and P2 is not a DNA sequence difference; rather it is the methylation state of a restriction site. It may be that the bands arrowed in the segregants derive from the P1 parent, but it is also possible that they derived from P2 where the DNA methylation state has been changed. If the methylation state of the restriction site does not (always) follow the parent-of-origin then clearly scoring the banding pattern may not properly reflect the allelic state.

Figure 3.2.1



P1, P2 : Parents,
S1 to S24 : segregants (RILs) from the cross between P1 and P2.
A1, A2 : an alternative way to assay the P1 band.
From Knox and Ellis (2001)[4]

This example illustrates how a marker assay may not give a reliable indication of the parent of origin of the corresponding DNA sequence. In the case illustrated what is assayed is the segregation of the determinant of the methylation state of the relevant restriction site. This may segregate in a normal Mendelian fashion or as some interaction between the locus corresponding to the restriction site and the determinant of its methylation state. If the segregation is not simple the marker will be ignored from the point of view of genetics, but if segregation is normal (or nearly so) the marker will likely be included in a genetic analysis.

This raises two important issues. First the genetic locus that determines the segregation pattern may not correspond to the location of the DNA sequence on which the assay is based. Secondly, if these two do correspond the assay may not be an entirely reliable indicator of the parent of origin.

2.3 Errors

The discussion above describes one type of error associated with DNA based genetic markers. There are many other potential types of error, and some of these are shared with other marker types. Errors may be trivial in their origin (mis-typed scores for example). Some errors can be eliminated by independent scoring of the data and re-checking the scores. In principle the fewer operations between an assay and its being recorded the fewer errors are likely to occur. The minimisation

[4] Knox M.R. and Ellis T.H.N. (2001) Stability and Inheritance of Methylation States at *Pst*I Sites in *Pisum*. Mol. Gen. Genet. 265: 497-507

of sources of error is important, but even where attempts to minimise errors have been exhaustive we cannot be sure that data is error free. Attempts to assess errors are an important part of any quantitative exercise. Determining marker order can be considered a way of dealing with data errors.

2.4 Marker order

Genetic mapping is simply a way of describing how allelic differences are distributed among offspring. However simple this is in concept, there are many practical difficulties. First of all, candidates for linked markers depend on statistical inference. Among the statistical approaches used to search for marker segregation patterns that are unexpectedly similar, are LOD scores and Chi square tests. For both these approaches it is necessary to be aware of their weakness. For example, in a Chi square analysis it is often considered that a value that would be expected by chance alone in 5% (or 1% etc) or fewer of pair-wise tests is a good indication of an association that is not due to chance alone.

For N markers there are $N(N\text{-}1)$ possible pairs, so for 10 markers there are 90 possible pairs, so at 95% confidence between 4 and 5 associations are expected by chance alone. Genetic mapping needs to sort out the difference between linked markers and those associated by chance alone. There are several approaches to overcoming this problem. The first is simply to set the threshold value rather high, so that chance associations are rare given the size of the data set. The second approach is to rely on expected properties of linked markers. Genetic linkage is a consequence of the linear arrangement of DNA sequences with respect to the formation of crossover events. So markers on genetic maps should obey linear rules of arrangement. This constraint does not apply to chance associations. Thus the construction of linear genetic maps is a test of the reliability of these associations or linkage. Deviations from linearity indicate problems with the underlying data.

To find the 'best' order of markers from a given data set is, in principle, simple to determine. All we need to do is look through all possible orders and find the one that proposes the smallest number of recombinants and also proposes the smallest number of double recombinants. This is the simplest hypothesis. However, there is a complication: given N markers there are $\frac{1}{2}N$! orders to look through, so for a modest number of markers the amount of work to do is prohibitively large (for 10 markers there are about 1.8 million orders, and for 20 about 10^{18}). This means that some system is needed to simplify the problem. Several mapping programs compute all possible

maps for markers taken three at a time: three-point mapping. This simply serves to eliminate a very large number of possible orders from the search for the simplest map. Other factors need to be taken into account, such as segregation ratio: adjacent markers should have similar segregation ratios, but distorted segregation may make markers appear to be linked when in fact they are not. Thus it is always important to consider the monogenic segregation ratio for a given marker. These brief comments are intended to serve as a background to genetic mapping, and experience with relevant mapping software[5] will be provided during the course.

2.5 Maps, and mapping functions

If we consider three markers A, B and C, if they are associated by chance alone then the recombination fraction between any pair is undefined, and there is no expected relationship between the recombination fractions between the three possible pairs. However, if these markers have a linear arrangement and the order is A - B - C then there are two intervals AB and BC between the markers. The compound interval AC between the external markers A and C has some relationship to the AB and AC intervals. In a genetic map we expect the recombination events to obey some rule such that the recombination between A and C will be more than between A and B or B and C. This simple relationship does not apply to markers associated by chance alone, where all these values could be the same. However, except for very small values of the recombination fraction, the recombination fraction between A and C is not expected to be the sum of the recombination fractions in the AB and BC intervals. We know that recombination is a fraction and that the maximum value is about 50 % (free association or no linkage), furthermore for ten successive intervals each with 10 % recombination we do not expect 100 % recombination between the extreme markers. So the recombination fractions themselves can not be additive.

Where the recombination fraction between markers A and B is written as r_{AB} (etc) we can say: $r_{AC} \neq r_{AB} + r_{BC}$. But we want to define a function, map distance (d), that is additive and is a function of r such that $d_{AC} = d_{AB} + d_{BC}$ Given the order $A - B - C$ where the fraction of recombinants in the interval $A - B$ (r_{AB}) is m and the fraction of recombinants in the interval $B - C$ (r_{BC}) is n then the fraction of non-recombinants in AC is

[5] Lander E.S. and Green P. (1987) Construction of multilocus linkage maps in human PNAS USA 84: 2363-2367

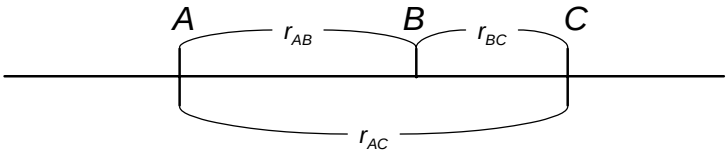
Lander E.S., Green P., Abrahamson J., Barlow A., Daly M.J., Lincoln S.E. and Newburg L. (1987) MAPMAKER: an interactive computer package for constructing genetic linkage maps of experimental and natural populations. Genomics 1: 174-181

Stam P. (1993) Construction of integrated genetic linkage maps by means of a new computer package: JoinMap. The Plant Journal 3: 739-744

(1-*m*)(1-*n*). The fraction of single recombinants in *AB* is *m*(1-*n*) and the fraction of single recombinants in *BC* is *n*(1-*m*). So the fraction of single recombinants in the combined interval *AC* is (1-*m*)*n* + *m*(1-*n*) = *m* + *n* -2*mn.*

We need to define *d* as a function of *r* such that $dAC = r_{AB} + r_{BC} - 2r_{AB}r_{BC}$. Haldane's function [$d = -½\ln(1-2r)$] satisfies this requirement as can be seen in figure 3.5.1.

Figure 3.5.1



fraction of recombinants in the interval AB is *m*
fraction of recombinants in the interval BC is *n*

Fraction of single recombinants is AB *m(1-n)*
Fraction of single recombinants is BC *n(1-m)*

Fraction of AC single recombinants *(1-m)n + m(1-n)*
= m +n -2mn

Addition rules:
$r_{AC} \neq r_{AB} + r_{BC}$

but
$d_{AC} = d_{AB} + d_{BC}$

Haldane's function*:*

$d = -½\ln(1-2r)$
$r = ½(1 - e^{-2d})$
$r_{AC} = ½(1 - e^{-2dAC})$
$= ½(1 - e^{-2dAB}\, e^{-2dBC})$
$= ½(1 - [(1-2r_{AB})(1-2r_{BC})])$
$r_{AC} = r_{AB} + r_{BC} - 2r_{AB}r_{BC}$
qv. *m +n - 2mn*

Note that Haldane's function fits the assumption that adjacent recombination events are independent. This is not always the case, and other map functions take account of interference between adjacent recombination events. (see de Vienne p 50 – 53 for the derivation of the Haldane mapping function and a discussion of interference.)

2.6 Linkage groups and chromosomes

The terms 'linkage group' and 'chromosome' are often used interchangeably. There is some justification for this, but they are really two different things.
A chromosome is literally a coloured thing, it is what you see when you stain some cells in an appropriate way and then look down a microscope to see the rod shaped structures that do marvellous things at mitosis and meiosis. The term is, logically, extended to the thing

that is stained in this way. My cells have chromosomes in them even though they are not stained.
A linkage group is a part of a data set. It corresponds to markers that are associated with each other according to a set of rules. It is not necessarily the case that all markers in a linkage group are observably 'linked', often they are in 'free segregation'. Markers at one end of a linkage group may well have no more association with markers at the other end than with markers on other linkage groups. However, you can always trace a series of connections between markers on a linkage group. On a linkage group with ten markers, where there is 10 % recombination between each adjacent marker, there is a chain of connection, and the extreme markers do not have 100 % recombination, but - as a property of the way probabilities are added - something approaching 50 %.
The markers on a linkage group are in some way derived from a chromosome (in the extended sense). Either they correspond to DNA sequences that are part of the chromosome or some derived property of that DNA sequence – a protein with a certain interesting mobility in a gel system, a flower colour ... The way these markers behave genetically: how they are distributed among offspring is a direct consequence of the way the chromosome behaves in meiosis and zygote formation. The properties of a linkage group are determined by chromosome behaviour and chromosome behaviour can be inferred from linkage group behaviour.
Of course all this had to be proved, and at the outset some serious geneticists thought the idea was nonsense – but that's another story.

2.7 Molecular markers and crop improvement

The purpose of this course is to explore, in some detail, the ways in which genetic markers can be applied to crop improvement. The first part of the course dealt with the issue of assessment of genetic diversity, and stressed the utility of molecular marker diversity in the choice of material for trait analysis. Part II of the course deals with the way molecular markers can be exploited in trait analysis both for understanding the inheritance of trait determinants and their manipulation in breeding. The value of molecular markers is clear for the understanding of trait genetics, but their deployment in a given breeding context is a matter of informed judgement. We hope that the present course will provide a firm foundation for this decision making.

In the source text, de Vienne states:

> *"In conclusion about marker assisted selection, it appears that markers can save time by selection in off-season generations, without agronomic evaluation, and above all they are irreplaceable for the management of recombinations, in order to accumulate favourable alleles as quickly as possible in a single genotype. To exploit their value fully, new schemes of*

recurrent selection or rather recurrent genotype construction must be devised".

Breeding is both a skill and an art, drawing on the breeder's knowledge, imagination and aspirations. Molecular markers are a tool the challenge is to deploy them with skill and good design.

2.8 Glossary

Expressivity: The intensity to which an organism expresses a phenotype

Haplotype: A set of linked alleles. Often this will refer to DNA sequence variation within a gene where one sequence variant may occur in the presence or absence of another. For two nucleotide positions this will generate four possible haplotypes:

1 **...A...C...**
2 **...G...C...**
3 **...G...T...**
4 **...A...T...**

A haplotype may arise by mutation. For example if the original allele is 1 above, and the mutation A > G occurs then the variant 2 will be generated. As this differs from the other alleles at one position it is not usually called a haplotype. However if a subsequent mutation occurs (C > T) then this must occur either in allele 1 or in allele 2, not both. This means we would expect to fine either haplotype 3 or haplotype 4. It is this coincidence of patterned allelic variation that makes haplotypes interesting. If the set of haplotypes 1, 2 and 3 arose by mutation haplotype 4 would arise either by independent mutation or (either G > A reversion from 3, or C > T mutation from 1) it could arise by recombination between alleles 1 and 2. If haplotypes represent variation at multiple sites then independent mutation at several is less likely than recombination. Thus haplotype structure in populations tells us a lot about which alleles have been present in the same individuals. That is it tells us about population substructure.

Interference: In the context of genetic mapping this refers to the positioning of crossovers with respect to each other. When chiasmata form, there are precursor structures that seem to combine or be coordinate to form a mature chaisma. This means that independent chiasmata tend not to occur immediately adjacent to each other. Chiasmata interfere with each other and with telomeres, but interference seems to ignore the centromere.
This means that there is a non-random distribution of recombination with respect to the chromosome.
Interference can (theoretically) also apply to chromatids. In the absence of chromatid interference when one chiasma forms between two chromatids of a bivalent, the chromatids involved in a second crossover are independent of the first selection

Orthology: Literally 'the same word'. Genes in two different species are said to be orthologous if they are descended from the same single gene in their most recent common ancestor. This is in fact difficult to

know for certain because the common ancestor may have had two different but very similar genes, and these different genes may be different progenitors of the two single genes present in the species being compared (in both of which the other second gene has been lost). A pair of genes descended from the same single gene, but present in the same genome (for example after a gene duplication event, or the formation of a polyploidy) are said to be paralogues. Paralogues and orthologous, are special cases of 'genes descended from a common ancestor' more generally known as homologous genes.

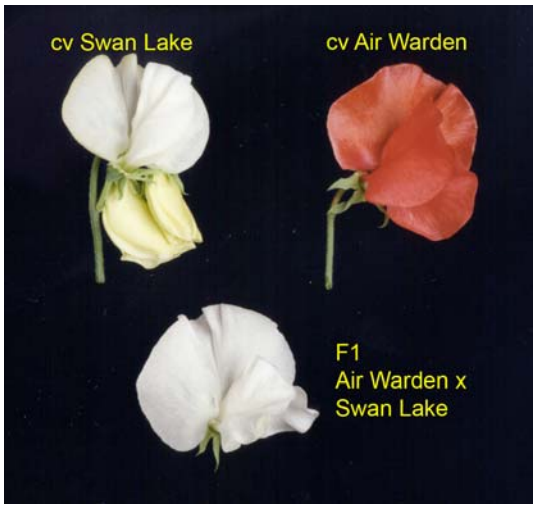
Penetrance: The proportion of organisms that show the effect of an allele.

Synteny: 'Holding together' – Synteny is the property of syntenous genes or genes in a syntenic region. This term has slightly different usage in animal vs plant genetics. In the former genes are considered syntenic if they are found in an equivalent segment of the genome of two species. In plant genetics the additional requirement is placed that syntenic genes should have the same relative order. In animal systems these are said to be collinear (with the confusing spelling). Plant geneticists tend to use collinear and syntenic as synonyms.

3. Genetic maps: the practicalities

3.1 Markers

Any segregating character can act as a genetic marker. Morphological characters, in the loose sense eg. flower colour, are easy to score and provide good genetic information, but often these characters exhibit dominance. Dominance is a genetic property, and is independent from the nature of the trait. For example, a dwarf trait such as that determined by the wheat *Rht* genes can be dominant[6]. In the example shown on the right, the F_1 of the cross between two inbred sweet pea (*Lathyrus odoratus*) lines shows that, in this case, the white flowered trait is dominant. In this case the F_2 segregated 3:1, white : coloured, and not all coloured flowers were red.



Not all morphological characters have clear dominance; in some cases the heterozygotes can be identified. This means that in an F_2 population more precise genotypic information is available. Marker types that allow the easy identification of heterozygotes include both isozyme markers and some types of DNA based markers.

In de Vienne's book many sources of molecular markers for genetic analysis are described in Chapter 1. The important point to note about DNA based markers is that there are potentially very many of these, and some differences between individuals in DNA sequence have no obvious consequence. Redundancy in the genetic code means that different sequences can encode the same polypeptide. Such differences may be completely silent, and the strength of selection is measured by the ratio of synonymous to non-synonymous differences. We know that most of the genome in eukaryotic organisms does not encode protein. DNA sequences in introns and in intergenic DNA are not necessarily subject to such intense selection as protein coding DNA. For this reason markers derived from such sequences are likely to be good sources of polymorphic markers.

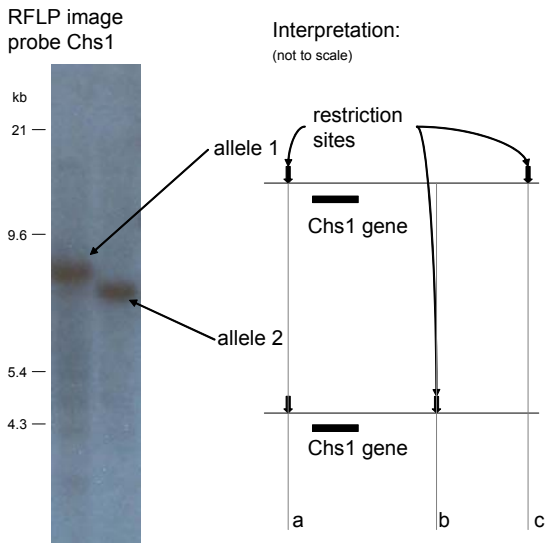
---

[6] J. Peng, D.E. Richards, N.M. Hartley, G.P. Murphy, K.M. Devos, J.E. Flintham, J. Beales, L.J. Fish, A.J. Worland, F. Pelica, D. Sudhakar, P. Christou, J.W. Snape, M.D. Gale and N.P. Harberd (1999). 'Green Revolution' genes encode mutant gibberellin response modulators. Nature 400: 256-261.

3.2 Dominant and co-dominant molecular markers

For molecular markers the idea of dominance may seem a little odd, but this simply reflects the way the different alleles are detected. If only one allele can be detected, and the other is inferred the marker is said to be dominant. If a marker is identified by the presence of a given band and its absence is the other, then if we see the band we know there is at least one 'band present' allele, but we don't know whether there is one or two of these. If the band is absent then we know there 'no band present' alleles, so we infer that there are two 'band absent' alleles. This seems reasonable, but it is an inference. Dominant molecular markers are thus much like dominant classical markers, when one allele is present the nature of the other is masked.

Co-dominant molecular markers are perhaps a little harder to understand because they push the idea of a locus or allele a little.



The picture to the right is a simple example of an RFLP. The hybridization probe (in this case a chalcone synthase gene) hybridises to different sized *EcoR*I fragments in two pea lines and these appear to be alleles. A heterozygote shows both bands and it is clear which band is which. In an F_2 the two parental bands and the heterozygote patterns segregate 1:1:2.
It is easy to score this RFLP as a co-dominant marker, but what is the allelic difference?
In the interpretation shown in the diagram, the DNA sequences at both alleles are almost identical. The exceptions are the restriction sites, and in principle could be single base differences. In allele 1 the restriction site at 'b' is missing, but we have no direct information that can tell us whether the restriction site at position 'c' is present in allele 2. In effect the assay is telling us whether the site at 'b' is present or not, and we can identify both possibilities. The allelism appears to be a single base difference identified by the restriction site present or absent from position 'b'.
What would be the consequence of an (extremely unlikely) crossover between the positions 'b' and 'c'. For one possibility (starting at the left of the sequence representing allele 1 and ending at the right of the sequence representing allele 2), we have no clear expectation for the

resulting band size. What appears to be a simple assay for a co-dominant marker is a little more complex than seems at first. Fortunately this type of crossover is exceptionally rare, so for all practical purposes these two dominant markers, tightly linked in repulsion phase (see below), behave as a single co-dominant marker.

3.3 Markers for comparative genetics

In the development of genetic markers that can be used to relate genetic maps of different species it is good to have a clear identifier for a genetic marker. It is useful if this is derived from a gene because it is often possible to identify the same gene in different species. In the case of *Rht* discussed in Peng et al (1999) the same gene and mutant phenotype has been identified in monocot and dicot species.

If we want to target genetic markers to genes, so that they can be identified in different species, or even wild relatives, then we have a problem. We have chosen a DNA sequence that is constrained in its evolution, and variants will be hard to find. We can overcome this to some degree by exploiting the structure of eukaryotic genes. The protein coding parts of genes (exons) are often interrupted by DNA sequences that correspond to parts of the mRNA that is removed (introns). These may have functions that are subject to purifying selection, but these often do not seem to depend precisely on their sequence and rather relate to approximate length or base composition. This means that introns can be good sources of markers because their sequence is variable and they are bounded by sequences that are conserved. This means that intron directed PCR is a good source of cross species genetic markers.

See intron directed PCR (see 4.5.4).

3.4 Scoring data and the codes used by mapping programmes

For the analysis of joint segregation of marker we need to know two things. The first is the parent-of-origin of the marker, and this has been discussed at length above. The second is the phase of the marker association. Alleles are said to be in 'coupling phase' when they are carried on the same homologue, but in 'repulsion phase' when they are carried on different homologues. This information is clear in the F_1 because the phase is determined by the parents. In the F_2 however this information is lost. One way round this problem is to use population types that avoid the need for this type of information, recombinant inbred or doubled haploid populations for example. In these populations the segregants are homozygous, so the worries about dominance are also avoided.

For dealing with these uncertainties the mapping programmes use codes for scoring and these need to be adopted. There are many possible ways of doing this but a common scheme, and the one used by Mapmaker is as follows:
You need to decide which parent is **a**, and which is **b**. Conventionally the first parent in the cross title is the female so the female parent could always be the **a** parent. However this is an important decision when scoring and analysing data from comparative mapping where there may be a parent common to many mapping populations.
For the F_2 data presented later in the manual (5.1.1) the parental line JI15 is **a**, and JI399 is **b**.

The codes for codominant markers are as follows:

A Homozygote for the allele from the **a** parent, ie. aa
B Homozygote for the allele from the **b** parent, ie. bb
H Heterozygote, ie. carrying both alleles from **a** and **b,** ie. ab
\- Missing data for an individual at the locus

There are additional codes for dominant markers, to allow for the heterozygotes being indistinguishable from either one of the homozygote class, they are as follows:

C Not a homozygote for allele parent **a**, ie. bb or ab
D Not a homozygote for allele parent **b**, ie. aa or ab

Worked example for round vs wrinkled seeds of pea.	
Round seeds carry the dominant *R* allele Wrinkled seeds are homozygous for the *r* allele In the cross between two inbred lines JI15 (round seeded *RR*) and JI1194 (wrinkled seeded *rr*) we can designate JI15 as parent **a** and JI1194 as parent **b**	
Code	Genotype
A	*RR*
B	*rr*
C	*Rr* or *rr*
D	*RR* or *Rr*
H	*Rr*
-	not known

This may seem complicated but is straightforward. The following examples using the F_2 population for pea where JI15 is parent **a**, and JI399 is parent **b**.
For a dominant marker from the **a** parent (JI15), the codes required are B and D, where B is the absence of the JI15 marker and D is the JI15 homozygote or the heterozygote, as for *T56p marker below:

```
*T56p
DDDDDDDDDBDBDDDDDDDDDBDBBDBDDBDBBDBBDDDDDBDDDDDBDDDDBDBDDDDDDDDBDBDDDDDBDBD
DDDBDBDDBDDBDBDBDDDDDDDDDDDDDBDDDDDDDDDDD
```

So we expect a 3:1 ratio of D:B (aa+ab:bb).

For a dominant marker from the **b** parent (JI399), the codes required are A and C, where A is the absence of the JI399 marker and C is the JI399 homozygote or the heterozygote, as for *S10m marker below:

```
*S10m
CACACCCCACCCCCCCCCCCCACCCACCCCACCCCCCCCCCACCAACCCCCACACACCAAACACAACCCCACCCCCC
-CCCCAAACACCCACCCC-CCCCCCCCCCCCCACCCCCCCCC
```

Again we expect a 3:1 ratio of C:A (bb+ab:aa).

For a codominant marker the codes are A, B and H, as for *Catheps (Cathepsin) below:

```
*Catheps    HBHABBHAAHAAHHAAAHAHHHAHAHHHBHHBHHAABAHAAAHAHA-
HHHHAHHHHABHABABHHAHBHABHABHHAAHAHBAAHHAHABAAHBAHH-HBHABHHHHHAABABAABBH
```

3.5 Coupling and repulsion phase - see de Vienne p53

Getting back to coupling and repulsion with dominant markers.

Coupling phase:

A series of linked markers that have the A and C mapping codes only are said to be **linked in coupling**; similarly for a series with the B and D codes only.

Repulsion phase:

A pair of linked markers where one has A and C and the other has the B and D codes are said to be **linked in repulsion**.

4. Practical Course: Molecular marker techniques for crop improvement

4.1 Safety information

SAFETY NOTE: wear gloves, labcoat and safety glasses throughout all of these procedures.

All of the liquid chemicals listed below are highly toxic and hazardous compounds. Take extreme care when handling these compounds and always wear a labcoat, gloves and safety glasses. On contact with skin/eyes wash immediately with water.

Acrylamide:Bisacrylamide (19:1) 40 % solution is bought ready made.
It is a neurotoxin and possible carcinogen; causes a tingling sensation on contact with skin.

Chloroform a carcinogenic and toxic solvent, readily absorbed through the skin, degreases and deproteinates the skin; dispense in the fumehood.

CTAB is harmful if swallowed, with a risk of serious damage to eyes: wear gloves and safety glasses.

Ethanol highly flammable

Ethidium bromide may cause heritable genetic damage.

Formaldehyde highly volatile carcinogenic and toxic, dispense stock in the fumehood.

Formamide a toxic and teratogen, dispense large volumes in the fume hood, small volumes in an aerated laboratory area.

Liquid nitrogen risk of cryogenic burns, always handle in a well aerated and open area.

Phenol a carcinogenic and toxic solvent, readily absorbed through the skin, causes severe and immediate burns, best treated initially with 10%PEG solution. Always dispense in fumehood.

Repelcote VS highly flammable, an irritant, use in the fumehood

Silver nitrate a possible carcinogen, causes burns, absorbed through the skin and stains.

Sodium thiosulphate irritant dispense stock in the fumehood.

TEMED Highly flammable, harmful by inhalation; causes burns.

All other chemicals in powder or crystal form that you are likely to encounter on this course will generally be irritants in one way or another, possibly causing irreversible effects on contact with the skin or on inhalation, always wear gloves when weighing, dispensing and handling solutions.

4.2 Buffers and solutions

DNA Extraction Buffer:
a)
3 X SSC
50mM EDTA pH 8
Store at RT
b)
500 mM NaCl
100 mM Tris pH 8.0
50 mM EDTA (same ish pH)
10 mM β Mercapto-ethanol (stock is 14M)
Store at RT

5 x RL Buffer
50 mM Tris-H acetate pH 7.5
50 mM Mg acetate
250 mM K acetate
25 mM DTT
Make from filter sterilized solutions and give a final filter sterilizations.
Store at -20°C

10 x PCR
500 mM KCl
100 mM Tris-HCl pH8.5
15 mM $MgCl_2$
1 mg/ml Gelatin
Add all components together; dispense into 20mls volumes and heat sterilise. Store at -20°C

10 x TBE
121 g Tris Base
51.3 g Boric Acid
3.7 g EDTA
Distilled Water to 1 litre
Dissolve components, sterilise by heating.
Store at room temperature (RT) 21

Acrylamide Mix 4.5%
420 g Urea
100 ml 10 x TBE
115 ml Acrylamide mix 40% (19:1)
Distilled Water to 1 litre

Dissolve the Urea in 400 ml of water by heating gently (1 minute/medium in a microwave oven, **do not over heat, this breaks down the Urea**); remove from microwave and stir till the crystals completely dissolve; once dissolved and while stirring add 100 ml of

10 x TBE and then add the 115 ml of acrylamide; make up to 1 litre with water. Store at 4ºC as this must be used **directly from chilled just before pouring a gel**.
NB make less, e.g. 250 ml, if you plan to pour only 2 to 3 gels.
Each gel requires 60 ml.

T 0.1E pH 8.0
10 mM Tris pH8
0.1 mM EDTA pH8
Make from heat sterilised components.
Store at room RT

20 X SSC
175.32 g NaCl
88.23 g Trisodium citrate
Distilled water to 1 litre
Store at RT

Acrylamide gel loading dye
98 % Formamide
0.025 % Bromophenol blue
0.025 % Xylene cyanol
10 mM EDTA pH8
Store at RT

Bind Silane A-174
Electran Stock solution:
40 ml of 100 % ethanol + 150 µl of Bind Silane.
Working solution:
40 ml of stock Bind Silane + 1 ml of 10 % acetic acid

Store all solutions, including stock, at 4ºC

Silver staining solutions

1. Developer: Dissolve 60 g sodium carbonate (make sure it is anhydrous and not too old) in 2 litre of distilled water and place at 4º C, for at least 4 hours before required.
2. Fixer: 10 % acetic acid (200 ml glacial acetic acid added to 1.8 litre distilled water).
3. Silver stain solution: 12 ml 1.01 N silver nitrate solution in 2 litre of distilled water; add 3 ml formaldehyde (40 % solution) and mix.
4. Developing solution: Immediately prior to developing the gel, add 300 µl of sodium thiosulphate solution (0.1001 N) and 3 ml of formaldehyde (40 % solution) to the pre-chilled sodium carbonate solution.

4.3.1 Standard units, prefixes and usage

In 1960, the 11th CGPM adopted a first series of prefixes and symbols of prefixes to form the names and symbols of decimal multiples and submultiples of SI units. Over the years, the list has been extended as summarized in the following table.

factor	prefix	symbol	factor	prefix	symbol
10^{24}	yotta-	Y	10^{-1}	deci-	d
10^{21}	zetta-	Z	10^{-2}	centi-	c
10^{18}	exa-	E	10^{-3}	**milli-**	**m**
10^{15}	peta-	P	10^{-6}	**micro-**	**µ**
10^{12}	tera-	T	10^{-9}	**nano-**	**n**
10^{9}	giga-	G	10^{-12}	**pico-**	**p**
10^{6}	mega-	M	10^{-15}	femto-	f
10^{3}	kilo-	k	10^{-18}	atto-	a
10^{2}	hecto-	h	10^{-21}	zepto-	z
10^{1}	deca-	da	10^{-24}	yocto-	y

Volumes, concentrations etc

The mole is the SI unit for the amount of a substance and one of the seven fundamental SI units. It is defined as the amount of substance of a system that contains as many elementary entities as there are atoms in 0.012 kilograms of carbon-12 (BIPM 1998, p. 97). It is abbreviated "mol," and the number of entities in a mole of substance is given by Avogadro's number (6.023×10^{23}).

Thus, mmol = 10^{-3} moles, µmol = 10^{-6} moles, etc

Note that some authors use upper case ‘m’, thus, pMol = 10^{-12} moles. This may lead to confusion as it is conventional to use uppercase ‘M’ for molarity (i.e. concentration in moles per litre).

Thus, 0.1 M = 0.1 moles per litre = 0.1 mol/l = 100 mM = 100 mmol/l etc.

Volume units are fractions of a litre. The more correct symbol for litre is a capital ‘L’, but lower case ‘l’ is more common. Thus ml = mL = milliliter = 10^{-3} litre

Note that a space should be introduced between the number and the symbol for units. It is not necessary to place a stop after the symbol and it is not necessary to pluralise. Thus, 10ml, 10 ml. and 10 mls are incorrect; 10 ml is correct.

4.3.2 Properties of oligonucleotides (primers)

You may need to calculate the molecular weight, melting temperature or some other property of an oligonucleotide which depends on its base sequence. Programs are available to help you do this. For example:

http://www.basic.northwestern.edu/biotools/oligocalc.html

Alternatively the companies from whom you purchase primers calculate the melting temperature and secondary structure when you input a sequence during on-line ordering. This gives you the opportunity to tailor the sequence to suit. On receiving the primers this same information is included in a useful table that also includes sequence information such as dilution suggestions.

4.4 Laboratory work

4.4.1. DNA preparation

4.4.1.1 Harvesting leaves for DNA preparations

On this course we will be working with leaves of *Lablab*, the method suggested below for this genus are more or less applicable to most crops. There are many other variations on this method, commercial kits are available and methods using CTAB (Appendix 1) can also be tried. However the method described here is straightforward and uses relatively inexpensive, easily available chemicals.
Leaves should be healthy, dry, without disease and preferably young.

Frozen material:
Large scale:
Label a plastic tube with the accession number, (Greiner 50 ml tubes will be used on the course) select a few leaves 3 to 4 (10-20 g of fresh tissue), place into the tube and immediately drop the tube into a container of liquid nitrogen.
These frozen samples can be treated on the same day, OR removed from the liquid nitrogen and stored at -20ºC.

Small scale:
As for large scale but choose one small leaf and place into 1.5ml plastic microtubes and scale down all components to suit the 1.5ml maximum volume of the tube.

Dried material:
As there is more degradation of DNA in drying leaves it is necessary to select two times the number of leaves compared to the frozen method.

Select leaves and place into brown paper bags, place the leaves as flat as possible and try not to have them overlapping. Do not seal the bags. The size of the paper bag is determined by the number of leaves. Keep in a dry place, eg. in a larger paper bag with silica gel at the bottom. Change the silica gel at regular intervals. Ideally this material should be prepared 3/4 weeks before DNA preparations are required.

4.4.1.2 Grinding leaves for DNA preparations
Suspension of plant tissue in buffer
Frozen material: requires a supply of liquid nitrogen and should be treated as follows:

1. Remove tubes of frozen leaves from the freezer and place into the liquid nitrogen container.

2. Handling one sample at a time, empty the contents of a tube into a mortar along with a small volume of liquid nitrogen. Grind the leaves to a fine powder.
3. When the outer most edge of powdered material starts to melt, ie. begins to turn dark green, add 10 mls of extraction buffer (1 ml per 1 g of tissue)* directly into the mortar and continue to grind until the whole bowl and contents begin to warm above freezing.
4. Add 100 µl of 20 % SDS to the extraction mixture, mix and then transfer the contents back into the original Greiner tube. Remember to add proportional volumes of SDS if the extraction buffer volume has exceeded 10 ml.
5. Proceed through steps 1 to 4 on the next sample.

Once the leaves are in the the extraction buffer the DNA content is safe from degardation by DNase as the EDTA concentration is inhibitory to enzyme activity. The SDS presence disrupts the cell wall, leaving nucleic acids free to further extraction.

Dried material: if these have been treated properly and dried slowly then they will grind down into a fine powder. If the leaves are slightly damp then it is possible to speed up the drying process prior to DNA extraction by placing the paper bags of leaves in a 40°C oven for 30 minutes. It is better to use oven drying just before DNA preparation and not immediately after harvesting as this heating process accelerates DNA degardation.

1. Grind dried leaves in a mortar to a fine powder.
2. Add 10 ml of extraction buffer* and continue to grind.
3. Add 100 µl of 20 % SDS to the extraction mixture, the same applies as 4 above if more than 10 ml of buffer have been required.
4. Mix the contents and transfer to a labelled 50 ml Greiner tube.
5. As for 5 above.

* It may be necessary to add more extraction buffer as this depends on the number and size of the leaves. The rule of thumb is to start with 10ml then add additional 5ml volumes if required.

4.4.1.3 DNA extraction

Both frozen and dried leaf material can now be treated identically using the following method of DNA extraction.

1. Shake the tube contents briefly and place the tubes in a 37°C waterbath for 5 minutes.
2. In the fume hood shake the tube contents briefly and add an equal volume of chloroform/IAA** (24:1) and shake again; balance tubes based on volume and centrifuge at 4000 rpm for 8 min (this is the first stage of deproteination).

3. Remove the upper aqueous phase, again do this in the fume hood, to a freshly labelled tube using the 5 ml Gilson pipette; try not to disturd the lower chloroform phase and the central green mass of leaf material. Repeat this for all the samples before proceeding.
4. Precipitate the nucleic acid by adding 2 volumes of 100 % ethanol into the decanted aqueous phase. Mix very gently by inverting the tube slowly. At this point it is possible to spool out the nucleic acid (for some crops like pea), but Lablab does not spool; in this case centrifuge at 4000 rpm for 10 minutes.
5. Pour off the ethanol carefully, (great care is needed here as the pellet will often start to move in which case it may be best to remove the last traces ot ethanol with a pipette) place the tube on its side and leave the pellet to dry.
6. Resuspend the pellet in 10 ml of TE pH8, add a further 5 ml if the pellet is reluctant to dissolve.
7. Degrade the RNA with addition of 2 µl of RNase A (100 mg/ml), incubate at 37°C for 5 minutes.
8. In the fume hood add 5 ml of phenol*** to each sample and 5 ml of chloroform/IAA (24:1) and shake; this phenol/chloroform treatment denatures the RNase and continues the deproteination process. Centrifuge at 4000 rpm for 8 minutes.
9. Decant the upper aqueous phase (ideally it is clear and not cloudy) to a fresh labelled tube; precipitate the DNA with the addition of 2 volumes of ethanol; gently invert and you may see the white strands of high molecular weight DNA. Centrifuge at 4000 rpm for 10 minutes and pour off the ethanol. The pellet will be either white in colour in which case the pellet will remain at the bottom of the tube or the pellet may be clear and more susceptible to movement. Once again take care when decanting and remove the last traces of ethanol with a pipette. Place the tube on its side and allow the pellet to dry.
10. Resuspend the pellet in 100 to 200 µl of TE pH8. If it is necessary to encourage resuspension do so using a P1000 blue tip with the end cut off, this opens the bore and prevents shearing of the DNA.
11. Spin down the tube contents with a brief centrifugation of less than 1 minute and transfer the DNA solution with an opened bore tip to a 1.5 ml microtube.
12. Give a final chlorform/IAA (24:1) extraction, this will remove any remaining traces of phenol. Make the volume to 300/400 µl with TE pH8 and add equal volume of chloroform/IAA (24:1), mix by inversion and spin at full speed for 5 minutes in micro-centrifuge.
13. Using an open bore P1000 blue tip, pipette off the upper aqueous phase to a fresh labelled tube and add 2 volumes of 100 % ethanol, mix gently, spin at full speed for 8 minutes.
14. Pour or pipette off the ethanol carefully leaving the pellet intact add 500 µl of 70 % ethanol and spin full speed for 5 minutes.

15. Remove the ethanol as in 14 above and leave the pellet to dry. Resuspend the pellet in 20 – 50 µl of TE pH8, or more if required, depends on the amount of DNA extracted. Aim for a DNA concentration of approximately 0.5 µg/µl. This looks viscous but not too fluid when the point of the tube is tapped, typically it is necessary to add 50 – 100 µl of TE.

** Chlorofrom is a hazardous solvent and melts plastic readily, take great care when handling and avoid chloroform on outside of tubes as it erases labels; always wear gloves when handling chloroform containing liquids.

*** Phenol is also a hazardous sovent that burns and is readily absorbed through the skin; always wear gloves when handling phenol containing liquids

4.4.1.4 Agarose gel assessment of DNA concentration

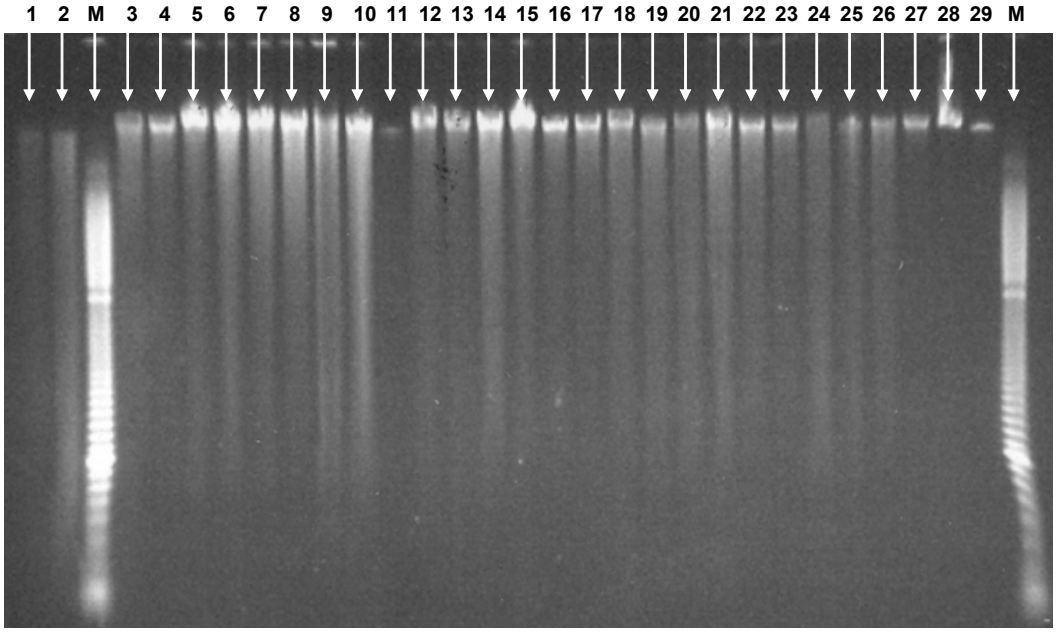
There are two simple methods of assessing DNA concentration, one is to take a dilution of your DNA in water and make a spectrophotometric reading at 260 nm, the other is to run a sample on agarose gel, stain with ethidium bromide and compare the concentration against known standards.
The preferred method for this course is agarose gel. This will show if the DNA is sheared, there is RNA remaining, will provide a DNA concentration and is more sensitive to weak concentrations of DNA. Typically 10 ng can be detected by ethidium bromide.

1. Using a 500 ml Duran bottle make up 250 mls of 0.8 % agarose in 1xTBE buffer, heat in a microwave oven on medium power until boiling and dissolved, leave the lid loose the whole time so that pressure cannot build up in the bottle; leave to cool t 65-80 ºC.
2. Prepare the gel casting tray with the desired number of wells and tape the end, pour the gel, not too deep, pour until the whole base is just covered with molten agarose and allow the gel to set for 1 hour.
3. Take 2/3 µl of each DNA preparation to 4 µl of SDW (sterile distilled water) and add 5 µl of orange G loading dye.
4. Prepare the standards using the lambda DNA at 0.5 µg/µl in the same way as in 2 above. The suggestion is to use 4 standards of 0.5, 0.25, 0.1 and 0.05 µg.
5. Remove the tape at the ends of the casting tray and place in the electrophoresis tank, pour in 1xTBE buffer and remove the comb.
6. Load samples and standards, segregate the samples into blocks of 15 or so separated by a DNA ladder; run at 100 V for 1-2 hr depending on the number of combs lined up in parallel.

7. Remove the gel from the casting tray to a deep tray containing 5 µl of 10 mg/ml ethidium bromide solution in 150 mls of 1xTBE and stain for 20/30 minutes; rinse the gel briefly with water.
8. Visualise under UV and photograph.
9. Store DNA at 4°C.
10. Assess the DNA concentration of each against the standards.

The photo below (figure 2.3.1) represents DNA preparations of *Lablab purpureus* from dried and frozen leaf material, run out on 0.8 % agarose in 1xTBE buffer; all samples were 3 µl loading.
Lanes 1 and 2 are dried samples from the UAS collection. L is the 100bp ladder. Lanes 3 to 27 are CPI, ILRI and UAS lines all from frozen leaves. Lanes 28 and 29 are the standards 0.5 and 0.05 µg of lambda DNA.

Figure 1



The photo shows most of the frozen preparations are of high molecular weight and are approximately or in excess of 0.5µg/µl; many lanes are overloaded. Lanes 1 and 2 from dried samples appear mostly sheared with no apparent high molecular weight DNA. Shearing of the DNA is best avoided where possible but a little can be tolerated where small PCR products are to be generated. All of the DNA samples illustrated are good enough for PCR templates (in this case AFLP templates were generated).

4.4.1.5 Preservation of DNA using FTA® paper (Whatman)

Fresh leaf material is pressed onto specially treated filter paper, leaf disks are made and can be stored at room temperature (25°C) for long periods. When required a disk is then taken through a purification

stage with there being sufficient DNA for PCR to be carried out directly. The protocol for using FTA paper, devised by Whatman, will be tried on this course, please see Appendix 2 for details, the protocol for AFLP using FTA card DNA is also included.

4.5 Molecular markers versus morphological markers

4.5.1 Why use molecular markers?

In de Vienne's book the issue of marker types is discussed in the Introduction and in Capter 1. Classical or morphological markers are very useful, but it is rare to find very many segregating in a given cross, and it is unusual to find them segregating in crosses between breeder's lines which are often selected for a particular overall form and growth habit. Molecular markers can vary extensively between plants with a similar morphology and physiology, provided they are not closely related by descent. Molecular markers have the advantage that they are abundant and have the potential to differ between plants with little or no consequence for their performance.

On the previous courses, experiments examined the molecular diversity of *Lablab* and AFLP was the marker system chosen. The use of the AFLP marker method enabled a rapid analysis of the germplasm available to us and introduced a reasonably complex marker system that involved some basic but complex molecular techniques.
AFLP are robust markers and can be used further in mapping studies, but the aim of this course is to introduce alternative marker systems that can be utilised in addition to AFLP.
On this course we will be using Simple Sequence Repeat (SSR) and exon spanning markers (intron-directed), both of which are PCR based and relatively straightforward to carry out. Both of these molecular marker methods are popular and widely utilised in mapping studies. These two marker types require primers to be designed from sequence information either from in house sequence analyses or collected via database searches. This latter option is fine if your species of interest has an abundance of public sequence information.
However, there is limited sequence availability in databases for *Lablab* (September 2005: about a dozen *Lablab* sequences from which primers can be designed) but we can use sequence information from related legumes. In fact, some of these studies have already been done and we can test whether or not the primers designed for other legumes will successfully amplify *Lablab* nuclear DNA. It is from these studies with three species in particular, the model legume *Medicago truncatula*, *Glycine max* and *Vigna unguiculata*, the latter two being more closely related to *Lablab*, that we will be testing primer combinations on this course.

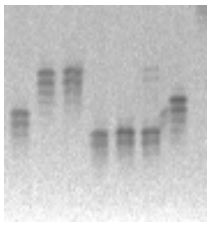
Two *Lablab* mapping populations have been generated at the UAS, they are:
1. HA-3 x Mac-1 (GL415 x GL404)
2. HA-3 x SR-L (GL415 x GL553)

Both populations are available at the F_2 generation. Preliminary analysis, involving the three parental lines of the crosses and also included Rongai, has already been carried out on a designated set of markers. However difficulties in obtaining polymorphic differences at the molecular level have been encountered during this preliminary marker analyses. There is yet a further set of primer combinations that can be tested

On the practical laboratory course the procedure for testing a typical set of markers, that will eventually lead to their use as loci on a map, will be demonstrated.

4.5.2.1 SSR markers

The de Vienne course book describes SSR markers on p 29 – 31. Primers are designed from the non-repeat regions of genomic DNA that flank the specific microsatellite repeat (Fig 10, De Vienne). The amplification products of these markers often produce a stutter of bands similar to the Figure 2 below.



239 bp

Figure 2: A typical microsatellite (SSR) marker stutter pattern on PAGE (Polyacrylamide Gel Electrophoresis), the bands are from a pea SSR marker, the lower set of bands are 239 bp. The polymormophic differences among 7 *Pisum* lines are clear.

The band stutter in each track, Figure 2, is usually attributed to replication slippage of the Taq polymerase during the PCR, this is explained and discussed by de Vienne. The leading major band in each track suggests that within the 7 *Pisum* lines there is polymorphic variation for this particular microsatellite repeat.

4.5.2.2 How does an SSR marker create a codominant marker

Using the SSR marker from Figure 2, a mapping population generated using the line 2 crossed with line 4 at F_2 would show, ideally in a 1:2:1 ratio:

a. some individuals with the larger band series from line 2

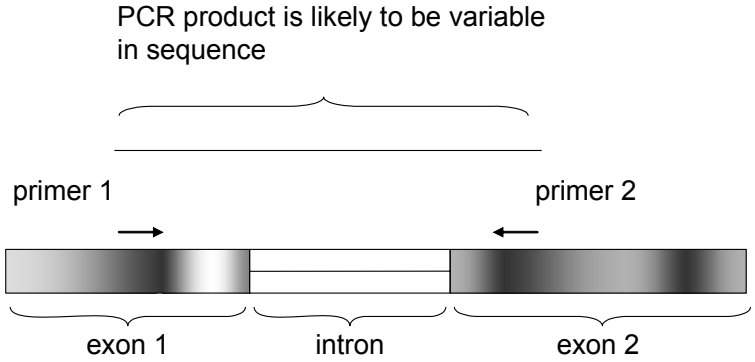
b. some individuals with the smaller band series from line 4
c. some individuals with both the larger and smaller band series, much like line 6, but the upper band series in this picture is a bit faint.

The set of SSR markers that have been tested for this course, for use with *Lablab*, are primer pairs from Wang [7], and are listed in Appendix 3; 32 primer pairs were tested in total, 18 designed from *M. truncatula* and 14 from *G. max*. Of this set 11 failed to amplify a band in the 4 *Lablab* lines tested. Those primer pairs that did amplify *Lablab* DNA showed there were no obvious polymorphic differences between the 3 *Lablab* parental lines and Rongai on agarose gels. Some of the successful primer combinations gave a multi band pattern with *Lablab* and a couple of these have been tested on PAGE and visualised using silver stain. There are a few polymorphic differences but these appear to be dominant marker types: much like those generated using AFLP.

The course work will continue with the primer screening techniques that have been used so far.

4.5.3 Intron-directed markers

These can also be called intron-directed PCR markers and their structure is illustrated in the figure below:



The darkness of a region is intended to indicate the degree to which its sequence is conserved between species or individuals.

[7] Wang et al. 2004 Transfer of simple sequence repeats (SSR) markers acros the legume family for germplasm characterisation and evaluation. Plant genetic Resource 2 (2), 107-119

Primers are designed to regions known to be well conserved, so the primers will work well in a range of plants, but the region amplified is likely to be variable in sequence.

This marker type also utilises sequence information from databases often from ESTs (expressed sequence tags). Primers are designed aiming to span an intron as above. Although the intron is missing from the EST, its location can be inferred by sequence comparisons to genomic DNA or gene models from related species. This may be selected sequencing of PCR products, whole genome sequence or BAC (bacterial amplified chromosome) end sequencing. The PCR products amplified from genomic DNA will contain both intron and exon sequence, and may be designed to contain more than one intron. A gene model is annotated sequence usually from a related species. As intron location is generally conserved over wide evolutionary distances this information can be used to design primers anticipating the positions of introns in the target species.

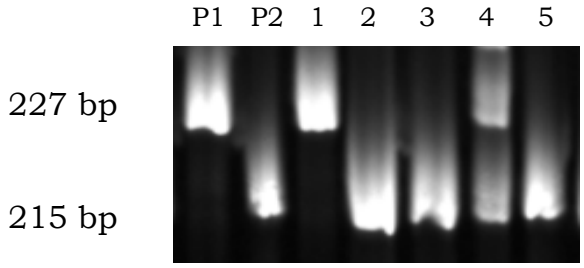


Figure 3

Figure 3 above shows the fragment pattern on PAGE from 5 individuals from an F_2 population of *Pisum sativum* with an exon directed marker for the gene Cathepsin; parental lines are P1 and P2, individual 4 is a heterozygote and is clearly distinguishable from the homozygotes, individuals 1, 2, 3 and 5, that have the presence of either the P1 or P2 allelic state for this marker.

A range of primers from *P. sativum* gene sequences were tested in the 3 *Lablab* parental lines and Rongai, none of which gave amplification.

In an ideal situation the intron spanning / exon directed markers also will be codominant. Preliminary experiments using 8 *M. truncatula* or *Vigna radiata* primers from Choi [8] using *Lablab* DNA were not encouraging. Of 8 primer combinations tested only 1 gave an amplification with *Lablab* DNA.

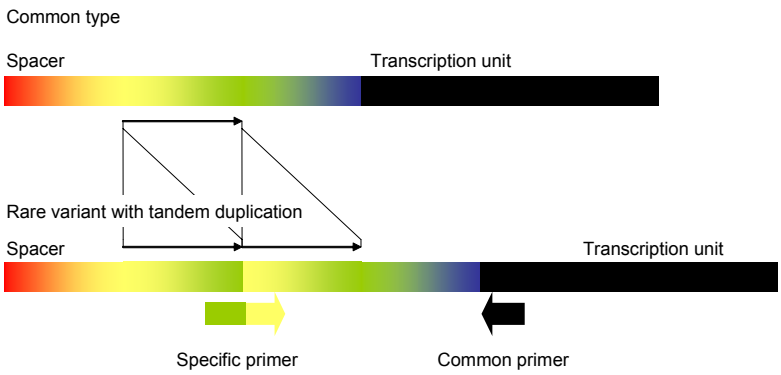
[8] Choi et al. 2004 A sequenced-based genetic map of *Medicago truncatula* and comparison of marker colinearity with *M. sativa*. Genetics 166: 1463 – 1502.

There are many more that potentially may provide markers for *Lablab* genetic mapping studies. Some of these provided by DJ Kim [9] (2005) will be tested during the course.

4.5.4 Allele specific PCR

Allele-specific PCR is a means of scoring single alleles at a locus independently, and so usually corresponds to a co-dominant marker (unless just one f the two loci is assayed). The point about allele specific PCR is that with careful primer design a simple PCR allows the identification of a specific allele. A similar effect can be obtained for example with CAPs markers, but these alternative methods require additional processing (eg digestion of the PCR product in CAPs).

One approach to allele specific amplification is to design primers around structural rearrangements in DNA sequences. For example in the amplification of a specific allele of the 5SrRNA genes in pea, a rare variant identifies one allele. This allele carries a length variant of the 5SrRNA gene repeat that includes a small duplication of sequence in the spacer region. Although this repeat contains sequences present in all other 5SrRNA gene repeats, a unique junction is generated such that a primer specific to the rare repeat can be identified. This is a slightly messy approach, but illustrates the power of primer design
(See figure:)

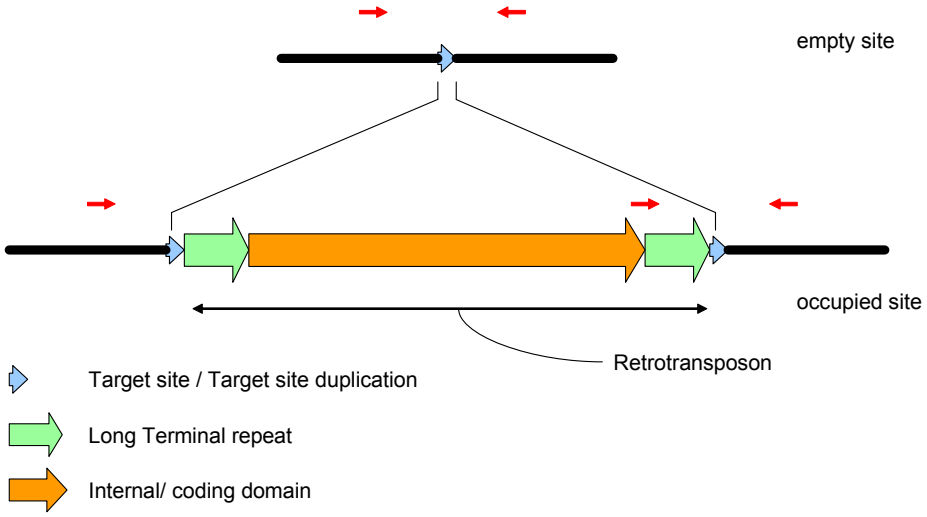


A second approach that has been used extensively in pea genetic diversity assessment and is the so-called RBIP (Retrotransposon Based *I*nsertion Polymorphism) marker system[10]; in this case three primers in a single PCR can detect the presence or absence of a retrotransposon insertion at a specific locus. (See figure below).

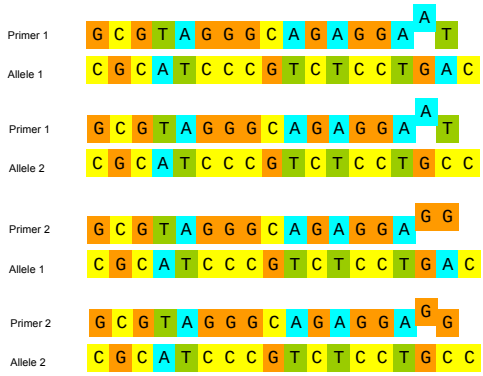
[9] DJ Kim 2005 Personal gift
[10] Flavell A.J., Knox M., Pearce S.R., and Ellis T.H.N. (1998) Retrotransposon-based insertion polymorphisms (RBIP) for high throughput marker analysis. Plant Journal 16: 643-650,

Coupled with a primer design that identifies SNPs (ARMS)[11] this procedure can be assayed in the same way as RBIP markers[12].

RBIP markers:



The principle of 'ARMS' primer design is that a single mismatch near the 3' end of a primer reduces the efficiency by which polymerase is able to synthesize an extended sequence from the primer, but that two mismatches (usually) reduces this to an undetectable amount. This means that mismatched primers can be used to generate allele specific PCR products.
(See figure).



In the example shown primer 1 can be extended with allele 1, but not allele 2 as a template, while primer 2 can be extended with allele 2, but not allele 1 as a template. Primer 1 or 2 in conjunction with a common primer will generate amplification products from only one allele. Note that the choice of mismatched primer is critical and

[11] Newton CR, Graham A, Heptinstall LE Powell SL Summers C Kalsheker N Smith JC and Markham AF (1989) Analysis of any point mutation in DNA. The amplification refractory mutation system (ARMS). Nucleic Acids Res. 17: 2503 - 2516

[12] Flavell A.J., Bolshakov V.N., Booth A., Jing R., Russell J., Ellis T.H.H. and Isaac P. (2003) A microarray-based high throughput molecular marker genotyping method - the Tagged Microrray (TAM) marker approach. Nucleic Acids Research 31:e115

should differ between the two primers to maximise fidelity. For further details see Newton et al (1989).

4.5.5 SNP markers
Single nucleotide polymorphisms (SNP) provide many opportunities for the generation of markers. Indeed several of the marker types discussed in this manual can be SNP assays. RFLPs for example may correspond to a single nucleotide difference. Current efforts in several labs seek to exploit oligonucleotide arrays as a way of assauing many SNPs in parallel. This way of exploiting SNPs requires extensive genome sequence information and so will not be discussed in the course.

One easy method of detecting SNP markers is by SSCP (described below *4.6.3.2*), this method will be demonstrated during the course. SNP also lend themselves to a range of high throughput automation methods such as fluorescent reader-based or Mass-spectrometry [10].

4.6 Parental screen using molecular markers

The following series of experiments with a set of primers from the sources described earlier (Wang et al 2004, Choi et al. 2004; DJ Kim 2005), designed from *M. truncatula, G. max*, *V. radiata, V. unguiculata*, *P. sativum* and *L. purpureus,* uses the standardised PCR conditions set out below. This regime was tested and used routinely in the preliminary experiments at JIC for use with *Lablab* DNA using primer sequences designed from these related legumes.
Positive amplification with control DNA, ie. DNA of the species from which primers were designed, was taken as successful PCR. The four *Lablab* lines were each tested in duplicate.

4.6.1 Initial primer screen with the Lablab parental lines: experimental procedure

a. Set up the PCR for testing with *Lablab* lines HA-3, Mac-1, SR-L, Rongai each should be carried out in duplicate and use the appropriate DNA as positive control; also include a negative control, this will show primer dimer. The sequence information for the primer combination is given in Appendix 3.

Table 1: PCR components

Components	**x 1 (µl)**	**Master Mix x 12 (µl)**
*15 ng Primer forward (7.5 ng/µl)	4	48
*15 ng Primer reverse (7.5 ng/µl)	4	48
200 µM each dNTP (1mM)	4	48

10 X PCR buffer	2	24
1U *Taq* Polymerase (5U/µl)	0.2	2.4
SDW to 18 µl	3.8	45.6
50 ng DNA** (50 ng/µl)	2	(216 = 12 x 18 µl)
Total PCR volume	20	

* Or use at 0.1 µM final concentration (2 µl of 1 µM in a 20 µl PCR volume).
** If the DNA concentration is initially at approximately 0.5 µg/µl, dilute tenfold in SDW.

b. The PCR cycling conditions uses a touch down regime as follows:

94°C/ 3 min
(94°C/ 30 secs; 50°C/30 secs; 72°C/1 min) repeat for 10 cycles reducing the annealing temperature by 0.5°C /cycle
(94°C/ 30 secs; 45°C/30 secs; 72°C/1 min) repeat 30 cycles
72°C/ 10 min
12°C/ 10 min
end

c. Divide the 20 µl PCR into 2 x 10 µl, ie. take 10 µl to a fresh set of microtubes and store at -20 °C for running on PAGE later, if agarose shows amplification

d. To the other 10 µl of PCR product add 5 µl of agarose loading dye, orange G

e. Load the whole 10 -15 µl sample onto a 1.5 % / 1 x TBE agarose gel, run at 100V for 1-2 hrs till the orange dye comes close to the end of the gel and stain with Ethidium bromide.

f. Take a photo and analyse the banding patterns observed

g. Make a table of results, both positive and negative, in Excel. Record whether or not there was amplification in the Controls and *Lablab*.

h. Combine all the data from all groups into one master Excel file.

4.6.2 Positive amplification and further testing

Those primer combinations where there has been amplification with the *Lablab* lines the saved 10 µl sample can be tested further. Either

this can be on 3 % agarose or 4.5% PAGE with silver staining (see Appendices 4 and 5) for details of these techniques).
The choice of method is dependent on band sizes obtained with *Lablab* and whether or not it is possible to resolve band size differences on agarose. Repeating on 3 % agarose is straightforward, follow steps *2.5.1* steps d - h above.

The preliminary experiments carried out at JIC suggested that 4.5 % PAGE was generally necessary. The method for visualising bands on silver stain gels is as follows:

a. to the 10 µl of saved sample add 8 µl of the acrylamide gel stop/loading buffer and denature the samples at 95ºC for 3 min, cool to 12ºC for 10 min. Either store at -20ºC or hold on ice and run 5-8 µl of each sample on denaturing PAGE (see Appendix 4)

b. Visualise bands after silver staining (see Appendix 5)

c. Make a hard copy of the gel and analyse the band pattern, record results.

d. Add the results to the Excel sheet in *2.5.1* g and h above.

4.6.3.1 Running the population on PAGE

Once polymorphic differences are observed the specific primer combinations responsible can be tested on the mapping population.

a. Repeat each PCR in 10 µl final volume, ie. reduce the component volumes in Table 1 above by half and scale up for the number in the population, to also include the parental lines each in duplicate

b. Repeat the steps *2.5.2* a – d for visualising on PAGE and silver staining.

4.6.3.2 SSCP – single strand conformation polymorphism

SSCP is an alternative PAGE method that can be used to visualise limited sequence variation, ie. SNP, a single base pair mis-match between two allelic sequences. This method utilises the migration characteristics of single stranded DNA and conformational polymorphic differences in secondary structure and mobility through a gel matrix. This method works optimally for fragments of less than 400 bp in length and is ideal for SNP (single nucleotide polymorphism) detection.

The method uses neutral polyacrylamide gels (non-denaturing) and they are run longer and at low voltage, preferably in a cold room, but good results can be had at 22-25ºC. Other than differing gel conditions, sample preparation and loading, and silver staining is the same as described earlier *2.5.2* steps a-d. Gel conditions for SSCP can be found in Appendix 6.

Properties of oligonucleotides (primers)

You may need to calculate the molecular weight, melting temperature or some other property of an oligonucleotide which depends on its base sequence. Programs are available to help you do this. For example:

http://www.basic.northwestern.edu/biotools/oligocalc.html

5. Statistics in genetic mapping

It is important to have all the scoring data in an Excel spreadsheet from which it can be analysed further and later converted to text files.

Before using a data series in mapping it is good practice to carry out some preliminary analysis of the data. The most obvious tests to carry out are:
1) A check that as many individuals as possible in the population have been scored for the marker, or at least that the majority of the marker scores are informative.

2) A test to be sure that the markers correspond to single genetic loci.

3) A test that the segregation of the marker is unbiased in the population. Sometimes it is necessary to accept markers with distorted segregation because regions of the genome may have biased segregation, and should not be ignored, but if segregation is distorted it is important to know about that.

5.1 Segregation ratios and the x^2 Test

Using the JoinMap mapping programme these calculations are automatically done within the programme. However with MapMaker this facility is not available but these calculations can be done using MS Excel. As we will be focusing mostly with Mapmaker on this course the next section describes how to carry out these calculations, plus there are 2 Exercises included in this section.

5.1.1 Using Excel

Open the file F2_data.xls found in 2005 course folder.

This file is split into 3 worksheets, choose the F2_scores worksheet.

The worksheet named F2_scores, contains the raw scoring data for an F_2 population (pea, JI15xJI399), there is a summary of experimental details at the top left hand corner; it is advisable to put information here for reference. This data set contains 153 markers (columns A and B), scored for the 120 individuals in the F_2 population (rows 5 – 157). The markers are scored using the appropriate codes necessary for the mapping programmes (see 3.4).

For future reference it is good practice to keep a worksheet for the raw data only and to set a protection (Tools, Protection, Protect sheet, give a password), so that the original data remains intact and cannot be

accidentally deleted or changed. From this master data set other worksheets can be opened for further analysis.

Open worksheet F2_analysis this is just a copy of the master worksheet F2_scores from which a series of statistical analyses can be carried out.

Exercise 1: Segregation ratios

a. From the F2_analysis worksheet take a batch of the data at a time and replace the Bs with 1. Use the SUM feature in Excel to add up the scores for B for the whole batch. Remember to replace 1 back to B before repeating the whole process again for D.
Repeat again for all the other data, ie. C and A.

b. Calculate the segregation ratios, for the dominant markers, ie. those that are D:B and C:A, a 3:1 ratio is expected.
The three codominant markers at the bottom are expected to be 1:2:1 for each of the homogygotes and heterozygote.

Exercise 2: x^2 Test

Using the x^2 we can test statistically whether the segregation ratios for a single marker fit the expected ratios for dominant and codominant markers.

The formula for calculating x^2 is below:

$$x^2 = \sum \frac{(O - E)^2}{E}$$

For an RI population a 1:1 ratio is expected regardless of the marker type. However for an F_2 population a dominant marker is expected in a 3:1 ratio and for a codominant marker it is 1:2:1. Below are two worked examples.

a) Markers from an RI population

The observed scores from an RI population of 100 individuals, for a particular marker, are A = 52, B = 37, and missing scores = 11. The expected ratio is (52 + 37)/2 = 44.5.

The x^2 will be:
$[((52 - 44.5)^2/44.5) + ((37- 44.5)^2)/44.5] = 2.53$

If the ratio is exactly 1:1 a x^2 of 0 will be obtained; the value in this case is 2.53 and with 1 degree of freedom is less than the 3.84 at the 5% level of significance (Table, Appendix 7, p = 0.05); this

marker data therefore has a segregation ratio that fits the expected at p = 0.05.

b) Markers from an F2 population: codominant

The observed score from an F_2 population of 120 individuals is: A = 22; H = 57; B = 30; missing scores = 11. There are 109 individuals with scoring information, so the expected ratios are 27.25 : 54.5 : 27.25. How close do the observed data come to this?

The x^2 will be:

$$[((22 - 27.25)^2/27.25) + ((57- 54.5)^2)/54.5 +((30 - 27.25)^2/27.25)]$$
$$= [1.01 + 0.11 + 0.28] = 1.4$$

At p = 0.05 and with two degrees of freedom this value of 1.4 is below 5.9 (Appendix Table 7) and the score for this marker fits a 1:2:1 ratio.

c) Markers from an F2 population: dominant

Using the expected ratio information given above for dominant markers work out the x^2 value for the following three markers:

i) A = 28 and C = 89
ii) B = 19 and D = 101
iii) B =21 and D = 97

Are the x^2 values significant at the p = 0.05 level?

Would you use these markers in a mapping analysis?

5.2 LOD scores

A LOD score is a similar type of test to a x^2, but it has a slightly different philosophy. The x^2 checks whether the observed data is significantly different from the expected value in a frequency distribution. A LOD score measures the support for a given statement against a null hypothesis given the data. This looks at the issue from a different angle. In practice LOD scores are usually used when thinking about linkage rather than monogenic segregation. LOD means the log of the odds ratio.

For example, in a recombinant inbred population we may observe that for two markers the fraction of lines that do not have the parental configuration of alleles is *R*. If the population size is *N* then there are

RN lines that have recombined the parental alleles and (1-*R*)*N* lines that have the parental configuration. We can let *x* equal *RN* and *y* equal (1-*R*)*N.*

We have two hypotheses: One, the null hypothesis that the loci are unlinked, and the alternative that they are linked with an intensity that gives a proportion *R* of recombinant inbred lines (RILs) that have recombined the parental alleles. The form of words is a bit complicated because *R* is not the recombination frequency, see later and don't worry.

If the two markers are unlinked, then the odds (Ou) of getting **exactly** *x* RILs that have recombined the parental alleles and **exactly** *y* RILs that have not recombined the Parental alleles is:

$Ou = (0.5^{x})(0.5^{y})$

If the markers are linked with the intensity proposed above we have the the odds (Ol)

$Ol = (R^{x})[(1-R)^{y}]$

The odds ratio is Ol/Ou and the LOD score is $Log_{10}(Ol/Ou)$

You can also calculate a x^2 for such linkage, and it is an informative exercise to plot the x^2 against the LOD score for a range of linkage intensities. There is some nicety in the statistics, but the two tests tell you more or less the same thing. LOD scores are fashionable because in human genetics most linkage values are calculated from (statistically) small families, so it is difficult to get significant tests. In multiple tests it is legitimate simply to add the LOD scores, and that's easy to do.
Note that the LOD score is not the likelihood of linkage, because there are many more ways of being unlinked than linked.

In de Vienne's book this is discussed in box 2.1. You will notice that his likelihood estimation includes a binomial coefficient. This is omitted above, because the same binomial coefficient applies to Ol and Ou, and they cancel when the odds ratio is used.

5.3 Basic Statistics and R

5.3.1 Introduction

R is free software. It can be used as an overpowered calculator, to carry out many standard statistical analysis, and to develop applications carrying out complex analyses in specific subject areas.

Many of these applications, and much other useful information are freely available from the Comprehensive R Archive network (CRAN) web site http://cran.r-project.org/. In particular, the Guide "An Introduction to R" is available under the documentation section and can be used to supplement the outline given here.

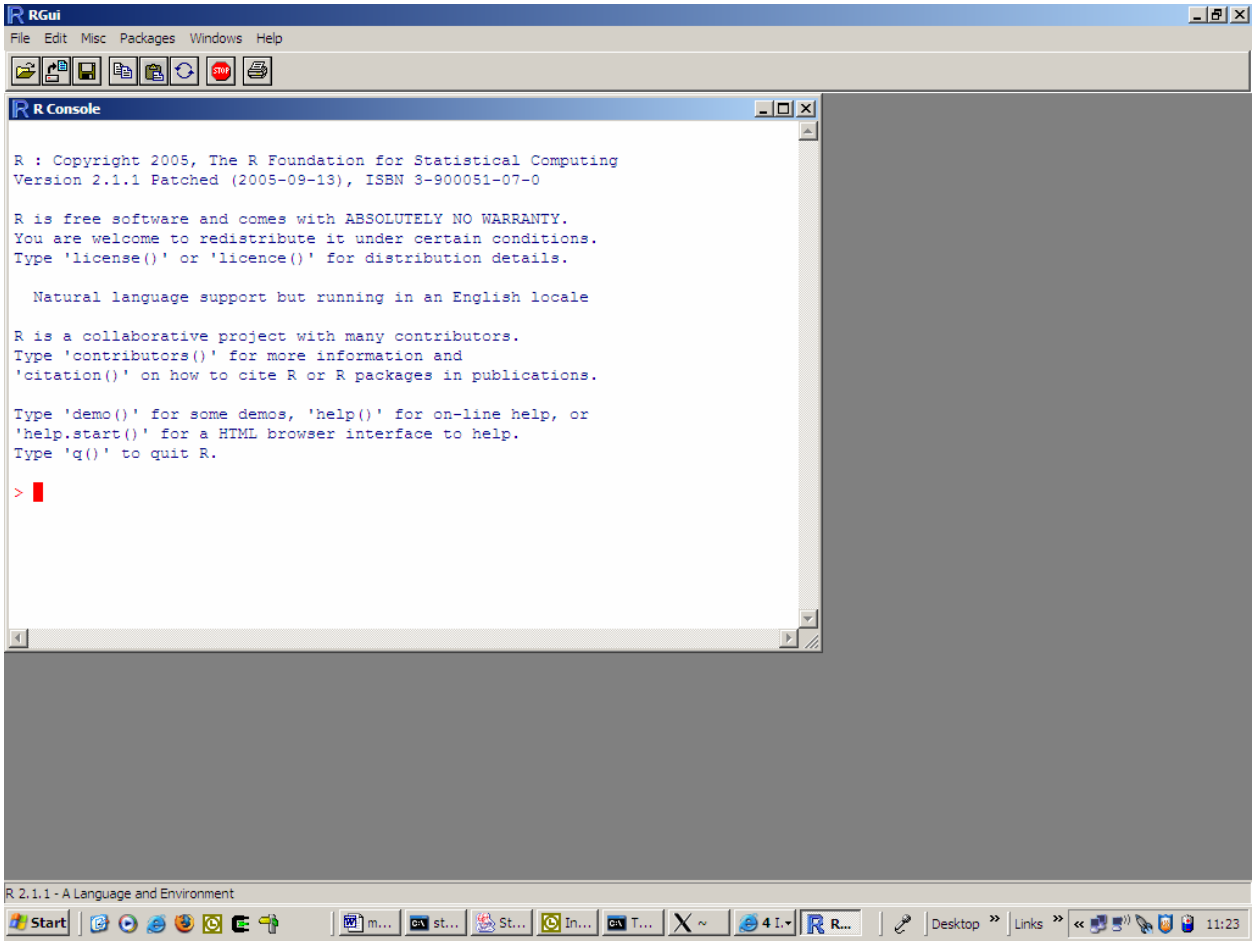
R is one among many excellent statistical packages, for example S, Splus, SAS, SPSS, STATA, to go through just one letter of the alphabet. Compared with the alternatives, it has its own set of strengths and weaknesses. The major strength is that it is free. This has led to an expanding user base and to the development of more and better applications. As a result R is increasingly used in larger commercial and academic establishments where cost is less of a concern: knowledge of R is now a useful and transferable skill to acquire. For many users, the major disadvantage of R is that commands need to be typed at a prompt, rather than selected from a menu. However, most commands are short, and as we shall see, much repetitive typing can be avoided.

The intention of this guide is to help the novice user get started. It will show you how to get data into R, carry out simple analyses, produce graphs, and save results. It is hoped that this will allow users to dip directly into the more complete and thorough guides such as "An Introduction to R" as required. This guide is a long way off being complete and many of the definitions and descriptions provided here are not wholly accurate. The hope is that the guide will allow you to carry out useful analyses as quickly as possible and provide a base from which additional knowledge can be acquired without too much extra effort.

5.3.2 Starting R

For this course, R has already been installed on your computer. For future use, it can be downloaded from http://www.r-project.org/ or from the CRAN web site mentioned above. Installation is straightforward, but if you are concerned, or have never installed software before, it would be best to get assistance from your IT department or from a more computer literate friend or colleague.

To start R, select R from the Windows Start menu. Precise details may vary slightly with installation. On my machine, I select "R.2.1.1 Patched." You should then see the screen shown below. All the usual Windows conventions about maximizing and minimizing, copying (Cntrl C) and pasting (Cntrl V) apply.



Commands are entered into the R console window – the window containing all the text. There are a few conventions that must be followed when working in this window. These are described in the next section.

5.3.3.Basic R syntax

R shows it is waiting for user input by displaying a prompt:

```
>
```

To enter a command, type alongside the prompt and complete by hitting the carriage return (Ret).

For example entering

```
> 9+1
```

will return

```
[1] 10
```

You are now fully trained to use R as a calculator. The syntax is essentially the same as used in, for example, Excel. Thus ^2 will square numbers, sqrt() will take the square root and so on. For example:

```
> sqrt(log10((9+1)^2))
```

will return the square root of 2.

```
[1] 1.414214
```

Note that R is case sensitive:

```
> sqrt(LOG10((9+1)^2))
Error: couldn't find function "LOG10"
>
```

If you make a mistake, it can be corrected by tapping the up arrow on the keyboard to bring back the last typed command, moving along the command line using the left and right arrows, deleting characters using the backspace or delete key, inserting correct numbers or text and hitting the carriage return key.

Multiple hits of the up arrow will bring back successively earlier commands. However, there is an easier way of finding and re-entering commands used some time previously, which shall be introduced later. Note also that it is possible to cut and paste from other windows. So R commands listed in another document, such as this one, can be copied and pasted into the R console and will run (after hitting carriage return).

Brackets

We have seen brackets used above in simple formulae – just as in Excel formulae. However, brackets are used more extensively in R: **all** commands include brackets. For example:

```
> quit()
```

will end your R session. However:

```
> quit
```

without the brackets returns:

```
function (save = "default", status = 0, runLast = TRUE)
.Internal(quit(save, status, runLast))
<environment: namespace:base>
```

This is the section of computer code which is run when you type `quit()`. Unless you are an enthusiast, it is not necessary to know what this means.

Equals symbols

R has four sets of symbols, all of which loosely mean equals, but which have differences in use. This can take some getting used to.

The most commonly used set is

```
<-
```

or sometimes

```
->
```

These symbols are typed using the dash (or minus sign) and the left or right arrow. They are used to assign results from one side of the arrow to the side to which the arrow is pointing.

For example:

```
> my_first_result <- sqrt(log10((9+1)^2))
```

just returns the prompt:

```
>
```

But then typing

```
> my_first_result
```

now returns:

```
[1] 1.414214
```

We have created a variable "my_first_result" and assigned the result of our calculation to it, using "<-" The same result would be achieved by

```
> sqrt(log10((9+1)^2))-> my_first_result
```

The entry stored in my_first_result is now available for additional manipulation. For example:

```
> my_first_result^2
[1] 2
```

The second "equals" symbol is the tilde ~ (above the # on most keyboards). This is used in R in statistical analysis to distinguish between x and y variables, that is to say between what is being analysed (eg the phenotype) and what factors and variates it is being analysed with (eg marker data and environmental factors or variates). So for example:

```
P~G+E
```

is R syntax to state that a variate P is to be explained by two variates or factors G and E. Equivalently, we can say that P is described by G+E. The exact context in which this syntax is used will be given later. For the time being, note that

```
P~E
```

represents simple linear regression of P on E (or a one way analysis of variance if E is factor like soil type rather than a variate like altitude). Interactions can also be included:

```
P~G+E+G:E
```

includes an interaction term between G and E. This can be abbreviated to

```
P~G*E
```

Yet more complex models, for example with nested factors, can be described by including bracketed terms.

The third “equals” symbol is “=”. This is most often used within commands to provide information about specific parameters. For example we shall come across :

```
xlab=”what you put here is used to label the x axis in a graph”
```

used in commands to generate graphs.

The final “equals” description is “==”. This truly means “is equal to”, and is used to test relationships: is A equal to B is written as

A==B

Related to == are:

!=	not equal to
>	greater than
<	less than
>=	greater than or equal to
<=	less than or equal to.

These, together with “==” itself, are entered here for completeness. They are not generally required to carry out standard statistical analyses described in this guide.

Continuation character.

If an R command is incomplete when the carriage return is depressed, you are prompted with a + to continue the command on the next line:

```
> sqrt(log10
+
```

The rest of the command can be entered after the + :

```
> sqrt(log10
+ ((9+1)^2))
[1] 1.414214
>
```

Text and numbers

Text is distinguished from numbers and from R commands by the use of quotes. Either single quotes – ‘ text ‘ or double quotes “ text “ will do, but the quotes must match.

Summary of syntax

Arithmetic:	+ - / * ^ ( ) just like Excel
Commands use brackets:	quit() is correct. quit() is wrong
Equals	<- stores a result ~ is described by = options in commands == logical equivalence
Text	use quotes

This introduction should provide sufficient syntax to get you going. Additional syntax is introduced, as required, in the discussion of specific commands and operations in the remainder of this document.

5.3.4 Getting data in and out

Reading data

There are many ways of entering the data you wish to analyse. This guide describes only one. It is straightforward and can be used for most datasets.

Firstly we need some example data: results from a yield trial in which the slightly obsessive plant breeder has named all his varieties after statistical packages.

```
          yield     height    SNP1 SNP2
stata     94.2   37      1    NA
sass      93.7   37      1    1
R         115.1  19      2    1
genstat   90.1   32      1    1
Splus     91.2   33      1    1
S         99.2   27      1    2
SPSS      77.1   24      1    2
minitab   95.5   39      2    2
BMDP      87.2   36      1    2
mstat          119.   27      2    2
```

Note that the first row has four fields – the column titles – and that successive rows have five fields – the data. The fields are separated by spaces and tables: for the methods we shall use to import data, R is not fussy about which, or about how many, you use. Exact alignment of the columns is not required: here the title row and the last row are misaligned. Finally note that the first entry for each row of data is a unique identifier for that row. Here, this identifier is text field, but it could be a number.

Files in this format can easily be set up by exporting data as a text file from Excel, or by cutting and pasting from Excel into Notepad or Wordpad. The most important thing to remember is to get the number of fields correct on each line. Note that this means there should be no spaces within column headings of text fields. If spaces are required, either enter the name within quotes (eg "`SNP 1`") or substitute the space with an underscore (eg `SNP_1`).

Finally, note the code "`NA`" for `SNP2`, entry `stata`. This is the special code to denote a missing values. Missing values cannot be left as blanks for our method of data input: R would treat them as part of the field separator.

On my computer the data printed above are stored in the text file:

```
C:\Documents and Settings\x9901006\My
Documents\India\Rdemo1.txt
```

To get the data into R, we first need to change the working directory or folder to
point at the location of the file. The simplest way of doing this is to click on File at the top of the R window. Select "Change dir…", and then browse until you find the correct directory – exactly as you would with any other Windows package. There is a command line method of doing this too, which needn't concern us. To check that R is indeed pointing at the correct directory:

```
> getwd()
[1] "C:/Documents and Settings/x9901006/My
Documents/India"
>
```

Note the directory is not displayed in the standard MS Windows manner: the separator delimiting subdirectories is a "/" and not a "\". R was originally developed for the UNIX operating system, where "/" is used as the separator for subdirectories. This is not a problem unless

we wish to enter directory names at the command line, in which case we must remember to use / and not \.

To read in the datafile, use the command `read.table`:

```
> dataset1<-read.table("Rdemo1.txt")
>
```

This reads the data into a special R structure called a data frame, the details of which need not concern us. Note that the file name must be enclosed in quotes. The data can be displayed by typing

```
> dataset1
```

which returns

```
        yield height SNP1 SNP2
stata      94.2     37    1   NA
sass       93.7     37    1    1
R         115.1     19    2    1
genstat    90.1     32    1    1
Splus      91.2     33    1    1
S          99.2     27    1    2
SPSS       77.1     24    1    2
minitab    95.5     39    2    2
BMDP       87.2     36    1    2
mstat     119.0     27    2    2
```

Displaying data

To display specific variables, we enter them by name as follows:

```
> dataset1$height
 [1] 37 37 19 32 33 27 24 39 36 27
>
```

To avoid the tedium of entereing `dataset1$` in front of every variable, we can use attach()

```
> height
Error: Object "height" not found
> attach(dataset1)
> height
 [1] 37 37 19 32 33 27 24 39 36 27
> detach()
> height
```

```
Error: Object "height" not found
>
```

Note the use of `detach()` to remove the link between the data frame and R. This can be useful to load data from multiple source files and switch between them.

Methods for displaying a selected set of variables are not very intuitive. The command `subset()` achieves this, but not very elegantly:

```
> subset(data1,select=c(SNP1,SNP2))
        SNP1 SNP2
stata 1    1   NA
sass       1    1
R          2    1
genstat    1    1
Splus      1    1
S          1    2
SPSS       1    2
minitab    2    2
BMDP       1    2
mstat      2    2
>
```

This syntax of this command is fairly intuitive apart from `c(SNP1,SNP2). c(…,…,…,)` is a method of concatenating data into a single entity and is used quite extensively. It can also be used as a method of entering small amounts of data directly into R.

Once data have been read into R, the vectors which contain each variable can be manipulated in the same manner as individual numbers:

```
harvest_index<-yield/height
> harvest_index
 [1] 2.545946 2.532432 6.057895 2.815625 2.763636
3.674074 3.212500 2.448718
 [9] 2.422222 4.407407
>
```

Exporting data and results

There are two simple methods of doing this. The simplest is to cut and paste from R to Word or Excel, exactly as you would for any other Windows application.

The second is to click on File, then select Save to File, then enter a file name – again as under most Windows applications. The complete set of commands and their responses is then available for subsequent editing and processing in the software of your choice.

5.3.5 Summary statistics

Time spent in reconnaissance is never wasted' – Napoleon (attrib).

It is always worthwhile to spend time scanning and summarising new datasets before starting formal statistical analysis. Simple methods such as studying the range, the distribution and the relationships between variables can often reveal unexpected structure or the presence of errors in a dataset.

Summaries of single variates

In R, the simplest method of generating a summary of data is:

```
> summary(data1)
    yield            height          SNP1
SNP2
 Min.   : 77.10   Min.   :19.00   Min.   :1.00   Min.
:1.000
 1st Qu.: 90.38   1st Qu.:27.00   1st Qu.:1.00   1st
Qu.:1.000
 Median : 93.95   Median :32.50   Median :1.00   Median
:2.000
 Mean   : 96.23   Mean   :31.10   Mean   :1.30   Mean
:1.556
 3rd Qu.: 98.28   3rd Qu.:36.75   3rd Qu.:1.75   3rd
Qu.:2.000
 Max.   :119.00   Max.   :39.00   Max.   :2.00   Max.
:2.000
                                                NA's
:1.000
>
```

The output shows:

```
Min.   :          the minimum value
1st Qu.:          the first quantile
Median :          the median
Mean   :          the sample average
3rd Qu.:          the third quantile
Max.   :          the maximum value.
```

The first quantile, the median and the third quantile give the values of the observations, ¼, ½, and ¾ of the way down a sorted list of each variable. These values, together with the minimum, maximum and average, give a simple assessment of the distribution of the traits. Of course, for the SNP data – SNP1 and SNP2 - this summary is a bit pointless, although the mean-1 is the allele frequency. The results for just a single variable can be given by

```
> summary(height)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 19.00  27.00  32.50  31.10  36.75  39.00
>
```

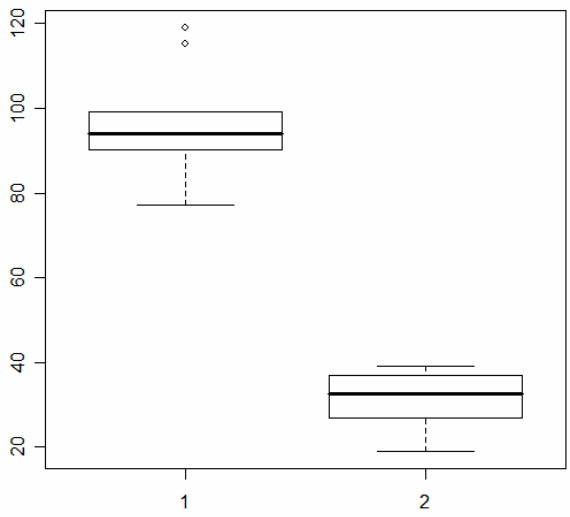
or for a selected set of variables by combing summary with subset:

```
> summary(subset(data1,select=c(yield,height)))
    yield           height
 Min.   : 77.10   Min.   :19.00
 1st Qu.: 90.38   1st Qu.:27.00
 Median : 93.95   Median :32.50
 Mean   : 96.23   Mean   :31.10
 3rd Qu.: 98.28   3rd Qu.:36.75
 Max.   :119.00   Max.   :39.00
>
```

A graphical equivalent of `summary` is the box plot, or box-and-whisker plot:

```
> boxplot(yield.2,height)
>
```

This command does not return a result in the R console. It opens up a graphics window displaying the graph shown below. We shall see later how to add titles to this and to other graphs, and how to save graphs as files.



The central bold line is the mean. The boxes show, approximately, the first and third quantiles. The lines extending from the boxes to the horizontal bars then show the distance to the maximum and minimum observations. However, any data viewed as outliers are plotted separately – as has happened here for yield. These plots must not be used a statistical test for outliers: in the example here there are far too little data (10 observations), to declare that two observations are aberrant.

The summary statistics given collectively by summary are also available as separate commands: listed below:

```
mean(x)
median(x)
quantile(x)
minimum(x)
maximum(x)
```

To run these commands, x should be substituted by the required variate name. If two variate names are given, the summary is over both variates:

```
> mean(height,yield)
[1] 32.5
>
```

Another quirk of R is shown below:

```
> mean(SNP2)
[1] NA
```

```
>
```

R does not necessarily ignore missing values. Because SNP2 has one missing value (remember the NA for the first data point), R does not believe it can therefore calculate an average. The rather irritating and longwinded way to cope with this is as follows:

```
> mean(SNP2,na.rm=T)
[1] 1.555556
>
```

Here, the additional logical variable na.rm (not available, remove) is set to the value T (for true): in English, remove the NA values before calculating the mean. A number of other commands also require this option. If a command returns NA, it is worth trying this option.

One dangerous exception to the default behaviour of R is the command length, which counts the number of entries in a vector:

```
> length(SNP2)
[1] 10
```

The value of 10 is returned: the count has included the missing value. The work around for this is shown below:

```
> length(SNP2)
[1] 10
> is.na(SNP2)
 [1]  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE FALSE
> !is.na(SNP2)
 [1] FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
TRUE  TRUE
> sum(!is.na(SNP2))
[1] 9
>
```

`is.na()` is the command to return the logicial value `TRUE` or `FALSE` depending on whether each value of a variate exists or not.

`!is.na()` switches this around to return TRUE if the value is NA (ie does not exist).

For arithmetic purposes, the logical `TRUE` has a value of 1 and `FALSE` has a value of 0, `sum()` returns the sum or total, so `sum(!is.na(SNP2))` returns the value of 9.

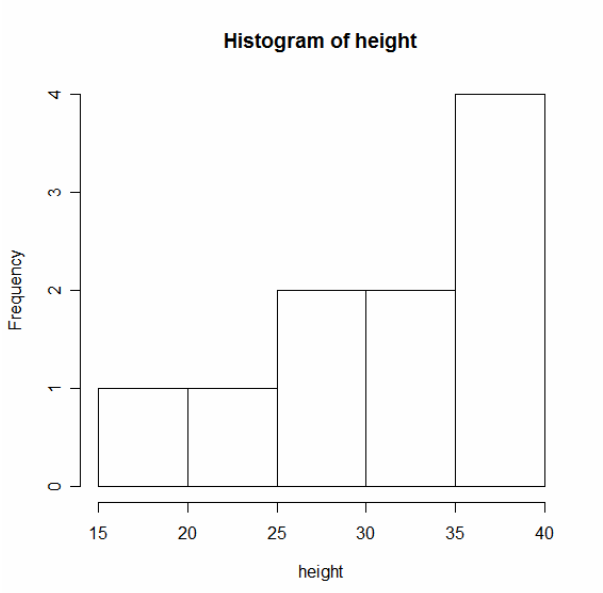
Some other useful summary commands are listed below:

`sum(x,na.rm=T)`	returns the total
`var(x,na,rm=T)`	returns the variance
`sd(x,na.rm=T)`	returns the standard deviation.

These commands have all been listed including the na.rm=T option. If the dataset is complete this option need not be included.

A simple way of viewing the distribution of a variable, and perhaps more informative than any set of summary statistics is to plot it:

```
> hist(height)
>
```



sorting data

Sorting of data and inspection of high and low values is also of great assistance in detecting erroneous data. In R this is carried out using the command `sort`:

```
> sort(yield)
 [1]  77.1  87.2  90.1  91.2  93.7  94.2  95.5  99.2
115.1 119.0
>
```

More generally, it is usual to sort a bock of data with respect to one or more columns. Routinely, this may be more easily achieved in Excel. In R, the command:

```
> yield[order(SNP1,SNP2)]
```

```
 [1]  93.7  90.1  91.2  99.2  77.1  87.2  94.2 115.1
95.5 119.0
>
```

gives a sorted list of yields. This is explained below:

First:

```
> order(SNP1,SNP2)
 [1]  2  4  5  6  7  9  1  3  8 10
```

returns the order of the data after sorting first by `SNP1` then by `SNP2`. Thus records two and four have the lowest ranking, corresponding to the genotypes 1,1. Records eight and ten come last, with the genotype 2,2. Missing data (NA) is ordered after data which exists, so record one (1,NA) is ranked after record nine (1,2) but before record three (2,1).

Additional syntax is also introduced with the square brackets [ ]. These are used to reference or index specific elements of a variable or factor. Thus

```
> yield[5]
[1] 91.2
```

returns the fifth entry of yield, while

```
> yield[c(5,9)]
[1] 91.2 87.2
```

returns the fifth and ninth elements. So

```
> yield[order(SNP1,SNP2)]
 [1]  93.7  90.1  91.2  99.2  77.1  87.2  94.2 115.1
95.5 119.0
```

returns `yield` in order specified by `order(SNP1,SNP2).`

Relationships between variates.

The correlation coefficient ranges from zero to one and measures the strength of the relationship between two variables:

```
> cor(yield,height)
[1] -0.3913837
```

We can enter the data frame name and generate a table of correlation coefficients in a single command. There are some traps for the unwary however:

```
> cor(data1)
Error in cor(data1) : missing observations in cov/cor
```

```
> cor(data1,use="complete.obs")
             yield      height       SNP1        SNP2
yield   1.00000000 -0.39403716  0.7602782 -0.07668775
height -0.39403716  1.00000000 -0.2386212  0.02780055
SNP1    0.76027819 -0.23862117  1.0000000  0.15811388
SNP2   -0.07668775  0.02780055  0.1581139  1.00000000


> cor(data1,use="pairwise.complete.obs")
             yield      height       SNP1        SNP2
yield   1.00000000 -0.39138365  0.7532037 -0.07668775
height -0.39138365  1.00000000 -0.2896914  0.02780055
SNP1    0.75320367 -0.28969140  1.0000000  0.15811388
SNP2   -0.07668775  0.02780055  0.1581139  1.00000000
>
```

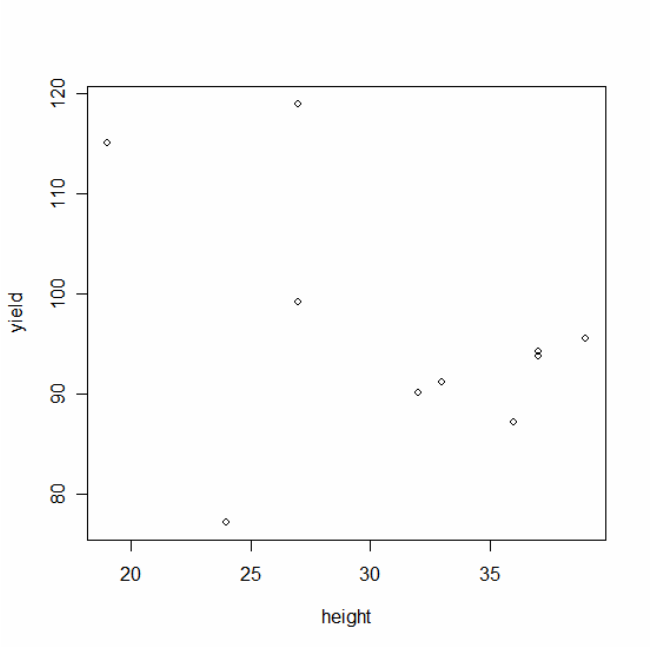
The first attempt fails because we failed to take into account the missing observation in SNP2. The second method uses "complete observations", that is to say only records with no missing data for any field. This is the default method for some commercial statistical software systems. In the example here, the first record of data is discarded. The third method does not discard complete records. It excludes from the analysis only those pairs of observations in which at least one of the pair is NA. In this example, correlations among yield, weight and SNP1 will be based on 10 paired observations, while correlations involving SNP2 will be based on 9. This option can be particularly useful with extensive sets of genotype data: even if marker calling rates are high, with multiple markers it may be rare for a single individual or line to have no missing data.

In passing, we note that the calculation of a correlation coefficient is not a particularly sensible or conventional way to study the relationship between two binary variables such as SNP1 and SNP2. More conventional would be to tabulate the data in a 2 x 2 contingency table. However, for SNP data, the squared correlation coefficient is one of the standardised measures of linkage disequilibrium between two loci, so the example given here has some justification (although a 2 x 2 table of observations would still be more informative and is demonstrated later on).

Correlation coefficients are a simple way of quantifying relationships between two variables. However, it is often better to visualise the data in a scatter plot:

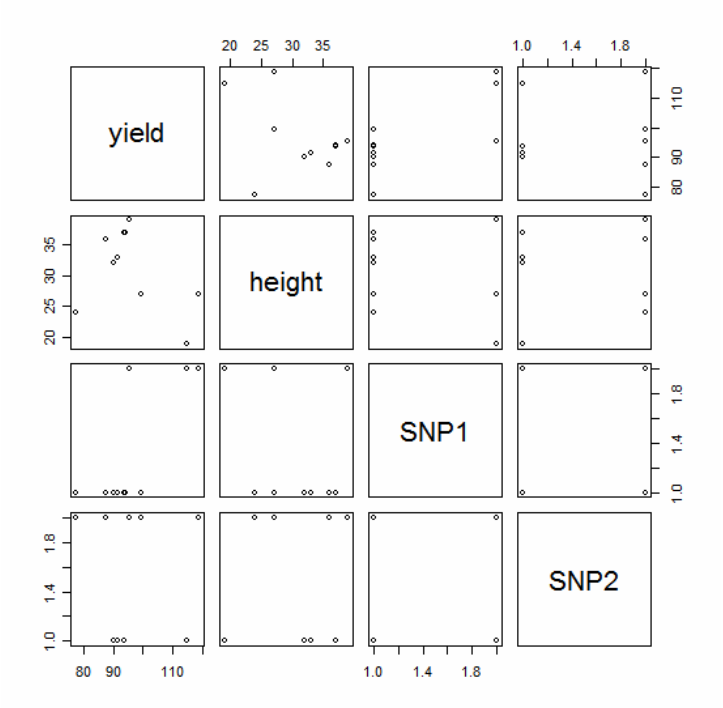
```
> plot(height,yield)
```

```
>
```



A particularly useful extension to this, which makes one forgive many of the quirks of R is:

```
> pairs(data1)
>
```



In the example shown here, only the plots for yield and height are of any practical use, but the ability to generate multiple scatter plots like

this in a single command is of great use in surveying patterns across large multivariate datasets. Trying doing this in Excel with 10 variables (requiring 45 plots).

5.3.6 Basic statistical analysis

5.3.6.1 The t-test

The t-test is a simple and robust method to test if the difference in means between two samples, or the difference between the mean of a sample and a known constant, is statistically significant. In other words, does the difference look too large to have occurred as a result of bad luck in selecting the samples for analysis.

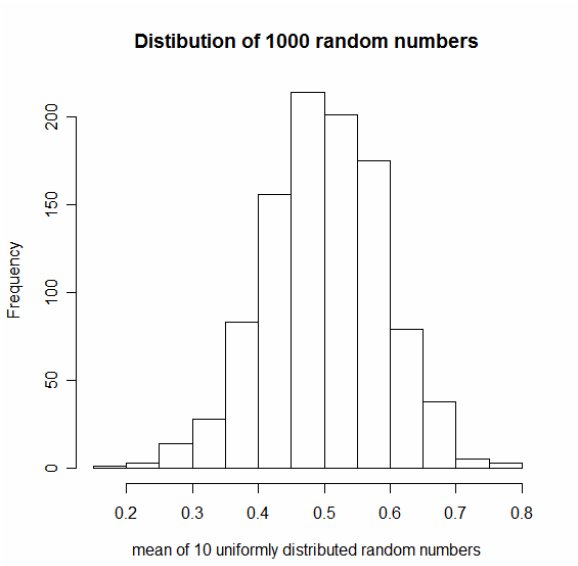
The test statistic is

$$t = \frac{\text{difference}}{\text{standard error of difference}}$$

For large sample sizes, a value of $t > 2$ will only occur by chance in about 5% of experiment. A value > 2 is therefore judged to be improbably large: the difference in means is declared to be statistically significant at the 5% level. In practice, R automatically calculates this probability more precisely.

The t-test assumes that the sampling error of the difference being tested is normally distributed. In real data sets, this condition is often met. Firstly, the trait being measured is itself often normally distributed, and secondly, even if the trait is has a non-normal distribution, mean trait values will follow close to normal distributions provided the sample size is moderately large (greater than about 10).

For example, the plot below shows the distribution of 1000 numbers. Each number was generated by taking the mean of 10 uniformly distributed random numbers. It is clear that although the original random numbers were very non-normal, the mean of a sample of 10 such numbers is pretty close to normal. In fact, in the early days of computing, normally distributed random numbers were often generated in this way. The tendency for the distribution of means to be normally distributed is called the Central Limit Theorem. It explains the popularity of the normal distribution in statistics and also the tendency for many traits in nature to be roughly normally distributed – for example if variation in a phenotype results from variation at multiple genes, the phenotype itself will often inevitably be normally distributed.



This histogram, and the random numbers from which it was built were generated in R, using the following two commands, which are presented below without explanation, but which show how much R can achieve with only a few commands.

```
> m1<-matrix(runif(10000),1000)
> hist(apply(m1,1,mean),
+ main="Distibution of mean of 10 random numbers",
+ xlab="mean of 10")
```

The t-test is very simply invoked in R. To test the difference in means between yield and height:

```
> t.test(yield,height)

        Welch Two Sample t-test

data:  yield and height
t = 14.5809, df = 13.649, p-value = 1.037e-09
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 55.5265 74.7335
sample estimates:
mean of x mean of y
    96.23     31.10

>
```

The output provides a value for t, a p-value to test the significance of the difference in means, and the means themselves. In addition. 95% confidence intervals are provided. These refer to the difference

between the two means. Statisticians can get quite hot under the collar about what, exactly, 95% confidence intervals actually are. We can state that over a long lifetime of calculating 95 % confidence intervals for parameter estimates, they will have included the true parameter value in 95 % of cases. It is best not to worry too much about this.

Note that the degrees of freedom (df) is 13.649 and not, as is usual a whole number. This is because the default setting for R is to assume that, whether or not the means of the two groups being tested are different, the variances themselves are different. In accounting for this we end up with fractional degrees of freedom. This is the Welch variant of the t-test – stated in the first line of the output.

To test whether the variances in the two groups are similar, we can use a variance ratio test, or F test – dividing one variance by the other and estimating whether the deviation from the expected value of 1 is attributable to chance or is indicative of something else,

```
> var.test(yield,height)

        F test to compare two variances

data:  yield and height
F = 3.5938, num df = 9, denom df = 9, p-value = 0.07035
alternative hypothesis: true ratio of variances is not
equal to 1
95 percent confidence interval:
  0.8926393 14.4684590
sample estimates:
ratio of variances
          3.593761

>
```

The F-ratio of 3.59, with 9 and 9 degrees of freedom is not significant (p-value 0.07). Note that if `var.test(height,yield)` was called, The F-ratio would be 0.278 (1/3.59) but the p-value would be unchanged.

Since the variances are not significantly different (ie they are homogeneous), t.test can be called in a form to take this into account:

```
> t.test(yield,height,var.equal=T)

        Two Sample t-test

data:  yield and height
```

```
t = 14.5809, df = 18, p-value = 2.069e-11
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 55.74562 74.51438
sample estimates:
mean of x mean of y
    96.23     31.10

>
```

This is the more usual form for the t test – as given in most text books. It is a more powerful test than Welch's variant, provided the variances are homogeneous. Note that the degrees of freedom is now integral.

The examples so far have been comparing yield with height. This has been for illustrative purposes only: not only does it make no biological sense, it is on shaky ground statistically too – the measurements are paired in the sense that each variety has been measured for both height and yield. There is therefore a chance that the measurements are correlated – growing conditions for one variety will affect both traits. A simple method to deal with this is analyse the differences between yield and weight for each individual. The mean difference is expected to be zero. The estimated mean difference and its standard error can be calculated and used to construct a t-test. This procedure is automated within R:

```
> t.test(height,yield,paired=T)

        Paired t-test

data:  height and yield
t = -12.6765, df = 9, p-value = 4.82e-07
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 -76.75259 -53.50741
sample estimates:
mean of the differences
                 -65.13
```

In this t-test for paired data, the test itself does not directly compare two means. An estimated mean is compared with an expected value (zero in this case) which is known without error. This is termed a one-sample t-test rather than the more typical two-sample test. In R it can be called explicitly by supplying a value for the known constant rather than the name of a second variable:

```
> t.test(yield,mu=31.10)

        One Sample t-test

data:  yield
t = 16.4852, df = 9, p-value = 4.956e-08
alternative hypothesis: true mean is not equal to 31.1
95 percent confidence interval:
  87.29265 105.16735
sample estimates:
mean of x
    96.23
```

The value of t has risen. This is because, although the difference between the values being tested has not changed, the standard error of the difference is less: the only source of error variation is from yield, we are falsely ignoring the error in the estimation of mean height.

A more biologically interesting comparison is whether the SNPs have a direct effect on yield. The problem here is that data for the two groups to be compared (yields for genotype 1 and yields for genotype 2) are no longer in separate variables. To tell R that data to be analysed are in one variate but are described by data in another, we use the tilde operator (~) introduced in the section on syntax:

```
> t.test(yield~SNP1,var.equal=T)

        Two Sample t-test

data:  yield by SNP1
t = -3.2387, df = 8, p-value = 0.0119
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 -33.351652  -5.610252
sample estimates:
mean in group 1 mean in group 2
       90.38571       109.86667

>
```

In spite of only three observations for genotype 2, the difference in means significant. SNP1 could lie in or close to a QTL for yield, or the significant result could be due to something we don't know about concerning the origins of the varieties under test. Disentangling trait-marker associations due to the presence of a closely linked QTL from other spurious causes of association is the challenge of association genetics.

For completeness, we note that variances are homogeneous within the two genotype results, and that SNP2 showed no significant association.

5.3.6.2 Linear Regression

We have already come across the command to correlate two traits: cor(). To fit a straight line to a data set we use the R command lm() – for linear model. Suppose we want to study the effect of height on yield:

```
> lm(yield~height)

Call:
lm(formula = yield ~ height)

Coefficients:
(Intercept)       height
    119.305       -0.742

>
```

The output is somewhat sparse. A feature of R, in contrast to many statistical packages is that by default it does not treat you to multiple pages of output from which you may only wish to extract a single figure. Here the output gives you the best fitting straight line: Yield = 119.305 – (0.742 x height). More output is available but we need to be explicit that we wish R to produce it. First we shall rerun the analysis, but save the results:

```
> yield_height_regression<-lm(yield~height)
>
```

No output is generated. We can produce some using summary():

```
> summary(yield_height_regression)

Call:
lm(formula = yield ~ height)

Residuals:
     Min       1Q   Median       3Q      Max
-24.3979  -4.9509   0.8878   4.4355  19.7280

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 119.3048    19.5654   6.098  0.00029 ***
```

```
height       -0.7420     0.6168  -1.203  0.26338
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1

Residual standard error: 12.19 on 8 degrees of freedom
Multiple R-Squared: 0.1532,     Adjusted R-squared:
0.04733
F-statistic: 1.447 on 1 and 8 DF,  p-value: 0.2634

>
```

The most interesting part of the output is given at the end: the F-statistic and p-value for the significance of the regression – not significant in this example. The Multiple R-Squared is the proportion of the total sum of squares accounted for by the regression. It is also the square of the correlation coefficient between yield and height. The Adjusted R-squared is the proportion reduction in variance after fitting the regression. Both these figures give an indication of how effective the regression has been in accounting for the observed variation: a significant regression does not imply that a relationship is particularly important. Equally, with very small experiments, large proportions of variation may be accounted for, but the regression is still non-significant. This is generally an indication that you should have designed a larger experiment.

A more conventional display of the regression analysis is given as:

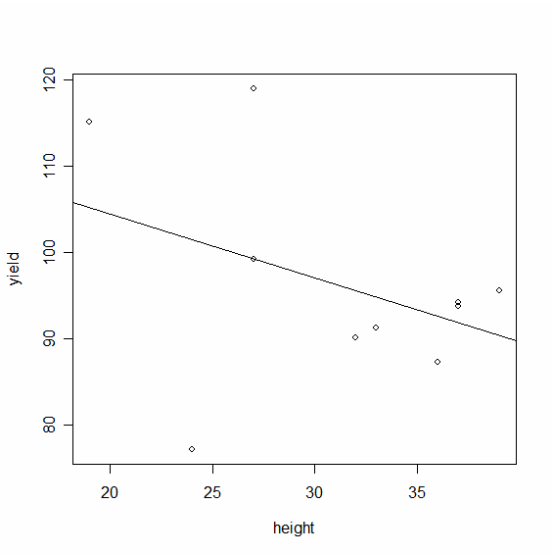
```
> anova(yield_height_regression)
Analysis of Variance Table

Response: yield
          Df  Sum Sq Mean Sq F value Pr(>F)
height     1  215.19  215.19  1.4471 0.2634
Residuals  8 1189.61  148.70
>
```

This uses the `anova()` command. To R, anova is the name given to a form of tabular output. Formally, the analysis of variance itself is just a particular type of multiple regression analysis, and that is exactly how R treats it, as we shall see shortly.

We can look at a plot of the data with our fitted line a follows:

```
> plot(height,yield)
> abline(yield_height_regression)
```



When using plot, give the name of the variate you want plotted on the x axis first. The additional command, `abline()`, adds the best fitting straight line. The fit is poor: it is not surprising the result is non-significant.

Finally, we look at how to extract residual values and fitted values from a regression. Large residual values for particular observations are often of use in searching for errors in data. Also, identification of the individuals or varieties responsible for large residuals may sometimes suggest some other factor which needs to be considered in the analysis. Fitted values and residuals are extracted as:

```
> fitted(yield_height_regression)
        1         2         3         4         5         6         7
8
 91.85247  91.85247 105.20765  95.56224  94.82029  99.27201 101.49788
90.36856
        9        10
 92.59442  99.27201


> resid(yield_height_regression)
          1           2           3           4           5
6
  2.3475313   1.8475313   9.8923510  -5.4622410  -3.6202865  -
0.0720133
          7           8           9          10
-24.3978767   5.1314403  -5.3944231  19.7279867
>
```

These could be saved to another variable, or plotted, in the usual manner. A plot of residuals against fitted values is often informative. If large residuals tend to be associated with large fitted values, for example, this indicates that the error variances are not homogenous and we treat our results with more caution. Residuals increasing with

the fitted values is often an indication that transforming the data to logarithms before analysis may be warranted.

5.3.6.3 Multiple regression

Multiple regression in R requires little more than simple linear regression. The `lm()` command is still used. All that is required is to specify a more complex model using the syntax described in the *Basic R Syntax* section of this guide:

```
> lm(yield~height+SNP2+SNP1)

Call:
lm(formula = yield ~ height + SNP2 + SNP1)

Coefficients:
(Intercept)       height         SNP2         SNP1
    90.5966      -0.4235      -4.7109      19.5607
```

Thus yield is predicted as:

90.59 -(0.42x height) –(4.72xSNP2) +(19.56xSNP1)

The effect of SNP1 is large compared to the mean.

Note this form of analysis if acceptable for SNPs or other binary markers provided the markers are coded numerically. 0 and 1 are ideal codes, but 0 can sometimes be confused with a missing value, especially if data are to be analysed in packages other than R. 1, 2 coding is also acceptable. Numeric coding with differences between alleles greater than one can make the analysis difficult to interpret and should be avoided. Numeric coding cannot be used at all for multiallelic markers, or for haplotype analyses, since it implies that alleles coded with a higher number are worth more than those with a lower number. We shall see how to account for this problem shortly.

Note the order in which the variates are supplied to `lm` does not affect the estimates.

```
> (lm(yield~SNP1+height+SNP2))

Call:
lm(formula = yield ~ SNP1 + height + SNP2)

Coefficients:
(Intercept)         SNP1       height         SNP2
    90.5966      19.5607      -0.4235      -4.7109
```

However, this is not the case for estimates of significance:

```
> anova(lm(yield~height+SNP1+SNP2))
Analysis of Variance Table

Response: yield
          Df Sum Sq Mean Sq F value  Pr(>F)
height     1 217.41  217.41  2.2842 0.19109
SNP1       1 659.08  659.08  6.9248 0.04645 *
SNP2       1  47.86   47.86  0.5028 0.50994
Residuals  5 475.88   95.18
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1


> anova(lm(yield~SNP1+height+SNP2))
Analysis of Variance Table

Response: yield
          Df Sum Sq Mean Sq F value  Pr(>F)
SNP1       1 809.36  809.36  8.5038 0.03316 *
height     1  67.12   67.12  0.7052 0.43933
SNP2       1  47.86   47.86  0.5028 0.50994
Residuals  5 475.88   95.18
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
>
```

In the first anova table, the sum of squares for SNP1 is 659.08. In the second analysis, the sum of squares for SNP1 is 809.36. The sum of squares for SNP2 and the residual sum of squares is identical in both analyses. In fact, the sum of squares for (SNP1 + height) is also identical in the two analyses. In unbalanced designs, such as this, where combinations of SNP1, SNP2 and height are not all equally represented, the results from the analysis of variance depend on the order in which the terms are represented in the model. However, there is an easy way to interpret this table. Taking output from the last analysis a line at a time:

```
> anova(lm(yield~SNP1+height+SNP2))
```

The effect for SNP1 is fitted first, and is found to be statistically significant:

```
          Df Sum Sq Mean Sq F value  Pr(>F)
SNP1       1 809.36  809.36  8.5038 0.03316 *
```

After fitting the SNP1 effect, height is fitted next:

```
height     1  67.12   67.12  0.7052 0.43933
```

The p-value of 0.44 for height is the significance for height, *after* accounting for any effect of SNP1. Finally, the effect of SNP2, after fitting both SNP1 and height effects is non-significant:

```
SNP2       1  47.86   47.86  0.5028 0.50994
```

This explains why the p-value for SNP2 is identical in both analyses. In both cases it is assessed the first fitting SNP1 and height effects. Whether these two effects are fitted first as height and then as SNP1 or vice-versa makes no difference: the total variation the two effects account for remains the same.

The significance of SNP1 from the first analysis:

```
          Df Sum Sq Mean Sq F value  Pr(>F)
SNP1       1 659.08  659.08  6.9248 0.04645 *
```

is testing the significance of SNP1 *after* accounting for variation in height, and this is slightly different to the significance assessed in the second analysis where SNP1 is the first term to be fitted.

In designed experiments – where different combinations of treatments and factors are usually equally represented, or balanced, the order in which terms are fitted usually makes no difference – the terms are said to be orthogonal. Balance not only has the property of making the terms orthogonal, it also makes the arithmetic very much easier. This was very important before the advent of readily available computers. However, the requirement for balance, solely from the point of view of data analysis, is now no longer required and many contemporary experimental designs (for example alpha-designs) are not balanced and would be impossible to analyse without a computer. A readable account of a contemporary approach to experimental design is given in Mead: “The Design of Experiments.”

In this example, in whatever order the effects are fitted, SNP1 is the only term to be significant. Selecting the order in which terms are fitted, and selecting which terms to include in the final model and which to excluded is something of an art, which we shall not develop here. There are formal methods to assist in this process. These too are not covered here, but are available within R. Generally, with genetic analysis, it is usual to account for variation attributable to causes other than genes or markers first, and then fit the genetic effects. In the example here, where the evidence suggests that only SNP1 is likely to be a genuine effect, it makes sense to test this by fitting the SNP1 effect last – after all other explanations of variation in yield have had their shot:

```
> anova(lm(yield~height+SNP2+ SNP1))
Analysis of Variance Table

Response: yield
          Df Sum Sq Mean Sq F value  Pr(>F)
height     1 217.41  217.41  2.2842 0.19109
SNP2       1   6.05    6.05  0.0636 0.81091
SNP1       1 700.88  700.88  7.3640 0.04209 *
Residuals  5 475.88   95.18
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
```

SNP1 is still statistically significant. Non-significant terms can be dropped from the model to leave:

```
 > anova(lm(yield~SNP1))
Analysis of Variance Table

Response: yield
          Df Sum Sq Mean Sq F value Pr(>F)
SNP1       1 796.97  796.97  10.489 0.0119 *
Residuals  8 607.84   75.98
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
```

Two things are of note from this analysis. Firstly, the p-value is identical to that we saw earlier when testing the effect of SNP1 on yield in a t-test. For a regression analysis on a single variate with only two values or classes, the two tests are equivalent. In fact

$t^2 = F$

In this case

-3.2387 ^2 = 10.49818.

The second item of note is that the total degrees of freedom in this final analysis (9) is one more than from the analyses containing all parameters (8). This is because of the missing value in SNP2. For any model which includes SNP2, all data for the variety containing the missing value is excluded. There is no easy way around this. It is of concern only if we are comparing models containing SNP2 to those without. In this case, because there is no hint that anything to do with SNP2 is significant, the effect on the analysis is minimal.

Note that `resid()`, `fitted()` and `summary()` work for multiple regression exactly as for simple linear regression.

5.3.6.4 The analysis of variance

In the t-test, we test if the difference between two treatment means is statistically significant. This is a special case of the Analysis of Variance in which we test if differences among multiple treatments are jointly statistically significant. We could analyse multiple treatments by carrying out multiple t tests, but with many tests, there is an increased risk that at least one test will be declared significant by chance alone – the so called problem of multiple testing. In addition the interpretation of results becomes increasingly complex. (There is, in fact an R command that automates this procedure and includes an adjustment for multiple testing: `pairwise.t.test`). The omnibus test for significance of all means, considered together, that the Analysis of Variance offers is therefore of great value.

The principal of the Analysis of Variance is that, in the absence of any genuine difference among means, the variability among those means can be predicted from the variability from observation to observation within each treatment. This argument is little more than saying that the variance of a mean is just the variance among the observations that contribute to that mean divided by the number of observations contributing to the mean:

$$V\bar{x} = Vx / n$$

The test for statistically significant differences among the means is therefore a variance ratio, or F test: the variance among treatment means is divided by the expected variance calculated within treatments. This is an oversimplification: differences in the number of observations within treatments must also be taken into account and with more complicated experimental designs the analysis is also more complicated, but the basic principal remains the same.

To illustrate the analysis of variance in R, we shall test whether there are statistically significant differences among haplotype frequencies in our small set of test data. Haplotypes for the two SNPs in our sample are given below.

```
Variety   SNP1 SNP2 haplotype
stata     1    NA   NA
sass      1    1    11
R         2    1    21
genstat   1    1    11
Splus     1    1    11
S         1    2    12
SPSS      1    2    12
minitab   2    2    22
BMDP      1    2    12
mstat     2    2    22
```

Haplotypes can be coded in various ways. The one adopted above is quite common for small numbers of SNPs. For haplotypes involving large numbers of SNPs, an interesting alternative is to code SNP alleles as 0 and 1, so that the haplotype is a binary number: 10010001 for example. This can be converted to a normal, base 10 number – 1+16+128 = 145 for 10010001. To avoid having haplotypes coded as 0, 1 is added to all numbers, so in the example the haplotype would be coded as 146. This has the advantage of taking up less space while retaining all the information about individual SNP alleles.

To generate our haplotype numbers, we could proceed as:

```
haplotype <- (SNP1*10+SNP2)
```

However, the haplotype would then be a number ranging from 11 to 22. The analysis would then fit a linear regression of yield on haplotype, which isn't what we want. We need to fix haplotype as a categorical variable or a factor.

```
> haplotype<-factor((SNP1*10+SNP2))
> haplotype
 [1] <NA> 11   21   11   11   12   12   22   12   22
Levels: 11 12 21 22
>
```

Note that the haplotype for the first entry is NA – since we don't know what SNP2 is for this variety, we cannot know what the haplotype is. Some care needs to be exercised in coding haplotypes by combining two fields in this manner. For example, combing the data as (SNP1+SNP2) will seemingly work but will generate a factor with three levels with the haplotypes "12" and "21". This warning applies equally

whether coding is carried out here or in some other package such as Excel.

The basic analysis of variance is now very straightforward:

```
> anova(lm(yield~haplotype))
Analysis of Variance Table

Response: yield
          Df Sum Sq Mean Sq F value Pr(>F)
haplotype  3 872.48  290.83  2.7554 0.1517
Residuals  5 527.74  105.55
>
```

In spite of the significant effect of SNP1 on yield that we detected earlier, there is no significant haplotype effect. Note that this does not mean that tests for association using haplotypes should be dismissed as having less power than single locus effects. Firstly, there are instances where the association directly attributable to ahaplotype (perhaps in linkage disequilibrium with another, undetected or ungenotyped causative genetic variant). Secondly, when testing many single SNPs for association, the risk of detecting at least one significant result by chance is increased – the problem of multiple testing. To overcome this, it is usual to increase the stringency required to declare statistical significance at any one test. One commonly used method is the Bonferroni correction: the p-value for significance is reduced from 0.05 to 0.05/(the number of independent tests). Methods for correcting for multiple testing and methods for analysing haplotypes or sets of multiple closely linked SNPs are active research topics.

To get more information from our analysis of variance, we can use all the methods introduced earlier for linear and multiple regressions: as stated before, the analysis of variance is just a special case of multiple regression. For example:

```
> summary(lm(yield~haplotype))

Call:
lm(formula = yield ~ haplotype)

Residuals:
         2          3          4          5          6
7          8          9         10
 2.033e+00 -6.203e-16 -1.567e+00 -4.667e-01  1.137e+01 -
1.073e+01 -1.175e+01 -6.333e-01  1.175e+01

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)   91.667      5.931  15.454 2.06e-05 ***
haplotype12   -3.833      8.388  -0.457    0.667
haplotype21   23.433     11.863   1.975    0.105
haplotype22   15.583      9.379   1.662    0.157
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1

Residual standard error: 10.27 on 5 degrees of freedom
Multiple R-Squared: 0.6231,     Adjusted R-squared: 0.397
F-statistic: 2.755 on 3 and 5 DF,  p-value: 0.1517

>
```

Here, the estimate for the intercept is the mean of the first haplotype: 11. The estimates for haplotypes 12, 21 and 22 are then the difference between the effect for 11 and the effect for the other three haplotypes. The mean effect for haplotype 12 is 91.667 – 3.833 = 87.834. This is identical to the simple mean of the three data entries with this haplotype, as it should be in this case.

We complete our survey of the Analysis of Variance with an example from a two-way analysis of variance:

```
> anova(lm(yield~factor(SNP1)*factor(SNP2)))
Analysis of Variance Table

Response: yield
                          Df Sum Sq Mean Sq F value
Pr(>F)
factor(SNP1)               1 809.36  809.36  7.6682
0.03940 *
factor(SNP2)               1  55.68   55.68  0.5275
0.50022
factor(SNP1):factor(SNP2)  1   7.45    7.45  0.0705
0.80114
Residuals                  5 527.74  105.55
```

No new R commands are required to carry out this analysis. The test for an interaction between SNP1 and SNP2 is essentially a test for a haplotype effect over and above any single locus effects detected by fitting the SNP1 and SNP2 main effects. This is a sensible analysis to carry out (apart from the ridiculously small size of the data set). However, with very large numbers of SNPs, there are a large number of pairwise interactions (300 with 25 SNPs for example), let alone any higher order interactions, and a direct analysis of haplotypes may be a better approach.

5.3.6.5 Categorical data – the chi-squared test

Suppose we wish to test if there is an association between SNP1 and SNP2. Such an association might arise because of the way the varieties included in the experiment have been selected, or because SNP1 and SNP2 are in linkage disequilibrium – they are so closely linked that insufficient meioses have occurred between the two to remove the original association generated when the SNPs were formed by mutation.
The standard method of analysis of such data is the contingency chi-squared test.

First we need to format our data into a table:

```
> table(SNP1,SNP2)
    SNP2
SNP1 1 2
   1 3 3
   2 1 2
```

The row and column headings here are confusing, so we shall change them:

```
> SNPtable <- table(SNP1,SNP2)
> rownames(SNPtable) <-c("allele1","allele2")
> colnames(SNPtable) <-c("allele1","allele2")
> SNPtable
         SNP2
SNP1      allele1 allele2
  allele1       3       3
  allele2       1       2
>
```

Note the data could also have been summarised by haplotype as:

```
> table(haplotype)
haplotype
11 12 21 22
 3  3  1  2
>
```

From this table, we can see that the allele frequency of allele 1 at SNP1 is 6/ 9 or 0.6667 and that the allele frequency of allele 1 at SNP2 is 4/9 or 0.4444. If these SNPs are behaving independently of each other, the allele carried by a variety at SNP1 will be independent of the allele carried by the same individual at SNP2. In this case, the predicted frequency of SNP1 allele1 , SNP2 allele 1 individuals (ie the frequency of 11 haplotypes) will be 0.4444 x 0.6667 or 0.2963. The predicted number of 11 haplotypes will be 0.2963 x 9 or 2.6667. This

same exercise can be carried out for each of the other three haplotype classes. If deviations between observed and expected numbers are sufficiently large, we draw the conclusion that genotypes at SNP1 and SNP2 are not independent of each other. The statistical test is a chi-squared with 1 degree of freedom, calculated as

$$\sum \frac{(O-E)^2}{E}$$

O represents the observed numbers and E the expected.

This chi-squared test is simply carried out as :

```
> chisq.test(table(SNP1,SNP2))

        Pearson's Chi-squared test with Yates' continuity
correction

data:  table(SNP1, SNP2)
X-squared = 0.0563, df = 1, p-value = 0.8125

Warning message:
Chi-squared approximation may be incorrect in:
chisq.test(table(SNP1, SNP2))
>
```

The results is not significant – p-value of 0.8125. The warning message is given because some cells have expected counts less than 5. Under this threshold, there is a chance that the chi-squared test will give misleading results. Note it is the expected count that matters, not the observed - which can legitimately be zero and often is when we are considering closely linked SNPs. That is to say we may observe only three or possibly only two of the four possible haplotypes. We can extract and examine the expected values without the need for the hand calculations described above:

```
> chisq.test(SNPtable)$expected
         SNP2
SNP1       allele1  allele2
  allele1 2.666667 3.333333
  allele2 1.333333 1.666667
Warning message:
Chi-squared approximation may be incorrect in:
chisq.test(SNPtable)
```

Observed counts can be extracted in the same manner:

```
> chisq.test(SNPtable)$observed
         SNP2
SNP1      allele1 allele2
```

```
  allele1       3       3
  allele2       1       2
Warning message:
Chi-squared approximation may be incorrect in:
chisq.test(SNPtable)
```

The expected count in all cells is less than five. In the days when these tests were calculated by hand, there was an approximate correction, Yates's correction, which could be used to take into account the potential failure of the test statistic to follow a chi-squared distribution when the expected numbers were low. It is common now to derive the distribution of the test statistic empirically, by repeated randomisation or permutation of the observed data, followed by recalculation of test statistic. The proportion of times the randomised test statistic is greater than or equal to the observed test statistic is then the empirical p-value. Here the randomisation procedure is very simple: the data for SNP2 are randomised over subjects, while the data for SNP1 are held constant. For a hi-squared test, R automates this procedure:

```
> chisq.test(SNPtable,simulate=T,B=1000000)

        Pearson's Chi-squared test with simulated p-value
(based on 1e+06
        replicates)

data:  SNPtable
X-squared = 0.225, df = NA, p-value = 1
```

The value of B, the number of randomisations can be set by the user. The default value is 2000. Here, after 1 million randomisations, no empirical chi-sq was smaller than the observed value, so the empirical p-value is 1!

An alternative test to the chi squared for contingency tables with small expected numbers is Fisher's exact test. This compares the probability of observing the actual 2x2 table, calculated from the multinomial distribution, with the cumulated probability of observing other, less likely, tables. This probability is also calculated from a multinomial distribution with the same marginal frequencies of effects (here the overall allele frequencies at SNP1 and SNP2). In R:

```
> fisher.test(SNPtable)

        Fisher's Exact Test for Count Data

data:  SNPtable
p-value = 1
```

```
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
   0.06060903 156.52286969
sample estimates:
odds ratio
  1.852496

>
```

Again, the p-value is 1. The odds ratio is an approximation to relative risk, which has some nice statistical properties, and is much favoured in medical statistics and in epidemiology. It need not concern us.

The contingency chi-squared test will easily accommodate larger tables – a 10 x 10 table for example. In the context of genetic markers, this could be used to test for association between pairs of microsatellites with multiple alleles. Fisher's exact test is computationally hard to calculate, even on a computer, for large contingency chi-squared tables (eg a 10 x 10 table) so empirial methods are often favoured. The calculations for a contrived 3 x 2 table is shown below:

```
> (table(SNP1,haplotype))
    haplotype
SNP1 11 12 21 22
   1  3  3  0  0
   2  0  0  1  2
>

> chisq.test(table(SNP1,haplotype))

        Pearson's Chi-squared test

data:  table(SNP1, haplotype)
X-squared = 9, df = 3, p-value = 0.02929

Warning message:
Chi-squared approximation may be incorrect in:
chisq.test(table(SNP1, haplotype))
>

> chisq.test(table(SNP1,haplotype),simulate=T)

        Pearson's Chi-squared test with simulated p-value
(based on 2000
        replicates)
```

```
data:  table(SNP1, haplotype)
X-squared = 9, df = NA, p-value = 0.0004998


> fisher.test(table(SNP1,haplotype))

        Fisher's Exact Test for Count Data

data:  table(SNP1, haplotype)
p-value = 0.03571
alternative hypothesis: two.sided

>
```

The discrepancy in p-value between Fisher's exact test and the chi-squared test with empirical p-value requires explanation. A simulation (outside R) of 1,000,000 tables showed that the observed test statistic was equalled in 35477 cases, but was never exceeded. If p-value is defined as the number of times the observed test is exceeded, the p-value is zero. If p-value is defined, following convention, as the number of times the observed test statistic is equalled or exceeded, the p-value is 0.035 – in line with the result from Fisher's exact test. Thus the simulated p-value and the exact p-value use different definitions. This must be regarded as a bug rather than a feature. In practice, therefore, if results from chisq.test with and without the simulate option are greatly different, caution is advised. Use Fisher's exact test where possible. Generally, we would expect empirical p-values to be higher (less significant) than the parametric values.

5.4 Graphs

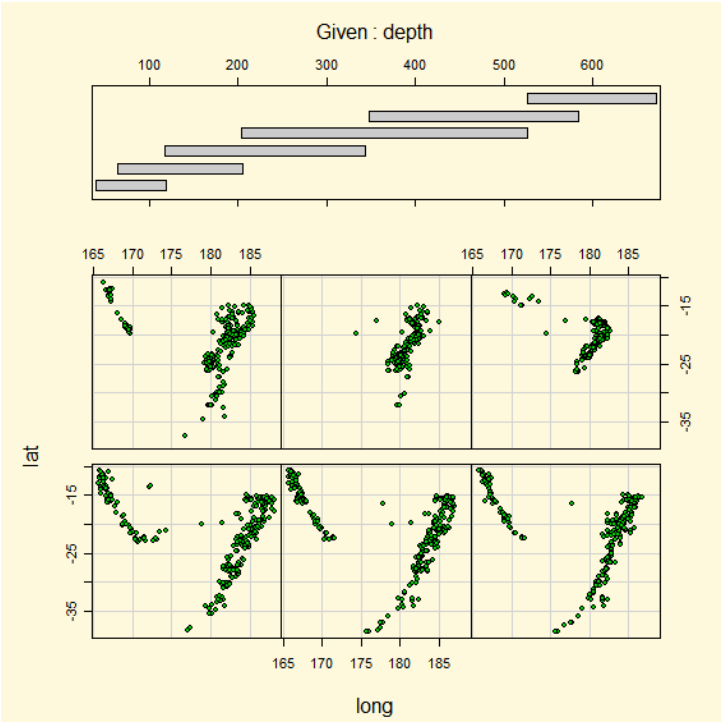
Graphical methods were introduced in context in the sections on summary statistics and on basic statistical analysis. The following methods have been used:

hist()	produces a histogram
plot()	produces a scatter graph
abline()	adds a line of best fit to a scatter graph
boxplot()	produces a box-and-whisker plot.
pairs()	produces a matix of scatter plots..

To show what R is capable of in skilled hands, type,

```
> demo(graphics)
```

The output below is just one of a series of examples. Methods for producing these plots are beyond the scope of this guide, however.



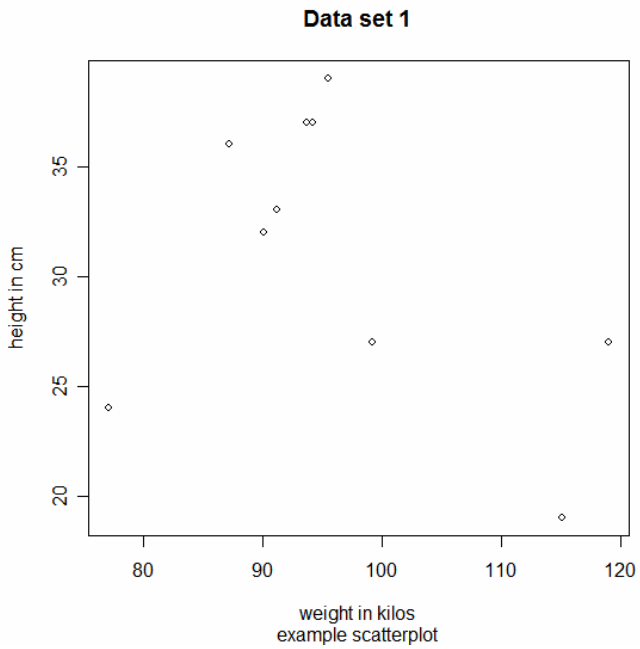
The simplest and quickest way to export graphs is to copy and paste. Click on the top of the graphical window to active it. Then copy the graph using control C (hold down the control key and then type C). Nothing will appear to happen. However, Control V will paste the graph into any appropriate Windows application: Word, Powerpoint or Paint for example. The graph can also be copied by selecting “File”, then “Copy to the clipboard”, then “as a bitmap” using the Windows menu system available at the top left of the R window. This same menu also allows you to save graphical files in pdf, jpeg or bitmap format.

Although R generates graphs very quickly and simply, the labelling and formatting are often not ideal. To alter graph titles, include the options:

main=”Main title”
sub=”subtitle”
xlab=”x-label”
ylab=”y-label”

For example

```
> plot(yield,height,main="Data set 1",sub="example
scatterplot",ylab="height in cm",xlab="weight in kilos")
>
```



5.5 Probability distributions

In this section we briefly describe how to use R to look up the p-value associated with the most commonly used test statistic and how to look up test statistics associated with p-values. This can also be done using functions in Excel. However, it may be convenient to do this from time to time in R. Also, although Excel is accurate over most of the range of possible p-values or test statistics, it is inaccurate at extreme values. Most of the time this doesn't matter. It can make a difference however, if a very large number of tests has been carried out and we are required to adjust for multiple testing. This occurs regularly in gene expression microarray experiments, for example.

Firstly the p-value associated with chi-squared.

```
> pchisq(3.84,1,lower.tail=F)
 [1] 0.05004352
```

The parameters given to `pchisq`, in order, are:

3.84 value of the test statistic.

1 the degrees of freedom

lower.tail If set to T (the default) the result is the cumulative distribution up to the value of the test statistic – 0.95 in the example. For significance testing we require the area of the upper tail: 1-0.95 and so set `lower.tail=F`.

To calculate a chi squared from a p-value we use the command `qchisq`:

```
> qchisq(0.05,1,lower.tail=F)
[1] 3.841459
```

The syntax is identical to that for `pchisq()`, so unfortunately we are required to include `lower.tail=F`.

Examples for the F distribution are shown below.

```
> pf(3.84,1,1000,lower.tail=F)
[1] 0.05032099

> qf(0.05,1,1000,lower.tail=F)
[1] 3.850775
```

The examples here are for 1 degree of freedom for the numerator and 1000 degrees of freedom for the denominator. The results are identical to those for a chi-squared test with 1 df. In fact, a chi-squared test with n degrees of freedom is identical to an F test with n degrees of freedom and a very large number of degrees of freedom (ideally infinite) in the denominator.

Values for probabilities associated with a normal distribution are :

```
> pnorm(1.96,lower.tail=F)
[1] 0.02499790
```

This probability is for a standardised normal distribution: with a mean of zero and a variance on 1. The probability is for a single tail of the distribution. Generally, we would require the result for a two tailed test – the probability of values higher than 1.96 and lower than -1.96. This probability is just double that for a single tail: 0.05 in this case. Again, this is the same value as for chi-squared with 1 degree of freedom. If a variate has a standardised normal distribution, the variate squared has a chi-squared distribution with 1 degree of freedom: $1.96^2 = -1.96^2 = 3.84$.

To derive the normal deviate associated with a specific probability:

```
> pnorm(1.96,lower.tail=F)
[1] 0.02499790
```

or

```
> pnorm(-1.96)
[1] 0.02499790
```

As ever, care is required to ensure specification of the correct tail.

Probabilities associated with normally distributed variables with different means and variances are produced by specifying the mean and variance.

`pnorm(q,mean=x,sd=y)` with the inverse function:
`qnorm(p,mean=x,sd=y)`

`lower.tail=F` can be added if required. The values for mean and standard deviation are now user specified (substitute for `x` and `y` in `pnorm` and `qnorm`). The default values are 0 and 1.

Finally, the t-test:

```
> pt(1.96,1000,lower.tail=F)
[1] 0.02513659

> pt(-1.96,1000)
[1] 0.02513659
```

We usually carry out a two sided t test – so we require the sum of the lower and upper tail probabilities: equal to two times the single tailed probability. The inverse function follows the usual format and nomenclature:

```
> qt(0.025,10000,lower.tail=F)
[1] 1.960201
```

5.6 Miscellany

Included here are some useful commands which have not so far been described.

`help(command)`

This opens a new window and provides help on the command. `help(lm)`for example, will give help on the linear modelling command that we have used extensively. The help is written in a terse and technical style however, which may be hard to understand. Nevertheless, it is useful to see what options are available with each command – for many of the commands used in this guide more are available than have been described. Often, sufficient of the output from `help` makes sense to be able to get a command working by trial and error. At the bottom of the output, there are often examples of the command's use: again not always easy to follow. The default argument

to `help` is `help`: `help()` is the same as `help(help)` and provides information about the `help` command itself.

```
history(x)
```

This opens up a window with a list of the most recently issued x commands. The default number is 25. These can be copied back into the R window and re-executed. The window with the output can be saved from the File menu to keep a record of commands issued during the R session.

`ls()` or equivalently `objects()`

Lists all the variables available in the current R session. This command can also be executed from the menu, selecting first “Misc”, then “List objects”.

To remove those that are no longer required:

```
rm(height,yield)
```

would remove the variables height and yield.

```
rm(list=ls(all.names=TRUE))
```

would remove all variables, although this can be more easily achieved from the windows menu by selecting “Misc” then “Remove all objects” random numbers

5.7 Saving work in progress

Work can be saved during a session by selecting “Save Workspace…” from the File menu and then following the prompts for a file name and location. The file extension for R is .Rdata . On resuming R, “Load Workspace…” can be selected to restore data and variables. You are also prompted to save you data when you exit R.

5.8 Exiting R

Either the R window can be closed, you can select “exit” from the File menu, or you can issue `quit()` from the command line.

5.9 Learn more

Much useful information and documentation is available on the R web site:
http://www.r-project.org/, including the R manual "An Introduction to R."

Note, the manuals, including "An introduction to R", are available directly from the Help menu from with R.

The book *Introductory Statistics with R,* Springer, 2002, ISBN 0-387-95475-9 is an excellent introduction both to R and to statistical analysis, with many simple examples.

Packages

Many are available from the Comprehensive R Archive network (CRAN) web site http://cran.r-project.org/. After downloading, these are easily installed from within R. Some of these are highly pertinent to plant genetics. An example, of which we have no direct experience is

qtl: Tools for analyzing QTL experiments
Analysis of experimental crosses to identify genes (called quantitative trait loci, QTLs) contributing to variation in quantitative traits.

Version:	1.00-17
Date:	7 Sep 2005
Author:	Karl W Broman and Hao Wu, with ideas from Gary Churchill and Saunak Sen and contributions from Brian Yandell
Maintainer:	Karl W Broman
License:	GPL version 2 or later

These packages generally come with their own manual, often detailed. That for the qtl package, for example, runs to 96 pages. Although the CRAN website is the first place to search for suitable packages, they are also found elsewhere and are often referred to in methodological publications or the methods sections of paper: programming in R is an expanding industry.

It is worth mentioning that our guide to the syntax and structures used in R has been very superficial. We have mentioned the data frame and little else. Knowledge of other structures – arrays, matrices, lists – should be acquired at some stage. They are explained in "An Introduction to R" but this book is not a page turner.

R's graphics capabilities are extensive, and have not been explored in detail.

5.10 List of commands described in this guide

General

attach	attaches a dataset to R for subsequent analyses
colnames	adds coumn names to a table
detach	attaches a dataset to R for subsequent analyses
demo	demonstration a command (not available for most commands)
getwr	returns the path to the working directory
help	returns help on a command
history	lists previously issued commands
`is.na()`	
length	returns the number of entries in a variate
ls	lists all active data tructures and variates
order	returns the order or a variate for use in a subsequent sort
quit	exit R
read.table	reads in data
rm	delete data structures and variates from R
rownames	adds row names to a table
sort	sorts data
subset	selected a subset of data for subsequent analysis
table	defines a table: used for input into contingency chi sq tests.

Graphical

abline	add the best fitting straight line to a scattergram
boxplot	produce a Boax-and-whisker plot
hist	plot a histogram
pairs	plot multiple scattergrams in a matrix format
plot	produce a scattergram

Statistics

cor	returns the correlation coefficient
mean	returns the mean of a variate
median	returns the median of a variate
minimum	returns the minimum of a variate
maximum	returns the amximum of a variate
quantile	returns the quantiles of a variate
sd	returns the standard deviationof a variate
sum	returns the sum of a variate
summary	summarise data
var	returns the mean of a variate

anova	return an anova table from a linear model
chisq.test	carry out a contingency chi-squared test
fisher.test	carry out a Fisher's exact test
fitted	returns the fitted values from a linear model
lm	define and execute a linear model
resid	returns the residuals from a linear model
t.test	one and two sample t-test
var.test	compare two variances by an F test

pchisq	returns the p-value of a chi-squared statistic
qchisq	returns a chi-squared statistic for a given probability
pf	returns the p-value of a F (variance ratio) statistic
qf	returns a F (variance ratio) statistic for a given probability
pnorm	returns the p-value for a normally distributed variate
qnorm	returns a normality distributed variate for a given probability
pt	returns the p-value of a t-test
qt	returns the t-test statistic for a given probability

6. Construction of genetic maps

6.1 Data files from Excel

Before copying files for mapping we can do some more editing. A copy of the marker names and scores can be seen in worksheet F2_MM.

From this worksheet we can remove unwanted data, eg. those individuals that have few/no scores or have x^2 values that do not fit the expected ratios. In this population, individuals 30 and 35 can be removed as they have no scores for any of the markers. The population size is now 118 and the number of markers is 153.

Editing of the marker titles also must be carried out, MapMaker is very fussy. Marker (locus) names for MM must be 8 or less characters and must start with * followed by a letter, eg. *T105p; must be of alphabetic or numeric character and not include \ or / or +; underscore is acceptable; all file titles must also be 8 or less characters; zero scores are - . Joinmap is less fussy but file titles must be 8 or less characters.

The mapping programs MapMaker (MM) and JoinMap (JM) both use text files from which the scoring data is read, conveniently the files are identical except for the header lines.

Exercise 3 File conversion: Excel to.txt

The next series of steps takes you through converting the Excel file to .txt. Proceed through the instructions below and make two .txt files for use with MapMaker and JoinMap.

1. Select the whole data set and COPY it, open a new Excel workbook and PASTE the sheet.
2. File, save as, choose CSV (Comma delimited) *.csv option
3. Then close the Excel workbook
4. Go to My Computer and open the .csv file in WORD
5. The commas need removing: go to Edit, Replace: put ,,, in the FIND box and put 5 spaces in the Replace box; then repeat this procedure and replace , with nothing
6. File, save as .txt file
7. Put in the approprite headings for either MapMaker or JoinMap, and call them either F2_MM.txt or F2_JM.txt, see below:

```
MapMaker
data type f2 intercross
118 153 0
*T56p
DDDDDDDDDBDBDDDDDDDDDBDBBDBDDBDBBDBBDDDDDBDDDDDBDDDDBDBDDDDDDDDBDBDDDDDDBDBD
DDDBDBDDBDDBDBDBDDDDDDDDDDDDDBDDDDDDDDDDDD
*T57p
DBDDBBDDDBDDDDDBDDDBDBBDBDDDBDDDDDBDDDDDDBDDBDDBDDDBDDDDBBDDDDDDDDDDDDDDDDD
DDDDDDDDDDDDDDDDDDDDBDDDDBBBDBDBDBDDBDDDD
*T58p
DBDBDBDBDDBDDDDDDDDDBDBBBDBDBDDDDDBDDDDDDDBBDBDBDBBBDDDDDDBDDBBDDDBDDDBBDD
BDBBBDBDBBDBDDBDBDDBDBDDDBDDDDDDDDDDDDDDD
```

For MM the first line has the generation and cross type. The second line has three numbers each separated by a single space; the first number (118) indicate the the population size, ie. the number of progeny for which there is data within the file; the second value (153) indicates the number of marker loci; the third value indicates the number of quantitative trait loci (QTL), in this case 0.

```
JoinMap
name = 15x399
popt = F2
nloc = 153
nind = 118
*T56p
DDDDDDDDDBDBDDDDDDDDDBDBBDBDDBDBBDBBDDDDDBDDDDDBDDDDBDBDDDDDDDDBDBDDDDDDBDBD
DDDBDBDDBDDBDBDBDDDDDDDDDDDDDBDDDDDDDDDDDD
*T57p
DBDDBBDDDBDDDDDBDDDBDBBDBDDDBDDDDDBDDDDDDBDDBDDBDDDBDDDDBBDDDDDDDDDDDDDDDDD
DDDDDDDDDDDDDDDDDDDDBDDDDBBBDBDBDBDDBDDDD
*T58p
DBDBDBDBDDBDDDDDDDDDBDBBBDBDBDDDDDBDDDDDDDBBDBDBDBBBDDDDDDBDDBBDDDBDDDBBDD
BDBBBDBDBBDBDDBDBDDBDBDDDBDDDDDDDDDDDDDDD
```

Similarly for JM but the format is slightly different and there is no need for a QTL number.

8. Now the files are ready to input to the mapping programmes. Make a copy of the ….MM.txt file in the Mapmaker folder on the C:\ drive.

Note. JoinMap will not be actively used during the course as this is not free software, however it will be demonstrated during the course. There is now currently a windows version v3 available. The flow chart below gives a brief outline:

6.2 Mapmaker (MM) Tutorial
You can find mapmaker at:
http://www.broad.mit.edu/ftp/distribution/software/mapmaker3

MapMaker[13], unlike JoinMap[7], is interactive and the operator has complete control over obtaining the 'best' marker order. It becomes difficult and extremely time consuming to deal with large data sets, of greater than 100 loci. However, Mapmaker provides good tables for analysis of two point data and of genotypes, both of which are extremely useful and informative. We will use the F2 data set from pea with the full set of 153 markers to construct a map.

Exercise 4: Running Mapmaker

The next series of commands is lengthy but will gradually build up a map. It is useful to have access to a printer and to some of the outputs from the programme as it proceeds.

Where to find the program: go to My Computer and select the C:\ drive, and the Mapmaker folder.
All the programs needed for the commands are within this folder. You will see there are many different types of files.

Select the 

icon, this starts the programme running.

Input the following commands:

1> prepare data F2_MM.txt	some text follows
2> load data F2_MM	
3> photo F2	makes file F2.out for session output
4> seq all	consider all marker loci
5> group	Min LOD 3, max distance 50

At this point MM splits the data into groups of linked loci based on 'two-point', or pairwise, linkage analysis based on a minimum LOD of 3 and a maximum distance of 50 (the default) and uses recombination frequency data.

[13] LANDER, E.S., P. GREEN, J. ABRAHAMSON, A. BARLOW, M. J. DALY, *et al.*, 1987 MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics **1:** 174-181
7 Van Ooijen JW, and RE Voorrips 2001. Joinmap® 3.0 Software for the calculation of genetic linkage maps. Plant Research International, Wageningen, The Netherlands.

You see below, from the F2.out file that MM has created 12 linkage groups with one marker in an unlinked group using the default criterion.

```
5> group
Linkage Groups at min LOD 3.00, max Distance 50.0

group1= T56p S55p T173p
-------
group2= T57p S48p T89m T126m T72m T185m
-------
group3= T58p T192m
-------
group4= T59p T78p T82p S19p S28p S31p S47p T21m T24m T60m T9m T11m T13m S18m
S29m S41m S51m S52m T121m T174m T176m T201m T189m T195m S60m S70m S77m S83m
S86m S99m CAG4m T64p T68p T70p T182p S68p S88p S95p S97p S98p CAG5p
-------
group5= T200p T77p S35p T87m S4m T40m T170m S75m S85m T187p S80p
-------
group6= T76p S7p S8p T108m T110m S10m S17m S46m T181m T188m S73m S79m S91m
T124p T163p T199p S58p S89p S93p S100p S102p B1_PDR1
-------
group7= T109p T28m T140m Catheps
-------
group8= T111p S1p S5p S30p S32p S42p T12m S22m S26m S34m S39m S44m S50m
T168m T169m T194m S64m S69m S104m T74p T166p T179p T197p S59p S103p CAG7p
-------
group9= S9p S12p S53m S84m S74p S94p
-------
group10= S21p S27p S36p S38p S54p S2m S20m S49m T165m T183m S67m S78m S105m
T66p T67p T73p T171p T202p S82p CAG1p
-------
group11= S45p T162m S90m T127p T159p T167p T198p S76p S107p
-------
group12= S3m S65p
-------
unlinked= T158pm
```

Perhaps you want to be less stringent and so reduce the number of linkage groups then use the commands as follows:

```
6> seq all
7> default linkage criteria 2 50
8> group
```

With the LOD reduced to 2 the number of linkage groups is reduced to 7 with one marker unlinked, see below.

```
8> group
Linkage Groups at min LOD 2.00, max Distance 50.0

group1= T56p S9p S12p S55p S53m S84m T173p S74p S94p
-------
group2= T57p S48p T89m T126m T72m T185m
-------
group3= T58p T192m
-------
group4= T59p T76p T78p T82p S7p S8p S19p S28p S31p S45p S47p T21m T24m T60m
T9m T11m T13m T108m T110m S10m S17m S18m S29m S41m S46m S51m S52m T121m
T162m T174m T176m T181m T201m T188m T189m T195m S60m S70m S73m S77m S79m
```

```
S83m S86m S90m S91m S99m CAG4m T124p T127p T64p T159p T68p T70p T163p T167p
T182p T198p T199p S58p S68p S76p S88p S89p S93p S95p S97p S98p S100p S102p
S107p CAG5p B1_PDR1
-------
group5= T200p T77p S21p S27p S35p S36p S38p S54p T87m S2m S3m S4m S20m S49m
T40m T165m T170m T183m S67m S75m S78m S85m S105m T66p T67p T73p T171p T187p
T202p S65p S80p S82p CAG1p
-------
group6= T109p T28m T140m Catheps
-------
group7= T111p S1p S5p S30p S32p S42p T12m S22m S26m S34m S39m S44m S50m
T168m T169m T194m S64m S69m S104m T74p T166p T179p T197p S59p S103p CAG7p
-------
unlinked= T158pm
```

Coincidentally pea has seven linkage groups but the LOD is low. Group 4, for example, has many markers that may not in fact be linked. To have confidence in the linkages it is better practice to have many linkage groups with fewer markers that can be linked together at a later stage. Smaller groups are also more manageable within MapMaker as you will see further on in this tutorial. So it is better to use minimum LOD scores of 3 or greater.

Choose the following:

9> seq all
10> default linkage criteria 4 50
11> group

```
11> group
Linkage Groups at min LOD 4.00, max Distance 50.0

group1= T56p S55p T173p
-------
group2= T57p S48p T89m T126m T72m T185m
-------
group3= T58p T192m
-------
group4= T59p T78p T82p S19p S28p S31p S47p T21m T24m T60m T9m T11m T13m S18m
S29m S41m S51m S52m T121m T174m T176m T201m T189m T195m S60m S70m S77m S83m
S86m S99m CAG4m T64p T68p T70p T182p S68p S88p S95p S97p S98p CAG5p
-------
group5= T200p T77p S35p T87m S4m T40m T170m S75m S85m T187p S80p
-------
group6= T76p S7p S8p S10m S17m S46m T181m T188m S73m S79m S91m T124p T163p
S58p S89p S93p S100p S102p B1_PDR1
-------
group7= T109p T28m T140m Catheps
-------
group8= T111p S1p S5p S42p T12m S22m S26m S34m S39m S44m S50m T168m T169m
T194m S64m S69m S104m T74p T166p T179p T197p S59p S103p
-------
group9= S9p S12p S53m S84m S74p S94p
-------
group10= S21p S27p S36p S38p S54p S2m S20m T165m T183m S78m S105m T66p T67p
T73p T171p T202p S82p CAG1p
-------
group11= S30p S32p CAG7p
-------
group12= S45p T162m S90m T127p T159p T167p T198p S76p S107p
-------
group13= T108m T110m
```

```
-------
group14= S49m S67m
-------
unlinked= S3m T199p S65p T158pm
```

It may be more useful to have the units in recombination frequencies rather than distance (cM). These can be adopted by changing the units as follows:

12> units rf
13> default linkage criteria 4 0.32
14> group

```
14> group
Linkage Groups at min LOD 4.00, max Distance 0.316

group1= T56p S55p T173p
-------
group2= T57p S48p T89m T126m T72m T185m
-------
group3= T58p T192m
-------
group4= T59p T78p T82p S19p S28p S31p S47p T21m T24m T60m T9m T11m T13m S18m
S29m S41m S51m S52m T121m T174m T176m T201m T189m T195m S60m S70m S77m S83m
S86m S99m CAG4m T64p T68p T70p T182p S68p S88p S95p S97p S98p CAG5p
-------
group5= T200p T77p S35p T87m S4m T40m T170m S75m S85m T187p S80p
-------
group6= T76p S7p S8p S10m S17m S46m T181m T188m S73m S79m S91m T124p T163p
S58p S89p S93p S100p S102p B1_PDR1
-------
group7= T109p T28m T140m Catheps
-------
group8= T111p S1p S5p S42p T12m S22m S26m S34m S39m S44m S50m T168m T169m
T194m S64m S69m S104m T74p T166p T179p T197p S59p S103p
-------
group9= S9p S12p S53m S84m S74p S94p
-------
group10= S21p S27p S36p S38p S54p S2m S20m T165m T183m S78m S105m T66p T67p
T73p T171p T202p S82p CAG1p
-------
group11= S30p S32p CAG7p
-------
group12= S45p T162m S90m T127p T159p T167p T198p S76p S107p
-------
group13= T108m T110m
-------
group14= S49m S67m
-------
unlinked= S3m T199p S65p T158pm
```

The groupings at 11> and 14> are the same, for MapMaker a maximum distance of 50 is equivalent to a maximum recombination frequency of 0.316.

Having decided on the final groupings number we can start to analyse the data further and construct a map and in so doing look at the recombination frequency and genotype data as we proceed.

Starting systematically with the group 1 that has 3 markers and that can be handled quickly and easily by Mapmaker:

15> seq group 1
16> LOD Table

```
16> Lod table

Bottom number is LOD score, top number is recombination fraction:

         T56p
               S55p

S55p     0.187
          6.16

T173p    0.018 0.177
         22.91  6.59
```

The table of pairwise recombination frequencies suggest that markers T56p and T173p have the closest linkage (0.018), T56p and T55p are the furthest apart (0.187) of the 3 markers, so the order could be T56p T173p T55p. The matching LOD scores, all greater than 3, also suggest we can have confidence in these recombination frequency values.
What will Mapmaker do?
With 3 markers we get three maps:

17> print names off — marker names to a number
18> error detection on — see manual
19> seq {1 33 121) — {means all combinations}
20> map

```
20> map
Map:  1                      Apriori
  Markers          Distance   Prob  Candidate Errors
    1  T56p        0.014 rf
  121  T173p       0.176 rf   1.0%  [#107 D-B-- 1.73] [#85 B-D-D 1.12]
   33  S55p       ----------
                    23.2 cM   3 markers   log-likelihood= -50.90
=========================================================================
Map: 2                       Apriori
  Markers          Distance   Prob  Candidate Errors
  121  T173p       0.009 rf
    1  T56p        0.179 rf   1.0%  [#85 D-B-D 2.37][#107 B-D-- 1.75]
   33  S55p       ----------
                    23.1 cM   3 markers   log-likelihood= -51.00
=========================================================================
Map:  3                      Apriori
  Markers          Distance   Prob  Candidate Errors
  121  T173p       0.142 rf
   33  S55p        0.151 rf   1.0%  [#9 D-B-D 1.23][#11 D-B-D 1.23] 15 more
    1  T56p       ----------
                    34.7 cM   3 markers   log-likelihood= -65.63
=========================================================================
```

From this data, map 3 is the longest and also has 17 cases of genotype switch (double crossover events). Map1 or 2 on the other hand are about the same length and both have only two cases of possible double crossover. Looking back at the recombination frequencies the best order was T56p T173p T55p which is map 1 above.

21> seq 1 121 33 this tells MM the fixed sequence
22> units cm centimorgans
23> map

```
23> map
=========================================================================
Map:                        Apriori
  Markers          Distance   Prob  Candidate Errors
    1  T56p          1.4 cM
  121  T173p        21.8 cM   1.0%  [#107 D-B-- 1.73][#85 B-D-D 1.12]
   33  S55p        ----------
                    23.2 cM   3 markers   log-likelihood= -50.90
=========================================================================
```

24> genotypes

```
24> genotypes
=========================================================================
Genotypes:
                  000000000000000000000000000000000000000000000000000
                  000000001111111111222222222233333333344444444445
                  123456789012345678901234567890123456789012345678890
                  --------------------------------------------------
 T56p             DDDDDDDDDBDBDDDDDDDDDBDBBDBDDBDBBDBBDDDDDBDDDDDBDD
          1.4 cM
 T173p      1.0% DDDDDDDDDBDBDDDDDDDDDBDBBDBDDBDBBDBBDDDDDBDDDDDBDD
         21.8 cM
 S55p             DDDDDDDDBBBBDDDD-DDDD-DDBDDDDDDBDDDBDBDDDDDDDDDB-D
                  --------------------------------------------------
           #Recs: 000000000000000000000000000000000000000000000000000

                  000000000000000000000000000000000000000000000000001
                  555555556666666666777777777788888888899999999990
                  123456789012345678901234567890123456789012345678890
                  --------------------------------------------------
 T56p             DDBDBDDDDDDDDBDBDDDDDDBDBDDDDBDBDDBDBDBDBDDDDDDDD
          1.4 cM                                    |
 T173p      1.0% DDBDBDDDDDDDDBDBDDDDDDBDBDDDDBDBDDBDBDBDBDDDDDDDD
         21.8 cM                                    |
 S55p             DDDDBDDDDDDDDBDBDDDDDDDDBBDDDBDBBDDDDBDBDBBDDDD-BD
                  --------------------------------------------------
           #Recs: 000000000000000000000000000000000000000000000000000

                  111111111111111111
                  000000001111111111
                  123456789012345678
                  ------------------
 T56p             DDDDDBDDDDDDDDDDD
          1.4 cM       |
 T173p      1.0% DDDDDBBDDDDDDDDDD
         21.8 cM       |
```

```
 S55p             DB-D-B-DDDDDDDDDB
                  ------------------
           #Recs: 00000000000000000

======================================================================
```

The output from the Genotypes command shows the individuals genotypes and crossover positions. This information was also given as probable candidate errors from the Map output (see 22> map). These scorings are almost certainly worth back checking to the original scores to check for typing/scoring error in the original data set for the particular individuals. This is a necessary part of the data checking process.

We have mapped group1, now we will move on to group2 that has 6 markers. This number of markers is too many for Mapmaker to handle simply as we have done for group1. For four or more markers the following set of commands should be used, also it best to work with print names off as we will be input lists of numbers:

25> print names off
26> seq all
27> group

```
27> group
Linkage Groups at min LOD 3.00, max Distance 0.316

group1= 1 33 121
-------
group2= 2 31 43 67 70 81
-------
group3= 3 84
-------
group4= 4 8 9 18 21 23 30 34 35 37 38 39 41 51 55 58 63 64 66 76 77 80 83 86 87
91 94 97 100 103 106 109 113 114 123 132 137 141 142 143 149
------- etc.....
```

28> error detection on
29> seq {2 31 43 67 70} {tells MM to look at combinations}
30> compare

```
30> compare

Best 20 orders:
1:    67 70 2 31 43   Like:  0.00
2:    67 2 70 31 43   Like: -0.03
3:    67 70 2 43 31   Like: -0.15
4:    67 2 70 43 31   Like: -0.31
5:    67 70 31 43 2   Like: -1.50
6:    2 67 70 31 43   Like: -1.57
7:    67 70 31 2 43   Like: -1.64
8:    67 70 43 2 31   Like: -2.18
9:    67 70 43 31 2   Like: -2.19
10:   67 2 31 70 43   Like: -2.23
11:   2 67 70 43 31   Like: -2.35
12:   2 67 31 70 43   Like: -2.42
```

```
13:   67 31 70 2 43   Like: -2.52
14:   67 31 70 43 2   Like: -2.66
15:   43 31 2 67 70   Like: -3.37
16:   31 43 2 67 70   Like: -3.42
17:   43 31 67 2 70   Like: -3.88
18:   67 31 2 70 43   Like: -4.23
19:   31 43 67 2 70   Like: -4.29
20:   2 43 31 67 70   Like: -4.77
order1 is set
```

31> seq order1 best suggested order from MM

32> map

```
32> map
================================================================
Map:                        Apriori
  Markers          Distance   Prob  Candidate Errors
   67  T126m       0.198 rf
   70  T72m        0.053 rf   1.0%  [#66 A-C-D 1.05]
    2  T57p        0.194 rf   1.0%  -
   31  S48p        0.074 rf   1.0%  -
   43  T89m       ----------
                    63.3 cM   5 markers   log-likelihood= -124.04
================================================================
```

33> seq 67 70 2 31 43 this is the fixed order

34> try 81

```
34> try 81

        81
       -------
      |-23.07 |
67    |       |
      |-25.39 |
70    |       |
      | -4.91 |
2     |       |
      | -0.14 |
31    |       |
      |  0.00 |
43    |       |
      | -0.15 |
      |-------|
INF   |-24.70 |
       -------
BEST  -129.44
```

The output is suggesting the best position for 81 9T185m) is between marker 31 (S48p) and 43 (T89m), so try this:

35> seq 67 70 2 31 81 43

36> map

```
36> map
================================================================
Map:                        Apriori
  Markers          Distance   Prob  Candidate Errors
   67  T126m       0.198 rf
   70  T72m        0.049 rf   1.0%  [#66 A-C-D 1.07]
    2  T57p        0.193 rf   1.0%  -
   31  S48p        0.063 rf   1.0%  -
   81  T185m       0.018 rf   1.0%  [#117D-C-A 2.21][#106D-C-A 1.55] 1 more
   43  T89m       ----------
```

```
                    63.2 cM   6 markers   log-likelihood= -129.44
=======================================================================
```

37> lod table

```
37> lod table
Bottom number is LOD score, top number is recombination fraction:

      67    70    2     31    81

70   0.206
      6.62

2    0.167 0.049
      3.30  5.02

31   0.300 0.142 0.177
      1.80  4.83  7.66

81   0.400 0.281 0.049 0.048
      0.57  3.38  4.40  5.80

43   0.389 0.288 0.049 0.048 0.025
      0.78  3.44  4.85  6.42 24.20
```

From the recombination frequency data this suggested order from MM does seem to be reasonable. Look at the genotypes:

38> genotypes

```
38> genotypes

Genotypes:
                  00000000000000000000000000000000000000000000000000
                  00000000011111111112222222222333333333344444444445
                  12345678901234567890123456789012345678901234567890
                  --------------------------------------------------
 T126m            CCCACCAACCCCACCCCCACCCCCACCCCACCCCAACCCACCCCCCCCCA
          25.2 cM       X                        X
 T72m        1.0% CCCACCCACCCAACCCCCACCCCCCAACCCCCCCCAACCCACCCCCCCCA
           5.1 cM
 T57p        1.0% DBDDBBDDDBDDDDDBDDDBDBBDBDDDBDDDDDBDDDDDDBDDBDDBDD
          24.4 cM
 S48p        1.0% DBDDBBDDDBDDDBDBBBDBDBBDBDDBBDDDDBBBDDDBDDDDDDDDBDD
           6.7 cM       XX                  X   X    XX   X
 T185m       1.0% CCAACCCCACAAACACCCACACCACAC-CCCACCCCCCCCACCACCCCCCA
           1.8 cM       XX                  X   X    XX   X
 T89m             CCAACCCCACAAACACCCACACCACACCCCCACCCCCCCCACCACCCCCCA
                  --------------------------------------------------
           #Recs: 00000032000000000000000000200030000220002000000000

                  00000000000000000000000000000000000000000000000001
                  55555555566666666667777777777888888888899999999990
                  12345678901234567890123456789012345678901234567890
                  --------------------------------------------------
 T126m            ACA---ACCCCAACAAACCCCCAAACAACCCCCACAC-CA------CCAA
          25.2 cM       X    XX  |X     XXX  X     X
 T72m        1.0% ACAACCCCCAACCAACCCCCAACCCAACCACCCCAACACACCCCAACCAA
           5.1 cM       O        |
 T57p        1.0% DBDDDDBBDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDBDD
          24.4 cM            O    O               OO
 S48p        1.0% DDDDDDBBDDDBDDDDBDDBDDDDDDDDDBDDBBDDDDDBDDDDDBDD
           6.7 cM    X        XX      |XXXX  X       X X X    X
 T185m       1.0% -CACCCCCC-ACCCAACCACACCCCAACCACCACCCCCCCCCCACACCAA
```

```
          1.8 cM    X        XX      |XXXX  X       X X X    X
T89m              ACACCCCCCAACCCAACCACCCCCCAACCACCACCCCCCCCCCCACACCAA
                  ----------------------------------------------------
          #Recs:  00020030000332013000123330030000032202020000200000

                  11111111111111111
                  00000000111111111
                  12345678901234678
                  ------------------
T126m             AACCCCCACA-ACCC-AC
         25.2 cM
T72m        1.0% AACCCCCACACACCCAAC
          5.1 cM
T57p        1.0% DDBBBDBDBDBDDBDDDD
         24.4 cM
S48p        1.0% DDBBBDDDDDBDDBDDDD
          6.7 cM XX   |          |
T185m       1.0% CCCCCCCAC-CACC-ACA
          1.8 cM XX   |          |
T89m              CCCCCACACCCACCAAAA
                  ------------------
          #Recs:  220000000000000010
```

The symbols within the genotype output are:

X = a crossover

0 = two crossovers on one interval

| = suggests a possible candidate error: given when error detection is on

"?" = an obligate recombinate which cannot be placed: also an indication of typing/scoring error

Exercise 5 marker order and map length

As an exercise for this group2 try placing 81, or any other marker, elsewhere in the list to see what happens.

Exercise 6 mapping some other linkage groups

Proceed through the other groups from the output at 14> above and map the groups as described. At any one time when using the > Try command only ever input two markers.

39>
40>
41>

6.3 Use of Mapchart with MM:

Mapchart, now incorporated as an integral part of Joinmap 3.0, is available free. Given a file of map order, marker name and distance, it

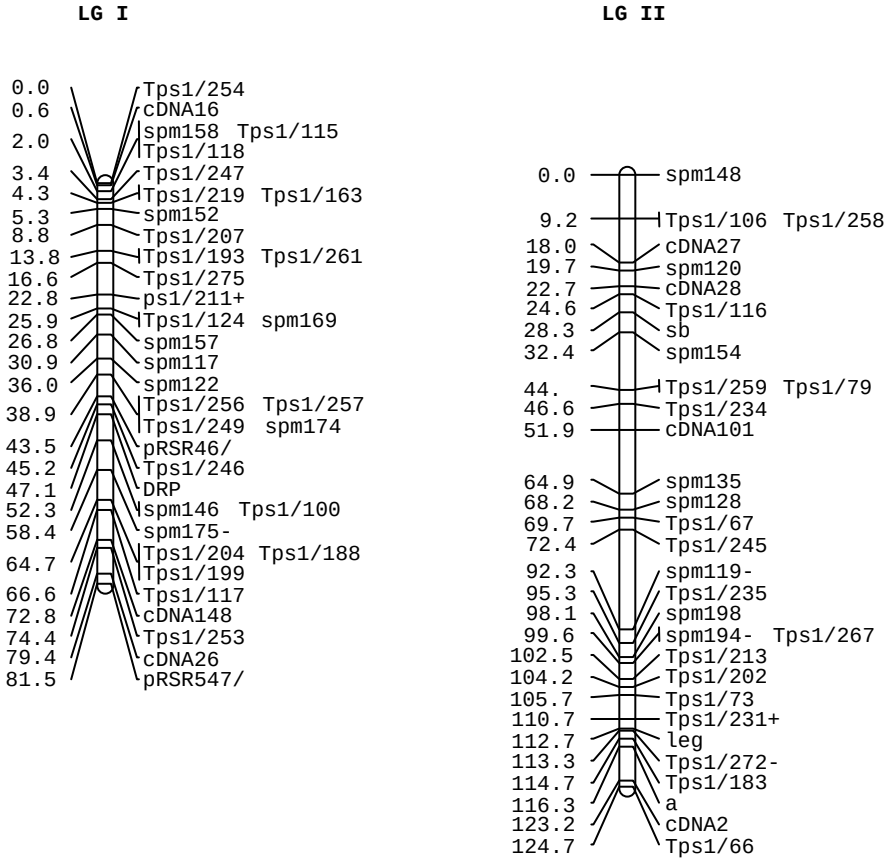
will draw the linkage groups. This can either be data from Mapmaker or earlier versions of Joinmap.

A text file format can be input to Mapchart and linkage groups drawn up. See below an example of the file format for the top and bottom of two linkage groups from the RI population for the cross JI15 x JI1194, as input to Mapchart.

```
Group 1(15x1194)
Tps1/254+              0.0
cDNA169                0.6
spm158-                2.0
Tps1/115-              2.0
.
.
.
cDNA148t              72.8
Tps1/253+             74.4
cDNA267               79.4
pRSR547/7             81.5


Group 2(15x1194)
spm148+                0.0
Tps1/106-              9.2
Tps1/258-              9.2
cDNA277               18.0
.
.
.
.
Tps1/183-            114.7
a                    116.3
cDNA24               123.2
Tps1/66+             124.7
```

From Mapchart the linkage groups can be copied to MS Powerpoint for editing: replacing groups with linkage group titles etc.


LG I
0.0 Tps1/254
0.6 cDNA16
2.0 spm158 Tps1/115
Tps1/118
3.4 Tps1/247
4.3 Tps1/219 Tps1/163
5.3 spm152
8.8 Tps1/207
13.8 Tps1/193 Tps1/261
16.6 Tps1/275
22.8 ps1/211+
25.9 Tps1/124 spm169
26.8 spm157
30.9 spm117
36.0 spm122
38.9 Tps1/256 Tps1/257
Tps1/249 spm174
43.5 pRSR46/
45.2 Tps1/246
47.1 DRP
52.3 spm146 Tps1/100
58.4 spm175-
64.7 Tps1/204 Tps1/188
Tps1/199
66.6 Tps1/117
72.8 cDNA148
74.4 Tps1/253
79.4 cDNA26
81.5 pRSR547/
LG II
0.0 spm148
9.2 Tps1/106 Tps1/258
18.0 cDNA27
19.7 spm120
22.7 cDNA28
24.6 Tps1/116
28.3 sb
32.4 spm154
44. Tps1/259 Tps1/79
46.6 Tps1/234
51.9 cDNA101
64.9 spm135
68.2 spm128
69.7 Tps1/67
72.4 Tps1/245
92.3 spm119-
95.3 Tps1/235
98.1 spm198
99.6 spm194- Tps1/267
102.5 Tps1/213
104.2 Tps1/202
105.7 Tps1/73
110.7 Tps1/231+
112.7 leg
113.3 Tps1/272-
114.7 Tps1/183
116.3 a
123.2 cDNA2
124.7 Tps1/66

You can also just use EXCEL to plot genetic maps manually. There are advantages and disadvantages to both.

7. Comparative genetic maps

Comparative genetic maps come in two basic forms. The first relates different maps within the same species, while the second seeks to compare genetic maps of different species.

7.1 Intraspecific comparisons

Comparisons of genetic maps within a species rely on the collection of data from the same genetic locus in different crosses. Ideally the use of a common parent in different crosses will simplify matters and ensure that for many loci the segregation is monitored for the same allele, but this information is not always available and the resulting difficulty is relatively simply overcome, for example by choosing the same DNA sequence for marker generation. This may simply be the use of the same primer combination in a PCR assay or the use of the same hybridisation probe for RFLP analysis. This does not guarantee that the same locus will be identified: there may be two or more loci that carry very similar DNA sequences, and different loci may be polymorphic in different crosses. The collation of data from several loci will help to confirm the inference.

In the collation of marker data from different crosses, with one parent in common, the degree of polymorphism for the markers in question is important. Consider three inbred lines A, B and C and the two crosses A x B and B x C. If the average frequency by which a given marker distinguishes any two lines is p, then the chance that it will distinguish the parents of both crosses is p^2. Thus for comparative mapping within a species the choice of parents and marker system is critical.

Microsatellite markers (4.4.2) are often selected as markers of choice for intra specific comparisons. There are essentially two reasons for this (i) the markers are generally co-dominant and (ii) microsatellite markers are highly polymorphic. Microsatellite markers have the disadvantage that they are often[14] derived from intergenic DNA sequences. In part this explains their high level of polymorphism, but unfortunately has the consequence that the primer binding sites are

[14] A minority of microsatellite markers are found in coding sequences.

also highly variable. This means that the primers have a limited taxonomic range of usefulness, and are often restricted to individual species. The variation in primer binding sites also provides an explanation for the frequency of null alleles for microsatellites in highly diverse material.

7.2 Interspecific comparisons

For inter-specific genetic map comparisons it is important that the markers used can be identified easily and unambiguously in all the species being compared.

Genes are the obvious choice of DNA sequences from which to derive genetic markers that can be compared across species. Gene phylogenies can be determined and their topology investigated with respect to species phylogenies. Sequence comparison using reciprocal BLAST analysis makes it possible to determine the most likely counterparts in the species being compared. This is not a trivial problem and the behaviour of DNA sequences in reciprocal BLAST analysis, together with their relative positions in genetic maps can be used in arguments about orthology[15]. There is a variety of marker methods that can be used to identify corresponding genes in different species. RFLPs were the first to be used, and remain an important source of evidence, but this method is much more demanding than PCR based approaches discussed below. Comparative RFLP maps were the basis for the determination of patterns of synteny among cereals[16].

PCR markers that have been designed based on gene sequences have been discussed in sections 4.4.4 and 4.4.5, so these methods, and their multiple variants, will not be discussed further.

There are multiple web-based information resources available that provide information of primers that work well in selected species groups, and the way to access these will be discussed in the course.

There are some features of the genome that may be of interest: the position of centromeres and telomeres for example, however these are more likely to be structures asked about rather than defining comparative genetic maps. These can be genetic markers, or identified by genetic markers but their location is determined with respect to a genetic map, leaving the problem of the identification of common markers.

It is worth bearing in mind that there are non-genetic approaches to studying genome synteny. One is the comparison of extended genome sequence deposited in public databases; another is the use of cytogenetic methods such as the *in situ* hybridisation of BAC clones to chromosome spreads. These can have very fine resolution: for example stretched fibre *in situ* hybridisation has a resolution of about a kilobase. However this approach simply provides information about

[15] see glossary
[16] Devos, K. M. & Gale, M. D. (2000) Plant Cell 12, 637–646.

the similarity between genomes. In the context of the present course we want to find out whether related traits are under equivalent control in different species. This can provide access to a wealth of genetic markers and basic understanding of the underlying biological processes.

8. Looking at data and mapping exercises

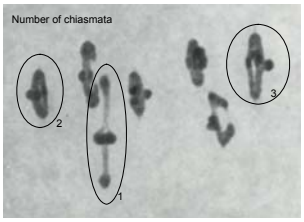
8.1. Interpretation of mapping data

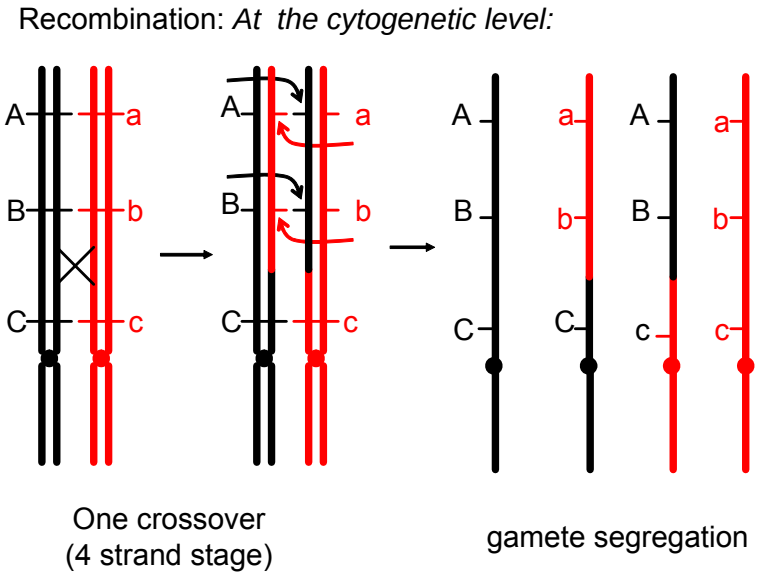
It is easy to forget that genetic maps are an interpretation rather than a representation of data. Marker data is collected and can be scrutinised in great detail to be sure there are no mistakes. The mapping programmes are a defined operation performed in the data so they seem a necessary *consequence.* A genetic map looks like a clear and simple representation of the data, but this suggestion is misleading. There is a great deal of interpretation hidden in the construction of a genetic map. This is illustrated by Isidore et al (2004)[17]

8.1.1 Map length

Recombination is a consequence of events that happen at meiosis, and the segregation of chromosomes into gametes followed by their (random) association in zygotes gives us the pattern of segregation of alleles among progeny. It was discussed above how markers (of whatever type) are an assay for the parent-of-origin of a genetic locus that tracks a chomosomal segment. There are many ways the assay can be misleading and we should be aware of this when examining the data.

Map length is a simple clue to how good a genetic map actually is at representing these meiotic events for a given population. The figure below illustrates the consequence of crossing over for gametes. The cytogenetic event that corresponds to a crossover is called a chiasma, plural chiasmata (see inset picture from J.S. Parker)





[17] Isidore et al (2003) Toward a Marker-Dense Meiotic Map of the Potato Genome: Lessons From Linkage Group I Genetics 165: 2107–2116

It is clear that after 1 crossover 50% of the gametes are recombinant. If a crossover happens in 1/N meioses then 2 out of 4N gametes will be recombinant. If we think about a chromosome where only one crossover occurs[18] but there are many different genetic markers positioned along the chromosome then the chance that there is a crossover between any two markers will be small, but there will always be one crossover between the extreme ends of the chromosome. The alleles of markers at the two extreme ends will be in the parental configuration 50% of the time and in the recombinant configuration in the other 50%. The recombination fraction between them is 0.5, and the map length is 50 cM. Work out what happens for two crossovers and if you are keen try more. You will see that the genetic map length corresponding to a chromosome with *x* crossovers is 50*x* centiMorgans.
If you know the cytogenetics if your population you know how long the genetic map should be. Usually genetic maps are longer than they should be and there while it is possible that the estimate of chiasma number is too low, it is likely that the marker data propose too many recombination events. This may be because of simple errors, but the Isidore et al (2003) paper clearly shows that this may not be the only reason. We will return to this point later, but for the moment the issue is what this means for the interpretation of that data from which a genetic map is constructed.

If a marker corresponds to a mis-score then it either proposes two additional recombination events or it may not be noticed because it (erroneously) extends a region of one parental genotype:

Suppose we have the code below describing the scores for an individual (note this diagram is on its side compared to the diagrams in section 6.

Marker number	1	2	3	4	5	6	7	8	9
Allele score	A	A	A	A	A	A	C	C	C

(1) Mis-score proposing no extra recombination events:

Marker number	1	2	3	4	5	6	7	8	9
Allele score	A	A	A	A	A	**C**	C	C	C

[18] In general crossovers are necessary for proper chromosome disjunction so it is usual for a chromosome to have at least one crossover. Usually there are more because an average of one would suggest some with zero and hence non-disjunction. An average of one crossover per chromosome arm is not far from the truth. Famously, in *Drosophila* males there is no crossing over: it is not clear (at least to me) how proper disjunction is achieved.

(2) Mis-score proposing two extra recombination events:

Marker number	1	2	3	4	5	6	7	8	9
Allele score	A	A	**C**	A	A	A	C	C	C

The mis-scores of the type shown in (1) should be fairly rare, so if we ignore them we can ask firly simple questions about mis-score rates and their consequences for map length. We will slightly exaggerate the effect on map length by ignoring (1) above.

Suppose there is a population of N gametes and we have the scores for m markers in a single linkage group[19]. Let's further suppose that there is an average of c crossovers per gamete (ie $2c$ crossovers per bivalent). This means that if we read the allele codes for each individual in turn we will count i instances where the scores change. There is a total of $i = c\,N$ of these. For 100 gametes and a fictitious chromosome with an average of one crossover per gamete (two per bivalent, one per arm) $i = 100$.
If the average distance between genetic markers is 5 centiMorgans (a low density map), then we know that there are 21 markers (20 intervals of 5 cM = 100 cM the length of a linkage group for a chromosome with an average of two chiasmata).
If the fraction marker scores in error is e then we have $2emN$ extra recombination events proposed by the data. To double the length of a chromosome $2emN = c\,N$, in this case $e = c/2eM$, in this case ~ 2.4%.
Marker scores an error rate of 2.4% are not great, but also not that bad given the number of operations AND the possibility that some of these 'errors' are inevitable.

Two things are clear:

1.2.1.1 Small error rates per marker score contribute greatly to excess map length.

1.2.1.2 Excess map length due to misscore increases with marker density. If map length increases with marker number that is a teltale sign of misscores.

8.1.2 Length distribution of non-recombined segments

Another way of looking at the validity of marker data is to examine the distribution of alleles. Often this is achieved by 'graphical genotyping' or 'colormapping'. These are intuitive methods, and have the virtue of focussing attention on the meaning of marker scores. This can also be examined analytically. Given the identity of the course sponsors it seems appropriate to deviate a little from standard approaches to consider the length distribution of non-recombined segments.

[19] linkage groups and chromosomes: see section 2.6

Linkage maps are usually represented as linear arrays of markers spaced by genetic distances. An alternative representation is to display sequences of marker intervals (or recombination bins) where there may or may not be an odd number of exchanges. This unusual representation has some useful properties that derive from being a sequence of two classes of object.

An interval can be designated 'a' for an odd number of exchanges or 'b' for zero or an even number of exchanges. Thus a single linkage group could be designated by a string such as 'aaabaaaabaa' where there are two recombination events at the intervals marked 'b'. While this is less intuitive than designating allele codes as in section 3.4, it has the advantage that the spacing between different interval types, and the statistics of this class of arrangement have been described by Mood (1940)[20], and in a slightly different context by Southern ()[21]. An isolated marker with where the allele is different from all surrounding alleles will be represented by 'bb', or where the number of 'a's is 0.

The expectation (E) of the number of runs (r) of 'a's of length N is given as:

$Er_{aN} = p_a^N p_b[(X-N-1)p_b + 2]$

where p_a is frequency 'a', p_b frequency of 'b' and X is the total number of 'a's and 'b's.

For a linkage group with M markers and an average of k crossovers it follows that the average value of N is M/(k+1). X is the total number of intervals summed over all individuals in the whole population. Typically this will be about 100 individuals, so X is typically much larger than N. Er_{aN}/X is the proportion (or frequency P) of runs (r) of type 'a' of length N, and where X is much larger than N we can say:

$P = p_a^N p_b^2$
so:

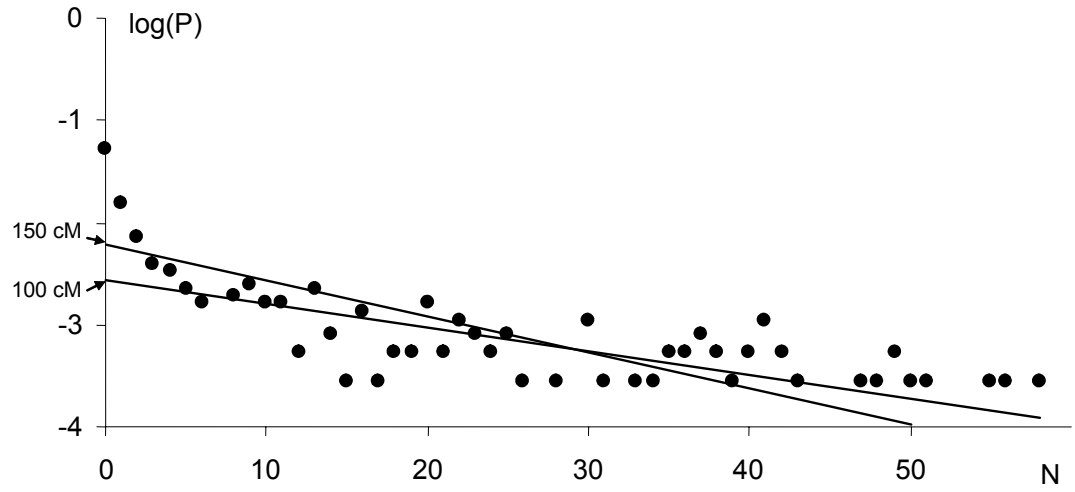
$\log(P) = 2 \log (p_b) + N \log (p_a)$

This means that a plot of the log of the frequency of a given run length plotted against run length (N) has a slope of $\log (p_a)$ and an intercept of $2 \log (p_b) = 2 \log (1-p_a)$.

A plot of this type is shown in the figure below. The scatter of points represents the observed distribution of lengths of non-recombinant intervals and the two lines plot the expected distribution for a linkage

[20] Mood A.M (1940) The distribution theory of runs. Annals Math Stat 11: 367-392
[21] Southern EM (1975) J.Long range periodicities in mouse satellite DNA. Molec. Biol. 94:51-69

group with this number of intervals where the map length is 100 or 150 cM respectively. We know from comparing genetic maps of several populations, and from cytogenetic observations, that the length of this linkage group should be about 100cM and certainly less than 150 cM.

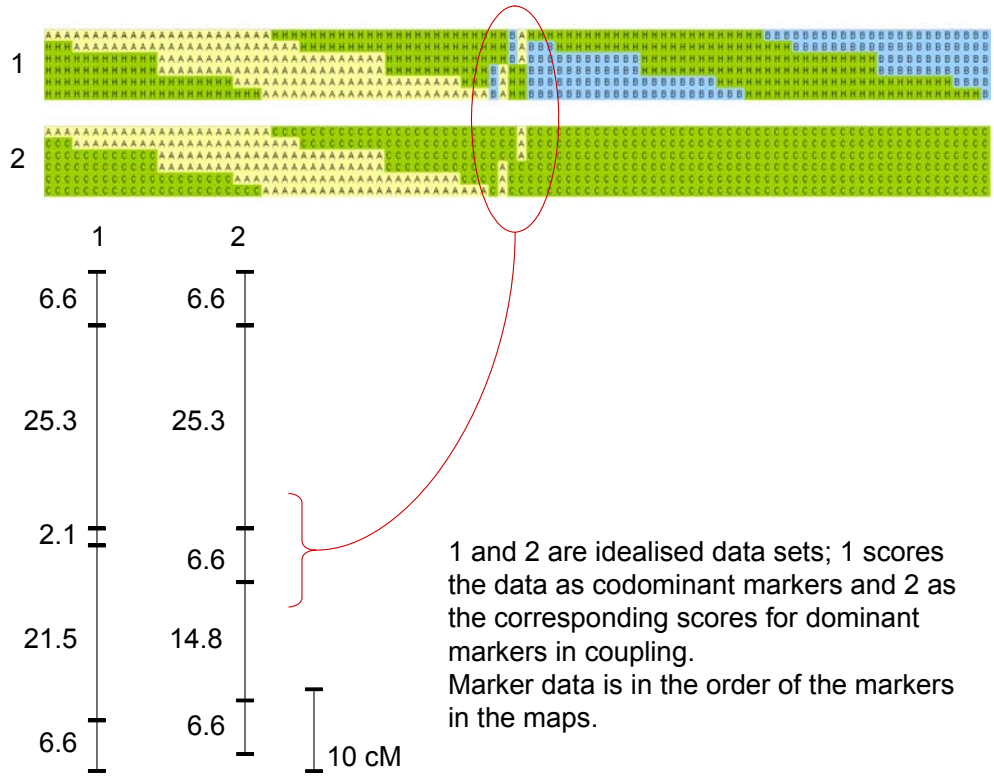


It is clear from the plot that there is an excess of the N=0 class, ie where there are recombination events in adjacent intervals, ie a 'singlet' or where an individual marker has a score different from its neighbours. In fact, for this data set, there seem to be about 10-fold more of these than expected, this is a strong indication of error. Interestingly the N=1 class is also significantly in excess. This is where two adjacent markers differ from those on either side, and is not expected from simple errors. The existence of this class suggests that something like the observations of Isidore et al (2003) [see section 8.1.1] is responsible for this excess.

8.1.3 Local order

The order of markers on a map can be related to the data set very simply by colouring the data scores in some simple scheme, and arranging the scores in the order of the markers along a linkage group. This has the grand name of a 'graphical genotype'. In the figure below an idealised data set is shown where the data is scored according to the scheme in section 3.4. In this case for the upper panel (1) the A scores are yellow, the B scores blue and the H scores green. In the lower panel (2) the data are exactly the same but with the condition that the score C is either B or H. The C scores are green and correspond to dominant markers score in coupling.
The corresponding linkage maps are illustrated as maps 1 and 2. There is some difference in the estimates of recombination distance because of the ambiguity from the dominant marker scores. Note that these can be either larger or smaller than the case for codominant scores.

The relative order of the two central markers is dependent on a very small number of individuals – four in the case of the codominant markers and just two for the dominant markers. A very small amount of mis-scoring could change the deduced relative order of these two markers, that is especially so for the dominant markers, even though the distance between these two markers appears greater.



This is a general property of close markers flanked by longer intervals. In QTL analysis, having the central 6.6 cM segment the wrong way round could distort the analysis and suggest that there are 2 QTL when in fact there is only one. An even distribution of markers may be more advantageous than maximising the marker number.

8.2. Recombination and segregation.

Genetic markers can be helpful in to specific tasks that are relevant to breeding: (i) fixing genotypes in early generations; and (ii) identifying recombinants that minimise linkage drag. This is de Vienne's '*management of recombinations*'.

8.2.1 Fixing genotypes in early generations

In an F2 population we expect 1 in 4 individuals to be homozygous for a single recessive allele. In practice breeding will often require the selection of genotypes that are homozygous for recessive alleles at multiple loci. This is de Vienne's accumulation of '*favourable alleles as quickly as possible in a single genotype*'. If we are interested in selecting homozygous recessives at *L* loci, then the population size

needs to be more than 4^L. In fact it needs to be about three times this size (see box).

If the chance of any one individual being what we want is p then:

The chance that :
one is not what we want is $1 - p$
none of N is what we want is $(1 - p)^N$
at least one of N is what we want is $1 - (1 - p)^N$

In this case we know $p = 1 / 4^L$ and we want $1 - (1 - p)^N > 0.95$

$1 - (1 - p)^N = 0.95$

$(1 - p)^N = 0.05$

$N\log(1 - p) = \log(0.05)$

$N = \log(0.05) / \log(1 - p)$

for L= 5, $p \sim 10^{-3}$, N ~ 3 000
for L= 10, $p \sim 10^{-6}$, N ~ 3 000 000

This means that finding an individual homozygous recessive at a significant number of loci requires huge population sizes. This is no surprise to a breeder, but what is of interest is that it may be possible, using molecular markers, to find individuals homozygous at a few loci that can be scored visually and then to find among these those that are heterozygous for the remaining desired loci. Thus in a two step process genotype construction is a possibility with a mixture of mass screening and marker analysis of a small number of selected individuals. Selection is replaced by genotype construction.

An interesting exercise is to consider whether there is sufficient recombination within a population for genotype construction.

8.2.2 Identifying recombinants that minimise linkage drag.
One problem with the strategy outlined above is that relatively few individuals are generated that have the desired combination of alleles. This means that the number of different recombination events that have generated the desired allelic combination is very small.

What fraction of F2 homozygotes for a recessive allele carry the entire parental chromosome of the donor of that allele? That's an unfair question, but suppose 1, 2 or 3 crossovers per meiosis and the answer can be derived for each case. Surprised? I was.

This means that if you are constructing a desired genotype form a wide cross (as for example in the introduction of a disease resistance trait) that you might well carry over a huge number of undesirable alleles at a range of loci not being monitored. This in turn has the likely consequence that a genotype constructed in this will perform poorly. For this reason alone it is hardly surprising that breeders choose to "cross the best with the best and hope for the best". What markers can do to help is twofold. First of all markers can be used to assess the scale of the damage done by introducing a recessive allele from a non-adapted background. Markers can tell you how far away from the target locus the crossovers actually were (if there were any). Secondly markers can be used to search for those individuals where recombination events were favourably located.

Deploying molecular markers does not help to generate the desired recombination events, but they do allow the development of a breeding strategy that reduces the number of plants to be examined to a manageable number.

9. Trait mapping

Much of what has been said above explains how trait mapping can be undertaken when the trait is a 'major gene' ie when there is a clear difference between the parental alleles. In this case the trait is simply another marker and can be mapped as such. This may seem unimportant – why bother if the difference is clear?

A good example is bread-making quality in wheat. This turns out to be a property of the alleles of the high molecular weight glutenins, so scoring these is a good predictor of bread-making quality. It is possible to select for this among F2 individuals individual even from half grains – but you can't make bread from this!

Having justified the idea that trait mapping might be of some value then there is some 'good tricks' to employ discussed briefly below.

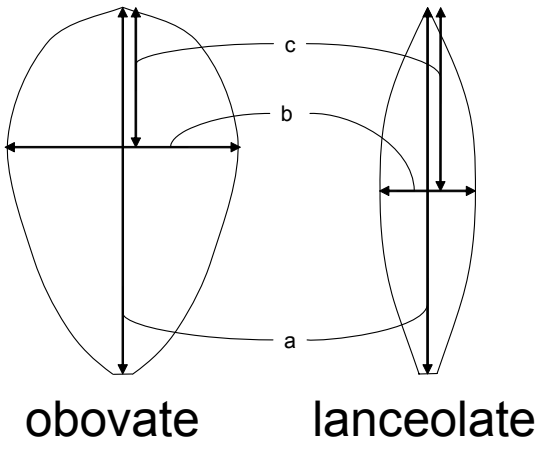
For example if we had a wheat recombinant inbred population where a major determinant of bread-making quality was segregating (and a lot of the variance was attributable to a single locus). Then we could determine the value of this parameter for each RIL. Now we can do several different things one is conventional or QTL mapping (section 10). Another is to identify the extreme types (the worst and best) we can then look for alleles that are more frequent in one than the other and thus identify markers linked to the determinant of the trait. This approach is useful if the difference is not large and there are multiple individuals of indifferent phenotype; ie that provide no information. The disadvantage of this approach is that there are few individuals in these extreme classes. Consequently we expect some differences between the groups by chance alone.

A commonly used approach is 'bulked segregant analysis'. This is common enough to have its own acronym and the procedure is explained by the name. Suppose we cross *TT* x *tt* and want to identify markers linked to *T* using an F2. We can prepare DNA from all the *tt* individuals (and from some that are either are Tt ot TT). We can make mixtures (bulks) of each type – usually bulks of 10 individuals. With a marker system (ideally a multiplex) we can look for markers present in the *TT* parent and *T_* bulk that are absent from the *tt* parent and *tt* bulk.
It is an interesting exercise to work out (roughly) how close the markers will be for a given number of bulks of 10.

10. QTL mapping

Approaches to QTL mapping are described in de Vienne's book, and these comments will not be reiterated here. However some general comments on the basis of QTL mapping follow.

QTL represent loci about which we know relatively little. It is probably easiest to think of them as weak alleles of single loci (but see section 2 – this is a contentious issue). In the discussion below the assumption is made that a QTL is identified by contrasting alleles of weak effect and that their consequence is best measured as a quantity.
For example leaves may be recognisably obovate or lanceolate, and these characters may segregate as discrete traits in some crosses. However, in a segregating population, there may be a range of intermediate leaf shapes between the two illustrated on the right. In this case it may be sensible to make some measurements such as the lengths 'a', 'b' and 'c' as illustrated. The values b/a and c/a give some description of leaf shape and represent quantitative traits.



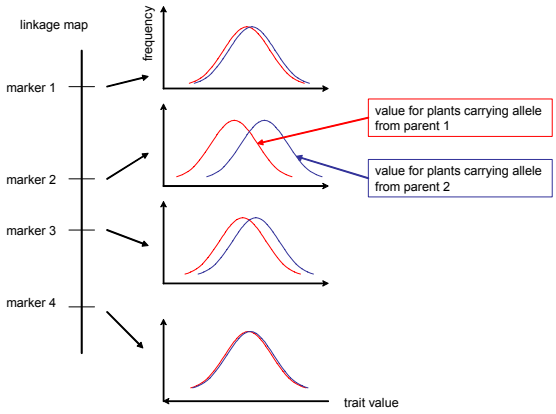
A QTL analysis would seek to explain the variation in these measures according to some genetic model. An important point to note is that a QTL analysis is not solely a measure of location of trait determinants, but will tell whether the growth determinants responsible for characteristics b/a or c/a are shared. That is, even if we are not particularly interested in determining a map location for these determinants, we might want to know whether they can vary independently. Perhaps this is not so important for leaf shape, but for characteristics such as seed aroma and susceptibility to insect

spoilage the issue may become important before launching a breeding strategy.

One simple way of looking at QTL mapping is to consider recombinant inbred or double haploid lines. These are all homozygous and can be used in replicated trials that can be very important for quantitative traits. For example it may be desired to measure field yield under drought stress, but the season may be wet. With RILs the experiment can be repeated, also multiple locations or randomized block trials can be undertaken. These are major advantages for the measurement of quantitative traits. If you look at de Vienne's book you will see that he points out some disadvantages that also need to be considered.

That said, let us consider a QTL analysis in an RI population. For each inbred line there is a measure of the trait quantity (v_i) and an associated error term. We can then, for each marker in turn, subdivide the population according to the two allelic states for that marker and ask whether there is a difference in the mean and variance of v for the two contrasting sub-populations. If there is statistical evidence for such a difference then we have evidence for the association of the marker and trait.

You have probably noticed that this single marker approach does not require a genetic map! That tells us that the single marker approach is missing some information.



Each assay is like one of the graphs in the illustration, but it is clear from the illustration that the way in which the frequency distributions change *along* the linkage map that there is a sense of position for the determinant of the trait. Probably the differences seen are the consequence of allelic differences at a single position. The order of the markers along a linkage group is a refinement to the model of single locus control, and a second approach – Interval Mapping – seeks to use the model of the linkage map to test where the genetic control of the trait determinant lies. Note that this uses information from marker scores, marker order and interval length. All of these may be error prone so it is important to be sure of the map and data scores. It is more important to have good coverage of the map and reliable data than to have large numbers of markers. These factors should be taken into account in assembling a data set to be analysed.

11. Disequilibrium and Association mapping.

11.1 Introduction

Linkage analysis has been very successful in plants. However, it has disadvantages. The accuracy with which QTL can be located depends, among other factors, on the number of meioses that are sampled. In populations of lines derived from F2s and backcrosses, unless large numbers of lines are grown, the number of meiosis can be quite small. As a result, particularly for traits of lower heritability and for smaller QTL effects, precision is often poor. If this problem is bad in plant genetics, it is dire in human genetics: in spite of massive effort, the number of confirmed QTL detected for complex traits (ie those which do not follow simple Mendelian inheritance) is very small. Recent effort in human genetics has, therefore, looked for alternative methods to locate genes. A method showing great promise is to exploit unobserved, historical recombinations through linkage disequilibrium mapping, also called association mapping. Briefly, this process relies on historical events, particularly mutations and population bottlenecks, to generate linkage disequilibrium (LD) between QTL and markers. Over successive generations, this LD declines at a rate proportional to the recombination fraction, such that in the present, strong marker-trait associations imply close proximity of the marker to a QTL.

This section describes the principles, methods and pitfalls of association mapping in more detail. First however, we need to establish some principles of population genetics to understand how LD originates and then decays.

11.2 Population genetics and linkage disequilibrium

11.2.1 Hardy-Weinberg equilbrium

Population genetics is the study of gene flow over time in populations. The founding principal of population genetics was established in 1908 and is now known after its co-discoverers as the Hardy-Weinberg equilibrium. Verbally, it can be stated as follows:

“The hereditary mechanism, of itself, does not change allele frequencies.”

An equivalent statement of the law would be:

“In the absence of mutation, selection and chance effects, no evolution will occur within a Mendelian population”.

These days, this law may seem self evident. When first describing it Hardy wrote: "*I am reluctant to intrude in a discussion concerning matters of which I have no expert knowledge, and I should have expected the very simple point which I wish to make to have been familiar to biologists.*"

Mathematically, the law amounts to saying that if the frequency of an allele in one generation is p, then the frequency in the next generation is also p.

In addition, for a diploid organism, if mating is also at random, genotype frequencies are also constant from generation to generation, with the values:

AA	Aa	aa
p^2	$2p(1-p)$	$(1-p)^2$

These genotype frequencies will not apply to a species like Lablab however which does not mate at random but is an inbreeder. However, if the rate of inbreeding is constant from generation to generation, genotypes frequencies will still be constant. Also, the more fundamental principle that allele frequencies are constant in the absence of any force acting to change them still applies.

11.2.2 Linkage disequilbrium

Now consider two biallelic loci in a randomly mating population, with alleles A and a at the first locus and alleles B and b at the second. Let the frequency of an allele be described as p_x, where x is one of A, a, B, c. There are four possible gamete types AB, Ab, aB and ab with frequencies p_{AB}, p_{Ab}, p_{aB} and p_{ab}. The frequencies of alleles and gametes at each loci can be set out in a contingency table:

	B	b	total
A	p_{AB}	p_{Ab}	p_A
a	p_{aB}	p_{ab}	$1-p_A$
total	p_B	$1-p_B$	1.0

From the Hardy-Weinberg law, we know that at equilibrium, p_A in one generation equals p_A in any other generation. The equivalent of the Hardy-Weinberg law for two loci is that at equilibrium:

$$p_{AB} = p_A \,.\, p_B$$

Equally, pAb = pA.(1- pB), and so on for the other gamete types.

In words, we can say that the frequency of a gamete type is the product of the frequency of the alleles carried on that gamete. This is called gametic phase equilibrium if the two loci are unlinked or linkage equilibrium (LD) if the two loci are on the same chromosome. Commonly, the term linkage disequilibrium is used to describe any sort of disequilibrium, whether the loci are linked or not. For linked loci, the gamete types (AB, Ab, aB, ab) are more generally referred to as haplotypes.

We define any departure from linkage equilibrium as:

$$D = p_{AB} - p_A \cdot p_B$$

Also

$$-D = p_{Ab} - p_A.(1-p_B)$$
$$-D = p_{aB} - (1-p_A).p_B$$
$$D = p_{ab} - (1-p_A).(1-p_B)$$

At equilibrium, D = 0.

D is called the coefficient of linkage disequilibrium. Verbally, it is the difference in frequency between actual gamete types and the expected frequency at equilibrium. Note there is no requirement for D to be positive. For example, a deficiency in AB gamete types would make D negative.

In the absence of equilibrium, frequencies of alleles and gametes can be written as follows:

	B	b	total
A	$p_{AB} + D$	$p_{Ab} - D$	p_A
a	$p_{aB} - D$	$p_{ab} + D$	$1-p_A$
total	p_B	$1-p_B$	1.0

In passing, note that a single parameter, D, is all that is required to account for the difference between actual and expected frequencies in this 2 x 2 contingency table. This illustrates why a 2 x 2 contingency chi-squared test has only 1 degree of freedom: it is testing the significance of a single parameter: D in this case.

Actual and equilibrium frequencies are compared side by side below.

	AB	Ab	aB	ab
At equilibrium	p_Ap_B	$p_A(1-p_B)$	$(1-p_A)p_B$	$(1-p_A)(1-p_B)$
No-equilibrium	$p_Ap_B +D$	$p_A(1-p_B) -D$	$(1-p_A)p_B -D$	$(1-p_A)(1-p_B) +D$

11.2.3 The interpretation of D

The coefficient of linkage disequilibrium, D, is hard to interpret. The values it can take are constrained by allele frequencies. A little trial an error will soon establish that certain combinations of allele frequency and D range will result in impossible negative haplotype frequencies. In fact, the maximum value of D, 0.25, is only possible at allele frequencies of $p_A = p_B = 0.5$. The minimum value, which also requires allele frequencies of 0.5 is -0.25. At other frequencies D must be greater than the maximum of $-p_Ap_B$ and $-(1-p_A)(1-p_B)$ and smaller than the minimum of $p_A-p_Ap_B$ and $p_B-p_Ap_B$.

To make D more easy to interpret two transformations are used to rescale it:
D' and |D'| and Δ^2 or r^2. These will be discussed after they are defined:

D' If D < 0, $D' = D/$ maximum $\{p_Ap_B, (1-p_A)(1-p_B)\}$

If D > 0, $D' = D/$ minimum $\{p_A(1-p_B), (1-p_A)p_B\}$

Δ^2 or r^2 $\Delta^2 = D^2/[p_Ap_B(1-p_A)(1-p_B)]$

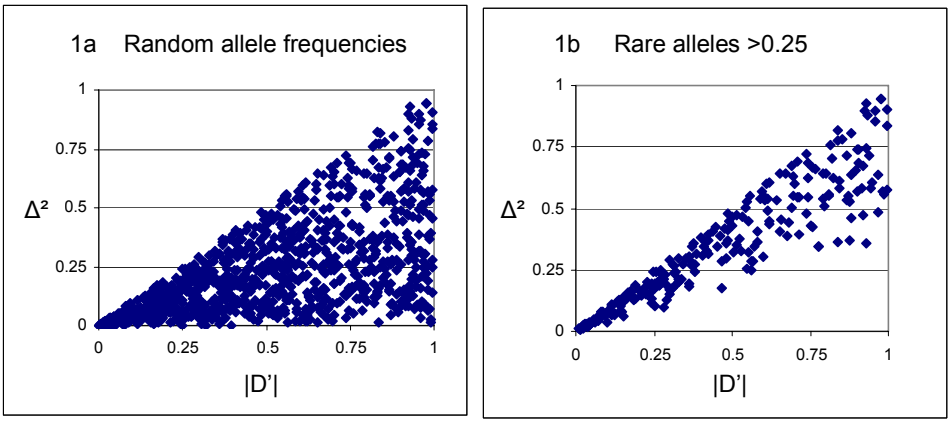
D' ranges from -1 to +1. Generally therefore, it is the absolute value of D' - |D'| that is quoted.

Δ^2 ranges from 0 to 1. It can be shown that it is the correlation coefficient between the two loci if the alleles are given numeric codes.

Both |D'| and Δ^2 have the advantage that they range from zero - representing perfect equilibrium to one - representing high linkage disequilibrium. |D'| will take a value of one when, of the four possible haplotypes, only three are observed. Δ^2 will take a value of one when, of the four possible haplotypes, only two are observed. In this case, the first locus is a perfect predictor of the second locus, and allele frequencies at the two loci will match. Since Δ^2 is a measure of predictability, it is useful for deciding appropriate marker densities and in studying the power of association to detect QTL. Figure 1a shows a plot of D' against Δ^2 for some simulated arbitrary values of p_A, p_B and D. Note that
for any value of |D'|, Δ^2 ranges from zero up to that value. Δ^2 is never greater than |D'|. |D'| is more likely to take high values at extreme allele frequencies. This effect can be seen more clearly in figure 1b, which plots the data from figure 1a after removing loci with allele frequencies less than 0.25. It can be seen that at intermediate allele frequencies, |D'| and Δ^2 measure much the same thing.

Figure 1

Comparison of measures of LD measures



11.2.4 The decay of linkage disequilibrium with time

Linkage disequilibrium is generally unstable: genetic recombination causes gamete frequencies to change towards their equilibrium values. Following random mating, in the absence of mutation, selection and chance effects - the same conditions required for Hardy-Weinberg equilibrium – the value of D in the next generation is:

$$D_1 = D_0 (1-\theta)$$

And therefore

$$D_t = D_0 (1-\theta)^t$$

θ is the recombination fraction between the two loci.
t is the number of generations since the start.

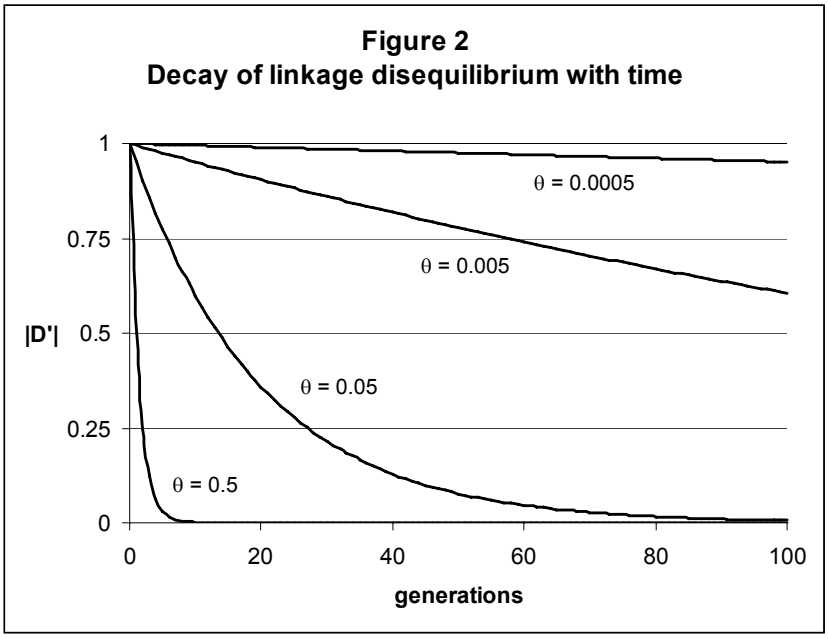
Linkage disequilibrium therefore decays quicker at higher recombination frequencies. For unlinked loci, the decay is at a rate of ½ per generation.

For close linkage and larger values of t, to a good approximation:

$$D_t \sim D_0 e^{-\theta t}$$

Thus recombination frequency and time are interchangeable – a halving of recombination fraction is compensated for by doubling the

number of generations. Figure 2 shows the decay in linkage disequilibrium over time at a series of recombination fractions.



Linkage disequilibrium decays very rapidly in the absence of linkage but persists for a very long time with very tight linkage. At $\theta = 0.0005$ the value is still 0.61 after 1000 generations.

11.2.5 The effect of inbreeding

In inbreeding species, the decay in linkage disequilibrium over time is reduced. In the most extreme case, if the population consists of a set of inbred lines with no intercrossing, there is no opportunity for recombination and linkage disequilibrium is fixed. If some outcrossing occurs however, linkage disequilibrium will decay although at a slower rate. The effect of inbreeding in pedigree breeding programmes is an interesting example. Here, because breeders are making crosses, there may be more recombination occurring in the domesticated crop than in nature. Assuming that all varieties are fully inbred, the formula for the rate of decay of LD

$$D_t = D_0 (1-\theta)^t$$

will still apply provided the definitions of θ and t are modified. t is no longer the generation time, but the cycle time: the time taken to produce a set of progeny lines from set of parents. θ is no longer the recombination fraction per generation, but the cumulative proportion of recombinants occurring from one cycle to another. This is $2r/(1+2r)$ where r is the true, generation-wise recombination rate. For closely

linked markers (<2 cM say), 2r/(1+2r) ~ 2r. With a cycle time of eight years (poor by current standards but reasonably accurate historically), the rate of decay of LD per generation is then roughly:

$$D_t \sim D_0 e^{-rt/4}$$

LD is decaying at about a quarter of the rate found in a truly randomly mating population with the same generation time. Of course this figure will be perturbed by the overlapping generation structure that breeders impose but it can act as a guide: in spite of the inbreeding nature of the many crop plants, LD is still expected to decay reasonably rapidly among cultivated varieties.

11.2.6 Linkage analysis and LD mapping compared

Linkage analysis, in its crudest form, locates a QTL by the increase in signal strength observed as markers and QTL get closer. This can be viewed as a test for the magnitude of D between the QTL and the marker locus. However, to locate QTL accurately in an F2 is difficult. Precision depends on how well we can detect differences in recombination fraction between QTL and adjacent markers. For example, with markers on top of the QTL ($\theta = 0$), 1 cM away ($\theta \sim 0.01$) and 10cM away ($\theta \sim 0.1$) the proportion of non-recombinant chromosomes is 1, 0.99 and 0.9 respectively. Detecting a difference in signal strength between these markers will require a large experiment unless the QTL is of large effect. Suppose now, that the F2 itself was randomly mated for 100 generations. The non recombinant chromosomes would be present at a frequency of 1, 0.36 and 0.0 respectively and we would be able to locate QTL quite precisely. Of course, carrying out 100 rounds of random mating in an F2 is impractical. However, in natural populations and non-experimental populations of crop plants, there will have been many rounds of recombination historically. If something in the past also generates linkage disequilibrium between QTL and marker loci, then mapping experiments could be carried out directly in non-experimental populations. We need, therefore, to consider the causes of linkage disequilibrium.

11.3 Causes of linkage disequilbrium

11.3.1 Mutation

Consider a single polymorphism with two alleles, A and a, segregating in any reasonably large population. Suppose a new mutation, B -> b say, occurs somewhere on a chromosome carrying the A allele. In the population as a whole there will be three haplotypes:

AB with a frequency very close to p_A
aB with a frequency very close to $1-p_A$
Ab the new mutant, carried on a single chromosome

There are four possible haplotypes in total, but only these three are observed, so |D'| = 1. In successive generations, assuming that the new b mutation is not lost from the population by drift but ultimately rises in frequency, the missing haplotype, ab, will be created by recombination. This can take a very long time for closely linked markers. For the majority of markers available for genotyping, mutation must have occurred a long time ago – many generations are required for allele frequencies to rise from a single copy to a frequency which makes genotyping worthwhile. The levels of linkage disequilibrium attributable to mutation will therefore only be high among very closely linked markers (or markers and QTL). Provided a sufficiently high marker density can be achieved, this situation is very favorable for association mapping.

In humans, it is common to find values of |D'| equal to 1 among very closely linked markers, often accompanied by high values of Δ^2. This indicates that little or no recombination has occurred among these markers. The pattern of LD in crop plants is less clear – there is less data available. Data is beginning to accumulate however. Among wild populations of *Arabidopsis*, an extensive survey has revealed that LD decays quickly – within 50 kb - even though this is an inbreeding species. (Nordborg et al. 2005).

11.3.2 Population bottlenecks, founder effects and drift.

A population bottleneck is an extreme reduction in population size. This might occur as a result of disease nearly wiping the population out, an environmental disaster or some other catastrophic event. A particular form of population bottleneck, a founder effect, occurs when a species colonizes a new niche or environment. Initially the population size can be extremely small – for a wild species only a few seeds might be carried to an island. For a crop species, only a few seeds or transplants may be introduced to establish the crop in a new country. Any restriction in population size will generate LD. An F2 can be regarded as an extreme case: the population is established from

two gametes a generation ago. As a result, levels of LD are at a maximum. However because linkage analysis occurs within a generation of the founding event, there has been little opportunity for LD to decay and it is hard to locate QTL accurately. Generally, the magnitude of LD generated by a bottleneck or founder effect is less extreme, but is still sufficient for association mapping. In crop plants, the activities of plant breeders themselves can result in population bottlenecks - the advent of a new disease or desired agronomic trait such as reduced height may result in a period of breeding in which only a small number of parental lines are used, or one or two lines are used very extensively.

In fact, any finite populations size generates some degree of LD, in the same way that genetic drift always causes some change in allele frequency, whatever the population size. For a population of constant size, a steady state is set up in which the expected value of Δ^2 is:

$$E(\Delta^2) = 1/(1+4N_e\theta)$$

N_e is the effective population size (not defined here). It is usually smaller than the actual populations size as a result of variability in true population size over generations and of variability in fertility from plant to plant.

11.3.3 Selection.

Selection on a trait will change allele frequencies at QTL determining the trait. In addition, allele frequencies will change at markers closely linked to the QTL. This is called hitchhiking. Its effect is to generate LD among markers around the region of selection. A region of increased LD, often accompanied by a reduced amount of polymorphism compared to other genomic regions, can be a signature of selection – a sign that a particular region has been subjected to selection pressure. Such regions have been identified in maize and *Arabidopsis.*

11.3.4 Migration and population admixture

If two populations, formerly isolated, are brought together, LD can be created. This is a result of allele frequency differences between the two source populations, which may have arisen through drift or through selection. For example:

haplotype	pop 1	pop 2	combined	expected difference
AB	0.04	0.64	0.34	0.25	0.09
Ab	0.16	0.16	0.16	0.25	-0.09
aB	0.16	0.16	0.16	0.25	-0.09

ab	0.64	0.04	0.34	0.35	0.09

In population 1, $p_A = 0.2$ and $p_B = 0.8$. In population 2, the frequencies are reversed. Within each population there is no linkage disequilibrium (for example $p_{aB} = p_a.p_B = 0.2 \times 0.8 = 0.16$ in population 1).

If the two populations are intermixed, without any crossing, the haplotype frequencies are just the average of the separate population frequencies. However, the allele frequencies are averaged too, such that $p_A=p_B=0.5$ and linkage disequilibrium is generated. In fact, D = 0.09, |D'| = 0.36 and $\Delta^2 = 0.13$.

With more modest rates of migration or gene flow from one population to another, the generation of disequilibrium is less severe. Provided migrants intermate with the host population, the disequilibrium will decay in successive generations.

Migration can be either an asset or a problem in association mapping. If population admixture is known to have occurred and if markers are available which discriminate, even imperfectly, between the two parental populations, then these markers can be used to map traits for which the populations differ. This is "admixture mapping". It is the population based equivalent of mapping in an F2: instead of two parental inbred lines, there are two parental populations. In human genetics there is considerable interest in this method, particularly in the USA: Afro-Americans are known to have about 10% European ancestry and are therefore a suitable group in which to map traits for which Africans and Europeans differ. In plants, there are no published accounts of admixture mapping, but suitable admixed populations may exist, for example in crosses between Flint and Dent maize or between Indica and Japonica rice.

Generally, migration is a problem. If we are trying to exploit linkage disequilibrium arising from mutation or an ancient bottleneck, recent migration introduces long range LD which can mask the marker-trait associations arising from close linkage which we wish to detect.

11.3.5 Summary

Linkage disequilibrium can arise from many causes. Current evidence shows that LD is generally higher between closely linked loci and that it declines with distance. However, instances of longer range LD do occur. There is therefore a major risk that associations between a QTL and a marker are not the result of close proximity but may arise from other causes which have not been taken into account. In practice, in any population, forces generating new LD and the decay of existing LD will both be occurring. Patterns of LD can therefore be complex. The

requirement for successful association mapping is to detect and correct for long range associations arising from recent events while locating close range LD arising from mutation and historical population bottlenecks.

Fortunately, recent research has supplied methods with help in this. These are introduced in the next section.

11.4 Experimental methods for association mapping

11.4.1 Association mapping in experimental populations

Linkage mapping and association mapping are not distinct methods. Both work by exploiting recombination. One relies on the detection of contemporary meioses in controlled crosses. The other relies on inferring historical meioses in uncontrolled populations.

In 1995, Darvasi and Soller proposed the Advanced Intercross as a means of improving the precision of QTL location. Here, F2 individuals are intermated for several generations before mapping. The successive rounds of recombination result in additional decay of LD. As a result, the precision of QTL location is increased. This approach has been extended to include populations set up with multiple parents and to take into account multipoint marker information (Mott et al. 2000) In mouse for example, a population originating with eight founder inbred lines has been intermated for at least 60 generations. This has resulted in individuals composed of very small tracts of chromosomes originating from multiple founders, which can be traced through a high density microsatellite map. In crop plants, the advantage of this approach is that a population can be set up containing lines which capture the majority of the variation available in the current breeding pool. Mapping would therefore be of direct relevance to lines derived from this pool. Although it would take time before such a population was sufficiently developed for mapping to start - probably around five years - the existence of such resources, which are cheap to set up and increase in value each generation, would be of great benefit to future mapping and breeding experiments.

11.4.2 Mapping in uncontrolled populations. I. The Transmission Disequilibrium Test

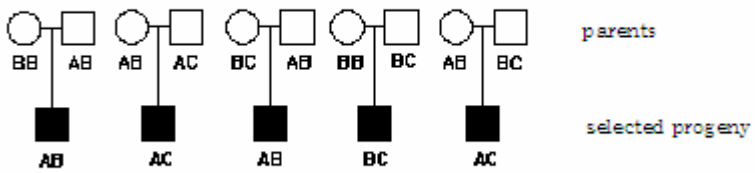
The ability to map QTL in uncontrolled or non-experimental populations - collections of breeders lines, old landraces, samples from natural populations - has great potential. Firstly, populations may exist in which LD decays more rapidly than in controlled crosses. Secondly, data sets may exist which have already been extensively phenotyped. This saves time and money. As we have seen, the

challenge is to distinguish QTL–marker associations arising from LD between closely linked markers from spurious background associations. The first and most robust method of achieving this was the transmission disequilibrium test, or TDT, introduced by Spielman et al. in 1993.

The TDT is a means of detecting linkage in the presence of association. Neither linkage alone, nor association alone, will generate a positive result. At its simplest, multiple families of the form given in figure 3 are collected.

Figure 3

The Transmission disequilibrium Test



Allele A is transmitted to offspring four times out of five.

The single progeny individual is usually selected for an extreme phenotype. In human genetics, for which the test was originally devised, this typically means they are affected by the disease under study. However, they can be selected for extreme values of any quantitative trait. Parents and progeny are genotyped, but only parents heterozygous for the marker under study are included in the analysis. From each parent, one allele must be transmitted to the progeny and one allele is not transmitted. Over all families, a tally is made of the number of transmissions and non-transmission. In the absence of linkage between QTL and marker, the ratio of transmission to non-transmission is expected to be 1:1. However, in the presence of linkage between the QTL and marker, this transmission ratio will be distorted from Mendelian expectation to an extent which depends on the strength of LD between the marker and QTL. This distortion is detected in a simple chi-squared test. The test can be regarded as a test for linkage which has increasing power with increasing LD. The power of the test also depends on how efficient selection of the extreme progeny has been in driving segregation at the QTL away from a 1:1 ratio.

This elegant test is extremely robust to the effects of population structure. It has been used extensively in human genetics and has been extended to study haplotype transmissions, quantitative traits and the use of information from extended pedigrees. It has two serious disadvantages. Firstly, it is extremely susceptible to an increase in false positive results generated by genotype error and biased allele calling. Secondly, a lot of phenotyping and genotyping effort is wasted

since only heterozygous parents are included in the analysis and three individuals are genotyped for each phenotype. As a result, its use in human genetics has declined over the last few years. Nevertheless, at some stage soon, its use in plant genetics will be published: it is already being used in the commercial sector.

Note that transmission is studied over a single generation. In plants, parental lines and progeny lines are usually separated by several generations. In this case the TDT is still a valid test, but is no longer completely robust: the process of pedigree breeding may introduce population structure.

11.4.3 Mapping in uncontrolled populations. II. Genomic control

Genomic control is a method of learning to live with background LD. It assumes that the effects which generate LD affect the whole genome rather than just specific regions. For example, migration or population admixture will generate LD between a trait and markers distributed over the whole genome. We can turn this process on its head. If we have sufficient markers distributed over the genome, we can detect population structure by studying the extent to which the distribution of the test statistic for association, estimated empirically from these markers, differs from the expected null distribution. To estimate the empirical distribution accurately would require many markers. However, it turns out that all we need to do is compare the observed mean test statistic with its expected value: 1.0 for a 1 degree of freedom chi-squared test. For this, we only need about 50 markers. Thus, if the average chi-squared at a set of 50 genome wide control markers is much greater than one, population structure is indicated.

Now assume that in addition to the these control markers, we also have one or more candidate loci which we wish to test for association with our trait. Our null-hypothesis is no longer that there is no association between trait and marker. Rather it is that there is no association over and above that expected as the result of population structure. To test this, we divide the observed chi-squared by the average value of chi-squared at the control markers. We can look up the p-value associated with the adjusted chi-squared in the usual manner.

$$\chi^2_{\text{genomic control}} = \frac{\chi^2_{\text{observed}}}{\Sigma\left(\chi^2_{\text{null markers}} / n\right)}$$

The procedure is illustrated in figures 4 and 5.

Figure 4

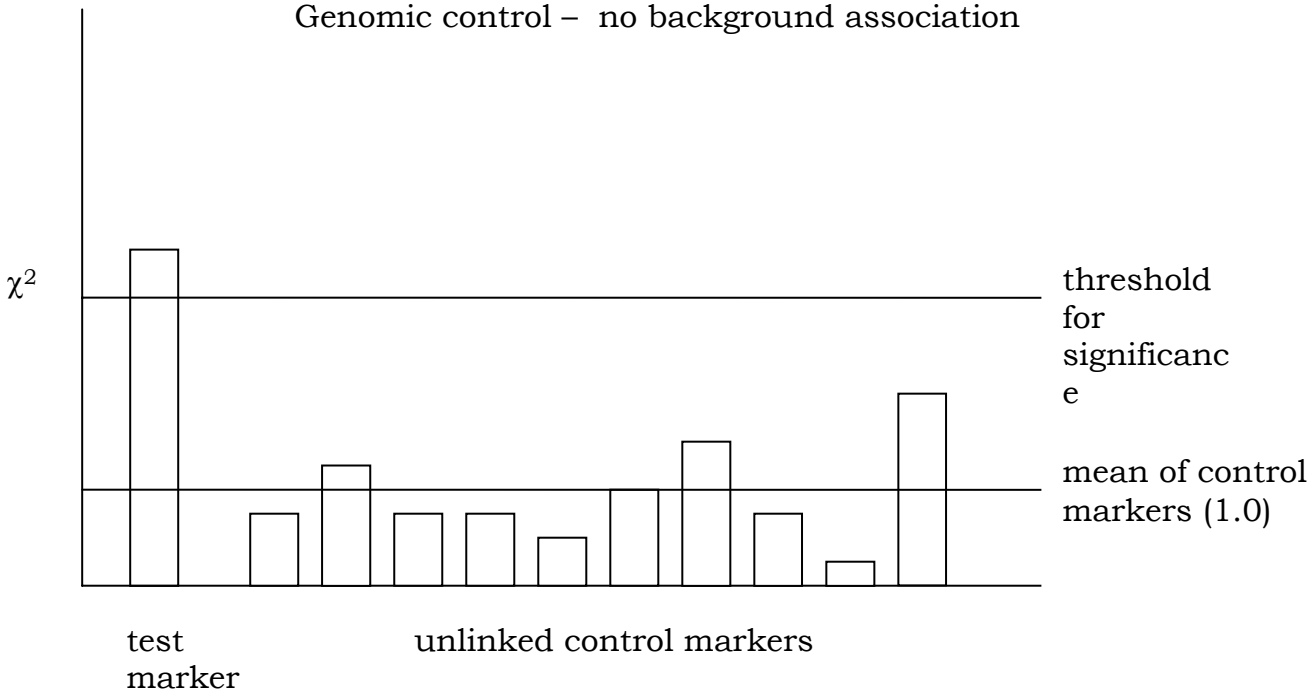
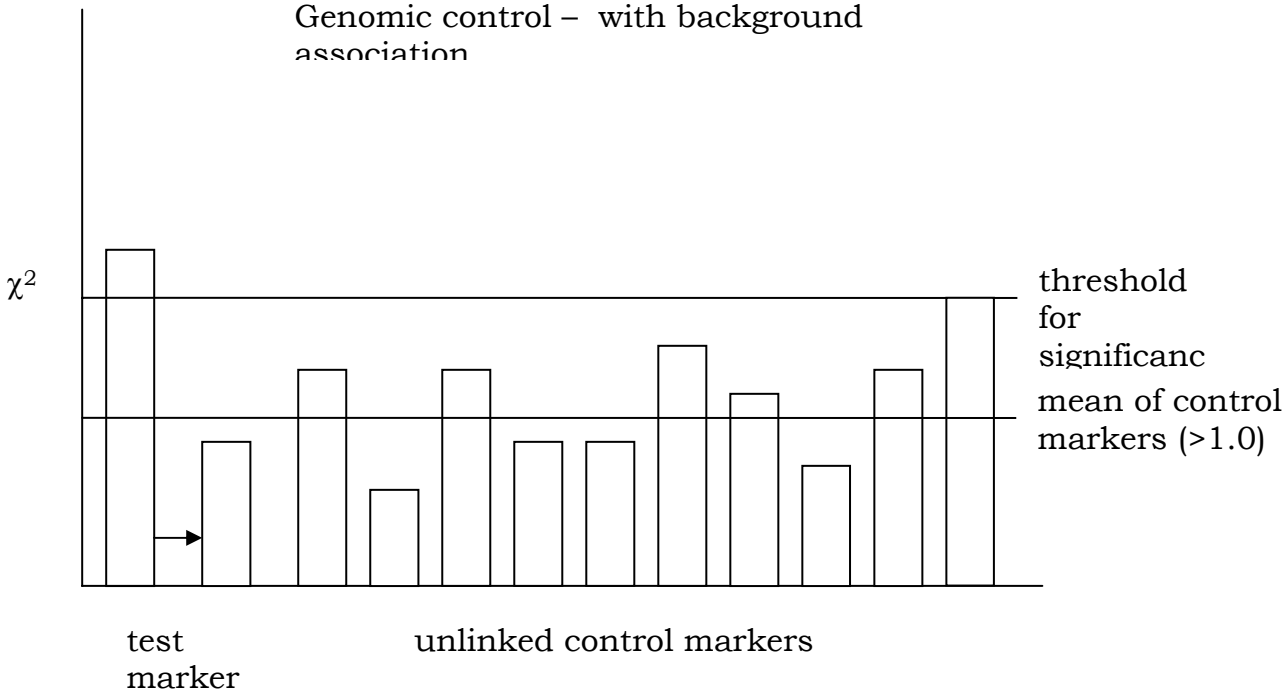


Figure 5



The proof that this remarkably simple adjustment is valid is difficult (Devlin and Roeder 1999, Bacanu et al. 2002). For quantitative traits,

the difference between trait means for each marker class is usually tested in a t-test. Provided the number of observation is reasonably large, if t is squared, it can be treated as a 1 degree of freedom chi-squared and the genomic control procedure can still be carried out. More recent work has suggested that greater accuracy is achieved by treating the test statistic as an F test with 1 df in the numerator and degrees of freedom in the denominator equal to the number of control loci (Devlin et al. 2004).

Genomic control is valid for any single degree of freedom test. Around 50 control markers are required. Preferably, these should loosely match the test marker in allele frequency. They should also be of a similar type – different types of markers may respond differently to the various forces that shape patterns of LD. It would not be wise, for example, to use microsatellites as control markers for SNPs.

More sophisticated versions of genomic control are also available too. With large numbers of candidate polymorphisms to test, the majority are not expected to be genuinely associated with the trait. In this case, procedures (invoking Bayesian statistics) and software are freely available to use the candidate markers as their own controls.

Example of genomic control

200 wheat varieties were scored for gain hardness and yield. A SNP at the *pinb* gene - a known major gene for grain hardness - was genotyped, together with 58 randomly distributed SSAP markers. Results are below:

	p-value at *pinb*	
trait	no genomic control	with genomic control
hardness	0.000	0.000
yield	0.723	0.865
lodging	0.008	0.053

Significance for grain hardness is detected with or without genomic control. The result for yield is relatively unaffected. The significant association with lodging, almost certainly a result of population structure, is lost after genomic control is applied.

11.4.4 Mapping in uncontrolled populations. III. Structured association

Structured association provides a sophisticated approach to detecting and controlling population structure. Again, additional markers are required which are randomly distributed across the genome. Just as for genomic control, the factors that generate population subdivision and structure are assumed to operate over the whole genome. In the

short term. these factors will have generated gametic and linkage disequilibrium among unlinked markers and loosely linked markers which has not yet had a chance to decay. In the simplest case, we may have unwittingly sampled individuals from several populations. In the absence of any intermating between populations, we would expect linkage equilibrium within these populations, but not across the whole dataset. If we knew how many populations were present, then by trial and error we could allocate individuals to them such that LD within populations was minimised. We could then carry out separate association tests within each population and pool the results. This is the approach that structured association takes. First, individuals are allocated to populations, then association testing is carried out within populations.

The allocation of individuals to populations must take into account two additional factors.

Firstly, we often do not know how many populations there are. However, reasonable estimates can frequently be attained: the allocation process can be repeated for different numbers and that which best fits the dataset can be selected. Thus it is possible to estimate the population number in addition to estimating to which population each individual belongs. Nevertheless, deciding on population number can be problematic: many workers report difficulties, and the manual accompanying the software acknowledges that the process can be somewhat heuristic.

Secondly, individuals from different populations have usually interbred, often several generations ago. The progeny of these hybrid individuals cannot be allocated to a single population. However, it is possible to estimate the proportion of ancestry attributable to each population, while still minimising the extend of LD within these populations.

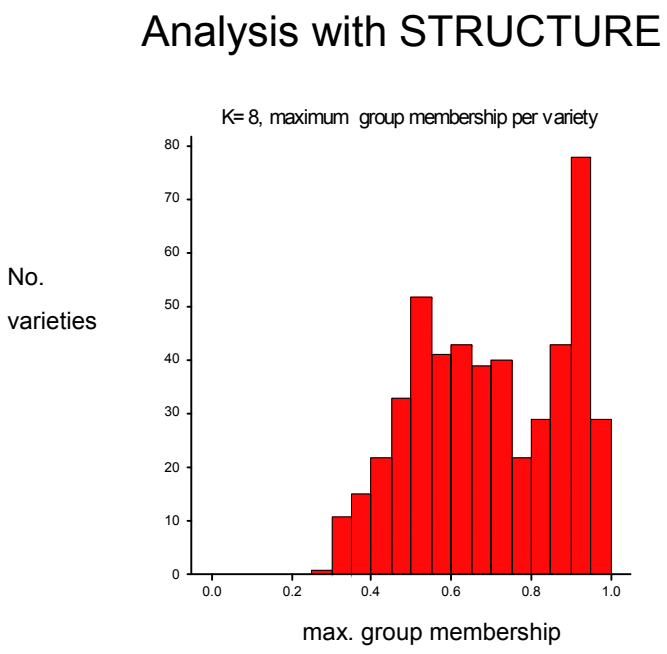
It is not possible to simply consider all possible partitions of individuals (and parts of individuals) into populations and pick the best. The computer program STRUCTURE uses sophisticated and computationally intensive methods to carry this process out. The software is easy to run, but care must be taken in setting the multiple input parameters required to run it. Running STRUCTURE will be an exercise for one of the practical sessions.

In practise, this method of allocating individuals to subpopulations has been enormously successful. The original paper describing the technique was published in 2000 and at the time of writing has been cited 565 (scholar.google.com). In crop plants, in addition to use in LD mapping, it is being used to study population subdivision within cultivated germplasm, processes in crop domestication, and relationships between landraces, wild species and modern varieties.

11.4.4.1 Example of STRUCTURE

500 old and modern European wheat varieties were sampled to study genetic variation within the group. As part of this study, 42 SSR markers were genotyped on all lines. Varieties were allocated into 10 subpopulations. Many lines showed high membership of a single population but most lines did not belong exclusively to a single subpopulation. Figure 6 is a histogram of maximum group membership. This bimodal distribution shows that roughly one third of the lines belong predominantly, but not exclusively, to one or other of the 10 populations, but two thirds of lines are highly admixed.

Figure 6



For some of the 500 lines, pedigree information was available. Among the lines with very high group membership, 46 parent-offspring pairs were identified. Group membership of these lines can be summarised as follows:

	Observed	Expected†
parent and progeny allocated to same subpopulation	34	5.8
parent and progeny allocated to different subpopulations	12	40.2

† Expected numbers are calculated on the basis of the estimated size of the subpopulations. When the number of subpopulations is large, it is more likely that a random pair of lines will belong to different populations rather than to the same population.

Many more parent-offspring pairs are allocated to the same population than expected. Predicted population membership is therefore genetic: it is transmitted from parent to offspring. This demonstrates that the method employed in STRUCTURE is working remarkably well in this dataset.

After individuals are allocated to populations, the association test takes place within populations only. For categorical traits such as disease resistance, or grain hardness, this can be carried out with the companion program to STRUCTURE: STRAT. For quantitative traits, there is currently no specific software available. However, the test can be carried out by logistic regression. Here, the genotype data are treated as the Y variable and the X variables are the predicted population memberships and the phenotype. To treat the genotype as a numeric variable, individuals carrying a selected allele are coded 1, and those not carrying it are coded 0. (For non-inbred diploids, the coding would be 0,1,2, with 1 representing the heterozygous class.)

During the model fitting process, the quantitative trait is fitted after the group membership effects. This process assesses the relationship between the quantitative trait and the marker, *after* adjusting the quantitative trait for any effects of group membership. (See the notes on R for the reasoning behind this approach.) The process of logistic regression is outlined in one of the practical sessions.

11.4.4.2 Some practical considerations

Association mapping in crop plants is a method whose use is beginning to increase. The resolving power of the method depends on how rapidly LD decays with distance. This will vary from species to species and from dataset to dataset. Patterns could be very different between populations of landraces, wild progenitors and modern cultivars. At the moment, there is very little data available in this area in plants: high densities of mapped markers and/or extensive sequence information is required to establish these patterns authoritatively. In human genetics, marker density is high enough, and decay of LD sufficiently rapid, that whole genome scans by LD mapping are now a reality. In crop plants, linkage analysis is

substantially easier than in humans and has been much more successful. However, the density of mapped markers in crop plants is quite low and genotyping costs are high. For these reasons it is hard to see LD mapping based whole genome scans of crop plants occurring in the foreseeable future. However, there is a niche for these methods in saturation genotyping of existing linkage regions. This will provide useful replication studies of these regions, often in adapted germplasm, and more importantly it is likely to increase the precision with which QTL can be located. By similar reasoning, methods in association genetics are ideal for candidate gene studies.

There are other advantages to using breeders lines or national collections of cultivars in association mapping experiments. Firstly, there is often a substantial collection of phenotypic data already in existence. Secondly, QTL discovery is carried out using germplasm of direct relevance to the crop, rather than in a sometimes esoteric F2 specifically set up for that purpose.

11.5 Analysis methods

The standard methods of analysing data for genetic association are very simple: the contingency chi-squared test for categorical traits and the t-test for quantitative traits. The preceding section discussed methods to control for population structure. In this section we mention some other problems of analysis, which have been glossed over so far.

11.5.1 Multiple alleles

In the absence of population structure, multiple alleles can be tested collectively for association by using a 2 x N contingency table for categorical traits and by an analysis of variance for quantitative traits. However, if many of the alleles are rare, these methods will loose power. A very rare allele adds one degree of freedom to the numerator of the test. It has virtually no chance of contributing towards a statistically significant result itself, but raises the threshold of the test statistic required for statistical significance. The test therefore looses power. For this reason, rare alleles should be pooled, or empirical significance tests should be used instead of looking up the p-value in tables.

A marker with multiple alleles cannot be tested in a t-test or a 2 x 2 contingency chi-squared test and so cannot simply be used in structured association or genomic control. The simplest method of testing is to recode a marker with N alleles into N pseudo-biallelic markers: where for each marker in turn, a pseudo marker is generated in which the two states are original-allele-present and original-allele-absent. These pseudo-markers can then be treated by

standard methods. Some form of adjustment for multiple testing is required, however.

11.5.2 Haplotype analysis

Analysis of haplotypes for association has two problems. Firstly, there are typically more than two haplotypes and we face the same problem, with the same solutions, as for multiple alleles. Secondly, it is often difficult to decide how to define haplotypes for analysis. For example, with six ordered SNPs, there are 20 contiguous combinations of SNPs with 124 possible haplotypes in total. This begins to present a serious multiple testing problem. There are many possible approaches to haplotype analysis. Within the scope of this manual, it is not possible to go into detail about them all. A brief overview of the most common approaches is given below. For the sake of brevity, we shall assume that all haplotypes are made up of biallelic SNPs.

1) Treat haplotypes as multiallelic single markers.

2) Ignore haplotypes and analyse the SNPs jointly in an analysis of variance. Interactions between SNPs can also be tested. These test for a haplotype effect over and above any effect attributable to the single SNPs. A simple example is given in the guide to R.

3) Assign evolutionary relationships to the haplotypes and take these into account during the analysis procedure. The more closely related the haplotypes, the more likely they are to share any QTL alleles, so the more disequilibrium they should display. For small numbers of SNPs, with no recombination, these sorts of patterns can be established by eye: a simple measure of the distance between two haplotypes is the number of mutations by which the two differ. Software is available which implements these types of analyses.

4) Look at patterns of linkage disequilibrium among the SNPs to see if there are contiguous blocks among which little or no recombination has taken place. Software exists to identify these blocks, but they are also often easily identified by creating a matrix of |D'| or Δ^2 in Excel and applying some conditional formatting. SNPs within each block are then analysed as an independent set.

5) Reduce the number of SNPs required for inclusion in a haplotype analysis by selecting tagging SNPS (tSNPs). The strength of LD can be so high that many SNPs in a region, often over half, are redundant in the sense that their

genotype can be accurately predicted on the basis of linear combinations of the remaining SNPs. Equally, haplotypes are often adequately tagged by only a few of the SNPs they carry. Software exists to identify these tagging SNPs, for example SNPtagger http://www.well.ox.ac.uk/~xiayi/haplotype/index.html. The identification of tagging SNPs can be used to reduce the number of SNPs in an analysis and so potentially increase power. It can also be used to save on genotyping costs: if tagging SNPs are identified in an initial survey of variation there is no need to genotype all SNPs in the future.

Clearly there are many possible approaches and methods, and research in this area is continuing. Choice, to some extent, depends on personal preference and familiarity. In any analysis however, it is recommended that the SNP by SNP analysis is included, that LD patterns among the SNPs are studied and that haplotype frequencies are estimated before any more sophisticated complex analyses is attempted.

11.5.3 Effects of allele frequency

Linkage disequilibrium generally declines with distance. However, even if a region is saturated with markers at a very close spacing, there is no guarantee that a QTL will be detected: we must also take into account allele frequencies at markers and QTL. If all our markers have allele frequencies close to 0.5, but the QTL is at a frequency of 0.1, the power to detect association will be reduced compared to the optimum case where allele frequencies at markers and QTL match. When selecting SNPs therefore, it is wise to keep an eye on the distribution of allele frequencies. To cover a broad spectrum of allele frequencies at the QTL, it is possible that a haplotype analysis or the use of multiallelic markers would be advantageous. This is not certain however, but depends on the evolutionary history of the region. For example, the high mutation rate of SSRs may make them ineffective in detecting ancient polymorphism at QTL.

11.6 Results in practice in crop plants

The first use of linkage disequilibrium to map a QTL in an unstructured population was in maize (Thornsberry et al. 2001). This group used structured association to confirm that polymorphisms in the *dwarf8* gene were associated with flowering time in a set of 92 maize inbred lines. 141 SSR markers were used to partition these lines into three subpopulations.

In subsequent years, association genetics approaches have been carried out, or advocated, in a range of crops including potato (Simko

2004), grass (Skøt et al. 2005), conifers (Neale & Savolainen 2004), barley (Kraakman et al. 2005) and wheat (Breseghello & Sorrells 2005). The wheat and barley papers report successful use of collections of cultivated varieties for mapping.

All these papers have relied on structured association. No published work, to date, has used genomic control. Although it is difficult to argue with success, there may be problems in applying this method to crop plants, in particular to collections of cultivars. In samples of such lines, it is possible that the predominant cause of population structure is the complex interrelationship between descendant and ancestral lines, both of which may be present. In this case, it is less clear that STRUCTURE is the correct tool for the job. Under these circumstances, genomic control may be a more effective method. However, to the best of our knowledge, the only group to have used genomic control as a method works within a commercial company, and they are not publishing their results. There are clearly many opportunities for research and development in this exciting area of crop genetics.

11.7 Appendix – software and resources

11.7.1 Software

The software for structured association is at:
http://pritch.bsd.uchicago.edu/software.html

The software for genomic control (written in R) is at:
http://wpicr.wpic.pitt.edu/WPICCompGen/genomic_control/genomic_control.htm

Software for multiparent advanced intercross lines:
http://www.well.ox.ac.uk/~rmott/happy.html

SNPtagger: haplotype tagging software:
http://www.well.ox.ac.uk/~xiayi/haplotype/index.html

A vast array of software for genetic analysis is available from
http://linkage.rockefeller.edu/soft/

11.7.2 General references available for free downloading

For anyone wishing to know more about population genetics, an excellent text book by J Felsenstein is available at:
http://evolution.genetics.washington.edu/pgbook/pgbook.html

The Hardy paper which started it all is available on-line from:

http://www.esp.org/foundations/genetics/classical/hardy.pdf

A good review of linkage disequilibrium in plants is:
Flint-Garcia SA, Thornsberry JM, Buckler ES 4th. (2004) Structure of linkage disequilibrium in plants. *Annu Rev Plant Biol.* **54**:357-74.
http://www.maizegenetics.net/publications/LDinPlants.pdf

A recent stunning review of the genetic structure of a model organism which illustrates what can be achieved with extensive molecular data is:

Nordborg M, Hu TH, Ishino Y, Jhaveri J, Toomajian C, Zheng H, Bakker E[2], Calabrese P, Gladstone J, Goyal R[1], Jakobsson M, Kim S, Morozov Y, Padhukasahasram B, Plagnol V, Rosenberg NA, Shah C, Wall JD, Wang J, Zhao K, Kalbfleisch T, Schulz V, Kreitman M, Bergelson J (2005)
The Pattern of Polymorphism in *Arabidopsis thaliana. PLoS Biol* **3(7)**: e196
http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1135296

Genomic control

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

Reich DA, Goldstein DB (2001), Detecting association in a case-control study while correcting for population stratification. *Genetic Epidemiology,* **20**:4-16

Bacanu S-A, Devlin B, Roeder K (2002), Association studies for quantitative traits in structured populations, *Genetic Epidemiology*, **22**:78-93

Devlin B, Bacanu S-A, Roeder K (2004), Genomic control in the extreme, Nature Genetics, **36**:1129-1130

Structured association

Pritchard JK, Stephens M, Donnelly P (2000), Inference of population structure using multilocus genotype data, *Genetics* **155**: 945-959

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000), Association mapping in structured populations, *American Journal of Human Genetics*
67:170-181

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**:1567-1587

Advanced Intercross lines

Darvasi A, Soller M (1995) Advanced Intercross Lines, an Experimental Population for Fine Genetic Mapping *Genetics* **141**:1199-1207

Mott R, Talbot CJ, Turri MG, Collins AC, Flint J (2000) A method for fine mapping quantitative trait loci in outbred animal stocks *PNAS* **97**: 12649-12654

Mott R, Flint J (2002) Simultaneous Detection and Fine Mapping of Quantitative Trait Loci in Mice Using Heterogeneous Stocks *Genetics* **160**:1609-1618

Admixture mapping

Smith MW, O'Brein SJ (2005) Mapping by admixture linkage disequilibrium: advances, limitiations and guidelines. *Nature Reviews in Genetics* **6:** 623-632
Transmission disequilibrium test

Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM) *American Journal of Human Genetics* **52**:506-516

Applications in crop plants

Breseghello F, Sorrells ME (2005) Association Mapping of Kernal Size and Milling Quality in Wheat (Triticum aestivum L.) Cultivars. *Genetics* [Epub ahead of print]

Kraakman AT, Niks RE, Van den Berg PM, Stam P, Van Eeuwijk FA (2004) Linkage disequilibrium mapping of yield and yield stability in modern spring barley cultivars. *Genetics* **168**:435-46.

Neale DB, Savolainen O (2004) Association genetics of complex traits in conifers.*Trends Plant Sci.* **9**:325-30

Simko I (2004) One potato, two potato: haplotype association mapping in autotetraploids *Trends in Plant Science* **9**:441-448

Skøt L, Humphreys MO, Armstead I, Heywood S, Skøt KP, Sanderson R, Thomas ID, Chorlton KH, Sackville Hamilton NR (2005) An

association mapping approach to identify flowering time genes in natural populations of Lolium perenne (L.) *Molecular Breeding* **15**:233-245

Thornsberry JM, Goodman MM, Doebley J, Kresovich S, Nielsen D, Buckler ES (2001) Dwarf8 polymorphisms associate with variation in flowering time. *Nat Genet.* **28:**286-9.

12. Appendices

Appendix 1: CTAB: An alternative method for DNA preparation

Genomic DNA extraction by the CTAB (Cetyltrimethylammonium bromide) method from the Laboratory of Molecular Biology
see:
Murray, M.G. and Thompson, W.F. (1980). Rapid isolation of high molecular weight plant DNA. Nucleic Acids Research. 8: 4321-4325.

Required Solutions and Preparation:

2 x CTAB buffer (5 0ml)
2 % w/v CTAB Powder (=1 g)
100 mM 5.0 ml 1 M Tris-HCl pH8
20 mM 2.0 ml of 0.5 M EDTA
1.4 M 14.0 ml of 5 M NaCl
Make the solution up to 50 ml with water

Also have avalable:

Add 4 ml of 2-Mercaptoethanol to each 1ml of 2xCTAB buffer required
Chloroform
Isopropanol
1 M $MgCl_2$ and 3 M NaOAc
70 % EtOH
1 x TE
3 eppendorf tubes per sample
Set water bath to 65°C

Protocol

1. Place 3 medium-large Arabidopsis rosette leaves in an eppen tube
2. Homogenise leaves in the eppendorf tube
3. Add 500 µl of 2 x CTAB/Mercaptoethanol buffer, keep on ice between samples
4. Incubate samples in 65°C water bath for 1 hour
5. Add 500 µl of Chloroform
6. Mix samples by inverting for 5 minutes at room temperature
7. Centrifuge for 10 minutes at room temperature (15,000 rpm)
8. Collect clear upper layer (400 µl) in a fresh eppendorf tube
9. Add 250 µl of isopropanol and invert to mix
10. Incubate at room temperature for at least 10 minutes
11. Centrifuge for 10 minutes at room temperature (15,000 rpm)
12. Discard supernatant (use yellow tips)
13. Add 320 µl of 1 x TE and put samples on ice
14. Ensure pellet is fully re-suspended
15. Add 40 µl of 1 M $MgCl_2$ and invert to mix

16. Incubate on ice for at least 10 minutes
17. Centrifuge for 10 minutes at 4°C (15,000 rpm)
18. Collect supernatant in fresh eppendorf tube
19. Add 40 µl of NaOAc
20 Add 250 µl of isopropanol
21. Incubate for at least 15 minutes at room temperature
22 Centrifuge for 10 minutes at room temperature (15,000 rpm)
23. Discard supernatant (use blue tips)
24 Add 1 ml of 70 % Ethanol and rotate tube to rinse pellet
25 Centrifuge for 5 minutes at room temperature (15,000 rpm)
26. Discard supernatant (use blue tips)
27 Quick spin at room temperature
28. Discard remaining supernatant (use yellow tips)
29. Vacuum dry for 3 minutes
30. Add 50 µl of 1 x TE and leave at room temperature for 10 minutes
31. Store DNA samples at 4°C.

An alternative DNA preparation procedure

1 Collect ca. 10-20g of fresh tissue. (Leaves are easiest.)

2 Freeze this tissue in liquid nitrogen. It is best to keep the tissue cold until frozen. (In the mortar use just enough liquid nitrogen to keep the leaves frozen without making the mortar too cold.)

Note: Liquid nitrogen poses problems because it is very cold, can expand explosively and excess nitrogen can lead to asphyxiation. In this method the main risk arises when collecting liquid nitrogen. The metal pipes delivering the liquid are extremeny cold and there is a risk of severe frostbite, especially if skin sticks to the cold metal: **Wear Gloves and a Face Mask**
When cooling mortars by adding liquid nitrogen there is a risk that the mortar may crack. These may also become extremely cold generating a risk of frostbite.
If the leaves have been frozen in a plastic centrifuge bottle, beware tha this may contain liquid nitrogen which will expand greatly on warming to room temperature: ensure that any expanding gas can be vented from the bottle.

3 Grind the frozen tissue in liquid nitrogen to a fine powder.

Wear Gloves

4 Allow to warm slightly before adding at least 1 ml of extraction buffer [1x EB] per 1 g of tissue and mixing thoroughly. Do this as the colour changes at the edge i.e. gets a little darker. Some older methods used 3x SSC, this buffer when mixed with ethanol generates an oily orecipitate that can interfere with the ability to pellet a DNA precipitate (but does not cause a problem wilt spooling the DNA precipitate)

5 Add 0.1ml of 20% SDS per 10ml to give 0.2% final w/v (of added buffer) and mix by grinding.

6 Transfer to an lidded centrifuge tube tube and add ca. 15ml of chloroform/amyl alcohol (24:1) per 10g of tissue and mix thoroughly. (ie about an equal volume)

It is best to transfer the aqueous mixture to a centrifuge and collect these on ice. When you have suficient to centrifuge take these to the fume hood for the addition of chloroform amyl alchol: Beware Chloroform is carcinogenic and should never be exposed to the air in the lab except within a fume hood. **Wear Gloves and eye protection.**

7 Centrifuge at 4K for 10mins at room temperature (Denley bench top centrifuge).

Balance the tubes by volume, do this for both phases.

8 The interface may be thick enough to allow the upper, aqueous, phase to be easily decanted off. Otherwise remove the aqueous layer carefully with a pasteur pipette.

Do this in the fume hood with a spill tray.
Dispose of the Chloroform/isoamyl alchol waste as a chlorinated solvent.
Leave chloroform contaminated glassware / plasticware to dry in the fume hood prior to disposal.

9 Layer 2 volumes of 100% ethanol onto the decanted aqueous phase.

10 Spool out DNA, (or pellet the precipitate as at 7) dry, and redissolve in a minimal volume (eg 0.5 - 1.0 ml) of 1xTE. Sometimes the DNA does not stick to the glass or plastic rod, so you may need to spin it.

11 Phenol extract with 0.5ml phenol, then take off the top aqueous layer. Occassionally I have known the phenol layer to be on top.

Wear Gloves and eye protection. Keep the phenol stock double contained. It is a good idea to collect roughly the volume you will need in a wide topped vessel (eg a beaker); this minimizes the risk of knocking over the container when distributing the phenol.
Note where the PEG mix for swabbing small phenol burns is located before you start. Check that there are other people around. If you use phenol after norml working hours, note where others are working lest you should need help.
Dispose of phenol waste in a designated container, and phenol contaminated plastic as a separate item for independent disposal.

12 Add 2 volumes of 100% ethanol and spool out the DNA, or precipitate by inversion.

13 Wash in 70% ethanol and dry thoroughly.

14 Redissolve in TE, and store at 4°C or frozen at -20°C

1 x EB is: 500 mM NaCl, 100 mM Tris pH 8.0, 50 mM EDTA (pH 8.0), 10 mM β Mercapto-ethanol (stock is 14M).

Prepare EB in the fume hood.

Appendix 2: Customer Developed FTA® Protocol

FTA® Technology

FTA Cards are impregnated with a patented chemical formula that lyses cell membranes and denatures proteins on contact. Nucleic acids are physically entrapped, immobilised and stabilised for storage at room temperature.
FTA Cards protect nucleic acids from nucleases, oxidation, UV damage and microbial and fungal attack.
Infectious pathogens in samples applied to FTA Cards are rendered inactive on contact. Samples collected on
FTA Cards and enclosed in a multi-barrier pouch can be shipped through the post making them an extremely useful tool for field collection of blood, plants or other specimens.
Indicating FTA Cards turn from pink to white on sample application and are recommended for clear or colourless samples. CloneSaver™ Cards are optimised for the room temperature collection and storage of plasmid DNA.

Handling Instructions

· Always wear gloves when handling FTA or CloneSaver Cards to avoid contamination of the Cards.
· Store unused FTA or CloneSaver Cards in a cool, dry place (avoid light and excessive humidity).
· Follow universal precautions when working with biological samples.
· FTA or CloneSaver Cards are non-toxic to humans.

Materials Required

· Whatman FTA Card – Indicating FTA Cards are recommended for use with clear samples. If applied to non-Indicating Cards, circle the application spot with a ballpoint pen or pencil.
· Whatman FTA purification reagent (cat no. WB120204).
· T 0.1 E buffer (10mM Tris-HCl, 0.1mM EDTA, pH 8.0).
· 2.0mm diameter Harris micropunch or other paper punch.
· Proteinase K (10 mg/ml).
· Ligation buffer (ATP 1 µL + Adaptors *EcoR*I and *Mse*1 + 1 unit of T4 ligase).
· **Restriction enzymes of choice.**
· Restriction ligation buffer (BSA + Water).
· 2.0mL micro centrifuge tube with spin basket insert (e.g. Whatman VectaSpin™).
· Micro centrifuge capable of speeds up to 12,000 x g.
· 37°C and 60°C incubator.
· Whatman FTA Protocol BD09 “Removing a Sample Disc from an FTA or CloneSaver Card for Analysis”.

The FTA Principle – Get it Right First Time, Every Time

FTA works by lysis of cells releasing the nucleic acid within the matrix of the Card, where the nucleic
acid will be entrapped among the cellulose fibres. Therefore the key step to ensure success is getting
DNA-containing cells into the FTA in the presence of moisture to activate the cell-lytic and DNAprotective chemicals.

Controls

It is recommended that internal standard controls are used during each analysis, these should include
the following:
· Negative control.
· Positive control of a known DNA standard

solution.

Whatman FTA Protocol BC01

Processing Protocol for Downstream AFLP Analysis of Sample DNA on an FTA® Card

Preparing an FTA Disc for DNA Analysis

1. Take two sample discs from the dried spot following the instructions provided in the protocol entitled “Removing a Sample Disc from an FTA or CloneSaver Card for Analysis”, protocol number BD09. For plant samples a 2.0mm disc is recommended.
2. Place discs in the bottom of a spin basket insert.
3. Place the spin basket into a 2.0mL micro centrifuge tube.
4. Add 500µL FTA purification reagent to the basket.
5. Incubate for 1 minute at room temperature.
6. Centrifuge at 6000 x g for 30 seconds.
7. Remove the basket and decant used reagent.
8. Return the basket to the micro centrifuge tube and repeat Steps 4-7 twice for a total of 3 washes with the FTA purification reagent.
9. Add 495µL FTA purification reagent and 5µL Proteinase K (10mg/mL) to the FTA discs in the basket.
10. Incubate for 1 hour at 60°C to remove residual Histones.
11. Centrifuge at 6000 x g for 30 seconds.
12. Remove the basket and decant used reagent/buffer.
13. Return the basket to the micro centrifuge tube.
14. Add 500µl of TE^{-1} buffer to the tube.
15. Centrifuge at 6000 x g for 30 seconds.
16. Remove the basket and decant used buffer.
17. Return the basket to the micro centrifuge tube.
18. Repeat steps 14-17 twice for a total of 3 washes with TE^{-1} buffer.
19. Transfer discs to a new tube.
20. Ensure that all the liquid has been removed before performing analysis.

Restriction Step

21. Add 2 restriction enzymes to the two FTA washed discs, a frequent cutter and a rare cutter (e.g. *EcoR*I and *Mse*1).
22. Add 4 units of each in restriction ligation buffer (BSA + Water) for a final volume 30µL.
23. Incubate for 1 hour at 37°C.
24. Mix solution by pipette.

Adapter Ligation Step

25. Draw off 30µL of restriction buffer leaving the discs behind.
26. Add to a fresh tube and add 10µL of ligation buffer (ATP 1µL, Adaptors EcoRI, MSE1 and 1 unit of T4 ligase).
27. Incubate for 3 hours or overnight at 37°C.

Preamplification Step

28. As per your standard protocols.

Selective Amplification Step

29. As per your standard protocols.

Technical Help

If you experience any problems with this protocol or wish to obtain additional information please contact Whatman Technical Service Team on the following regional numbers. Alternatively, please visit www.whatman.com for additional product information and further contact details.
India +91 22 529 7035 – ask for technical service

For Additional Protocol Information Please Visit
www.whatman.com

Appendix 3: PCR-based marker primer sequences

Wang et al 2004	Number	Accession	Origin_EST	Repeat	Size (bp)	Sequence FORWARD 5' to 3'	Sequence REVERSE 5' to 3'
Marker type SSR	1	AL370549	Medicago	AC(11)	**198**	CGTCCCGATATCGTCAACTT	CCACCACGACACATGTTACC
	2	**BF650979**	Medicago	AT(28)	**179**	TTGTGGAAGGAACAACTCTGG	GAAACCGGCATGATTAAGAC
	3	**BF647899**	Medicago	AT(34)	**266**	CTGTCAACAAGGGGTTAGGTG	TGCATCTACACCCAAAACAA
	4	**AQ842128**	Medicago	TA(23)	**209**	TCAATGCTGATGCCATTTTC	TCGCGTATTATAGCACAACA
	5	AI94357	Medicago	TC(25)	**183**	TCTCAATTCCCCAACTTGCT	TCTCCTTCACCCATCTTTGC
	6	**AW256794**	Medicago	TC(17)	**192**	GTCATCGAAGGCCAAAACAC	GTTTGCGAGAAACACCGATT
	7	**AA660488**	Medicago	TC(19)	**194**	TTGCATTATTTTCCTTTTTGACC	AACCCACAACCCAAAAATCA
	8	**AW584539**	Medicago	ACA(8)	**204**	TTGATGGGCAATACATGTCG	GTTGAAGGAAGGTGGTGGT
	9	**AW586959**	Medicago	ACA(10)	**222**	CGAGAATCATCGTAATTGGACA	CGAAGTTCAATGGCATCAGA
	10	**AW775229**	Medicago	AGC(8)	**222**	TACTGGGGTGATGCAAGACA	CAATACCCAGAGGAGCAGCT
	11	**BF005356**	Medicago	AGG(8)	**193**	CTTCAATTGTCACCGCCTCT	CTTATCTCGTCGTCCTCATC
	12	BF649209	Medicago	CCA(7)	**200**	AAGAGGCGGAGAGTGAGGTT	GGTAAGAGAACGAGCGAGG
	13	AW684360	Medicago	CGA(8)	**183**	TGTCATGGCGTCTCAAACC	CCTAACGCAGGAGAAGGAG
	14	**AW685679**	Medicago	GCC(5)	**197**	ACCTCACCTCACCTCCCTTT	GATCATCTGGGTTTCGCAAG
	15	AI974841	Medicago	TCT(11)	**169**	TCACCACCAAACCCCAAC	TGGCAATGCTACAAGCCTAA
	16	**AW688216**	Medicago	AGTG(9)	**211**	CACGAGGGATTGTTGTTTGA	GGAGCAGTAGGGTTGCATCT
	17	**AW127626**	Medicago	GTTT(7)	**191**	CATTTTGAAGGAAGGAAGAAGG	ATTTGGAAGCGGAATGTGAA
	18	**AW688861**	Medicago	CAACT(7)	**195**	TTGTTGTGTGGCTTCTTTGG	AAACCAACCACCTGTGTTGA
	19	AG81	Soybean	AG(13)	**105**	ATTTTCCAACTCGAATTGACC	TCATCAATCTCGACAAAGAA
	20	**AW186493**	Soybean	CTT(13)	**219**	GCGGTGATCCGTGAGATG	GCGGAAAGTAGCACCAAGAC
	21	**GMENOD2B**	Soybean	AT(17)	**164**	TAGGCAAAAGACTAAAAGAGTA	GCATGTCATTTTGATTGA
	22	AG48	Soybean	AG(18)		CAGAAACCTGAAATCTTCACC	CTTGGGTTTTTTTATGGGTTC
	23	**AG50b**	Soybean	AG(19)		ATAAATTGGAAGATGTGTTGGC	TACTGATGTGGATTCTCCCA
	24	AG93	Soybean	AG(17)		TCCATGCATGTATACTCCACC	TCATATGCCACAGGTTTTGT
	25	**BE347343**	Soybean	GA(18)	**244**	GCGCAAGCACTGAATGTCA	GCGTCACTAACACCTATAAC
	26	SoyPRP1	Soybean	ATT(20)	**141**	CGTGCCAAATTACATCA	TGATGGGAACAAGTACATAA
	27	**AF186183**	Soybean	ATT(22)	**204**	GCGTATTTTGGGGGATTTTGAACA	GCGTTTCTCTTCTTATTCTTT
	28	AW277661	Soybean	ATT(23)	**247**	GCGCATGGAGCATCATCTTCATA	GCGAGAAAACCCAATCTTTA

	29		**GMABAB**	Soybean	ATT(25)	**159**	CAAAACATAAAAAAGGTGAGA	AAGAACCACACTAATATTATT
	30		BE801128	Soybean	CAA(13)	**172**	GCGACAGTTCTCCACTCTTC	GCGCCCCTTATAGATTTGTA
	31		**AW508247**	Soybean	CTT(10)	**153**	GCGCCCAATCCCAATCTCAC	GCGAAGCCAATAAATGATAA
	32		AM620774	Soybean	CTT(9)	**152**	GCGATTTCCCCTCTTACTC	GCGAAAAACCAAGTTC
Choi et al 2004 Marker type Intron-directed	33	DK224R	AQ917190	Medicago		**110**	CGAAACAATAATCACAAAACAAATCAG	ATCTTGTTTATATGTGTTTGT
	34	DK225L	AQ917191	Medicago		**270**	TGTCCTTGCTTCTTATCCTTCCTTCA	AGCAGCACAACAACTTACAA
	35	DK302L	AQ917144	Medicago		**230**	GCATGGAAATAGTTTGGGTTAGTAGTTAGT	CTGATAAATGCATATTTTCAA
	36	DK353L	AQ917298	Medicago		**330**	CCATGCCATGGAAGGGTGTTT	GCAAGAACCAGATACCCTTG
	37	DK369R	AQ917327	Medicago		**310**	GGAACGTGGAGTTGTTGATGGTATTAT	GATGTAAAAACCTTTACACTT
	38	DK379L	AQ917338	Vigna radiata		**410**	AGCTTGTTGAGGTGGAAGGAAGTC	GTGTGTATGAGTGTCGTAAG
	39	DK413L	AQ917375	Medicago		**200**	TGATTGACCCCTGCTTTGATGCT	GTCAGGTTTGTTGTTGTTTT
	40	DK427R	AQ917398	Medicago		**500**	CCAAACAAGGAAAAGTGTTGGTGTCA	ATGAGAAACTTTTGAAATTTA
SRS sequences of Lablab Marker type Intron-directed	41	MET_1	AB176566	Lablab		**380**	TGT CTG GCT GTG GGT GTG G	AGA GCT TTT GAA CTT GTA
	42	MET_2	AB176567	Lablab		**250**	AAT GTC TTG CTG CGG TGG	AGC TCA CTT GCA AGT ACA
	43	fril	AF067417	Lablab		**310**	TAC AGT GCT TCC TGA ATG GG	ACA AAC AAC ATA CAA GTA
	44	5SrDNA	AY583516	Lablab		**380**	CGT GTG TTG AGA GGG AGG G	AGA ACA AGC TCG TGG GAA
	45	pDLL_1	AY049046	Lablab		**250**	CTT CAT GCT ACT TTT TCT TCT GGG	ACA AAC ACA TTG TGC AGG
	46	pDLT_1	AY049047	Lablab		**310**	ATG GTG GTG TTA AAG GTG TGC	TGC AAG GTT CGT AGC AGA

The 192 primer sequences from DJ Kim (2005 largely based on *Vigna unguiculata* sequence are available electronically.

Appendix 4: PAGE – PolyAcrylamide Gel Electrophoresis - preparation:

A 4.5% polyacrylamide stock solution can be prepared in advance and stored at 4°C (see Buffer List). The glass plates need to be thoroughly cleaned and silanised before pouring the polyacrylamide gel.

Plate preparation:

Both of the plates are cleaned using detergent, rinsed with de-ionised water and dried with paper towels. If you are preparing plates for the first time[22] decide on the working side, i.e. the side that will come into contact with the gel and keep to this by marking the outer surface with a diamond marker or some sticky tape.
Repelcote silane treatment: take the larger plate to the fumehood, tip a small amount of Repelcote onto a paper towel and spread, working in even sweeps across the whole plate. Leave for a few minutes to dry. Give an ethanol wipe over the whole surface to evenly spread the silane and a second ethanol wipe to thoroughly polish the plate; give a final polish with a dry piece of paper towel. Rinse with de-ionised water and dry with paper towels.
Bind silane treatment: prepare this solution as instructed by the company (See Buffer & Solutions section). The smaller of the two plates will be given this treatment. Dispense 500 µl of the working solution into the middle of the plate and spread evenly over the whole surface, working in even sweeps. Allow to dry. This can be done on a work bench; as this is less volatile compared to the repel silane a fumehood is not necessary. (Note that fresh gloves need to be worn each time so that there is no cross contamination of silane solutions).
Ensure spacers are clean and dry. Align to the edges of the larger plate and place the smaller plate on top (take care not to touch the working side with gloves). Square off the corners and secure both plates together at the sides with thermostable tape at 2 to 3 places down each edge.

Pouring the Polyacrylamide gel and preparation for sample loading:

Dispense 60 ml of 4.5% polyacrylamide gel mix from the chilled stock (see Buffer List) into a pouring bottle, add 30 µl of chilled TEMED (N,N,N'N' – Tetramethyl ethylenenediamine) and 400 µl of chilled 10% ammonium persulphate. Mix by inverting the bottle gently.
Open the nozzle and allow the polyacrylamide to flow slowly and gently between the plates; you have about 10 minutes (depends on ambient temperature) before the acrylamide begins to polymerise.

[22] New plates need to be seasoned - this means silanising the plates and pouring/setting polyacrylamide gel a couple of times before using with important reactions.

Place the comb into position immediately after pouring the gel; ensure that air bubbles are not trapped; clip the plates at the comb end:
Choice of comb:
a castellated well forming comb is put in place; it is removed once the gel has polymerised and the pre-formed wells flushed free of urea just before pre-running the gel and before loading samples
a sharkstooth comb is placed into position with the straight edge in contact with the gel during polymerisation. This comb is inverted to form the wells and left in place while loading samples and running the gel.
Allow the gel to polymerise for 1.5 hours. While the gel sets it is important to start to prepare for silver staining by making the **DEVELOPER** as follows: **Dissolve 60 g sodium carbonate (make sure it's anhydrous and not too old) in 2 litre distilled water and refrigerate at 4°C.**
Rinse the well/comb region carefully under a flow of de-ionised water, and rinse away the gel on the outside of the plate.
Remove the comb by sliding it out horizontally, continue to rinse the well area.
Dry the plates and mount onto the electrophoresis unit, the longer plate outermost; pour 500 ml of 1 x TBE buffer (See Buffer List) into the top reservoir; flush out the well area using a syringe to remove any loose acrylamide. This is very important as acrylamide stuck along the edge of the smaller plate will interfere with sample loading.
If using a sharkstooth comb place it into position so that the points of the comb just touch the gel, placing it level with the edge of the longer plate and fill the bottom reservoir with 500 ml of 1 x TBE buffer.

IMPORTANT SAFETY NOTE: **If you have poured a sub-standard polyacrylamide gel, i.e. has too many bubbles, then leave the gel to polymerise before discarding the acrylamide**
Sample loading:

Load 3 µl of each sample (keep these on ice while loading) take care not to touch/move the comb.
Run at 1500 - 1600 V for 1 to 2 hours until the darker blue (bromophenol blue) dye runs off.
Unmount the gel/plates, remove the tape; using the plastic Wonder Wedge to separate the two plates while still hot. The gel should remain in complete contact with the smaller, bind silaned, plate.

Appendix 5: Silver staining of the gel:

Make up the fixer: (10 % acetic acid) 1.8 litre distilled water and 200 ml glacial acetic acid in 2 litre pot. Pour into a shallow tray and immerse the gel/plate to fix for half an hour, shaking gently. Meanwhile make the silver stain solution: 12 ml 1.01 N silver nitrate solution in 2 litre of distilled water; add 3 ml formaldehyde (40 % solution) and mix.

Tip the fixer back into the pot (and save); wash the gel 3 times in fresh distilled water. Carry out the 4th rinse on a shaker for approximately 10 minutes or until "greasiness" has gone from the gel.

Pour off the rinse water and add the silver stain solution. Leave on a shaker for half an hour.

Set up a tray containing approx 3 litre of distilled water, have a timer and a piece of A4 white paper in a clear plastic bag close at hand. Immediately prior to developing the gel add 300 µl of sodium thiosulphate solution (0.1001 N) and 3 ml of formaldehyde (40 % solution) to the pre-chilled sodium carbonate solution. Pour this developing solution into a tray.

*****THE NEXT FEW STEPS HAVE TO BE FOLLOWED QUICKLY AND CAREFULLY SO MAKE SURE YOU HAVE EVERYTHING SET UP READY*****

Remove the gel from the silver stain and rest it ***ON*** the tray containing the 3 litres of water (do not put it into the water yet). Tip the stain back into its pot; it can be used for up to five more gels. Rinse remnants of stain from the tray with de-ionised water.

Set a timer for 10 seconds. Start the timer and quickly lower the gel into the water. Agitate several times so that all excess silver stain is removed from the gel surface. When 10 seconds is up, ***QUICKLY*** drain the gel and place it in the developing solution. Initially, tip the tray to ensure developer covers the gel evenly; then as the larger fragments begin to develop immerse up and down holding the top end of the plate. This way the middle / smaller sized fragments remain in the developer longer. The pre-cooling of the sodium carbonate also greatly slows down the rate of development.

Use the piece of white paper to check for the progress of band development. Stop the reaction when bands near the bottom of the gel start to show (i.e. 70 bp marker on the Ladder), by adding the 2 litre of fixer saved from earlier. Agitate the gel plate vigorously until bubbling ceases.

Soak the gel in a tray of distilled water for 10 mins and leave to dry overnight, standing vertically.

Photocopy or scan the image on the plate. Alternatively, use duplicating film as follows: in the dark room, under safe light, place a sheet of duplicating film on the bench with the grey/emulsion side uppermost. Position the plate so that the gel is in contact with the film. Turn on an anglepoise lamp held at an arm's length above the film and move lamp around for 7 to 9 seconds (timing depends on how dark the gel is from the staining/developing). Turn off the lamp and develop the film as normal using developer and fixer; rinse the film in water, hang up and allow to air dry.

Appendix 6: PAGE for SSCP gels

This is PAGE in non-denaturing conditions, ie. no urea and low voltage during gel running that extends for more than 12 hours, generally overnight.

Glass plates are prepared exactly as for denaturing PAGE (Appendix 4); but the gel mix is different:

SSCP gel mix, components for 60 ml:

MDE gel mix	15 ml	(Cambrex Bio Science, Rockland)
10 x TBE	3.6 ml	
SDW	41.4	

Mix together then just before pouring the gel add 240 µl of 10 % APS and 24 µl of TEMED and pour gel immediately and insert comb. Polymerisation is about 1 hr.

Denature PCR samples as normal, and place on ice. Load 5 - 8 µl for each sample. Run the gel for 15 -18 hr at 4 Watts, this depends on band sizes, and is preferrably run in a cold room or at least in an air conditioned room.

To visualise bands silver staining can be carried out as normal (see Appendix 5).

Appendix 7: Table of Chi-square statistics

df	P = 0.05	P = 0.01	P = 0.001
1	3.84	6.64	10.83
2	5.99	9.21	13.82
3	7.82	11.35	16.27
4	9.49	13.28	18.47
5	11.07	15.09	20.52
6	12.59	16.81	22.46
7	14.07	18.48	24.32
8	15.51	20.09	26.13
9	16.92	21.67	27.88
10	18.31	23.21	29.59
11	19.68	24.73	31.26
12	21.03	26.22	32.91
13	22.36	27.69	34.53
14	23.69	29.14	36.12
15	25.00	30.58	37.70
16	26.30	32.00	39.25
17	27.59	33.41	40.79
18	28.87	34.81	42.31
19	30.14	36.19	43.82
20	31.41	37.57	45.32
21	32.67	38.93	46.80
22	33.92	40.29	48.27
23	35.17	41.64	49.73
24	36.42	42.98	51.18
25	37.65	44.31	52.62
26	38.89	45.64	54.05
27	40.11	46.96	55.48
28	41.34	48.28	56.89
29	42.56	49.59	58.30
30	43.77	50.89	59.70
31	44.99	52.19	61.10
32	46.19	53.49	62.49
33	47.40	54.78	63.87

34	48.60	56.06	65.25
35	49.80	57.34	66.62
36	51.00	58.62	67.99
37	52.19	59.89	69.35
38	53.38	61.16	70.71
39	54.57	62.43	72.06
40	55.76	63.69	73.41
41	56.94	64.95	74.75
42	58.12	66.21	76.09
43	59.30	67.46	77.42
44	60.48	68.71	78.75
45	61.66	69.96	80.08
46	62.83	71.20	81.40
47	64.00	72.44	82.72
48	65.17	73.68	84.03
49	66.34	74.92	85.35
50	67.51	76.15	86.66
51	68.67	77.39	87.97
52	69.83	78.62	89.27
53	70.99	79.84	90.57
54	72.15	81.07	91.88
55	73.31	82.29	93.17
56	74.47	83.52	94.47
57	75.62	84.73	95.75
58	76.78	85.95	97.03
59	77.93	87.17	98.34
60	79.08	88.38	99.62
61	80.23	89.59	100.88
62	81.38	90.80	102.15
63	82.53	92.01	103.46
64	83.68	93.22	104.72
65	84.82	94.42	105.97
66	85.97	95.63	107.26
67	87.11	96.83	108.54
68	88.25	98.03	109.79
69	89.39	99.23	111.06

70	90.53	100.42	112.31
71	91.67	101.62	113.56
72	92.81	102.82	114.84
73	93.95	104.01	116.08
74	95.08	105.20	117.35
75	96.22	106.39	118.60
76	97.35	107.58	119.85
77	98.49	108.77	121.11
78	99.62	109.96	122.36
79	100.75	111.15	123.60
80	101.88	112.33	124.84
81	103.01	113.51	126.09
82	104.14	114.70	127.33
83	105.27	115.88	128.57
84	106.40	117.06	129.80
85	107.52	118.24	131.04
86	108.65	119.41	132.28
87	109.77	120.59	133.51
88	110.90	121.77	134.74
89	112.02	122.94	135.96
90	113.15	124.12	137.19
91	114.27	125.29	138.45
92	115.39	126.46	139.66
93	116.51	127.63	140.90
94	117.63	128.80	142.12
95	118.75	129.97	143.32
96	119.87	131.14	144.55
97	120.99	132.31	145.78
98	122.11	133.47	146.99
99	123.23	134.64	148.21
100	124.34	135.81	149.48

Appendix 8: Flowchart to run JoinMap v3:

Starting a new project with Joinmap.exe

File... New Project

A window opens give a filename, eg. test, 2 files are created: one is a folder called test.jmd; the other is the JM icon test.jmp

File.....Prepare data

A window asks for the data file, i.e. the scores .txt file in JM format

It also asks for a name to a .loc file; this file will be read by JM from now on

File ....Load data

The .loc file is loaded

In the left hand information panel a yellow icon with P and a file name appears

The right hand panel has tabs for data analysis:

Data tab has the .loc file
Locus genotype frequency gives the segregation of each marker, the Chi square and its significance. This is a very useful list and can be copied and pasted into Excel for segregation ratio analysis. Markers with extreme segregation distortion can be eliminated from the mapping data set.
It is useful to take time to analyse all of the data before proceeding to the mapping; after each selection hit the **Calculate** button.

Options.....Calculate options

The window offers the chance to alter factors, eg. choice of Haldane or Kosambi mapping functions.

LOD groupings (tree) tab and Calculate

A tree of markers grouped by LOD score appears in the middle window, this can be extended and lengthened by clicking in the +/- squares.
Choose the groups to map by highlighting to pink, one at a time, deselecting once completed.

Population...Create groups for mapping

Choose an icon highlight to pink

Group....Calculate map

Stick diagram maps are produced as well as a log file of linkage data

Finally after proceeding through each group, select all the Map icons you wish to include and select Map Join next to File in main menu. All linkage groups are displayed and can be printed off.

Opening an existing project

File.....Open project

JM icon .jmp

Appendix 9: Demonstation of Genomic Control in association analysis:

This exercise will guide you through the analysis of a small dataset for marker-trait association, adjusting for population structure by the method of genomic control. All analyses are carried out in R, although they are sufficiently simple that they could also be carried out in Excel.

The dataset we are using is part of the Gediflux project, also used in the demonstration of the program STRUCTURE, and present here in the file "GC demo.xls". There are 156 wheat varieties, two quantitative traits (protein content and endosperm texture), one categorical trait (grain hardness), one candidate SNP (pinb_a – as used in the STRUCTURE example) and 46 SSAP control markers. These SSAPs have been selected to have a minor allele frequency greater than 0.1. Note that there is a modest amount of missing data, and that this has been coded "NA" in preparation for input into R. Note also that grain hardness and all genotype data has been coded 0/1.

Data input and quality checking

Read the data in R. Either export data from Excel to your own text file or use the file "gc.txt".

```
gcdata<-read.table("gc.txt")
attach(gcdata)
summary(gcdata)
```

Remember you may need to point R at the directory containing the file first.

It is worth summarising and plotting the trait data – use the graphical and summary commands available in R. Note the relationship between endosperm texture and grain hardness. (Hint – try `boxplot`). Note also one extreme value for protein content. Ordinarily, this would require checking and consideration given to removing it from the analyses. Here we shall include it.

Test for significance of any association between pinb_a and the three quantitative traits using a t test and chisq test as appropriate.

```
t.test(Protein_Content~Pinb_a)
t.test(Endosperm_Texture~Pinb_a)
chisq.test(hardness,Pinb_a)
```

Note that chisq.test does not require that a table is first made from hardness and Pinb, for example:

```
chisq.test(table(hardness,Pinb_a))
```

However, it is worth creating and displaying this table, since it will explain the warning message that R provides in response to chisq.test statement.

Genomic control – Endosperm texture

To carry out genomic control we need to carry out each of these statistical tests on every SSAP marker, then calculate the average value.

This could be done by:

```
t.test(Endosperm_Texture~SSAP1)
t.test(Endosperm_Texture~SSAP3)
t.test(Endosperm_Texture~SSAP4)
```

etc., followed by writing down the test statistic, then calculating their average. However, a little more advanced R can ease the workload.

Try:

```
attributes(t.test(Endosperm_Texture~SSAP4))
```

R works on "objects". The results of an analysis are one sort of object. A dataset is another type of object. These objects have attributes and the command "`attributes`" allows us to see what they are.

Note the first attribute of our t.test is "statistic". The value associated with this can be displayed as:

```
t.test(Endosperm_Texture~SSAP4)$statistic
```

In general, `object$attibute` will access the value(s) associated with that attribute.

What this allows us to do is display the value of the t test statistic without the rest of the information which `t.table` provides. It is worth experimenting with some of the attributes of `t.table` to see what other information can be accessed directly.

We now wish to accumulate values for each t test in turn. Try this:

```
> t_result<-t.test(Endosperm_Texture~SSAP1)$statistic
> t_result[2]<-t.test(Endosperm_Texture~SSAP3)$statistic
> t_result[3]<-t.test(Endosperm_Texture~SSAP4)$statistic
> t_result
```

The final command should display the test statistic values for each of the three t tests. Note that the first call is
`t_result<-t.test……`

and not

`t_result[1]<-t.test……`

The first call creates a new data structure `t_result`. The second call assumes the data structure `t_result` already exists and replaces the current value, whatever that may be, with a new value. If `t_result` does not already exist, an error is generated.

Once `t_result` is in existence, however, `t_result[x]<-value` will append (or replace) the x^{th} value. Interestingly, if there are three values in `t_result`, then

`t_result[100]<-0`

will create 96 missing data values and then add 0 as the value of the 100^{th} entry: try it.

We now wish to increment `t_result` automatically, without editing the index numbers and the SSAP names.

First:

`attributes(gcdata)`

This gives row and column numbers and corresponding names to our data table. The rows, columns and single cells of the data table can be accessed directly, either using their index numbers or index names. For example:

```
gcdata["93","Endosperm_Texture"]
gcdata[6,2]
```

should return the same value. (Note this provides a simple method for changing or deleting (set to `NA`) individual values of our data from within R – for example we could remove our outlying value for protein content.

From the call to `attributes(gcdata)` we can see that columns 5-50 contain the SSAP data.

Verify that

```
t.test(Endosperm_Texture~SSAP4)$statistic
```

and

```
t.test(gcdata[,2]~gcdata[,7])$statistic
```

give the same result. Not that `[,2]` and `[,7]` refer to the entire second and seventh columns of `gcdata`. `[2,]` and `[7,]` would refer to the second and seventh rows.

We are now in a position to automate the reference to each column of SSAP data for each t test:

```
> t_result_endo<-NA
> for(i in 5:50) t_result_endo[i-4]<-
t.test(Endosperm_Texture~gcdata[,i])$statistic
> t_result_endo
```

The new command here is: "`for(i in 5:50)`" which will repeatedly execute the subsequent statement 46 times – for all values of i from 5 to 50 inclusively. Since the reference to columns of `gcdata` in the t test is made using "i" rather than a specified column number, the t test is carried out consecutively for all 46 SSAP markers.

Remember that to start things off, we need to create t_result_endo[1]: in this case with a missing data value "NA". This is changed on the first call to `t.test`. Also note that the index for `t_result_endo` is `[i-4]` and not `[i]` since otherwise we would create 50 entries for `t_result_endo`: the first 4 with the value `NA`.

This may seem somewhat complicated, but in practice things are not too bad. One works towards the correct result through trial and error: mistakes are corrected by editing and re-executing commands until they work as expected.

Finally, we need the average value of t. However, sometimes t is negative and sometimes t is positive – arbitrarily depending on whether the SSAP allele is associated with increasing or decreasing the phenotype. We could avoid this difficultly by taking the absolute value of the t-tests, but shall avoid this difficulty by working on the square of t, which should be approximately distributed as a chi-square with 1 df and is more appropriate for genomic control.

Provided the number of observations is reasonably large – say 50 or more – this is justified.

```
mean(t_result_endo^2)
```

All the hard work is now done. You should now be able to:

Calculate t^2 for Endosperm texture and Pinb_a, and establish the significance.

Adjust this t^2 by dividing by the mean t^2 at the control markers, and look up the significance.

Compare the two results. Is endosperm texture associated with the pinb_a SNP in this dataset? Has the result been affected by population admixture?

You can now repeat this exercise for the other quantitative trait – protein content.

Other traits
You can also study the association between pinb_a and grain hardness. For this categorical trait, the association test is a contingency chi-square and not a t test.

1) Calculate the chi-sq between hardness and pinb_a.

2) Calculate the chi-sq between hardness and each of the SAPP markers

(Hint: `chisq.test(A,B)$statistic` will provide the test statistic for the contingency chi-squared test between A and B).

3) Calculate the average chi-sq at the control markers.

4) Divide the result from (!) by the results from (3).

5) Test for significance before and after adjustment for genomic control.

Advanced problems.

The t-test we have carried out has assumed that the variance within each group is unequal (the R default), both for the candidate and for the control markers. Would we be better off assuming variances to be equal? If variances are different at the candidate locus, what is this telling us?

Appendix 10: Demonstration of STRUCTURE and its use in association analysis.

This exercise is to show you how to get data into the computer program STRUCTURE, run the program and analyse the results. STRUCTURE is easy to run but has its pitfalls. This guide should not be used as a substitute for the STRUCTURE manual, which gives more detailed guidance.

The test dataset we are using is part of the EC funded Gediflux project to study crop diversity in Europe. There are 113 European wheat varieties, one quantitative trait, two candidate polymorphisms and 42 SSR markers. The quantitative trait is endosperm texture – an important component of bread making quality, The two polymorphisms are pinb_a – a SNP in the pin_b gene, and an unmapped SSAP marker. The 42 SSRs are distributed as one per chromosome arm (bread wheat is 2n=6x=42). We wish to test for association at the two candidates correcting for the effects of population structure if required. All data are in the spreadsheet "structure strat demo.xls"

Running Structure

Data input

In Explorer, locate the file "structure.bat" and double click to run it.

The file containing the SSR data for analysis is "data_for_structure.txt". This was created by cutting and pasting from Excel. Note, however, that the only data in the first line of the text file are the SSR marker names. Although columns of data in addition to the SSRs can remain in the datafile, they must not be given column names. It is sometimes convenient to leave the column of phenotypes and candidates in the dataset, since they may be wanted for other analyses. They have been unnecessarily left in here to show how they can be handled.

Structure guides you through data input process using a Windows style import wizard.

To start the data import window:

`File, New Project…`

`Step 1`
Enter the project name and select a directory and the input file. This process can be a fiddle – sometimes you have to click once to select,

and sometimes twice. Trial and error should see you through, however.

`Step 2`
Enter information about the data. If you've forgotten this, clicking on "`Show data file format`" may help.

`Ploidy`
For a collection of inbred lines enter 1 here to treat the data as haploid.

`Number of loci`
Enter 42 – the number of SSRs. We do not want the candidate SNPs included in the analysis. This will be sorted out in step 3.

`Missing data value`
Enter 0. It is extremely important to get this right – otherwise the missing values will be included as additional alleles with unpredictable results.

`Step 3`
This formats data input for any additional columns or rows in the file.

Select the `row of marker names` – our file includes this.

We are not using map data – we only have one marker per chromosome arm: they are effectively unlinked. Linkage information can be included in the analysis for loosely linked markers. Map distances between markers can included at the top of the input file. See the STRUCTURE manual for details. Generally, good results are obtained treating markers as unlinked.

`Phase information`: irrelevant here because we are dealing with a haploid organism! Generally, with outbred polyploid individuals, this box would need ticking if the phase was known. See the manual for further details.

`Data file stores data for individuals in a single line`. This also does not apply to haploids. With diploids, data may be entered on a single line as:

1 1 2 2 3 3 4 4 for four loci

or as:
1 2 3 4
1 2 3 4

The second format is the default for STRUCTURE.

Step 4

Individual ID for each individual
We have a unique variety code as the first column of our data, so select this.

Putative population of origin for each individual.
With a-priori knowledge of the population of origin of each individual, this can be entered and included in the analysis. This can work extremely well if, for example, you have a set of lines of known origin, and a set of unknown or admixed individuals. The data from the known lines is effectively used as a training set to classify the unknown lines. Here, leave it blank.

USEPOPINFO selection flag
Used in conjunction with the previous column to identify which individuals to use in the training set and which are to be classified without prior information. See the Structure manual for more details. Leave blank here.

Phenotype information.
Do not select this. STRUCTURE is expecting a column describing a categorical trait rather than a quantitative trait. This is more standard for human genetics where 2 = affected, 1 = unaffected and 0 = unknown. These could be input into the companion program, STRAT. We shall analyse our quantitative trait using R.

Other extra columns.
We have one quantitative phenotype and two candidate loci, which are not part of the STRUCTURE analysis, so enter 3.

Click on Finish, Check the details are correct, then click Proceed. STRUCTURE will test the format of the data and you will be prompted to correct any errors.

If the data input is successful, a spreadsheet-like display of the input data should be returned.

Parameter Sets

Before running STRUCTURE, we need to supply some input parameters. This is deceptively easy. However, the default parameters are not necessarily suitable for all datasets. See the manual for further details. One approach to selecting sensible parameter sets is to find a good publication working on a similar dataset to your own and copy from there.

Again there is an input wizard to guide you through the process. Select:

`Parameter Set, New…`

`Run Length`
STRUCTURE uses Monte Carlo Markov Chain (MCMC) methods. To run successfully, the program iterates many times. There is generally an initial "burn-in" period during which the program settles down, and then a further period in which the program runs and results are generated. The longer these periods are, the more reliable the results. In practice, the numbers selected are determined by the power of your computer and your patience. Generally, select a burn in and a run length of 100,000 at least. In this demonstration, for reasons of speed, select 10,000 for each box

`Ancestry Model`
Stick with the defaults. See manual for more details

`Allele frequency model`
Stick with the defaults. See manual for more details

`Advanced`
In addition to the defaults, select `Print Q-hat.` This writes the membership of each individual in each of the inferred populations to a separate file: useful for subsequent analyses.

Finally, click OK and give your parameter set a name.

Running the program

We are now ready to run the program. To establish how many cryptic populations we have, STRUCTURE is run multiple times, varying the population number. Select:

`Project, Start a Job`

Select the name of your new project. In a session running STRUCTURE you may create several different parameter sets (eg with different run times) and you would be prompted with a list here.

"K" is the number of populations. We shall run from 1 to 3.

To check for stability / repeatability of the STRUCTURE run, it is advisable to replicate each run several times,. We shall have only two replicates – for reasons of time.

Remember, when running STRUCTURE on your own data, you would typically use a longer burn-in, a longer run-time, test more population numbers, and carry out more repeat runs.
Click `Start`
Structure will take a few minutes to run.

On completion, before inspecting the output, start another job – this will take longer to run, so start now to save time:

Select Parameter Set, Modify current sets…

Increase the burn-in and number of MCMC Reps to 100,000 and save the parameter set with a new name.

Select `Project, Start a Job` as before, select the new parameter set name, select K from 8 to 8, 1 replication only, and run.

Now return to consider the results of you runs for K = 1…3. You can do this while the K=8 job is running, although you might find the response from your computer a little sluggish.

Select: `View, Simulation Summary`

This presents a table of summary information from each run. The most important column here is the fourth: Ln P(D).

Ln P(D) gives the posterior probability of the population number. Hopefully, you will see that this increases (gets less negative) as population number increases, but that values are reasonably close within replicate runs. Ideally, with more time, one would continue to increase K to find the value at which Ln P(D) was maximised. In practice, this is not always possible: runs are unstable (see below), they take to long, or K can continue to increase to improbable values. Some compromise is required. The manual describes the problems of deciding on an appropriate value of K in more detail.

In the left hand side of the screen, select one of the K = 3 runs. The right hand side screen should now change to display the results for that run. One can select various graphical displays of the results. It is worth exploring and experimenting with these. The "`Data plot`" options are very useful to check on the stability of the runs. In particular the plot of Log Likelihood against the number of iterations should be seen to stabilise during the burn-in and then fluctuate around a constant value during the run. There should be no trend upwards or downwards during this period (which would indicate a longer burn-in was required). The data plot of Ln P(D) may fluctuate

initially, but should settle to a constant value by the end of the run, with no increasing or decreasing trend.

The "`Bar plot`" shows population membership for each individual in the dataset. This can be sorted by maximum population membership to give a cleaner display (`Sort by Q`). This plot is often seen in publications using STRUCTURE.

The "`Triangle plot`" shows group membership for any pair of populations, plus the residual pooled membership for the remaining populations, all in a single graph. Generally, if the dataset has population structure, and STRUCTURE has detected it, these plots will show clusters of individuals in the corners of the triangle (they come from that particular population), with some individuals scattered along the sides and in the body of the triangle (they are admixed between two or more populations). If this pattern is not seen, then you should suspect that there is no population structure, or none has been detected. For the gediflux dataset, you might see some evidence of population structure in the triangle plots with K=3 (it will vary from run to run). We know from the values for Ln P(D) that at least three subpopulations are present. However, 10,000 iterations is too few for this dataset, especially once K is increased much beyond 3.

To look at the results from a larger population number and longer run-time, you have already short set off a job with K = 8 and a 100,000 burn in and run. If this hasn't finished yet, go for a cup of coffee.

K=8, is possibly still too low for this dataset, and 100,000 may still be too few iterations, but hopefully it will give stable results. (After the job has finished, check to see).

To exit STRUCTURE, close the program in the same way as any other Windows application.

Testing for association

Go to Explorer. Go to the folder where the STRUCTURE data file was kept. You should see that additional folders have been created, corresponding to the project names you entered in STRUCTURE. Each of these in turn will have two subdirectories, one called `PlotData` and on called `Results`. Within the `Results` folder you will find pairs of files of type `name_q` and `name_f`.

The `name_f` file has full results and parameters for each STRUCTURE run. The `name_q` files gives reduced output, containing the variety code, and then population membership for each individual – ideal for

additional statistical analysis. Locate the file for the STRUCTURE run with K = 8.

This file can be read directly into R. R recognises that there are no column headers and generates it's own – V1-V9 in this case. V2-V8 are the population memberships for each line, and V1 is the variety code.

In R, change to the correct directory then:

```
struct<-read.table("run2_run_1_q")
attach(struct)
```

your file name will be different, but will still end `_q`

An interesting plot is given by:

```
> hist(pmax(V2,V3,V4,V5,V6,V7,V8,V9))
>
```

`pmax` (stands for parallel maximum) takes the maximum across the specified fields for each row of the dataset in turn. In this case this gives the maximum population membership for each variety in turn. Subject to STRUCTURE having run successfully, you should see that about 1/3 to 1/2 of the varieties have a maximum group membership >0.8: they are reasonably pure and not admixed. The remaining varieties do not have any strong individual group membership but are admixed across populations. (On the complete dataset, and with long run-times, this distribution looks very bimodal.) If the histogram looks unclear, try:

```
hist(pmax(V2,V3,V4,V5,V6,V7,V8,V9),breaks=20)
```

to increase the number of subdivisions in the histogram.

A researcher or breeder with knowledge of the origins of the varieties may be able to make sense of the populations by studying which varieties fall into which population.

Now load and attach the candidate marker and phenotype data. Use the spreadsheet "`structure strat demo.xls`" to create a file suitable to import into R the variety codes, the phenotype and the two candidate SNPs. Alternatively, use the file "`trait_and_candidate.txt`" which has already been created for this purpose. Remember you may need to change directory again to locate the file.

```
assoc<-read.table("trait_and_candidate.txt")
```

```
attach(assoc)
```

We can test for association very simply with a t test:

```
t.test(Endosperm_Texture~SSAP10)
t.test(Endosperm_Texture~Pinb_a)
```

Results are not corrected for population structure effects of course. For this we require logistic regression. This is a statistical technique which is suited to the analysis of binomially distributed variables where, since any single observation is either "0" or "1", the errors attached to the observation cannot be treated as normally distributed. Each observation is treated as having a probability p of turning up as a 1 and a probability (1-p) of turning up as 0. These probabilities are then modelled directly.

Informally, logistic regression fits the model:

Log[p/(1-p)] = regression parameters + noise.

It is therefore similar to standard regression in which the model

Y = regression parameters + normally distributed noise

Log[p/(1-p)] is the *log odds* or *logit*. It is used because it has some statistically and mathematically convenient properties. In epidemiology, where logistic regression is used extensively, it can also often be interpreted in terms of risk.

Logistic regression is a member of a family of models which have error distributions other than the normal, but which related to it. Together, these are called generalised linear models. Ordinary linear regression is a member of this family.

Practically, in R, logistic regression is fitted using the command "`glm`". This command has virtually identical syntax to the "`lm`" command for ordinary regression.

In the current context, the benefit of logistic regression is that we can regard our candidate locus, suitably coded, as the binary outcome variable, and model this in terms of population group membership and of phenotype. This is the reverse of the standard procedure in which we model the phenotype in terms of the genotype, but this reversal is irrelevant to the detection of association.

Appendix 11: Notes on data handling and error.

The effect of error

We have already seen that genotype errors are often present and can increase map lengths. Even low error rates can have very serious consequences for linkage and association analyses. Moreover, not only to genotype errors generally reduce power, they can sometimes, especially in some forms of association analysis, increase the number of false positive discoveries.

There has been little formal study of the effects of phenotype error on data analysis. If making an error in recording a phenotype is independent of the true phenotype, the effect will be to increase the variance, and possibly alter the mean, of the subset of data containing the errors. This is particularly of concern for any experiment which involves selection, since the error carrying data will tend to be concentrated in the extremes of the phenotypic distribution. This was studied by computer simulation by Mackay and Caligari (1999). With high error rates of 1%, it is actually possible that response to selection is reduced as intensity of selection is increased: an unfortunate result for any plant breeder. Moreover, routine rejection of extreme values generally reduces response to selection too. The conclusion is that it is very important to be vigilant and guard against errors. In the context of mapping experiments, the effect of errors is very clearly seen in bulked segregation analysis – where extremes are selected prior to genotyping. However, in classic mapping experiments too, phenotypic errors will reduce power.

Observed error rates.

There is much anecdotal evidence about the frequency and nature of errors in agricultural research. For example:

Numbers transposed: 17.4 entered as 71.4
Missing plots entered as disease free rather than as missing data.
Misplaced decimals
Hitting the three instead of the decimal point on data entry: a mistake I often make.

In the medical world, where errors are potentially much more damaging to health, their frequency has been studied. An interesting review in medicinal biochemistry is given at
http://www.jr2.ox.ac.uk/bandolier/band47/b47-6.html

Unfortunately, in agriculture and plant breeding, there is little or no quantitative data on error rates.

In 1999, Clifford Thomas Ltd., a publishing company, quoted me a maximum undetected error rate of 0.00001 per alphanumeric key for double entered data. I have no current quoted error rate. (Much UK data entry is now carried out in India!) The error rate per operator is therefore at least the square root of this, roughly 0.3% per punched key, or roughly 1% per three digit number. (This is a minimum estimate because for an error to be undetected, not only do both operators have to make an error, but they have to make the same error.) With double entered data the error rate per three digit number will only be 0.003%. As an experiment, I once entered 2000 randomly distributed three digit numbers and made 11 mistakes: quite close to the commercially quoted rate.

These figures demonstrate both the high error rates expected in the absence of data checking, and the effectiveness of double data entry.

Good practice in data entry.

Copy data by hand as little as possible.

All data entered into a computer must be checked.

Double data entry (by different individuals) is the gold standard.

Second best:

After data entry, print out the data. Someone reads data out from the original scoring sheets, and someone else checks the computer version. It is preferable to use individuals other than the person who first entered the data for this process: each individual tends to make their own set of idiosyncratic mistakes.

Third best:

You enter and check your own data. This is inevitable for small amounts of data.

Data are never entered without any checking.

Databases, spreadsheets and Excel.

For large volumes of data, it is best to enter and store data in a database – Access is perfectly adequate for most projects. Bigger multi centre, multi-user datasets may require something like Oracle: expensive and beyond the scope of this course.

For most Excel or some other spreadsheet is fine. Double data entry is still the preferred method for entry, however. Unfortunately, it is extremely easy to create and propagate errors within spreadsheets. An

interesting review from the business side, containing many horror stories is given by Kruch and Sheetz (2001)

Standardise on names – use of capitals letters, abbreviations, spaces, hyphens and so on. This is particularly important if different individuals are entering data into the same database or spreadsheet.

Although Excel is case insensitive – "VARIETY" and "variety" are treated as identical - this is not the case for many programs, including many statistical packages. Thus you may find yourself analysing more varieties than you thought you had in your field trial.

It is extremely easy to overwrite data unintentionally. With multi-user databases, this can be controlled by the way the database is set up – such that only a limited number of individuals are capable of changing or appending data. In Excel, it is worth copy protecting your source data so that it cannot be changed accidentally – available from the "Tools/Protection" menu.

Always back up your data.

Another pitfall is the presence of leading and trailing spaces around text fields. "VARIETY " (with a trailing space) and "VARIETY" will appear the same, but they are treated as different in Excel, in sorting and in formulae.

Beware of spaces in names in EXCEL. These can lead to problems or unnecessary complications when exporting to other programmes. It is simple to use the underscore character "_" in place of spaces. This can easily be replaced later, in report writing.

Related to problems with trailing spaces, but potentially more damaging: blank cells and cells containing only spaces are impossible to distinguish by eye in Excel. However, they are treated differently. A blank cell multiplied by a number will give a "0" but a cell with a space will give a "#VALUE!". If you are deriving percentages, or transforming data in some other way in Excel, this can give serious problems. If in doubt, the formula "=TRIM(cell reference)*1 will return "#VALUE!" on blank cells. Alternatively, you can use Excel find and replace, available from the edit menu, to find blank cells and replace them with a visible code, or to find cells containing spaces and replace them with a blank cell.

Once final word about Excel. Don't use "*" to denote missing data. They are impossible to find and replace within Excel. "*" is treated as a meta-character, or wild-card, which matches every cell. "NA" is a good alternative – especially for subsequent analysis within R. Some

programs may require a user specified number to be used to represent missing values: -99.999 or similar is common.

Finally, it is worth keeping an audit trail of changes and procedures you have carried out during data entry. Although tedious at the time, it can be very helpful and time saving later on when you are trying to remember what you did to your data, and why. More advanced databases will contain a complete audit trail – tracking all changes made to the data.

Detecting errors

However good your data entry and data handling procedures, it is inevitable that some errors will still get through. Even with automatic data capture, errors occur: recording machines go wrong, balances are incorrectly tared and so on. Data should always be checked for possible errors after entry, whatever its origins. Simple checks can reveal many errors. The most damaging errors manifest themselves as extreme values so are relatively easy to detect:

Sort each field and check:

- The largest couple of values
- The smallest couple of values
- The range
- The expected number of records.

Also check that means / variances / CVs are more or less as you would expect. Histograms of each trait are easy to generate and can be highly informative. Scattergrams of pairs of traits can also be revealing – in particular for data pairs which are not outliers for any single trait, but appear to lie outside the joint distribution of pairs of traits. Equally, a record with extreme values for both traits in a pair is unlikely to be a data input or recording error. The data may be aberrant, and may still ultimately require removal from the analysis, but the observations are probably genuine.

Check any possible relationships you can think of – even if they are not relevant to the analyses you are planning. For example, if you have scores recorded over time, check that they increase or decrease as expected: leaf number or plant height would usually be expected to increase during vegetative growth for example. Here it is important to distinguish between errors of measurement – heights of 26 cm followed by 25 cm a week later are probably indicative of slow growth plus difficulty in measuring plant height in the field with any accuracy. Plant height of 62 cm followed by 26 cm is most likely an error during data entry.

Some checks on the measuring processes themselves may also be possible: if samples are processed through a quality test sequentially, a trend in quality over the order in which the samples are processed may indicate drift in the measuring instrument over time.

With replicated field trials, in addition to checks on the raw observations, one has the luxury of studying residuals from the analysis: i.e. the error terms after the effects of variety and replicate (and QTL) are subtracted from the single observation. Any good statistical software will generate and plot these values.

Duplicate Samples

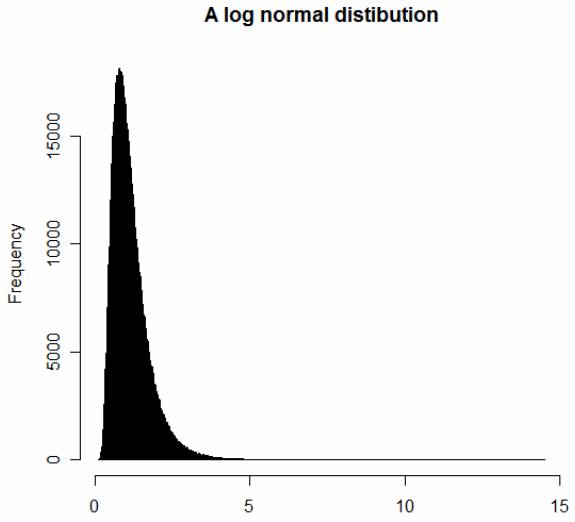
Check for duplicated records. Records with multiple identical phenotypes might be correct of course - it depends on the nature of the phenotypes. Records with multiple identical genotypes might be correct too, but are possible duplicate samples – reflecting a mix up of seed or DNA, or sometimes the existence of a single variety with two names or accession codes. The identification of these duplicate samples generally require multiple highly heterozygous markers such as SSRs: half a dozen SNPs is inadequate. If the varieties are all accessions from a single population, then the probability that of a pair of accessions are identical at all genotypes by chance can be calculated. If this probability is very low, yet these pairs are observed, then there is something wrong. In practice, with multiple genotypes, even if two samples are identical, genotype errors mean that the samples may differ at one or two loci. A very simple method is to count the number of loci at which every pair of accessions differ and plot a histogram of the distribution. If it looks bimodal – with one or more pairs having very high identity – then these are potential duplicates. A more sophisticated graphical method is available in the software GRR: http://www.sph.umich.edu/csg/abecasis/GRR/

GRR is targeted at detecting errors in human pedigrees with genotype data from many SSR markers. It will still pick out the potential duplicates in more modest plant data, however. With sufficient genotype data – currently a very rare luxury for most plant geneticists – this software will discriminate in a very appealing visual manner between full-sibs, half-sibs, parent-offspring relationships and sometimes more distant relationships too.

Duplicate samples are not wasted: comparing their data allows an assessment of genotype error rates.

Normalility

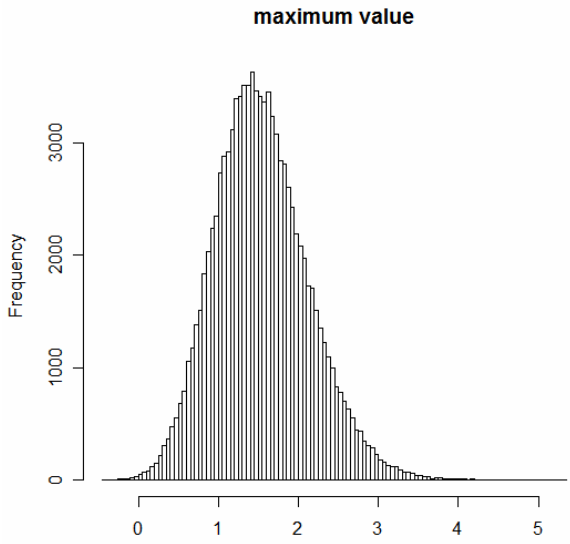
Most quantitative trait analysis assumes that the traits are normally distributed, or more accurately, that the errors of the trait measurements (after fitting QTL and other fixed effects) are normally distributed. Fortunately, many quantitative traits do fall reasonably closely into this category. Sometimes, however, we may need to consider transforming our raw data to a different scale of measurement before analysing the data. The most commonly used is to take logarithms of the data. A sample of 1 million log normal random numbers is plotted below. Note the long tail. If you data looks skewed, you should try analysing the logarithms of your data. Other transformations also exist, and there are other methods of handling non-normally distributed data. However, we would advise that if no simple transformation such as the logarithmic appears to work, you should take advice.



Removal of probable errors.

Finally, we must consider what to do with any aberrant data that is identified. The first thing is to check back to the original scoring sheets and field records (which consequently must not be thrown away, however engrained with mud and sweat they become). This may help resolve many errors. There remains the problem about what to do with outliers for which no obvious cause can be found. There are no hard and fast rules. Really aberrant data will be deleted. Sometimes, if the cause of the aberrant observations can be found, in can be included as a covariate in the analysis - for example if excess pesticide was applied to one section of a field trial, possibly resulting in stunted growth, then an additional factor could be included in the analysis to account for this. Sometimes, a cause for the aberrant observation can be found and the most sensible course of action is to remove the data point. There may remain, however, a number of data points which are deviant enough to cause disquiet, but for which no obvious cause can be found. Should these be removed or kept in the

analysis? Empirically, most plant breeders would opt to remove them and many statisticians would vote to keep them. The statistican's point of view is that it is all too easy to remove outlying data, giving a spurious sense of accuracy and reliability to the analysis of the remaining data, which may even then be biased towards some preferred result. This is illustrated in the following example. Generating 100,000 lots of 10 random numbers, with a mean of zero and a standard deviation of one, then plotting the maximum value from each lot of ten in a histogram, we get:



This slightly skewed distribution shows that in these random samples of size ten, the largest value is frequency surprisingly large. For example, 22% of these maxima have a value > 1.96, although only 2.5% of single observations would be expected to be at least as large as this. So by concentrating on large values and eliminating them from the analysis, we will be throwing away much good data. There are formal methods for deciding whether data should be deleted or retained in an analysis, and there is a whole field of Robust Statistics which deals with methods for analysing messy data. In the end, however, it is a matter of judgement as to whether data should be retained or not. Nevertheless, the statisticians warning should be heeded: don't get too overenthusiastic about removing data. A frequently advocated compromise is to analyse data with the problem observations removed and with them kept in, and see what difference it makes. The best solution is to take care not to introduce errors in the first place, and to take care to detect and correct as many as possible.

References

Mackay, IJ, Caligari, PDS 1999. Major Errors in Data and Their Effect on Response to Selection. *Crop Sci* **39**:697-702

Kruck SE, Sheetz SD 2001. Spreadsheet Accuracy Theory *Journal of Information Systems Education* **12:**93-108