New Phytol. (1997), **137**, 165–177

Markers and mapping: we are all geneticists now

By NEIL JONES[1,3]* HELEN OUGHAM[2,3] AND HOWARD THOMAS[2,3]

[1] *Institute of Biological Sciences, University of Wales Aberystwyth, Ceredigion SY23 3DD*
[2] *Institute of Grassland and Environmental Research, Plas Gogerddan, Aberystwyth, Ceredigion SY23 3EB, UK*
[3] *Aberystwyth Cell Genetics Group*

(*Received 19 May 1997; accepted 18 July 1997*)

SUMMARY

This is a review of genetic mapping with molecular markers aimed at the non-specialist who wishes to use, or at least grasp the concepts behind, this powerful analytical tool. Restriction fragment length polymorphisms (RFLPs) are defined and used to illustrate the different aspects of mapping. The principles of segregation, recombination and linkage are considered and related to the idea of a molecular marker map. A description of a typical mapping population and how it is analysed follows. Traits to be mapped are divided into those controlled by 'major' genes and those governed by quantitative trait loci (QTLs). Exploitation of the map for marker-assisted selection, gene cloning and synteny comparisons is discussed, as are some of the limitations to the usefulness of molecular marker maps. Finally other marker systems are introduced, namely minisatellites or variable number tandem repeats (VNTRs); randomly amplified polymorphic DNA (RAPDs); microsatellites or simple sequence repeats (SSRs); and amplified fragment length polymorphisms (AFLPs).

Key words: RFLP, QTL, RAPD, AFLP, SSR, VNTR.

INTRODUCTION

Molecular markers and marker mapping are part of the intrusive 'new genetics' that is thrusting its way into all areas of modern biology, from genomics to breeding, from transgenics to developmental biology, from systematics to ecology, and even, perhaps especially, into plant and crop physiology. Now that we have the capacity to isolate and clone genes, and to map quantitative trait loci, geneticists and physiologists have passed through their courtship phase and gone into serious partnership. We have the technology, and we can glimpse the prize of making that vital connection between the gene and the character, but there are still many obstacles hindering consummation.

One of the difficulties is that the genetic science of molecular markers and their mapping is a complete mystery to many people (including some geneticists). One can get lost in the language, or be tied in knots over the genetical concepts, or end up just befuddled by the black box which holds the software. Notwithstanding these barriers, the fact remains that to put physiology on the map there is an absolute need to 'first find your gene'. The physiologist would argue that, almost by definition, a gene is identified by a change in its function, which is true; but it's also a truism that until we can map genes, or have clear signposts to their location in a linkage group, we cannot do much with them – which is where molecular markers come in.

WHAT ARE MOLECULAR MARKERS?

Molecular markers (DNA markers) reveal neutral sites of variation at the DNA sequence level. By 'neutral' is meant that, unlike morphological markers, these variations do not show themselves in the phenotype, and each might be nothing more than a single nucleotide difference in a gene or a piece of repetitive DNA. They have the big advantage that they are much more numerous than morphological markers, and they do not disturb the physiology of the organism.

Restriction enzymes, electrophoretic separation of DNA fragments, Southern hybridization, the polymerase chain reaction (PCR), and labelled probes are the tools that allow us to access and to use

* To whom correspondence should be addressed.
E-mail: rnj@aber.ac.uk

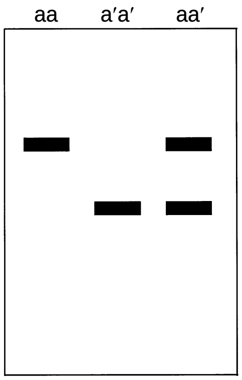


Figure 1. Southern hybridization pattern with a single probe using DNA from plants with three RFLP genotypes at one locus. Track aa is from the homozygote for the larger RFLP allele, a′a′ for the genotype homozygous for the smaller allele and aa′ for the heterozygote. The co-dominance of RFLPs allows for all three genotypes at a single locus to be scored.

these markers. In this discussion, concepts and principles will be developed with reference to just one class of molecular marker (RFLPs), and to plants that are normally diploid. Other marker systems will be covered at the end.

Restriction fragment length polymorphisms

Restriction enzymes cut DNA at restriction sites. Each different restriction enzyme recognizes a specific and characteristic nucleotide sequence. Because even a single nucleotide alteration can create or destroy a restriction site, mutations cause variation in the number of sites. Thus there is variation – or polymorphism – between individuals in the positions of cutting sites and the lengths of DNA between them, resulting in restriction fragments of different sizes. Since the genome of most plants contains between 10^8 and 10^{10} nucleotides, changes in even a small proportion of these can yield a large number of potential DNA markers (Paterson, Tanksley & Sorrells, 1991). A particular restriction enzyme, say a four-base cutter, will generate a whole range of fragment sizes, and when the DNA digest is run out on an agarose gel it will form a smear with the larger pieces at the +ve end and the smaller at the −ve. The range of fragment lengths will be different for different restriction enzymes: a six-base cutter will generate fewer, and on the average larger-sized, fragments than a four-base cutter.

A small piece of cloned genomic DNA, from the same sample of DNA, will match the whole or part of *one* of the fragments in our smear, and if we label this cloned bit with a radioactive or chemical tag it will serve as a probe in a Southern hybridization and will detect the *single* fragment with which it has sequence homology. Figure 1 presents the band pattern that might result. A DNA sample from one plant may show a single band, because the two fragments from a diploid are homozygous, with restriction sites at identical places, and the probe detects both of them at the same place in the Southern blot. A second plant might give a variant of the same fragment that differs in length, because it is homozygous for a mutation which has either destroyed one of the restriction sites or else created a new one within the original fragment. A third plant – it could be an F_1 hybrid between plants 1 and 2 – will show two bands, corresponding in size to the bands from plants 1 and 2, since we are now looking at the heterozygote. Thus we can speak about three different forms of this particular locus, that is the place in the chromosome concerned where our fragment is located, and as there are three forms the locus is polymorphic – a restriction fragment length polymorphism (RFLP) – (Tanksley *et al.*, 1989).

The usefulness of RFLPs

The two different-sized fragments are alleles of one locus. The locus itself is identified by the probe used to detect it, and takes the name or number of that probe. The RFLP is a marker, and it can be used in genetic analysis like any other marker which has alleles identifying a locus; although we note also that the RFLP is co-dominant since we can distinguish all three morphs. This makes the RFLP more informative than a morphological marker with full dominance, where we can only identify two phenotypes: (AA or Aa) and aa.

RFLPs arise as mutations that alter restriction sites, but the events giving rise to them, over evolutionary time, are as stable as the mutations giving any other form of allelic variation; that is, they are constant for all practical purposes. It follows that we might find large numbers of such markers, depending only on the level of polymorphism in a population and the availability of probes. In the numbers game this puts us orders of magnitude ahead of classical markers (such as isoenzymes and morphological features) in our capacity to detect selectively-neutral allelic variation, and therefore far ahead also in the resolving power of our genetics.

WHAT IS MAPPING?

Mapping is putting markers in order, indicating the relative genetic distances between them, and assigning them to their linkage groups on the basis of the recombination values from all their pairwise combinations. To explain mapping we need to refresh ourselves about the genetic concepts of segregation and recombination, illustrated with classical Mendelian markers showing full dominance. Dominant and recessive alleles are given as upper and lower case letters respectively.

Segregation and recombination

As a result of meiosis, two alleles of a locus will segregate (separate from one another) with equal frequencies into the gametes. If a and A are two such alleles, then a diploid individual heterozygous at this locus (genotype Aa) will give gametes half of which are A and half of which are a. Similarly alleles b and B at a separate locus will segregate fifty-fifty into the gametes. If the a/A locus and the b/B locus are unlinked (that is, are on different chromosomes) then the alleles will undergo independent segregation, giving four possible combinations in the gametes: AaBb ⇒ AB, Ab, aB, ab. The simplest way to follow such events, and to introduce recombination, is first to make a cross between two homozygous parents (P_1 and P_2). The offspring of this cross are referred to as the first filial (F_1) generation:

$$P_1\,\mathrm{AABB} \times \mathrm{aabb}\,P_2 \Rightarrow F_1\,\mathrm{AaBb}$$

Next we carry out a testcross between F_1 and the double-recessive parent P_2:

F_1 AaBb × aabb testcross parent

The F_1 segregates to give four kinds of gametes (AB, Ab, aB, ab). The phenotypes of the testcross progeny tell us the genotypes of the gametes:

Testcross progeny

AB ab	Parental type
Ab ab	Recombinant
aB ab	Recombinant
ab ab	Parental type

The four classes of testcross progeny will occur in equal numbers. The two phenotypes that differ from P_1 and P_2, those phenotypically Ab and aB, are the recombinants; and with independent segregation these will comprise 50 % of the testcross progeny.

On the other hand, if the genes are linked (that is, on the same chromosome) the recombinants will only arise when crossing over occurs between them, and then their frequency will be $< 50\,\%$, as a rule. Why 50 %? Because crossing over happens at the four-strand stage of meiosis, and only involves two of the four chromatids. Therefore the maximum crossover value we can get for linked genes is 50 %, and this will only occur when the loci are far apart, like at opposite ends of the chromosomes, so that there is always at least one crossover point (chiasma) between them (Fig. 2).

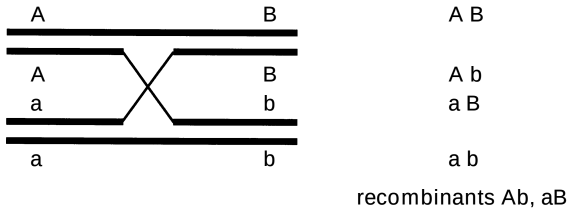


Figure 2. Diagram of a bivalent at the four-strand (diplotene) stage of meiosis, showing how a single chiasma involves only two of the four chromatids and can lead to a maximum of 50 % recombination for genes at opposite ends of the chromosomes. When the two loci are closer together chiasma formation will not always occur, and recombination will be $< 50\,\%$.

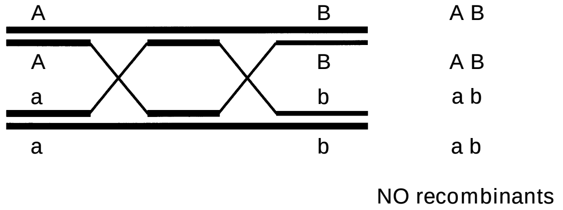


Figure 3. Diagram of a bivalent at the four-strand (diplotene) stage of meiosis, showing how double crossovers involving the same pair of chromatids go undetected as recombinants, and thus underestimate genetic distance.

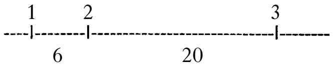
Recombination is the process by which new combinations of parental genes or characters arise and, as seen above, it can occur by independent segregation of unlinked loci or by crossover between loci that are linked. The percentage of a sample of testcross progeny that are recombinants is the recombination frequency or crossover value. This figure gives us an estimate of the distance between two loci in a chromosome, on the assumption that the probability of crossing over is proportional to the distance between the loci.

Recombination and linkage maps

The recombination value for a pair of loci from a segregating backcross population is:

$$\frac{\text{no. recombinants} \times 100}{\text{total no. progeny}} = \text{say, } \frac{18}{300} = 6\,\%$$

Suppose the recombination between loci 1 and 2 = 6 %, that between loci 2 and 3 = 20 %, and that between 1 and 3 = 24 %, then we can order the loci along the chromosome:



One percent recombination = one arbitrary map unit (centimorgan, or cM), and notice that in our map the genetic distances are not additive: 6 + 20 = 26 is the true distance between markers 1 and 3 (not 24). The underestimate based on the recombination between 1 and 3 is due to double (or multiple) crossovers, which go undetected as recombinants (Fig. 3). It is for this reason that maps are built up by adding small intervals. Markers that map together as one linkage group do so because they are all located in a single chromosome. The number of different linkage groups that we eventually find, given enough markers, will correspond to the basic chromosome number of the species.

We also have to appreciate that what we are working with is 'genetic distance' (genetic map), based on recombination frequency. In cases where crossovers are clustered in certain regions, rather than being randomly distributed, then the genetic map will be a distortion of the physical distances separating loci on the chromosomes.

WHAT IS A MOLECULAR MARKER MAP?

Molecular markers, as we have explained for RFLPs, are alleles of loci at which there is sequence variation in DNA that is neutral in terms of phenotype. The alleles are detected using probes, which are pieces of radiolabelled DNA with sequence homology to the marker fragments. Crosses can be made between parent lines which differ for these alleles to give heterozygous F_1 hybrids, and these F_1s can be used to produce a segregating population, from which to calculate recombination values between the marker loci, and thus to make a genetic map in the same way as we have described above for classical gene loci.

The mapping population

The simplest way to make an RFLP map is to make crosses between homozygous lines which reveal allelic differences for selected probes. The F_1 hybrids are then used in various ways to complete the mapping population:

(i) F_1s can be used to produce doubled haploids. Plants are regenerated from pollen (which is haploid) and treated to restore the diploid condition in which every locus is homozygous. Since the pollen population has been generated by meiosis, the doubled haploids represent a direct sample of the segregating gametes.
(ii) The F_1 plants can be backcrossed (testcrossed) to one of the parents to give a segregating backcross population.
(iii) F_1s can be selfed, or crossed in pairs, to give a segregating F_2 population.
(iv) Recombinant inbred lines can be derived from the F_1 population, and represent an 'immortal' or permanent mapping family.

By one means or another a mapping population will be produced which comprises the parent plants, the F_1 and a segregating population (Fig. 4), and all three generations then have to be scored with a large number of probes to determine their genotypes and to calculate recombinant values for pairs of markers. DNA samples are prepared from all plants in the mapping population, and the probes are applied to follow the inheritance of the RFLPs. Clearly the range of markers that can be used will depend on the degree of divergence between the parents going into the cross, and the number and qualities of the probes that have been made.

Making the probes

Probes are generally prepared from genomic DNA or cDNA from the same species as the mapping population (homologous probes), or as heterologous probes from a closely- (or even distantly-) related species. Standard molecular biology manuals give many protocols for making probes (Sambrook, Fritsch & Maniatis, 1989). Here is a typical procedure. Genomic DNA is extracted and restricted with a methylation-sensitive enzyme like *Pst* I which generally does not cut within regions of highly repetitive DNA. This is important because a probe from repetitive DNA might hybridize with very many fragments and give an uninformative smear, whereas probes derived from unique sequences generally give discrete bands. The digest is fractionated on the basis of fragment length, and DNA sequences in the size range 500–4000 bp are recovered and cloned into plasmids. When labelled genomic DNA is hybridized to dot blots of clones, weak signals indicate which plasmids carry unique sequences. Clones are then further selected using Southern blots to genomic DNA to sort out those giving only one or two informative bands from those which give several. The final selection is for the clones that show a polymorphism with the parents used for producing the mapping family. In practice, several combinations of probes and restriction enzymes will be available for a given species, generating a large number of RFLPs. Not only are these markers abundant, they are also stable, convenient, unaffected by the environment, and detectable in all tissues and at all stages of development.

Making the map

The data from the mapping population are produced by probing Southern blots and then classifying the plants for their RFLP pattern (Fig. 4).

The example in the diagram in Figure 4 is a highly simplified scheme with only 12 backcross progeny. It shows the outcome in two separate gels for probes 1 and 2. In the case of probe 1 we see that the heterozygous F_1 segregates its two alleles in equal numbers (idealized numbers of 6 of each) and that these combine with the single allele from P_1 to give six of each of two kinds of backcross progeny in our sample. Probe 2 behaves in the same way with the same DNA samples from the same plants, except

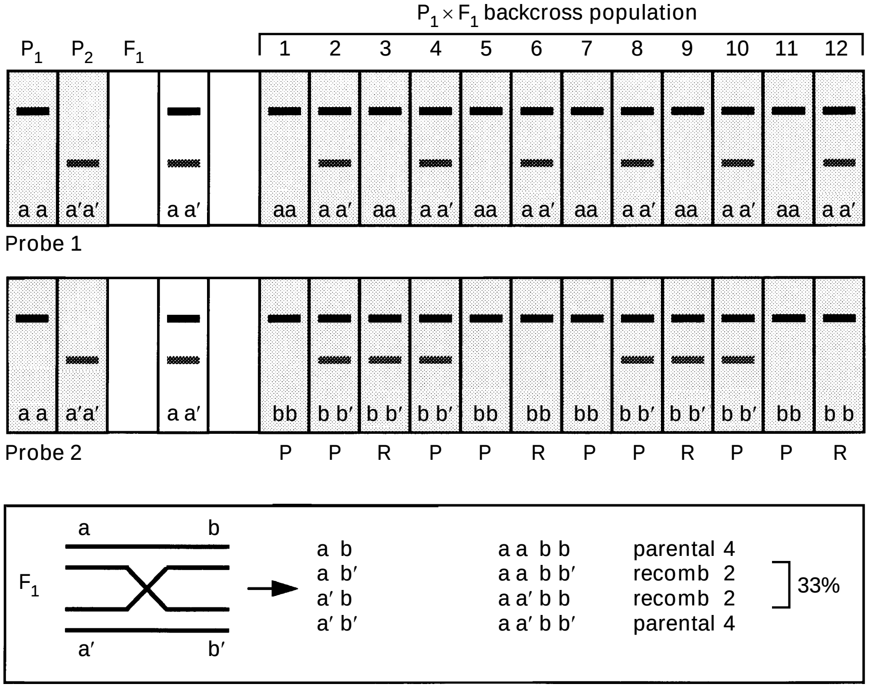


Figure 4. Simplified procedure for RFLP mapping using a backcross. The mapping population consists of parents (P_1, P_2), the F_1 and the backcross progeny. RFLP alleles at two different loci are identified by probes 1 and 2, and the recombinants are the genotypes which have three bands across both gels. The lower part of the figure shows how crossing over between the two loci generates recombinants.

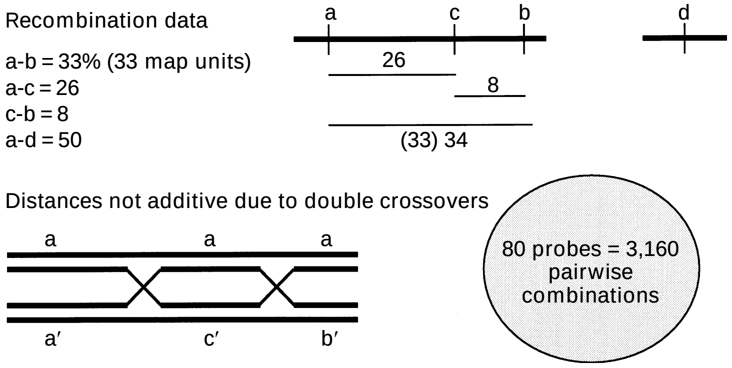


Figure 5. Use of recombination data to produce a genetic map. To make an RFLP map it is necessary to calculate recombination values for a large number of pairwise combinations of loci and then to find the best fit of these values into linkage groups. This procedure can only be accomplished with the aid of a computer program.

that the band patterns are now different. The lower part of Figure 4 explains how the patterns from the two probes are compared to calculate recombination between the two loci detected by probes 1 and 2. The recombinants are all of those which have three bands across the two panels; the other patterns are parental types. Four recombinants out of 12 backcross progeny = 33 % recombination. In the same way many other probes are used, and the data are then analysed, making all possible pairwise combinations (Fig. 5).

If we use $n = 80$ probes, which is a realistic number, then we have to deal with 3160 $((n-1) \times (n/2))$ pairwise comparisons in order to make the best fit for our linkage map. This task requires the analytical power of a computer, and there are software packages available to carry out this task. It is easy to imagine how the data from a mapping population can be entered using a simple binary code and an identifier for each probe. The outcome will be a molecular marker map, of which there are several real-life examples in this volume.

The map of one chromosome might look something like this:

```
--|-----|------|--|-------|---|-|----|-----|-----|--------|--|---|----|--|---|----|---|-----
```

where each vertical line denotes the map position of a locus named after its probe.

Once it has been constructed, what is the use of such a detailed map describing the relative position of large numbers of neutral DNA sequences?

PUTTING TRAITS ON THE MAP

The answer to usefulness is that we now have numerous extra signposts which can point to genes of interest. Instead of having a virtually featureless map of, for example, isoenzymes and morphological markers, as we may have had before, we have a wealth of detail filling in all the gaps. But in order to make use of this new potential for genetic resolution, the adaptive, morphological, developmental or other trait that we seek to analyse must be put onto the same map, so that its precise location can be read with respect to the RFLP signposts. This requires a screening method for the trait to be available. We can then use these signposts to point us and to lead us to the genes of interest, be it for selecting or for isolating and cloning.

To put a given gene onto a molecular marker map there must be phenotypic variation for the trait controlled by that gene within the mapping population. For example, a population might include polymorphism for alleles at a particular flower colour locus. These alleles will segregate together with particular RFLP markers. By computing linkage values between alleles at that locus and the RFLPs, the pigmentation gene can be included in the map. Exactly the same approach can be applied to loci controlling traits such as disease resistance or morphological markers.

'Major' genes

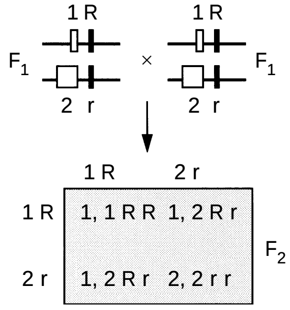
Breeders and other applied geneticists use the term major gene to describe a gene which is inherited in a Mendelian manner and whose allelic forms give qualitatively distinct phenotypes. Mapping of such genes is a relatively simple exercise. For example, in a mapping population segregating for presence or absence of mildew resistance we may discover that the resistance locus always segregates together with a certain RFLP. An example based on the analysis of an F_2 population is shown in Figure 6 (Wricke, Dill & Senft, 1996).

The data in Figure 6 tell us that the resistance gene and the locus encoding the RFLP are so close that they map to the same location. This RFLP then becomes a very useful tight marker or gene tag for resistance. To find a truly coincident marker for a gene of interest is fortunate. Usually we have to work with a nearby tag – say, within 5 cM or so. In this case we would have to learn to live with the loss through recombination of up to 5% of resistant genotypes if selection relied on only one such neighbouring RFLP marker.

In many plants, identifying specific markers linked to a gene of interest is made more difficult by substantial genetic variation throughout the rest of the genome – this problem is greatest in outbreeding species. A way round the problem is to use the approach called bulked-segregant analysis (Michelmore, Paran & Kesseli, 1991). Suppose the gene of interest can confer an apomictic phenotype (Hayward, unpublished). A large number of plants displaying apomixis is collected from the population, bulked together, and DNA is extracted. The same is done for a large sample of plants in the population not exhibiting apomixis (i.e. reproducing sexually). The two DNA preparations so obtained represent the 'average' genome of plants either with, or without, the apomixis allele at the locus of interest, and each can now be treated as though it came from a single plant. The DNA samples are subjected to RFLP analysis using a range of markers as described above. Although the two groups of plants (apomict and sexual) would each have been genetically variable at a wide range of loci, the only *consistent* difference between them would have been at the apomixis locus, and therefore a marker which reveals a polymorphism between the two samples will be linked (more or less tightly) to the locus of interest. Once the marker has been identified by bulked-

Part sample of an F_2 of 29 plants

Plant	resistance	codominant probe
11	1	1, 2
12	9	2, 2
13	1	1, 2
14	6	2, 2
15	1	1, 2
16	1	1, 2
17	5	2, 2
18	5	2, 2
19	1	1, 1
20	1	1, 1



	1 R	2 r
1 R	1, 1 R R	1, 2 R r
2 r	1, 2 R r	2, 2 r r

3:1 for resistance (RR, Rr = score 1): susceptibility (rr = scores 4-9)

Figure 6. Scheme showing how an RFLP allele is used as a tightly-linked marker to screen a segregating F_2 rye population for a gene for mildew resistance (based on data from Wricke, Dill & Senft, 1996).

segregant analysis, it can be applied to individual plants to screen for the trait. Although the procedure is described here with reference to a single Mendelian gene, it is also applicable to the quantitative trait loci discussed in the next section.

Many physiological processes are modulated by major genes. Experimentally induced mutations in model species such as *Arabidopsis* have been valuable in mapping such genes. If a screening procedure for a particular non-lethal physiological aberration can be devised, it is usually possible to find individuals displaying the corresponding phenotype in a population derived from mutagenised parents, and to place the mutant locus on the molecular linkage map (Koornneef, Alonso-Blanco & Peeters, 1997). This approach has successfully cracked the problem of the cellular and molecular mechanisms underlying some of the most recalcitrant aspects of plant form and function, such as floral development or hormone action (Weigel, 1995).

Quantitative trait loci

A major gene trait is ‘digital’ – in most cases, the character is either expressed or not. But many traits, particularly those of significance for crop physiology, are ‘analogue’ and are observable in a segregating population as a more or less continuous range of behaviour between extremes which may even lie outside the mean range of the parents. Examples of characters showing such a quantitative mode of inheritance include yield, stress acclimation, size and so on. The genes that contribute to these complex phenotypes will usually be several in number (polygenes) and may be linked only in the physiological, but not the genetic sense. Molecular marker maps can help us to resolve these complex characters into their contributing quantitative trait loci (QTLs).

QTLs cannot be mapped in the standard way that we have described for RFLPs or major genes, because the individual loci cannot be identified. The principle of QTL mapping is to associate the QTLs with RFLP markers, in their inheritance, and thereby to identify them by having their map locations. The procedure is statistical, and the only reason that it works at all is that we have a detailed molecular map in the first place with which to reference our QTLs. The idea is summarized in Figure 7.

Suppose plant height is the character of interest. A mapping population will be established by crossing two parent lines that are divergent in height, as well as in their RFLP markers. RFLP alleles and QTLs for height will then segregate in the progeny. For simplicity let us suppose we have a single QTL which comprises a cluster of more or less adjacent genetic elements interacting to give quantitative control of stature. Let us consider the following possibilities for this single, albeit complex, QTL in relation to a nearby molecular marker, a more distant marker, one that is remote but will linked, and an unlinked marker (Fig. 7):

(i) *RFLP locus a is tightly linked to the QTL*. Assume that, in the F_1, the linkage is such that RFLP allele a is on the same (homologous) chromosomal DNA strand as the height alleles for short, and RFLP allele a′ is on the same homologue as height alleles for tall. The F_1 plants will be intermediate for height and heterozygous aa′. In the backcross progeny the RFLP a allele will segregate together with height alleles for short, and the a′ allele will segregate with height alleles for tall. If we then plot height distributions separately for the plants carrying the a and a′ alleles we will find two distinct height distribution curves (Fig. 7, locus 1).

(ii) *RFLP locus b is closely linked to the QTL*. In this case the F_1 will be heterozygous in the same way, now for bb′, but the height alleles will be some small distance away and crossing-over will cause recombinants to occur which will associate the RFLP alleles and height alleles in new combinations. We will now get *most* of our b alleles segregating with height alleles for short, and a few of them with height alleles for tall, and contrariwise for b′. When we now plot our height distributions for segregating classes b and b′, the two curves will show more overlap (Fig. 7, locus 2).

(iii) *RFLP locus c distantly linked to the QTL*. By the same argument, a high level of crossing over in the F_1 will make many new associations between the RFLP and the height alleles, and the two curves will be almost coincident (Fig. 7, locus 3).

(iv) *RFLP locus d unlinked to QTL*. Where the molecular marker is on a different linkage group (chromosome) from the QTL, there will be independent segregation of the RFLP and the height alleles in the heterozygous F_1, and we will have a single curve of height distribution (Fig. 7, locus 4).

The concept is simple enough; we associate our QTL with our molecular markers, in their inheritance, and determine the map location. In reality we would have data relating the QTL effects to many RFLP loci, and gradations of effect according to the strength of linkages. Without losing the concept in the complexity let us just say that we could plot the likelihood according to effect against locus positions (Fig. 7*b*) and then map our QTL in relation to its nearest markers. If the distributions shown in Figure 7 accounted for all of the variation in height then we could say that we have a single QTL mapping between loci 1 and 2. It is more likely, however, that we would have to account for height variation by having several QTLs; but notwithstanding this outcome, we would still have an idea of how many loci are involved. This opens up the possibility of dissecting the trait genetically and of using our markers for selection.

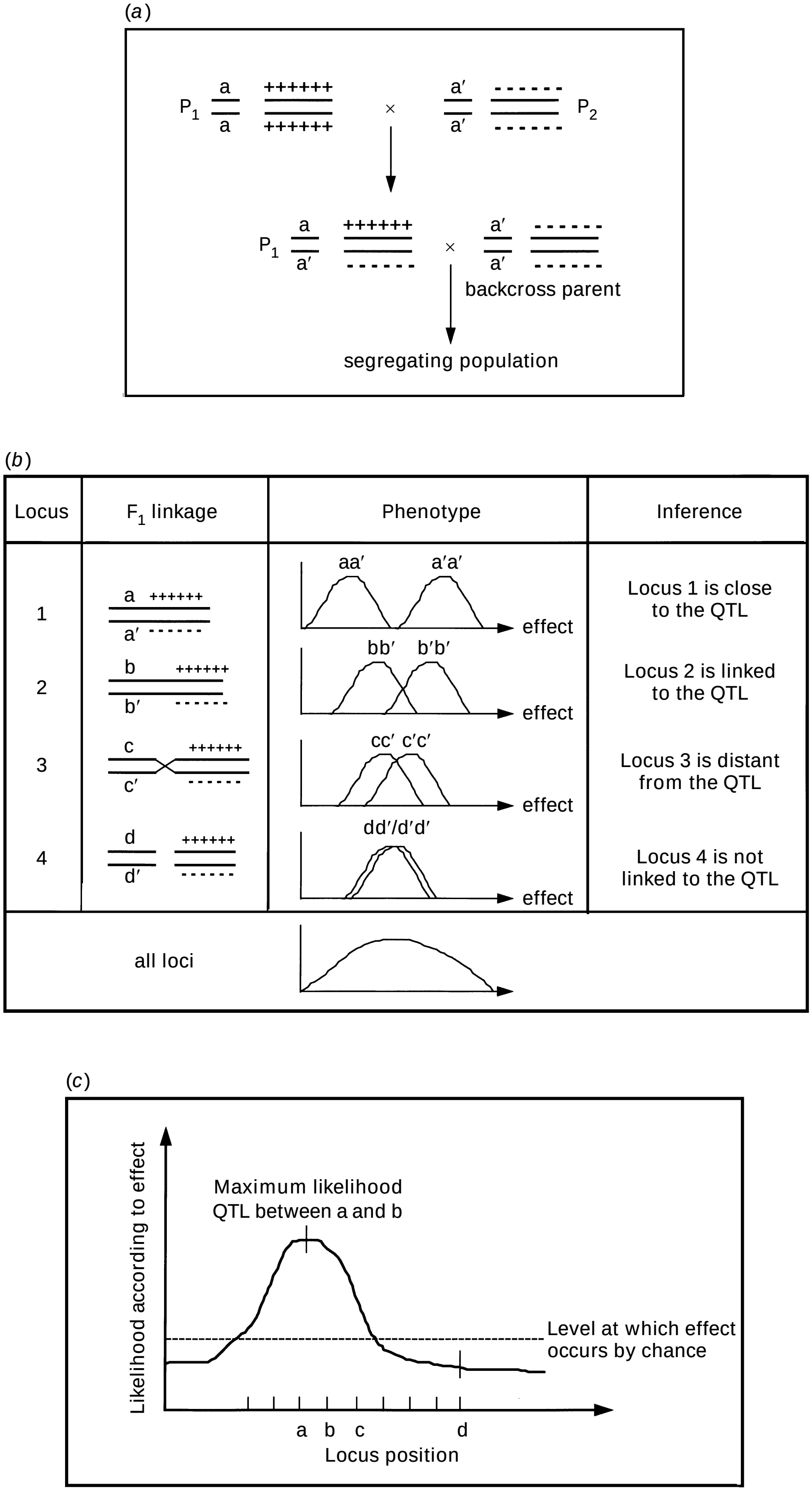


Figure 7. Composite diagram of the procedure for mapping a quantitative trait locus (QTL). (*a*) A mapping population is established by crossing parents which are divergent for their RFLP markers and for the quantitative character concerned (plant height). The heterozygous F_1 is then backcrossed to one of the parents to give the segregating population. (*b*) The linkage between the QTL and various marker loci can then be ascertained by the way in which height distribution patterns are associated with the segregation of the two alleles at each locus. (*c*) The map position of the QTL is determined as the maximum likelihood from the distribution of likelihood values (ratio of likelihood that the effect occurs by linkage: likelihood that the effect occurs by chance) calculated for each locus. (Figure based on an idea by Glynis Giddings.)

A word about transgressive segregation. This topic is also addressed in this volume by Bachmann & Hombergen (1997). The term is used to describe the appearance, in progeny, of characters which quantitatively fall outside the boundaries defined by the phenotypes of the parents in the cross. How can it be accounted for on the basis of QTLs? Consider a trait such as height, governed by – let us say – three loci, each having two allelic forms (A or a; B or b; C or c). Suppose further that in each instance the uppercase allele confers increased height and the lowercase, reduced height. If the parents (P_1 and P_2) in a cross are genetically ABc and abC respectively, then P_1 will be taller than P_2 – but some of the progeny will be ABC and hence taller than P_1, and some, with the genetic makeup abc, will be shorter than P_2. By screening individual plants for markers tightly linked to the desired allelic forms, it is possible to select for those whose phenotypes will lie at the extremes of the height distribution curve. This principle can be applied whether the trait is specified by many QTLs or only three, or (if the QTLs do not contribute equally to expression of phenotype) even two.

READING THE MAP

So, finally we have arranged our molecular markers into linkage groups and assigned major genes and QTLs to their map locations. Where could reading this map take us?

Marker assisted selection

The discussion above considered an RFLP tag on a major gene for disease resistance. If a resistant parent selected to carry this RFLP marker were used in a breeding programme, the progress of the resistance gene through the generations could be followed simply by RFLP screening of, for example, DNA from seedlings – a far less time-, space- and resource-demanding process than carrying out full-scale disease sensitivity trials at each stage. The same approach can be applied to a QTL of interest. It can also work for 'negative selection', where the aim is to eliminate an undesirable trait – one that, for example, might hitch a ride with a useful character that a breeder wants to transfer from a wild species into a related crop plant. The value of markers increases when we wish to combine several characters at once, leading very efficiently to the production of advanced elite lines essentially through 'parallel' rather than more conventional progressive ('serial') processing of genetic information (the term pyramiding has been coined to describe this strategy). This kind of marker-assisted selection is central to the improvement strategies for the world's major crops and, together with the promise of transgenics, may be the best hope we have for meeting humanity's food requirements in the next century (Lee, 1995).

Ongoing technological developments are now simplifying selection procedures to the level where they can be routinely used by plant breeders. In the case of resistance genes, for instance, it is now possible to design primers for PCR reactions which will amplify the particular RFLP allele which segregates with the resistance gene (Fig. 6). We can use these primers to screen populations of plants and to find those which give PCR amplification products. This way of using the map is much simpler, more practical and more economical than full-scale RFLP analysis. Thus we are at the point of 'reading' the DNA map of our crop plants, and then using nothing more than a simple kit to identify and to exploit useful genes. As said in the 'Introduction' – first find your gene.

Gene cloning

Putting a major gene or complex trait on the molecular map also offers the prospect of isolating the corresponding DNA locus, using nearby markers as jumping-off points. If a genomic library is available, it may be possible to move from the clone carrying the RFLP tag to the clone with the gene of interest through a series of intermediate overlapping clones by chromosome walking. If a locus is saturated with markers, we can even go straight to the gene and achieve chromosome landing (Tanksley, Ganal & Martin, 1995). The availability of vehicles such as YACs and BACs (yeast, bacterial artifical chromosomes) for cloning large fragments of genomic DNA, and the production of libraries that relate directly to fully contiguous and highly saturated molecular maps for species such as rice and *Arabidopsis* has made map-based cloning a practical proposition for gene isolation. The limitations of genetic maps for identifying the physical location of genes are discussed later.

Synteny

Molecular marker mapping has strengthened our realization that, in several taxonomic groups of crop plants, e.g. the gramineae which share the same common basic chromosome number, the linkage groups and the individual chromosome maps look very similar. When we take out the repetitive DNA and compare the maps for single-copy sequences (essentially RFLPs) we find that they are syntenic. This means that even between crops as diverse as wheat and rice the genes we are interested in are basically the same in both species, and they line up into maps which are very similar (Moore *et al.*, 1995). The added value to mapping is that not only can we use the same set of RFLP probes across wide species gaps, but we can transfer map information, even entire maps, from one species to another. It becomes possible to know the location of genes of interest in, for example, wheat by reading the marker map for rice (the model graminaceous species). Increasingly too we are beginning to relate genetic maps to the physical dimensions and organisations of chromosomes, and this greatly enhances prospects for gene isolation and manipulation.

LIMITATIONS TO MOLECULAR MARKER MAPS

Molecular mapping is amongst the most powerful tools available to the modern biologist, but it does no service to the technique to ignore the tricky practical and theoretical obstacles to making and exploiting a molecular marker map. Conceptually, arranging genes on a map is facile: if two loci are linked physically by a stretch of chromosome within which crossing-over can occur, then their proximity is measured by their likelihood of becoming separated during meiosis. But this relation makes a number of suppositions. In particular, it requires the two loci to be *functionally independent*, and crossing-over at any point on the chromosome between them to occur *randomly*. Neither assumption is safe in practice.

Genetic distance is not physical distance

Genetic maps do not tell us which linkage groups correspond to which chromosomes or how the markers within a linkage group relate to the physical structure of the chromosome. It is common in grasses, cereals and many other plants that recombination does not occur with equal frequency across the whole genome. Under these circumstances, a marker may appear tightly linked to a gene of economic value, but in reality be many kilobases away in the actual chromosome. For marker-assisted selection, this tight linkage will serve the purpose anyway; but for marker-assisted cloning the desired gene might be too far away from the marker to be reached using the marker probe. It is thus desirable to saturate the map with as many markers as possible and, for cloning purposes, to integrate genetic (linkage) and physical (chromosomal) maps.

To assign linkage groups to specific chromosomes, use can be made of various chromosomal stocks, such as trisomics, monosomics, addition lines, translocations and deletions, which give modified segregation patterns and expose chromosome-specific markers. A particularly useful account of physical mapping in barley, using *in situ* hybridization (ISH) to chromosomes, has recently been presented by Pedersen, Giese & Linde-Laursen (1995). These authors show how mapping of single- and low-copy genes by ISH can provide ‘anchor sites’ for integrating the physical and genetic maps.

Loci are not always independent in their action

In relating a phenotype to a plant’s genetic profile obtained using markers, it is important to bear in mind that many – perhaps most – characters are controlled by more than one genetic locus, and that these loci may or may not be linked. This does not apply to quantitative traits only. Take the case of an enzyme used in isoenzyme analysis. It might have two subunits encoded by genes on different chromosomes, each of which has three possible active allelic forms. There are thus nine different subunit compositions possible for the functional enzyme, each of which might exhibit a different electrophoretic mobility. To screen a population for individuals producing a particular one of the isoenzyme forms, it will be necessary to obtain the right pair of markers each of which is linked to the right allele at one of the two discrete loci encoding the enzyme subunits. Another well characterized example of interacting but often unlinked loci is the group of genes known as homeotic, whose products determine the form of plant organs such as flowers (see, for example, Mena *et al*., 1995). Expression of two or more homeotic genes in the right cells at the right time is essential for normal flower development; knocking out one gene has a different effect from knocking out another. Finally, it should be noted that loci which interact to control one trait may interact differently, or not at all, to control another (see also Bachmann & Hombergen, 1997). For example, drought tolerance can be influenced by genes regulating rooting depth, stomatal density and accumulation of water-soluble carbohydrate (Thomas, 1997), so all these genes would appear as QTLs affecting drought tolerance. On the other hand, carbohydrate accumulation might also be an important component of nutritional quality in a forage, and therefore the gene controlling it would be a QTL affecting quality; but rooting depth and stomatal density would not be relevant to nutritional quality, and these genes would not appear as QTLs when mapping this trait.

Mapping populations are just that

The decision has to be made very early on in a mapping programme: which population is most suitable for constructing the map? However carefully a population is selected, its shortcomings may become apparent later on. The most common limitation is that the mapping population may not be polymorphic for a trait of interest or the marker(s) linked to it, or both. Consequently it will be necessary to design a cross that generates a population segregating for that trait, to bring the population into the programme and to merge maps using additional markers. Another problem is that markers developed in one population of one species might not always be transferable to other closely-related species, or even other populations; this is a greater problem with PCR-based techniques (described below) than with RFLPs. A further consideration is the variable stability of QTLs between different mapping families. And in constructing the 'definitive' saturated map for any species (plant, animal or human), the philosophical question arises: what popultion is most truly representative of that species? This question becomes almost impossible to answer for species like maize which have very plastic genomes (McClintock, 1978), and the answer must be decided on pragmatic grounds to suit the needs of the ecologist, plant breeder or molecular biologist concerned.

OTHER MARKER SYSTEMS

We have used RFLPs to explain how molecular markers are used in mapping and how the maps generated from then can be applied. In fact this publication is full of examples of the making, and modes of exploitation, of molecular marker maps.

In addition to RFLPs, other neutral marker systems which can be used in much the same way have been, and continue to be, developed. Anything which can reveal allelic variation has the potential to be used as a marker for mapping or for DNA profiling, and the same principles described above will apply. Since this contribution is to do with explaining the concepts of map-making and map-reading, rather than reviewing markers, it is appropriate to deal only briefly with some of the other marker systems.

Minisatellites or variable number tandem repeats

RFLPs are probe-based markers, since we need a labelled probe in order to detect the polymorphism. The other well-known probe-based marker is the variable number tandem repeat (VNTR), or minisatellite, first discovered in humans by Alec Jeffreys in the 1980s, and used extensively and sensationally in forensic DNA profiling (Jeffreys, Wilson & Thein, 1985). Minisatellites consist of tandem arrays of short repeated sequences highly dispersed throughout the genome at numerous loci. They are embedded in unique flanking sequences, and the loci are hypervariable in terms of their number of repeat units. Not only are there many different loci, but multi-allelic forms of single loci exist as well at the population level due to unequal crossing over. A consensus sequence for different loci means that a 'polycore' probe can be constructed which can detect up to 30 loci simultaneously, to give a detailed 'DNA fingerprint'. VNTRs are particularly useful in vertebrates, but probes also exist which can be used to produce low-resolution fingerprints in plants, and these find their application in cultivar identification.

The remaining marker systems considered are not probe-based but instead rely on variations of the polymerase chain reaction technique.

Randomly amplified polymorphic DNA

It was discovered (Williams *et al.*, 1990) that a single PCR primer of about 10 arbitrary nucleotides in length will find homologous sequences in DNA, by chance, and will amplify several different regions of a genome. The primer amplifies pieces of DNA of between 200 and 2000 kb long, which lies between two inverted copies of itself, one copy binding to each strand of the DNA. Statistically, priming occurs once in every million base pairs. During the PCR reaction a set of fragments of differing sizes will be generated, and because the fragments have been amplified there is enough DNA to be visualized by staining with ethidium bromide. In general, for the average-sized genome between 5 and 10 fragments will be amplified to produce discrete DNA-banding patterns. Polymorphisms arise because sequence variation in the genome alters the primer binding sites. Randomly amplified polymorphic DNAs (RAPDs) are therefore dominant markers due to their presence/absence at particular loci, and they will segregate from a heterozygous diploid as Mendelian alleles. RAPDs are much simpler and less expensive to work with than RFLPs because no prior knowledge of sequences is reqired and there is no need for radioactive probes. Many different primers can be made, and there is virtually no limit to the numbers of RAPDs in a genome. RAPDs can be used for mapping, but because of the random nature of their generation, and short primer length, they cannot easily be transferred between species. They are most often used as species-specific markers for diversity and phylogenetic studies, e.g. genome relationships in Triticeae (Wei & Wang, 1995). Their main disadvantages are poor reliability and reproducibility, and their sensitivity to experimental conditions (Karp, Seberg & Buiatti, 1996).

Microsatellites or simple sequence repeats

Plant genomes contain large numbers of simple sequence repeats (SSRs), or microsatellites, of < 6 bp which are tandemly repeated and widely scattered at many hundreds of loci throughout the chromosome complement. Typically they may be dinucleotides (AC)*n*, (AG)*n*, (*AT*)*n*; trinucleotides (TCT)*n*, (TTG)*n*; tetranucleotides (TATG)*n* and so on, where *n* is the number of repeating units within the microsatellite locus. In addition to occurring at many different loci, they can also be polyallelic. (AT)*n* dinucleotides are the most abundant type of SSR in plants (Ma, Roder & Sörells, 1996). The methodology used to isolate an SSR at a particular locus starts with the construction of a small-insert genomic library. The library is then screened with a number of microsatellite probes to identify inserts carrying SSRs. The inserts are then sequenced and primers are chosen which match unique flanking sequences for particular loci. PCR amplification is used to generate DNA banding patterns on a gel and to reveal the polymorphism based on different numbers of repeats at the two alleles of a locus. The marker thus has the advantage of being codominant. In addition they are simple, PCR-based and extremely polymorphic, and highly informative due to the number and frequency of alleles detected and to their ability to distinguish between closely-related individuals. They find application as markers for mapping, cultivar identification, protecting germplasm, determination of hybridity, analysis of genepool variation, and as diagnostic markers for traits of economic value (Powell, Machray & Provan, 1996). Microsatellites are, however, expensive to establish, they have a long development time and they need specific primers.

Amplified fragment length polymorphism

The amplified fragment length polymorphism (AFLP) method combines the use of restriction enzymes with PCR amplification of fragments, and detects fragment length polymorphisms (Frijters *et al*., 1995). The first step in the generation of AFLPs is to double-digest genomic DNA with two restriction enzymes. A rare cutter such as *Pst*I cuts in non-methylated DNA and is used to create a bias towards low-copy fragments, and a frequent cutter such as *Mse*I then produces the smaller fragments with an average length of *c*. 256 bp. The use of frequent cutter enzymes only would generate too many fragments for gel electrophoresis. Next, a specific short DNA sequence is linked to one end of the fragment, and a different sequence added to the other. These sequences, together with the adjacent restriction sites serve as binding sites for PCR primers. The primers are designed to match the two different added sequences, and they also carry short extensions of 1–3 nucleotides to bring about selective amplification of those fragments with complementary 1–3 nucleotide sequence. Three kinds of fragments result: Type I are fragments with rare cutter ends only, and these are rare and negligible; Type II have one rare cutter and one frequent cutter end; and Type III have two frequent cutter ends. The Type II fragments are tagged with biotin on their rare cutter end, and this tagging then allows for their separation by affinity to avidin molecules bound on magnetic beads. Thus only the Type II fragments are used in the PCR amplification. The AFLP system is technically difficult and expensive to set up, but it detects a large number of loci, reveals a great deal of polymorphism and produces high complexity DNA fingerprints which can be used for identification and for high resolution mapping and marker assisted cloning.

UNCHARTED TERRITORY

Physiologists are now fully familiar with the power of the gene expression approach (messenger RNA, cDNA, differential and subtractive cloning and so on) to analysing plant processes. But in general they are less au fait with what mapping offers, and the benefits of bypassing gene expression and going straight to the genome. Mapping should now be counted as one of the weapons in the physiologist's armoury. Conversely, physiologists can comfort themselves with the knowledge that without their skills, a molecular marker map bears the same relation to reality as does a road map to a real functioning road: it's a useful guide, but you can't travel very far on it. Geneticists and physiologists will need to work together to turn the abstractions of the map into real biology and agronomy.

ACKNOWLEDGEMENTS

The authors are grateful to Glynis Giddings for her generous assistance, particularly in devising the explanation for QTL mapping shown in Figure 7, and to Catherine Howarth for helpful comments. The Institute of Grassland & Environmental Research receives financial support from the Biotechnology & Biological Sciences Research Council.

REFERENCES

Bachmann K, Hombergen E-J. 1997. From phenotype via QTL to virtual phenotype in *Microseris* (Asteraceae): predictions from multilocus marker genotypes. *New Phytologist* **137**: 9–18.

Frijters A, Pot J, Peleman J, Kuiper M, Zabeau M. 1995. AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Research* **23**: 4407–4414.

Jeffreys AJ, Wilson V, Thein SL. 1985. Hypervariable 'minisatellite' regions in human DNA. *Nature* **314**: 67–73.

Karp A, Seberg O, Buiatti M. 1996, Molecular techniques in the assessment of botanical diversity. *Annals of Botany* **78**: 143–149.

Koornneef M, Alonso-Blanco C, Peeters AJM. 1997. The genetic approach in plant physiology. *New Phytologist* **37**: 1–8.

Lee M. 1995. DNA markers and plant breeding programs. *Advances in Agronomy* **55**: 265–344.

Ma ZQ, Roder M, Sörells ME. 1966. Frequencies and sequence characteristics of di-, tri-, and tetra-nucleotide microsatellites in wheat. *Genome* **39**: 123–130.

McClintock B. 1978. Mechanisms that rapidly reorganize the genome. *Stadler Genetics Symposia* **10**: 24–47.

Mena M, Mandel MA, Lerner DR, Yanofsky MF, Schmidt RJ. 1995. A characterization of the MADS-box gene family in maize. *Plant Journal* **8**: 845–854.

Michelmore RW, Paran I, Kesseli RV. 1991. Identification of markers linked to disease resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions using segregating populations. *Proceedings of the National Academy of Science USA* **88**: 9828–9832.

Moore G, Devos KM, Wang Z, Gale MD. 1995. Grasses line up and form a circle. *Current Biology* **5**: 737–739.

Paterson AH, Tanksley SD, Sorrells ME. 1991. DNA markers in plant improvement. *Advances in Agronomy* **46**: 39–90.

Pedersen C, Giese H, Linde-Laursen I. 1995. Towards an integration of the physical and the genetic chromosome maps of barley by *in situ* hybridization. *Hereditas* **123**: 77–88.

Powell W, Machray GC, Provan J. 1996. Polymorphism as revealed by simple sequence repeats. *Trends in Plant Science* **1**: 215–222.

Sambrook J, Fritsch EF, Maniatis T. 1989. *Molecular cloning: a laboratory manual. 2nd edn*. New York: Cold Spring Harbor.

Tanksley SD, Ganal MW, Martin GB. 1995. Chromosome landing – a paradigm for map-based gene cloning in plants with large genomes. *Trends in Genetics* **11**: 63–68.

Tanksley SD, Young ND, Paterson AH, Bonierbale MW. 1989. RFLP mapping in plant-breeding – new tools for an old science. *Bio-technology* **7**: 257–264.

Thomas H. 1997. Drought resistance in plants. In: Basra AS, Basra RK, eds. *Mechanisms of Environmental Stress Resistance in Plants*. Amsterdam: Harwood Academic Publishers, 1–42.

Wei J-Z, Wang RR-C. 1995. Genome- and species-specific markers and genome relationships of diploid perennial species in Triticeae based on RAPD analyses. *Genome* **38**: 1230–1236.

Weigel D. 1995. The genetics of flower development: from floral induction to ovule morphogenesis. *Annual Review of Genetics* **29**: 19–39.

Williams JGK, Kubelic AR, Livak KJ, Rafalsky JA, Tingey SV. 1990. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acids Research* **18**: 6531–6535.

Wricke G, Dill P, Senft P. 1996. Linkage between a major gene for powdery mildew resistance and an RFLP marker on chromosome 1R of rye. *Plant Breeding* **115**: 71–73.