Review TRENDS in Plant Science Vol.12 No.2

Methods for linkage disequilibrium mapping in crops

Ian Mackay and Wayne Powell

NIAB, Huntingdon Road, Cambridge, CB3 0LE, UK

Linkage disequilibrium (LD) mapping in plants detects and locates quantitative trait loci (QTL) by the strength of the correlation between a trait and a marker. It offers greater precision in QTL location than family-based linkage analysis and should therefore lead to more efficient marker-assisted selection, facilitate gene discovery and help to meet the challenge of connecting sequence diversity with heritable phenotypic differences. Unlike family-based linkage analysis, LD mapping does not require family or pedigree information and can be applied to a range of experimental and non-experimental populations. However, care must be taken during analysis to control for the increased rate of false positive results arising from population structure and variety interrelationships. In this review, we discuss how suitable the recently developed alternative methods of LD mapping are for crops.

Linkage disequilibrium mapping: methods developed for human genetics find applications in crops

Linkage disequilibrium (LD) mapping, also known as association mapping or association analysis, detects and locates quantitative trait loci (QTL) based on the strength of the correlation between mapped genetic markers and traits (see Glossary). It relies on the decay of LD, initially present in a population, at a rate determined by the genetic distance between loci and the number of generations since it arose (Box 1). Over a series of generations, in an unstructured population (a randomly mating population with no complicating factors such as population subdivision and immigration), only correlations between QTL and markers closely linked to the QTL will remain, facilitating fine mapping. However, most populations have some degree of structure or subdivision and the simple relationship between strength of correlation and meiotic distance does not apply: correlations between unlinked loci often occur. Recently, methods of LD mapping that adjust marker-trait associations for these spurious associations have been introduced. Originally developed for human genetics [1,2], these methods and their derivatives are now being applied to crops; driven by the development of cheaper, higher density molecular markers. Successful use will lead to more efficient marker-assisted selection, facilitate gene discovery and help to meet the challenge of connecting sequence diversity with heritable phenotypic differences.

In this review, we first describe the relationship between family-based linkage analysis and LD mapping. We then outline the methods currently available and the opportunities and challenges of LD mapping in crops. A review of practical results in crops can be found be found in Ref. [3].

Family-based linkage mapping and LD mapping compared

Family-based linkage (FBL) mapping can be regarded as a special case of LD mapping in which LD is generated by establishing a population from a small number of founders in the recent past. An F2 population, for example, is derived from a single F1 plant. The meiotic process and an appropriate experimental design ensure that the strength of the correlation between a marker and trait is proportional to the genetic distance of the marker from the QTL, with the correlation between unlinked loci being zero. The precision of the QTL location depends on detecting differences in the recombination fraction (θ) between QTL and adjacent markers. In an F2 with markers located 0, 1 and 10 cM away from a QTL, the proportion of non-recombinant chromosomes is roughly 1.00, 0.99 and 0.90, respectively. Detecting a difference in signal strength between these markers requires a large experimental population. If the F2 was randomly mated for 100 generations, then the frequencies of non-recombinant chromosomes are 1.00, 0.68 and 0.50, respectively [4], and QTL could be located more precisely. In natural populations of crop plants, or among collections of cultivars, there have often been many rounds of historical recombination. LD mapping exploits this historical recombination and provides opportunities for fine mapping that are difficult to achieve through family-based linkage analysis. However, for QTL detection, rather than location, FBL mapping is usually more powerful. In this case, the lack of recombination between a QTL and linked marker increases the power of detection. For these reasons, it is unlikely that LD mapping will supersede FBL mapping: the two approaches are complimentary.

FBL and LD mapping also differ in their dependency on allele frequency in the population being mapped. In populations of plants derived from an F2, QTL are either not segregating, or are segregating at a frequency of 0.5 (ignoring selection and drift). Careful choice of parents, for example selecting phenotypic extremes, is therefore required to ensure that the population is segregating for most QTL for the trait of interest. LD mapping generally

Corresponding author: Powell, W. (wayne.powell@niab.com).
Available online 16 January 2007.

 1360-1385/$ – see front matter doi:10.1016/j.tplants.2006.12.001

Glossary

Admixture: intermingling of individuals from genetically different populations.
Analysis of variance: a method to test the statistical significance of differences among several categories, rather than just two; in which case a *t* test is usually used.
Candidate polymorphisms: polymorphisms that have not been chosen at random to test for trait association, but for which prior knowledge exists: they might be in a known linkage region or, for example, in a gene predicted to affect the phenotype.
CentiMorgan (cM): a measure of genetic distance, additive over loci. At small values, the distance in cM and the recombination fraction (×100) are nearly identical.
Chi-squared test: a widely used test of statistical significance.
Consanguinity or kinship: close genetic relationships between individuals.
Drift: the change in allele frequency over time that results from sampling variation from generation to generation.
False negative: the declaration of an outcome as statistically non-significant, when the effect is actually genuine.
False positive: the declaration of an outcome as statistically significant, when there is no true effect.
Family-based linkage analysis: a method of mapping in which the co-inheritance of markers and traits is related to known genetic relationships between members of the same family or pedigree.
Haplotype: a set of genetic markers located on the same chromosome that are sufficiently closely linked and that tend to be inherited as a unit.
Landrace: an old cultivated form of a crop, potentially adapted to local growing conditions, but unimproved by contemporary plant breeding.
Linkage disequilibrium (LD): the non-random association of alleles at separate loci located on the same chromosome (see Box 1).
Logistic regression: a form of regression analysis in which the dependent variable is either 1 or 0, denoting presence or absence. Commonly used in human genetics and epidemiology with 1 denoting diseased individuals and 0 healthy or control individuals. It can also be used to regress the presence or absence of a particular allele at a locus onto a phenotype, as an alternative to the *t* test.
Mapping: the process of locating a genetic variant on a chromosome. Coarse mapping will only locate a variant within a broad interval. Fine mapping increases precision, ultimately enabling the identification of the functional polymorphism(s) responsible.
Mapping population: a set of individuals or lines, typically derived from an F2 or a backcross, which are used to construct genetic maps and to detect and locate QTL on those maps by family-based linkage analysis.
Marker: an identifiable location on a chromosome.
Microsatellite: repetitive lengths of short DNA sequences used as genetic markers.
Multiple regression: regression analysis in which there are multiple independent variables. In LD mapping, these could be multiple markers, within the same or different genes.
Multiple testing: in an experiment involving many candidate polymorphisms, many statistical tests will be carried out. A consequence of this multiple testing is that it is more likely that a false positive result will be declared by chance. Modified methods of significance testing can control the expected number of false positive results.
Non-experimental population: a population not established specifically to map markers or QTL. It is not necessarily a natural population. For example, it could be a collection of breeders' lines.
Population structure: the non-random distribution of genotypes among individuals within a population.
Population subdivision: the partition of a population into subgroups such that most mating occurs within subgroups.
Quantitative trait locus (QTL): a polymorphic site contributing to the genetic variability of a quantitative trait.
Recombination fraction: the fraction of meiotic events that show recombination between a pair of loci.
Single nucleotide polymorphism (SNP): a polymorphism involving a change in only a single nucleotide.
Stepwise selection: a set of methods in which the best subset of all independent variables available for multiple regression is selected. Ideally, only those variables that have an effect on the dependent variable are selected and all others are rejected. In LD mapping, this approach attempts to separate markers affecting a trait from those that do not.
Structured population: a population in which mating does not occur at random.
***t* test**: a test for the statistical significance of a difference between two means.

samples lines from a pre-existing population with multiple founders. The greater range of genetic material in such a population makes it more likely that multiple QTL will be segregating for multiple traits. However, allele frequencies at QTL and markers will also vary. The ease of detecting QTL depends greatly on QTL allele frequency; rare alleles have low powers of detection. Detection is also more likely if QTL and marker allele frequencies match. Therefore, in LD studies, it is wise to ensure that the full range of marker allele frequencies is covered. Moreover, if prior knowledge suggests that QTL allele frequencies are rare (for example, a rare trait might show Mendelian inheritance) then LD mapping is unlikely to be successful and FBL mapping is preferred.

For LD mapping to be possible, LD must be present in the population under study. Causes of LD are outlined in Box 2.

Methods for LD mapping

Multiparent Advanced Generation Intercross

In the Advanced Intercross [5], F2 individuals are intermated for several generations before mapping. The successive rounds of recombination cause LD to decay and the precision of QTL location to increase. This approach has now been extended to include populations with multiple parents, to take into account information from multiple linked markers [6,7], and to prioritize candidate polymorphisms [8]. Its resolution and power are reviewed in Ref. [9]. The multiparent advanced generation intercross (MAGIC) was first proposed and applied to mice [6] and is described as 'heterogenous stock'. Recent successes are described in Ref. [10]. In both crops and animals, an advantage of the method is that a population can be established containing lines that capture the majority of the variation available in the gene pool. Although it might take several years before these populations are suitable for fine mapping, they are cheap to set up and their value as mapping resources increases with each generation. In plants, MAGIC can be used to combine coarse mapping with low marker densities on lines derived from an early generation, with fine mapping using lines derived from a more advanced generation of crossing and a higher marker density. If such populations were established now, they would be well placed to exploit the advances in genomics technology and reduction in genotyping and sequencing costs predicted to occur in the next few years [11–13].

The Transmission Disequilibrium Test and derivatives

The ability to map QTL in collections of breeders' lines, old landraces or samples from natural populations has great potential. In these populations, LD often decays more rapidly than in controlled crosses. Furthermore, phenotypic data often already exist, saving time and money. 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 (TDT) introduced by Richard Spielman *et al.* in 1993 [14].

The TDT provides a way of detecting linkage in the presence of disequilibrium [14]. Neither linkage alone nor disequilibrium alone (i.e. between unlinked markers) will generate a positive result so the TDT is an extremely robust way of controlling for false positives. At its simplest, multiple families consisting of two parents and a single progeny are collected, as shown in Figure 1.

Box 1. Linkage disequilibrium

Principles of detecting and quantifying linkage disequilibrium

Linkage disequilibrium is the non-random association of alleles at separate loci located on the same chromosome. If one locus has alleles A and a with frequencies p_A and $1 - p_A$, and a second has alleles B and b with frequencies p_B and $1 - p_B$, then at equilibrium, even though the loci are linked, the expected haplotype frequencies are the product of the constituent allele frequencies. For example, using the AB haplotype:

$$p_{AB} = p_A \times p_B$$

We define any departure from this state of linkage equilibrium as:

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

At equilibrium, D = 0.

D is the coefficient of linkage disequilibrium. It can be difficult to interpret: its range depends on allele frequency and it is not symmetrical about zero. It is therefore usually rescaled to give it a range from 0 to 1.

The decay of linkage disequilibrium with time

Recombination causes gamete and haplotype frequencies to change towards their equilibrium values. Following random mating, in the absence of mutation, selection and chance effects, the value of the coefficient of linkage disequilibrium, D, in successive generations is:

$$D_{t+1} = D_t (1 - \theta)$$

and therefore

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

θ is the recombination fraction between the two loci.
t is the number of generations of random mating since the start.
D is the coefficient of linkage disequilibrium.
LD decays quicker at higher recombination frequencies. For unlinked loci, the decay is at a rate of 0.5 per generation.
For close linkage and larger values of t:

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

Thus recombination frequency and time are interchangeable – a halving of the recombination fraction is compensated for by a doubling of the number of generations. Figure I shows the decay in LD over time at a series of recombination fractions. LD decays rapidly in the absence of linkage but persists for a long time with tight linkage.

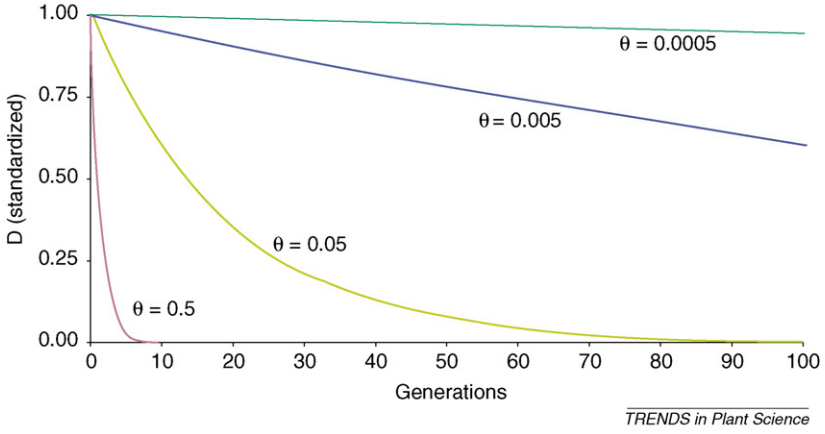

Figure I. Decay of linkage disequilibrium with time for four different recombination fractions (θ). For unlinked loci, $\theta = 0.5$ and LD decays rapidly within a small number of generations. For closely linked loci, the decay in LD is extremely slow. Abbreviation: D = coefficient of linkage disequilibrium.

The single progeny in each family is usually selected for an extreme phenotype. In human genetics this typically means they are affected by the disease under study. Parents and progeny are genotyped, but only parents heterozygous at the marker locus are included in the analysis. From each parent, one allele must be transmitted to the progeny and one is not transmitted. Over all families, a count is made of the number of transmissions and non-transmissions. In the absence of linkage between QTL and marker, the expected ratio of transmission to non-transmission is 1:1. In the presence of linkage it is distorted to an extent that depends on the strength of LD between the marker and QTL. The distortion is tested in a chi-squared test. Power depends on the strength of LD and on the effectiveness of selection of extreme progeny in driving segregation away from expectation.

This elegant test is extremely robust to the effects of population structure, but is susceptible to an increase in false positive results generated by genotype error and biased allele calling [15]. This risk can be reduced by modelling genotype errors and missing data in the analysis [16–18], or by comparing the transmission ratio for extreme phenotypes with that for control individuals or for the opposite extreme. The TDT has been extended to study haplotype transmissions, quantitative traits, the use of sib pairs rather than parents and progeny, and information from extended pedigrees. TDT and other family-based association tests are reviewed in Ref. [19].

In crops, parental and progeny lines are usually separated by several generations of gametogenesis rather than by one. In this case, the TDT is still valid, but might no longer be so robust: the process of breeding might itself distort segregation patterns. A family-based association test that is applicable to plant breeding programs has recently been proposed [20]. The authors point out that for candidate gene studies, this method is more cost effective than the alternative methods described below given that no additional control markers are required. However, some power will be lost because only progeny derived from F1s known to have a heterozygous marker genotype are informative.

Genomic control

Population structure arising from recent migration and population admixture will generate LD between a trait and markers distributed over the whole genome. This can be

Box 2. Causes of linkage disequilibrium

Mutation

Immediately after a mutation occurs, it is in LD with all other loci: the new mutation only occurs on a single haplotype. In successive generations, recombination causes LD to decay as new haplotypes are created, but this process takes a long time for closely linked markers. Most of the polymorphisms we observe are old: many generations are required for allele frequencies to rise to a frequency at which we detect them. Therefore, most pairs of polymorphic loci show little LD originating from mutation unless closely linked.

Population bottlenecks, founder effects and drift

A population bottleneck is an extreme reduction in population size. It causes loss of variation and increased LD. A founder effect is a special case, occurring when a species colonizes a new environment. The number of founders can be extremely small – only a few seeds might need to be introduced to establish the crop. Most crop plants underwent at least one bottleneck during domestication. The activities of plant breeders themselves can result in bottlenecks – the introduction of a new disease resistance or agronomic trait might result in a period of breeding in which a small number of parental lines are used extensively. Indeed, any finite population size generates some degree of LD, just as genetic drift changes allele frequencies.

Selection

Directional selection changes allele frequencies at QTL determining the selected trait. Allele frequencies will also change at closely linked markers. This process, called hitchhiking, generates LD among markers around the selected locus [37,38]. A region of increased LD, often accompanied by reduced polymorphism, can indicate a history of directional selection. Similarly, a region of increased LD and increased polymorphism can result from balancing selection. Such regions have been identified, for example, in maize and *Arabidopsis* [39,40].

Migration and population admixture

If two populations, differing in allele frequency, are brought together, LD is created. Less extreme population admixture or migration also generates LD. If population admixture is known to have occurred and if markers are available that discriminate, even imperfectly, between the parental populations, then these markers can be used to map traits for which the populations differ. This is 'admixture mapping' [41,42] and has been applied in plants [43,44].

More typically, migration and admixture are a problem for LD mapping. The long-range LD they introduce mask the marker-trait associations arising from the close linkage that we wish to detect.

detected by studying whether the distribution of the test statistic for association, estimated empirically from a set of genome-wide distributed markers, differs from the expected null distribution. This is the basis of genomic control (GC) [21,22]. To estimate the empirical distribution accurately would require many markers. However, all that is required is to estimate the mean test statistic and compare it with its expected value (1.0 for a 1 degree of freedom chi-squared test) for which only ~50 markers are needed [23]. If the average chi-squared at a set of 50 control markers is much greater than 1.0, population structure is indicated.

For any candidate marker, the null-hypothesis is now no longer absence of association between it and the trait. Rather, it is that there is no association above the background level resulting from population structure. To test for this, we simply divide the observed chi-squared between the candidate and trait by the average chi-squared at the control markers and look up the *p*-value of the adjusted chi-squared in the usual manner.

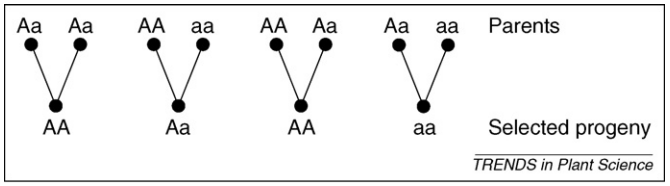

Figure 1. The transmission disequilibrium test. In the simplest case, progeny are selected for an extreme phenotype and transmissions to the progeny from heterozygous parents counted. In the case shown, there are four heterozygous parents from which allele 'A' is transmitted three times and allele 'a' once. This frequency is compared with the 1:1 ratio expected in the absence of linkage disequilibrium between the marker and linked QTL.

GC is valid for any single degree of freedom test. Preferably, the control markers should loosely match the test marker in allele frequency, but this is not crucial [22].

For quantitative traits, the difference between trait means for each marker class is usually tested in a *t* test. Provided the number of observations is reasonably large, t^2 is distributed as a 1 degree of freedom chi-squared and GC 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 one degree of freedom (df) in the numerator and degrees of freedom in the denominator equal to the number of control loci [24].

More sophisticated versions of GC are available. With large numbers of candidate polymorphisms to test, the majority are not expected to be genuinely associated with the trait. In this case, procedures and software are available in which, in effect, the candidate markers act as their own controls. GC has also been extended to control for bias in accuracy of genotyping between DNA samples from different origins [25] and to tests with >1 df [26].

GC also corrects for unknown kinship among collections of lines [21]. The presence of related lines can greatly increase the frequency of false positives. For many crop datasets this will be the greatest source of bias.

The correction of the false positive rate using GC comes at a cost: power is always decreased. This loss of power can be great in cases of extreme population subdivision [27]. Furthermore, because loci can vary in their differentiation between populations, the uniform adjustment of GC might be insufficient for some candidate polymorphisms and overcorrect at others [28].

Structured association

Structured association (SA) provides a sophisticated approach to detecting and controlling population structure [29–31]. Again, additional markers are required, randomly distributed across the genome. Just as for GC, recent migration and population admixture are assumed to generate LD among unlinked and loosely linked markers that has yet to decay fully. However, we expect the parental populations themselves to be in linkage equilibrium. By trial and error one could allocate the individuals in our sample to parental populations such that disequilibrium within populations was minimized. One could then include information on population membership in the test of association. This is the approach taken for SA. First individuals are allocated to populations, then this information is used to control for population membership in the test of association [29–31].

To allocate individuals to populations we need to know in advance how many populations there are. If unknown, this can be estimated: the allocation process is repeated for different possible numbers and the best fitting selected. Nevertheless, deciding on population number can be problematic.

The computer program STRUCTURE [29] uses computationally intensive methods to partition individuals into populations. Many individuals or lines will not belong uniquely to one, but will be the descendents of crosses between two or more ancestral populations. STRUCTURE also estimates the proportion of ancestry attributable to each population.

Following allocation of individuals to populations, the test for association is carried out in a model fitting exercise. Here, the principle is that variation attributable to population membership is accounted for first, using estimates of population membership from STRUCTURE, and then the presence of any residual association between the marker and phenotype is tested. For example, to test for association between a quantitative trait and a microsatellite, the trait is first regressed on the estimated coefficients of population membership and then on the marker – coded as a factor as if in an analysis of variance [32].

SA is effective in detecting and adjusting for the presence of population structure, but does not deal with consanguinity within populations. Recently, Ed Buckler's group introduced a method in which population membership is estimated using STRUCTURE and kinship among varieties is estimated empirically from a second set of control markers [33]. The analysis takes into account both population structure and the correlation between individuals that results from their relationships. This method is implemented in the software TASSEL[*].

Logistic regression

Recent simulations suggest that multiple stepwise logistic regression is robust to the effect of population structure in its own right [27]. Here disease status (affected or unaffected) was used as the outcome variable in a logistic regression on multiple null and candidate markers. Stepwise multiple logistic regression gave false positive rates close to the desired significance level with little loss of power. The authors propose that logistic regression using null markers as covariates is a less conservative (fewer false negatives) method than GC, but with a lower requirement for additional markers than SA. To date, the method has not been tested on crops and has not been adapted for quantitative traits. However, multiple regression with stepwise selection has been applied to barley to consider the joint effect of multiple marker-trait associations [34].

Principal component analysis

A method termed EIGENSTRAT has recently been proposed [28]. It is based on principal component analysis (PCA) across a large number of biallelic control markers with a genome wide distribution. The PCA summarizes the variation observed across all markers into a smaller number of underlying component variables. These can be interpreted as relating to separate, unobserved, sub-populations from which the individuals in the dataset (or their ancestors) originated. The loadings of each individual on each principal component describe the population membership or the ancestry of each individual. However, these estimates are not ancestral proportions (values can be negative) in the same way that estimates of ancestry from STRUCTURE are. The loadings are used to adjust individual candidate marker genotypes (coded numerically) and phenotypes for their ancestry. The adjusted values are independent of estimated ancestry so a statistically significant correlation between an adjusted candidate marker and adjusted phenotype is therefore evidence of close linkage of a trait locus to the marker.

The EIGENSTRAT approach is similar to that of SA but is less dependent on assessing the number of ancestral populations. Although each principal component is attributed to a separate population, the analysis is robust to the number included in the analysis, provided this is sufficiently large to capture all true population effects.

EIGENSTRAT was developed for analysing human datasets, which have high-density genotyping and low levels of population differentiation. Many crops have much higher levels of population differentiation than those found in human datasets and often only low densities of markers are available. In addition, EIGENSTRAT does not cope with close kinships. The authors suggest identifying these by other means and then selecting the largest subset of unrelated individuals. However, they also suggest combining EIGENSTRAT with GC to control for residual confounding. It is possible that such use of GC would also account well for kinship. EIGENSTRAT, unlike SA, will not readily handle multiallelic markers. However, a microsatellite with 10 alleles could be coded as 10 biallelic loci, all in complete LD. An analysis of human data showed that EIGENSTRAT was little affected by LD among >3 million single nucleotide polymorphisms (SNPs). It is possible, therefore, that EIGENSTRAT will be applicable to more modest numbers of microsatellite genotypes, suitably coded, but this remains to be demonstrated. The method shows great promise but additional research is required to establish its suitability for crops.

Haplotype analysis

LD mapping can be extended to consider multiple markers simultaneously. For closely linked markers, haplotype analysis can offer advantages over single marker-by-marker analysis [35]. There are many possible approaches and methods and research in this area is continuing. Within the scope of this review, it is not possible to discuss these. The simplest approaches are:

- Test each haplotype in turn against a pool of all others. This converts a system of n haplotypes to one of n biallelic loci. Analysis is then straightforward but adjustment for multiple testing is required.
- Ignore haplotypes but analyse the constituent markers and their interactions jointly. A significant interaction is evidence of a haplotype effect over and above any effect attributable to the single markers.

[*] Zhang, Z. *et al.* (2006) TASSEL 2.0: a software package for association and diversity analyses in plants and animals. *Plant & Animal Genomes XIV Conference*, 14th–18th January 2006, San Diego, California, USA (http://www.maizegenetics.net/bioinformatics/tassel/screenshots/TASSEL_PAGXIV.jpg).

Recommendations and conclusions

The substantial quantities of phenotype data already in existence from the variety trials of breeders and the variety testing organizations are valuable resources for LD mapping. For example, a genome-wide survey of associations with yield and yield stability components has been carried out in barley [34] using historic data. To generate novel phenotypic data for mapping traits such as stability of yield would usually be prohibitively expensive. Moreover, QTL are detected in germplasm of direct relevance to the crop. Unfortunately, all methods currently available for controlling population structure in such collections have weaknesses. For ease of application and low marker density requirement we favour GC, even though it can be conservative: in the long run, false negative results are less damaging than false positives. With higher marker densities, the more intensive methods of SA and EIGENSTRAT should have greater power. However, even here GC can have a role: to confirm that these more sophisticated approaches have worked.

The resolving power of LD mapping depends on how rapidly LD decays with genetic distance. This varies between populations of landraces, wild progenitors and modern cultivars as a result of the diverse history to which crop plants have been subjected since their domestication [36]. In some populations, LD will decay so rapidly that they are best suited for fine mapping, whereas in others the decay might be so slow that whole genome scans are practical. In crops where collections of contemporary, historical and wild material exist, selection of different sets of lines might permit both fine and coarse mapping [36]. However, in most crops, marker density is currently too low for genome scans. Before attempting these, power calculations should demonstrate that, given the rate of decay of LD in the population to be studied, the density of markers and their allele frequency distribution are adequate to detect linked QTL accounting for specified proportions of the phenotypic variation. Population size is also important. An LD mapping experiment will almost always have lower power than a FBL mapping experiment of equivalent size: if 100 lines are just sufficient for a FBL study, they will be too few for LD mapping.

For these reasons we believe that the best use of LD mapping is to refine the location of QTL identified in FBL and candidate gene studies. Longer term, prospects for high-throughput genotyping and resequencing might make whole-genome scans by LD mapping more feasible. The challenge is to identify and create the appropriate populations so that computational, analytical and profiling advances can be rapidly harnessed by the crop science community. For this purpose, the MAGIC approach is ideal: highly diverse, no population structure, and suitable for both fine and coarse mapping. We believe that MAGIC populations should be established now in all crops.

References

1 Bodmer, W.F. (1986) Human genetics: the molecular challenge. *Cold Spring Harb. Symp. Quant. Biol.* 51, 1–13
2 Duggan, D. *et al.* (2005) Recent developments in genomewide association scans: a workshop summary and review. *Am. J. Hum. Genet.* 77, 337–345
3 Gupta, P.K. *et al.* (2005) Linkage disequilibrium and association studies in higher plants: present status and future prospects. *Plant Mol. Biol.* 57, 461–485
4 Winkler, C.R. *et al.* (2003) On the determination of recombination rates in intermated recombinant inbred populations. *Genetics* 164, 741–745
5 Darvasi, A. and Soller, M. (1995) Advanced intercross lines, an experimental population for fine genetic mapping. *Genetics* 141, 1199–1207
6 Mott, R. *et al.* (2000) A method for fine mapping quantitative trait loci in outbred animal stocks. *Proc. Natl. Acad. Sci. U. S. A.* 97, 12649–12654
7 Mott, R. and Flint, J. (2002) Simultaneous detection and fine mapping of quantitative trait loci in mice using heterogeneous stocks. *Genetics* 160, 1609–1618
8 Yalcin, B. *et al.* (2005) Using progenitor strain information to identify quantitative trait nucleotides in outbred mice. *Genetics* 171, 673–681
9 Valdar, W. *et al.* (2006) Simulating the collaborative cross: power of quantitative trait loci detection and mapping resolution in large sets of recombinant inbred strains of mice. *Genetics* 172, 1783–1797
10 Valdar, W. *et al.* (2006) Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nat. Genet.* 38, 879–887
11 Metzker, M.L. (2005) Emerging technologies in DNA sequencing. *Genome Res.* 15, 1767–1776
12 Syvänen, A-C. (2005) Toward genome-wide SNP genotyping. *Nat. Genet.* 57, S5–S10
13 Macdonald, S.J. *et al.* (2005) A low-cost open-source SNP genotyping platform for association mapping applications. *Genome Biol.* 6, R105
14 Spielman, R.S. *et al.* (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.* 52, 506–516
15 Mitchell, A.A. and Chakravarti, A. (2003) Undetected genotyping errors cause apparent overtransmission of common alleles in the transmission/disequilibrium test. *Am. J. Hum. Genet.* 72, 598–610
16 Gordon, D. *et al.* (2001) A transmission/disequilibrium test that allows for genotyping errors in the analysis of single-nucleotide polymorphism data. *Am. J. Hum. Genet.* 69, 371–380
17 Gordon, D. *et al.* (2004) A transmission disequilibrium test for general pedigrees that is robust to the presence of random genotyping errors and any number of untyped parents. *Eur. J. Hum. Genet.* 12, 752–761
18 Allen, A.S. *et al.* (2003) Informative missingness in genetic association studies: case-parent designs. *Am. J. Hum. Genet.* 72, 671–680
19 Laird, N.M. and Lange, C. (2006) Family-based designs in the age of large-scale gene-association studies. *Nat. Rev. Genet.* 7, 385–394
20 Stich, B. (2006) A new test for family-based association mapping with inbred lines from plant breeding programs. *Theor. Appl. Genet.* 113, 1121–1130
21 Devlin, B. and Roeder, K. (1999) Genomic control for association studies. *Biometrics* 55, 997–1004
22 Reich, D.E. and Goldstein, D.B. (2001) Detecting association in a case-control study while correcting for population stratification. *Genet. Epidemiol.* 20, 4–16
23 Bacanu, S-A. *et al.* (2002) Association studies for quantitative traits in structured populations. *Genet. Epidemiol.* 22, 78–93
24 Devlin, B. *et al.* (2004) Genomic control in the extreme. *Nat. Genet.* 36, 1129–1130
25 Clayton, D.G. *et al.* (2005) Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat. Genet.* 37, 1243–1246
26 Zheng, G. *et al.* (2006) Robust genomic control for association studies. *Am. J. Hum. Genet.* 78, 350–356
27 Setakis, E. *et al.* (2006) Logistic regression protects against population structure in genetic association studies. *Genome Res.* 16, 290–296
28 Price, A.L. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38, 904–909
29 Pritchard, J.K. *et al.* (2000) Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959
30 Pritchard, J.K. *et al.* (2000) Association mapping in structured populations. *Am. J. Hum. Genet.* 67, 170–181

31 Falush, D. *et al.* (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164, 1567–1587
32 Aranzana, M.J. *et al.* (2005) Genome-wide association mapping in *Arabidopsis* identifies previously known flowering time and pathogen resistance genes. *PLoS Genet* 1, e60
33 Yu, J. *et al.* (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* 38, 203–208
34 Kraakman, A.T. *et al.* (2004) Linkage disequilibrium mapping of yield and yield stability in modern spring barley cultivars. *Genetics* 168, 435–446
35 Buntjer, J.B. *et al.* (2005) Haplotype diversity: the link between statistical and biological association. *Trends Plant Sci.* 10, 466–471
36 Caldwell, K.S. *et al.* (2006) Extreme population-dependent linkage disequilibrium detected in an inbreeding plant species, *Hordeum vulgare*. *Genetics* 172, 557–567
37 Maynard Smith, J. and Haig, J. (1974) The hitchhiking effect of a favourable gene. *Genet. Res.* 23, 23–35
38 Barton, N. (2000) Genetic hitchhiking. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 355, 1553–1562
39 Palaisa, K. *et al.* (2004) Long-range patterns of diversity and linkage disequilibrium surrounding the maize Y1 gene are indicative of an asymmetric selective sweep. *Proc. Natl. Acad. Sci. U. S. A.* 101, 9885–9890
40 Tian, D. *et al.* (2002) Signature of balancing selection in Arabidopsis. *Proc. Natl. Acad. Sci. U. S. A.* 99, 11525–11530
41 Darvasi, A. and Shifman, S. (2005) The beauty of admixture. *Nat. Genet.* 37, 118–119
42 Smith, M.W. and O'Brien, S.J. (2005) Mapping by admixture linkage disequilibrium: advances, limitations and guidelines. *Nat. Rev. Genet.* 6, 623–632
43 Hu, Z-M. (2005) Detection of linkage disequilibrium QTLs controlling low-temperature growth and metabolite accumulations in an admixed breeding population of Leymus wildryes. *Euphytica* 141, 263–280
44 Lexer, C. *et al.* (2007) Admixture in European Populus hybrid zones makes feasible the mapping of loci that contribute to reproductive isolation and trait differences. *Heredity* 98, 74–84