Annu. Rev. Plant Biol. 2003. 54:357–74
doi: 10.1146/annurev.arplant.54.031902.134907

STRUCTURE OF LINKAGE DISEQUILIBRIUM IN PLANTS*

Sherry A. Flint-Garcia
Department of Genetics, North Carolina State University, Raleigh, North Carolina 27695; email: saflintg@unity.ncsu.edu

Jeffry M. Thornsberry and Edward S. Buckler IV
USDA-ARS, Plant Science Research Unit, Raleigh, North Carolina 27695; Department of Genetics, North Carolina State University, Raleigh, North Carolina 27695; email: jthornsb@unity.ncsu.edu, buckler@statgen.ncsu.edu

Key Words gametic phase disequilibrium, allelic association, association mapping, recombination, quantitative trait loci (QTL)

■ **Abstract** Future advances in plant genomics will make it possible to scan a genome for polymorphisms associated with qualitative and quantitative traits. Before this potential can be realized, we must understand the nature of linkage disequilibrium (LD) within a genome. LD, the nonrandom association of alleles at different loci, plays an integral role in association mapping, and determines the resolution of an association study. Recently, association mapping has been exploited to dissect quantitative trait loci (QTL). With the exception of maize and *Arabidopsis*, little research has been conducted on LD in plants. The mating system of the species (selfing versus outcrossing), and phenomena such as population structure and recombination hot spots, can strongly influence patterns of LD. The basic patterns of LD in plants will be better understood as more species are analyzed.

CONTENTS

INTRODUCTION

One hallmark of twentieth-century genetics will be the tremendous strides made in understanding how individual genes control simple traits (phenotypes). However, the fruits of the revolution in molecular genetics will likely be seen in this century, when the genes and alleles that control complex traits [quantitative trait loci (QTL)] are identified and understood. Currently, F_1-derived mapping populations are the key tool for identifying the genetic basis of quantitative traits. An alternative is to use natural populations to map traits by means of association analysis. Association analysis, or linkage disequilibrium (LD) mapping, has been used extensively to dissect human diseases, most notably Alzheimer's disease (2) and cystic fibrosis (27). This approach has recently been extended to plants, thereby increasing mapping resolution substantially over the current capabilities of standard mapping populations. Association analysis has the potential to identify a single polymorphism within a gene that is responsible for the difference in phenotype. In addition, many plant species have high levels of diversity for which association approaches are well suited to evaluate the numerous alleles available.

LD plays a central role in association analysis. The distance over which LD persists will determine the number and density of markers, and experimental design needed to perform an association analysis. For these reasons, it is important to understand LD and to determine the extent of LD in the species under investigation. In this review we describe LD, summarize what is known about variation in LD among species, and comment on the application of LD in the dissection of quantitative traits. We also discuss future key issues surrounding LD that will be important for its application.

What Is LD?

LD is also known as gametic phase disequilibrium, gametic disequilibrium, and allelic association. Simply stated, LD is the "nonrandom association of alleles at different loci." It is the correlation between polymorphisms [e.g., single nucleotide polymorphisms (SNPs)] that is caused by their shared history of mutation and recombination. In a large, randomly mated population with loci segregating independently, but in the absence of selection, mutation, or migration, polymorphic loci will be in linkage equilibrium (10). In contrast, linkage, selection, and admixture will increase levels of LD.

The terms linkage and LD are often confused. Although LD and linkage are related, they are distinctly different. Linkage refers to the correlated inheritance of loci through the physical connection on a chromosome, whereas LD refers to the correlation between alleles in a population. The confusion occurs because tight linkage may result in high levels of LD. For example, if two mutations occur within a few bases of one another, they undergo the same pressures of selection and drift through time. Because recombination between the two neighboring bases is rare, the presence of these SNPs is highly correlated and the tight linkage will result in high LD. In contrast, SNPs on separate chromosomes experience different selection pressures and independent segregation, so these SNPs have a much lower correlation or level of LD.

How Is LD Measured?

A variety of statistics have been used to measure LD. Delvin & Risch (4), and more recently Jorde (26), have reviewed the relative advantages and disadvantages of each statistical approach. Here, we introduce the two most common statistics for measuring LD: r^2 and D'. Consider a pair of loci with alleles A and a at locus one, and B and b at locus two, with allele frequencies π_A, π_a, π_B, and π_b, respectively. The resulting haplotype frequencies are π_{AB}, π_{Ab}, π_{aB}, and π_{ab}. The basic component of all LD statistics is the difference between the observed and expected haplotype frequencies,

$$D_{ab} = (\pi_{\mathrm{AB}} - \pi_{\mathrm{A}}\pi_{\mathrm{B}}).$$

The distinction between these statistics lies in the scaling of this difference.

The first of the two measures, r^2, also described in the literature as Δ^2, is calculated as

$$r^2 = \frac{(D_{ab})^2}{\pi_{\mathrm{A}}\pi_{\mathrm{a}}\pi_{\mathrm{B}}\pi_{\mathrm{b}}}.$$

It is convenient to consider r^2 as the square of the correlation coefficient between the two loci (21). However, unless the two loci have identical allele frequencies, a value of 1 is not possible. Statistical significance (*P*-value) for LD is usually calculated using either Fisher's exact test (12) to compare sites with two alleles at each locus, or multifactorial permutation analysis (56) to compare sites with more than two alleles at either or both loci.

Alternatively, the LD statistic D' (34) is calculated as

$$|D'| = \frac{(D_{ab})^2}{\min(\pi_{\mathrm{A}}\pi_{\mathrm{b}},\, \pi_{\mathrm{a}}\pi_{\mathrm{B}})} \text{ for } D_{ab} < 0;$$

$$|D'| = \frac{(D_{ab})^2}{\min(\pi_{\mathrm{A}}\pi_{\mathrm{B}},\, \pi_{\mathrm{a}}\pi_{\mathrm{b}})} \text{ for } D_{ab} > 0.$$

D' is scaled based on the observed allele frequencies, so it will range between 0 and 1 even if allele frequencies differ between the loci. D' will only be less than

1 if all four possible haplotypes are observed; hence, a presumed recombination event has occurred between the two loci.

The statistics r^2 and D' reflect different aspects of LD and perform differently under various conditions. Figure 1 presents three scenarios of how linked polymorphisms may exhibit different levels of LD. Figure 1*a* shows an example of absolute LD, where the two polymorphisms are completely correlated with one another. An instance when absolute LD can develop is when two linked mutations occur at a similar point in time and no recombination has occurred between the sites. In this case, the history of mutation and recombination for the sites is the same. Other processes, such as genetic drift, can produce similar patterns, as discussed below. Both r^2 and D' have a value of 1 in this scenario. Figure 1*b* shows an example of LD when the polymorphisms are not completely correlated, but there is no evidence of recombination. One way this type of LD structure can develop is when the mutations occur on different allelic lineages. This situation can reflect the same recombinational history, but different mutational histories. This is the situation in which r^2 and D' act differently, with D' still equal to 1, but where r^2 can be much smaller. Figure 1*c* shows an example of when polymorphisms are in linkage equilibrium. If the sites are linked, then equilibrium could be produced through a recombination event between the two sites. In this case, the recombinational history differs for the various haplotypes, but the mutational history is the same. Hence, both r^2 and D' will be zero.

Although neither r^2 nor D' perform extremely well with small sample sizes and/or low allele frequencies, each has distinct advantages. Whereas r^2 summarizes both recombinational and mutational history, D' measures only recombinational history and is therefore the more accurate statistic for estimating recombination differences. However, D' is strongly affected by small sample sizes, resulting in highly erratic behavior when comparing loci with low allele frequencies. This is due to the decreased probability of finding all four allelic combinations of low frequency polymorphisms even if the loci are unlinked. For the purpose of examining the resolution of association studies, we generally favor the r^2 statistic, as it is indicative of how markers might correlate with the QTL of interest.

There are two common ways to visualize the extent of LD between pairs of loci. LD decay plots are used to visualize the rate at which LD declines with genetic or physical distance (Figure 2). Scatter plots of r^2 values versus genetic/physical distances between all pairs of alleles within a gene, along a chromosome, or across the genome are constructed. Alternatively, disequilibrium matrices are effective for visualizing the linear arrangement of LD between polymorphic sites within a gene or loci along a chromosome (Figure 3). It should be noted that LD decay is unpredictable. Both plot types highlight the random variation in LD owing to a variety of forces discussed below.

What Affects LD?

Because allele frequency and recombination between sites affect LD, most of the processes observed in population genetics are reflected in LD patterns.

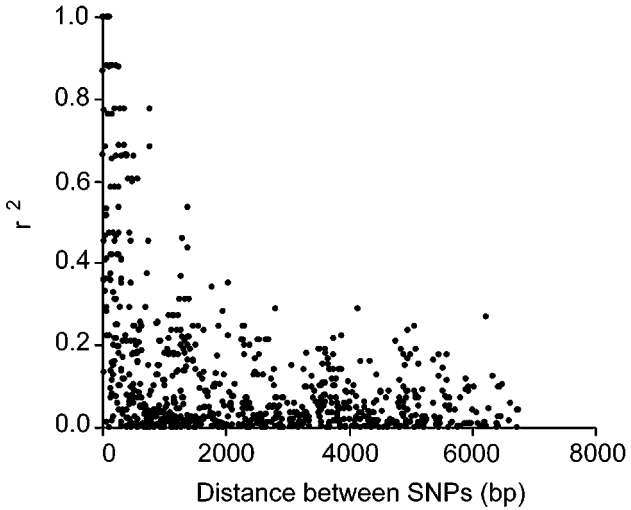

Figure 2 Linkage disequilibrium (LD) decay plot of *shrunken 1* (*sh1*) in maize. LD, measured as r^2, between pairs of polymorphic sites is plotted against the distance between the sites. For this particular gene, LD decayed within 1500 bp. Data from Reference 48.

Mutation provides the raw material for producing polymorphisms that will be in LD. Recombination is the main phenomenon that weakens intrachromosomal LD, whereas interchromosomal LD is broken down by independent assortment. Population size also plays an important role. In small populations, the effects of genetic drift result in the consistent loss of rare allelic combinations, which increase LD levels. When genetic drift and recombination are at equilibrium,

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

where N is the effective population size and c is the recombination fraction between sites (56).

Population mating patterns and admixture can strongly influence LD. Generally, LD decays more rapidly in outcrossing species as compared to selfing species (36; see discussion below). This is because recombination is less effective in selfing species, where individuals are more likely to be homozygous, than in outcrossing species. Admixture is gene flow between individuals of genetically distinct populations followed by intermating. Admixture results in the introduction of chromosomes of different ancestry and allele frequencies. Often, the resulting LD extends to unlinked sites, even on different chromosomes, but breaks down rapidly with random mating (41).

LD can also be created in populations that have recently experienced a reduction in population size (bottleneck) with accompanying extreme genetic drift (7). During a bottleneck, only few allelic combinations are passed on to future

generations. This can generate substantial LD. In human genetic studies, populations that have undergone severe bottlenecks (e.g., Finnish and Afrikaner populations) have been used in LD mapping a number of disease traits (19). Selection, which produces locus-specific bottlenecks, also causes LD between the selected allele at a locus and linked loci. Moreover, selection for or against a phenotype controlled by two unlinked loci (epistasis) may result in LD despite the fact that the loci are not physically linked.

DISSECTING TRAITS

Historically, linkage analysis was used to measure the genetic proximity of loci to each other, to map qualitative traits, and, more recently, to map QTL. In plants, most of these types of cosegregation analyses have been conducted in highly structured populations with known pedigrees, such as F_2 populations. However, these populations have two major limitations. First, the limited number of recombination events results in poor resolution for quantitative traits. Second, only two alleles at any given locus can be studied simultaneously (Figure 4). In order to increase the resolution of mapping populations, large recombinant inbred line populations that have undergone several rounds of random mating have been created for several plant species [e.g., the intermated B73 $\times$ Mo17 (IBM) population in maize (33)]. These rounds of mating increase the potential number of recombination events. Despite these efforts, the resolution for many QTL is still several centimorgans, corresponding to hundreds of genes. Additionally, the low number of alleles sampled per locus in each population makes it difficult to examine the full range of genetic diversity available for many plant species.

An increasingly common method of refining the identification of QTL is the production of near isogenic lines (NILs) and positional cloning. Technical limitations, such as the lack of contiguous coverage and the large amounts of repetitive DNA in the genomes of many plant species, prevent the successful implementation of positional cloning by means of chromosome walking. Aside from these technical issues, positional cloning may not be efficient at identifying genes responsible for complex traits. This is due in part both to the difficulty of developing NILs for loci that explain less than 20% of the variance and to constraints created by only using two alleles. The majority of genes cloned via positional cloning explain large portions of the phenotypic variation, e.g., "*fruit weight2.2*" in tomato, "*teosinte branched1* (*tb1*)" in maize, "*heading date1*" in rice, and *FRIGIDA* and *CRYPTOCHROME2* in *Arabidopsis* (6, 8, 13, 25, 59). The productin of NILs is also time consuming, especially for long-generation species. However, when nothing is known about the genes in a particular pathway, positional cloning may be the best option.

Linkage analysis has not been successful in fine-scale mapping of disease loci in humans because construction of organized pedigrees from controlled breeding crosses is not possible. Even when studying families with high occurrence of a disease, it is often difficult to find direct evidence of genetic recombination between

polymorphic sites. The medical community turned to association analysis because there were too few meioses in most families to finely map diseases.

Association analysis, also known as LD mapping or association mapping, is a population-based survey used to identify trait-marker relationships based on LD. Unlike linkage analysis, where familial relationships are used to predict correlations between phenotype and genotype, association methods rely on previous, unrecorded sources of disequilibrium to create population-wide marker-phenotype associations (22). Genetic diversity is evaluated across natural populations to identify polymorphisms that correlate with phenotypic variation. Association analysis is extremely powerful because the individuals that are sampled do not have to be closely related, which harnesses all of the meiotic and recombination events between those individuals to improve resolution [reviewed in (9, 29, 31)]. Because of these recombination events, only markers in LD with a disease or trait of interest will associate with the disease or trait. Association analysis was responsible for the identification and cloning of the cystic fibrosis gene (27), the diastrophic dysplasia gene (19), and one of the major Alzheimer's factors (2).

Association analysis recently emerged as a powerful tool to identify QTL in plants. The first association study of a quantitative trait based on a candidate gene was the analysis of flowering time and the *dwarf8* (*d8*) gene in maize (53). This putative transcription factor has been implicated as playing a role in the "Green Revolution" varieties of wheat (39), and the *Arabidopsis* ortholog has been shown to play a role in regulating flowering time variation (58). In the association study, variation in *d8* was evaluated for association with flowering time and plant height in 92 maize inbred lines. Nine polymorphisms, including a MITE insertion in the promoter region and a two–amino acid deletion adjacent to the SH2-like domain, a potentially key binding domain in this putative transcription factor, were found to associate with flowering time. The distribution of polymorphisms indicated that selection had occurred at this locus to produce earlier flowering maize. The two–amino acid deletion had an estimated effect of reducing time to flowering by 7–11 days. LD decayed rapidly such that no association was found between flowering time and *tb1*, located just 1 cM from *d8*.

The *sugary1* (*su1*) gene in maize is responsible for the production of naturally occurring varieties of sweet corn. In a recent survey of allelic diversity at this locus, association methods were used to map the mutation to a single nucleotide (57) even though approximately 150 mutations were segregating within the 12 kb of this locus. There was little recombination within the locus, but the association survey found diverse alleles that were key to resolving the functional mutation to a single nucleotide. Biochemical and molecular studies confirmed that the identified nucleotide is the functional cause (5).

The first association study to attempt a genome scan in plants was conducted in sea beet (*Beta vulgaris* ssp. *maritima*), a wild relative of sugar beet (*Beta vulgaris* ssp. *vulgaris*) (18). Growth habit in beet, whether the plant requires vernalization prior to bolting, is determined by a single gene, the bolting (*B*) gene. Two of 440 genome-wide AFLP markers were significantly associated with the *B* gene.

However, when three markers that were located within 1.5 cM of the *B* gene were tested, only one showed weak association.

Association analysis can also be used to rule out a candidate gene as the gene underlying a QTL. In *Arabidopsis*, the *GLABROUS1* (*GL1*) locus was one of six genes identified as candidates for trichome initiation and density. *GL1* is a member of the R2R3-MYB class of transcription factors, which have homology to animal MYB DNA binding domains transcription factors. Hauser et al. (20) studied trichome density and sequence polymorphism across *GL1* in 28 *A. thaliana* accessions and several mutant (glabrous) lines. A cladistic analysis approach was used (51) to avoid statistical power problems related to the nonindependence of polymorphisms within *GL1*. When analysis of variance was used to analyze three different cladistic levels, *GL1* did not significantly associate with trichome density. The association analysis suggests that it is unlikely that GL1 plays a major role in natural variation for trichome density, although it could play a minor or epistatic role.

LD IN ANIMAL SYSTEMS

LD in Humans

LD has been studied extensively in humans (*Homo sapiens*) and has been reviewed recently by Pritchard & Przeworski (40). There is tremendous heterogeneity in human LD estimates because of differences in loci, marker types (microsatellites versus SNPs), sample populations, and chromosome type (sex chromosomes versus autosomes). In general, studies indicate that LD extends over large distances ranging from 60 kb (46) to 500 kb (50) (Table 1). However, over smaller regions (<10 kb) LD does not follow theoretical expectations, perhaps owing to gene

TABLE 1

Species	Mating system	LD range[a]	Reference
Human	Outcrossing		
Nigerian		5 kb	(46)
European		80 kb	(46)
Cattle	Outcrossing[b]	10 cM	(11)
Drosophila	Outcrossing	<1 kb	(35)
Maize	Outcrossing		
Diverse maize		1 kb	(52)
Diverse inbred lines		1.5 kb	(48)
Elite lines		>100 kb	(45)
Arabidopsis	Selfing	250 kb	(37)
Sugarcane	Outcrossing/vegetative propagation	10 cM	(23)

[a]The LD value provided is estimated where $r^2 = 0.10$.

[b]There is an extreme sex bias in terms of number of breeding males versus breeding females.

conversion, the nonreciprocal transfer of genetic information resulting from the mismatch repair system (44). These conflicting reports are likely the results of the evolutionary and recombinational history of the various regions of the genome. As expected, LD levels were higher for sex chromosomes than autosomes because recombination only occurs between X chromosomes in females (50). Large differences in LD estimates exist between human populations that have major differences in population history such as reproductive isolation or have undergone recent bottlenecks. The Out-of-Africa model suggests that ancestors of modern-day humans emigrated from Africa. This model is supported by the discovery that LD extends longer distances in Northern European populations than in more diverse Nigerian populations (46).

LD in Other Animal Systems

Although LD has been examined most comprehensively in human populations, LD studies have also been conducted in cattle (*Bos taurus*) and fruit flies (*Drosophila melanogaster*). Extensive LD between microsatellites in the Dutch black and white dairy cattle population (11) has been reported, extending in the range of tens of centimorgans, even to unlinked markers. It has been concluded that most of the LD observed in this population could be accounted for by bottlenecks caused by the globalization of semen trading. It is estimated that the top ten ranked sires account for 40% of the inseminations. Similar patterns may be observed in some elite varieties of important agricultural crops, where modern elites are the result of a small number of ancestors (see discussion below).

In *Drosophila*, LD decays within a few kilobases in the *delta* gene (35) and within 1 kb in the *su*(*s*) and *su*(w^a) regions on the X chromosome (32). The effects of bottlenecks on LD are also seen, where LD decays more rapidly in the diverse African populations than it does in non-African populations (54).

LD IN PLANT SYSTEMS

Although LD has been studied extensively in animal systems, little research has been conducted regarding LD in plant systems. Among plants, most of the LD research has been carried out in maize and *Arabidopsis*.

LD in Maize

In maize (*Zea mays* ssp. mays), several studies have been conducted to investigate LD over a wide range of population and marker types. The patterns of LD vary substantially with the population chosen. Tenaillon et al. (52) investigated sequence diversity at 21 loci on chromosome 1 in a diverse group of maize germplasm. LD, measured as r^2, decreased to less than 0.25 within 200 bp on average. Analysis of interlocus LD revealed little LD between loci, despite the fact that all the loci are located on the same chromosome.

In a similar study, Remington et al. (48) examined intragenic and genome-wide LD between SNPs in a diverse set of 102 inbred lines representing a sample of the genetic diversity commonly used in public-sector breeding worldwide. In a survey of six candidate genes, intragenic LD decayed rapidly ($r^2 < 0.1$ within 1500 bp). This rapid decay has also been observed in an additional set of 15 genes (J.M. Thornsberry, S.R. Whitt, L.M. Wilson, S.A. Flint-Garcia, S.A. Andaluz & E.S. Buckler, unpublished data). However, occasionally loci have been found that deviate from this rapid decay of LD. One gene that had little LD decay over 12 kb was the *su1* locus in maize. A recent analysis of selection suggests that this locus has been a recent target of selection during the domestication process (57), which is likely the source of the extensive LD observed. LD was also measured between 47 microsatellite markers to estimate LD across the genome (48). Greater levels of LD were detected between these markers than were detected using SNPs. This suggests that the rapidly evolving microsatellites may track recent population structure better than the relatively older SNPs.

Rafalski (45) reported that LD extends to greater than 100 kb for the *adh1* and *y1* loci in elite maize populations, which have an even more narrow germplasm base than the inbred lines reported by Tenaillon et al. (52) and Remington et al. (48). In this same elite germplasm, Rafalski and collaborators also looked at LD decay over a 300–500 bp range for 18 genes and found virtually no LD decay (1).

Labate et al. (30) examined LD between RFLP loci in the two synthetic populations that have been randomly mated for many generations. These populations were derived from 12–16 progenitor inbred lines. Each original population has undergone recurrent selection for 12 generations. It is interesting to note that the populations responded differently to selection. One population substantially increased in LD over the 12 generations, whereas the other decreased.

There are several explanations for why the LD patterns are so different between these samples. First, most of the diversity in maize is descended from an extremely variable outcrossing wild relative with large effective population sizes. Most of the observed recombinant haplotypes were probably generated before domestication of this wild relative. Hence, the different rates of LD decay reflect differing levels of population bottleneck, i.e., the progression from diverse landraces to diverse inbreds to elite inbreds. Additionally, the LD reported between loci 100 kb apart likely includes recombinationally inactive repetitive regions of the genome, which are not present in the other studies.

LD in *Arabidopsis*

The LD pattern in *Arabidopsis* (*Arabidopsis thaliana*) is a sharp contrast to the pattern in maize. As expected, LD extends much farther in *Arabidopsis* because it is a highly selfing species (36). Hagenblad & Nordborg (17) sequenced 14 short fragments from a 400 kb region of the flowering time locus *FRIGIDA*. They found that LD decayed within 250 kb, equivalent to 1 cM. Strong LD was seen between sites that were closely linked. Analysis of 163 genome-wide SNPs in 76 accessions also revealed that LD decayed within 250 kb (37).

To investigate founder effects in Arabidopsis, LD was investigated around the disease locus *RPM1* in several recently founded populations isolated in Michigan (37). In these populations, LD extended tens of centimorgans, on the order of megabases. This extensive LD may be due to the limited number of recombination events that have occurred over the past 200 years.

LD in Other Plant Species

Sugarcane (*Saccharum* spp.) exhibits extensive long-range LD, approximately 10 cM (23). This is not surprising considering the bottleneck in the breeding history of modern sugarcane cultivars. The majority of modern cultivars were derived from the interspecific cross between *S. officinarum* and *S. spontaneum*, followed by multiple backcrosses to *S. officinarum*. Because sugarcane is propagated vegetatively, the resulting cultivars generally resulted from fewer than 10 meioses since the first interspecific cross. In this study, LD was investigated between RFLP loci in 59 cultivars. The majority of the locus pairs in significant LD were physically linked on the same chromosome. However, 14% of the cases of significant LD involved loci on different chromosomes. Jannoo and colleagues believe that the overall estimate of LD may be exaggerated because of the polyploid nature of sugarcane (23). They also point out that most of the pairs of loci in LD are derived from an *S. spontaneum* parent reflecting the phenomenon of homeologus pairing.

CURRENT ISSUES RELATED TO LD

Currently, the basic structure of LD is understood for two plant species. There are still many issues that need to be better studied and resolved before LD can be used routinely to dissect complex traits.

Population Structure

Association testing or LD mapping has been used with mixed results in studies of human genetic diseases and quantitative traits of *Drosophila* and has not been used in plant systems until recently. The reluctance to use this technique in plant systems and the mixed results seen in animal systems is due in large part to the effects of population structure. The presence of population stratification and an unequal distribution of alleles within these groups can result in nonfunctional, spurious associations (28). Highly significant LD between polymorphisms on different chromosomes may produce associations between a marker and a phenotype, even though the marker is not physically linked to the locus responsible for the phenotypic variation (41).

The classical example of interference by population structure involved a study of the occurrence of type 2 diabetes in the Pima and Papago Native American tribes from southern Arizona (28). Researchers found a correlation between a particular haplotype at the immunoglobulin G locus and reduced incidence of diabetes.

Further analysis showed that the diabetic individuals had a lower proportion of European ancestry relative to controls and that the important immunoglobulin haplotype was more prevalent in Europeans than in Native Americans. When the analysis was restricted to individuals with similar proportions of European ancestry, the association was no longer detected, suggesting that the original association was due to the effects of population admixture.

Similar population structure exists in many plant systems. The complex breeding history of many important crops and the limited gene flow in most wild plants have created complex stratification within the germplasm, which complicates association studies (49). Association tests that do not attempt to account for the effects of population structure must be viewed with skepticism. However, recent developments in statistical methodologies make it possible to properly interpret the results of association tests. All of these methods rely on the use of independent marker loci to detect stratified populations and to correct for them (41). These methods assume that population structure has similar effects on all loci. The method developed by Reich & Goldstein (47) examines the association of a moderate number of unlinked genetic markers with a given phenotype. The strength of these associations is then compared with the association of a candidate gene. Pritchard et al. (42, 43) have developed an approach that incorporates estimates of population structure directly into the association test statistic. In the first empirical application of these methods, the Pritchard approach was modified for use with quantitative traits and used to study flowering time in maize (53). In this study, a suite of polymorphisms in the maize *dwarf8* gene was significantly associated with variation in flowering time. The incidence of false positives created by population structure was reduced by up to 80% as a result of the Pritchard method. Using these statistical methods in an association test allowed researchers to improve their resolution from the level of a 20-cM region to that of an individual gene. The methodological advances that estimate the effects of population structure–induced linkage disequilibria should allow the use of association testing in a much wider context, enabling the use of this very powerful technique.

Selfing Versus Outcrossing Species

Nordborg & Donnelly (38) showed that effective recombination rate is related to the degree of selfing that a species exhibits. This is because recombination is less effective in selfing species where individuals are more likely to be homozygous at a given locus than in outcrossing species. Although physical recombination may occur more often in selfing species, recombination is rarely between distinct alleles; hence, the amount of effective recombination is fairly low. The effective recombination rate, c, is related to selfing fraction, s, as follows:

$$c = 1 - \frac{s}{2 - s}.$$

For species that self two thirds of the time, the effective recombination rate is reduced by half, relative to an obligate outcrossing (i.e., self-incompatible) species. There is a 10-fold reduction in effective recombination for species that are 95% selfing and a 50-fold reduction for species that are 99% selfing. This relationship between recombination and selfing can extend to LD. Because effective recombination is reduced severely in highly selfing species, LD will be more extensive. As mentioned above, LD is proportional to the recombination fraction. In coalescent simulations, high levels of selfing greatly increase levels of LD (36). For example, in outcrossing species ($s = 0$) LD will decay within 500 bp, but for highly selfing species ($s = 0.95$) LD may extend to 10 kb.

The effect of a mating system on LD is exemplified in the comparison of maize and *Arabidopsis*, where a 250-fold difference in LD is observed. LD in maize typically decays to r^2 within several kilobases, whereas LD persists for hundreds of kilobases in *Arabidopsis*. These differences in LD between selfing and outcrossing species have crucial implications with respect to association analysis. The number of markers needed to cover the genome is determined by the extent of LD and, therefore, will differ for selfing and outcrossing species. Tenaillon et al. (52) estimated that, in maize, a marker density of one SNP per 100 to 200 bp is needed to maintain power in association analysis. In contrast, polymorphic markers are needed only every 50 kb to cover the genome in *Arabidopsis* (36). These estimates from maize and *Arabidopsis* demonstrate the massive effect that a mating system can have on experimental design. To date, no studies have been conducted in soybean to estimate the extent of LD, but LD is likely to extend great distances owing to its selfing history and limited germplasm base (3).

One must be cautious, however, when predicting the structure of LD based on the present-day mating system because the mating system may have changed significantly, whether by natural evolutionary processes or by human intervention. For example, the cultivated soybean (*Glycine max*) and its ancestor, *Glycine soja*, differ significantly in their outcrossing rates. The self-pollinating *G. max* has an outcrossing rate of approximately 1%, whereas *G. soja* outcrosses at an average rate of 13% (16). For low-resolution mapping, cultivated varieties may be best, whereas fine mapping would be more effective in populations of wild relatives. Based on observed selfing rates, wild relatives should have at least 11-fold higher resolution. Because selfing rates can change rapidly, it is necessary to empirically determine the LD structure before employing association-based methods.

Recombination Hot Spots

A major unresolved question is how genome structure and the rate of recombination affect the structure of LD across the genome. It is generally accepted that different regions of genomes undergo different rates of recombination. Jeffreys et al. (24) documented the relationship between LD decay and hot spots in recombination for the major histocompatibility complex (MHC) in humans. Using SNPs, they found

that the MHC region had large blocks with very high LD that were interrupted by areas where LD broke down rapidly. This LD structure reflects the historical effects of recombination in the region. The authors then used sperm typing to estimate recombination frequency across the MHC regions to map the hot spots. Recombination rates varied by more than 1000-fold. Many of these hot spots coincide with sites where LD breaks down rapidly, suggesting that these regions of cold- and hot-spots may persist over evolutionary time. Furthermore, the hot spots were clustered in groups of 60 to 90 kb, but the clusters did not necessarily correspond to genes.

In maize, there is extensive evidence for tremendous heterogeneity in rates of recombination across the genome (55). One study examined a 140-kb region with four genes and substantial amounts of repetitive DNA and retroelements (60). The recombination rates per basepair varied almost 1000-fold across this region, with virtually no recombination events in the repetitive regions. However, the correlation between genes and recombination events was not perfect: Two genes were hot spots, one gene was not, and one nongenic region was a hot spot. There is also evidence that gene-rich stretches are likely to have more recombination than methylated, gene-poor regions (15). One reason for decreased recombination in various regions is that the retrotransposon composition can be entirely different between two alleles, as illustrated by the *bronze1* locus in maize (14).

To date, the direct connection between the present locations of hot spots and structure of LD produced through evolution has not been demonstrated in plants. However, it is likely that this connection does exist, as in humans. This suggests that predicting LD levels between two sets of polymorphisms based solely on physical distance will be problematic. For example, two sites at either end of a 5-kb gene might have very little LD if the gene is a hot spot, whereas two sites on either side of 100 kb of retrotransposons could have very high levels of LD. The design of LD mapping experiments and placement of SNPs will require a thorough understanding of how these hot spots are dispersed.

LD and the Future of Genome Dissection

Association approaches have been the main application of LD, but the nature of LD in the population determines what type of association approach can be conducted. The rate of LD decay determines whether genome scans versus candidate gene–based association approaches can be used. In genome scans, markers are distributed across the genome to evaluate all genes simultaneously. For example, the human genome may require 70,000 markers, *Arabidopsis* 2000 markers, diverse maize landraces 750,000 markers, but only 50,000 markers for elite maize lines. For species other than *Arabidopsis*, this is an unwieldy number of markers, although technological improvements in the foreseeable future will likely enable the scoring of the necessary number of markers. However, more problematic than the genotyping is the large number of resources needed for phenotyping and statistical issues. When scoring 50,000 SNPs across the genome, there is a large

multiple-test problem, as thousands of independent tests are being conducted. Correcting these multiple tests would require extremely low *P*-values for each independent test. Statistical significance in a genome scan could only be obtained with large sample sizes of thousands of individuals for QTL that explain modest amounts of variation.

There are two ways to circumvent this problem: Either populations with greater levels of LD can be chosen, or the analysis can be restricted to candidate gene regions. By choosing a bottlenecked population, one can substantially increase genome-wide LD. Many human geneticists have used this approach, focusing on bottlenecked human populations (19). The limitation of this approach is that the appropriate populations must be identified, and by their nature, these bottlenecked populations will only contain a subset of the total variation. This approach of finding bottlenecked populations could work well in high diversity/low LD species such as maize, where Rafalski (45) suggested that elite germplasm with its high levels of LD would be ideal for low-resolution association approaches. Again, it is necessary to point out that novel alleles outside the elite germplasm will not be identified.

The candidate gene–association approaches rely on combining multiple lines of evidence to restrict the numbers of genes that are evaluated. Genome sequencing, comparative genomics, transcript profiling, low-resolution QTL analysis, and large-scale knockouts all provide opportunities to develop and refine candidate gene lists. These approaches are powerful at identifying candidate genes, but not at evaluating allelic affects. The candidate gene approach can substantially reduce the amount of genotyping required, but most importantly, it can reduce the multiple issues created by testing thousands of sites across the genome. The statistical issues in combining these disparate types of evidence have not been resolved.

In plants, another way to conduct a genomic scan is to use F_1-derived mapping populations. These populations are efficient for doing a genome scan, as often only a few hundred markers are needed. Because only two alleles are being evaluated, these populations will have more statistical power to evaluate the effect of a chromosomal region in comparison to association mapping. Additionally, there is more statistical power to evaluate epistasis. The advantages of association mapping in terms of resolution, speed, and allelic range are complementary to the strengths of F_2-based QTL mapping, namely, marker efficiency and statistical power.

SUMMARY

Plant genomics is beginning to allow the merger of molecular and biochemical approaches with quantitative genetics, and LD will likely play a key role in this merger. Association mapping may play an important role in identifying and evaluating the basis of quantitative variation in a wide range of species. But the key to designing and carrying out these association analyses is a thorough understanding of the LD structure across a wide range of species.

The *Annual Review of Plant Biology* is online at http://plant.annualreviews.org

LITERATURE CITED

1. Ching A, Caldwell KS, Jung M, Dolan M, Smith OS, et al. 2002. SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. *Biomed. Central Genetics* 3:19–32
2. Corder EH, Saunders AM, Risch NJ, Strittmatter WJ, Schmechel DE, et al. 1994. Protective effect of apolipoprotein-E type-2 allele for late-onset Alzheimer-disease. *Nat. Genet.* 7:180–84
3. Delannay X, Rodgers DM, Palmer RG. 1983. Relative genetic contributions among ancestral lines to North American soybean cultivars. *Crop Sci.* 23:944–49
4. Delvin B, Risch N. 1995. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* 29:311–22
5. Dinges JR, Colleoni C, Myers AM, James MG. 2001. Molecular structure of three mutations at the maize *sugary1* locus and their allele-specific phenotypic effects. *Plant Physiol.* 125:1406–18
6. Doebley J, Stec A, Hubbard L. 1997. The evolution of apical dominance in maize. *Nature* 386:485–88
7. Dunning AM, Durocher F, Healey CS, Teare MD, McBride SE, et al. 2000. The extent of linkage disequilibrium in four populations with distinct demographic histories. *Am. J. Hum. Genet.* 67:1544–54
8. El-Assal S, Alanso-Blanco C, Peeters A, Raz V, Koornneef M. 2001. A QTL for flowering time in *Arabidopsis* reveals a novel allele of *CRY2*. *Nat. Genet.* 29:435–40
9. Ewens WJ, Spielman RS. 2001. Locating genes by linkage and association. *Theor. Popul. Biol.* 60:135–39
10. Falconer DS, Mackay TF. 1996. *Introduction to Quantitative Genetics*. Essex, UK: Longman Group Ltd. 464 pp.
11. Farnir F, Coppieters W, Arranz J-J, Berzi P, Cambisano N, et al. 2000. Extensive genome-wide linkage disequilibrium in cattle. *Genome Res.* 10:220–27
12. Fisher RA. 1935. The logic of inductive inference. *J. R. Stat. Soc. Ser. A* 98:39–54
13. Frary A, Nesbitt TC, Grandillo S, van der Knaap E, Cong B, et al. 2000. *fw2.2*: a quantitative trait locus key to the evolution of tomato fruit size. *Science* 289:85–88
14. Fu H, Dooner HK. 2002. Intraspecific violation of genetic colinearity and its implications in maize. *Proc. Natl. Acad. Sci. USA* 99:9573–78
15. Fu H, Zheng Z, Dooner HK. 2002. Recombination rates between adjacent genic and retrotransposon regions in maize vary by 2 orders of magnitude. *Proc. Natl. Acad. Sci. USA* 99:1082–87
16. Fujita R, Ohara M, Okazaki K, Shimamoto Y. 1997. The extent of natural cross-pollination in wild soybean (*Glycine soja*). *J. Hered.* 88:124–28
17. Hagenblad J, Nordborg M. 2002. Sequence variation and haplotype structure surrounding the flowering time locus *FRI* in *Arabidopsis haliana*. *Genetics* 161:289–98
18. Hansen M, Kraft T, Ganestam S, Säll T, Nilsson N-O. 2001. Linkage disequilibrium mapping of the bolting gene in sea beet using AFLP markers. *Genet. Res.* 77:61–66
19. Hästbacka J, de la Chapelle A, Kaitila I, Sistonen P, Weaver A, Lander E. 1992. Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. *Nat. Genet.* 2:204–11
20. Hauser M-T, Harr B, Schlötterer C. 2001. Trichome distribution in *Arabidopsis thaliana* and its close relative *Arabidopsis lyrata*: molecular analysis of the candidate gene *GLABROUS1*. *Mol. Biol. Evol.* 18:1754–63
21. Hill WG, Robertson A. 1968. Linkage disequilibrium in finite populations. *Theor. Appl. Genet.* 38:226–31

22. Jannink J-L, Bink MC, Jansen RC. 2001. Using complex plant pedigrees to map valuable genes. *Trends Plant Sci.* 6:337–42
23. Jannoo N, Grivet L, Dookun A, D'Hont A, Glaszmann JC. 1999. Linkage disequilibrium among modern sugarcane cultivars. *Theor. Appl. Genet.* 99:1053–60
24. Jeffreys AJ, Kauppi L, Neumann R. 2001. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat. Genet.* 29:217–22
25. Johanson U, West J, Lister C, Michaels S, Amasino R, Dean C. 2000. Molecular analysis of *FRIGIDA*, a major determinant of natural variation in *Arabidopsis* flowering time. *Science* 290:344–47
26. Jorde LB. 2000. Linkage disequilibrium and the search for complex disease genes. *Genome Res.* 10:1435–44
27. Kerem BS, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, et al. 1989. Identification of the cystic fibrosis gene: genetic analysis. *Science* 245:1073–80
28. Knowler WC, Williams RC, Pettitt DJ, Steinberg AG. 1988. $Gm^{3-5,13,14}$ and type 2 diabetes mellitus: an association in American Indians with genetic admixture. *Am. J. Hum. Genet.* 43:520–26
29. Kruglyak L. 1999. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat. Genet.* 22: 139–44
30. Labate JA, Lamkey KR, Lee M, Woodman W. 2000. Hardy-Weinberg and linkage equilibrium estimates in the BSSS and BSCB1 random mated populations. *Maydica* 45:243–55
31. Lander ES, Schork NJ. 1994. Genetic dissection of complex traits. *Science* 265: 2037–48
32. Langley CH, Lazzaro BP, Phillips W, Heikkinen E, Braverman JM. 2000. Linkage disequilibria and the site frequency spectra in the *su*(*s*) and *su*(w^a) regions of the *Drosophila melanogaster* X chromosome. *Genetics* 156:1837–52
33. Lee M, Sharopova N, Beavis WD, Grant D, Katt M, Blair DL, Hallauer AR. 2002. Expanding the genetic map of maize with the intermated B73 × MO17 (IBM) population. *Plant Mol. Bio.* 48:453–61
34. Lewontin RC. 1964. The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* 49:49–67
35. Long AD, Lyman RF, Langley CH, Mackay TF. 1998. Two sites in the *Delta* gene region contribute to naturally occurring variation in bristle number in *Drosophila melanogaster*. *Genetics* 149:999–1017
36. Nordborg M. 2000. Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization. *Genetics* 154:923–29
37. Nordborg M, Borevitz JO, Bergelson J, Berry CC, Chory J, et al. 2002. The extent of linkage disequilibrium in *Arabidopsis thaliana*. *Nat. Genet.* 30:190–93
38. Nordborg M, Donnelly P. 1997. The coalescent process with selfing. *Genetics* 146: 1185–95
39. Peng J, Richards D, Hartley N, Murphy G, Devos K, et al. 1999. 'Green revolution' genes encode mutant gibberellin response modulators. *Nature* 400:256–61
40. Pritchard JK, Przeworski M. 2001. Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* 69:1–14
41. Pritchard JK, Rosenberg NA. 1999. Use of unlinked genetic markers to detect population stratification in association studies. *Am. J. Hum. Genet.* 65:220–28
42. Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155:945–59
43. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. 2000. Association mapping in structured populations. *Am. J. Hum. Genet.* 67:170–81
44. Przeworski M, Wall JD. 2001. Why is there so little intragenic linkage disequilibrium in humans? *Genet. Res.* 77:143–51
45. Rafalski A. 2002. Applications of single nucleotide polymorphisms in crop genetics. *Curr. Opin. Plant Biol.* 5:94–100

46. Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, et al. 2001. Linkage disequilibrium in the human genome. *Nature* 411: 199–204
47. Reich DE, Goldstein DB. 2001. Detecting association in a case-control study while correcting for population stratification. *Genet. Epidemiol.* 20:4–16
48. Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, et al. 2001. Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proc. Natl. Acad. Sci. USA* 98: 11479–84
49. Sharbel TF, Haubold B, Mitchell-Olds T. 2000. Genetic isolation by distance in *Arabidopsis thaliana*: biogeography and postglacial colonization of Europe. *Mol. Ecol.* 9:2109–18
50. Taillon-Miller P, Bauer-Sardiña I, Saccone NL, Putzel J, Laitinen T, et al. 2000. Juxtaposed regions of extensive and minimal linkage disequilibrium in human *Xq25* and *Xq228*. *Nat. Genet.* 25:324–28
51. Templeton AR, Boerwinkle E, Sing CF. 1987. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and analysis of Alcohol Dehydrogenase activity in *Drosophila*. *Genetics* 117:343–51
52. Tenaillon MI, Sawkins MC, Long AD, Gaut RL, Doebley JF, et al. 2001. Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). *Proc. Natl. Acad. Sci. USA* 98: 9161–66
53. Thornsberry JM, Goodman MM, Doebley J, Kresovich S, Nielsen D, et al. 2001. *Dwarf*8 polymorphisms associate with variation in flowering time. *Nat. Genet.* 28: 286–89
54. Wall JD. 2001. Insights from linked single nucleotide polymorphisms: what we can learn from linkage disequilibrium. *Curr. Opin. Genet. Dev.* 11:647–51
55. Weil CF. 2002. Finding the crosswalks on DNA. *Proc. Natl. Acad. Sci. USA* 99:5763–65
56. Weir BS. 1996. *Genetic Data Analysis II*. Sunderland, MA: Sinauer. 376 pp.
57. Whitt SR, Wilson LM, Tenaillon MI, Gaut BS, Buckler ES. 2002. Genetic diversity and selection in the maize starch pathway. *Proc. Natl. Acad. Sci. USA.* 99:12959–62
58. Wilson RN, Heckman JW, Somerville CR. 1992. Gibberellin is required for flowering in *Arabidopsis-thaliana* under short days. *Plant Physiol.* 100:403–8
59. Yano M, Katayose Y, Ashikari M, Yamanouchi U, Monna L, et al. 2000. *Hd1*, a major photoperiod sensitivity quantitative trait locus in rice, is closely related to the *Arabidopsis* flowering time gene *CONSTANS*. *Plant Cell* 12:2473–83
60. Yao H, Zhou Q, Li J, Smith H, Yandeau M, et al. 2002. Molecular characterization of meiotic recombination across the 140-kb multigenic a1-sh2 interval of maize. *Proc. Natl. Acad. Sci. USA* 99:6157–62

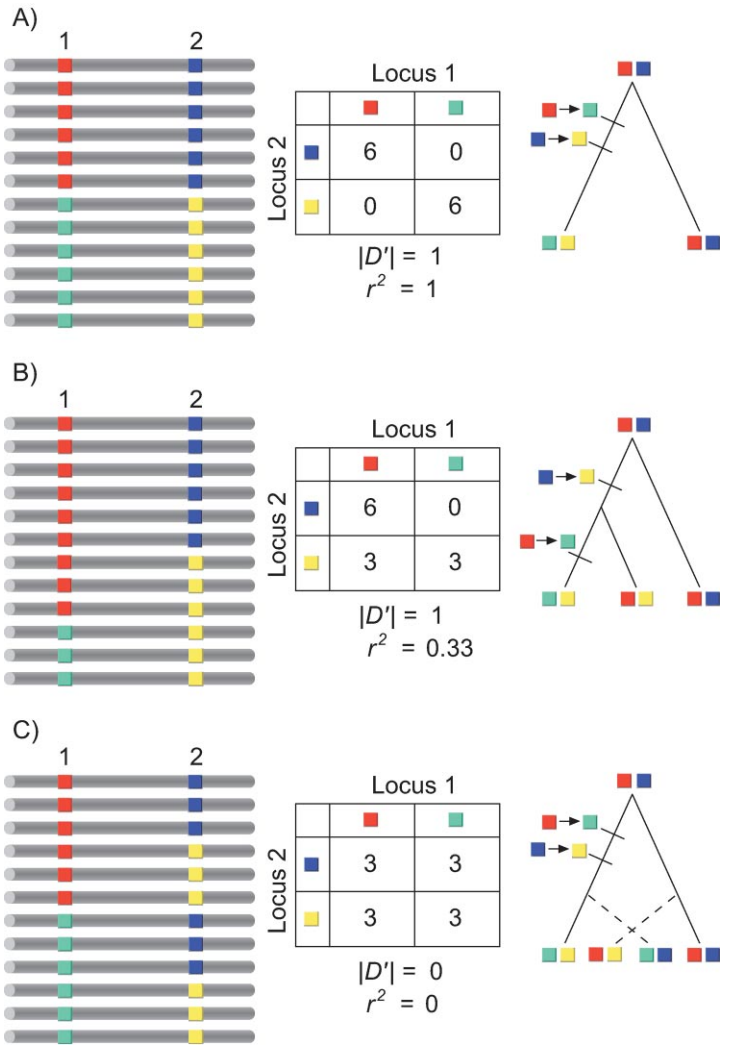

Figure 1 Hypothetical scenarios of linkage disequilibrium (LD) between linked polymorphisms caused by different mutational and recombinational histories demonstrating the behavior of the r^2 and D' statistics. Images in the *left* column represent the allelic states of two loci. The *middle* column represents the 2×2 contingency table of haplotypes and the resulting r^2 and D' statistics. The *right* column represents a possible tree responsible for the observed LD present. (*A*) Absolute LD exists when two loci share a similar mutational history with no recombination. Both r^2 and D' equal 1. (*B*) LD can result when mutations occur on different lineages without recombination between the loci. Notice the large difference in measures of LD as calculated by r^2 and D'. (*C*) Linkage equilibrium is produced when there is recombination between loci, regardless of mutational history. In this situation, both r^2 and D' equal 0. Adapted from Rafalski (44).

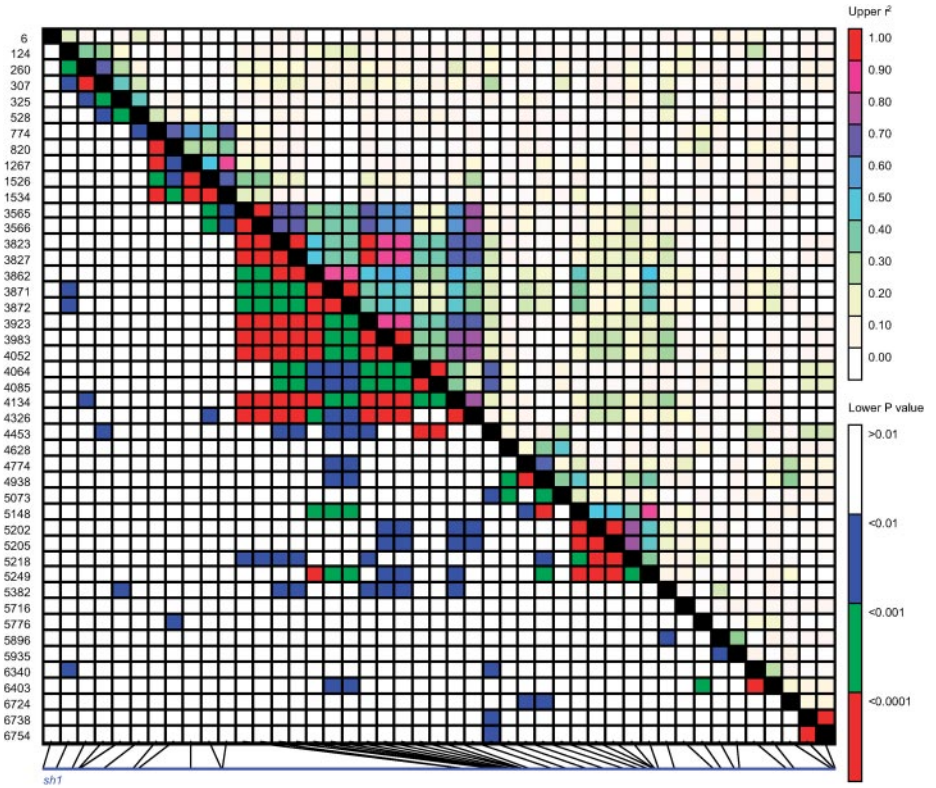

Figure 3 Disequilibrium matrix for polymorphic sites within *sh1*. Polymorphic sites are plotted on both the X-axis and Y-axis. Pairwise calculations of linkage disequilibrium (LD) (r^2) are displayed above the diagonal with the corresponding *P*-values for Fisher's exact test displayed below the diagonal. Coloration is indicative of the corresponding *P*-value or r^2 values from the bars on right. Notice that some blocks of LD do persist over larger distances within the gene, which do not necessarily correspond to tight linkage.

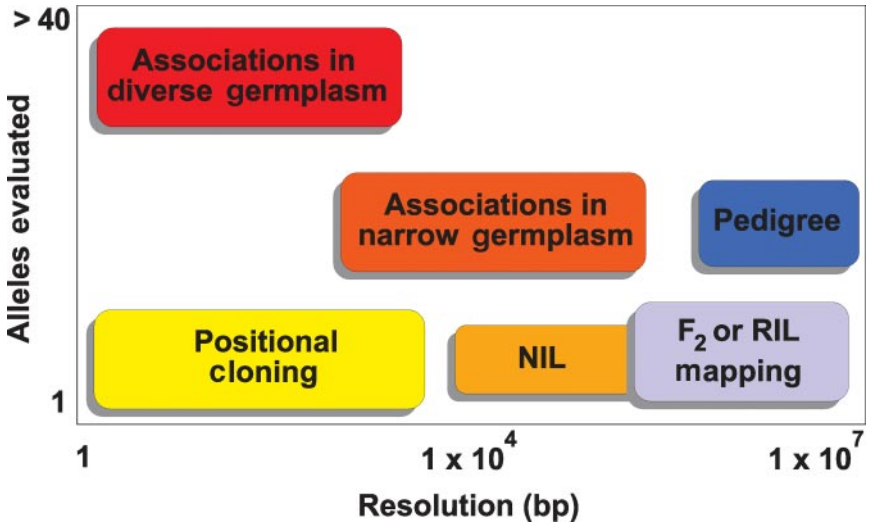

Figure 4 Schematic comparison of various types of analyses for range of alleles sampled versus resolution. Traditional quantitative trait loci (QTL) mapping methods are often limited by the number of alleles sampled and their resolution. Association analyses of diverse germplasm, on the other hand, are perfectly suited for sampling a wide range of alleles with high resolution.