Genetics and Molecular Biology, 31, 4, 805-814 (2008)
Copyright © 2008, Sociedade Brasileira de Genética. Printed in Brazil
www.sbg.org.br

Review

Linkage disequilibrium, genetic association mapping and gene localization in crop plants

Karim Sorkheh[1], Lyudmyla V. Malysheva-Otto[2], Michelle G. Wirthensohn[3], Saeed Tarkesh-Esfahani[1] and Pedro Martínez-Gómez[4]

[1]*Department of Agronomy and Plant Breeding, Faculty of Agriculture, Shahrekord University, Shahrekord, Iran.*
[2]*Institute of Plant Genetics and Crop Plant Research, Gatersleben, Germany.*
[3]*The University of Adelaide, School of Agriculture, Food and Wine, Waite Campus, Australia.*
[4]*Department of Plant Breeding, CEBAS-CSIC, Murcia, Spain.*

Abstract

DNA-based molecular markers have been extensively utilized for a variety of studies in both plant and animal systems. One of the major uses of these markers is the construction of genome-wide molecular maps and the genetic analysis of simple and complex traits. However, these studies are generally based on linkage analysis in mapping populations, thus placing serious limitations in using molecular markers for genetic analysis in a variety of plant populations. Therefore, alternative approach has been suggested, linkage disequilibrium-based association analysis which detects and locates quantitative trait loci (QTL) by the strength of the correlation between a trait and a marker. Although association analysis has already been used for studies on genetics of complex traits in humans, its use in plants has newly started. In the present review, we describe what is known about variation in linkage disequilibrium (LD) and summarize published results on association studies in crop plant species. We give a list of different factors affecting LD, and discuss the current issues of LD research in plants. Later, we also describe the various uses of LD in crop plants research and summarize the present status of LD researches in different plant genomes. Finally, future key issues about the application of these studies on the localization of genes in these crop plants have been also discussed.

Key words: linkage disequilibrium, LD, association mapping, population structure, QTLs, localization of genes, marker-trait associations.

Received: October 8, 2007; Accepted: May 16, 2008.

Introduction

Linkage disequilibrium (LD) is defined as a non-random association of alleles at separate loci located on the same chromosome (Mackay and Powell, 2007). The presence of LD is a prerequisite for association mapping where the LD extent or the physical size of LD blocks, that is chromosomal regions across which all pairs of adjacent loci are in LD (Stich, 2006), determines the marker density required for association mapping. Genome-wide association studies are currently exploited for mapping of disease genes in human genetics (see, for example, The Wellcome Trust Case Control Consortium, 2007). In crop plants, the potential of exploiting LD to detect marker-trait associations was recently investigated for maize (reviewed by Yu and Buckler, 2006; Belo *et al.*, 2008), wheat (Ravel *et al.*, 2006; Rhone *et al.*, 2007; Tommasini *et al.*, 2007), barley (Kraakman *et al.*, 2004; Kraakman *et al.*, 2006; Malysheva-Otto and Röder, 2006; Rostoks *et al.*, 2006), sorghum (Hamblin *et al.*, 2004), ryegrass (Skøt *et al.*, 2005; Xing *et al.*, 2007), soybean (Hyten *et al.*, 2007) and rice (Garris *et al.*, 2003). The published results suggest that association mapping is a valuable additional tool in the search for the detection of novel genes or QTLs for important agronomic characteristics. The extensive application of this approach in crop plants is to be expected in the long term as a result of establishing of the novel high-throughput genotyping and sequencing technologies (Mackay and Powell, 2007; Oraguzie *et al.*, 2007).

In contrast to QTL mapping, where typically biparental crosses with contrasting genotypes are used, in the case of association studies a collection of cultivars, lines, or

Send correspondence to Karim Sorkheh. Department of Agronomy and Plant Breeding, Faculty of Agriculture, Shahrekord University, Shahrekord, P.O. Box 115, Iran. E-mail: karim_sorkheh2000@yahoo.com.

landraces are genotyped with densely spaced markers. In plant genetics, using a collection of cultivars has a number of advantages over the use of a bi-parental cross. Firstly, in the population a broader genetic variation in a more representative genetic background will be available. This implies that one is not limited to the marker and trait loci that happen to differ between two parents (Kraakman *et al.* 2006). Secondly, LD mapping may attain a higher resolution, because of the use of all meioses accumulated in the breeding history. Thirdly, historic phenotypic data on cultivars can be used to link markers to traits, without the need for new trials with special mapping populations. The methodology for associating markers and traits in a collection of cultivars is still under development (Jannink and Walsh, 2002; Yu and Buckler, 2006; Mackay and Powell, 2007).

In order to identify marker-trait associations, LD has to occur in the plant germplasm. LD may increase due to selection in a population, for instance when an important trait is regulated by multiple loci, or due to recent introductions of genotypes. Factors contributing to the increase of LD include also small population size, inbreeding, genetic isolation between lineages, population subdivision, low recombination rate, population admixture, genetic drift and epistasis. While factors like outcrossing, high recombination rate, high mutation rate, gene conversion, etc., lead to a decrease/disruption in LD. The factors affecting LD have been extensively discussed in a number of papers (Ardlie *et al.*, 2002; Jannink and Walsh, 2002; Weiss and Clark, 2002; Flint-Garcia *et al.*, 2003; Gaut and Long, 2003; Gupta *et al.*, 2005; Kim *et al.*, 2007), and were also recently listed by Rafalski and Morgante (2004).

LD will tend to decay with genetic distance between the loci under consideration, because genetically distant loci are more likely to have recombined in the past than tightly linked loci. In populations, for any pair of linked polymorphic loci LD decreases over generations, because of accumulation of recombinations. Finally the loci will be in linkage equilibrium (LE), *i.e.* alleles are not preferentially paired anymore. The process of decrease of LD to reach LE depends on the opportunities of genetic recombination between the allele pairs of the loci under consideration. For effective recombination double heterozygotes are required, and these are much more common in allogamous than in autogamous plant species. Therefore, LD will tend to be more obvious after repeated inbreeding, as in autogamous species, than in out-crossing species. If LD estimates are supposed to be used for association analysis, the understanding of the factors affecting LD is particularly relevant, because one needs to rule out the possibility of LD caused by factors other than linkage.

Association studies based on correlations between alleles at different sites or LD can provide high resolution for the identification of genes that contribute to phenotypic variation in natural populations. This approach has a 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. Therefore it is important to understand LD and to determine the extent of LD in the species under investigation.

In this review we describe what is known about variation in Linkage Disequilibrium in crop plant species and summarize published results on genetic association mapping studies. Future key issues about the application of these studies on the localization of genes in these crop plants will also be discussed.

Linkage Disequilibrium Measurement, Visualization and Scope of Variation

The different measures for estimating the level of LD including the statistical tests for the significance of these measures have largely been described in recent reviews on LD in plants (Flint-Garcia *et al.*, 2003; Gaut and Long, 2003; Gupta *et al.*, 2005). The basic component of all LD statistics is the difference between the observed and expected haplotype frequencies at polymorphic loci, and the mathematical formulas for calculations can be found in Flint-Garcia *et al.* (2003). Briefly LD is calculated pairwise between two polymorphic sites; and the most frequently used LD measures are D' and r^2. The D' is the standardized disequilibrium coefficient which mainly measures recombinational history and is therefore useful to assess the probability of historical recombination in a given population. The r^2 is essentially the correlation between the alleles at two loci; it summarizes both recombinational and mutational history and is useful in the context of association studies. Both parameters vary in the interval from 0 to the value of 1.

Most LD calculations stand the linkage disequilibrium coefficient D, for which the layout and notation are shown in Table 1. Consider two loci A and B, each locus having two possible alleles: A_1 and A_2 at locus A, and B_1 and B_2 at locus B. The allele frequencies are denoted as p and naturally represent only sample estimates of some underlying population parameters, which are mostly unknown unless the total populations have been scored. There are four possible allele combinations between these two loci, which could represent the four possible types of gametes in a sexually reproducing organism. If the two loci are physically linked on the same chromosome, this array specifically represents the four haplotypes, but this does not have to be the case. If the two loci assort completely independently (*i.e.* linkage equilibrium) the gametic frequencies are calculated by the products of the allele frequencies; for example, the frequency of a haplotype carrying allele A_1 at the first locus, and allele B_1 at the second locus is given

Table 1 - Association between two alleles at each of two loci, showing the actual gametic frequencies and the expected gametic frequencies when the loci are in linkage disequilibrium. The marginal frequencies represent the allele frequencies (Mueller and Andrioli, 2004).

			Alleles at locus B		
			B_1	B_2	
	A_1		A_1B_1	A_1B_2	
		Actual	P_{A1B1}	P_{A1B2}	P_{A1}
Alleles at locus A		Expected	$P_{A1}P_{B1}$	$P_{A1}P_{B2}$	
	A_2		A_2B_1	A_2B_2	
		Actual	PA_2B_1	PA_2B_2	P_{A2}
		Expected	PA_2PB_1	PA_2PB_2	
			P_{B1}	P_{B2}	1

by the product P_{A1} P_{B1}. A simple and basic component of many disequilibrium measures is the difference (D) between the actual gametic frequency and the expected gametic frequency when the loci are independent (see review Mueller 2004).

To visualize or depict the extent of LD one can present a plot of LD decay which shows how LD declines with genetic (centiMorgans, cM) or physical (base pairs, bp) distance (Figure 1). Alternatively it is possible to construct the Disequilibrium Matrix which shows all loci in LD with corresponding probabilities (Figure 2). The matrix can cover the whole genome or distinct genomic locus. Nowadays quite a few software packages are available to calculate LD and to visualize its variation across the selected chromosome region or across the genome. These allow to calculate multilocus LD for bi-allelic as well as for multiallelic loci. In plant studies the most widely used programmes are TASSEL, HAPLOVIEW as well as statistical programmes developed within research groups (Kraakman *et al.*, 2004; Melchinger, 1996).

During the last five years the information about the LD pattern variation in different crop plant species was gradually accumulated. Most of the studies described LD in maize, wheat and barley, besides single reports on rice, ryegrass, soybean, sugarcane and sorghum. The extent of genome coverage in these studies varied from short distances as a few hundred base pairs up to genetic regions as huge as tens of centiMorgans (cM) or genome-wide. In general, the extent of LD varies greatly along the genome, so averages, while useful to know, may not reflect the local extent of LD. This makes the estimates of the number of markers needed more problematic. In addition, there are large variations in recombination frequency along the genome (lower near centromeres) which will affect LD in these regions.

Evaluation of LD in the maize genome revealed for diverse maize inbred lines rapid decay within 1 cM up to values of $r^2 < 0.05$ when assessed with intragenic SNPs, but a much higher genome-wide LD levels when assessed with SSRs (Remington *et al.*, 2001). For the commercial elite inbred lines LD blocks as long as 100 kb were detected (Ching *et al.*, 2002). At the physical distance level LD detected with SNPs, persisted over only 1 kb for *PSY2* locus (a putative phytonene synthase), but extended up to 600 kb in the region surrounding the maize phytonene synthase

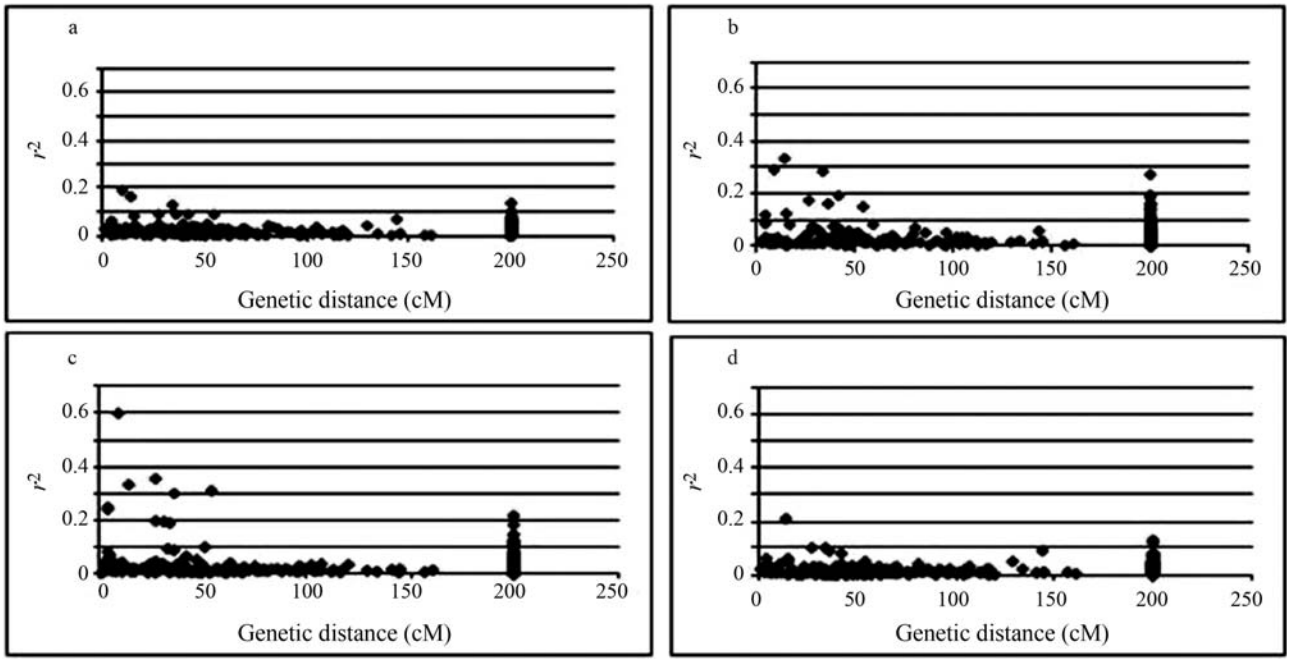

Figure 1 - The pattern of LD for 48 SSR loci in dependence on the population structure. Plots of LD represented by r^2 against genetic distance (in centiMorgan) in the global population of 953 accessions (a), 565 European accessions (b), 207 European 2-rowed spring accessions (c), and in the random set of 200 accessions (d). Pairs of loci mapped to different chromosomes were assigned to 200 cM (reproduced from Malysheva-Otto *et al.*, 2006a).

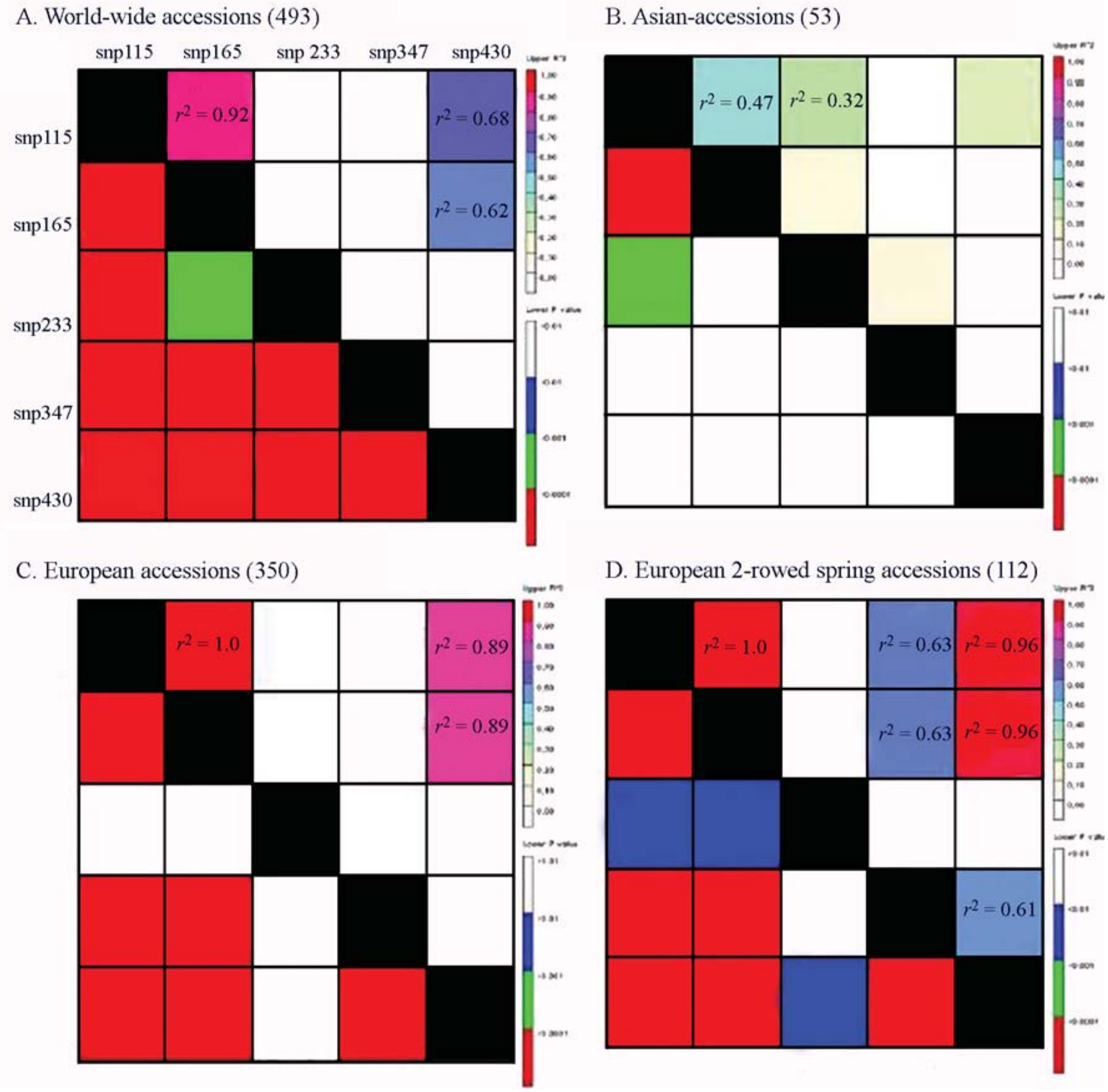

Figure 2 - Haplotype frequencies of *Bmy1* in barley accessions originating from various geographic regions worldwide. EU - Europe, AS - Asia, AF - Africa, NE - Near East, AM - America (reproduced from Malysheva-Otto and Röder, 2006b).

gene *Y1* (Palaisa *et al.*, 2004). In the other study LD across 800 kb around *Y1* and across 500 kb for maize alcohol dehydrogenase gene *adh1* was reported (Jung *et al.*, 2004). Persistent high levels of LD ($r^2 > 0.2$) were also shown over the whole genomic loci about 3.5 kb long of the maize *PAL* gene (Andersen *et al.*, 2007).

In wheat, genome-wide studies revealed LD extended up to 10 cM with mean $r^2 = 0.18$ (evaluated with SSRs, Maccaferri *et al.*, 2004) or mean with $r^2 = 0.133$ (Breseghello and Sorrells, 2006). However, intrachromosomal LD was much higher with $r^2 = 0.551$ (chromosome 2D) or $r^2 = 0.909$ (chromosome 5A) and decayed within a distance of < 1 cM (2D) and < 5 cM (5A) (Breseghello and Sorrells, 2006). In the study of the bread and durum wheat cultivars Somers *et al.* (2007) observed that only a small fraction of the locus pairs approximately 2-3 cM apart showed r^2 values > 0.2, but few loci at longer distances showed high levels of LD with $r^2 = 0.7$ and 1.0 at 25.5 and 41.2 cM, respectively. In subpopulations LD extended for longer distances and even higher r^2 values were detected (Somers *et al.*, 2007). Studies of LD across specific genomic regions were carried out for the region surrounding the yellow rust resistance gene Yr17 (Rhone *et al.*, 2007) and for the region of QSng.sfr-3BS, a major QTL for resistance to *Stagonospora nodorum glume* blotch on chromosome 3B (Tommasini *et al.*, 2007). In the experimental wheat populations maintained under dynamic management conditions a strong LD ($r^2 > 0.6$) was preserved over several generations in the zone surrounding Yr17 gene for a distance of 20 cM (Rhone *et al.*, 2007). In the study of LD across the QSng.sfr-3BS region (Tommasini *et al.*, 2007), LD dropped to $r^2 < 0.2$ within less than 0.5 cM in 44 diverse varieties, while it extended about 30 cM with $r^2 > 0.2$ in 240 RILs (LD was based on SSR and STS markers).

In cultivated barley genome-wide LD extended from 10 cM to 15 cM when evaluated with SSRs (Malysheva-Otto *et al.*, 2006), AFLP markers (Kraakman *et al.*, 2004) or SNPs (Rostoks *et al.*, 2006), and the pattern of LD was extremely population dependent. Substantial intralocus LD in barley was measured across a contiguous 212 kb region of four gene loci surrounding the hardness locus (Caldwell *et al.*, 2006), and within the 132 kb-long physical contiguity of barley gene Hv-eIF4E and flanking region, which confers resistance to the barley yellow mosaic virus (BYMV) complex (Stracke *et al.*, 2003). In the region surrounding the hardness locus, mean LD values of $r^2 > 0.2$ (Caldwell *et al.*, 2006) were detected, while over the physical contiguity of BYMV resistance locus the mean value of r^2 was > 0.4 (Stracke *et al.*, 2003). However, LD varied abruptly within the region, and at the genetic level dropped to $r^2 = 0.3$ within 1 cM. In wild barley an excess of interlocus LD was observed by analysing 18 genes in 25 accessions, and LD levels were lower than in maize (Morrell *et al.*, 2005).

Single reports are available about genome-wide or intralocus LD in other crops. Genome-wide scans with RFLP loci showed a decay of LD to values of $r^2 < 0.05$ within 10 cM in sugarcane (Flint-Garcia *et al.*, 2003) and within 50 cM in sorghum (Hamblin *et al.*, 2004). In the natural populations of perennial ryegrass at a genome-wide scale using AFLP genetic markers, the majority of the linked pairs were in significant LD within genetic distance of 4.37 cM with r^2 values not exceeding $r^2 = 0.12$, but two pairs were more than 20 cM apart (Skøt *et al.*, 2005).

The evaluation of the intragenic LD was performed in ryegrass (Xing *et al.*, 2007), soybean (Hyten *et al.*, 2007) and rice (Garris *et al.* 2003). In ryegrass nucleotide polymorphism analysis for 11 expressed disease resistance candidate (R) genes using about 1 kb genomic fragments for each of the genes revealed low intragenic LD with $r^2 < 0.2$ for most R genes, and rapid LD decay within 500 bp (Xing *et al.*, 2007).

The structure of LD in soybean germplasm was analyzed across three genomic regions up to 574 kb long which were located in different linkage groups (Hyten *et al.*, 2007). In the wild ancestor of soybean, *G. soja*, LD did not extend past 100 kb, with r^2 values slightly over 0.1; however, in the three cultivated *G. max* groups, LD extended up to 574 kb and higher r^2 values were detected (Hyten *et al.*, 2007).

Finally, in the case of fruit crops references about LD variation are very scarce. One of the first approach is the work of Aranzana *et al.* (2007) in peach using SSR markers indicating an estimated LD over 125 kb.

Application of Linkage-Disequilibrium Based Genetic Association Mapping in Crop Plants

Linkage disequilibrium can be used for a variety of purposes in crop plant genomics research. One of the major current and future uses of LD in plants would be to study marker-trait association (without the use of a mapping population) followed by marker-assisted selection (MAS). Another important application is its use in the studies of population genetics and genetic diversity in natural populations and germplasm collections and in crop improvement programmes.

Marker-trait association in crop plants is generally conducted through linkage analysis, utilizing methods like t-test, simple regression analysis and QTL interval mapping (for a discussion of these methods see, Hackett, 2002). Limitations of these methods have also been widely discussed (Darvasi *et al.*, 1993; Hästbacka *et al.*, 1994; Melchinger, 1996; Mackay, 2001; Hackett, 2002). The limitations of linkage analysis approach imposed by the availability of mapping populations have largely been overcome in LD-based association mapping, which can be applied to germplasm bank collections, synthetic populations, and elite germplasm. Genetic association mapping or linkage disequilibrium mapping is a method that relies on linkage disequilibrium to study the relationship between phenotypic variation and genetic polymorphisms (Breseghello and Sorrells, 2006).

For a study of marker-trait association using LD, the methods may differ for discrete traits and quantitative traits, although sometimes quantitative traits may also be treated as discrete traits. Two procedures that have been commonly used for mapping of discrete traits (disease genes) in humans are (i) case-control (CC) and (ii) transmission/disequilibrium test (TDT) (Spielman and Ewens, 1996; Allison, 1997). Similar (but not identical) approaches have also been used in crop plants (see review Gupta *et al.*, 2005). For instance, one such study involving discrete traits in plants was recently conducted in maize (Palaisa *et al.*, 2003), in which 78 out of 81 informative SNP and *InDel* polymorphisms in *Y1* gene were found associated with endosperm color when genotyped over a set of 41 yellow/orange endosperm lines and 34 white endosperm lines. The methodology used in this study is comparable to that used in CC studies in humans. In the research of Kumar *et al.* (2004) conducted in radiata pine, 200 full sib families were used to study the marker-trait associations. In this study the parental genotypes were also considered during analyses (Kumar *et al.* 2004), so that the method can be compared with TDT in humans.

The use of LD for mapping of QTLs for a quantitative trait is more challenging, but is also more rewarding, because it allows more precise locating of the position of a QTL controlling the trait of interest. When comparing linkage analysis and LD mapping for QTL detection, it is revealed that linkage mapping is more useful for genome-wide scan for QTLs, while LD mapping gives more precise location of an individual QTL. One may therefore like to use linkage analysis for preliminary location of QTLs and then use LD for more precise location (Mackay, 2001; Glazier *et al.*, 2002). LD between a single marker and a QTL can be measured by regression analysis, where the data on the trait is regressed on the individual marker genotypes, so that significant regressions will identify the markers associated with the phenotype (Remington *et al.*, 2001).

However, since this association of marker can sometimes be due to reasons other than linkage, further analysis is needed to select markers that are really associated with the trait due to close linkage. Therefore this regression of the trait on the marker genotype is sometimes examined by testing two adjacent markers for their association with the trait. In other cases, the effect of marker haplotypes on the trait through regression analysis can be estimated. Haplotypes having similar marker alleles (identical by descent), and associated with similar phenotypic effect should carry a QTL (Meuwissen and Goddard, 2000). Locating of a precise position within a very small chromosome region is possible through LD, but not through linkage analysis, since recombination within such a small region may not be available in an examined finite population (Mackay, 2001). The neutral theory of evolution holds that the majority of polymorphisms observed within and among species are selectively neutral or at least nearly so (Tajima, 1989). Neutrality makes mathematical modelling easy, giving a natural null model. Features, like selection, migration and demographic history can then be viewed as perturbation of a standard neutral model.

Genetic association mapping is a new approach which takes into account thousands of polymorphisms to evaluate for QTL effect and is more efficient as compared to linkage analysis because it does not require generation of segregating populations/large numbers of progeny (Oraguzie *et al.*, 2007; Belo *et al.*, 2008). However, association mapping is only capable of identifying phenotypic effects of alleles with reasonably high frequency in the population under investigation. Rare alleles usually cannot be evaluated because of lack of power (not enough individuals carrying this allele). So, for such alleles classical biparental mapping can be more appropriate.

The efficiency of association mapping is significantly influenced by the population structure. The presence of population stratification and an unequal distribution of alleles facilitate mapping and identification of the underlying causes of quantitative trait variation in plants. Subgroups can result in non-functional, spurious associations. 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 (Pritchard and Rosenberg, 1999).

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 (Sharbel *et al.*, 2000). 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 assume that population structure has similar effects on all loci and rely on the use of independent marker loci to detect stratified populations and to correct for them (Pritchard and Rosenberg, 1999). Pritchard *et al.* (2000) have developed an approach that incorporates estimates of population structure directly into the association test statistic. The essential idea of the method is to decompose a sample drawn from a mixed population into several unstructured subpopulations and test the association in the homogeneous subpopulations. The methods have been applied to association analyses in humans (Rosenberg *et al.*, 2002; Cardon and Bell, 2001) and crop plants, with modified test statistics being used to deal with quantitative traits (Thornsberry *et al.*, 2001; Belo *et al.*, 2008). In the study of flowering time locus in maize a suite of polymorphisms in the maize *dwarf8* gene was significantly associated with variation in flowering time (Thornsberry *et al.*, 2001). The incidence of false positives created by population structure was reduced by up to 8% 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. In the other research whole genome scan association mapping was used to identify loci with major effect on oleic acid content in maize kernels, and molecular marker at about 2 kb from a fatty acid desaturase, *fad2*, was associated with the differences in the phenotype (Belo *et al.*, 2008). 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.

The other method developed by Reich and Goldstein (2001) 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.

Nowadays there exists a handful of published software to assess the association of marker loci with traits. The most commonly used statistics include logistic regression with the possibility of structured associations implemented in TASSEL General Linear Model (Yu and Buckler, 2006; TASSEL: http://www.maizegenetics.net), a multiple regression model combined with the estimates for the false discovery rate suggested by Kraakman *et al.* (2006), and an unified mixed-model approach described by Yu *et al.* (2006) and implemented in TASSEL Mixed Linear Model or in SAS v9.1.2 (Ehrenreich *et al.*, 2007).

In addition, in any organism, LD can be used for identifying genomic regions, which have been regarded as the targets of natural selection (both directional selection and balancing selection), during evolutionary process (Gupta *et al.*, 2005; Ross-Ibarra *et al.*, 2007). Adaptive selection can leave one of two signatures on a gene region through genetic hitchhiking (Ross-Ibarra *et al.*, 2007). Directional selection can reduce levels of polymorphism through the rapid fixation of a new adaptive mutation. Balancing selection can increase levels of polymorphism when two or more alleles are maintained longer than expected under a neutral model. For example, if a polymorphism maintained by balancing selection is old, it will have enhanced sequence variability in the flanking regions, which may be used as a `signature of selection'. Due to inherent difficulties, only very few such studies have been conducted in the past, but more studies will certainly be conducted in future. One of the difficulties in such studies is caused by similar pattern of genetic variation expected due to natural selection on one hand and population demographic history (size, structure and mating pattern) on the other, while selection affects specific sites, demography affects the entire genome (Zhang *et al.*, 2002; Somers *et al.*, 2007).

In crop plants, efforts have also been made to identify genomic regions or genes, which were the targets of selection during domestication and subsequent selective breeding. For instance, QTLs for agronomic traits that were selected during domestication were identified through QTL interval mapping (Paterson *et al.*, 1995; Peng *et al.*, 2003, Pozzi *et al.*, 2004), even when functions of these genomic regions are unknown. For instance, in a study in maize, as many as 501 genes were screened using 75 EST-SSRs, to obtain signatures of selection. Fifteen of these 75 EST-SSRs gave some evidence of selection (Vigouroux *et al.*, 2002). In another study in maize, variability seems to have been reduced in a short regulatory region that lies 5' upstream of the teosinte branched1 (tb1) locus (Clark *et al.*, 2004). Large differences in the pattern of polymorphism between genomic regions are also seen in barley (Lin *et al.*, 2001).

Linkage Disequilibrium 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* 2,000 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 different 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 (Hästbacka *et al.*, 1992). 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 (2002) 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 F1-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 F2-based QTL mapping, namely, marker efficiency and statistical power.

One of the major uses of LD-based association analysis in future will be the study of marker-trait associations, leading to MAS, which was discussed earlier in this review. The approach will be particularly useful in forest trees, where mapping populations can not be easily generated, but MAS will prove extremely useful. For this purpose, LD will also facilitate development of functional markers (FMs), which are the perfect markers for marker-trait association (see Andersen and Lübberstedt, 2003; Gupta *et al.*, 2005; Simko *et al.* 2004).

Genetic and physical maps of genomes, based on molecular markers have now been constructed in all major crops. The work on the construction of LD maps in humans has already started, but the construction of LD maps for plant genomes has yet to start. In humans, LD maps of small regions of the genome or those involving mapping of disease genes relative to molecular markers have been constructed successfully. In due course of time such mapping will be attempted in plants too. These LD maps will make use of molecular markers that flank marker intervals delimited on the basis of estimations of LD, the distances being represented as LD units (Zhang *et al.*, 2002). LD mapping theory extends the estimation of covariance D for a random sample of haplotypes or diplotypes (disomic genotypes) to the association probability ρ = D/Q (1-R), where D is an estimation of LD (see above), Q is the frequency of the rarest and therefore putatively the youngest allele, and R is the frequency of the associated marker allele (Maniatis *et al.*, 2002). The estimates of these three parameters will be utilized for LD mapping. The software's ALLASS (allele association) and LDMAP VERSION 0.1, March 2002 (both developed by Andrew Collins at the University of Southampton, UK) are recommended for use in constructing LD maps.

Concluding Remarks

Linkage disequilibrium has been extensively utilized for a variety of purposes including mapping of disease QTLs in humans, but its use in plants has just begun. With the availability of high density maps in a number of crop plants, as well as whole genome sequencing in model plants like *Arabidopsis* and rice, and the sequencing of gene rich regions in crops like sorghum, maize, barley and wheat, we are at the threshold of utilizing the LD based genetic association mapping in crop plants in a wide range. Facilitate mapping and identification of the underlying causes of quantitative trait variation in plants. This approach will be used in various plant genomes for construction of LD maps, for study of marker-trait association both independently and in combination with linkage analysis and for the study of population genetics and evolution in nature as well as under domestication. Association mapping will facilitate gene mapping and identification of the underlying causes of quantitative trait variation in plants. Future studies of LD in crop plants will also elucidate further the structures of plant genomes and will also facilitate the use of marker-assisted selection (MAS) and map-based cloning of genes for difficult traits.

Acknowledgments

The authors gratefully acknowledge the help of the Head of the Department of Genetics and Plant Breeding of the Shahr-e-Kord University in Iran for providing facilities. In addition, careful reading and critical anonymous reviewers helped in the improvement of the manuscript.

References

Allison DB (1997) Transmission disequilibrium tests for quantitative traits. Am J Hum Genet 60:676-690.

Andersen JR and Lübberstedt T (2003) Functional markers in plants. Trends Plant Sci 8:1360-1385.

Andersen JR, Zein I, Wenzel G, Krützfeldt B, Eder J, Ouzunova M and Lübberstedt T (2007) High levels of linkage disequilibrium and associations with forage quality at a *Phenylalanine Ammonia-Lyase* locus in European maize (*Zea mays* L.) inbreds. Theor Appl Genet 114:307-319.

Aranzana MJ, Howad W and Arús P (2007) The extent of linkage disequilibrium in peach: A first approach. XII EUCARPIA Fruit Section Symposium, Zaragoza, pp 157.

Ardlie KG, Kruglyak L and Seielstad M (2002) Patterns of linkage disequilibrium in the human genome. Nat Rev Genet 3:299-309.

Beló A, Zheng P, Luck S, Shen B, Meyer DJ, Li B, Tingey S and Rafalski A (2008) Whole genome scan detects an allelic variant of *fad2* associated with increased oleic acid levels in maize. Mol Genet Genomics 279:1-10.

Breseghello F and Sorrells MS (2006) Association mapping of kernel size and milling quality in wheat (*Triticum aestivum* L.) cultivars. Genetics 172:1165-1177.

Caldwell KS, Russell J, Langridge P and Powell W (2006) Extreme population-dependent linkage disequilibrium detected in an inbreeding plant species, *Hordeum vulgare*. Genetics 172:557-567.

Cardon LR and Bell JI (2001) Association study designs for complex diseases. Nat Rev Genet 2:91-99.

Ching A, Caldwell KS, Jung M, Dolan M and Smith OS (2002) SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. BMC Genetics 3:19-32.

Clark RM, Linton E, Messing J and Doebley JF (2004) Patterns of diversity in the genomic region near the maize domestication gene tb1. Proc Nat Acad Sci USA 101:700-707.

Darvasi A, Weintreb A, Minke V, Weller S and Soller M (1993) Detecting marker QTL linkage and estimating QTL gene effect and map location using a saturated genetic map. Genetics 134:943-951.

Ehrenreich IM, Stafford PA and Purugganan MD (2007) The genetic architecture of shoot branching in *Arabidopsis thaliana*: A comparative assessment of candidate gene associations *vs.* quantitative trait locus mapping. Genetics 176:1223-1236.

Flint-Garcia SA, Thornsberry JM and Buckler ES IV (2003) Structure of linkage disequilibrium in plants. Annu Rev Plant Biol 54:357-374.

Garris AJ, McCouch SR and Kresovich S (2003) Population structure and its effects on haplotype diversity and linkage disequilibrium surrounding the xa5 locus of rice *Oryza sativa* L. Genetics 165:759-769.

Gaut BS and Long AD (2003) The lowdown on linkage disequilibrium. Plant Cell 15:1502-1506.

Glazier AM, Nadeau JH and Aitman TJ (2002) Finding genes that underlie complex traits. Science 298:2345-2349.

Gupta PK, Rustgi S and Kulwal, PL (2005) Linkage disequilibrium and association studies in higher plants: Present status and future prospects. Plant Mol Biol 57:461-485.

Hackett CA (2002) Statistical methods of QTL mapping in cereals. Plant Mol Biol 48:585-599.

Hamblin MT, Mitchell SE, White GM, Gallego J, Kukatla R, Wing RA, Paterson AH and Kresovich S (2004) Comparative population genetics of the panicoid grasses: Sequence polymorphism, linkage disequilibrium and selection in a diverse sample of *Sorghum bicolor*. Genetics 167:471-483.

Hästbacka J, de la Chapelle A, Kaitila I, Sistonen P, Weaver A and Lander E (1992) Linkage disequilibrium mapping in isolated founder populations: Diastrophic dysplasia in Finland. Nat Genet 2:204-11.

Hästbacka J, de la Chapelle A, Mahtani M, Clines G, Reeve-Daly MP, Daly M, Hamilton BA, Kusumi K, Trivedi B, Weaver A, *et al.* (1994) The diastrophic dysplasia gene encodes a novel sulfate transporter: Positional cloning by fine-structure linkage disequilibrium mapping. Cell 78:1078-1087.

Hyten DL, Choi I-Y, Song Q, Shoemaker RC, Nelson RL, Costa JM, Specht JE and Cregan PB (2007) Highly variable patterns of linkage disequilibrium in multiple soybean populations. Genetics 175:1937-1944.

Jannink J-L and Walsh B (2002) Association mapping in plant populations. In: Kang MS (ed), Quantitative Genetics, Genomics and Plant Breeding. CAB International, New York, pp 59-68.

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

Kim S, Plagnol V, Hu TT, Toomajian C, Clark RM, Ossowski S, Ecker JR, Weigel D and Nordborg M (2007) Recombination and linkage disequilibrium in *Arabidopsis thaliana*. Nat Genet 39:1151-1155.

Kraakman ATW, Martínez F, Mussiraliev B, van Eeuwijk FA and Niks RE (2006) Linkage disequilibrium mapping of morphological, resistance, and other agronomically relevant traits in modern spring barley cultivars. Mol Breed 17:41-58.

Kraakman ATW, Niks RE, Van der Berg PMMM, Stam P and Van Eeuwijk FA (2004) Linkage disequilibrium mapping of yield and yield stability in modern spring barley cultivars. Genetics 168:435-46.

Kumar S, Echt C, Wilcox PL and Richardson TE (2004) Testing for linkage disequilibrium in the New Zealand radiata pine breeding population. Theor Appl Genet 108:292-298.

Lin JZ, Brown AHD and Clegg MT (2001) Heterogeneous geographic patterns of nucleotide diversity between two alcohol dehydrogenase genes in wild barley *Hordeum vulgare* subspecies spontaneum. Proc Natl Acad Sci USA 98:531-536.

Maccaferri M, Sanguineti MC, Noli E and Tuberosa R (2004) Population structure and long-range linkage disequilibrium in a durum wheat elite collection. In: Plant & Animal Genomes XII Conference, San Diego, pp 416.

Mackay L and Powell W (2007) Methods for linkage disequilibrium mapping in crops. Trends Plant Sci 12:57-63.

Mackay TFC (2001) The genetic architecture of quantitative traits. Annu Rev Genet 33:303-39.

Malysheva-Otto L and Röder MS (2006) Haplotype diversity in the endosperm specific β-amylase gene βmy1 of cultivated barley (*Hordeum vulgare* L.). Mol Breed 18:143-156.

Malysheva-Otto L, Ganal MW and Röder MS (2006) Analysis of molecular diversity, population structure and linkage disequilibrium in worldwide survey of cultivated barley germplasm (*Hordeum vulgare* L.). BMC Genetics 7:6.

Maniatis N, Collins A and Xu C-F (2002) The first linkage disequilibrium (LD) maps: Delineation of hot and cold blocks by diplotype analysis. Proc Natl Acad Sci USA 99:2228-2233.

Melchinger AE (1996) Advances in the analysis of data on quantitative trait loci. In: Chopra VL, Singh RB and Verma A (eds), Proceedings 2nd International Crop Science Congress, New Delhi, pp 773-791.

Meuwissen THE and Goddard ME (2000) Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. Genetics 155:421-430.

Morrell PL, Toleno DM, Lundy KE and Clegg MT (2005) Low levels of linkage disequilibrium in wild barley (*Hordeum vulgare* ssp. s*pontaneum*) despite high rates of self-fertilizaton. Proc Natl Acad Sci USA 102:2442-2447.

Mueller JC and Andreoli C (2004) Plotting haplotype-specific linkage disequilibrium patterns by extended haplotype Linkage disequilibrium for different scales and applications homozygosity. Bioinformatics 20:786-787.

Oraguzie NC, Rikkerink EHA, Gardiner SE and Silva HN de (2007) Association Mapping in Plants. Springer, Tokio and New York, 277 pp.

Palaisa K, Morgante M, Tingey S and Rafalski A (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 USA 101:9885-9890.

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

Paterson AH, Lin Y-R, Li Z, Schertz KF, Doebley JF, Pinson SRM, Liu S-C, Stansel JW and Irvine JE (1995) Convergent domestication of cereal crops by independent mutations at corresponding genetic loci. Science 269:1714-1718.

Peng J, Ronin Y, Fahima T, Röder MS, Li Y, Nevo E and Korol A (2003) Domestication quantitative trait loci in *Triticum dicoccoides*, the progenitor of wheat. Proc Natl Acad Sci USA 100:2489-2494.

Pozzi C, Rossini L, Vacchietti A and Salamini F (2004) Gene and genome changes during domestication of cereals. In: Gupta PK and Varshney RK (eds), Cereal Genomics. Kluwer Academic, Springer Netherlands, pp 165-198.

Pritchard JK and Rosenberg NA (1999) Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 65:220-28.

Pritchard JK, Stephens M, Rosenberg NA and Donnelly P (2000) Association mapping in structured populations. Am J Hum Genet 37:170-181.

Rafalski A (2002) Applications of single nucleotide polymorphisms in crop genetics. Curr Opin Plant Biol 5:94-100.

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

Ravel C, Praud S, Murigneux A, Linossier L, Dardevet M, Balfourier F, Dufour PH, Brunel D and Charmet G (2006) Identification of Glu-B1-1as a candidate gene for the quantity of high-molecular-weight glutenin in bread wheat (*Triticum aestivum* L.) by means of an association study, Theor Appl Genet 112:738-743.

Reich DE and Goldstein DB (2001) Detecting association in a case-control study while correcting for population stratification. Genet Epidemiol 20:4-16.

Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doebley J, Kresovich S, Goodman MM and Buckler ES IV (2001) Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc Nat Acad Sci USA 98:11479-11484.

Rhone B, Raquin AL and Goldringer I. (2007) Strong linkage disequilibrium near the selected Yr17 resistance gene in a wheat experimental population. Theor Appl Genet 114:787-802.

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA and Feldman MW (2002) Genetic structure of human populations. Science 298:2381-2385.

Ross-Ibarra J, Morrell PL and Gaut BS (2007) Plant domestication, a unique opportunity to identify the genetic basis of adaptation. Proc Nat Acad Sci USA 104 (suppl. 1):8641-8648.

Rostoks N, Ramsay L, MacKenzie K, Cardle L, Bhat PR, Roose ML, Svensson JT, Stein N, Varshney RK, Marshall DF, *et al.* (2006) Recent history of artificial outcrossing facilitates whole-genome association mapping in elite inbred crop varieties. Proc Natl Acad Sci USA 103:18656-18661.

Sharbel TF, Haubold B and Mitchell-Olds T (2000) Genetic isolation by distance in *Arabidopsis thaliana*: Biogeography and postglacial colonization of Europe. Mol Ecol 9:2109-2118.

Simko I, Haynes KG, Ewing EE, Costanzo S, Christ BJ and Jones RW (2004) Mapping genes for resistance to *Verticillium albo-atrum* in tetraploid and diploid potato populations using haplotype association tests and genetic analysis. Mol Genet Gen 271:522-531.

Skøt L, Humphreys MO, Armstead I, Heywood S, Skøt KP, Sanderson R, Thomas ID, Sanderson R, Chorlton KH and Hamilton NRS (2005) An association mapping approach to identify lowering time genes in natural populations of *Lolium perenne* (L.). Mol Breed 15:233-245.

Somers DJ, Banks T, Depauw R, Fox S, Clarke J, Pozniak C and McCartney C (2007) Genome-wide linkage disequilibrium analysis in bread wheat and durum wheat. Genome 50:557-567.

Spielman RS and Ewens WJ (1996) The TDT and other family based tests for linkage disequilibrium and association. Am J Hum Genet 59:983-989.

Stich B (2006) A new test for family-based association mapping with inbred lines from plant breeding programs. Theor Appl Genet 113:1121-1130.

Stracke S, Perovic D, Stein N, Thiel T and Graner A (2003) Linkage disequilibrium in barley. 11th Molecular Markers Symposium of the GPZ, http://meetings.ipkgatersleben.de/moma2003/index.php.

Tajima F (1989) Statistical method for testing the neutral mutational hypothesis by DNA polymorphism. Genetics 123:585-595.

The Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 147:661-678.

Thornsberry JM, Goodman MM, Doebley J, Kresovich S, Nielsen D and Buckler ES IV (2001) Dwarf8 polymorphisms associate with variation in flowering time. Nat Genet 28:286-289.

Tommasini L, Schnurbusch T, Fossati D, Mascher F and Keller B (2007) Association mapping of *Stagonospora nodorum* blotch resistance in modern European winter wheat varieties. Theor Appl Genet 115:697-708.

Vigouroux Y, McMullen M, Hittinger CT, Houchins K, Schulz L, Kresovich S, Matsuoka Y and Doebley J (2002) Identifying genes of agronomic importance in maize by screening microsatellites for evidences of selection during domestication. Proc Natl Acad Sci USA 99:9650-9655.

Weiss KM and Clark AG (2002) Linkage disequilibrium and the mapping of complex human traits. Trends Genet 18:19-24.

Xing Y, Frei U, Schejbel B, Asp T and Lübberstedt T (2007) Nucleotide diversity and linkage disequilibrium in 11 expressed resistance candidate genes in *Lolium perenne*. BMC Plant Biol 7:43.

Yu J and Buckler ES (2006) Genetic association mapping and genome organization of maize. Curr Opin Biotechn 17:155-160.

Yu J, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, *et al.* (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38:203-208.

Zhang W, Collins A, Maniatis N, Tapper W and Morton NE (2002) Properties of linkage disequilibrium LD maps. Proc Natl Acad Sci USA 99:17004-17007.

Associate Editor: Everaldo Gonçalves de Barros

License information: This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.