IJPS INTERNATIONAL JOURNAL OF PLANT SCIENCES
e ISSN-0976-593X Volume **10** | Issue 1 | January, 2015 | 85-94
DOI: 10.15740/HAS/IJPS/10.1/85-94
Visit us - *www.researchjournal.co.in*

A REVIEW

Association mapping : a useful tool

■ **HEMANT SAHU, JAIRAM AMADABADE AND NAMRATA DHIRI**

SUMMARY

Future advances in plant genomics will make it possible to scan a genome for polymorphisms associated with qualitative and quantitative traits. Association mapping, also known as linkage disequilibrium mapping. Association mapping has the promise of higher mapping resolution through exploitation of historical recombination events at the population level that may enable gene level mapping on non-model organisms where linkage-based approaches would not be feasible. Association mapping utilizes ancestral recombinations and natural genetic diversity within a population to dissect quantitative traits and is built on the basis of the linkage disequilibrium concept. One of the working definitions of linkage disequilibrium (which here on will be referred to as LD) is the non-random co-segregation of alleles at two loci. In contrast to linkage-based studies, LD-based genetic association studies offer a potentially powerful approach for mapping causal genes with modest effects. The commonly used LD measure, *D* or D^1 (standardized version of *D*). Besides D, a various different measures of LD (D^1, r^2, D^2, *D**, *F, X* and¢) have been developed to quantify LD. Softwares packages measuring LD, such as "graphical overview of linkage disequilibrium" (GOLD), power marker, and TASSEL (Trait Analysis by association Evolution and Linkage) are available. Applications of AM has been extended from model plant arabidopsis to field crops. The increasing number of AM studies in crop species indicates the potential of this approach in all plant species in near future.

Key Words : Linkage disequilibrium, Association mapping, Genetic diversity, Polymorphism

How to cite this article : Sahu, Hemant, Amadabade, Jairam and Dhiri, Namrata (2015). Association mapping : a useful tool. *Internat. J. Plant Sci.*, **10** (1): 85-94.

Article chronicle : Received : 29.09.2014; **Accepted :** 25.12.2014

Association mapping, also known as linkage disequilibrium mapping, is a relatively new and promising genetic method for complex trait dissection. Association mapping has the promise of higher mapping resolution through exploitation of historical recombination events at the population level, that may enable gene level mapping on non-model organisms where linkage-based approaches would not be feasible (Risch and Merikangas 1996; Nordborg and Tavare,2002). Association mapping utilizes ancestral recombinations and natural genetic diversity within a population to dissect quantitative traits and is built on the basis of the linkage disequilibrium concept (Geiringer,1944; Lewontin and Kojima, 1960). One of the working definitions of linkage disequilibrium (which here on will be referred to as LD) is the non-random co-segregation of alleles at two loci. In contrast to linkage-based studies, LD-based genetic association studies offer a potentially powerful approach for mapping causal genes with modest effects (Hirschhorn and Daly, 2005). Association mapping focuses on associations within populations of unrelated individuals. In other words, the time to most recent common ancestor (MRCA) of any

MEMBERS OF THE RESEARCH FORUM

Author to be contacted :
HEMANT SAHU, Department of Genetics and Plant Breeding, Indira Gandhi Krishi Vishwavidyalaya, RAIPUR (C.G.) INDIA

Address of the Co-authors:
JAIRAM AMADABADE, Department of Genetics and Plant Breeding, G.B. Pant University of Agriculture and Technology, PANTNAGAR (UTTARAKHAND) INDIA

NAMRATA DHIRI, Department of Genetics and Plant Breeding, Indira Gandhi Krishi Vishwavidyalaya, RAIPUR (C.G.) INDIA

given two individuals from a population of unrelated individuals would be greater than that of a pedigree population. This is what makes LD mapping suitable for fine-scale mapping, there will have been more opportunities for recombination to take place over several generations, between many alleles, in a species, while there can be only a few generations of recombination present in pedigree populations. Increase in the rate of recombination will lead to reshuffling of the chromosomal segments into smaller pieces. This will lead to reduction of the LD in short distances around loci and lead to significant co-occurrence (*i.e.* LD) between only loci physically close, allowing high resolution. Whereas pedigree studies work with recombination events in few generations that enable exchange between chromosomes at the order of mega bases, association studies deal with segmental exchanges measured in kilobases (Paterson *et al.*, 1990; Stuber *et al.*, 1992 and Thornsberry *et al.*, 2001).

Genetic mapping of causative variants :

The main goal of genetic mapping is to detect neutrally inherited markers in close proximity to the genetic causatives or genes controlling the complex quantitative traits. Genetic mapping can be done mostly in two ways.

- Using the experimental populations (also referred to as "biparental" mapping populations) that is called QTL-mapping as well as "genetic mapping" or "gene tagging,"
- Using the diverse lines from the natural populations or germplasm collections that is called LD-mapping or "association mapping."

Linkage analysis can be done firstly, the experimental populations such as F_2, back cross (BC), double haploid (DH), recombinant inbred line (RIL), and near isogenic line (NIL) populations, derived from the genetic hybridization of two parental genotypes with an alternative trait of interest, need to be developed. Secondly, these experimental populations including a large number of progenies or lines are measured for the segregation of a trait of interest in the different environmental conditions. Thirdly, a set of polymorphic DNA markers, differentiating the parental genotypes and segregating in a mapping population, need to be identified and genotyped. For that, usual practice is that, first, the parental genotypes are screened, and if markers are polymorphic over the parents, then, all individuals of a mapping population are genotyped with these polymorphic molecular markers. Once genotypic data of a mapping population is ready, marker data are used to construct the framework linkage maps, representing the order (position) and linkage (a relative genetic distance in cM) of used molecular markers along the linkage groups or segments of particular chromosomes. Now these linkage map are statistically correlated with phenotypic characteristics of individuals of a mapping population and QTL regions affecting a trait of interest. The precision of QTL-mapping largely depends on the genetic variation (or genetic background) covered by a mapping population, the size of a mapping population, and a number of marker loci used. These marker tags are the most effective tools in a crop improvement that allows the mobilization of the genes of interest from donor lines to the breeding material through marker-assisted selection (MAS). Although traditional QTL-mapping will continue being an important tool in gene tagging of crops, but it has several limitation such as :

- Overall is very costly
- Low resolution
- Hampering the fine mapping, is associated with the availability of only a few meiotic events to be used that occurred since experimental hybridization in a recent past.

Association mapping as an alternative approach :

These limitations, however, can be reduced with the use of "association mapping". Turning the gene-tagging efforts from biparental crosses to natural population of lines (or germplasm collections), and from traditional QTLmapping to linkage disequilibrium (LD)-based association study became a powerful tool in mapping of the genes of interest. This leads to the most effective utilization of *Ex situ* conserved natural genetic diversity of worldwide crop germplasm resources. LD refers to a historically reduced (nonequilibrium) level of the recombination of specific alleles at different loci controlling particular genetic variations in a population. This LD can be

Table 1: Difference b/w AM and linkage mapping

	Linkage mapping		Association mapping
1.	Family based - development of mapping population is required	1.	Families or unrelated
2.	Only few recombination's are considered	2.	Historic mutations and recombination's are considered
3.	Only two alleles will be considered at time	3.	More than two alleles per locus can be studied simultaneously
4.	Low resolution	4.	High resolution
5.	Species/population specific	5.	Not specific for species/population
6.	Few markers for genome coverage	6.	Many markers for genome coverage
7.	Weak design	7.	Powerful design
8.	Duration is more	8.	Less duration
9.	Good for initial detection; poor for fine-mapping	9.	Poor for initial detection; good for fine mapping

detected statistically, and has been widely applied to map and eventually clone a number of genes underlying the complex genetic traits in humans.

The advantages of population-based association study :

- Availability of broader genetic variations with wider background for marker-trait correlations
- Many alleles evaluated simultaneously
- Higher resolution mapping because of the utilization of majority recombination events from a large number of meiosis throughout the germplasm development history,
- Possibility of exploiting historically measured trait data for association, and
- No need for the development biparental populations that makes approach time saving and cost-effective.

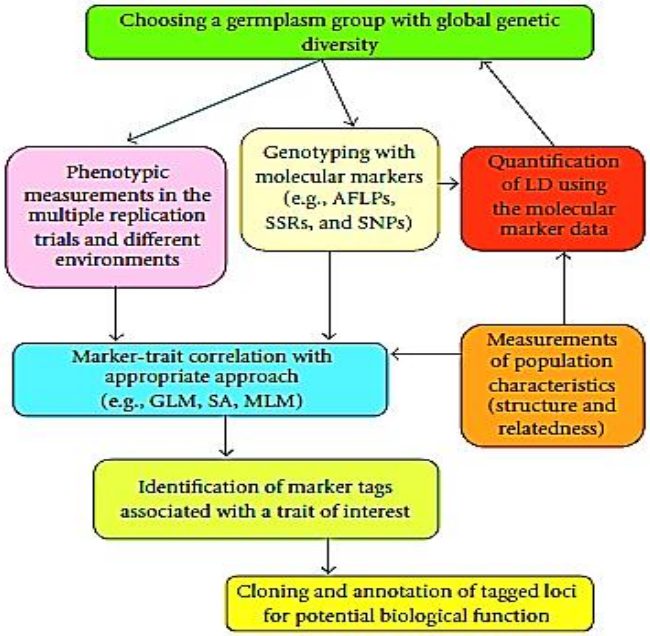

Fig. 1: The scheme of association mapping for tagging a gene of interest using germplasm accessions.
Note that the outlined scheme may vary based on population characteristics and methodology chosen for association study

Steps in association mapping (Fig. 1):

- Selection of a group of individuals from a natural population or germplasm collection with wide coverage of genetic diversity;
- Recording or measuring the phenotypic characteristics (yield, quality, tolerance, or resistance) of selected population groups, preferably, in different environments and multiple replication trial design;
- Genotyping a mapping population individuals with available molecular markers;
- Quantification of the extent of LD of a chosen population genome using a molecular marker data;
- Assessment of the population structure (the level of genetic differentiation among groups within a sampled population individuals) and kinship (co-efficient of relatedness between pairs of each individuals within a sample) and
- Based on information gained through quantification of LD and population structure, correlation between phenotypic and genotypic/haplotypic data is established with the application of an appropriate statistical method that reveals "marker tags" positioned within close proximity of targeted trait of interest. As a starting point for association mapping, it is important to gain knowledge of the patterns of LD for genomic regions of the "target" organisms and the specificity of the extent of LD among different populations or groups to design and conduct unbiased association mapping.

Linkage disequilibrium (LD) :

Linkage equilibrium (LE) is a random association of alleles at different loci and equals the product of allele frequencies within haplotypes. In contrast, LD is a nonrandom association of alleles at different loci, describing the condition with nonequal (increased or reduced) frequency of the haplotypes in a population at random combination of alleles at different loci. LD is not the same as linkage, although tight linkage may generate high levels of LD between alleles. Usually, there is significant LD between more distant sites or sites located in different chromosomes, caused by some specific genetic factors. The concept of LD was first described by Jennings in 1917, and its quantification (*D*) was developed by Lewtonin in 1964. The simplified explanation of the commonly used LD measure, *D* or *D*_ (standardized version of *D*), is the difference between the observed gametic frequencies of haplotypes and the expected gametic haplotype frequencies under linkage equilibrium :

$[D = P(AB) - P(A)P(B) = P(AB)P(ab) - P(Ab)P(aB)]$
Besides *D*, a various different measures of LD (D_-, r^2, D^2, D^1, *F*, *X* (2) and ¢)

Have been developed to quantify LD. Choosing the appropriate LD measures really depends on the objective of the study, and one performs better than other in particular situations and cases; however, D_- and r^2 is the most commonly used measures of LD.

D_- is informative for the comparisons of different allele frequencies across loci and strongly inflated in a small sample size and low-allele frequencies, therefore, intermediate values of D_- is dangerous for comparative analyses of different LD studies and should be verified with the r^2 before using for quantification of the extent of LD. The r^2, the square of the co-

rrelation co-efficient between the two loci have more reliable sampling properties than *D*_ with the cases of low allele frequencies. The r^2 is affected by both mutation and recombination while *D*_ is affected by more mutational histories. Considering the objective, the most appropriate LD quantification measure needed for association mapping is r^2 that is also an indicative of marker-trait correlations. The r^2 value varies from 0 to 1 and it will be equal to 1 when only two haplotypes are present. The r^2 value of equal to 0.1 (10%) or above considered the significant threshold for the rough estimates of LD to reveal association between pairs of loci.

For two biallelic loci, D' and r^2 have the following formula:

$$D' = \frac{[D]}{D_{max}}$$

D' is constrained between - 1 and + 1.
D' = 1 (perfect positive LD between SNP alleles)
D' = 0 (linkage equilibrium between SNP alleles)
D' = -1 (perfect negative LD between SNP alleles)
D' = 0.87 (strong positive LD between SNP alleles)
D' = 0.12 (weak positive LD between SNP alleles)
For LD It should be ≥ 0.5

where,

$D_{max} = \min (P_A P_b, P_a P_B)$ if $D > 0$

$D_{max} = \min (P_A P_B, P_a P_b)$ if $D < 0$

$$r^2 = \frac{D^2}{P_A P_a P_B P_b}$$

For LD It should be > 0.1

$$0 \le r^2 \le 1$$

Visualization of L D :

Graphical display of pairwise LD between two loci is very useful to estimate the LD patterns measured using a large number of molecular markers.

LD triangle or Heatmap:

Pairwise LD can be depicted as a colour-code triangle plot, based on significant pairwise LD level (r^2 and *p*-value as well as *D*_) that helps to visualize the block of loci (red blocks) in significant LD. The large red blocks of haplotypes along the diagonal of the triangle plot indicate the high level of LD between the loci in the blocks, meaning that there has been a limited or no recombination since LD block formations. There is freely available specific computer software, "graphical overview of linkage disequilibrium" (GOLD), to depict the structure and pattern of LD. Some other software packages measuring LD such as "Trait analysis by association, Evolution and Linkage" (TASSEL) and PowerMarker have also similar graphical display features. The strong block-like LD structures are of a great interest in association mapping which simplifies LD mapping efforts of complex traits. LD blocks are very useful in association mapping when sizes are calculated, which suggest the needs for the minimum number of markers to efficiently cover the genome-wide haplotype blocks in association mapping (Fig. 2).

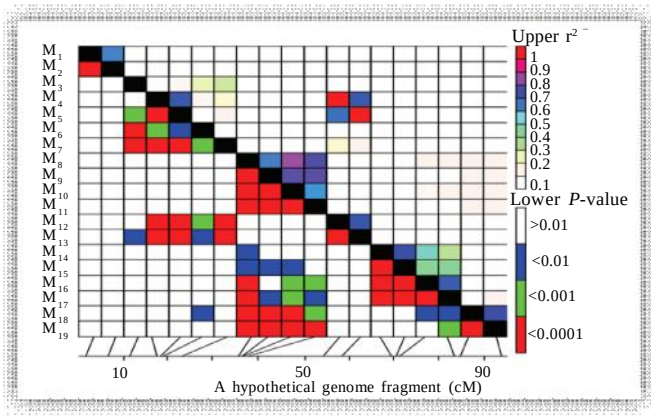

Fig. 2 : The TASSEL generated triangle plot for pairwise LD between marker sites in a hypothetical genome fragment.
Pairwise LD values of polymorphic sites are plotted on both the *X*- and *Y*-axis; above the diagonal displays r^2 values and below the diagonal displays the corresponding *p*-values from rapid 1000 shuffle permutation test. Each cell represents the comparison of two pairs of marker sites with the colour codes for the presence of significant LD. Coloured bar code for the significance threshold levels in both diagonals is shown. The genetic distance scale for a hypothetical genome fragment was manually drawn. Note: this is for demonstration purposes only and does not have any real impact or correspond to any genomic fragment of an organism

LD decay plots :

To estimate the size of these LD blocks, the r^2 values (alternatively, *D*_ can also be used) usually plotted against the genetic (cM) or weighted (bp) distance referred to as a "LD decay plot". One can estimate an average genome-wide

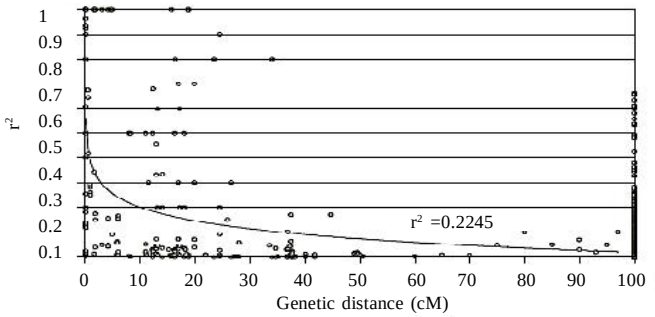

Fig. 3 : Linkage disequilibrium (LD) decay plot depicted from the LD values of a hypothetical marker.
The data are to demonstrate a measure of an average genome-wide LD block sizes. A pairwise LD values (r^2) are plotted against a genetic distance. Inner fitted trend line is a nonlinear logarithmic regression curve of r^2 on genetic distance. LD decay is considered below r^2 = 0.1 threshold and based on trend line it is around 38–40 cM in above plot. A pairwise LD between unlinked marker loci is assigned to 100 cM distance point

decay of LD by plotting LD values obtained from a data set covering an entire genome against distance. When such a LD decay plot generated, usual practice is to look for distance point where LD value (r^2) decreases below 0.1 or half strength of $D\_$ ($D\_ = 0.5$) based on curve of nonlinear logarithmic trend line. This gives the rough estimates of the extent of LD for association study, but for more accurate estimates, highly significant threshold LD values ($r^2 \geq 0.2$) are also used as a cutoff point. The decrease of the LD within the genetic distance indicates that the portion of LD is conserved with linkage and proportional to recombination (Fig. 3).

Out of which mutation and recombination are the key factors affecting LD significantly. Increased LD is the result of new mutations, population structure, autogamy, genetic isolation, admixture, genetic drift, small founder population size, epistasis, genomic rearrangement, selection and kinship, whereas higher rates of recombination and mutation, recurrent mutations, gene conversion and outcrossing significantly decrease LD. Theoretically, kinship creates LD between genetically linked loci but it can also create LD between genetically unlinked loci when predominant parents are included in the population. The population structure (existences of distinctly clustered subdivisions in a population) and population admixture are the main factors to

Table 2: Factors affecting LD: it can be divided into two types

	Factors increasing LD		Factors decreasing LD
1.	Mutation	1.	Gene conversion
2.	Mating system (self-pollination)	2.	Out crossing
3.	Population structure	3.	Recurrent mutation
4.	Relatedness (kinship)	4.	High recombination
5.	Small founder population size or genetic drift		
6.	Admixture		
7.	Selection		
8.	Epistasis		
9.	Low recombination rate		

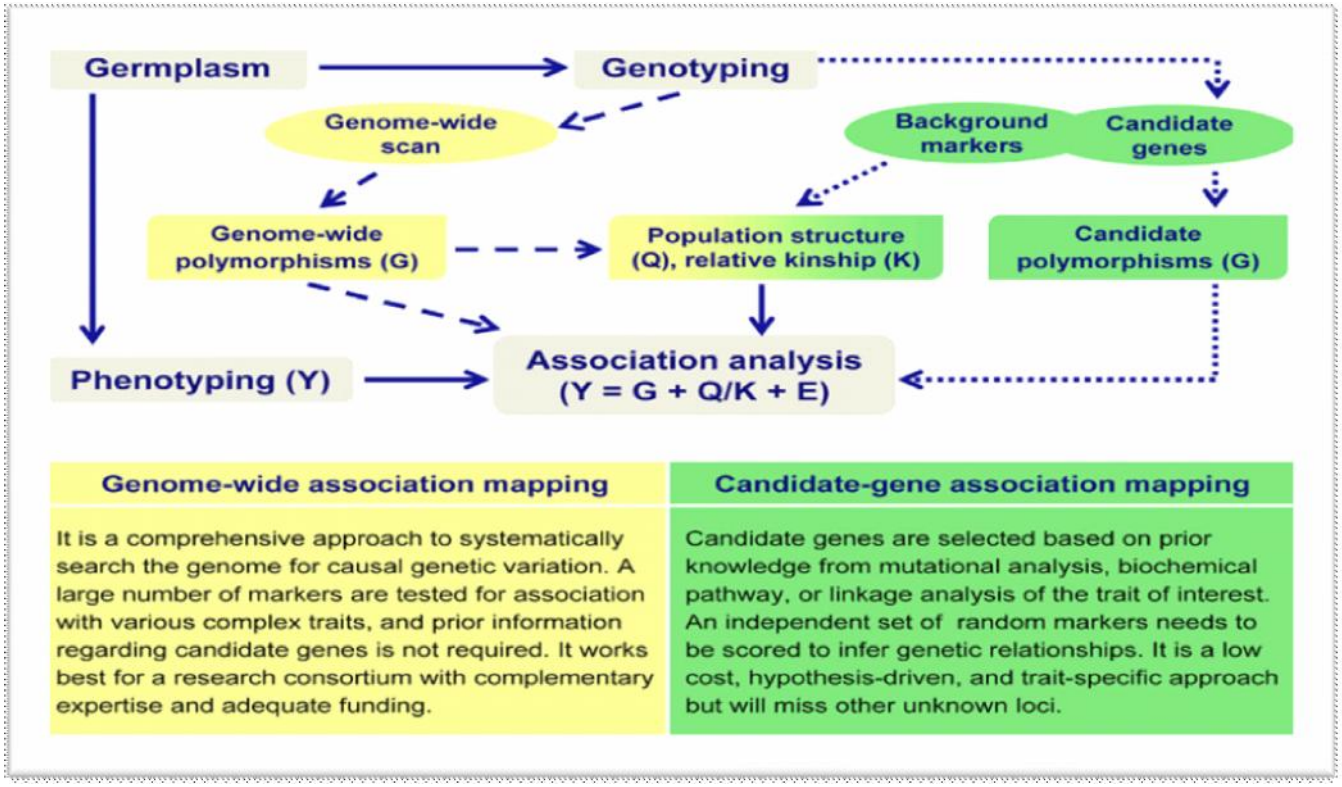

Fig. 4: Types of association mapping

create such an LD between unlinked loci. Theoretically, relatedness generates LD between linked loci, yet it might also generate LD between unlinked loci pairs when predominant parents exist in germplasm groups. There is evidence that relatedness caused LD between linked and unlinked loci in an equal proportion in maize germplasm The other factors such as genetic drift or bottlenecks might have also generated LD in a genome. In contrast, less extensive level of LD (means that LD quickly decays within a short distance) requires many markers to tag a gene of interest, but in high resolution (fine mapping). Hence, choosing a population with low or high level of LD depends on the objective of association mapping study. LD generated by selection, population structure, relatedness, and genetic drift might be theoretically useful for association mapping in specific situations and population groups that reduces number of markers needed for association mapping, but requires serious attention to control factors affecting LD. The extensive level of LD (long stretched LD) reduces the number of markers required for marker-trait association but lowers the mapping resolution (coarse mapping). Conversely, less extensive LD (short stretched LD) needs relatively more number of markers to mine a gene but increase mapping resolution. Selection of a population with LD level higher or lower depends on the objective of mapping study.

Minor alleles (present in less than 10% individuals) largely inflate LD values (Caldwell *et al.*, 2006). Hence, in LD quantification and AM, markers with minor allele frequency are :

- Replaced with missing values (Barnaud *et al.*, 2006; Breseghello and Sorrels, 2006)
- Pooled into common allele class (Hamblin *et al.*, 2004) or
- Removed before analysis (Hamblin *et al.*, 2004; Kraakman *et al.*, 2004; Caldwell *et al.*, 2006; Kraakman *et al.*, 2006).

Types of association mapping :

Genome wide association mapping :

Search whole genome for causal genetic variation. A large number of markers are tested for association with various complex traits and it doesn't require any prior information on the candidate genes.

Candidate gene association mapping :

Dissect out the genetic control of complex traits, based on the available results from genetic, biochemical, or physiology studies in model and non-model plant species (Mackay, 2001). Requires identification of SNPs between lines within specific genes.

Statistical analysis of association mapping :

- Basic – Liner regression, ANOVA, t-test or Chi - square test
- The transmission disequilibrium test
- Genomic control
- Structured association
- Principal components analysis (PCA)
- Haplotype analysis
- Mixed-linear model approach Q+K model.

The transmission disequilibrium test and derivatives :

The first and most robust method of achieving AM was the transmission disequilibrium test (TDT) introduced by Richard Spielman *et al*. in 1993. The TDT provides a way of detecting linkage in the presence of disequilibrium. Neither linkage alone nor disequilibrium alone (*i.e.* between unlinked markers) will generate a positive result so the TDT is an extremely robust way of controlling for false positives. In the absence of linkage between QTL and marker, the expected ratio of transmission to non-transmission is 1:1. In the presence of linkage it is distorted to an extent that depends on the strength of LD between the marker and QTL. The distortion is tested in a Chi-squared test.

Genomic control :

Population structure arising from recent migration and population admixture will generate LD between a trait and markers distributed over the whole genome. This can be detected by studying whether the distribution of the test statistic for association differs from the expected Null distribution. This is the basis of genomic control (GC). To estimate the empirical distribution accurately would require many markers. If the average Chi-squared at a set of 50 control markers is much greater than 1.0, population structure is indicated. GC also corrects for unknown kinship among collections of lines. The presence of related lines can greatly increase the frequency of false positives. For many crop datasets this will be the greatest source of bias.

Structured association :

Structured association (SA) provides a sophisticated approach to detecting and controlling population structure. Again, additional markers are required, randomly distributed across the genome. However, we expect the parental populations themselves to be in linkage equilibrium. By trial and error one could allocate the individuals in our sample to parental populations such that disequilibrium within populations was minimized. First individuals are allocated to populations, then this information is used to control for population membership in the test of association. SA is effective in detecting and adjusting for the presence of population structure, but does not deal with consanguinity within populations. Recently, Ed Buckler's group introduced a method in which population membership is estimated using STRUCTURE and kinship among varieties is estimated empirically from a second set of control markers. The analysis takes into account both population structure and the

correlation between individuals that results from their relationships. This method is implemented in the software TASSEL*.

Logistic regression :

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

Principal component analysis :

It is based on principal component analysis (PCA) across a large number of biallelic control markers with a genome wide distribution. The PCA summarizes the variation observed across all markers into a smaller number of underlying component variables. These can be interpreted as relating to separate, unobserved, sub-populations from which the individuals in the dataset (or their ancestors) originated. The loadings are used to adjust individual candidate marker genotypes (coded numerically) and phenotypes for their ancestry. The adjusted values are independent of estimated ancestry so a statistically significant correlation between an adjusted candidate marker and adjusted phenotype is, therefore, evidence of close linkage of a trait locus to the marker.

Haplotype analysis :

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

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

Estimation of LD using markers :

The quantification methodology of LD, perfectly suitable for biallelic codominant type of markers (majorly, single nucleotide polymorphisms (SNPs) and now largely extended to multiallelic simple sequence repeats-SSRs), has been well developed and used in human, animal, and plant populations. LD quantification using dominant markers (such as random amplified polymorphic DNAs-RAPD and amplified fragment length polymorphisms-AFLPs) is poorly explored and usually subject to wrong perception and interpretation. However, many underrepresented plant species, like forest trees, or other crops with limited genomic information largely rely on dominant type of markers such as RAPDs a Furthermore, even with codominant, and multiallelic SSR markers lacking historical pedigree information, are genotyped. Misassignment of allelic relationships of loci is the concern in association analysis. To avoid such a challenging cases :

- One might select only single band SSR loci and code a dataset as a codominant marker type,
- Alternatively, multiple-band SSRs with unknown allelic relationship may be scored as a dominant marker taking each band as an independent marker locus (uniquely) with a clear size band separationnd AFLPs.

There are also a number of reports where dominantly coded (present versus absent) marker data of RAPD, RFLP, AFLP, "candidate gene" (CAPs) and SSRs were successfully used in genome-wide LD analyses and LD-based association mapping in plants. Although a dominant type of coding has limited statistical power compared to co dominant markers in population based analyses because of missing heterozygote

Table 3: Softwares used in AM

Sr. No.	Software	Focus	Website
1.	STRUCTURE 2.3	Population structure	*http://pntch.bsd.uchicago.edu/software.html*
2.	BAPS 5.0	Population structure	*http://web.abo.fi/fac/mnf/mate/jc. software/bapc.html*
3.	mStruct	Population structure	*http://www.cs.cmu.edu/suyash/mstruct.html*
4.	Haploview 4.2	Haplotype analysis and LD	*http://www.broad.mit.edu/mpg/haploview/*
5.	TASSEL	Stratification LD and AM	*http://www.maizegenetics.net*
6.	GenStat	Stratification LD and AM	*http://www.vsni.co.uk/*
7.	JMP genomics	Stratification LD and structured AM	*http://www.jmp.com/software/genomics*
8.	SVS 7	Stratification LD and AM	*http://goldenhelix.com*

information. Dominant-type markers can be a useful tool to estimate the kinship co-efficients between individuals. It is recommended that a mixture of codominant and dominant markers should be used to better characterization of a genetic structure of a population.

Implication of LD quantification for AM :

- LD more quickly declines in outcrossing plant species than highly self-pollinating plants
- The extent of LD varies across the genomic regions
- LD measures differ per marker systems
- LD blocks in narrow-based germplasm groups are longer than broad-based germplasm groups in plants
- Population characteristics and biological behaviour have serious impact on pattern and structure of LD.

Power of AM :

The power of association mapping is the probability of detecting the true associations within the mapping population size. Power to detect associations depends on :

- Sample size and experimental design
- Accurate phenotypic evaluations.
- Genotyping,
- Genetic architecture.

The power of AM can be increased by better data recording and analysis and increasing population size. For AM study in the presence of population structure Pritchard *et al.* (2000) established a useful technique for structured association (SA). Structured association (SA) uses Bayesian approach (Marttinen and Corander, 2010) to search sub-populations using Q matrix to avoid false positives. Population

Table 4 : AM studies in plants

Species	Germplasm	Trait	Marker system	Reference
Arabidopsis	Diverse accessions	Flowering time/ pathogen resistance	Sequences	Aranzana *et al.* (2005)
	Diverse accessions	Multiple traits	SSRs/SNPs	Ersoz *et al.* (2007)
	Natural accessions	Flowering time	SNPs	Brachi *et al.* (2010)
	Diverse accessions	Climate-sensitive QTL	SNPs	Li *et al.* (2010)
	Landraces	Downy mildew	SNPs	Nemri *et al.* (2010)
Maize	Inbred lines	Aluminum tolerance	SNPs	Krill *et al* . (2010)
	Inbred lines	Drought tolerance	SNPs	Lu *et al.* (2010)
	Inbred lines	Northern leaf blight	SNPs	Poland *et al.* (2011)
	Inbred lines	Southern leaf blight	SNPs	Kump *et al.* (2011)
	Inbred lines	Leaf architecture	SNPs	Tian *et al.* (2011)
Teosinte	Landraces	Domestication-related genes	SNPs	Weber *et al.* (2009)
Wheat	Cultivars	Kernel size, milling quality	SSRs	Breseghello and Sorrells (2006)
	Diverse accessions	Aluminum resistance	DArT	Raman *et al.* (2010)
	Breeding lines	Stem rust resistance	DArT	Yu *et al.* (2011)
	Diverse accessions	Flowering time	SNPs	Rousset *et al.* (2011)
Barley	Inbred lines	Growth habit	SNPs	Rostoks *et al.* (2006)
	Cultivars	Anthocyanin pigmentation	SNPs	Cockram *et al.* (2010)
	Breeding lines	Winterhardiness	SNPs	Von zitzewitz *et al.* (2011)
Oat	Diverse cultivars	Agronomic and kernel quality traits	AFLPs	Achleitner *et al.* (2008)
Rice	Landraces	Heading date, plant height and panicle length	SSrs	Wen *et al.* (2009)
	Landraces	Multiple agronomic traits	SNPs	Huang *et al.* (2010)
Canola	Diverse accessions	Leaf traits, flowering time and phytate content	AFLPs	Zhao *et al.* (2007b)
	Diverse accessions	Oil content	SSRs	Zou *et al.* (2010)
Soybean	Breeding lines	Iron deficiency chlorosis	SSRs	Wang *et al.* (2008)
Cotton	Diverse cultivars	Fibre quality	SSRs	Abdurakhmonov *et al.* (2009)
Peanut	Diverse accessions	Seed quality traits	SSRs-SNPs	Wang *et al.* (2011)
Sugar beet	Inbred lines	Sugar content and yield	SSRs	Stich *et al.* (2008)
	Inbred lines	Multiple traits	SNPs	Wurschum *et al.* (2011)
Alfalfa	Cultivars	Biomass yield and stem composition	SSRs	Li *et al.* (2011a)

SNPs: Single nucleotide polymorphisms, SSRs : Simple sequence repeats, DArT : Diversity arraya technology, AFLPs: Amplified fragment lengment length polymorphisms

structure (Q-matrix) and kinship co-efficient (K-matrix) can be estimated in subpopulations using the programme STRUCTURE (Pritchard and Wen, 2004). Recently, Yu *et al.* (2006) established another approach called a mixed linear model (MLM) to bloc structure information (Q-matrix) and kinship information (K-matrix) in AM analysis. Later on, the Q+K MLM model performed better even in highly structured population of Arabidopsis as compared to any other model that used Q- or K-matrix alone.

Conclusion :

Association mapping after its successful application in human genetics has found its way in plant genetics to help decipher complex quantitative traits. applications of AM has been extended from model plant Arabidopsis to field crops such as rice, wheat, maize, barley, sugarcane and forage grasses. The increasing number of AM studies in crop species indicates the potential of this approach in all plant species in near future. Furthermore, advancements to develop more cost-effective sequencing technologies for efficient genome sequencing of crop plants will certainly accelerate progress in genome-wide association studies (Anonymous, 2007) discovering rare and common alleles (Estivill and Armengol, 2007) and epigenomic information about the trait of interest. This will increase the power of LD-based association mapping for discovering true associations to facilitate its effective utilization in crop breeding programs. Association mapping holds an important and rapidly expanding niche in quantitative trait mapping studies, along with linkage mapping and positional cloning and it is likely that this niche will continue to expand over the next decade.

REFERENCES

Abdurakhmonov, A.Y. and A. Abdukarimov (2008). Application of association mapping to understanding the genetic diversity of plant germplasm resources. *Internat. J. Plant Genom*., 1–18.

Abdurakhomonov, I.Y., Kohel, R.J., Saha, S., Pepper, A.E., Yu, J.Z., Buriev, Z.T., Abdullaev, A., Shermatov, S. Jenkins, J.N. Scheffler, B. and Abdukarimov, A. (2007). Genomewide linkage disequilibrium revealed by microsatellite markers and association study of fibre quality traits in cotton. *Proceedings of the 15th Plant and Animal Genome Conference*, San Diego, CALIFORNIA, USA.

Al-Maskri, A.Y., Sajjad M. and Khan, S.H. (2012). Association mapping: a step forward to discovering new alleles for crop improvement. *Internat. J. Agric. Biol.*, **14** (1) : 153–160.

Anonymous (2007). The wellcome trust case control consortium, genomewide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**: 661–678.

Aranzana, M., Kim, S., Zhao, K., Bakker, E., Horton, M., Jacob, K., Lister, C., Molitor, J., Shindo, C., Tang, C., Toomajian, C., Traw, B., Zheng, H., Bergelson, J., Dean, C., Marjoram, P. and Nordborg, M. (2005). Genome-wide association mapping in arabidopsis identifies previously known flowering time and pathogen resistance genes. *PLoS Genetics*, **1**(5): 60-65.

Barnaud, A., Lacombe, T. and Doligez, A. (2006). Linkage disequilibrium in cultivated grapevine, *Vitis vinifera* L. *Theor. Appl. Genet*., **112** (4) : 708–716.

Brachi, B., Faure, N., Horton, M., Flahauw, E., Vazquez, A., Nordborg, M., Bergelson, J., Cuguen, J. and Roux, F. (2010). Linkage and association mapping of *Arabidopsis thaliana* flowering time in nature. *PLoS Genet.*, **6**(5): 940-947.

Breseghello, F. and Sorrells, M.E. (2006). Association mapping of kernel size and milling quality in wheat (*Triticum aestivum* L.) cultivars. *Genetics*, **172** (2) : 1165–1177.

Caldwell, K.S., Russell, J., Langridge, P. and Powell, W. (2006). Extreme population-dependent linkage disequilibrium detected in an inbreeding plant species, *Hordeum vulgare*. *Genetics*, **172** (1) : 557–567.

Cannon, G.B. (1963). The effects of natural selection on linkage disequilibrium and relative fitness in experimental populations of *Drosophila melanogaster*. *Genetics*, **48** (9) : 1201–1216.

Ehrenreich, I.M., Stafford, P.A. and Purugganan, M.D. (2007). The genetic architecture of shoot branching in Arabidopsis thaliana: a comparative assessment of candidate gene associations vs. quantitative trait locus mapping. *Genetics*, **176** (2) : 1223–1236.

Ersoz, E., Yu, J. and Buckler, E. (2007). Application of linkage disequilibrium and association mapping in crop plants, In: *Genomics-assisted crop improvement*, pp.97-119. Springer, ISBN 987-1-4020-6294-0, Dordrecht. The NETHERLANDS.

Estivill, X. and Armengol, L. (2007). Copy number variants and common disorders: filling the gaps and exploring complexity in genome-wide association studies. *PLoS Genetics*, **3** (1) : 1787–1799.

Flint-Garcia, S.A., Thornsberry, J.M. and Buckler, E.S. (2003). Structure of linkage disequilibrium in plants. *Annl. Rev. Plant Biol*., **54**: 357–374.

Gupta, P.K., Rustgi, S. and Kulwal, P.L. (2005). Linkage disequilibrium and association studies in higher plants: present status and future prospects. *Plant Molec. Biol*., **57** (4) : 461–485.

Hamblin, M.T., Mitchell, S.E., White, G.M., Gallego, J., Kukatla, R., Wing, R.A., Paterson, A.H. 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** (1) : 471–483.

Hirschhorn, J. N. and Daly, M.J. (2005). Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.*, **6** : 95–108.

Karim S., Lyudmyla, V. Malysheva-Otto, Michelle, G., Wirthensohn, Tarkesh-Esfahani, S. Kraakman, A.T.W., Niks, R.E., Van Den Berg, P.M.M.M., Stam, P. and Van Eeuwijk, F.A. (2004). Linkage disequilibrium mapping of yield and yield stability in modern spring barley cultivars. *Genetics*, **168** (1) : 435–446.

Mackay, I. and Powell, W. (2007). Methods for linkage disequilibrium mapping in crops. *Trends Plant Sci*., **12** (2) : 57–63.

Marttinen, P. and Corander, J. (2010). Efficient Bayesian approach for multilocus association mapping including gene-gene interactions. *BMC Bioinformatics*, **11**: 443.

Nemri, A., Atwell, S., Tarone, A., Huang, Y., Zhao, K., Studholme, D., Nordborg, M. and Jones, J. (2010). Genome-wide survey of arabidopsis natural variation in downy mildew resistance using combined association and linkage mapping. *Proc. National Acad. Sci. United States America*, **107** (22): 10302-10307.

Oraguzie, N.C., Wilcox, P.L., Rikkerink, E.H.A. and De Silva, H.N. (2007). Linkage disequilibrium. *In*: (Ed.), *Association Mapping in Plants*, pp. 11-39. Springer, NEW YORK, U.S.A.

Paterson, A.H., DeVerna, J.W., Lanini, B. and Tanksley, S.D. (1990). Fine mapping of quantitative trait loci using selected overlapping recombinant chromosomes, in an interspecies cross of tomato. *Genetics,* **124** (3) : 735–742.

Pritchard, K.J. and Wen, W. (2004). *Documentation for structure software*. The University of Chicago Press, Chicago, Ill, U.S.A.

Risch, N. and Menikangas, K. (1996). The future of genetic studies of complex human diseases. *Science,* **273** (5281) : 1516-1517.

Skøt, L., Humphreys, J., Humphreys, M.O., Thorogood, D., Gallagher, J., Sanderson, R., Armstead, I.P. and Thomas, I.D. (2007). Association of candidate genes with flowering time and water-soluble carbohydrate content in *Lolium perenne* (L.). *Genetics*, **177** (1) : 535–547.

Spielman, R., McGinnis, R. and Ewens, W. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American J. Human Genet.,* **52**(3): 506-516.

Stuber, C.W., Lincoln, S.E., Wolff, D.W., Helentjaris, T. and Lander, E.S. (1992). Identification of genetic factors contributing to heterosis in a hybrid from two elite maize inbred lines using molecular markers. *Genetics,* **132** (3) : 823–839.

Thornsberry, J.M., Goodman, M.M., Doebley, J., Kresovich, S., Nielsen, D. and Buckler, E.S. (2001). Dwarf 8 polymorphisms associate with variation in flowering time. *Nat. Genet*., **28** (3) : 286–289.

Wang, W., Thornton, K., Berry, A. and Long, M. (2002). Nucleotide variation along the Drosophila melanogaster fourth chromosome. *Science*, **295** (5552) : 134–137.

Yu, J., Pressoir, G., Briggs, W.H., Bi, I.V., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Holland, J.B., Kresovich, S. and Buckler, E.S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.*, **38** (2) : 203–208.

Zhu, Y.L., Song, Q.J., Hyten, D.L., Van Tassell, C.P., Matukumalli, L.K., Grimm, D.R., Hyatt, S.M., Fickus, E.W., Young, N.D. and Cregan, P.B. (2003). Single-nucleotide polymorphisms in soybean. *Genetics*, **163** (3) : 1123–1134.

10th Year
★★★★★ of Excellence ★★★★★