Chapter 13
Applications of Linkage Disequilibrium and Association Mapping in Maize

Elhan S. Ersoz, Jianming Yu, and Edward S. Buckler

13.1 Introduction

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). While linkage analysis is based upon detection of non-random association between a genotype and a phenotype in well-characterized pedigrees, association mapping focuses on associations within populations of *unrelated* individuals. In general, chromosomes sampled from *unrelated* individuals in a population will be much more distantly related than those sampled from members of traditional pedigrees. In other words, the time to most recent common ancestor

Elhan S. Ersoz
Institute for Genomic Diversity, Cornell University, Ithaca, NY 14853, USA

Jianming Yu
Department of Agronomy, Kansas State University, Manhatten, KS 66506, USA

Edward S. Buckler
USDA-ARS, US Plant, Soil and Nutrition Laboratory, Ithaca, NY 14853, USA
Institute for Genomic Diversity, 159 Biotechnology, Cornell University, Ithaca, NY 14853, USA
e-mail: esb33@cornell.edu

A.L. Kriz, B.A. Larkins (eds.), *Molecular Genetic Approaches to Maize Improvement*
Biotechnology in Agriculture and Forestry, Vol. 63

(MRCA) of any 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 megabases, association studies deal with segmental exchanges measured in kilobases (Paterson et al. 1990; Stuber et al. 1992; Thornsberry et al. 2001).

13.2 What is Linkage Disequilibrium and How is it Related to Association Mapping Studies

The term *linkage disequilibrium* was first introduced back in the late 1940s to describe the degree of non-random association between pairs of loci. In the absence of demographic effects that might confound the LD patterns, LD summary statistics such as r^2 can be used to define the level of co-occurrence of alleles at two loci (Hill and Robertson 1968). When r^2 is zero, alleles at two loci do not co-occur more frequently than would be expected under random sampling. r^2 approaches its maximum of 1 as alleles at two loci show more frequent co-occurrence within the population sample examined. There are various other LD statistics that can be used for this purpose (Hedrick 1987) all of which aim to estimate the predictive value of a marker locus on another locus that is displaying non-zero LD with it (if LD statistic is zero, two loci examined have zero predictive value for each other).

Association mapping uses these properties of the measures of pairwise LD statistics to infer the predictive value of a marker locus for the association of the chromosomal region where it resides with the phenotype. The high-LD chromosomal region around a marker locus defines the predictive range of a certain genetic marker. If LD within this genomic range is complete, any polymorphism within this range will have the same predictive value for the association with the phenotype. Hence, as a result of a significant marker–phenotype association, it can be concluded that the causative polymorphism resides within this high LD region around the marker locus.

With respect to association mapping, the most significant aspect of LD is its predictive properties over the haplotype it resides in. However, the extent of LD (in base pairs) within species and even within individual genomes is highly variable, and therefore most reliably estimated empirically (Long and Langley 1999).

Theoretical estimation of the levels of LD for realistic population models that does not satisfy the assumptions of the Wright-Fisher model is complex. The hardship is mostly due to the large number of interrelated factors involved in the formation of patterns of LD, including but not limited to genetic drift, population admixture, and natural selection (Pritchard and Przeworski 2001; Wall and Pritchard 2003).

The statistical power of associations is determined by the extent of LD with the causative polymorphism, as well as sample size used for the study (Long and Langley 1999; Wang and Rannala 2005). If LD decays too fast within a region, a large number of markers would be required to scan target regions of a genome. On the other hand, if LD decays too slowly, the size of the haplotype blocks would be too large to unambiguously reveal underlying causative locus. In other words, the decay of LD over physical distance in the study population determines the marker density required and the level of resolution that may be obtained in an association study.

13.2.1 How to Estimate LD

There are several summary statistics proposed for estimation of LD (Hedrick 1987); however, the most commonly used summary statistic within the association study framework is known as r^2 (Hill and Robertson 1968; Lewontin 1988). Conceptually and mathematically r is the Pearson's (product moment) *correlation coefficient* of the correlation that describes the predictive value of the allelic state at one polymorphic locus on the allelic state at another polymorphic locus, where r^2 is the squared value of correlation coefficient that is also called *coefficient of determination*. r^2 explains the proportion of a sample variance of a response variable that is *explained* by the predictor variables when a linear regression is performed.

Lewontin's D is another summary statistic for LD that is commonly used. D describes the difference between the coupling gamete frequencies and repulsion gamete frequencies at two loci. From D a second measure of LD, that is normalized D', can also be estimated. Even in samples taken from populations at equilibrium under neutrality, variances of LD summary statistics are typically large, but D' has the lowest variance (Hedrick 1987). However, estimation using D' may generate erratic and unreliable results when low frequency alleles or small sample sizes are used for the analysis. It is advisable to collapse the alleles using an allele frequency cut-off prior to estimation of LD statistics D and D'.

Other than these commonly used summary statistics for LD, there are also likelihood-based methods that investigate probability of independence between pairs of sites using two-locus sampling distributions, rather than calculating a summary statistic for LD. These methods, usually referred to as model-based LD estimators, also provide means of estimating population recombination parameter $4Nc$

under a neutral equilibrium model from nucleotide sequence data (Golding 1984; Hudson 1985, 2001) or generating other model-based estimates of LD for comparisons with observed patterns (Mueller 2004) under various population structure and demographic history scenarios. Although the estimation of LD through these methods is more computationally intensive compared to pairwise-LD estimation methods, they are extensively used for evolutionary and population genetic studies as well as investigations into the domestication of various crop plant species (Wright et al. 2005; Wright and Gaut 2005).

13.2.2 Interpretation of LD Data

Estimating LD from empirical data is a straightforward procedure; however, interpretation of the results of LD analysis and extrapolation of this information to the genome may be more complex. It is important to estimate the rate of decay of LD with physical distance to be able to extrapolate information gathered from a small collection of sampled loci to the whole genome investigated. This extrapolation is essential for association mapping study design since it may be used to determine the marker density required for scanning previously unexplored regions of the genome as well as the maximum resolution that can be achieved for genotype–phenotype associations in the study population.

The levels of LD are expected to be highly variable across the genome due to several factors, such as variation in recombination rate and selection. For reliable results, this variation needs to be taken into account when designing experiments to exploit LD. Variation in rate of recombination across the genome is a key factor that contributes to the variance observed in patterns of LD. A number of researchers have focused on the distance at which average r^2 is reduced to 0.10 as a reasonable point to conclude that there is minimal LD to detect associations with complex traits. The reasoning for this r^2cut-off is as follows: in a complex trait a large quantitative trait locus (QTL) may only explain approximately 10% of the phenotypic variation. If a marker only explains 10% of the total QTL variation, then the marker will only explain 1% of the phenotypic variation. Detection of locus effects that cause larger than 1% phenotypic variation requires exponentially increasing population sizes, and therefore such small effects would be considered undetectable in a moderate size study population.

To maintain sufficient power for dissection of complex traits through association studies, the choice of marker density and population size are of importance. Not only high enough marker density to screen and target region(s) at blocks of greater LD (i.e. $r^2 > 0.8$) but also large-sized populations are required in order to achieve sufficient power. Current human genetic studies focus on genome scans aiming for much higher LD (e.g. $r^2 > 0.80$) (Barrett and Cardon 2006), and are developing haplotype-based approaches that can help capture more variants (Pe'er et al. 2006).

13.2.3 LD in Maize

Studies on rates of decay of LD in various plant taxa (Flint-Garcia et al. 2003) such as maize (*Zea mays* ssp. *mays*) (Remington et al. 2001b; Ching et al. 2002; Tenaillon et al. 2002; Palaisa et al. 2003), barley (*Hordeum vulgare*) (Caldwell et al. 2004, 2006), *Arabidopsis thaliana* (Nordborg et al. 2002, 2005), sorghum (*Sorghum bicolor*) (Hamblin et al. 2005) and durum wheat (*Triticum durum*) (Maccaferri et al. 2005) indicate tremendous variation in the extent of LD. This variation is mostly due to founder effect followed by genetic drift that leads to unequal number of effective recombinations in species sub-populations. Selfing also plays an important role (Nordborg 2000).

The population sample effect is clearly observed in maize, where LD decays within 1 kb in land races (Tenaillon et al. 2001), in approximately 2 kb in diverse inbred lines (Remington et al. 2001b) and can extend up to 100–500 kb in commercial elite inbred lines (Ching et al. 2002; Jung et al. 2004). One key issue in comparing distances within genes and between genes is that recombination occurs very rarely outside of genes, so LD can extend for great distances in retroposon regions.

LD decay can also vary considerably from locus to locus. For example, significant LD was observed up to 4 kb for the Y1 locus (encoding phytonene synthase), but was seen at only 1 kb for PSY2 (a putative phytonene synthase) in the same maize population (Palaisa et al. 2003). A more recent study showed that LD for some haplotypes extends over 800 kb around Y1 (Palaisa et al. 2004). The Y1 case is a clear example of strong selection, with a decade-long period tremendously reducing the diversity linked to the key polymorphism, which created very extensive LD.

13.3 Association Populations and Statistics

There are five main stages for association studies: (1) selection of population samples; (2) determination of the level and influence of population structure on the sample; (3) phenotyping the population sample for traits of interest; (4) genotyping the population, for either candidate genes/regions or as a genome-wide scan; and (5) testing the genotypes and phenotypes for their associations (Fig. 13.1).

The choice of association test is the last step of the study and is mostly dependent on the previous steps, according to the characteristics of the population that was used to collect the genotypic and phenotypic data (Lewis 2002; Breseghello and Sorrells 2006a, b). Furthermore, possible complications due to population structure in the study sample may adversely affect the association test results. The influence of population structure on each association study depends on the relatedness between sampled individuals in the studied population (Fig. 13.2, Fig. 13.3). Therefore, the populations amenable for association studies may be classified according to the level of relatedness between the individuals forming the association population.

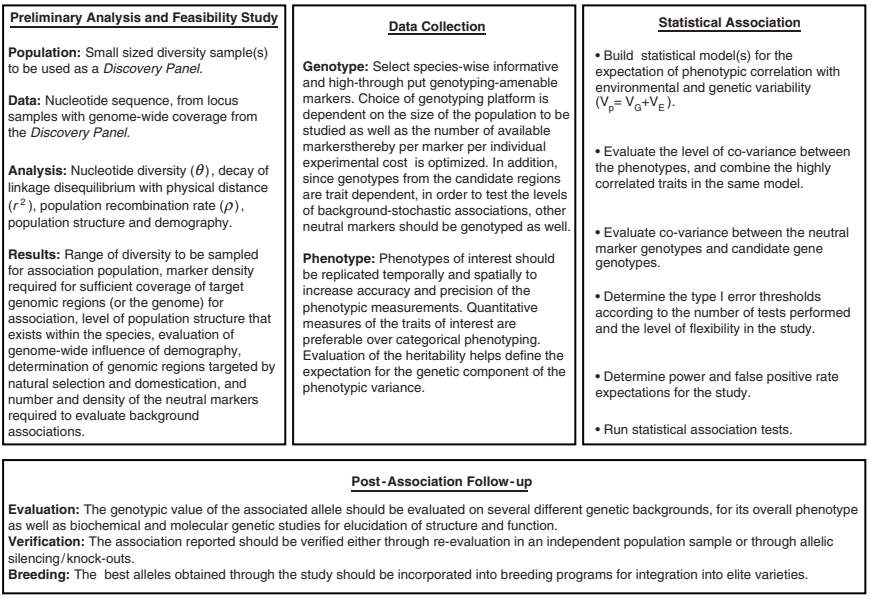

Fig. 13.1 The steps employed during an association study

In the following subsections, we will first discuss the influences of population structure on various association study designs, followed by examples of control on its influences by accounting for the relatedness between individuals forming the association population.

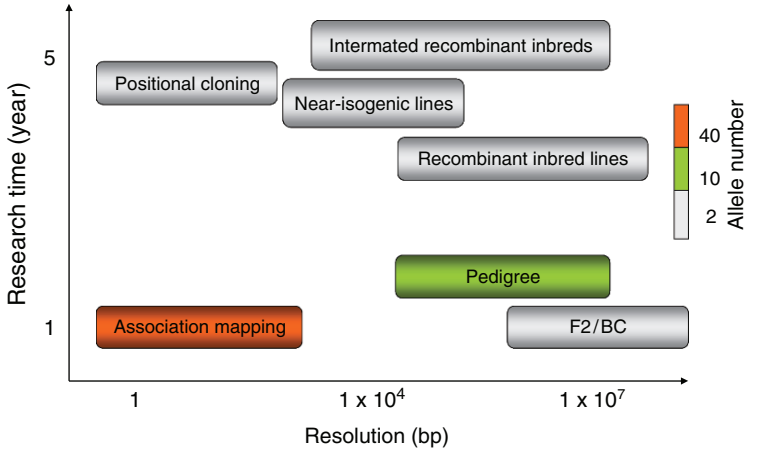

Fig. 13.2 Schematic comparison of various methods for identifying nucleotide polymorphism trait association in terms of resolution, research time and allele number

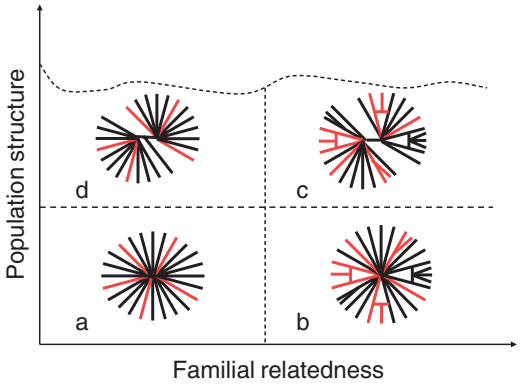

Fig. 13.3 Schematic diagram of the different types of population encountered in association mapping studies. Examples and relevant statistical methods for the analysis of the different population types are described. *a* Ideal sample with subtle population structure and familial relatedness (e.g. F2 population or synthetic), regression and genomic control (GC). *b* Family-based sample (e.g. extended pedigree), transmission disequilibrium test, quantitative transmission disequilibrium test, GC and mixed model (pedigree-based coancestry matrix and relative kinship matrix). *c* Sample with population structure (e.g. maize landraces), structured association (SA) and GC. *d* Sample with both population structure and familial relationships (e.g. maize association panel), SA, GC and mixed model (population structure (Q) plus relative kinship matrix (K))

13.3.1 Population Structure

The most important constraint to the use of association mapping for crop plants is unidentified population substructuring and admixture due to factors such as adaptation or domestication (Thornsberry et al. 2001; Wright and Gaut 2005). Population structure creates genome-wide LD between unlinked loci. When the allele frequencies between sub-populations of a species are significantly different, due to factors such as genetic drift, domestication or background selection, genetic loci that do not have any effect whatsoever on the trait may demonstrate statistical significance for their co-segregations with a trait of interest. Provided that a large number of neutral markers are available for estimation of genome-wide effects of structure, it is possible to statistically account for such effects in association data analysis (Yu et al. 2006b).

In cases where population structuring is mostly due to population stratification (Pritchard 2001; Bamshad et al. 2004), three methods are often acknowledged to be suitable for statistically controlling the effects of population stratification on association tests: (1) genomic control (GC) (Devlin and Roeder 1999; Devlin et al. 2001, 2004); (2) structured association (SA) method, including two extensions that are modified for the type of association study – case control (the SA-model) (Pritchard et al. 2000b) and quantitative trait association study (the Q-model) (Thornsberry et al. 2001; Camus-Kulandaivelu et al. 2006); and (3) the unified mixed model approach (Q+K) (Yu et al. 2006b).

The first method suggested for statistically controlling population structure was GC, which assumes that population structuring has equivalent effects on all loci genome-wide. In the GC method, a small random set of markers (e.g., polymorphisms unlikely to affect the trait of interest) are used to estimate influence of population structure on the association test statistics (*inflation factor*), such that the significance of the association statistic (*P* value) estimated is adjusted to account for population structure. The general principle of GC is to use individual genomes from the sample to estimate the levels of confounding due to substructure and more direct relatedness, such as familial relationship, in the study, and scale the final significance level of the association reported accordingly (Devlin et al. 2001).

Structured association methodology utilizes marker loci unlinked to the candidate genes under investigation to infer *sub-population membership*. The application of structured association to qualitative and quantitative traits is done using the appropriate model, depending on the trait and population type, with either SA or Q models, respectively. In application of SA for quantitative trait association (Q-model), a two-stage procedure is constructed, where for the first stage each subject's probability of membership in each sub-population is estimated (Pritchard et al. 2000a, b), and then in the next stage a test of association is conducted using sub-population membership as a variable for the association model tested (Pritchard et al. 2000b); then, in the next stage, a test of association is conducted using sub-population membership. In case-control studies, the probability of the SNP frequency distribution based on population structure is compared between the case and control samples. For quantitative traits, the population structure estimates are used as co-variates in the regression model that defines the correlation of the genotype with the phenotype (Thornsberry et al. 2001; Camus-Kulandaivelu et al. 2006).

In the unified mixed model approach (aka Q+K model) of Yu et al. (2006b), a large set of random markers that can provide genome-wide coverage are used to estimate population structure (Q) and relative kinship matrix (K), which are fit into a mixed-model framework to test for marker-trait association. In the unified mixed-model approach, each of the factors that may confound association analysis, that is familial relatedness between individuals (K) and relatedness due to population structure (Q), are considered as independent variables within the species population. In order to account for the combined effects of such relatedness factors, they are included as covariates in the regression model that defines the correlation between genotype and phenotype during association testing.

The genetic makeup of the study population that was used to collect genotypic and phenotypic data defines the model and type of association statistics to be used for association tests. This will be discussed further in the next section.

13.3.2 Classic Association Populations

If the individuals forming the study population are *effectively* unrelated, the study population may be considered a random sample of individuals from species

populations and is therefore equivalent to any natural population. The relatedness amongst the individuals forming the population can be either estimated using pedigrees (Emik and Terrill 1949) or inferred using molecular markers (Lynch and Ritland 1999; Wang 2002; Blouin 2003; Oliehoek et al. 2006). These individuals can be either selected from originally natural populations or subselected from selections included in breeding programs, to form a classic association population. Selecting individuals from breeding programs offers the advantage of easy incorporation into future breeding programs; however, the number of lineages incorporated in the association study becomes limited (Breseghello and Sorrells 2006a, b).

All the previously mentioned statistical methods for population structure inferences are applicable to the classic association populations; however, the Q+K model has the widest base of applicability across all structured association study designs in natural populations.

In plants, so far the focus has been on quantitative traits in natural populations. In maize, using diverse inbred lines, it was possible to select a sample of 102 lines with relatively few closely related individuals by sampling across the world's breeding programs (Remington et al. 2001b; Thornsberry et al. 2001). However, as larger samples were gathered to increase statistical power to over 300 maize lines it became extremely difficult to find samples that match the structure expected in natural populations (Flint-Garcia et al. 2005). These are the cases where the combined natural and family-based approaches are most powerful (Yu et al. 2006a). In *Arabidopsis* (Nordborg et al. 2005), natural samples were collected from around the world, but because of strong population structure and selfing, these samples in many respects behave more like families for association mapping purposes (Aranzana et al. 2005). Association studies with some tree species are more likely to fall into the model of effectively unrelated individuals (Thumma et al. 2005; Gonzaléz-Martinéz et al. 2006). Most crop plant studies will probably fall on a continuum between natural and family-based association populations.

13.3.3 Family-Based Association Populations

If the association population is a collection of unrelated families, instead of single unrelated individuals, it is possible to perform a joint linkage and association analysis on the population, that potentially can be more informative on the trait of interest than either approach alone (Holte et al. 1997; Karayiorgou et al. 1999). For instance, in human genetics, where the association populations are collections of parent–offspring trios, two types of study design are considered: transmission disequilibrium tests (TDTs) (Spielman et al. 1993; Allison 1997; Rabinowitz 1997; Monks et al. 1998; Fulker et al. 1999) and family-based association tests (FBATs) (Laird et al. 2000; Lake et al. 2000; Horvath et al. 2001; Lange et al. 2003; Herbert et al. 2006; Laird and Lange 2006). Stich et al. (2006) modified the QTDT algorithm to test its applicability to inbred plant populations, and developed a model named the Quantitative Inbred Pedigree Disequilibrium Test (QIPDT), for analysis of joint linkage and association data from crop plant populations. Another family-based

population design that was essentially developed for crop and livestock breeding is the Henderson's Mixed Model Approach (Henderson 1975), which is generally known for its applications in best linear unbiased predictors (BLUPs). Family-based association study design investigates co-segregation and linkage simultaneously (Spielman et al. 1994).

A long-standing mixed model method has been used by animal scientists to analyze the data from extended pedigree in dairy and beef cattle breeding programs (Henderson 1975, 1976, 1984). The superiority of the mixed model lies in its incorporation of the phenotypic observations from relatives of an individual in the estimation of the breeding value of that individual. The amount of information that is incorporated depends on the heritability of the trait and the genetic relationships (traditionally defined by pedigree information) among individuals. Naturally, this method has been extended to quantify the single gene effect while accounting for the pedigree relationship (Kennedy et al. 1992), and is applicable to association mapping with family-based association populations. Taking this mixed model framework, Yu et al. (2006b) suggested replacing the pedigree-based co-ancestry with a marker-based relative kinship (K) to account for the relatedness among individuals.

This unified mixed model approach is demonstrated to be the most powerful statistic compared to all the rest of the statistics for the family-based association studies and those studies falling between classical and family-based designs. The flexibility and generality of this approach allow association studies to be carried out on any population without the restriction on the specific family structure.

13.3.4 Special Association Populations

Recently, the field of plant association genetics pioneered the use of a new type of association population, designed to incorporate advantages of both linkage-based and LD-based quantitative trait dissection approaches in association studies, in a stronger design than transmission-disequilibrium test (TDT) design. This builds on some of the joint linkage-association approaches encountered in cattle breeding (Meuwissen and Goddard 1997; Blott et al. 2003). Nested association populations (NAM) are developed through controlled crosses between a diverse selection of unrelated individuals according to a breeding scheme that aims to shuffle alleles in diverse samples either across backgrounds or against a reference background, while keeping track of number and locations of the recombination events that shuffle the parental chromosomes (Yu et al. 2006a). The subsequent generations of progeny of the crosses can then be used as association populations. A population generated according to this described scheme not only provides tremendous power to the statistical tests of association, but also enables the projection of genotype information from the parents to the progeny, optimizing genotyping cost for large studies. The cross design is expected to effectively reduce many of the effects of admixture and population structure on the association population. For such populations, a two-step procedure for associations is suggested.

The two-stage study design of nested association mapping requires deep sequencing or genotyping of the parents for SNP identification across the genome, followed by lower density genotyping in the progeny in order to infer the locations of the recombination breakpoints during the crosses. Once the recombination breakpoints are localized and the recombination blocks are traced back to the contributing parent, the haplotype information from the parents can be directly projected on the progeny genome, without further need for genotyping within these blocks.

This design scheme enables the researcher to utilize the advantages of both linkage-based and LD-based genetic mapping approaches. It provides genome-wide coverage with high resolution and is performed on an experimental cross that is robust to genetic heterogeneity, with representation of several alleles per loci in a large population.

Because of the balanced design, straightforward multiple regression approaches can be applied (Yu et al. 2006a) for association testing. Currently, availability of such nested association populations are reported for maize (Yu et al. 2006a) and loblolly pine (Baltunis et al. 2005; Kayihan et al. 2005; Ersoz 2006). Further statistical methods that are going to utilize and combine information from both parent and progeny generations for NAM-type populations are currently under development.

These mentioned association population structures represent the continuum of LD levels from low in classic association populations towards high in biparental breeding populations. Nested association populations that are similar to heterogenous intermated populations (Niebur et al. 2004) fall in the mid-range of this continuum with moderate levels of LD and linkage.

13.4 False Positives and Power of Association

One of the major concerns of association mapping studies is the statistical power of the association testing, since, as it stands, there is a trade-off between the power and accuracy of reporting associations due to false positives. The major determinant of the levels of false positives and power of associations is the level of population structure in the association population.

A false positive (type I error) occurs when a test incorrectly reports that it has found a positive result where none really exists. The classical definition of type I error is an incorrect rejection of the null hypothesis – accepting the alternative hypothesis even though the null hypothesis was true. The second functional biological definition of false positives is also used in association studies. In this framework, false positives arise not only due to the failure of the statistical test performed, but also in cases where the statistical test is valid and the association exists but it is an association with population structure instead of the trait of interest. Population structure can lead to identification of loci that generate statistically significant but biologically invalid associations solely due to their tight correlation with population structure. However, if the population structure in an association study is properly dealt with, this is not expected to be a source of false positives.

Traditionally, type I error rate (α) for multiple testing is controlled with the Bonferroni correction. The Bonferroni correction in general is conservative and leads to power loss for detection if the polymorphisms are in LD and/or the traits are correlated with one another.

Another statistical method suggested for control of multiple testing is the false discovery rate (FDR) procedure. The FDR is the proportion of positive results that are actually false positives versus the whole set of positive results obtained from a statistical test. The procedure can be used to estimate a cutoff for a particular FDR (Benjamini and Hochberg 1995) or an FDR for a particular cutoff (Storey 2002; Storey and Tibshirani 2003). FDR approaches may be most appropriate when multiple traits are being compared or when the markers are not in extensive LD (Chen and Storey 2006). Essentially based on the relative costs of false positives on further follow-up research, appropriate FDRs should be determined and used.

A third procedure that can be applied for multiple testing correction is the permutation test (Churchill and Doerge 1994; Doerge and Churchill 1996), which controls for the genome-wide error rate (GWER). The permutation test has the ability to estimate effects on significance levels caused by the use of correlated markers as well as correlated traits. In this approach, the trait values are permuted relative to the genotypic data. These permutation approaches are appropriate ways to control the GWER; however, they can be quite conservative if one expects numerous QTLs. Recently, the GWER_k approach of Chen and Storey (2006) incorporating a more liberal balance of true and false positives provides a reasonable avenue.

Other than the statistical methods proposed, it is also possible to non-parametrically estimate the FDR through comparison of distributions of P values against a set of markers of known influence and a set of random markers scored on the same association population, with simulations. The probability of false associations is simply the ratio of the proportion of significant associations detected in the random set to the proportion of significant associations detected in the simulated set of known influence loci. This method provides a fast and rigorous way of estimating FDR if a set of random markers has been scored on the association population. Since random markers are required to estimate population structure, this method should be applicable for association testing in most cases.

The power of a statistical test is the probability that the test will reject a false null hypothesis. Some of the relevant parameters that can affect the power of association studies are, but are not limited to, (1) the type of association test – single marker or haplotype based; (2) the multiplicity control method; (3) the population-structure control method; (4) genetic architecture of the trait; (5) population size; (6) marker density; and (7) type of populations used for associations – family based or effectively unrelated (Long and Langley 1999).

Simulation studies that investigate the power of the association tests for the candidate gene association approach report that 300 individuals in a natural population provide enough power to detect *repeatable* associations when population structure is controlled properly (Long and Langley 1999; Thornsberry et al. 2001; Camus-Kulandaivelu et al. 2006; Yu et al. 2006a). These power estimates are based on candidate gene studies, where there are few SNPs being evaluated relative to the entire

genome. Genome scan-type association studies are rapidly becoming feasible, but for such studies the population sample size required to obtain sufficient power will be larger. The exact population size required will depend on the LD structure for the population. Population sizes of 1000 to 5000 genotypes will likely be sufficient in most cases.

The power of association will be low if the trait is highly correlated with population structure. Statistical controls for population structure under such circumstances would result in false negatives. An example of such a case is demonstrated for maize and *Arabidopsis* flowering time traits (Aranzana et al. 2005; Flint-Garcia et al. 2005). The reason for flowering time and population structure to be correlated is that flowering time is an adaptive trait that largely defines the structure. The Q+K model can produce somewhat better results in these situations (Yu et al. 2006b), but in general a different sample or genetic design is required to work with traits that are tightly correlated with population structure. From a study of 60 traits on a maize diversity panel of 302 inbred lines, the only traits that showed strong relationship with structure were two flowering time-related traits.

Three studies using different germplasm have analyzed maize flowering time and the *dwarf8* (*d*8) gene (Thornsberry et al. 2001; Andersen et al. 2005; Camus-Kulandaivelu et al. 2006). These studies highlight the difficulties of studying traits related to population structure. In all three studies, when population structure is ignored, highly significant associations between the traits and polymorphisms in *d*8 are detected that are often much more significant than any of the random markers. It is clear that the putatively functional allele is segregating with a very high allele frequency in some populations, while it is represented at very low frequencies in other populations. This is exactly what would be expected if flowering time is under diversifying selection between the various sub-populations. Furthermore, upon application of standard corrections for managing population structure (Q), the *d*8-flowering time association is significant for some samples but not for others, in all three studies. Essentially, there is low statistical power to evaluate candidate genes that are involved in the clinal adaptation and/or creation of population structure. While empirical significance estimates obtained through contrasting the significances of the candidates with large numbers of random markers, the most effective approach for this type of trait may be specially constructed association populations with balanced designs.

13.5 Phenotyping and Genotyping Strategies for Association Testing

As in all other quantitative genetic studies, the success of an association study is heavily dependent on the accurate evaluation of the phenotype of interest. The within-population variation observed for genotypes and phenotypes for an association is much greater than that found in most bi-parental mapping populations. While greater variation is preferable when aiming for higher resolution and allele mining,

it can pose problems for accurate evaluation of this variation in a meaningful way in a single environment.

The inherent variation observed in phenotypic trait measurement, when combined with the substantial genetic variation included in some association studies, requires careful experimental design to acquire quality data. In addition, evaluations in multiple environments with controls and unbalanced designs may be required. In our experience with maize, we found that evaluating the germplasm in short-day environments facilitated some trait evaluation by reducing photoperiod effects between lines. Additionally, we found that evaluating the germplasm in testcrosses (F1 hybrids) has reduced the phenotypic range to a manageable level. Since each of these approaches interact with the genetic architectures of the traits, future studies will be needed to fully understand the tradeoffs of various study design approaches.

In the association study design, genotyping is required for inferences both on the genotype/phenotype associations and on the population structure and demography. The first aim of querying candidate regions for polymorphisms is best achieved by genotyping SNPs within these candidate regions. The second aim of gathering information on population-specific phenomena, such as structure, linkage, demography and kinship, can be achieved through genotyping neutral background markers, such as SNPs on non-coding regions and SSRs (simple sequence repeats) distributed evenly throughout the genome.

All genetic markers can be used for investigating association; however, SNPs potentially have the most utility compared to other genetic markers. Various assays were developed for detection of known and unknown SNPs. Some are relatively easy to implement and low in cost, while others are developed for high-volume screening at substantial cost. As the cost of genotyping diminishes, genome-wide scans of all available polymorphisms in a species' genome are becoming rapidly feasible and preferable over targeted SNP genotyping approaches. SSR markers have historically been useful in association studies and do have high information content, but they may be difficult to find in candidate gene regions and they are several-fold more expensive to score than SNPs.

For the purposes of inferences on the population history, genotype information from a large number of neutral marker loci is required. We are using the term neutral marker loosely here to indicate the non-candidate loci, i.e. the loci that were *not* designated as candidate loci that can putatively influence a trait of interest. The density of the markers required should be scaled to provide genome-wide coverage. Simulation studies suggest 100 SSR or 200 SNP markers would suffice to get a reasonable estimate of population structure and relatedness for most crop plants (J. Yu and E.S. Buckler, unpublished results).

When targeting candidate loci for association studies, the greatest statistical power is achieved when the marker and QTL have equal allele frequencies (Abecasis et al. 2001) in the study population. This is due to the opportunity created for maximal linkage and LD, since robust detection of associations requires that the marker and trait loci are in phase. If there is no knowledge of the QTL frequency distribution a priori, the best alternative is to choose markers with a wide range of allele frequencies that are likely to mimic the QTL mutation rate. Some SSRs probably mutate

faster and have a different frequency distribution than QTLs, which may make them less useful for association mapping. SNPs with a wide range of allele frequencies are most likely to be informative. In order to maximize the information content of SNPs, a large number of them can be chosen to scan a particular genomic region, and this can be achieved with numerous algorithms available for choosing SNPs (Daly et al. 2001; Johnson et al. 2001; Patil et al. 2001; Gabriel et al. 2002; Ackerman et al. 2003; Ke and Cardon 2003; Sebastiani et al. 2003; Zhang and Jin 2003; Halldorsson et al. 2004; Forton et al. 2005).

Whether the trait of interest has a binary or quantitative phenotype, it is also of interest for the association study design. When a binary trait is being investigated, case-control-type populations are required for association analysis, where equivalent sized sub-populations of individuals that display the phenotype of interest (cases) and do not display the phenotype of interest (controls) are queried for allelic association of genetic loci with the case and control phenotypes in a statistically significant manner. The statistical test performed is simply an hypothesis test that asks whether or not the allelic frequency distribution of a locus is the same or different for a given locus between the two sub-populations. Bulk segregant analysis (BSA)-type (Michelmore et al. 1991) bulked sample genotype screening methods for all the available marker loci may facilitate candidate gene and association discovery for binary traits (Shaw et al. 1998). The challange of case-control type studies is to make sure that the case and control groups are comparable in terms of their genetic makeup. Most of the statistical methods aim to detect and correct for the effects of population statification and ancestry differences between the case and control groups (Pritchard et al. 2000b; Price et al. 2006).

13.6 Association Mapping in Crop Plants

The motivations for attempting association mapping in different crop plants are highly variable. For historically well-studied crop plants, such as maize and rice, the major motivation for the association approach is dissection of complex traits at very high-level resolution, as well as allele mining from natural genetic diversity resources. For other organisms where there is insufficient or few genetic resources, the major motivation is functional marker development and identification of molecular markers tightly linked to the trait locus for marker assisted selection and breeding practices. Thus, each association study stands alone for its own motivations and should be evaluated for its utility and success based on the initial motivations and aims.

The association mapping approach requires extensive infrastructure development and preliminary studies to determine population structure and LD (Fig. 13.1). Once the preliminary data and infrastructure for association mapping for a species are available, several association studies on various plant taxa report successful results for tests of associations between candidate locus genotypes and various complex phenotypes (Table 13.1).

Table 13.1 Association studies that report significant results. SA Structured association; MLM mixed linear model

Species	Population type	Association method	Trait	References
Zea mays	Diverse inbred lines	SA (Q model)	Flowering time	Thornsberry et al. 2001; Andersen et al. 2005; Camus-Kulandaivelu et al. 2006
		SA (Q model)	Kernel composition Starch pasting properties	Wilson et al. 2004
		SA (Q model)	Maysin synthesis	Szalma et al. 2005
		Case-control	Carotenoid content	Palaisa et al. 2004
		MLM (Q+K model)	Carotenoid content	Harjes et al. 2008
Zea mays	Diverse inbred lines	Haplotype tree scanning	Sweet taste	Tracy et al. 2006

In the model organism *Arabidopsis*, the association mapping practice is mostly motivated by generating proof of concept, identification of QTLs involved in adaptation, and additional alleles to supplement other mutagenesis approaches. The candidate-gene association study at the *CRY2-Cryptochrome2* locus reported diverse functional alleles (Olsen et al. 2004). In their first attempt at a genome-wide association study in *Arabidopsis*, Aranzana et al. (2005) reported identification of previously known flowering time (*FRI* locus) and three known pathogen-resistance genes.

In maize, all reported association studies so far have targeted candidate genes with known mutant phenotypes and are motivated by high resolution mapping and allele mining purposes. For instance, *d*8locus with flowering time (Thornsberry et al. 2001; Andersen et al. 2005; Camus-Kulandaivelu et al. 2006), *bt2* (*brittle2*), *sh1* (*shrunken1*) and *sh2* (*shrunken2*) with kernel composition, *ae1* (*amylose extender1*) and *sh2* (*shrunken2*) with starch pasting properties (Wilson et al. 2004) and sweet taste (Tracy et al. 2006), *a*1(*anthocyaninless1*) and *whp1* (*whitepollen1*) genes with maysin synthesis (Szalma et al. 2005), and *lyc-e* (*lycopene epsilon cyclase*) gene with carotenoid content (Harjes et al. 2008) are studies that report very high resolution associations, as well as localizing the causative polymorphism within 1–2 kb of the marker loci reported. In maize, very little is known about association mapping from a genomic scale, mostly due to incomplete genomic sequence and very rapid decay of LD. At the *Y*1 locus a relatively large genomic context was examined. *Y*1 is a key gene in carotenoid production in maize (Buckner et al. 1990, 1996), and through an association study (Palaisa et al. 2003) the allelic variation was traced down to multiple independent insertions in the *Y*1 promoter region that cause up-regulation of the downstream *Y*1 gene. At this locus, associations were also shown to extend to neighboring genes (Palaisa et al. 2004), albeit with weaker significances. This extended LD is mostly the result of breeding efforts

in the twentieth century that specifically targeted this simple Mendelian inherited trait. The extended LD at the *Y*1 locus is likely to be one of the most extensive in the maize genome, effective over hundreds of kilobases, while other domestication loci, *tb1*(*teosinte branched 1*) (Lukens and Doebley 2001) and *tga* (*teosinte glume architecture*) (Wang et al. 2005), show LD that extends over tens of kilobases. However, it should be emphasized that *tb1* and *tga* domestication loci demonstrate patterns of reduced diversity as well as extended LD, indicating that the estimates of LD are not as efficient as they are at *Y*1. Furthermore it is plausible to assume that not all of the selection events may have similar LD patterns to that of the *Y*1 locus.

Another motivation for the association approach is the opportunity to unify the elite germplasm resource of an organism through investigation of the breeding material. In an association study, Breseghello and Sorrells (2006b) investigated wheat kernel size and milling quality in an elite germplasm collection of soft-winter wheat from eastern USA. They identified three candidate regions on chromosomes 2D, 5A and 5B that are significantly associated with these traits (Breseghello and Sorrells 2006a). This study clearly demonstrates how results obtained from association mapping-based genetic trait dissection studies can be utilized for marker-assisted selection.

13.7 Conclusions

So far, map-based cloning approaches have been reported to successfully isolate 12 major-effect QTLs and nine small-effect QTLs (Price 2006). The time from QTL mapping to positional cloning is estimated to be between 5 and 10 years, while sufficient marker resolution for QTL cloning through association mapping can be achieved within 2–3 years. Furthermore, there is a substantial lag between QTL discovery and marker assisted crop improvement practices dedicated to verification of the presence and stability of QTL in traditional linkage-based studies. In a well-designed association study, some of the results can be immediately applied to marker-assisted improvement.

The true large-scale applications of association mapping will become apparent as multiple species begin to have marker densities sufficiently high for whole genome scan by association mapping. Currently, several research groups are working on whole genome scan approaches in half a dozen species that have whole genome sequences available, and there are at least 50 more species whose genome sequences will be completed in the near future.

The goal of association mapping in many crop plants is to identify key genes controlling various traits and mine the best alleles from diverse germplasm for incorporation into elite breeding material. Traditionally, genetic markers were mostly used for trait improvement through several breeding-based approaches, such as marker assisted selection (MAS), marker assisted breeding (MAB) and mapping as you go (MAYG) (Podlich et al. 2004), as well as QTL

cloning/transformation-based approaches (Remington et al. 2001a). Association mapping has the potential to provide numerous useful alleles to these marker-assisted breeding programs. Marker-assisted breeding programs using association data are now underway in numerous plant breeding companies. In the next few years, we will also witness applications of association mapping and MAS for public 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

Abecasis GR, Cookson WO, Cardon LR (2001) The power to detect linkage disequilibrium with quantitative traits in selected samples. Am J Hum Genet 68:1463–1474

Ackerman H, Usen S, Mott R, Richardson A, Sisay-Joof F, Katundu P, Taylor T, Ward R, Molyneux M, Pinder M, Kwiatkowski DP (2003) Haplotypic analysis of the TNF locus by association efficiency and entropy. Genome Biol 4:R24

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

Andersen JR, Schrag T, Melchinger AE, Zein I, Lubberstedt T (2005) Validation of Dwarf8 polymorphisms associated with flowering time in elite European inbred lines of maize (*Zea mays* L.). Theor Appl Genet 111:206–217

Aranzana MJ, Kim S, Zhao K, Bakker E, Horton M, Jakob K, Lister C, Molitor J, Shindo C, Tang C, Toomajian C, Traw B, Zheng H, Bergelson J, Dean C, Marjoram P, Nordborg M (2005) Genome-wide association mapping in Arabidopsis identifies previously known flowering time and pathogen resistance genes. PLoS Genet 1:e60

Baltunis BS, Huber DA, White TL, Golfard B, Stelzer HE (2005) Genetic effects of rooting loblolly pine stem cuttings from a partial diallel mating design. Can J Forest 35:1098–1108

Bamshad M, Wooding S, Salisbury BA, Stephens JC (2004) Deconstructing the relationship between genetics and race. Nat Rev Genet 5:598–609

Barrett JC, Cardon LR (2006) Evaluating coverage of genome-wide association studies. Nat Genet 38:659–662

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate – a practical and powerful approach to multiple testing. J R Stat Soc Series B 57:289–300

Blott S, Kim JJ, Moisio S, Schmidt-Kuntzel A, Cornet A, Berzi P, Cambisano N, Ford C, Grisart B, Johnson D, Karim L, Simon P, Snell R, Spelman R, Wong J, Vilkki J, Georges M, Farnir F, Coppieters W (2003) Molecular dissection of a quantitative trait locus: a phenylalanine-to-tyrosine substitution in the transmembrane domain of the bovine growth hormone receptor is associated with a major effect on milk yield and composition. Genetics 163:253–266

Blouin JD (2003) DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. Trends Ecol Evol 18:503–511

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

Breseghello F, Sorrells M (2006b) Association analysis as a strategy for improvement of quantitative traits in plants. Crop Sci 46:1323–1330

Buckner B, Kelson TL, Robertson DS (1990) Cloning of the y1 locus of maize, a gene involved in the biosynthesis of carotenoids. Plant Cell 2:867–876

Buckner B, Miguel PS, Janick-Buckner D, Bennetzen JL (1996) The y1 gene of maize codes for phytoene synthase. Genetics 143:479–488

Caldwell KS, Langridge P, Powell W (2004) Comparative sequence analysis of the region harboring the hardness locus in barley and its colinear region in rice. Plant Physiol 136: 3177–3190

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

Camus-Kulandaivelu L, Veyrieras JB, Madur D, Combes V, Fourmann M, Barraud S, Dubreuil P, Gouesnard B, Manicacci D, Charcosset A (2006) Maize adaptation to temperate climate: relationship between population structure and polymorphism in the Dwarf8 gene. Genetics 172:2449–2463

Chen L, Storey JD (2006) Relaxed significance criteria for linkage analysis. Genetics 173:2371–2381

Ching A, Caldwell KS, Jung M, Dolan M, Smith OS, Tingey S, Morgante M, Rafalski AJ (2002) SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. BMC Genet 3:19

Churchill GA, Doerge RW (1994) Empirical threshold values for quantitative trait mapping. Genetics 138:963–971

Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES (2001) High-resolution haplotype structure in the human genome. Nat Genet 29:229–232

Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55:997–1004

Devlin B, Roeder K, Wasserman L (2001) Genomic control, a new approach to genetic-based association studies. Theor Popul Biol 60:155–166

Devlin B, Bacanu SA, Roeder K (2004) Genomic control to the extreme. Nat Genet 36:1129–1130; author reply 1131

Doerge RW, Churchill GA (1996) Permutation tests for multiple loci affecting a quantitative character. Genetics 142:285–294

Emik LO, Terrill CE (1949) Systematic procedures for calculating inbreeding coefficients. J Hered 40:51–55

Ersoz ES (2006) Candidate gene-association mapping for dissecting fungal disease resistance in loblolly pine. PhD Dissertation in Genetics, University of California, Davis

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

Flint-Garcia SA, Thuillet AC, Yu J, Pressoir G, Romero SM, Mitchell SE, Doebley J, Kresovich S, Goodman MM, Buckler ES (2005) Maize association population: a high-resolution platform for quantitative trait locus dissection. Plant J 44:1054–1064

Forton J, Kwiatkowski D, Rockett K, Luoni G, Kimber M, Hull J (2005) Accuracy of haplotype reconstruction from haplotype-tagging single-nucleotide polymorphisms. Am J Hum Genet 76:438–448

Fulker DW, Cherny SS, Sham PC, Hewitt JK (1999) Combined linkage and association sib-pair analysis for quantitative traits. Am J Hum Genet 64:259–267

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D (2002) The structure of haplotype blocks in the human genome. Science 296:2225–2229

Geiringer H (1944) On the probability theory of linkage in Mendelian heredity. Ann Math Stat 15(1)25–57

Golding GB (1984) The sampling distribution of linkage disequilibrium. Genetics 108:257–274

Gonzaléz-Martinéz SC, Wheeler N, Ersoz ES, Neale DB (2006) Association genetics in *Pinus taeda* L.I. Wood property traits. Genetics 175:399–409

Halldorsson BV, Bafna V, Lippert R, Schwartz R, De La Vega FM, Clark AG, Istrail S (2004) Optimal haplotype block-free selection of tagging SNPs for genome-wide association studies. Genome Res 14:1633–1640

Hamblin MT, Salas Fernandez MG, Casa AM, Mitchell SE, Paterson AH, Kresovich S (2005) Equilibrium processes cannot explain high levels of short- and medium-range linkage disequilibrium in the domesticated grass *Sorghum* bicolor. Genetics 171:1247–1256

Harjes CE, Rocheford TR, Bai L,, Brutnell TP, Kandianis CB, Sowinski SG, Stapleton AE, Vallabhaneni R, Williams M, Wurtzel ET, Yan J, Buckler ES (2008) Natural genetic variation in lycopene epsilon cyclase tapped for maize biofortification. Science 319(5861):330–333

Hedrick PW (1987) Gametic disequilibrium measures – proceed with caution. Genetics 117:331–341

Henderson CR (1975) Best linear unbiased estimation and prediction under a selection model. Biometrics 31:423–447

Henderson CR (1976) Simple method for computing inverse of a numerator relationship matrix used in prediction of breeding values. Biometrics 32:69–83

Henderson CR (1984) Application of linear models in animal breeding. University of Guelph, Guelph

Herbert A, Gerry NP, McQueen MB, Heid IM, Pfeufer A, Illig T, Wichmann HE, Meitinger T, Hunter D, Hu FB, Colditz G, Hinney A, Hebebrand J, Koberwitz K, Zhu X, Cooper R, Ardlie K, Lyon H, Hirschhorn JN, Laird NM, Lenburg ME, Lange C, Christman MF (2006) A common genetic variant is associated with adult and childhood obesity. Science 312:279–283

Hill WG, Robertson A (1968) Linkage disequilibrium in finite populations. Theor Appl Genet 38:226–231

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

Holte S, Quiaoit F, Hsu L, Davidov O, Zhao LP (1997) A population based family study of a common oligogenic disease – part I: association/aggregation analysis. Genet Epidemiol 14:803–807

Horvath S, Xu X, Laird NM (2001) The family based association test method: strategies for studying general genotype–phenotype associations. Eur J Hum Genet 9:301–306

Hudson RR (1985) The sampling distribution of linkage disequilibrium under an infinite allele model without selection. Genetics 109:611–631

Hudson RR (2001) Two-locus sampling distributions and their application. Genetics 159:1805–1817

Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, Ueda H, Cordell HJ, Eaves IA, Dudbridge F, Twells RC, Payne F, Hughes W, Nutland S, Stevens H, Carr P, Tuomilehto-Wolf E, Tuomilehto J, Gough SC, Clayton DG, Todd JA (2001) Haplotype tagging for the identification of common disease genes. Nat Genet 29:233–237

Jung M, Ching A, Bhattramakki D, Dolan M, Tingey S, Morgante M, 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

Karayiorgou M, Sobin C, Blundell ML, Galke BL, Malinova L, Goldberg P, Ott J, Gogos JA (1999) Family-based association studies support a sexually dimorphic effect of COMT and MAOA on genetic susceptibility to obsessive-compulsive disorder. Biol Psychiatry 45:1178–1189

Kayihan GC, Huber DA, Morse AM, White TL, Davis JM (2005) Genetic dissection of fusiform rust and pitch canker disease traits in loblolly pine. Theor Appl Genet 110:948–958

Ke X, Cardon LR (2003) Efficient selective screening of haplotype tag SNPs. Bioinformatics 19:287–288

Kennedy B, Quinton M, Vanarendonk J (1992) Estimation of effects of single genes on quantitative traits. J Anim Sci 70:2000–2012

Laird NM, Lange C (2006) Family-based designs in the age of large-scale gene-association studies. Nat Rev Genet 7:385–394

Laird NM, Horvath S, Xu X (2000) Implementing a unified approach to family-based tests of association. Genet Epidemiol 19:S36–S42

Lake SL, Blacker D, Laird NM (2000) Family-based tests of association in the presence of linkage. Am J Hum Genet 67:1515–1525

Lange C, Lyon H, DeMeo D, Raby B, Silverman EK, Weiss ST (2003) A new powerful non-parametric two-stage approach for testing multiple phenotypes in family-based association studies. Hum Hered 56:10–17

Lewis CM (2002) Genetic association studies: design, analysis and interpretation. Brief Bioinform 3:146–153

Lewontin RC (1988) On measures of gametic disequilibrium. Genetics 120:849–852

Lewontin RC, Kojima K (1960) The evolutionary dynamics of complex polymorphisms. Evolution 14:458–472

Long AD, Langley CH (1999) The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. Genome Res 9:720–731

Lukens L, Doebley J (2001) Molecular evolution of the teosinte branched gene among maize and related grasses. Mol Biol Evol 18:627–638

Lynch M, Ritland K (1999) Estimation of pairwise relatedness with molecular markers. Genetics 152:1753–1766

Maccaferri M, Sanguineti MC, Noli E, Tuberosa R (2005) Population structure and long-range disequilibrium in a durum wheat elite collection. Mol Breed 15:271–290

Meuwissen TH, Goddard ME (1997) Estimation of effects of quantitative trait loci in large complex pedigrees. Genetics 146:409–416

Michelmore RW, Paran I, Kesseli RV (1991) Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. Proc Natl Acad Sci USA 88:9828–9832

Monks SA, Kaplan NL, Weir BS (1998) A comparative study of sibship tests of linkage and/or association. Am J Hum Genet 63:1507–1516

Mueller J (2004) Linkage disequilibrium for different scales and applications. Brief Bioinfo 5:355–364

Niebur W, Rafalski JA, Smith OS, Cooper M (2004) New directions for a diverse planet. Proceedings of the 4th International Crop Science Congress, Brisbane

Nordborg M (2000) Linkage disequilibrium, gene trees, selfing: an ancestral recombination graph with partial self-fertilization. Genetics 154:923–929

Nordborg M, Tavare S (2002) Linkage disequilibrium: what history has to tell us. Trends Genet 18:83–90

Nordborg M, Borevitz JO, Bergelson J, Berry CC, Chory J, Hagenblad J, Kreitman M, Maloof JN, Noyes T, Oefner PJ, Stahl EA, Weigel D (2002) The extent of linkage disequilibrium in *Arabidopsis thaliana*. Nat Genet 30:190–193

Nordborg M, Hu TT, Ishino Y, Jhaveri J, Toomajian C, Zheng H, Bakker E, Calabrese P, Gladstone J, Goyal R, Jakobsson M, Kim S, Morozov Y, Padhukasahasram B, Plagnol V, Rosenberg NA, Shah C, Wall JD, Wang J, Zhao K, Kalbfleisch T, Schulz V, Kreitman M, Bergelson J (2005) The pattern of polymorphism in *Arabidopsis thaliana*. PLoS Biol 3:e196

Oliehoek PA, Windig JJ, van Arendonk JA, Bijma P (2006) Estimating relatedness between individuals in general populations with a focus on their use in conservation programs. Genetics 173:483–496

Olsen KM, Halldorsdottir SS, Stinchcombe JR, Weinig C, Schmitt J, Purugganan MD (2004) Linkage disequilibrium mapping of Arabidopsis CRY2 flowering time alleles. Genetics 167:1361–1369

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

Palaisa K, Morgante M, Tingey S, 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

Paterson AH, DeVerna JW, Lanini B, Tanksley SD (1990) Fine mapping of quantitative trait loci using selected overlapping recombinant chromosomes, in an interspecies cross of tomato. Genetics 124:735–742

Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP, Nguyen BT, Norris MC, Sheehan JB, Shen N, Stern D, Stokowski RP, Thomas DJ, Trulson MO, Vyas KR, Frazer KA, Fodor SP, Cox DR (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science 294:1719–1723

Pe'er I, Chretien YR, de Bakker PI, Barrett JC, Daly MJ, Altshuler DM (2006) Biases and reconciliation in estimates of linkage disequilibrium in the human genome. Am J Hum Genet 78:588–603

Podlich D, Winkler C, Cooper M (2004) Mapping as you go. Crop Sci 44:1560–1571

Price AH (2006) Believe it or not, QTLs are accurate! Trends Plant Sci 11:213–216

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904–909

Pritchard JK (2001) Deconstructing maize population structure. Nat Genet 28:203–204

Pritchard JK, Przeworski M (2001) Linkage disequilibrium in humans: models and data. Am J Hum Genet 69:1–14

Pritchard JK, Stephens M, Donnelly P (2000a) Inference of population structure using multilocus genotype data. Genetics 155:945–959

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000b) Association mapping in structured populations. Am J Hum Genet 67:170–181

Rabinowitz D (1997) A transmission disequilibrium test for quantitative trait loci. Hum Hered 47:342–350

Remington DL, Ungerer MC, Purugganan MD (2001a) Map-based cloning of quantitative trait loci: progress and prospects. Genet Res 78:213–218

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

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516–1517

Sebastiani P, Lazarus R, Weiss ST, Kunkel LM, Kohane IS, Ramoni MF (2003) Minimal haplotype tagging. Proc Natl Acad Sci USA 100:9900–9905

Shaw SH, Carrasquillo MM, Kashuk C, Puffenberger EG, Chakravarti A (1998) Allele frequency distributions in pooled DNA samples: applications to mapping complex disease genes. Genome Res 8:111–123

Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52:506–516

Spielman RS, McGinnis RE, Ewens WJ (1994) The transmission/disequilibrium test detects cosegregation and linkage. Am J Hum Genet 54:559–560; author reply 560–553

Stich B, Melchinger AE, Piepho H-P, Heckenberger M, Maurer HP, Reif JC (2006) New test for family-based association mapping with inbred lines from plant breeding programs. Theor Appl Genet 113(6):1121–1130

Storey JD (2002) A direct approach to false discovery rates. J R Stat Soc Series B 64(3):479–498

Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. P Natl Acad Sci USA 100:9440–9445

Stuber CW, Lincoln SE, Wolff DW, Helentjaris T, Lander ES (1992) Identification of genetic factors contributing to heterosis in a hybrid from two elite maize inbred lines using molecular markers. Genetics 132:823–839

Szalma SJ, Buckler ES, Snook ME, McMullen MD (2005) Association analysis of candidate genes for maysin and chlorogenic acid accumulation in maize silks. Theor Appl Genet 110:1324–1333

Tenaillon MI, Sawkins MC, Long AD, Gaut RL, Doebley JF, Gaut BS (2001) Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). Proc Natl Acad Sci USA 98:9161–9166

Tenaillon MI, Sawkins MC, Anderson LK, Stack SM, Doebley J, Gaut BS (2002) Patterns of diversity and recombination along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). Genetics 162:1401–1413

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

Thumma BR, Nolan MF, Evans R, Moran GF (2005) Polymorphisms in cinnamoyl CoA reductase (CCR) are associated with variation in microfibril angle in *Eucalyptus* spp. Genetics 171:1257–1265

Tracy WF, Whitt SR, Buckler ES (2006) Recurrent mutation and genome evolution: example of *Sugary1* and the origin of sweet maize. Crop Sci 46:1–7

Wall JD, Pritchard JK (2003) Haplotype blocks and linkage disequilibrium in the human genome. Nat Rev Genet 4:587–597

Wang H, Nussbaum-Wagler T, Li B, Zhao Q, Vigouroux Y, Faller M, Bomblies K, Lukens L, Doebley JF (2005) The origin of the naked grains of maize. Nature 436:714–719

Wang J (2002) An estimator for pairwise relatedness using molecular markers. Genetics 160:1203–1215

Wang Y, Rannala B (2005) In silico analysis of disease-association mapping strategies using the coalescent process and incorporating ascertainment and selection. Am J Hum Genet 76:1066–1073

Wilson LM, Whitt SR, Ibanez AM, Rocheford TR, Goodman MM, Buckler ES (2004) Dissection of maize kernel composition and starch production by candidate gene association. Plant Cell 16:2719–2733

Wright SI, Gaut BS (2005) Molecular population genetics and the search for adaptive evolution in plants. Mol Biol Evol 22:506–519

Wright SI, Bi IV, Schroeder SG, Yamasaki M, Doebley JF, McMullen MD, Gaut BS (2005) The effects of artificial selection on the maize genome. Science 308:1310–1314

Yu J, Holland JB, McMullen MD, Buckler ES (2006a) Genetic design and statistical power of nested association mapping in maize genetics. Nat Genetics 178:539–551

Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES (2006b) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38:203–208

Zhang K, Jin L (2003) HaploBlockFinder: haplotype block analyses. Bioinformatics 19:1300–1301