Published July, 2008



# Status and Prospects of Association Mapping in Plants

Chengsong Zhu, Michael Gore, Edward S. Buckler, and Jianming Yu*

## Abstract

There is tremendous interest in using association mapping to identify genes responsible for quantitative variation of complex traits with agricultural and evolutionary importance. Recent advances in genomic technology, impetus to exploit natural diversity, and development of robust statistical analysis methods make association mapping appealing and affordable to plant research programs. Association mapping identifies quantitative trait loci (QTLs) by examining the marker-trait associations that can be attributed to the strength of linkage disequilibrium between markers and functional polymorphisms across a set of diverse germplasm. General understanding of association mapping has increased significantly since its debut in plants. We have seen a more concerted effort in assembling various association-mapping populations and initiating experiments through either candidate-gene or genome-wide approaches in different plant species. In this review, we describe the current status of association mapping in plants and outline opportunities and challenges in complex trait dissection and genomics-assisted crop improvement.

LARGE-SCALE GENOME-WIDE association analyses of major human diseases have yielded very promising results, corroborating findings of previous candidate-gene association studies and identifying novel disease loci that were previously unknown (The Wellcome Trust Case Control Consortium, 2007). The same strategy is being exploited in many plant species thanks to the dramatic reduction in costs of genomic technologies. In contrast to the widely used linkage analysis traditional mapping research in plants, association mapping searches for functional variation in a much broader germplasm context. Association mapping enables researchers to use modern genomic technologies to exploit natural diversity, the wealth of which is known to plant geneticists and breeders but has been utilized only on a small scale before the genomics era. Owing to the ease of producing large numbers of progenies from controlled crosses and conducting replicated trials with immortal individuals (inbreds and recombinant inbred lines, RILs), association mapping in plants may prove to be more promising than in human or animal genetics. In the current review,

C. Zhu and J. Yu, Dep. of Agronomy, Kansas State University, 2004 Throckmorton Hall, Manhattan, KS 66506; M. Gore, Dep. of Plant Breeding and Genetics, Cornell University, Ithaca, NY 14853; Edward S. Buckler, USDA-ARS and Institute for Genomic Diversity, Cornell University, Ithaca, NY 14853. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the USDA. Received 11 Feb. 2008. *Corresponding author (jyu@ksu.edu).

**Abbreviations:** AB-QTL, advanced backcross QTL;AFLP, amplified fragment length polymorphism; GC, genomic control; IL, introgression library; K, kinship matrix; lcyE, lycopene epsilon cyclase; LD, linkage disequilibrium; NAM, nested association mapping; oligo, oligonucleotide; PCA, principal component analysis; Q, population structure; QTDT, quantitative transmission disequilibrium test; QTLs, quantitative trait loci; RAPD, random amplified polymorphic DNA; RILs, recombinant inbred lines; SA, structured association; SBE, single base extension; SFP, single feature polymorphism; SNPs, single nucleotide polymorphisms; SSRs, simple sequence repeats.

Published in The Plant Genome 1:5–20. Published 16 July 2008.
doi: 10.3835/plantgenome2008.02.0089

677 S. Segoe Rd., Madison, WI 53711 USA
An open-access publication

we focus on presenting association mapping as a new strategy for genetic dissection of complex traits, steps to initiate an association mapping study, and common methods in genotyping, phenotyping, and data analysis. Interested readers may also refer to previous reviews on other technical aspects such as linkage disequilibrium, population structure, and statistical analysis (Ersoz et al., 2008; Flint-Garcia et al., 2003; Yu and Buckler, 2006).

## WHY ASSOCIATION MAPPING?

### New Tool

The phenotypic variation of many complex traits of agricultural or evolutionary importance is influenced by multiple quantitative trait loci (QTLs), their interaction, the environment, and the interaction between QTL and environment. Linkage analysis and association mapping are the two most commonly used tools for dissecting complex traits (Fig. 1). Linkage analysis in plants typically localizes QTLs to 10 to 20 cM intervals because of the limited number of recombination events that occur during the construction of mapping populations and the cost for propagating and evaluating a large number of lines (Doerge, 2002; Holland, 2007). While hundreds of linkage analysis studies have been conducted in various plant species over the past two decades (Holland, 2007; Kearsey and Farquhar, 1998), only a limited number of identified QTLs were cloned or tagged at the gene level (Price, 2006). Association mapping, also known as linkage disequilibrium (LD) mapping, has emerged as a tool to resolve complex trait variation down to the sequence level by exploiting historical and evolutionary recombination events at the population level (Nordborg and Tavare, 2002; Risch and Merikangas, 1996). As a new alternative to traditional linkage analysis, association mapping offers three advantages, (i) increased mapping resolution, (ii) reduced research time, and (iii) greater allele number (Yu and Buckler, 2006). Since its introduction to plants (Thornsberry et al., 2001), association mapping has continued to gain favorability in genetic research because of advances in high throughput genomic technologies, interests in identifying novel and superior alleles, and improvements in statistical methods (Fig. 2).

Based on the scale and focus of a particular study, association mapping generally falls into two broad categories (Fig. 3), (i) candidate-gene association mapping, which relates polymorphisms in selected candidate genes that have purported roles in controlling phenotypic variation for specific traits; and (ii) genome-wide association mapping, or genome scan, which surveys genetic variation in the whole genome to find signals of association for various complex traits (Risch and Merikangas, 1996). While researchers interested in a specific trait or a suite of traits often exploit candidate-gene association mapping, a large consortium of researchers might choose to conduct comprehensive genome-wide analyses of various

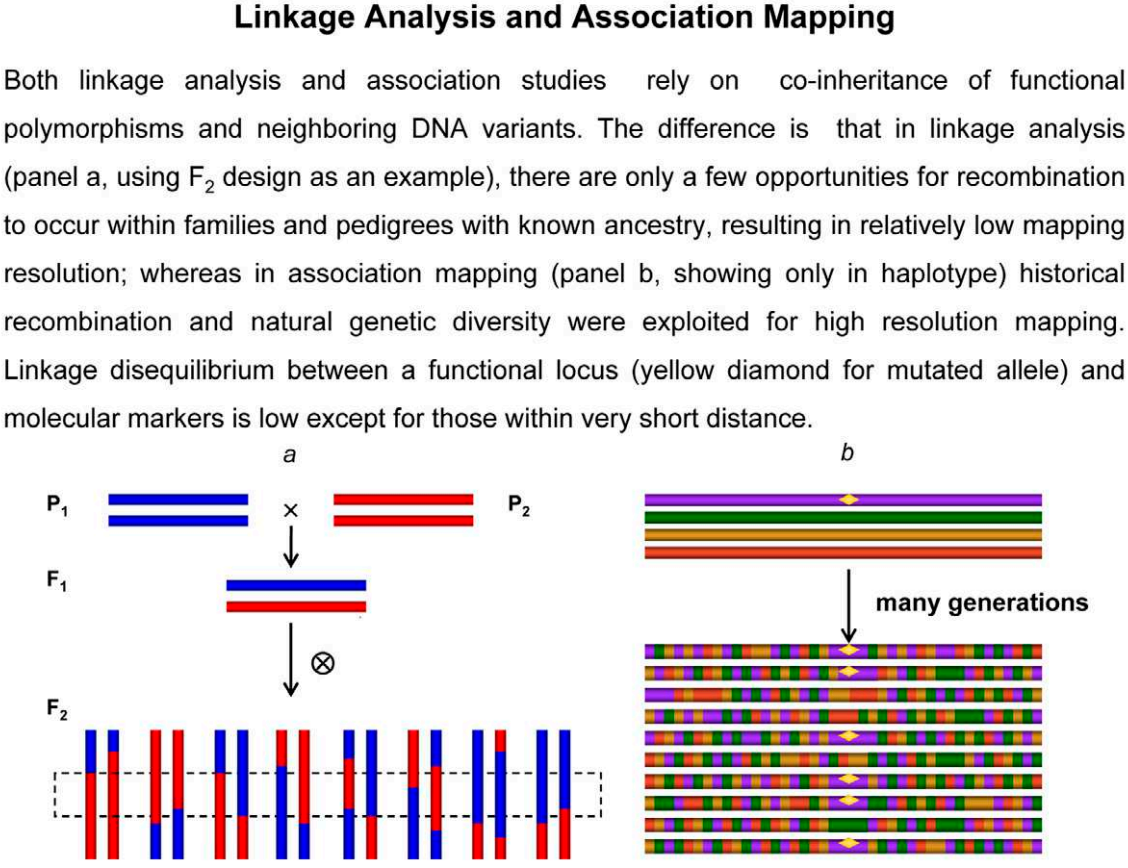



Figure 1. Schematic comparison of linkage analysis with designed mapping populations and association mapping with diverse collections.

traits by testing hundreds of thousands of molecular markers distributed across the genome for association.

## Genomic Technology

Advances in high-throughput genotyping and sequencing technologies have markedly reduced the cost per data point of molecular markers, particularly single nucleotide polymorphisms (SNPs) (Hirschhorn and Daly, 2005; Syvanen, 2005). For candidate-gene association mapping, information regarding the location and function of genes involved in either biochemical or regulatory pathways that lead to final trait variation often is required. Fortunately, due to the availability of annotated genome sequences from several model species and the general application of genomic technology (e.g., sequencing, genotyping, gene expression profiling, comparative genomics, bioinformatics, linkage analysis, mutagenesis, and biochemistry), a whole host of candidate gene sequences for various complex traits is now available for further association analysis. On the other hand, as it becomes affordable to identify hundreds of thousands of SNPs through resequencing a core set of diverse lines and genotype these SNPs across a larger number of samples, researchers are moving toward genome-wide association analyses of complex traits. For example, the *Arabidopsis* HapMap provided a powerful catalog of genetic diversity with more than 1 million SNPs (i.e., on average one SNP every 166 bp) (Clark et al., 2007), a rate about 11-fold higher than that of human populations (Hinds et al., 2005).

Not too long ago, our capacity to conduct even a thorough linkage analysis study with a few hundred molecular markers was limited by the cost of genotyping. Now, a new question faced by many researchers is "How can I take advantage of the high-throughput genomic technologies?" Obviously, association mapping is one approach that heavily leverages these emerging genomic technologies, with sequencing, resequencing, and genotyping as the intermediate steps to the final goal of linking functional polymorphisms to complex trait variation.

## Natural Diversity

Association mapping harnesses the genetic diversity of natural populations to potentially resolve complex trait variation to single genes or individual nucleotides. Conventional linkage analysis with experimental populations derived from a bi-parental cross provides pertinent information about traits that tends to be specific to the

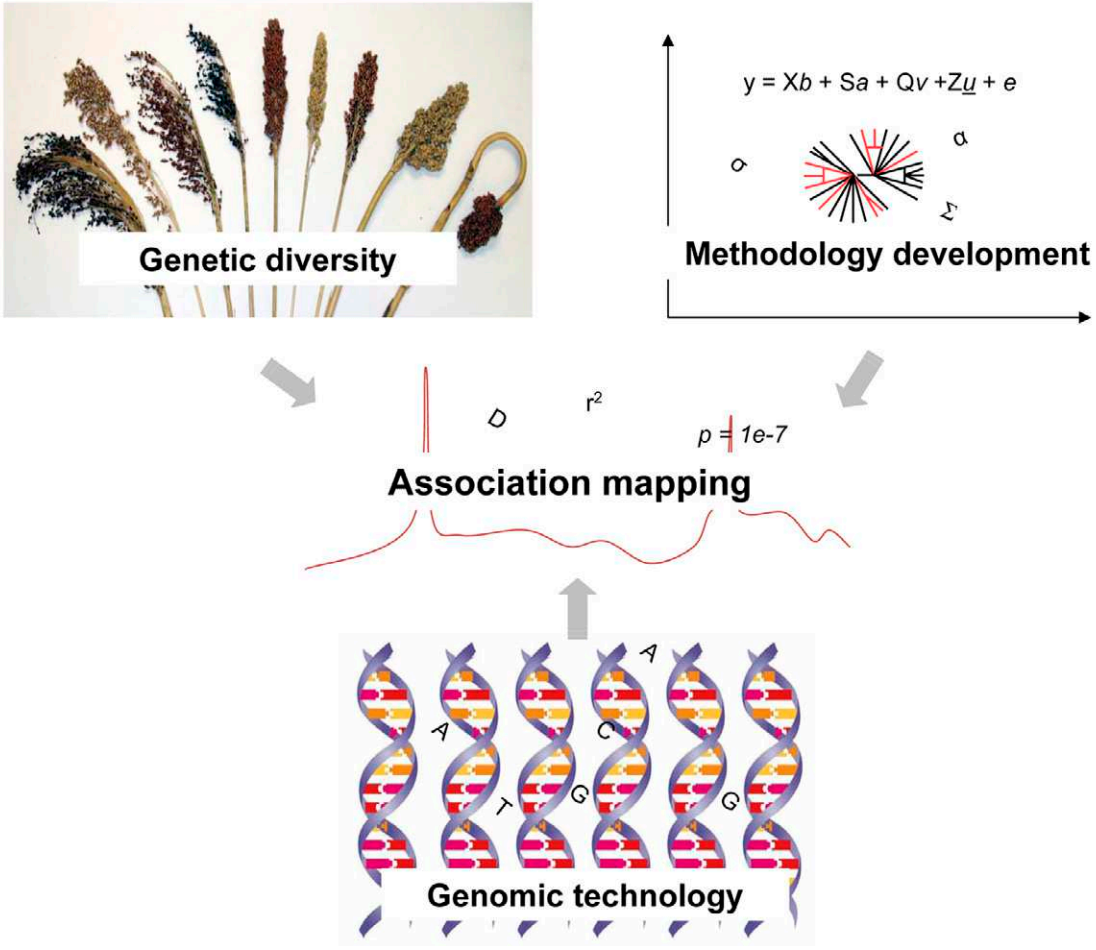
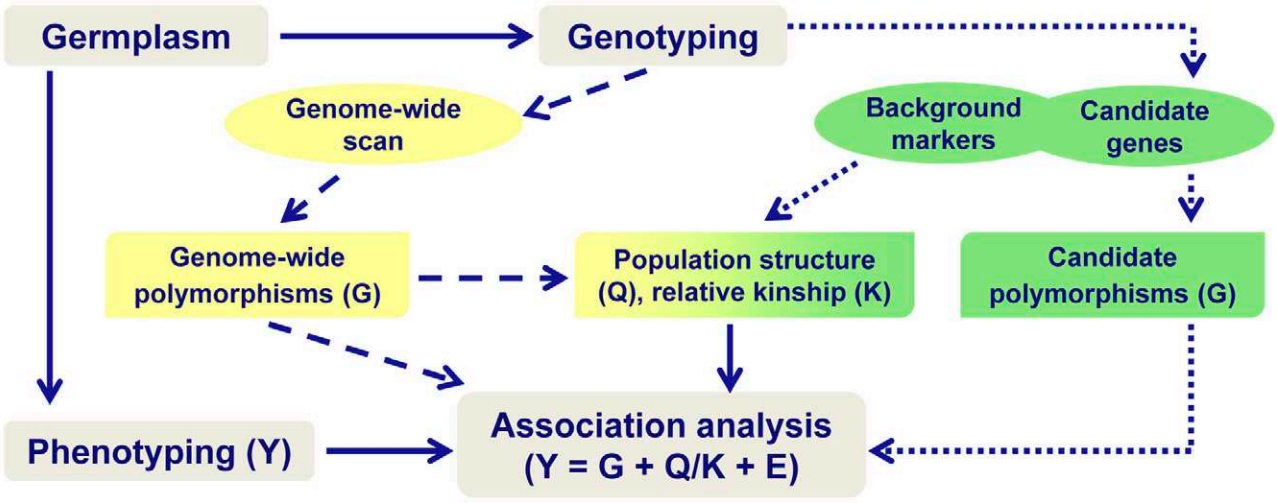



Figure 2. Main driving forces of the current interest in association mapping. Genomic technologies for high-throughput genome sequencing and genotyping made it more affordable to obtain a large amount of marker data across a large diversity panel for complex trait dissection and superior allele mining. Methodology development alleviated the issue of false positives due to population structure.




| Genome-wide association mapping | Candidate-gene association mapping |
|---|---|
| It is a comprehensive approach to systematically search the genome for causal genetic variation. A large number of markers are tested for association with various complex traits, and prior information regarding candidate genes is not required. It works best for a research consortium with complementary expertise and adequate funding. | Candidate genes are selected based on prior knowledge from mutational analysis, biochemical pathway, or linkage analysis of the trait of interest. An independent set of random markers needs to be scored to infer genetic relationships. It is a low cost, hypothesis-driven, and trait-specific approach but will miss other unknown loci. |

Figure 3. Schematic diagram and contrast of genome-wide association mapping and candidate-gene association mapping. The inclusion of population structure (Q), relative kinship (K), or both in final association analysis depends on the genetic relationship of the association mapping panel and the divergence of the trait examined. E stands for residual variance.

same or genetically related populations, while results from association mapping are more applicable to a much wider germplasm base. The ability to map QTLs in collections of breeding lines, landraces, or samples from natural populations has great potential for future trait improvement and germplasm security. With regard to exploring natural diversity, advanced backcross QTL (AB-QTL) and introgression library (IL) are well-known strategies for mining alleles from exotic germplasm to improve the productivity, adaptation, quality, and nutritional value of crops (Tanksley and McCouch, 1997; Zamir, 2001). Association mapping is complementary to AB-QTL and IL in that it is an additional tool for evaluating extant functional diversity in each crop species on a much larger scale (Breseghello and Sorrells, 2006a; Flint-Garcia et al., 2003).

## Methodology Development

Conventional linkage mapping in plant species, including single marker analysis, interval mapping, multiple interval mapping, and Bayesian interval mapping, is well developed and validated (Doerge, 2002; Zeng, 2005). In contrast, little effort has been made to develop robust methods of association mapping in plant species. False positives generated by population structure have long been regarded as a hurdle to association mapping and it has been difficult to replicate significant results in independent studies and follow up on detected signals with costly molecular and biochemical analyses. Given the geographical origins, local adaptation, and breeding history of assembled genotypes in an association mapping panel, these non-independent samples usually contain both population structure and familial relatedness (Yu and Buckler, 2006). Recently, several statistical methods have been proposed to account for population structure and familial relatedness, structured association (SA) (Falush et al., 2003; Pritchard and Rosenberg, 1999; Pritchard et al., 2000a), genomic control (GC) (Devlin and Roeder, 1999), mixed model approach (Yu et al., 2006), and principal component approach (Price et al., 2006). The essence of these approaches is to use genotypic information from random molecular markers across the genome to account for genetic relatedness in association tests either explicitly (e.g., SA and mixed model) or through ad hoc adjustment (e.g., GC). With these methods, the issue of false positives generated by population

structure can now be dealt with accordingly (Price et al., 2006; Yu et al., 2006; Zhao et al., 2007).

### Current Status

So far, a series of research papers focusing on LD and association mapping have been published, spanning more than a dozen plant species (Table 1). Many major crops, such as maize (*Zea mays,* L.), soybean (*Glycine max* (L.) Merr.), barley (*Hordeum vulgare* L.), wheat (*Triticum aestivum* L.), tomato (*Lycopersicon esculentum* Mill.), sorghum (*Sorghum bicolor* (L.) Moench), and potato (Solanum *tuberosum* L.), as well as tree species such as aspen (*Populus tremula* L.) and loblolly pine (*Pinus taeda* L.), have been studied. Many questions still demand further study as we attempt to gain a better grasp of the various genetic and statistical aspects of association mapping. For example, should one choose a highly pedigreed group of individuals from breeding programs or a diverse collection of germplasm bank accessions? Does one need to be concerned about false positives due to population structure? What is the appropriate analysis method? Should one start a candidate-gene or genome-wide association analysis? Are cryptic genetic relationships adequately estimated by random markers? We offer our opinions on some of these questions in the following sections.

## HOW TO INITIATE ASSOCIATION MAPPING?

### Species and Germplasm

Before initiating association mapping, researchers should carefully consider all genetic aspects of the species and the associated germplasm available. The ploidy level of individuals from a plant species whose genetics are not well characterized should be evaluated, particularly if the assembled population contains wild accessions obtained from a germplasm bank. This helps to avoid the difficulty in differentiating the effects of functional polymorphisms from that of allele dosage. Because the task of assembling and studying an association mapping population requires a long-term commitment, it is worthwhile to examine various genetic tools available for a given species. Are there groups of scientists who have been conducting genetics, physiological, or biochemical studies within the species? What are the available molecular markers that have been

**Table 1. Examples of association mapping studies in various plant species.**

| Plant species | Populations | Sample size | Background markers | Traits | References |
|---|---|---|---|---|---|
| Maize | Diverse inbred lines | 92 | 141 | Flowering time | (Thornsberry et al., 2001) |
| | Elite inbred lines | 71 | 55 | Flowering time | (Andersen et al., 2005) |
| | Diverse inbred lines and landraces | 375 + 275 | 55 | Flowering time | (Camus-Kulandaivelu et al., 2006) |
| | Diverse inbred lines | 95 | 192 | Flowering time | (Salvi, 2007) |
| | Diverse inbred lines | 102 | 47 | Kernel composition<br>Starch pasting properties | (Wilson et al., 2004) |
| | Diverse inbred lines | 86 | 141 | Maysin synthesis | (Szalma et al., 2005) |
| | Elite inbred lines | 75 | | Kernel color | (Palaisa et al., 2004) |
| | Diverse inbred lines | 57 | | Sweet taste | (Tracy et al., 2006) |
| | Elite inbred lines | 553 | 8950 | Oleic acid content | (Belo et al., 2008) |
| | Diverse inbred lines | 282 | 553 | Carotenoid content | (Harjes et al., 2008) |
| *Arabidopsis* | Diverse ecotypes | 95 | 104 | Flowering time | (Olsen et al., 2004) |
| | Diverse ecotypes | 95 | 2553 | Disease resistance<br>Flowering time | (Aranzana et al., 2005)<br>(Zhao et al., 2007) |
| | Diverse accessions | 96 | | Shoot branching | (Ehrenreich et al., 2007) |
| Sorghum | Diverse inbred lines | 377 | 47 | Community resource report | (Casa et al., 2008) |
| Wheat | Diverse cultivars | 95 | 95 | Kernel size, milling quality | (Breseghello and Sorrells, 2006b) |
| Barley | Diverse cultivars | 148 | 139 | Days to heading, leaf rust, yellow dwarf virus, rachilla hair length, lodicule size | (Kraakman et al., 2006) |
| Potato | Diverse cultivars | 123 | 49 | Late bright resistance | (Malosetti et al., 2007) |
| Rice | Diverse land races | 105 | | Glutinous phenotype | (Olsen and Purugganan, 2002) |
| | Diverse land races | 577 | 577 | Starch quality | (Bao et al., 2006) |
| | Diverse accessions | 103 | 123 | Yield and its components | (Agrama et al., 2007) |
| *Pinus taeda* | Unstructured natural population | 32 | 21 | Wood specific gravity, late wood | (Gonzalez-Martinez et al., 2006) |
| | Lines | 435 | 288 | Microfibril angle, cellulose content | (Gonzalez-Martinez et al., 2007) |
| Sugarcane | Diverse clones | 154 | 2209 | Disease resistance | (Wei et al., 2006) |
| Eucalyptus | Unstructured natural population | 290 | 35 | Microfibril angle | (Thumma and Nolan, 2005) |
| Perennial ryegrass | Diverse natural germplasms | 26 | 589 | Heading date | (Skøt et al., 2005) |
| | Diverse natural germplasms | 96 | 506 | Flowing time, water soluble carbohydrate | (Skøt et al., 2007) |

developed for this species? What is the current status of linkage analysis for the targeted traits?

Choice of germplasm is critical to the success of association analysis (Breseghello and Sorrells, 2006a; Flint-Garcia et al., 2003; Yu et al., 2006). Genetic diversity, extent of genome-wide LD, and relatedness within the population determine the mapping resolution, marker density, statistical methods, and mapping power. Generally, plant populations amenable for association studies can be classifiable into one of five groups (Yu and Buckler, 2006; Yu et al., 2006), (i) ideal sample with subtle population structure and familial relatedness, (ii) multi-family sample, (iii) sample with population structure, (iv) sample with both population structure and familial relationships, and (v) sample with severe population structure and familial relationships. Due to local adaptation, selection, and breeding history in many plant species, many populations for association mapping would fall into category four. Alternatively, we can classify populations according to the source of materials, germplasm bank collections, synthetic populations, and elite germplasm (Breseghello and Sorrells, 2006a).

## Linkage Disequilibrium

Linkage disequilibrium, or gametic phase disequilibrium, measures the degree of non-random association between alleles at different loci. The difference between observed haplotype frequency and expected based on allele frequencies is defined as $D$.

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

where $p_{AB}$ is the frequency of gamete AB; $p_A$ and $p_B$ are the frequency of the allele A and B, respectively. In absence of other forces, recombination through random mating breaks down the LD with $D_t = D_0(1 - r)^t$, where $D_t$ is the remaining LD between two loci after $t$ generations of random mating from the original $D_0$. Several statistics have been proposed for LD, and these measurements largely differ in how they are affected by marginal allele frequencies and small sample sizes (Hedrick, 1987). Both $D'$ (Lewontin, 1964) and $r^2$ (Hill and Robertson, 1968) have been widely used to quantify LD. For two bi-allelic loci, $D'$ and $r^2$ have the following formula:

$$D' = \frac{|D|}{D_{\max}}$$

$$where \quad D_{\max} = \min(p_A p_b, p_a p_B) \; if \; D > 0;$$

$$D_{\max} = \min(p_A p_B, p_a p_b) \; if \; D < 0$$

$$r^2 = \frac{D^2}{p_A p_a p_B p_b}$$

One undesirable feature of $D$ is that its range is determined by the allele frequency. For this reason the $D'$ statistic was developed to partially normalize the $D$ value with respect to the maximum value possible for the allele frequencies and has a range between 0 and 1. The $r^2$ statistic is the same as the squared value of the Pearson's (product moment) correlation coefficient and has an expectation of $1/(1+4Nc)$, where $N$ is the effective population size and $c$ is the recombination rate in morgan (Hill and Robertson, 1968).

In terms of identifying SNPs or haplotypes significantly associated with phenotypic trait variation, $r^2$ is the most relevant LD measurement. Typically, $r^2$ values of 0.1 or 0.2 are often used to describe the LD decay. If a true functional polymorphism contributes a fraction of the total trait variation, $h^2_q$, and has a LD value of $r^2$ with another SNP, then the trait variation that can be explained by this SNP will be $r^2 \times h^2_q$. A similar inference cannot be made using $D$ or $D'$. An empirical example was recently reported, in which the significance level of association between the phenotype and SNPs followed the $r^2$ plot of the most likely functional SNP and other adjacent SNPs, but not the $D'$ plot (Ducrocq et al., 2008).

Though LD is affected by many factors (Ardlie et al., 2002), LD due to linkage is the net result of all the recombination events that occurred in a population since the origin of an allele by mutation, providing a greater opportunity for recombination to take place between any two closely linked loci than what is in linkage analysis (Holte et al., 1997; Karayiorgou et al., 1999). Among other factors, the reproduction mode of a species partly determines the level of LD in a diverse population (Flint-Garcia et al., 2003). Generally, LD extends to a much longer distance in self-pollinated crops, such as wheat, than in cross-pollinated species, such as maize, and LD generated by population structure within the sample needs to be accounted for in the analysis to avoid spurious results. Detailed reviews on LD in plant species have been given previously (Ersoz et al., 2008; Flint-Garcia et al., 2003). Genome-wide LD determines the mapping resolution and marker density for a genome scan. If LD decays within a short distance, mapping resolution is expected to be high, but a large number of markers are required. On the other hand, if LD extends a long distance, sometimes in cM, then mapping resolution will be low, but a relatively small number of markers are required. A graphical view of LD can be presented either as a LD decay plot of $D'$ or $r^2$ over physical or genetic distance or as in a linear arrangement of LD between polymorphic sites within a gene or loci along a chromosome (Bradbury et al., 2007; Flint-Garcia et al., 2003).

## Community Resources

As sequencing and genotyping costs continue to decrease, we expect to see more genome-wide association mapping studies in plants than in animals because of the relatively low cost of creating and maintaining inbred lines, shared seed, and evaluation in multiple environments. In several plant species, diverse germplasm panels are being established for whole-genome association analysis (Caldwell et al., 2006; Hamblin et al., 2006; Nordborg et al., 2005; Yu and Buckler, 2006). In addition to a diversity panel of 300 maize inbred lines (Flint-Garcia et al., 2005), a large-scale maize QTL mapping population comprised of 5000 RILs derived from the crosses

of a common parent with each of 25 diverse founders is available (www.panzea.org; verified 27 May 2008) (Yu et al., 2008). This common platform will enable researchers to efficiently exploit numerous genetic, genomic, and systems biology tools. In sorghum, a diversity panel of 377 inbred lines was assembled for association mapping (Casa et al., 2008). All major cultivated races (i.e., tropical lines from diverse geographic and climatic regions) in sorghum and important U.S. sorghum breeding lines and their progenitors were included. The Barley Coordinated Agricultural Project (BarleyCAP) was initiated to genotype approximately 3000 SNPs across 3840 lines contributed from 10 barley breeding programs, including progenies of pedigree programs and a collection of diverse barley genotypes (Muehlbauer, 2006). This project involves multiple institutions and multi-disciplinary cooperation. In wheat, four regional association mapping populations are being assembled to accommodate both winter and spring types and grain hardness (Mark Sorrells, personal communication, 2008). This effort is in addition to the existing soft winter wheat panel (Breseghello and Sorrells, 2006b). Community germplasm resources not only allow researchers to integrate studies of mutual interests but also allow a deeper understanding and dissection of complex traits. Therefore, community efforts should be emphasized more while conducting association analysis.

## GENOTYPING FOR ASSOCIATION MAPPING

### Background Markers

In association studies, a set of unlinked, selectively neutral background markers scaled to achieve genome-wide coverage are employed to broadly characterize the genetic composition of individuals. Background genetic markers are useful in assigning individuals to populations (Pritchard and Rosenberg, 1999), preventing spurious associations if population structure and relatedness exist (Pritchard et al., 2000b; Thornsberry et al., 2001; Yu et al., 2006), and estimating kinship and inbreeding (Lynch and Ritland, 1999). Random amplified polymorphic DNA (RAPD) (Williams et al., 1990) and amplified fragment length polymorphism (AFLP) (Vos et al., 1995) markers can serve as background markers, but almost all RAPD and AFLP markers are dominantly inherited and thus demand special statistical methods if used to estimate population genetic parameters (Falush et al., 2007; Ritland, 2005). Conversely, codominant microsatellites, or simple sequence repeats (SSRs), and SNPs are more revealing (i.e., no allelic ambiguity) than their dominant counterparts and, therefore, are more powerful in estimating population structure (Q) and the relative kinship matrix (K).

Because SSR markers are multiallelic, reproducible, PCR-based, and generally selectively neutral they have been the predominant molecular marker in kinship and population studies. Semi-automated systems exist for the multiplexed detection and sizing of fluorescent-labeled SSR products with internal size standards; thus greatly increasing both the allele size accuracy and genotyping throughput (Mitchell et al., 1997). Nascent polymorphic SSR alleles are mostly spawned from the slipped strand mispairing (i.e., slippage) of allelic tandem repeats during DNA replication (Levinson and Gutman, 1987). In theory, the highly mutagenic process of slippage can generate an unlimited number of SSR alleles, but longer SSR allele sizes are more likely to be eliminated by natural selection (Li et al., 2002). The same slippage phenomenon that results in highly polymorphic SSR loci also is the basis of size homoplasy, a situation when SSR alleles are identical in size but not identical by descent (Viard et al., 1998). If alleles have a high mutation rate and strong size constraint, SSR size homoplasy could be problematic when estimating genetic parameters in a large population (Estoup et al., 2002).

Due to higher genome density, lower mutation rate, and better amenability to high-throughput detection systems, SNPs are rapidly becoming the marker of choice for complex trait dissection studies. Either single marker assays or multiplexes in scalable assay plates and microarray formats can be used to score SNPs. The selection of a specific genotyping technology is dependent on both the number of SNP markers and individuals to be scored (Kwok, 2000; Syvanen, 2005). The mutation rate per site per generation is several times lower than the SSR mutational rate per generation (Li et al., 2002; Vigouroux et al., 2002). Therefore, on a per-site basis, due to SNPs' predominantly biallelic nature they are less informative than multiallelic SSRs. Because the expected heterozygosity of individual SNPs is lower, more SNP than SSR background markers are needed to reach a reasonable estimate of population structure and relatedness for most crops. This should not be considered a shortcoming because SNPs are more widely distributed throughout the genome and are several-fold less expensive to score than SSRs.

### Candidate Genes

Candidate-gene association mapping is a hypothesis-driven approach to complex trait dissection, with biologically relevant candidates selected and ranked based on the evaluation of available results from genetic, biochemical, or physiology studies in model and non-model plant species (Mackay, 2001; Risch and Merikangas, 1996). Because SNPs offer the highest resolution for mapping QTL and are potentially in LD with the causative polymorphism they are the preferential candidate-gene variant to genotype in association studies (Rafalski, 2002). Candidate-gene association mapping requires the identification of SNPs between lines and within specific genes. Therefore, the most straightforward method of identifying candidate gene SNPs relies on the resequencing of amplicons from several genetically distinct individuals of a larger association population. Fewer diverse individuals in the SNP discovery panel are needed to identify common SNPs, whereas many more are needed to identify rarer SNPs. Promoter, intron, exon, and

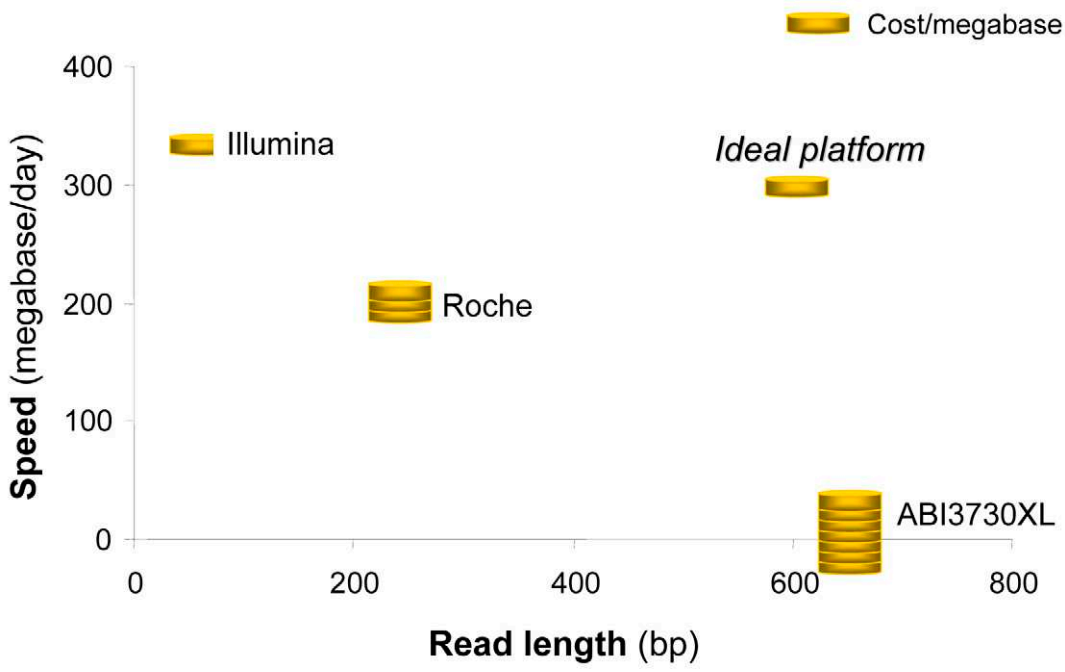



Figure 4. Comparison of sequencing platforms for high-throughput SNP discovery. Adapted from (Salisbury, 2007). Comparison is based on performance of Illumina/Solexa's Genetic Analyzer, Roche/454's GS FLX, and Applied Biosystems' ABI3730XL.

5′/3′-untranslated regions are all reasonable targets for identifying candidate gene SNPs, with non-coding regions expected to have higher levels of nucleotide diversity than coding regions. The rate of LD decay for a specific candidate gene locus dictates the number of SNPs per unit length (e.g., kb) needed to identify significant associations (Whitt and Buckler, 2003). Therefore, the number and base-pair length of amplicons required to sufficiently sample a candidate gene locus is almost entirely dependent on LD and SNP distribution, with a higher density of SNP markers needed in regions of relatively low LD and high nucleotide diversity.

It is not essential to score every candidate gene SNP. Because a key objective of this approach is to identify SNPs that are causal of phenotypic variation, those with a higher likelihood to alter protein function (coding SNPs) or gene expression (regulatory SNPs) should be a top priority for genotyping (Tabor et al., 2002). However, the biological function of SNPs, if any, for the most part is unknown or not easily discerned. In cases of ambiguity where there are blocks of several SNPs in significant LD, an alternative strategy is to select and score a small fraction of SNPs (tag SNPs) that capture most of the haplotype block structure in candidate-gene regions (Johnson et al., 2001). Genotyping tag SNPs is more cost effective and, if properly designed, does not result in a significant loss of statistical testing power (Kui et al., 2002). In most cases, allele resequencing in diploid inbred lines (homozygous loci) allows for the direct determination of haplotypes. Reconstructing haplotypes from SNP data in heterozygous and polyploid (ancient or modern) individuals is more challenging, as statistical algorithms are needed to resolve phase ambiguities (Simko, 2004; Stephens et al., 2001) and transmission tests are needed to confirm orthologous relationships (Cogan et al., 2007).

Candidate-gene selection is straightforward for relatively simple biochemical pathways (e.g., starch synthesis in maize) or well characterized pathways (e.g., flowering time in *Arabidopsis*) that have been resolved mainly through genetic analysis of mutant loci (natural or induced). But for complex traits such as grain or biomass yield, the entire genome could potentially serve as a candidate (Yu and Buckler, 2006). Most candidate-gene studies investigating a single pathway or trait in a crop species have genotyped less than 100 SNPs in a population of 100 to 400 individuals (Table 1) (Ersoz et al., 2008). In these studies, Sanger sequencing and single base extension (SBE) assays were the predominant technologies used to score candidate gene SNPs. Advantages of SBE assays over Sanger sequencing are reflected in their lower reagent costs, enhanced resolution of heterozygous genotypes, and better suitability to multiplex detection on higher-throughput, lower cost analytical platforms (Syvanen, 2001).

## Whole-Genome Scan

If whole-genome association scans are to be conducted in crops, an important first step is to use high-capacity DNA sequencing instruments or high-density oligonucleotide (oligo) arrays to efficiently identify SNPs at a density that accurately reflects genome-wide LD structure and haplotype diversity. The appropriateness of a DNA sequencing platform (Fig. 4) for SNP discovery depends on the number of SNPs required for effective whole-genome scans in an association population. For

example, the extensive LD in 95 *Arabidopsis* accessions and 102 elite barley inbred lines made it possible to association test a low number of evenly spaced SNPs discovered via capillary-based Sanger sequencing and still achieve a medium level of genome-wide mapping resolution (Aranzana et al., 2005; Rostoks et al., 2006). Alternatively, tens to hundreds of thousands of SNP markers are required for powerful whole-genome scans in crops with low LD and high haplotype diversity, such as maize and sunflower. In such a scenario, the 454-GS FLX (Margulies et al., 2005) and Illumina 1 G Genome Analyzer (www.illumina.com; verified 28 May 2008) are ideal platforms for identifying scores of SNPs through short read resequencing of allelic fragments from several genetically diverse individuals. After SNPs are identified, different array-based platforms can be used to genotype thousands of tag SNPs in parallel.

A high quality whole-genome reference sequence is extremely valuable in construction of a SNP haplotype map from short reads produced by the 454 and Illumina sequencing platforms. This is because short reads are more easily assembled by aligning to a preexisting genome reference sequence compared to de novo assembly. Also, a reference genome is useful in masking repetitive and paralogous sequences, as the orthology of high copy sequences is difficult to determine unless candidate SNPs are genetically mapped. Because the base calling accuracy of 454 and Illumina is presently lower than that of Sanger sequencing, emphasis should be placed on calling SNPs that have multiple read support ( ≥2× coverage/allele/individual). The newness and expense of next-generation sequencing technologies have limited their wide-spread implementation for SNP discovery in crops. Recently, a 454-based transcriptome sequencing method was used in maize to identify more than 36,000 candidate SNPs between two maize inbred lines (Barbazuk et al., 2007). This 454-SNP study is a promising step toward development of numerous genome-wide SNP markers in a highly diverse crop species with a rapid breakdown of LD, but more importantly lays the framework for identifying SNPs based on sequencing of random genomic fragments.

The simultaneous discovery and genotyping of allelic variation with high-density oligo expression arrays designed from a reference sequence is based on the concept that a perfectly matched target binds to a 25-bp oligo feature with greater affinity than a mismatched target (Borevitz et al., 2003; Winzeler et al., 1998). If an individual feature on an array shows a significant and repeatable difference in hybridization intensity between genotypes, it can serve directly as a polymorphic marker or single feature polymorphism (SFP). Expression arrays hybridized with total genomic DNA allow for highly accurate scoring of several thousand SFPs in the relatively small genomes of ~135-Mb *Arabidopsis* (Borevitz et al., 2003) and ~430-Mb rice (Kumar et al., 2007). Whole-genome, genome complexity reduction, and gene enrichment target preparation methods are only modestly successful for detecting SFPs in larger retrotransposon-rich plant genomes (Gore et al., 2007; Rostoks et al., 2005). Notable limitations are that SFPs tend to be less heritable (i.e., lower quality) than SNPs and map unknown polymorphisms only at 25-bp resolution. If scored at very high density and moderate accuracy, SFPs are potentially powerful tools to detect associations in crop genomes with extensive LD (Kim et al., 2006) and relatively low levels of repetitive DNA.

In a whole-genome resequencing-by-hybridization approach championed by Perlegen Sciences (Mountain View, CA), high-density arrays consisting of tiled, overlapping 25-bp oligos are used to identify SNPs and other polymorphisms in a hybridized target genome at single base pair resolution (Borevitz and Ecker, 2004; Mockler et al., 2005). Tiling arrays were used to construct a haplotype map by essentially resequencing 20 diverse *Arabidopsis* genomes and cataloging more than 1 million nonredundant SNPs (Clark et al., 2007). Only 27% of the total polymorphisms were scored in a given ecotype due to ineffective SNP detection in highly polymorphic regions. Tiling array projects are in progress to identify SNPs in multiple rice lines (McNally et al., 2006) and score 250,000 tag SNPs in an association panel of 1000 *Arabidopsis* ecotypes. It is still an open question as to whether resequencing-by-hybridization on tiling arrays will come to fruition as a routine SNP discovery platform for crop genomes that predominantly contain repetitive DNA, extensive sequence duplications, or high nucleotide diversity.

## PHENOTYPING FOR ASSOCIATION MAPPING

### Field Design

The importance of phenotyping has not received as much attention as genotyping. While accuracy and throughput of genotyping have dramatically improved, obtaining robust phenotypic data remains a hurdle for large-scale association mapping projects. Because association mapping often involves a relatively large number of diverse accessions, phenotypic data collection with adequate replications across multiple years and multiple locations is challenging. Efficient field design with incomplete block design (e.g., $\alpha$-lattice), appropriate statistical methods (e.g., nearest neighbor analysis and spatial models), and consideration of QTL × environmental interaction should be explored to increase the mapping power, particularly if the field conditions are not homogenous (Eskridge, 2003). This type of study is challenging because direct empirical proof of the importance of field design requires comprehensive studies with different levels of homogeneity in field conditions, as well as strong collaborations between geneticists and statisticians (Kent Eskridge, personal communication, 2007). The increase in power of detecting QTLs with repeated measurements is well known and also has been demonstrated by simulation studies in mapping with pedigree-based breeding germplasm (Arbelbide et al., 2006; Yu et al.,

2005). Nevertheless, the importance of phenotyping has started to receive its deserved attention, as exemplified by the Symposium on Advances in Phenotyping held by the Crop Science Society of America in 2006 (http://a-c-s.confex.com/crops/2006am/techprogram/S2649.HTM; verified 28 May 2008).

Given the diverse nature of an association mapping panel, it is also critical to consider the influence of flowering time on the expression of other correlated traits. It might be worthwhile to block a field by flowering time if traits of interest are dependent on developmental transitions. Other issues that need be considered in phenotyping include photoperiod sensitivity, lodging, and susceptibility to prevalent pathogens because these traits affect the measurement of other morphological or agronomic traits at field condition.

### Data Collection

Collection of high quality phenotypic data is essential for genetic mapping research. Association mapping studies often are long-term projects, with phenotyping being conducted over years in multiple locations (Flint-Garcia et al., 2005). In this framework, any newly discovered candidate gene polymorphism can always be tested for association with existing phenotypic data. Also, transitioning from a candidate-gene to a genome-wide approach should be seamless if the original association mapping panel was constructed in a manner such that other complex traits can be evaluated and robust phenotypic data were collected along the way.

To ensure that high quality data are obtained from a wide range of conducted experiments, each researcher should assess the quality of the experiment for which they are responsible. Specific information about the experiment, such as check performance and environmental growth conditions (field or greenhouse), should be included as an annotation to the experiment in the trait database. In established programs, bar-coding systems and scanner-based data collection greatly facilitate the data collection process (www.maizegenetics.net; verified 28 May 2008).

For data storage and bioinformatics of large projects in association mapping, different models have been developed including the Genomic Diversity and Phenotype Data Model (GDPDM) schema (http://www.maizegenetics.net/gdpdm; verified 28 May 2008) used by the maize diversity group (www.panzea.org), and Germinate schema (http://bioinf.scri.ac.uk/germinate/wordpress; 28 May 2008) used by the BarleyCAP project (www.barleycap.org; verified 28 May 2008).

## STATISTICAL ANALYSIS

### Methods

The basic statistics for association analysis, under an ideal situation, would be linear regression, analysis of variance (ANOVA), *t* test or chi-square test. However, as population structure can generate spurious genotype–phenotype associations, different statistical approaches have been designed to deal with this confounding factor. For family-based samples, the transmission disequilibrium test (TDT) (Spielman et al., 1993) is used to study the genetic basis for human disease, whereas the quantitative transmission disequilibrium test (QTDT) is employed in the dissection of quantitative traits (Abecasis et al., 2000; Allison, 1997). To address the issue of population structure in population-based samples, GC and SA are the two most common methods utilized in both human and plant association studies. With GC, a set of random markers is used to estimate the degree that test statistics are inflated by population structure, assuming such structure has a similar effect on all loci (Devlin and Roeder, 1999). By contrast, SA analysis first uses a set of random markers to estimate population structure (Q) and then incorporates this estimate into further statistical analysis (Falush et al., 2003; Pritchard and Rosenberg, 1999; Pritchard et al., 2000a). Modification of SA with logistic regression has been used in previous association studies (Thornsberry et al., 2001; Wilson et al., 2004), and a general linear model version of this method is implemented in the software TASSEL (Bradbury et al., 2007).

A unified mixed-model approach for association mapping that accounts for multiple levels of relatedness was recently developed (Yu et al., 2006). In this method, random markers are used to estimate Q and a relative kinship matrix (K), which are then fit into a mixed-model framework to test for marker-trait associations. As this mixed-model approach crosses the boundary between family-based and population-based samples, it provides a powerful complement to currently available methods for association mapping (Zhao et al., 2007).

Principal component analysis (PCA) has long been used in genetic diversity analysis and was recently proposed as a fast and effective way to diagnose population structure (Patterson et al., 2007; Price et al., 2006). The PCA analysis summarizes variation observed across all markers into a smaller number of underlying component variables. These principle components could be interpreted as relating to separate, unobserved subpopulations from which the individuals in the dataset (or their ancestors) originated. The loadings of each individual on each principal component describe the population membership or the ancestry of each individual. Replacing Q with PCA in the mixed model shows some promise (Weber et al., 2008; Zhao et al., 2007), but additional research is required to establish its suitability for crop species.

### Sample Size and Number of Background Markers

Sample size for association mapping remains relatively small. In many recent association mapping studies, only about 100 lines were investigated (Table 1). To explain this in the context of genetic variation of a population, we compare the sample size of linkage analysis and association mapping. The sample size for many linkage

analysis studies in plants involves about 250 individuals ($F_2$, $BC_1$, RIL, etc.) with a homogenous, bi-parental genetic background (Bernardo, 2002). The genetic variation within an association-mapping panel is usually much greater than of linkage populations. Unless the functional locus has a very large effect and tested markers are in high LD with this locus, it will be difficult to identify marker-trait associations with a small population, regardless of whether the candidate-gene or genome-scan approach is used. Our preliminary simulations with empirical maize data show that a large sample size is required to obtain high power to detect genetic effects of moderate size.

The number of background markers required to accurately estimate genetic relationships is a common issue that needs to be addressed in candidate-gene association mapping studies. The number of required markers is much higher for biallelic SNPs than for multiallelic SSRs. We argue that a good starting point for the number of needed SSR markers is about four times the chromosome number of that species, which translates to two markers per chromosome arm. Of course, length of the chromosome, diversity of the species, diversity of the particular sample, and cost and availability of different marker systems also will impact the number of background markers used in a study.

### Software

A variety of software packages are available for data analysis in association mapping (Table 2). TASSEL is the most commonly used software for association mapping in plants and is frequently updated as new methods are developed (Bradbury et al., 2007). In addition to association analysis methods (i.e., logistic regression, linear model, and mixed model), TASSEL is also used for calculation and graphical display of linkage disequilibrium statistics and browsing and importation of genotypic and phenotypic data. STRUCTURE software typically is used to estimate Q (Pritchard et al., 2000a). The Q is an $n \times p$ matrix, where $n$ is the number of individuals and $p$ is the number of defined subpopulations. SPAGeDi software is used to estimate K among individuals (Hardy and Vekemans, 2002). K is an $n \times n$ matrix with off-diagonal elements being $F_{ij}$, a marker-based estimate of probability of identity by descent. The diagonal elements of K are one for inbreds and $0.5 \times (1 + F_x)$ for noninbred individuals, where $F_x$ is the inbreeding coefficient. EINGENSTRAT software is used to estimate PCs of the marker data and correct test statistics resulting from population stratification (Price et al., 2006). Other software commonly used in human association mapping includes Merlin (Abecasis et al., 2002) and QTDT (Abecasis et al., 2000).

SAS software (SAS Institute, 1999) or R (Ihaka and Gentleman, 1996) often are used by advanced researchers with programming skills as the platform to develop various methods. ASREML (Gilmour et al., 2002) and MTDFREML (Boldman et al., 1993) are two of several software packages used in animal genetics in mixed model analysis of data from a very large number of individuals.

## PERSPECTIVES

### Sequencing and Genotyping

The advent of next-generation sequencing platforms is a challenge to the reigning dominance of modern Sanger-based capillary sequencers. Aside from the 454 GS FLX and Illumina 1G Genome Analyzer, other highly parallel sequencing platforms such as Applied Biosytems' Supported Oligonucleotide Ligation and Detection system (SOLiD) (Shendure et al., 2005) and Helicos BioSciences' HeliScope (Braslavsky et al., 2003) are poised to begin competing for market share. Use of these and forthcoming next-generation sequencers for resequencing and directed genotyping applications will eventually become commonplace as the length and accuracy of their sequence reads improve, especially since the cost per Mb will undoubtedly continue to decline (Fig. 4). Already, DNA bar coding with unique oligo tags allows highly multiplexed genotyping-by-sequencing of alleles from multiple individuals in a single 454 sequencing run (Binladen et al., 2007; Meyer et al., 2007; Parameswaran et al., 2007), and paired end read sequencing on a 454 GS-FLX has led to mapping of structural variants in the human genome (Korbel et al., 2007).

Recently, two new strategies were developed to significantly improve the efficiency of targeted gene

**Table 2. Common statistical software packages for association mapping.**

| Software package | Focus | Website | Comment |
|---|---|---|---|
| TASSEL | Association analysis | http://www.maizegenetics.net | Free, LD statistics, sequence analysis, association mapping (logistic regression, linear model, and mixed model) |
| SAS | Generic | http://www.sas.com | Commercial, standard software widely used in data analysis and methodology work |
| R | Generic | http://www.r-project.org/ | Free, convenient for simulation work for researches with good programming and statistics background |
| STRUCTURE | Population structure | http://pritch.bsd.uchicago.edu/structure.html | Free, widely used for population structure analysis |
| SPAGeDi | Relative kinship | http://www.ulb.ac.be/sciences/ecoevol/spagedi.html | Free, genetic relationship analysis |
| EINGENSTRAT | PCA, association analysis | http://genepath.med.harvard.edu/~reich/Software.htm | Free, PCA was proposed as an alternative for population structure analysis |
| MTDFREML | Mixed model | http://aipl.arsusda.gov/curtvt/mtdfreml.html | Free, mixed model analysis for animal breeding data, also can be used for plant data |
| ASREML | Mixed model | http://www.vsni.co.uk/products/asreml | Commercial, mixed model analysis for animal breeding data, also can be used for plant data |

sequencing. The first approach combines multi-gene amplification and massively parallel sequencing (Dahl et al., 2007). In this approach, selector technology is used to amplify candidate genes in a highly multiplexed and target-specific fashion; this is followed by the 454 sequencing. This technology was demonstrated to have a lower cost and greater sequence depth per target than whole-genome sequencing and is well suited for resequencing specific genomic regions. The second approach combines array-based hybridization enrichment and ultra-high-throughput sequencing (Albert et al., 2007; Hodges et al., 2007; Okou et al., 2007; Porreca et al., 2007). In this approach, a high-density custom oligodexoynucleotide array can be designed to capture the desired fraction of the genome. After hybridization, the captured fragments are eluted and processed into fragments suitable for ultra-high-throughput sequencing.

Currently, the scientific community's formidable goal is to develop a technology that is capable of resequencing an entire mammalian-sized genome for $1000 (Service, 2006). When, not if, such a monumental technical advance is finally achieved, the next question will be how to bioinformatically catalog and statistically analyze thousands to millions of whole-genome sequences in crop association mapping studies.

## Genome Scans and Candidate Genes

Association studies with high density SNP coverage, large sample size, and minimum population structure offer great promise in complex trait dissection. To date, candidate-gene association studies have searched only a tiny fraction of the genome. The debate of candidate genes versus genome scans is traced to the original milestone paper of Risch and Merikangas (1996). As genomic technologies continue to evolve, we would certainly expect to see more genome-wide association analyses conducted in different plant species. So far, there have been few successful results from candidate-gene association mapping. But for many research groups, starting with candidate-gene sequences and background markers will provide a firm understanding of population structure, familial relatedness, nucleotide diversity, LD decay, and many other aspects of association mapping. Afterward, this knowledge can be built on through comprehensive genome scans with intensive sequencing and high-density genotyping.

Another reason for the promising but still limited success found in the candidate-gene approach is the way candidate genes were selected. Obviously, many candidate genes were discovered though comparisons of severe mutants and the wild-type lines. We do not have a strong understanding of naturally occurring effects of alleles at such loci. Even if the loss-of-function allele results in a significant phenotypic change, we can only expect that mild mutations would have a somewhat modest effect on the phenotype; those changes, in turn, could be detected with the assembled association mapping population. Moreover, both the frequency and effect of the allele affect whether variation explained by a locus is detectable. A skewed allele frequency would make it difficult to detect an association even though the candidate gene polymorphism is truly underlying the phenotypic variation.

## Nested Association Mapping

Ultimately, it is desirable to have both candidate-gene and genome-wide approaches to exploit in a species along with traditional linkage mapping. Joint linkage and linkage disequilibrium mapping have been proposed as a fine mapping approach in theory (Mott and Flint, 2002; Wu and Zeng, 2001; Wu et al., 2002) and demonstrated in practice (Blott et al., 2003; Meuwissen et al., 2002). Nested association mapping (NAM), as currently implemented in maize, could be an even more powerful strategy for dissecting the genetic basis of quantitative traits in species with low LD (Yu et al., 2008). For other crop species, different genetic designs (e.g., diallel, design II, eight-way cross, single round robin, or double round robin) could be used to accommodate the level of LD, practicality of creating the population and phenotyping a large number of RILs, and resources available (Churchill et al., 2004; Rebai and Goffinet, 2000; Stich et al., 2007; Verhoeven et al., 2006; Xu, 1998). In essence, by integrating genetic design, natural diversity, and genomics technologies, the NAM strategy allows high power, cost-effective genome scans, and facilitates community endeavors to link molecular variation with complex trait variation.

## Mapping and Breeding

The most commonly studied trait has been flowering time (Table 1), a trait that is heavily influenced by population structure. As we gain a better handle on genetic relatedness within association mapping panels, many other complex traits with agronomic importance are expected to be examined such as carotenoid content, disease resistance, and seed quality, besides general plant architecture traits.

Association mapping with pedigree-based germplasm is likely to pinpoint superior alleles that have been captured by breeding practices and facilitate marker-assisted selection. The approach of in silico mapping, in which association mapping is conducted with existing phenotypic, genotypic, and pedigree data generated from plant breeding programs (Arbelbide et al., 2006; Parisseaux and Bernardo, 2004; Yu et al., 2005), is complementary to the association mapping with assembled germplasm. Association mapping with diverse germplasm can identify superior alleles that were not captured by breeding practices and support introgression of these alleles into elite breeding germplasm. In a recent candidate-gene association mapping study, lycopene epsilon cyclase (lcyE) locus has been identified to alter flux down alpha-carotene versus beta-carotene branches of the carotenoid pathway among diverse maize inbred lines (Harjes et al., 2008). The association findings were further verified through linkage mapping, gene expression analysis, and mutagenesis. Because the correlation between β-carotene and grain color (scaled as shade of yellow) is low within

diverse maize germplasm, germplasm screening and direct selection of favorable lcyE alleles with the identified markers will enable breeders to more effectively produce maize lines with higher provitamin A level than screening and selection based on grain color.

Findings from these gene- or genomic region-targeted approaches can be further incorporated into two selection strategies, parental selection and marker-assisted pedigree selection. For parental selection, mixed model is used to calculate the breeding values of existing inbreds to aid the selection of parents for crossing (Bernardo, 2002; Bernardo, 2003). Within segregating breeding populations (e.g., $F_2$, $BC_1$, or three-way cross), marker-assisted recurrent selection (MARS) (Bernardo and Charcosset, 2006; Johnson, 2004) and genome-wide selection (GS) (Bernardo and Yu, 2007) can be implemented.

In summary, association mapping platforms are being developed for multiple plant species. Empirical studies from these established association mapping panels will generate valuable information for future mapping panel assembly and a better understanding of various genetic and statistical aspects of association mapping. Theoretical studies that closely track empirical results will provide valuable general guidelines for association mapping. Genetic diversity and phenotyping are expected to gain further attention, as researchers become more aware of their importance. Eventually, we will move toward researching traits, in addition to flowering time or plant height, that have economic and evolutionary values. Superior allele mining for trait improvement will be greatly facilitated by synergy among various research groups involved in different aspects of association mapping.

## Acknowledgments

This project is supported by the National Research Initiative (NRI) Plant Genome Program of the USDA Cooperative State Research, Education and Extension Service (CSREES) (2006-03578) (JY). We acknowledge other funding support from USDA-ARS (ESB), United States National Science Foundation (DBI-9872631 and DBI-0321467) (ESB), Kansas Grain Sorghum Commission (JY), and the Targeted Excellence Program of Kansas State University (JY).

## References

Abecasis, G.R., L.R. Cardon, and W.O. Cookson. 2000. A general test of association for quantitative traits in nuclear families. Am. J. Hum. Genet. 66:279–292.

Abecasis, G.R., S.S. Cherny, W.O. Cookson, and L.R. Cardon. 2002. Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. Nat. Genet. 30:97–101.

Agrama, H.A., G.C. Eizenga, and W. Yan. 2007. Association mapping of yield and its components in rice cultivars. Mol. Breed. 19:341–356.

Albert, T.J., M.N. Molla, D.M. Muzny, L. Nazareth, D. Wheeler, X. Song, T.A. Richmond, C.M. Middle, M.J. Rodesch, C.J. Packard, G.M. Weinstock, and R.A. Gibbs. 2007. Direct selection of human genomic loci by microarray hybridization. Nat. Methods 4:903–905.

Allison, D.B. 1997. Transmission-disequilibrium tests for quantitative traits. Am. J. Hum. Genet. 60:676–690.

Andersen, J.R., T. Schrag, A.E. Melchinger, I. Zein, and T. Lübberstedt. 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, M.J., S. Kim, K. Zhao, E. Bakker, M. Horton, K. Jakob, C. Lister, J. Molitor, C. Shindo, C. Tang, C. Toomajian, B. Traw, H. Zheng, J. Bergelson, C. Dean, P. Marjoram, and M. Nordborg. 2005. Genome-wide association mapping in arabidopsis identifies previously known flowering time and pathogen resistance genes. PLoS Genet 1:e60.

Arbelbide, M., J. Yu, and R. Bernardo. 2006. Power of mixed-model QTL mapping from phenotypic, pedigree and marker data in self-pollinated crops. Theor. Appl. Genet. 112:876–884.

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

Bao, J.S., H. Corke, and M. Sun. 2006. Microsatellites, single nucleotide polymorphisms and a sequence tagged site in starch-synthesizing genes in relation to starch physicochemical properties in nonwaxy rice (*Oryza sativa* L.). Theor. Appl. Genet. 113:1185–1196.

Barbazuk, W.B., S.J. Emrich, H.D. Chen, L. Li, and P.S. Schnable. 2007. SNP discovery via 454 transcriptome sequencing. Plant J. 51:910–918.

Belo, A., P. Zheng, S. Luck, B. Shen, D.J. Meyer, B. Li, S. Tingey, and A. Rafalski. 2008. Whole genome scan detects an allelic variant of fad2 associated with increased oleic acid levels in maize. Mol. Genet. Genomics 279:1–10.

Bernardo, R. 2002. Breeding for Quantitative Traits in Plants. Stemma Press, Woodbury, MN.

Bernardo, R. 2003. Parental selection, number of breeding populations, and size of each population in inbred development. Theor. Appl. Genet. 107:1252–1256.

Bernardo, R., and A. Charcosset. 2006. Usefulness of gene information in marker-assisted recurrent selection: A simulation appraisal. Crop Sci. 46:614–621.

Bernardo, R., and J. Yu. 2007. Prospects for genomewide selection for quantitative traits in maize. Crop Sci. 47:1082–1090.

Binladen, J., M.T.P. Gilbert, F.P. Bollback, C. Bendixen, R. Nielsen, and E. Willerslev. 2007. The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. PLoS ONE 2:e197.

Blott, S., J.J. Kim, S. Moisio, A. Schmidt-Kuntzel, A. Cornet, P. Berzi, N. Cambisano, C. Ford, B. Grisart, D. Johnson, L. Karim, P. Simon, R. Snell, R. Spelman, J. Wong, J. Vilkki, M. Georges, F. Farnir, and W. Coppieters. 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.

Boldman, K.G., L.A. Kriese, L.D. Van Vleck, and S.D. Kachman. 1993. A manual for the use of MTDFREML: A set of programs to obtain estimates of variances and covariances. ARS, USDA, Washington, DC.

Borevitz, J.O., and J.R. Ecker. 2004. Plant genomics: The third wave. Annu. Rev. Genomics Hum. Genet. 5:443–477.

Borevitz, J.O., D. Liang, D. Plouffe, H.S. Chang, T. Zhu, D. Weigel, C.C. Berry, E. Winzeler, and J. Chory. 2003. Large-scale identification of single-feature polymorphisms in complex genomes. Genome Res. 13:513–523.

Bradbury, P.J., Z. Zhang, D.E. Kroon, T.M. Casstevens, Y. Ramdoss, and E.S. Buckler. 2007. TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633–2635.

Braslavsky, I., B. Hebert, E. Kartalov, and S.R. Quake. 2003. Sequence information can be obtained from single DNA molecules. Proc. Natl. Acad. Sci. USA 100:3960–3964.

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

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

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

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

Casa, A.M., G. Pressoira, P.J. Brown, S.E. Mitchell, W.L. Rooney, M.R. Tuinstrac, C.D. Franks, and S. Kresovicha. 2008. Community resources and strategies for association mapping in sorghum. Crop Sci. 48:30–40.

Churchill, G.A., D.C. Airey, H. Allayee, J.M. Angel, A.D. Attie, J. Beatty, W.D. Beavis, J.K. Belknap, B. Bennett, W. Berrettini, A. Bleich, M. Bogue, K.W. Broman, K.J. Buck, E. Buckler, M. Burmeister, E.J. Chesler, J.M. Cheverud, S. Clapcote, M.N. Cook, R.D. Cox, J.C. Crabbe, W.E. Crusio, A. Darvasi, C.F. Deschepper, R.W. Doerge, C.R. Farber, J. Forejt, D. Gaile, S.J. Garlow, H. Geiger, H. Gershenfeld, T. Gordon, J. Gu, W. Gu, G. de Haan, N.L. Hayes, C. Heller, H. Himmelbauer, R. Hitzemann, K. Hunter, H.C. Hsu, F.A. Iraqi, B. Ivandic, H.J. Jacob, et al. 2004. The collaborative cross, a community resource for the genetic analysis of complex traits. Nat. Genet. 36:1133–1137.

Clark, R.M., G. Schweikert, C. Toomajian, S. Ossowski, G. Zeller, P. Shinn, N. Warthmann, T.T. Hu, G. Fu, D.A. Hinds, H. Chen, K.A. Frazer, D.H. Huson, B. Scholkopf, M. Nordborg, G. Ratsch, J.R. Ecker, and D. Weigel. 2007. Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana. Science 317:338–342.

Cogan, N.O., M.C. Drayton, R.C. Ponting, A.C. Vecchies, N.R. Bannan, T.I. Sawbridge, K.F. Smith, G.C. Spangenberg, and J.W. Forster. 2007. Validation of in silico-predicted genic SNPs in white clover (*Trifolium repens* L.), an outbreeding allopolyploid species. Mol. Genet. Genomics 277:413–425.

Dahl, F., J. Stenberg, S. Fredriksson, K. Welch, M. Zhang, M. Nilsson, D. Bicknell, W.F. Bodmer, R.W. Davis, and H. Ji. 2007. Multigene amplification and massively parallel sequencing for cancer mutation discovery. Proc. Natl. Acad. Sci. USA 104:9387–9392.

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

Doerge, R.W. 2002. Mapping and analysis of quantitative trait loci in experimental populations. Nat. Rev. Genet. 3:43–52.

Ducrocq, S., D. Madur, and A. Charcosset. 2008. Key impact of Vgt1 on flowering time adaptation in maize: Evidence from association mapping and ecogeographical information. Genetics 178:2433–2437.

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

Ersoz, E.S., J. Yu, and E.S. Buckler. 2008. Applications of linkage disequilibrium and association mapping in crop plants. p. 97–120. *In* R. Varshney and R. Tuberosa (ed.) Genomic assisted crop improvement: Vol. I: Genomics approaches and platforms. Springer Verlag, Germany.

Eskridge, K.M. 2003. Field design and the search for quantitative trait loci in plants. Available at: http://www.stat.colostate.edu/graybillconference2003/Abstracts/Eskridge.html; verified 20 May 2008.

Estoup, A., P. Jarne, and J.-M. Cornuet. 2002. Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. Mol. Ecol. 11:1591–1604.

Falush, D., M. Stephens, and J.K. Pritchard. 2003. Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. Genetics 164:1567–1587.

Falush, D., M. Stephens, and J.K. Pritchard. 2007. Inference of population structure using multilocus genotype data: Dominant markers and null alleles. Mol. Ecol. Notes 7:574–578.

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

Flint-Garcia, S.A., A. Thuillet, J. Yu, G. Pressoir, S.M. Romero, S.E. Mitchell, J. Doebley, S. Kresovich, M.M. Goodman, and E.S. Buckler. 2005. Maize association population: A high-resolution platform for quantitative trait locus dissection. Plant J. 44:1054–1064.

Gilmour, A.R., B.J. Gogel, B.R. Cullis, S.J. Welham, and R. Thompson. 2002. ASReml user guide release 1.0. VSN International Ltd., Hemel Hempstead, UK.

Gonzalez-Martinez, S.C., E. Ersoz, G.R. Brown, N.C. Wheeler, and D.B. Neale. 2006. DNA sequence variation and selection of Tag single-nucleotide polymorphisms at candidate genes for drought-stress response in *Pinus taeda* L. Genetics 172:1915–1926.

Gonzalez-Martinez, S.C., N.C. Wheeler, E. Ersoz, C.D. Nelson, and D.B. Neale. 2007. Association genetics in *Pinus taeda* L. I. Wood property traits. Genetics 175:399–409.

Gore, M., P. Bradbury, R. Hogers, M. Kirst, E. Verstege, J. van Oeveren, J. Peleman, E. Buckler, and M. van Eijk. 2007. Evaluation of target preparation methods for single-feature polymorphism detection in large complex plant genomes. Crop Sci. 47:S135–S148.

Hamblin, M.T., A.M. Casa, H. Sun, S.C. Murray, A.H. Paterson, C.F. Aquadro, and S. Kresovich. 2006. Challenges of detecting directional selection after a bottleneck: Lessons from sorghum bicolor. Science 173:953–964.

Hardy, O.J., and X. Vekemans. 2002. SPAGeDi: A versatile computer program to analyse spatial genetic structure at the individual or population levels. Mol. Ecol. Notes 2:618–620.

Harjes, C.E., T.R. Rocheford, L. Bai, T.P. Brutnell, C.B. Kandianis, S.G. Sowinski, A.E. Stapleton, R. Vallabhaneni, M. Williams, E.T. Wurtzel, J. Yan, and E.S. Buckler. 2008. Natural genetic variation in lycopene epsilon cyclase tapped for maize biofortification. Science 319:330–333.

Hedrick, P.W. 1987. Gametic disequilibrium measures: Proceed with caution. Genetics 117:331–341.

Hill, W.G., and A. Robertson. 1968. Linkage disequilibrium in finite populations. Theor. Appl. Genet. 38:226–231.

Hinds, D.A., L.L. Stuve, G.B. Nilsen, E. Halperin, E. Eskin, D.G. Ballinger, K.A. Frazer, and D.R. Cox. 2005. Whole-genome patterns of common DNA variation in three human populations. Science 307:1072–1079.

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

Hodges, E., Z. Xuan, V. Balija, M. Kramer, M.N. Molla, S.W. Smith, C.M. Middle, M.J. Rodesch, T.J. Albert, G.J. Hannon, and W.R. McCombie. 2007. Genome-wide in situ exon capture for selective resequencing. Nat. Genet. 39:1522–1527.

Holland, J.B. 2007. Genetic architecture of complex traits in plants. Curr. Opin. Plant Biol. 10:156–161.

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

Ihaka, R., and R. Gentleman. 1996. R: A language for data analysis and graphics. J. Comput. Graph. Stat. 5:299–314.

Johnson, G.C.L., L. Esposito, B.J. Barratt, A.N. Smith, J. Heward, G. Di Genova, H. Ueda, H.J. Cordell, I.A. Eaves, F. Dudbridge, R.C.J. Twells, F. Payne, W. Hughes, S. Nutland, H. Stevens, P. Carr, E. Tuomilehto-Wolf, J. Tuomilehto, S.C.L. Gough, D.G. Clayton, and J.A. Todd. 2001. Haplotype tagging for the identification of common disease genes. Nat. Genet. 29:233.

Johnson, R. 2004. Marker-assisted selection. Plant Breed. Rev. 24:293–309.

Karayiorgou, M., C. Sobin, M.L. Blundell, B.L. Galke, L. Malinova, P. Goldberg, J. Ott, and J.A. Gogos. 1999. Family-based association studies support a sexually dimorphic effect of COMT and MAOA on genetic susceptibility to obsessive-compulsive disorder- Extending the Transmission Disequilibrium Test (TDT) to Examine Genetic Heterogeneity. Biol. Psychiatry 45:1178–1189.

Kearsey, M.J., and A.G. Farquhar. 1998. QTL analysis in plants; where are we now? Heredity 80:137–142.

Kim, S., K. Zhao, R. Jiang, J. Molitor, J.O. Borevitz, M. Nordborg, and P. Marjoram. 2006. Association mapping with single-feature polymorphisms. Genetics 173:1125–1133.

Korbel, J.O., A.E. Urban, J.P. Affourtit, B. Godwin, F. Grubert, J.F. Simons, P.M. Kim, D. Palejev, N.J. Carriero, L. Du, B.E. Taillon, Z. Chen, A. Tanzer, A.C.E. Saunders, J. Chi, F. Yang, N.P. Carter, M.E. Hurles, S.M. Weissman, T.T. Harkins, M.B. Gerstein, M. Egholm, and M. Snyder. 2007. Paired-end mapping reveals extensive structural variation in the human genome. Science 318:420–426.

Kraakman, A.T.W., F. Martínez, B. Mussiraliev, F.A. v. Eeuwijk, and R.E. Niks. 2006. Linkage disequilibrium mapping of morphological, resistance, and other agronomically relevant traits in modern spring barley cultivars. Mol. Breed. 17:41–58.

Kui, Z., P. Calabrese, M. Nordborg, and S. Fengzhu. 2002. Haplotype block structure and its applications to association studies. Power and Study Designs. Am. J. Hum. Genet. 71:1386.
Kumar, R., J. Qiu, T. Joshi, B. Valliyodan, D. Xu, and H.T. Nguyen. 2007. Single feature polymorphism discovery in rice. PLoS ONE 2:e284.
Kwok, P.Y. 2000. High-throughput genotyping assay approaches. Pharmacogenomics 1:95–100.
Levinson, G., and G.A. Gutman. 1987. Slipped-strand mispairing: A major mechanism for DNA sequence evolution. Mol. Biol. Evol. 4:203–221.
Lewontin, R.C. 1964. The Interaction of Selection and Linkage. I. General considerations; heterotic models. Genetics 49:49–67.
Li, Y.-C., A.B. Korol, T. Fahima, A. Beiles, and E. Nevo. 2002. Microsatellites: Genomic distribution, putative functions and mutational mechanisms: A review. Mol. Ecol. 11:2453–2465.
Lynch, M., and K. Ritland. 1999. Estimation of pairwise relatedness with molecular markers. Genetics 152:1753–1766.
Mackay, T.F. 2001. The genetic architecture of quantitative traits. Annu. Rev. Genet. 35:303–339.
Malosetti, M., C.G. van der Linden, B. Vosman, and F.A. van Eeuwijk. 2007. A mixed-model approach to association mapping using pedigree information with an illustration of resistance to *phytophthora infestans* in potato. Genetics 175:879–889.
Margulies, M., M. Egholm, W.E. Altman, S. Attiya, J.S. Bader, L.A. Bemben, J. Berka, M.S. Braverman, Y.-J. Chen, Z. Chen, S.B. Dewell, L. Du, J.M. Fierro, X.V. Gomes, B.C. Godwin, W. He, S. Helgesen, C.H. Ho, G.P. Irzyk, S.C. Jando, M.L.I. Alenquer, T.P. Jarvie, K.B. Jirage, J.-B. Kim, J.R. Knight, J.R. Lanza, J.H. Leamon, S.M. Lefkowitz, M. Lei, J. Li, K.L. Lohman, H. Lu, V.B. Makhijani, K.E. McDade, M.P. McKenna, E.W. Myers, E. Nickerson, J.R. Nobile, R. Plant, B.P. Puc, M.T. Ronan, G.T. Roth, G.J. Sarkis, J.F. Simons, J.W. Simpson, M. Srinivasan, K.R. Tartaro, A. Tomasz, K.A. Vogt, G.A. Volkmer, S.H. Wang, Y. Wang, M.P. Weiner, P. Yu, R.F. Begley, and J.M. Rothberg. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376.
McNally, K.L., R. Bruskiewich, D. Mackill, C.R. Buell, J.E. Leach, and H. Leung. 2006. Sequencing multiple and diverse rice varieties. Connecting whole-genome variation with phenotypes. Plant Physiol. 141:26–31.
Meuwissen, T.H., A. Karlsen, S. Lien, I. Olsaker, and M.E. Goddard. 2002. Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. Genetics 161:373–379.
Meyer, M., U. Stenzel, S. Myles, K. Prufer, and M. Hofreiter. 2007. Targeted high-throughput sequencing of tagged nucleic acid samples. Nucleic Acids Res. 35:e97.
Mitchell, S.E., S. Kresovich, C.A. Jester, C.J. Hernandez, and A.K. Szewc-McFadden. 1997. Application of multiplex PCR and fluorescence-based, semi-automated allele sizing technology for genotyping plant genetic resources. Crop Sci. 37:617–624.
Mockler, T.C., S. Chan, A. Sundaresan, H. Chen, S.E. Jacobsen, and J.R. Ecker. 2005. Applications of DNA tiling arrays for whole-genome analysis. Genomics 85:1–15.
Mott, R., and J. Flint. 2002. Simultaneous detection and fine mapping of quantitative trait loci in mice using heterogeneous stocks. Genetics 160:1609–1618.
Muehlbauer. 2006. Barley coordinated agricultural project proposal. Available at: http://barleycap.cfans.umn.edu/ (verified 20 May 2008).
Nordborg, M., and S. Tavare. 2002. Linkage disequilibrium: What history has to tell us. Trends Genet. 18:83–90.
Nordborg, M., T.T. Hu, Y. Ishino, J. Jhaveri, C. Toomajian, H. Zheng, E. Bakker, P. Calabrese, J. Gladstone, R. Goyal, M. Jakobsson, S. Kim, Y. Morozov, B. Padhukasahasram, V. Plagnol, N.A. Rosenberg, C. Shah, J.D. Wall, J. Wang, K. Zhao, T. Kalbfleisch, V. Schulz, M. Kreitman, and J. Bergelson. 2005. The pattern of polymorphism in Arabidopsis thaliana. PLoS Biol. 3:e196.
Okou, D.T., K.M. Steinberg, C. Middle, D.J. Cutler, T.J. Albert, and M.E. Zwick. 2007. Microarray-based genomic selection for high-throughput resequencing. Nat. Methods 4:907–909.
Olsen, K.M., and M.D. Purugganan. 2002. Molecular evidence on the origin and evolution of glutinous rice. Genetics 162:941–950.
Olsen, K.M., S.S. Halldorsdottir, J.R. Stinchcombe, C. Weinig, J. Schmitt, and M.D. Purugganan. 2004. Linkage disequilibrium mapping of arabidopsis CRY2 flowering time alleles. Genetics 167:1361–1369.
Palaisa, K., M. Morgante, S. Tingey, and A. Rafalski. 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.
Parameswaran, P., R. Jalili, L. Tao, S. Shokralla, B. Gharizadeh, M. Ronaghi, and A.Z. Fire. 2007. A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing. Nucl. Acids Res. 35:e130.
Parisseaux, B., and R. Bernardo. 2004. In silico mapping of quantitative trait loci in maize. Theor. Appl. Genet. 109:508–514.
Patterson, N., A.L. Price, and D. Reich. 2007. Population structure and eigenanalysis. PLoS Genet 2:e90.
Porreca, G.J., K. Zhang, J.B. Li, B. Xie, D. Austin, S.L. Vassallo, E.M. LeProust, B.J. Peck, C.J. Emig, F. Dahl, Y. Gao, G.M. Church, and J. Shendure. 2007. Multiplex amplification of large sets of human exons. Nat. Methods 4:931–936.
Price, A.H. 2006. Believe it or not, QTLs are accurate! Trends Plant Sci. 11:213–216.
Price, A.L., N.J. Patterson, R.M. Plenge, M.E. Weinblatt, N.A. Shadick, and D. Reich. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38:904–909.
Pritchard, J.K., and N.A. Rosenberg. 1999. Use of unlinked genetic markers to detect population stratification in association studies. Am. J. Hum. Genet. 65:220–228.
Pritchard, J.K., M. Stephens, and P. Donnelly. 2000a. Inference of population structure using multilocus genotype data. Genetics 155:945–959.
Pritchard, J.K., M. Stephens, N.A. Rosenberg, and P. Donnelly. 2000b. Association mapping in structured populations. Am. J. Hum. Genet. 67:170–181.
Rafalski, A. 2002. Applications of single nucleotide polymorphisms in crop genetics. Curr. Opin. Plant Biol. 5:94.
Rebai, A., and B. Goffinet. 2000. More about quantitative trait locus mapping with diallel designs. Genet. Res. 75:243–247.
Risch, N., and K. Merikangas. 1996. The future of genetic studies of complex human diseases. Science 273:1516–1517.
Ritland, K. 2005. Multilocus estimation of pairwise relatedness with dominant markers. Mol. Ecol. 14:3157–3165.
Rostoks, N., J.O. Borevitz, P.E. Hedley, J. Russell, S. Mudie, J. Morris, L. Cardle, D.F. Marshall, and R. Waugh. 2005. Single-feature polymorphism discovery in the barley transcriptome. Genome Biol. 6:R54.
Rostoks, N., L. Ramsay, K. Mackenzie, L. Cardle, P.R. Bhat, M.L. Roose, J.T. Svensson, N. Stein, R.K. Varshney, D.F. Marshall, A. Graner, T.J. Close, and R. Waugh. 2006. Recent history of artificial outcrossing facilitates whole-genome association mapping in elite inbred crop varieties. Proc. Natl. Acad. Sci. USA.
Salisbury, M. 2007. Next-gen sequencing: The waiting game. Genome Technol. 72:26–28.
Salvi, S. 2007. Conserved non-coding genomic sequences associated with a flowering-time quantitative trait locus in maize. Proc. Natl. Acad. Sci. USA 104:11376–11381.
SAS Institute. 1999. SAS/STAT user's guide. Version 8 SAS Institute, Inc, Cary, NC.
Service, R.F. 2006. GENE SEQUENCING: The race for the $1000 genome. Science 311:1544–1546.
Shendure, J., G.J. Porreca, N.B. Reppas, X. Lin, J.P. McCutcheon, A.M. Rosenbaum, M.D. Wang, K. Zhang, R.D. Mitra, and G.M. Church. 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309:1728–1732.
Simko, I. 2004. One potato, two potato: Haplotype association mapping in autotetraploids. Trends Plant Sci. 9:441.
Skøt, L., J. Humphreys, M.O. Humphreys, D. Thorogood, J. Gallagher, R. Sanderson, I.P. Armstead, and I.D. Thomas. 2007. Association of

candidate genes with flowering time and water-soluble carbohydrate content in *Lolium perenne* (L.). Genetics 177:535–547.
Skøt, L., M.O. Humphreys, I. Armstead, S. Heywood, K.P. Skøt, R. Sanderson, I.D. Thomas, K.H. Chorlton, and N.R.S. Hamilton. 2005. An association mapping approach to identify flowering time genes in natural populations of *Lolium perenne* (L.). Mol. Breed. 15:233–245.
Spielman, R.S., R.E. McGinnis, and W.J. Ewens. 1993. Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am. J. Hum. Genet. 52:506–516.
Stephens, M., N.J. Smith, and P. Donnelly. 2001. A new statistical method for haplotype reconstruction from population data. Am. J. Hum. Genet. 68:978–989.
Stich, B., J. Yu, A.E. Melchinger, H.P. Piepho, H.F. Utz, H.P. Maurer, and E.S. Buckler. 2007. Power to detect higher-order epistatic interactions in a metabolic pathway using a new mapping strategy. Genetics 176:563–570.
Syvanen, A.C. 2001. Accessing genetic variation: Genotyping single nucleotide polymorphisms. Nat. Rev. Genet. 2:930–942.
Syvanen, A.C. 2005. Toward genome-wide SNP genotyping. Nat. Genet. 37:S5–S10.
Szalma, S.J., E.S. Buckler, M.E. Snook, and M.D. McMullen. 2005. Association analysis of candidate genes for maysin and chlorogenic acid accumulation in maize silks. Theor. Appl. Genet. 110:1324–1333.
Tabor, H.K., N.J. Risch, and R.M. Myers. 2002. Candidate-gene approaches for studying complex genetic traits: Practical considerations. Nat. Rev. Genet. 3:391–397.
Tanksley, S.D., and S.R. McCouch. 1997. Seed banks and molecular maps: Unlocking genetic potential from the wild. Science 277:1063–1066.
The Wellcome Trust Case Control Consortium. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3000 shared controls. Nature 447:661–678.
Thornsberry, J.M., M.M. Goodman, J. Doebley, S. Kresovich, D. Nielsen, and E.S. Buckler. 2001. Dwarf8 polymorphisms associate with variation in flowering time. Nat. Genet. 28:286–289.
Thumma, B.R., and M.F. Nolan. 2005. Polymorphisms in cinnamoyl CoA reductase (CCR) are associated with variation in microfibril angle in Eucalyptus spp. Genetics 173:1257–1265.
Tracy, W.F., S.R. Whitt, and E.S. Buckler. 2006. Recurrent mutation and genome evolution: Example of Sugary1 and the origin of sweet maize. Crop Sci. 46:S49–S54.
Verhoeven, K.J., J.L. Jannink, and L.M. McIntyre. 2006. Using mating designs to uncover QTL and the genetic architecture of complex traits. Heredity 96:139–149.
Viard, F., P. Franck, M.-P. Dubois, A. Estoup, and P. Jarne. 1998. Variation of microsatellite size homoplasy across electromorphs, loci, and populations in three invertebrate species. J. Mol. Evol. 47:42–51.
Vigouroux, Y., J.S. Jaqueth, Y. Matsuoka, O.S. Smith, W.D. Beavis, J.S.C. Smith, and J. Doebley. 2002. Rate and pattern of mutation at microsatellite loci in maize. Mol. Biol. Evol. 19:1251–1260.
Vos, P., R. Hogers, M. Bleeker, M. Reijans, T. van de Lee, M. Hornes, A. Frijters, J. Pot, J. Peleman, and M. Kuiper. 1995. AFLP: A new technique for DNA fingerprinting. Nucleic Acids Res. 23:4407–4414.
Weber, A., R.M. Clark, L. Vaughn, J.D.J. Sánchez-Gonzalez, J. Yu, B.S. Yandell, P. Bradbury, and J.F. Doebley. 2008. Major regulatory genes in maize contribute to standing variation in Teosinte (*Zea mays* ssp. parviglumis). Genetics 177:2349–2359.
Wei, X.M., P.A. Jackson, C.L. McIntyre, K.S. Aitken, and B. Croft. 2006. Associations between DNA markers and resistance to diseases in sugarcane and effects of population substructure. Theor. Appl. Genet. 114:155–164.
Whitt, S.R., and E.S. Buckler. 2003. Using natural allelic diversity to evaluate gene function. Methods Mol. Biol. 236:123–140.
Williams, J.G.K., A.R. Kubelik, K.J. Livak, J.A. Rafalski, and S.V. Tingey. 1990. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Res. 18:6531–6535.
Wilson, L.M., S.R. Whitt, T.R. Rocheford, M.M. Goodman, and E.S. Buckler, 4th. 2004. Dissection of maize kernel composition and starch production by candidate gene association. Plant Cell 16:2719–2733.
Winzeler, E.A., D.R. Richards, A.R. Conway, A.L. Goldstein, S. Kalman, M.J. McCullough, J.H. McCusker, D.A. Stevens, L. Wodicka, D.J. Lockhart, and R.W. Davis. 1998. Direct allelic variation scanning of the yeast genome. Science 281:1194–1197.
Wu, R., and Z.B. Zeng. 2001. Joint linkage and linkage disequilibrium mapping in natural populations. Genetics 157:899–909.
Wu, R., C.X. Ma, and G. Casella. 2002. Joint linkage and linkage disequilibrium mapping of quantitative trait loci in natural populations. Genetics 160:779–792.
Xu, S. 1998. Mapping quantitative trait loci using multiple families of line crosses. Genetics 148:517–524.
Yu, J., and E.S. Buckler. 2006. Genetic association mapping and genome organization of maize. Curr. Opin. Biotechnol. 17:155–160.
Yu, J., M. Arbelbide, and R. Bernardo. 2005. Power of in silico QTL mapping from phenotypic, pedigree, and marker data in a hybrid breeding program. Theor. Appl. Genet. 110:1061–1067.
Yu, J., J.B. Holland, M.D. McMullen, and E.S. Buckler. 2008. Genetic design and statistical power of nested association mapping in maize. Genetics 178:539–551.
Yu, J., G. Pressoir, W.H. Briggs, I. Vroh Bi, M. Yamasaki, J.F. Doebley, M.D. McMullen, B.S. Gaut, D.M. Nielsen, J.B. Holland, S. Kresovich, and E.S. Buckler. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. 38:203–208.
Zamir, D. 2001. Improving plant breeding with exotic genetic libraries. Nat. Rev. Genet. 2:983–989.
Zeng, Z.B. 2005. QTL mapping and the genetic basis of adaptation: Recent developments Genetica:25–37.
Zhao, K., M.J. Aranzana, S. Kim, C. Lister, C. Shindo, C. Tang, C. Toomajian, H. Zheng, C. Dean, P. Marjoram, and M. Nordborg. 2007. An arabidopsis example of association mapping in structured samples. PLoS Genet 3:e4.