Association Breeding Strategies for Improvement of Self-Pollinated Crops

Mark E. Sorrells
Cornell University, Department of Plant Breeding and Genetics, Ithaca, NY 14853, http://plbrgen.cals.cornell.edu/ Email mes12@cornell.edu

Abstract

Association breeding strategies utilize phenotypic and genotypic information to increase gain from selection and reduce selection cycle time. Because breeding programs are dynamic, complex genetic entities, marker / phenotype relationships must be evaluated frequently. In contrast to a biparental-cross population, association mapping can be conducted directly on the breeding material, thus facilitating the practical use of information in a crop improvement program. Because there is more genetic variation in a breeding program than in a biparental cross, phenotypic variation and marker polymorphism are much higher. Genotypic data can be combined with phenotypic data from routine screening and variety trial evaluations to facilitate selection for low heritability traits. Novel alleles can be identified and the relative allelic value can be assessed as often as necessary. To minimize statistical error, correction for population structure is critical in a collection of genotypes, especially in a breeding program where relationships are highly variable. Three approaches are elaborated that utilize molecular marker information for crop improvement: 1) association breeding: crossing/ selection/ testing program, 2) marker-assisted recurrent selection (MARS), and 3) genome-wide selection. Increasingly efficient breeding methods will continue to be developed for identifying and evaluating allelic effects on a large scale so that breeders can assemble desirable alleles in superior varieties.

Media Summary

As plant breeders expand our knowledge of how genes evolve and interact to produce an infinite range of phenotypes, new strategies will emerge for manipulating genetic variation to the benefit of humankind.

Key Words

Association analysis, molecular markers, linkage disequilibrium, genotype x environment interaction, breeding methods, quantitative trait loci

Introduction

Plant breeders play a critical role in the utilization of biological diversity for the benefit of humankind. A multidisciplinary approach is required to successfully develop superior crop germplasm and varieties. Crop improvement is driven by new knowledge and technological advances that increase the efficiency of accurately selecting unique phenotypes and genotypes for target environments. Molecular markers have enhanced the resolution of genome mapping in those species without a complete genome sequence and contributed to our understanding of the genetic control of important traits. With the dramatic improvements in genotyping technologies, phenotyping has become a bottleneck and new methods and technologies that increase the efficiency of phenotyping and data analysis will contribute significantly to crop improvement. This review will cover strategies and approaches for effectively exploiting association analyses for crop improvement. Much of the information is general and can be applied to all crops, however many of the examples will draw on the maize and Triticeae literature. Other recent reviews include Jannink et al. (2001), Gupta et al. (2005), Breseghello and Sorrells (2006a), Ersoz et al. (2007), MacKay and Powell (2007) or Yu and Buckler (2006).

Linkage Disequilibrium

Estimates of the level of genetic diversity and historical relationships among germplasms are very useful for association mapping (AM) and the exploitation of genetic variation in cultivated species. Linkage disequilibrium (LD), or non-random association of alleles at adjacent loci throughout the genome within a population forms the
basis for AM strategies. In biparental crosses, LD is maximized. The power of association analysis is affected by
the patterns of LD, the extent of LD in the genome, and the variation in LD from one population to another.
Several factors affect LD including mating system, recombination rate, population structure, population history,
genetic drift, directional selection and gene fixation (reviewed by Gaut and Long 2003). Linkage disequilibrium
estimates for self-pollinated crops such as wheat and barley indicated that LD decays over 5 to 40 cM, a much
slower rate than reported for outcrossing species such as maize. The LD between alleles at two loci can be
measured using quantitative metrics such as \( D, D' \) (Lewontin 1964), \( r^2 \) (Hill and Robertson 1968), or other
statistics (e.g. Hedrick 1987; Pritchard and Przeworski 2001). The \( D' \) statistic is partially normalized with
respect to allele frequencies. A commonly used statistic is the squared value of the Pearson’s (product moment)
correlation coefficient, \( r^2 \), that is a measure of the proportion of the variance of a response variable explained by
a predictor variable (Hill and Robertson 1968). Although \( r^2 \) is also affected by gene frequency, the intuitive
nature of \( r^2 \) facilitates the interpretation of marker densities and association analyses. If a causative DNA
polymorphism, or quantitative trait nucleotide (QTN), is assumed to contribute a fraction of the total variation in
a quantitative trait, we can estimate the fraction of the variance explained by a marker in LD. For example, if the
QTN has a heritability of \( h^2_Q \) (i.e., it explains that fraction of the phenotypic variance), then the fraction of the
phenotypic variance explained by a marker in LD is \( r^2 \times h^2_Q \).

LD varies widely within a genome, among different populations, and among species (e.g. Remington et al.
2001; Flint-Garcia et al. 2003; Breseghello and Sorrells 2006b; Maccaferri et al. 2006; Rostoks et al. 2006;
Ersoz et al. 2007). Never the less, the genome-wide LD is of interest as a general guide to marker density that
may be required for whole-genome AM and each population must be evaluated on a case-by-case basis.
Typically, \( r^2 \) values for all pairwise linked or syntenic markers are plotted against either map distance or
physical distance. Values of \( r^2 = 0.1-0.2 \) are sometimes chosen arbitrarily as an indicator of statistically
significant LD. This level of LD, however, would indicate that the closest marker only captures 10 to 20% of the
phenotypic variation resulting from a causal polymorphism. This is insufficient to generate detectable
association between the marker and the phenotype.

**Relevance of Germplasm to Association Studies**

Germplasm can be broadly classified into three categories: exotic accessions from germplasm bank collections,
intermating populations, and elite lines. The genetic expectations and applications of these classes of germplasm
differ widely (Breseghello and Sorrells 2006a). Exotic accessions from a germplasm bank may be used to screen
high heritability traits, whereas elite cultivars and lines are typically evaluated for low heritability traits in
replicated, multi-environment trials. Intermated progenies of a segregating population are often evaluated each
generation or cycle and depending on the recurrent selection method and traits. The genetic expectations for an
exotic core collection are low LD, low to medium population structure, and high allelic diversity. In early
generations of a segregating population, LD is high and declines with additional cycles of intermating and
selection. Breeders often use exotic germplasm as a source of novel alleles in a marker-assisted backcross
scheme whereas elite lines are intermated and marker-assisted selection is used in the segregating progenies in a
forward-breeding strategy. Intermated segregating populations can balance power and precision for association
analysis and would allow mapping of quantitative traits with increasing resolution through cycles of intermating.

**Association Breeding**

Breeding progress is based on i) the discovery and generation of genetic variation for agronomic traits, ii)
development of genotypes with new or improved attributes due to superior combinations of alleles at multiple
loci, and iii) accurate selection of rare genotypes that possess new improved characteristics. Association analysis
can be integrated into breeding programs by incorporating genotype information to facilitate marker-assisted
selection of parents and segregating populations. Germplasm is in a constant state of flux in a breeding program.
Consequently, frequent evaluation of marker / phenotype relationships is required. In contrast to biparental cross
populations, association mapping can be conducted directly on the breeding material greatly facilitating the
practical use of information in a crop improvement program. Genotypic data can be combined with phenotypic
data from routine screening and variety trial evaluations to facilitate selection for low heritability traits. Another
important advantage for a breeding program is that novel alleles can be identified and the relative allelic value can be assessed as often as necessary. To minimize statistical error, correction for population structure is critical in a collection of genotypes, especially in a breeding program where relationships are highly variable. The estimation of gene effects using molecular markers is susceptible to errors resulting from sampling variance and systematic biases. Also, population structure can contribute to increased false positive associations. In a typical breeding program, there is likely to be a wide range in genetic relationships among the genotypes as well as heterogeneity requiring the use of covariates for population structure and kinship (Yu et al. 2006; Crossa et al. 2007) in a mixed linear model.

There are three approaches elaborated that utilize molecular marker information for crop improvement: 1) association breeding: crossing/selection/testing program, 2) marker-assisted recurrent selection (MARS), and 3) genome-wide selection. In a typical breeding program, selected genotypes are crossed to produce new populations that are subject to phenotypic and/or genotypic selection. Those materials are either intermated or inbred to produce new populations or inbred lines that are evaluated in replicated, multi-environment trials. The breeder uses the trial information to select elite parents that re-enter the hybridization program. In association breeding, genotypic data (preferably whole genome coverage) and the appropriate analyses are incorporated to validate previously mapped marker/trait associations and potentially identify new ones. This information is used to estimate allelic value at selected loci (or all loci in genome-wide selection) and then create a genotypic value index for each genotype and trait (Land and Thompson 1990; Christopher et al. 2007). The marker assisted recurrent selection method involves the improvement of an F2 or F1-derived doubled haploid (DH) population using one generation of phenotypic and genotypic evaluation to identify marker/trait associations followed by multiple cycles of recurrent selection using only allelic values at the selected marker loci (Johnson 2001 2004; Bernardo and Yu 2007). The genetic gain of MARS over phenotypic selection has been studied through computer simulation in maize (Bernardo and Charcosset 2006; Bernardo et al. 2006). The genome-wide selection method involves marker-assisted selection in which selections are based on all markers across the entire genome rather than just those showing significant effects. Phenotypic and genotypic information is combined to produce breeding values of all the markers that are then fitted as random effects in a linear model. Individuals in subsequent recurrent selection generations are then selected based on the sum of those breeding values (Meuwissen et al. 2001, 2002). Bernardo and Yu (2007) compared MARS to genome-wide selection in simulations involving a population of 144 individuals from which 4 individuals were selected in cycles 1 and 2. For MARS, a selection index based on all selected markers was calculated as suggested by Lande and Thompson (1990). For GWS, the best linear unbiased predictor (BLUP) of breeding values was estimated by fitting all the markers as random effects and imposing a convenient assumption of equal variances based on cycle 0 performance. Genome-wide selection resulted in a larger response to selection than MARS. Depending on the number of QTL and the heritability, the response to genome-wide selection was 6-18% higher than MARS with the biggest advantage for complex traits with low heritability.

More efficient methods will continue to be developed for identifying and evaluating allelic effects on a large scale so that breeders can assemble desirable alleles in superior varieties. As we expand our knowledge of how genes evolve and interact to produce the nearly infinite range of phenotypes, new opportunities to manipulate genetic variation to the benefit of humankind will arise.

References