Scanning for yield: high-throughput discovery of candidate agronomic loci for marker-assisted selection in maize

RATIONALE AND SIGNIFICANCE

Importance of maize

Maize is a natural resource of fundamental national importance, vital for food, livestock feed, and fuel production. Furthermore, maize is by far the most valuable agricultural crop in the United States: in 2007, annual production in the U.S. was worth more than $50 billion dollars (USDA 2009), almost half the value of U.S. crude oil production the same year (EIA 2009) and double the value of soybean, the next most important field crop (USDA 2009). With a growing human population and an increased demand for alternative fuels, maize production must keep pace. Maize yield has steadily increased over the last three decades (Fig. 1), but continued efforts are necessary to maintain or improve this yield trajectory. An increase in the acreage planted to maize may accommodate some of this need, but it is clear that this is not a feasible long-term solution: in the last 10 years, for example, only ~30% of the total increase in yield can be attributed to increased acreage (USDA 2009), and every day more arable land is lost to development. Some gains in yield may also be won by improved farming practices, but the majority of yield increase over the past 70 years can be directly attributed to breeding efforts (Duvick 1992), and breeding must remain of central importance in order to meet increased yield demands.

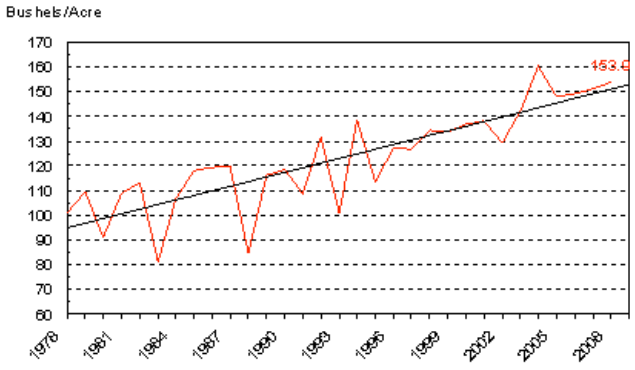

Figure 1. U.S. Corn Yield. Data from USDA (2009).

Improved breeding methods

Continued breeding efforts and novel approaches to breeding are essential to continue increase in yield. Marker-assisted selection (MAS) has become an important tool of modern plant breeding, allowing breeders to combine traits of interest without the need for additional, costly phenotyping, pedigree analysis, or even a detailed functional understanding of the molecular basis of a trait. The success of MAS relies on the availability of markers that associate with agronomic traits. Genetic mapping has been enormously successful in identifying markers for use in MAS. But traditional mapping approaches such as QTL and association mapping have a number of drawbacks that may limit their utility or generality (see below).

Selection mapping can circumvent some of these problems by utilizing changes in allele frequency to identify markers that have been, or are tightly linked to, the target of historical selection. We propose a novel approach to selection mapping that takes into account the complex relatedness of breeding populations to identify candidate agronomic loci (CAL) for MAS in maize. We further propose to directly test the utility of these CAL by comparing their performance in yield trials to MAS with random loci and to line selection on yield alone. The

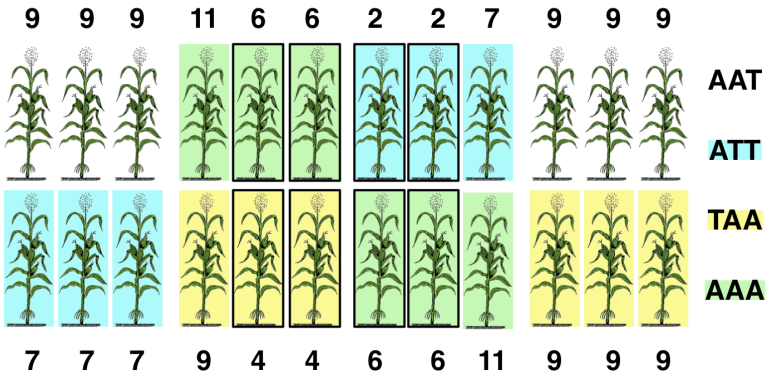

Figure 2. Hypothetical yield trial. Rows of four inbred varieties (colors) crossed to a common tester in a hybrid yield trial. Shown are their yield (top and bottom) and genotypes (right) at three candidate loci. Poor environment lowers the yield of eight rows (solid border) in the center of the field. As a result, although the A allele is associated with higher yield at each locus, the white variety (mean yield = 9) appears phenotypically superior to the green variety (mean yield = 7.67).

resulting list of CAL, along with the software and genotyping tools we will develop, will allow breeders to select lines using useful genetic markers rather than phenotype alone, accelerating the progress of yield improvement (Fig. 2). Moreover, the mapping approach and methods developed here should be easily extendible to virtually any group of lines in any crop species with sufficient genomic resources to generate dense, genome-wide genotype data, regardless of the availability of pedigree information; we expect that our selection mapping approach will prove to be an important advance in breeding methods for a number of crop species.

INTRODUCTION

Maize evolution and breeding

The evolution of maize as a crop can be divided into three main periods: domestication, improvement, and modern breeding (Fig. 3). Evolutionarily, *Zea mays* is a young species, having diverged from other lineage in the genus as little as 150-300,000 years ago (Ross-Ibarra *et al.* 2009). Maize, *Zea mays* ssp. *mays*, was domesticated approximately 9,000 years ago from its wild ancestor, the teosinte *Zea mays* ssp. *parviglumis* (Piperno *et al.* 2009, Matsuoka *et al.* 2002). Domestication involved numerous dramatic changes to plant morphology and architecture (reviewed in Doebley 2004). Maize is an outcrossing crop, and, for most of its cultivated history, mass selection on open-pollinated populations (landraces) was the only form of maize breeding. Starting in the early to mid 20th century, maize breeders began to focus on hybrids of inbred parental lines in order to take advantage of heterosis. During this period of improvement, the change from open-pollinated to inbred lines almost certainly brought about selection for new characteristics (c.f. Yamasaki *et al.* 2005), including heterosis and for individual inbreds to serve primarily male or female function. Since the development of hybrid maize 50-60 years ago, modern breeding has shifted from double-cross hybrids to single-cross hybrids, and breeders have focused considerable efforts on the development of inbred lines within distinct heterotic groups (most notably, stiff-stalk and non-stiff stalk). These differences, combined with advances in management (increased planting density, increased use of inputs such as nitrogen fertilizer, etc.), have undoubtedly led to novel

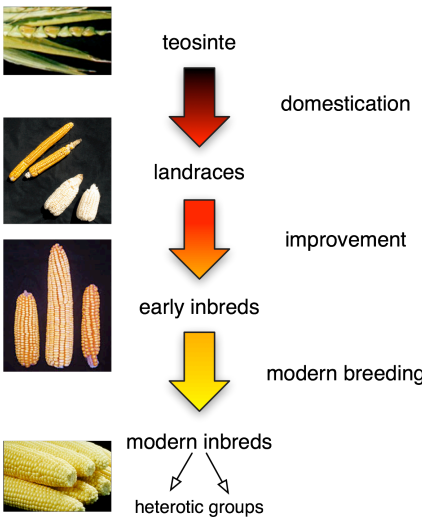

Figure 3. Maize evolution

selective regimes resulting in additional allelic change in modern lines (c.f. Duvick *et al.* 2004, Feng *et al.* 2006).

Genetics of yield

Yield is perhaps the quintessential quantitative trait, controlled by a large number of genetic factors and affected by numerous other traits, including plant architecture, inflorescence morphology, flowering time, or resistance to drought and disease. The yield potential of individual loci is strongly affected not only by genetic background but by interaction with environmental variance as well. Because of the relatively low heritability of yield, it would be preferable to identify the genetic factors that underly yield-correlated traits, such as those mentioned above. However, the traits that are most important for yield improvement in any given cross or in any set of environments are both difficult to identify and difficult to measure. This suggests that new mapping methods are required to identify the many and various factors that influence yield in different backgrounds and environments. Ideally, such a method would identify the loci with largest impact on yield, benefiting from the generations of yield trials and line selections that many hundreds of breeders have conducted.

Yield is the single most important agronomic trait in maize, and has continued to increase from initial maize domestication up through modern inbred varieties; selection for increased yield must have been a powerful force affecting allele frequencies in maize population. Because of the quantitative nature of yield, the variety of environments and genetic backgrounds used in modern maize breeding, and the observed continual increase in yield, there must be a large number of loci for which genetic variance for yield is present in even the most elite collection of modern inbred lines. Continued gains in yield will thus rely on our ability to combine these favorable alleles in breeding populations.

Traditional mapping strategies

Traditionally, two distinct mapping approaches have been used to identify candidate agronomic loci. These methods are based on what has been termed a "top-down" approach, beginning with a phenotype of interest and then identifying causative genomic regions with Quantitative Trait Locus (QTL) and association or Linkage Disequilibrium (LD) mapping (Ross-Ibarra *et al.* 2007). The most successful method for finding these genes has been QTL mapping, but LD methods are rapidly gaining favor in the plant genomics community.

Quantitative Trait Locus (QTL) mapping

QTL mapping was the first – and is still the most widely used – method available for localizing the genetic basis of a trait (e.g. Sax 1923). Coupled with molecular markers, QTL mapping of agronomic traits has been enormously successful, permitting the identification of loci (actually, typically large chromosomal regions) that underly such diverse traits as fruit morphology (Frary *et al.* 2003), drought tolerance (Tuberosa and Salvi 2006), disease resistance (Young 1996), and domestication-related traits (Ross-Ibarra 2005). Moreover, favorable traits identified in QTL studies can be efficiently combined with MAS (Dekkers and Hospital 2002, Ashikari *et al.* 2005). But QTL mapping is not without its limitations. The most obvious limitation is the time-intensive process of developing crosses and mapping populations, often requiring many generations of careful backcrossing or selfing to establish mapping lines. Of even greater import, however, is the fact that the results of QTL analysis are often dependent on the environment in which the population is grown (Paterson *et al.* 1991) as well as the parental lines

used in the cross (Doebley and Stec 1991, Li *et al.* 2006a); both of these problems negatively affect the generality of results from QTL experiments. In one recent study in maize, for example, a QTL identified in one population was discovered to have the opposite phenotypic effect when introgressed into a second population (Bouchez *et al.* 2002). A number of statistical issues pose challenges for QTL analysis as well, the most important of which is the limited power to accurately estimate the number and size of QTLs – an observation that has become known as the Beavis Effect (Beavis 1994, 1998). Though this latter limitation has not proven problematic for cloning genes of large phenotypic effect (e.g. Shomura *et al.* 2008, Li *et al.* 2006b, Uauy *et al*. 2006, Frary *et al.* 2000, Doebley *et al.* 1997), statistical power poses a major concern for more classical quantitative traits such as yield that are likely to be determined by a larger number of loci of smaller effect.

LD mapping

In the hope of overcoming some of the limitations of QTL analysis, plant researchers have moved toward LD or association mapping as an additional means to identify genomic regions that contribute to phenotypes. The primary advantage of association approaches is that they can rely on population samples; there is no need for crosses and the production of large numbers of progeny. In addition to circumventing the need to develop crosses and mapping lines, a population sample may contain many more informative meioses (i.e., much more recombination) than a traditional QTL mapping population, thus allowing for increased mapping resolution. Though a relatively new approach, LD mapping has already shown promise (e.g. Breseghello and Sorrells 2006, Malysheva-Otto and Roder 2006, Thornsbery *et al*. 2001) and continued efforts will certainly increase its utility (Yu *et al*. 2008). Like QTL approaches, however, there are several features of experimental design that need to be carefully considered when undertaking LD mapping. First, distinguishing true associations from statistical noise requires large sample sizes, both for statistical power and to correct for multiple tests (Long and Langley 1999, Macdonald and Long 2004). Another design challenge is sample origin: geographic structure, interrelatedness due to shared pedigree, or other departures from simple randomly mating populations can result in spurious associations in which a genotype is associated with a particular geographic region or founder genotype rather than the phenotype of interest. This is especially problematic for phenotypes that vary by geographic region, such as flowering time or photoperiod sensitivity. The difficulties inherent to LD mapping are reflected in the literature, where false positives (Aranzana *et al.* 2005) and failure to reproduce associations are not uncommon (see discussion in McCarthy *et al.* 2008).

Further disadvantages – phenotype requirements and allele frequencies

A drawback to both QTL and LD mapping is that both are tied to a phenotype. Both methods assume that one knows *a priori* the phenotype of interest and can measure the phenotype accurately. For example, while certain components of yield can be readily measured (e.g., ear length), it is likely that many factors (e.g., drought resistance or tolerance of planting density) impact on yield in a given environment, and it is difficult *a priori* to identify all of these traits for phenotyping. And though one can also measure more complex phenotypes like yield per plot, such measurements are strongly affected by non-genetic factors and measurement error that negatively impact the ability to associate phenotype with genotype. Finally, both QTL and LD mapping can only associate phenotype to loci polymorphic in a particular population or cross – if

a functional allele is at very high (or very low) frequency, then it is unlikely to be polymorphic in parents of a QTL cross, and LD approaches may have limited power to identify associations.

Selection mapping to identify candidate agronomic loci

LD and QTL mapping take a "top-down" approach to identifying genes of interest, choosing a phenotype of interest and, in a single assay, associating marker alleles with an experimental measure of phenotype. In contrast, selection mapping uses a "bottom-up" approach, identifying loci that have been associated with an advantageous phenotype over a number of generations by scanning for the signal of selection – such as an increase in allele frequency – in a chronological comparison of marker data. Compared to QTL or LD mapping, selection mapping has a number of advantages for identifying loci associated with complex traits such as yield. First, selection mapping does not require measurement of a phenotype and thus avoids the associated experimental error. Second, by comparing progenitor and derived populations, selection mapping can identify alleles that are at very high (or low) frequency in extant populations and would be difficult to identify with other approaches. Third, like LD mapping, selection mapping does not require the time-consuming construction of mapping populations. Finally, simulations suggest selection mapping may be quite powerful for sample sizes much smaller than those needed for either QTL or LD mapping (Teshima *et al.* 2006), especially for recent, strong selection like modern crop improvement (e.g. Palaisa *et al.* 2003, Olsen *et al.* 2006).

Although the method is relatively new, selection mapping has already been applied to identifying loci of interest in *Arabidopsis* (Toomajian *et al.* 2006) and a number of crops, including sunflower (Chapman *et al.* 2008), sorghum (Casa *et al.* 2005), and rice (Olsen *et al.* 2006). In one particularly relevant example, Pswarayi *et al.* (2008) performed QTL mapping in both legacy and modern barley lines, followed by selection mapping between legacy and modern lines. Of 11 yield-associated QTL found to change frequency significantly over time, 9 showed an increase suggestive of positive selection.

Examples of selection mapping in maize

Several studies have applied a limited selection mapping approach to identify loci of interest in maize. In one of the first studies to apply selection mapping to any crop, Stuber and Moll (1972) and Stuber *et al.* (1980) tracked the change in allele frequencies at 8 allozyme loci during 10 cycles of selection for increased yield. After identifying alleles that increased in frequency in response to selection, they investigated the efficiency of using their alleles in MAS for increased yield (Stuber *et al.* 1982). In spite of the extremely small number of markers used, Stuber *et al.* (1982) reported gains from MAS equivalent to 1.5-2 cycles of selection on yield alone.

Selection mapping has also been used to identify a number of maize loci of agronomic interest, including loci associated with oil content in the Illinois long-term selection experiment (Sughroe and Rocheford 1994), endosperm color (Palaisa *et al.* 2003), domestication-related traits (Vigouroux *et al.* 2002, Vigouroux *et al.* 2005), and quantitative disease resistance (Wisser *et al.* 2008). In the largest selection mapping project to date, resequencing data from >1,000 loci were used to identify genomic regions associated with domestication (Wright *et al.* 2005) and improvement (Yamasaki *et al.* 2005, Yamasaki *et al.* 2008). The Wright *et al.* study alone yielded a more comprehensive list of candidate loci than all previous QTL analyses combined. The study also suggested that as many as 4% of loci across the maize genome have undergone

selection during domestication and improvement. Importantly, these authors demonstrated that their selection mapping approach had a reasonable false positive rate (Yamasaki *et al.* 2008), and that the loci they identify statistically associate with known QTL (Wright *et al.* 2005). Perhaps most significantly, further analysis revealed that the candidate loci identified by selection mapping are over-expressed in the maize ear, likely the tissue most changed during domestication and improvement (Hufford *et al.* 2007).

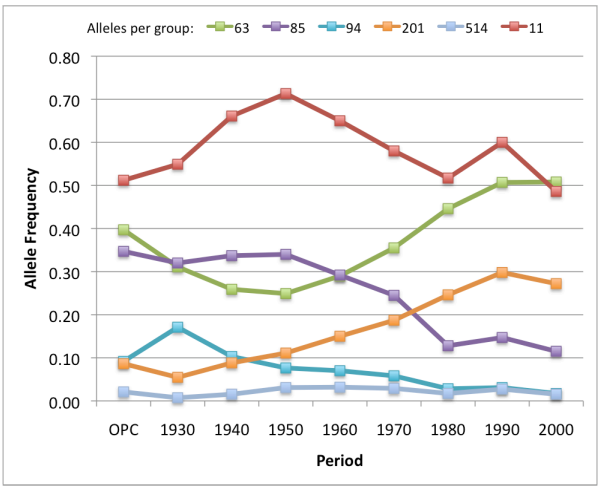

Figure 4. Change in allele frequency over time in maize. Duvick *et al.* (2004) clustered alleles at 98 SSR loci into groups based on patterns of frequency change, concluding that observed patterns of change (especially in the green and orange groups) are consistent with specific genomic regions having contributed to breeding success. OPC = open-pollinated crosses.

Finally, a pair of recent studies has made use of chronological sampling of popular maize lines not dissimilar to that proposed in this grant. Duvick *et al.* (2004) measured the change in frequency of alleles at 98 microsatellite loci in a stratified sample covering 70 years of maize breeding (Fig. 4). While these authors did not perform a selection mapping analysis on individual loci or correct for relatedness among lines, their results do show significant changes in allele frequencies over time, a pattern they argue is evidence that multiple genomic regions have contributed differentially to the breeding success of modern lines. In a follow-up analysis, Feng *et al.* (2006) utilized 361 simple sequence repeat (SSR) loci to compare allele frequencies between samples of early 20th century and modern inbred lines. Using pedigree information, they identified 24 SSR alleles that may have been targeted by selection during maize breeding. Notably, many of these alleles were unique to one or the other of the two heterotic groups (stiff-stalk, non-stiff-stalk) of modern inbred lines sampled.

Clearly, there are a number of advantages to selection mapping as a method for identifying loci of agronomic interest, and selection mapping has been shown to be an effective methodology in maize. We propose to take advantage of recent developments in genotyping technologies to perform a dense, genome-wide scan for selection across a large sample of chronologically stratified maize lines.

APPROACH

The proposed project has two major objectives aimed at identifying agronomically important alleles for use in marker-assisted selection in maize.

Objective 1: Identify candidate agronomic loci (CAL) by scanning the maize genome for alleles that have increased in frequency due to artificial selection

Objective 2: Test breeding utility of identified CAL in experimental yield trials.

Objective 1: Identify CAL

We define a candidate agronomic locus (CAL) as a marker-tagged region of the genome that has been, or is tightly linked to, the target of selection during maize evolution. Because our goal is

to provide marker loci for use in marker-assisted-selection, the identity and function of individual CAL, though obviously of interest, is outside the scope of this project. To identify CAL, we propose to harness genome-wide SNP genotyping in a large panel of chronologically stratified samples of *Zea mays* (Fig. 3, Table 1). We will develop statistical methodologies that account for the unusual nature and complex relatedness patterns of samples from breeding lines, using these methods to identify alleles that have increased in frequency due to artificial selection in each of three time periods during maize improvement.

Sample selection

We will genotype a total of 444 samples (Table 1), including 432 samples stratified across four chronological periods: 96 teosintes (*Z. mays* ssp. *parviglumis*) representing the wild ancestor of domesticated maize; 96 landraces, open-pollinated varieties cultivated until the use of hybrid maize in the 1930's; 96 legacy inbred lines, representing the first public inbred lines for breeding; and 144 modern inbreds, including both public and off-patent proprietary inbred lines from the late 20th century. A variety of teosinte lines will be chosen to represent the geographical distribution of *Z. mays* ssp. *parviglumis*. To increase the utility of our marker data, the majority of teosinte accessions sampled will be accessions included in previous association studies (Briggs *et al.* 2007, Weber *et al.* 2007, Weber *et al.* 2008), for which seed is publicly available. Landrace and inbred lines will be chosen to be both representative of the diversity of maize lineages and of direct interest to modern breeding. We will use results from literature analyses (e.g. Vigouroux *et al.* 2008), pedigree information (Gerdes *et al.* 1993, Smith *et al.* 2004, Smith *et al.* 2006), and available marker data (www.panzea.org) to help choose lines based on diversity analyses. We will additionally consult with breeders (e.g. letter of support from Dr. Anand Pandravada) to identify lines of most interest for modern breeding efforts. The postdoctoral associate, Dr. van Heerwaarden, is currently working on software for the identification of core diversity panels in maize; it is expected that a modified version of this software will be useful in choosing lines for analysis here. Importantly, in our selection of 144 modern inbreds, we will split the lines roughly in half, sampling from the two major heterotic groups used in modern breeding: stiff stalk and non-stiff stalk maize. Because breeding and selection has been largely independent in these two groups, this sampling strategy will allow identification of CAL important in only one of the heterotic groups as well as loci important in modern maize as a whole. We will also track the breeding program of origin and relative maturity group for each sample, as selection regimes may have differed among breeding programs and across geographic regions.

Type	Number
Teosinte	96
Landraces	96
Legacy Inbreds	96
Modern Inbreds	144
Test Crosses	12
Total	**444**

Table 1. Sampling scheme for genotyping. NAM founder lines will not be genotyped; data for these lines are

In addition to the 432 samples we will genotype to identify CAL, we will also include the founders of the nested association mapping (NAM) population (Yu *et al.* 2008), for which resequencing data including all or most of the polymorphisms genotyped here is publicly available (http://www.maizegenetics.net/maize-hap-map). Finally, to ensure that single nucleotide polymorphisms (SNPs) genotyped in the panel are inherited in a Mendelian manner (i.e., are not polymorphisms between paralogous loci) we will genotype six hybrid offspring from each of two crosses between inbreds in our panel. These parent-offspring trios provide the potential to detect SNPs subject to non-Mendelian patterns of transmission that would be much less useful for MAS.

Genotyping

Genotyping Platform

We propose to utilize the Infinium HD genotyping platform (Illumina Inc.), a chip-based technology that can efficiently and inexpensively genotype 60-200,000 single nucleotide polymorphisms (SNPs) in 12 samples (www.illumina.com). Porcine, equine, and several human versions of these chips are already commercially available. As part of a consortium of interested stakeholders, we are currently working with representatives from Illumina to design an Infinium HD chip for maize, scheduled to be released in the Summer of 2009. The final number and selection of SNPs has not yet been determined (see discussion in budget justification): an array of 120,000 SNPs from predominantly genic regions of the genome seems the most likely scenario, but we note that even the most sparse array of 60,000 SNPs could place approximately two polymorphic markers in each of the annotated genes in the maize genome. Importantly, most or all of the SNPs used in the development of the chip will come from panels of extant inbred lines. This selection approach has enormous advantages from a breeding perspective, as it will only sample alleles polymorphic in inbred material. Unlike alleles found only in wild relatives or exotic material, CAL identified using this platform can thus be immediately incorporated into breeding programs without extensive backcrossing and with reduced potential for linkage drag.

Data collection and curation

Samples will be genotyped at the UC Davis Genome Center DNA Technologies Core, which has extensive prior experience genotyping with Infinium Chips and a robotic high-throughput system for optimal genotyping efficiency (see letter of support from Dr. Charles Nicolet). The Genome Center provides a bioinformatics pipeline to infer genotypes from the raw Illumina data, which, along with the raw data, are sent directly to the customer. The process of inferring genotypes is not without error, however, as most software operates under the assumption that genotypes are the result of random mating. This is an important concern for samples of inbred lines that likely deviate significantly from the Hardy-Weinberg genotypic frequency expected under random mating. We will utilize both custom Perl programs and published error-detection software (Toleno *et al.* 2007) to identify loci that are potentially problematic, and each of these loci will then be manually inspected to make genotype calls. To ensure the highest degree of data quality, we have budgeted up to 500 hours of undergraduate research assistance in manually checking the genotype calls; problematic loci will be reviewed individually by the PI or postdoctoral associate.

Statistical methods to identify CAL

Both population genetic and pedigree methods have been used to identify loci of interest in selection mapping experiments. There are several features of maize breeding populations, however, that make these methods less than ideal to identify CAL in this study.

Identifying selection

A number of authors have developed approaches for detecting selection in genome-wide scans of genetic marker variability (e.g. Sabeti *et al.* 2002, Schlötterer 2002, Tang *et al.* 2007), and others have developed statistical models of allele frequency change under artificial selection (Waples 1989, Keightley and Bulfield 1993, Turelli and Barton 1994, Kim and Stephan 1999, Nuzhdin et al 2007). Most of these approaches, however, assume either large, constant-size populations, or

are specific to frequency change in controlled mapping populations. Most additionally assume that alleles are drawn at random from the population and are thus representative of the true distribution of allele frequencies in the sampled population. Modern maize varieties are not random samples from a panmictic population, but exhibit complex patterns of relatedness due to shared pedigree and founder effects during decades of breeding (Smith *et al.* 2004). Moreover, breeding populations differ in terms of size, crossing scheme, and number of generations of selection. Finally, many genotyping technologies (including the one utilized in this proposal) only genotype alleles previously ascertained to be polymorphic in a SNP discovery panel, thus violating model assumptions about the distribution of allele frequencies (Nielsen *et al.* 2004, Clark *et al.* 2005). These model violations may lead to erroneous conclusions; particularly worrisome is the likelihood of an increased false positive rate in identifying loci of interest – one recent method suggests a false positive rate of up to 90% when equilibrium assumptions are violated (Jensen *et al.* 2005).

Pedigree approaches

One alternative to methods based on explicit population genetic models has been to make use of information on pedigrees relating lines of interest to their founders (Feng *et al.* 2006). This approach relies on sampling founder lines and then using pedigree information to estimate the expected frequency of an allele in a sample of modern lines of interest. The pedigree approach has several shortcomings for detecting CAL. First, the approach assumes complete knowledge of pedigree for the lines of interest as well as the ability to sample specific founding lines – requirements that are unlikely to be met for a broad sampling of modern lines, and may be difficult to meet for many individual lines of interest. Second, pedigree data, when present, are never a perfect representation of genetic parentage (Bernardo 1993, Crepieux *et al.* 2004), and ignore factors such as genetic drift or recurrent selection within purportedly inbred lines (Nelson *et al.* 2008). Third, the analysis presented often assumes genetic markers are independent – while this works well for a small number of markers, for truly genome-wide datasets such as ours, linkage disequilibrium between markers may lead to spurious identification of CAL. Fourth, a meaningful amount of genetic diversity exists even within inbred lines, such that genotyping a single individual may misrepresent the allelic makeup of a line as much as 10% of the time (Gethi *et al.* 2002). Finally, SNP-based estimates of relatedness can be more accurate than pedigree (Zhao *et al.* 2007) and have been demonstrated to outperform pedigrees when predicting the breeding value of an individual (Hayes *et al.* 2009). Most notably, maize yield has specifically been shown to be more strongly correlated with genetic differentiation at marker loci than with pedigree (Smith *et al.* 1990).

A general approach

Appropriate methods to identify CAL must go beyond conventional approaches that fail to account for ascertainment bias, demographic (breeding) history, founder effects, and errors in haplotype estimation. We propose to develop a more general statistical method to identify loci that have increased in frequency across chronologically stratified samples of lines. As in most selection mapping methodologies, we wish to identify alleles that show evidence of selection – in this case, alleles that have increased in frequency in derived groups compared to their progenitor lines. Most selection mapping approaches, however, do not allow incorporation of the complex patterns of relatedness inherent in samples from breeding programs. Recent advances in association methods (Yu *et al.* 2005, Price *et al.* 2006) have provided new means to correct for population structure and relatedness. However, these methods were designed to detect

association between single markers and extant phenotypes, and were not meant to detect the multi-locus effects caused by directional selection over many generations. We therefore propose to develop a more general approach that combines features of selection mapping and association methods. Essentially, we will follow an adjusted association mapping scheme in which we test for significant relationships between SNP loci and a quantitative phenotype. In our method, however, "phenotype" is an errorless ordinal variable with levels defined by the derived states of the different lines (i.e. newness). Using this approach, we will be able to detect selected SNPs as those that show a significant change in frequency between progenitor and derived populations (i.e. are associated with "newness"), while correcting for both population structure and relatedness. We will enhance this approach by incorporating information on adjacent SNPs in order to harness information on changes in the site frequency spectrum and LD at linked sites. Genotype data will be analyzed using a sliding window to take account of the fact that patterns of relatedness due to founder effects or breeding may differ throughout the genome. Because our dense genotype data allow for extremely precise estimates of relatedness and structure, this combined approach should resolve many of the problems with both pedigree-based and population-genetic methods, obviating the need for assumptions about the number of generations, the distribution of allele frequencies, ascertainment schemes, patterns of intermating, population size, or historical contribution of individual lines.

Validating the method

Analysis of simulated data

As a first test of our statistical methodology, we will assess the power and accuracy of our methodology by analyzing simulated data sets with known selection regimes. We will utilize forward population genetic simulation to create idealized data sets in which pedigree and selection are known exactly. We will test a wide variety of situations including varying selection intensity, relatedness among modern lines, founder effects, and linkage. Such tests will identify scenarios in which our method behaves sub-optimally, allowing us to modify the method appropriately. We note that both Dr. van Heerwaarden and Dr. Ross-Ibarra have prior experience writing forward simulations (Hollister *et al.* 2009, Piñeyro-Nelson *et al.* 2009, van Heerwaarden *et al.* 2009).

Comparison to other genome scan methods

Once we have analyzed our genotype data, another measure of the utility of our approach is to compare results from our method to those of other genome scan approaches to detect selection. Specifically, we will apply a comparison of historical to actual recombination rates (O'reilly *et al.* 2008) and a SNP-based assessment of the site frequency spectrum (Nielsen *et al.* 2004, Nielsen *et al.* 2005) to scan the data for evidence of selection. We suspect both approaches will identify numerous regions suggestive of selection, including many or most of the CAL identified by our methodology. But because maize breeding populations violate several of the assumptions of these models (see above), we expect that both will identify a number of false positives as well. Comparison among of these results to our own will nonetheless help verify the CAL we describe, and potentially identify cases in which our methodology is too conservative.

Comparison to other QTL

Our proposed selection mapping experiment will likely identify a large number of CAL. Given previous estimates that as many as 4% of maize loci have been the targets of selection during

domestication or improvement (Wright *et al.* 2005), we might expect that 30,000 loci * 4% = 1,200 loci have been the targets of selection. As further validation of our method, we will compare our proposed set of CAL to a comprehensive list of yield-associated loci culled from the QTL and LD mapping literature in maize and related crops. Because the SNPs included in our genotyping array will be genetically and physically mapped to the maize genome, we can directly compare the overlap between CAL we identify and those from the maize, sorghum, and perhaps rice genomes. Evidence that the list of CAL we identify is enriched for previously identified loci for yield or other agronomic traits will provide strong support for the effectiveness of our mapping approach. Nonetheless, because of the genome-wide scope of our project and our use of a mapping approach that does not rely on the direct measurement of phenotype, we will undoubtedly identify many CAL not previously found in the literature.

Software and chip development

In order to make the tools developed here broadly available, we will develop software and genotyping resources for use by breeders or other researchers. We will write software for the analysis of genotypic data using the novel statistical methods developed here, making the program and source code freely available under an open-access license. This will allow breeders or other researchers to quickly and easily apply our methodology to other samples of chronological lines, or to modify and extend our methods. Moreover, to facilitate the utility of the list of CAL we identify, we will select the most promising CAL from our statistical and field analyses (see below), keeping only those loci that also provided clean data with few technical problems. We will make this shortened list of CAL, along with primer information, readily available as a package for development of low-density genotyping chips such as Illumina's Golden Gate assay. The availability of a smaller set of CAL pre-screened for utility in a broad set of lines will enable breeders to quickly and cheaply genotype lines for a set of loci of interest without having to develop genotyping assays or deal with the complex bioinformatics of large genomic datasets.

Objective 2: Experimentally test CAL

The first objective of this proposal will produce a list of CAL. This list alone is a useful tool for plant breeders, as it will be enriched for loci that have undergone selection during maize improvement. However, a list of CAL is of much greater utility if breeders have reason to believe such loci are likely to produce measurable results in terms of yield gain. Although we compare our method to simulation results and previous lists of identified loci, demonstrating the association of our list of CAL with yield remains the ultimate validation of our approach. Extensive testing across multiple generations and environments is outside of the scope of this proposal; we will focus on yield trials sufficient to provide initial validation both of our mapping method and the utility of the CAL we identify. We propose a two-year field trial, including nearly half of our lines, to test the association between CAL and yield.

Field testing facilities

All of the plant growth and field trials will be performed at the nearby Monsanto field experiment station in Woodland, CA. (see letter of support from Dr. Anand Pandravada). The ca. 200 acre field station is designed for yield trials, and has all of the required infrastructure, including mechanized planting, harvesting, shelling, irrigation, and extensive nursery and

greenhouse space. In addition to field and nursery space, Dr. Pandravada has graciously offered assistance in performing crosses, seed increases, and gathering yield data for our trials.

Seed increase and initial testing

In the first season, we will plant short observation rows of 13 plants for all of our domesticated lines (landraces, legacy inbreds, and modern inbreds). This will insure that our field trials include maize lines that grow well at the Woodland field site, with appropriate flowering time, water use efficiency, etc. Basic qualitative phenotypic data on growth, flowering, and yield will be used to rule out lines which are not suitable for further field testing.

Crosses and hybrid yield trial

Based on our initial observations, we will choose ~200 lines for further field testing. Each of these lines will be crossed to one of two standard tester lines (male or female) included in our sampling (one likely scenario would be to use the common inbred lines B73 and Mo17 as testers). The resulting 200 hybrid lines will be planted in the field in year two. Each hybrid will be planted in 3 replicate plots of 2 rows (44 plants per row) in a randomized block design. Yield will be estimated for each replicate based on the dry weight of the kernels produced.

Association analysis and line selection

The results from the yield trial will be used to choose 3 groups of lines for additional testing in year 3. The first group of 12 lines will be chosen by total yield alone. The second and third group of lines will be chosen on the basis of an association analysis. In each case a set of marker loci will be used to perform a standard single-marker association analysis (such as an F-test) of the yield data from year 2, and a group of 12 lines will be chosen to maximize representation of alleles that associate with yield. Group 2 will use a set of markers chosen from our list of CAL for the association, and group 3 will use a random set of SNPs from the original genotyping array.

Crosses and elite hybrid yield trial

The association analysis in year 2 will result in at most 36 lines, depending on the degree of overlap among the 3 groups of lines. Additional modern inbred lines will be chosen randomly to bring the total number of lines for year 3 to ~100. Each lines will then be crossed to the same tester used in year 1 in addition to two new tester lines, such that the 3 tester lines represent 1 landrace, 1 legacy inbred, and 1 modern inbred. These hybrids will be planted in the field (as per year 2, but with 2-3 replicates depending on space and budgetary constraints) in year 3.

Validation of CAL and test of combining ability

If our CAL provide an added benefit for breeding programs, then the group of lines chosen using the CAL for association analysis in year 2 should outperform both of the other groups of lines in terms of yield in year 3. Additionally, year 3 data will allow testing of the importance of combining ability. General combining ability, or the average performance of a line in hybrid crosses (Sprague and Tatum 1942), is an important trait in breeding programs, and has likely been under selection during maize improvement. If our CAL reflect this selection, then the lines chosen using our CAL should show greater general combinability than lines chosen using either random SNPs or yield alone.

Due to the limited scope of our field trials, not all of the CAL we identify may be tested, and not all of the CAL tested may significantly associate with yield in the lines used or in the particular

environment of our field trial. Because of the quantitative nature of yield, not all CAL may be important in every line or environment. If, however, our CAL predict yield more accurately than random SNPs or phenotype alone, we will have demonstrated the utility of our method for breeding and MAS.

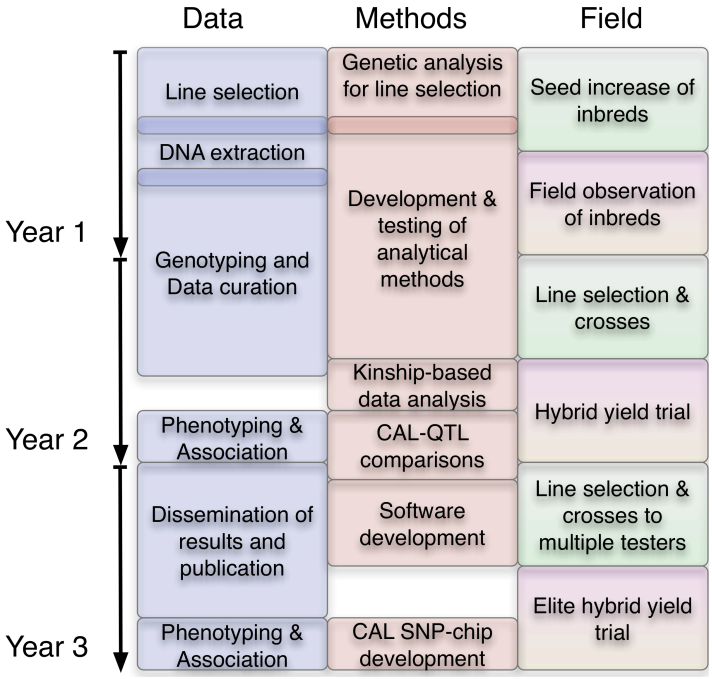

Figure 5. Project Timeline

Research Timeline

A timeline of the proposed research is presented in graphical format in Figure 5, and the division of labor is shown in Table 2. In the first year, Dr. Ross-Ibarra will work closely with Dr. van Heerwaarden on the development of a robust statistical framework for the analysis of genotype data, along the lines described above. Both will also work on the preliminary selection of lines based on public genotype data, inclusion in other studies or association panels, and availability. Dr. Pandravada at Monsanto will provide advice on lines of particular interest to breeding programs. Dr. van Heerwaarden will perform DNA extraction for the lines to be genotyped, and DNA will be sent to the Genome Center at UC Davis for genotyping at the end of the first year. Seed increase of all lines will be performed at the Monsanto field station in Woodland, followed by planting an observation row of each line the summer of the first year.

In year two, as the genotyping data is received, we will curate the data, checking genotype calls, and develop a database for rapid and efficient use of the data. We will also continue working on method development, testing the method against simulated data and other published genome scans. By the end of year 2 we plan to have completed the statistical analysis of the genotype data and comparison to lists of QTL from the literature. In year two the best lines from the observation rows will be crossed to a single tester (male or female), and the F1 hybrids planted in the field for the first yield trial.

In the final year of funding we will focus on publishing our methods and statistical software, as well as the results of the genotype analyses, including analysis of models of maize evolution, patterns of LD, and the list of identified CAL. Year 3 will also see the final yield trial and validation of the breeding utility of the CAL, as well as the choice of CAL for development of low density genotyping chips.

BROADER IMPACTS

Training

The proposed project will help train students in both bioinformatics and plant breeding. The graduate student employed by the grant will work closely with both Dr. Ross-Ibarra and Dr.

	PI	postdoctoral associate	graduate student	student workers	Woodland Field Station
Field					
Initial line selection	X	X			X
Seed increase			X		X
Field observation			X		X
Line selection for crosses	X		X		X
Crosses			X	X	X
Association analysis	X	X	X		
Phenotyping/yield measurement			X	X	X
Selection Mapping					
DNA extraction		X			
Genotyping	X				
Statistical methods	X	X			
Data curation		X		X	
Data analysis	X	X			
Comparison to QTL		X			
SNP-chip design	X				

Table 2. Participant responsibilities

Anand Pandravada at Monsanto, learning the practical and analytical aspects of field testing, yield trials, selection mapping, and association mapping. Undergraduate workers at the field station will be trained in cross-pollination and other breeding practices. Undergraduate students working on genotype calls will be encouraged to participate as well in undergraduate research coursework, in which they will work under the guidance of Dr. Ross-Ibarra or the postdoctoral associate to develop a research project in line with the goals of the overall program. Dr. Ross-Ibarra has a good history of undergraduate mentorship, including a current mentorship underway developing Perl software to analyze SNP data in maize.

Benefits to maize breeding and genomics

List of CAL

The main product of our proposal will be a list of candidate agronomic loci (CAL) for yield in maize. This list of SNPs, their context sequences, and their position on the physical and genetic map, will be made available in publications and via the Internet. The primary purpose of such a list is to allow other breeders to jump directly to Objective 2 of our project – utilizing the CAL discovered here as markers in yield trials for their crosses of interest. This should save time, money, and augment the power of MAS for increased yield.

SNP Genotyping Chip

The list of CAL will be of use to breeders or researchers who wish to pursue individual loci or create their own genotyping procedure. But for many breeders, the availability of a pre-designed genotyping chip would save significant time, effort, and cost. We estimate that the cost to genotype 384 CAL in 30 lines using the Illumina Golden Gate Assay, for example, would be only ~$3,500. We will choose a set of CAL based on statistical and field validation as well as technical ease of use for genotyping. We will provide as a package the primer and other information necessary for the development of such a chip, allowing other workers to simply transfer this information to industry or genotyping centers for chip design and processing, saving the time and cost of developing a genotyping assay from scratch.

Relatedness data

Genome-wide SNP data will allow much more accurate estimates of relatedness among our lines than currently available from molecular data or pedigrees. These relatedness measures are directly useful to breeders, and can be incorporated into association analyses. In order to

facilitate use of this data, in addition to the genotypes themselves, we will also make the relatedness matrix of all lines publicly available.

Maize genomics

Our study brings together a unique combination of large sample size with whole-genome diversity data. This combination allows for a number of descriptions of the maize genome not feasible with smaller data sets, including, for example, copy number variation identified by screening for groups of SNPs that are over- or under-represented in specific lines, fine-scale patterns of LD and historical recombination, and genome-wide identification of SNPs useful in differentiation alleles or lines or the development of other markers such as RFLPs.

Dissemination of data and methods

We will disseminate our method and results widely, both in presentations at conferences and in peer-reviewed publication. In addition to publishing our results, we will make our list of CAL (and related information including map position, primers, etc.), the raw genotype data, and the source code for our analysis all publicly available via the Internet.

Software development

The statistical methodologies for line selection and identification of CAL will be developed as a software package made publicly available via the Internet. We will provide documentation and support for the software, making the methods developed here as broadly available as possible to breeders and other researchers. Both Dr. Ross-Ibarra and Dr. van Heerwaarden have established a history of making their code publicly available. Current software from both can be found at the Ross-Ibarra lab website, http://www.rilab.org/.

Application to other breeding programs

The present proposal is focused on maize, but the analytical and technical approaches (and software) developed here should be appropriate to similar genotype-based studies in any breeding program. By developing a general approach to account for relatedness and population structure among the lines sampled, our statistical approach should be utilizable in virtually any breeding program for which genome-wide genotyping is feasible. Importantly, because we are taking a general approach to selection mapping, our methods will be equally applicable to taxa for which pedigree information is missing or inadequate.