Scanning the weather: high-throughput discovery of agronomic loci for advanced maize breeding to address climate change

RATIONALE AND SIGNIFICANCE

Importance of maize

Maize is a natural resource of fundamental national importance, vital for food, livestock feed, and fuel production. Maize is by far the most valuable agricultural crop in the United States: in 2010, annual production in the U.S. was worth more than $66 billion dollars (USDA 2011), or ~60% of the value of domestic crude oil production in 2009 (EIA 2010) and nearly double the value of soybean, the second most important field crop ($39 billion, USDA 2011). 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, 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: from 1999-2009, 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 of course be won by improved farming practices, but the majority of yield increase can be directly attributed to breeding efforts (Duvick 1992, 2005), and breeding must remain of central importance in order to meet increased yield demands.

Climate change

Changing climatic conditions pose a serious threat to our ability to continue increasing maize yield. Historical analyses suggests that climate change over the last 30 years has already dramatically impacted maize yields worldwide, retarding the gains from breeding and management (Lobell *et al.* 2011a). While these losses have been comparatively mild in the temperate climate of the United States, the same models suggest that change of even 1ºC could negatively impact yields by as much as 5-17% (Lobell *et al.* 2003, Lobell *et al.* 2011a). Some models even predict that U.S. maize yields could be 30-46% below current levels by the end of the century (Schlenker and Roberts 2009). Given predictions for long-term changes in mean temperature (Fig. 1 A-C), it is reasonable to conclude that the impacts of climate change on maize yield could translate into economic losses of billions of dollars. Moreover, these estimates do not take into account effects of the likely yield loss in the rest of the world (Lobell *et al.* 2011b) on prices and supply. Substantial efforts will be needed to preserve US maize production and increase or maintain yields, and in particular breeding efforts to adapt US maize to a warming climate must be a priority (Troyer 2004).

While future climates will undoubtedly differ in a number of important ways, we focus in this proposal on adaptation to changing temperature. Neither of the other major changes predicted as a result of climate change — elevated CO_2 and changes in rainfall — are likely to be as important for maize yields. Though elevated CO_2 may actually prove beneficial for yields of some crops, models suggest the effect on maize, a C4 plant, to be relatively minor (Lobell *et al.* 2011a). Similarly, analyses of past climate change find significant negative effects of

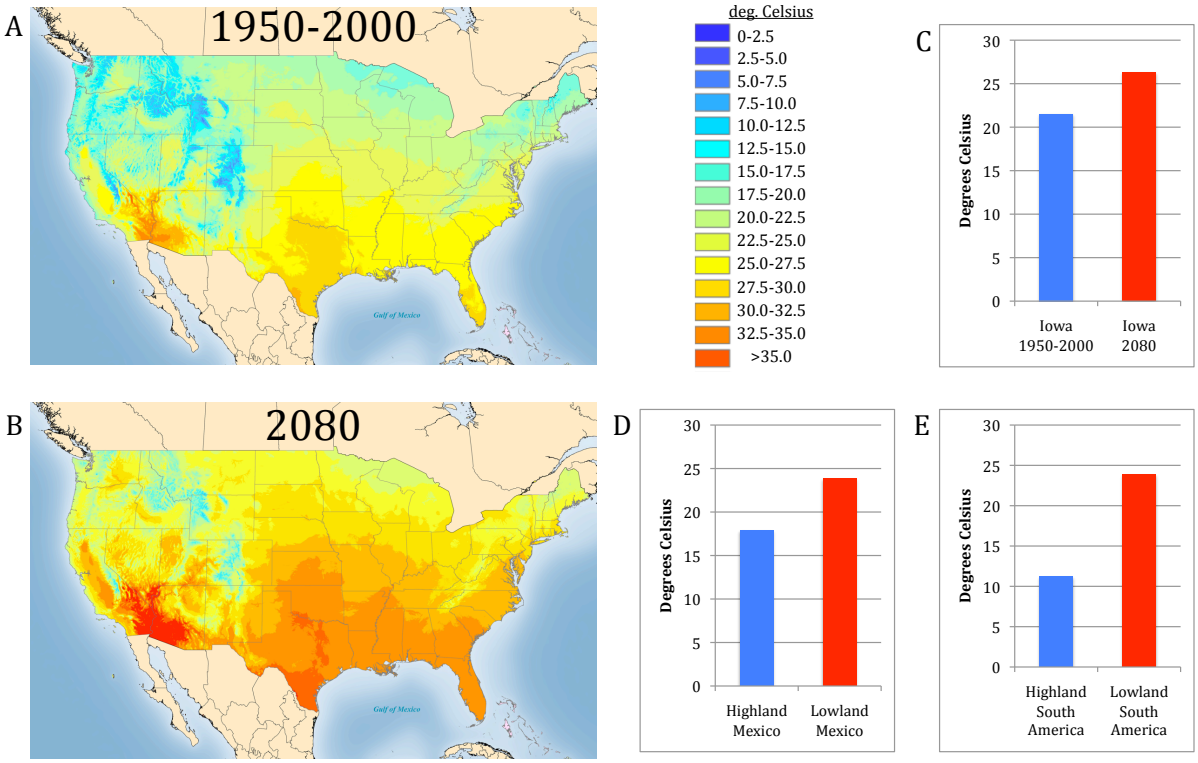

Figure 1. Climate comparisons. **A)** Mean August temperatures representative of the period 1950-2000 (data source: worldclim.org, Hijmans *et al.* 2005). **B)** Predicted mean August temperatures for 2080 (SRES A1 Emission Scenario, CCCMA-CGCM31 Climate Model, data source: CIAT, ccafs-climate.org). **C-E)** Mean current and predicted August temperatures in central Iowa **(C)** and mean current August temperatures at landrace sampling locations in Mexico **(D)** and South America **(E)**.

temperature on maize yield but much weaker effects of precipitation (Lobell *et al.* 2003, Roberts and Schlenker 2009, 2010). Increasing temperature, on the other hand, is likely to have strong, nonlinear effects on yield (Lobell *et al.* 2011b, Schlenker and Roberts 2009), as temperature extremes have an inordinate impact on overall yield.

Improved breeding methods

Novel approaches to molecular plant breeding have been proposed as a solution to changing climes (Takeda and Matsuoka 2008), and some models suggest that breeding or other technological advances could meaningfully mitigate yield loss (Li *et al.* 2011). Both marker-assisted selection (MAS) and transgenic approaches have become important tools of modern plant breeding, allowing breeders to combine traits of interest without the need for additional, costly phenotyping, pedigree analysis, or (in the case of MAS) a detailed functional understanding of the molecular basis of a trait. The success of both approaches, however, depend on our ability to identify useful markers and alleles. Genetic mapping has been effective at identifying markers for use in MAS and candidates for transgenic methods. But traditional mapping approaches such as quantitative trait locus (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 to extend our previous work on selection mapping for yield to identify candidate agronomic loci (CAL) that will prove useful for molecular breeding approaches to adapt maize to changing climates. Our approach takes advantage of several millennia of adaptation of traditional maize varieties to different climate regimes, natural experiments of grand scale, using selection mapping methods to identify the targets of selection in the independent adaptation of maize to highland environments in both Mexico and South America. Importantly, differences between highland and lowland environments mirror the predicted effects of future climate change in the US (Fig. 1 C-E).

The resulting list of CAL, along with the bioinformatic tools we will develop, will provide new opportunities for molecular breeding, accelerating the progress of adaptation to climate change and thus ensuring continued yield increases. Moreover, thanks to rapid advances in sequencing technology, the selection mapping approach proposed here should be easily extendible to virtually any crop species with sufficient germplasm resources; a modified method would be possible even for crops without a reference genome. We therefore expect that our approach will prove to be an important advance in breeding methods for several crop species.

INTRODUCTION

Maize diffusion and adaptation

Despite its humble beginnings as a wild grass, maize spread rapidly across the globe to become the world's top-producing crop. Wild members of the genus *Zea*, which diverged relatively recently (Ross-Ibarra *et al.* 2009), are endemic to a small region stretching from northern Mexico to Central America. The direct wild ancestor of maize, *Zea mays* ssp. *parviglumis*, occurs only along the low-elevation slopes of the Sierra Madre Occidental in the southwestern corner of Mexico (Sánchez-González and Ruiz-Corral 1997). After its domestication from ssp. *parviglumis* ~9,000 before present (BP; Matsuoka *et al.* 2002, Piperno *et al.* 2009), maize spread from the lowlands of Southwest Mexico (van Heerwaarden *et al.* 2011), rapidly diffusing across the Americas. By ~6,000 BP, maize had adapted to the high elevations of central Mexico and spread to the lowlands of South America (Piperno 2006); by 4,000 BP maize was being grown at high altitudes in the Andes (Perry *et al.* 2006). After European contact with the New World, maize continued its world-wide diffusion, spreading quickly across Europe and subsequently to Asia and Africa. Maize is now the world's most broadly cultivated crop, currently grown on six continents in 162 countries and territories (FAOSTAT, 2009) in a distribution spanning 90° of latitude (from Chile to Canada) and more than 3000m of elevation (Tenaillon and Charcosset 2011). During expansion to such varied regions, maize encountered — and adapted to — extremes of temperature, day length, precipitation, and soil types (Troyer 2004).

There is clearly tremendous potential to harness the allelic variation present among maize populations for modern breeding, taking advantage of genotypes molded by selection to adapt cultivated maize to new and changing environments. In particular, the fact that maize has

independently adapted multiple times to similarly extreme environments (e.g., the highlands of the Mexican Central Plateau and the Andes) presents a unique opportunity to identify adaptive loci for use in maize breeding. While these environments represent cooler, rather than warmer climates, identifying the loci underlying such adaptation in multiple populations should nonetheless be a powerful approach to finding CAL (see Approach below).

Genetics of adaptation

While variation in several well-known maize phenotypes (branching, glume architecture, endosperm color) is largely controlled by single loci (*tb1, tga1,* and *y1* respectively), most complex traits have been found to be highly polygenic. For instance, recent genome-wide association studies (GWAS) of leaf morphology, flowering time, and disease resistance have all found tens of QTL of small effect (Buckler *et al.* 2009, Tian *et al.* 2011, Kump *et al.* 2011). These authors suggested that the outcrossing mating system of maize may explain the highly quantitative genetic architecture of adaptive traits in contrast to selfing species like rice and *Arabidopsis* where fewer loci are involved. Adaptation to varying temperature during maize diffusion post-domestication is thus likely to have also involved a diffuse genetic architecture. For example, tassel blasting and leaf firing, traits related to heat tolerance, have previously been shown to be polygenic in nature (Frova and Sari-Gorla 1994, Bai 2003) and temperature is known to affect many aspects of plant growth and development (e.g., seed germination, photosynthesis, respiration; Wahid *et al.* 2007).

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 QTL and association or Linkage Disequilibrium (LD) mapping (Ross-Ibarra *et al.* 2007).

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 (typically limited in resolution to 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. A more serious limitation 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.* 2006); 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).

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 population samples usually contain many more informative meioses (i.e., much more recombination) than a traditional QTL mapping population, thus allowing for increased mapping resolution. LD mapping has already proven to be a powerful approach in maize (Buckler *et al.* 2009, Kump *et al.* 2011, Tian *et al.* 2011) and other crops (Breseghello and Sorrells 2006, Huang *et al.* 2010). Like QTL approaches, however, there are challenges to LD mapping. First, distinguishing true associations from statistical noise requires large sample sizes, both for statistical power in detecting associations and when correcting 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 results for genotypes 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 temperature, flowering time or photoperiod sensitivity. This latter difficulty is well 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).

Shared disadvantages – phenotype requirements and allele frequencies

A limitation 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 it accurately. For complex traits such as yield or temperature adaptation, though certain components could be readily measured (e.g., ear length or germination at higher temperatures), it is likely that many factors (e.g., drought resistance, growth rate, or photoperiod) may play important roles, 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

LD and QTL mapping take a “top-down” approach to identifying genes of interest, choosing a phenotype of interest and 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 signals

of selection – such as a difference in allele frequency among populations or the loss of diversity and increased LD around a selected site. Compared to QTL or LD mapping, selection mapping has a number of advantages for identifying loci associated with complex traits. First, selection mapping does not require measurement of a phenotype and thus avoids the associated experimental error. Second, using comparisons of multiple populations, selection mapping can identify alleles that are at very high (or low) frequency which 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 the relatively recent, strong selection we might expect for adaptation to new environments.

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 (Molina *et al.* 2011, He *et al.* 2011).

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 first large-scale efforts at selection mapping, Wright *et al.* (2005) and Yamasaki *et al.* (2005) used resequencing data from >1,000 loci to identify genomic regions associated with domestication and improvement. 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. More recent work has successfully expanded this approach to include the comparison of whole-genomes (Vielle-Calzada *et al.* 2009, Gore *et al.* 2009). Recent work from our group (see below) has extended both the scale (Hufford *et al.* In Review) and statistical methodology (van Heerwaarden *et al.* 2010, van Heerwaarden *et al.* In Prep) of selection mapping in maize.

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 sequencing technologies to

perform a dense, genome-wide scan for selection across multiple population comparisons in order to find loci relevant to climate change.

PROGRESS TO DATE

The three years of our initial grant are focused on identifying CAL for yield by selection mapping across a chronological sample of North American maize lines. We are currently in year two of the grant, and have already made progress in genotyping and analysis, with several papers published, in review, or in preparation. During year one we selected a list of 400 US lines, dividing these into landraces, pre-1950's inbreds, pre-1980's inbreds, and former patent variety protection (ex-PVP) lines. All of these lines were genotyped on the Illumina 55K SNP array. During this time we also worked on methods, including publication of an application of principal component analysis to detect structure using linked markers (van Heerwaarden *et al.* 2010) and methods for estimating drift and ancestry using SNP data (van Heerwaarden *et al.* 2011).

In year one of the grant we also embarked on a collaboration with other maize researchers to describe variation in cultivated and wild maize based on >100 fully resequenced genomes. Our group led the population genetic analysis of the data, using selection mapping approaches to identify hundreds of new CAL (Hufford *et al.* In Review; Fig. 2).

The second year of the grant has primarily involved data analysis. Our analyses of the collaborative resequencing data are currently in review (Hufford *et al.* In Review, Chia *et al.* In Review), and a manuscript is in preparation on analysis of selection and coancestry in our 400 genotyped lines (van Heerwaarden *et al.* In Prep). The resequencing analysis identified hundreds of CAL that appear to have been selected across most inbred maize, including loci in the

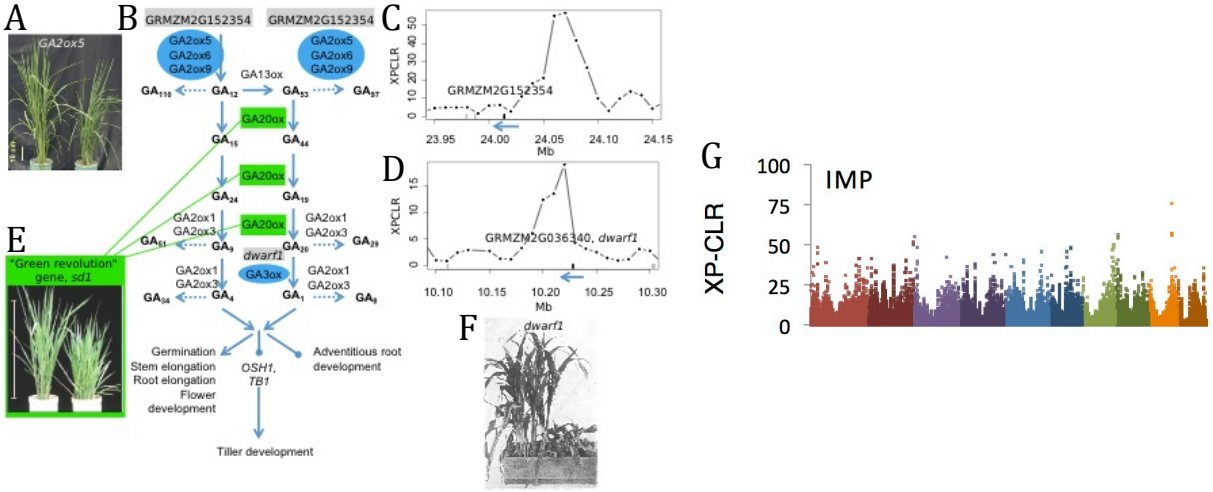

Figure 2. Selection screens in maize **A)** *GA2ox5* mutant in rice. **B)** The gibberellin biosynthesis pathway showing maize CAL (grey background) and their rice orthologs (blue). **C-D)** Likelihood of selection (XP-CLR) in the genomic regions around example CAL. **E)** High-yielding rice variety IR8 has a mutation in *GA20ox* **F)** *dwarf1* mutant in maize. **G)** XP-CLR for modern breeding across the maize genome. Chromosomes are shown in different colors. Panels A,B,E and F adapted from the literature (Emerson 1912, Sasaki *et al.* 2002, Lo *et al.* 2008, Ryu *et al.* 2009)

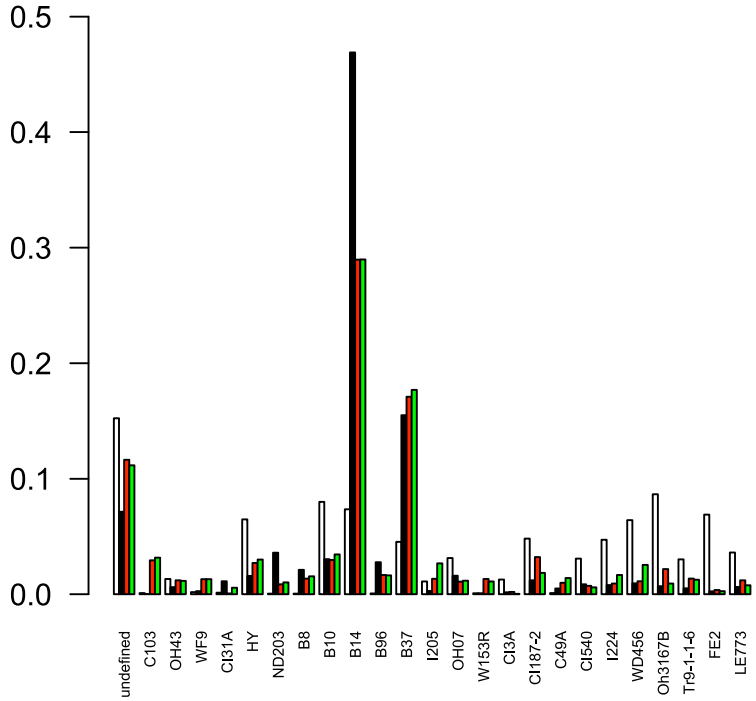

Fig. 3. Bar plot of the genomic distribution of basal ancestry in the different time categories (white: pre-1950's; black: pre-1980's, red: ex-PVP; green: selected regions).

Best	Worst
B14	P8
B37	C102
A334	H49
I205	K4
IDT	MT42
W22	A6
B164	66
T8	N6
W182B	W9
CI187-2	MO1W

Table 1. Ten best and ten worst lines, determined by comparison to an idealized genotype made up entirely of CAL identified in selection scans of North American breeding lines.

gibberelin biosynthesis pathway (Fig. 2), known for the rice "green revolution" gene, *sd1* (Sasaki *et al.* 2002). Our analysis of North American breeding material has developed novel haplotype-based methods to assign ancestry and estimate the contribution of founder breeding lines to later material (Fig. 3). Because we have also identified >600 CAL using selection mapping approaches, our ancestry analysis also allows us to parse out the contribution of founder lines to regions which have experienced selection (Fig. 3, green bars). Comparison of founder lines to an idealized genotype created from a combination of all selected alleles further allows us to rank founder lines in terms of their value to breeding programs; table 1 shows the 10 best and 10 worst lines resulting from this analysis.

Our validation of the identified CAL is progressing as well. Though our original collaborator at Monsanto has left the nearby Woodland field station, Monsanto has continued the collaboration at additional sites (seed increase in greenhouses in Nebraska and crossing at field sites in Hawaii). We have chosen a subset of 60 of our lines to cross to a single tester (207) for analysis of yield of the F1s. Because the genotypes of all the parents are known, no additional genotyping will be needed. Initial seed increases were successful, but the first attempt at crosses was not successful for all lines; an additional planting and crosses are planned for later this year. In addition to the yield tests through our Monsanto collaboration, Dr. Rita Mumm (U. Illinois) has kindly offered to share her yield trial data with us (see attached email). These data, from a full diallele cross of 12 of our genotyped lines (B73, Mo17, and 10 ex-PVP), should provide a valuable additional resource for validating the CAL identified via our selection mapping approach.

APPROACH

The proposed project has two major objectives aimed at identifying candidate agronomic loci (CAL) for use in marker-assisted selection for adaptation to climate change in maize.

Objective 1: Identify CAL by scanning the maize genome for loci that show evidence of parallel adaptation to climate

Objective 2: Validate CAL in growth chamber experiments using association analysis

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 adaptation to new climate regimes. Because our goal is to provide marker loci for use in marker-assisted-selection, detailed analysis of the function of individual CAL, though obviously of interest, is outside the scope of this project. To identify CAL, we propose to harness full-genome resequencing of open-pollinated populations of *Zea mays*. We will take advantage of natural "replicate experiments," in which maize has independently adapted to new climates on multiple occasions. Using population genetic analyses to scan for the targets of selection in each of these occasions, we will identify loci that have played a role in successful adaptation to new climate. Comparison of these loci with existing variability in temperate and tropical breeding material will augment the power of our approach and identify material immediately useful for molecular breeding.

Sample selection

Maize was initially domesticated in the lowlands of Mexico and subsequently diffused into the highlands of the Mexican Central Plateau, the lowlands of Central and South America, and the Andean highlands. This diffusion involved two separate adaptations of lowland material to proximate highland environments: 1) lowland Mexico to highland Mexico, and 2) lowland South America to the Andes. Our resequencing panel will sample from representative landraces of these two evolutionary "experiments" in highland adaptation. In an unrelated project, we have previously genotyped 94 open-pollinated maize landraces (24 lowland Mexican, 24 highland Mexican, 23 lowland South American, 23 Andean) at >55,000 single nucleotide polymorphisms (SNPs; Fig. 4) using the Illumina 55K array. We have used principal component analysis to identify a representative core set of ten accessions per geographic region for full-genome resequencing (a total of 40 genomes). Tissue of these individuals is already available in the laboratory; use of the same tissue will allow estimation of error rates via comparison between the Illumina 55K SNP array and our resequencing data.

Both Mexican and Andean highland populations have colonized colder high altitude environments from a warm-adapted lowland ancestor (Fig. 1 D-E), paralleling the adaptation of modern US temperate lines to colder climes. It is thus the case that the population genetic signals of selection we identify in our study may often be the result of selection for adaptation to colder climates. This should not pose a problem for our selection mapping approach, however,

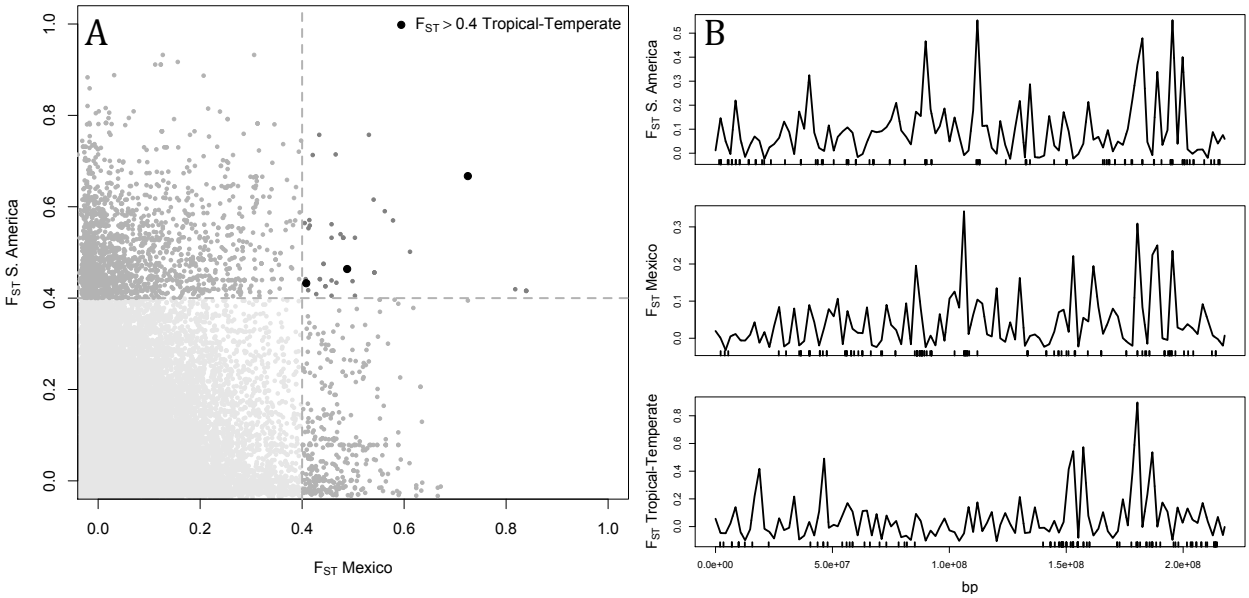

Figure 4. Analysis of allele frequency differentiation (F_{ST}) in landraces and inbreds using the Illumina 55K SNP array. A) F_{ST} between highland and lowland landraces. The 3 large black points represent SNPs that are highly differentiated in both comparisons and between temperate and tropical maize. B) F_{ST} shown across chromosome 5. Line shows a lowess fit of all the data, hatches show the position of the 2.5% most differentiated SNPs in each comparison.

as we expect most CAL to have alleles that are differentially adapted to different temperatures: identification of an allele selected in highland environments thus also identifies the alternate allele as one adapted to warmer environments. Even in a scenario in which the lowland/tropical allele confers no advantage over the highland/temperate allele under warmer conditions, a CAL would at worst represent a locus that harbors variation which can confer adaptation to different temperatures, implying that other functional variants at that locus could prove agronomically useful.

Sequencing

While we have already genotyped a set of highland and lowland individuals using the Illumina 55K SNP array, this platform is not well suited to selection mapping in outbred landraces. While inbred lines of maize from a single breeding program will exhibit considerable LD such that markers may be informative about selection over broad regions, LD decays rapidly in outbred landraces, and a higher marker density is thus needed to effectively cover the majority of the genome. One consequence of this is shown in Fig. 4B, where comparisons of highland to lowland in Mexico and South America to temperate and tropical inbreds all appear to show elevated differentiation (implying selection) in a common region on the distal end of chromosome 5, but our genotyping panel fails to highlight individual SNPs as likely CAL. In fact, Gore *et. al* (2009) concluded that tens of millions of SNPs would be necessary to successfully "tag" most loci in the genome with a marker in high LD. While marker density is improved through high-throughput genotyping systems (e.g., Elshire *et al.* 2011), even these do not approach the marker density suggested by Gore *et al.* (2009). We have shown, however, that full-genome resequencing — which generated more than 55 million SNPs (Chia *et al.* In

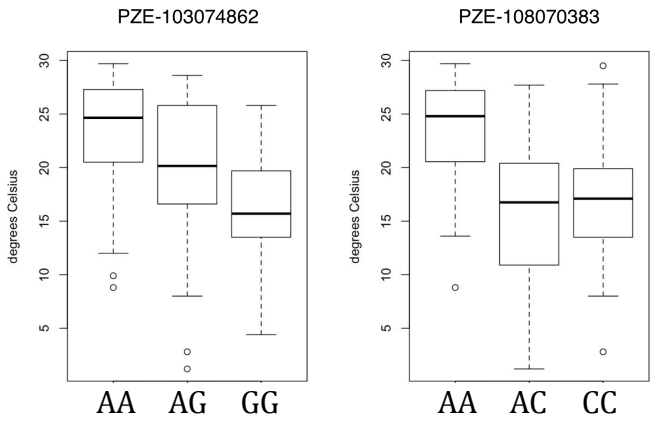

Fig. 5. Mean August temperature of collection locations of 94 landrace accessions in Mexico and S. America. Accessions are divided into genotypes for each of two SNPs showing strong differentiation between highland and lowland environments and between tropical and temperate inbred lines (see Fig. 4A).

Review) — provides sufficient marker density for effective selection mapping in even outbred material (Hufford *et al.* In Review).

As part of our preliminary investigations for this project we have obtained a small grant from the UC Davis Genome Center to resequence one highland Mexican landrace. We are generating 100bp paired-end reads in a single lane of the Illumina HiSeq 2000 sequencing system. This platform should provide us with approximately 30 Gigabases of sequence, or 13X coverage of the 2.3Gb maize genome. Using these preliminary data, we will hone our read-mapping and SNP-calling pipelines and confirm that a single lane of sequence provides adequate coverage to genotype open-pollinated, heterozygous material with an acceptably low error rate. As we describe in Chia *et al.* (In Review), reads will be mapped against the B73 maize reference sequence (RefGenV2) using the Bowtie and Novoalign algorithms, retaining only uniquely mapped reads. Variants identified relative to the B73 reference will be filtered using population-genetic-based quality control criteria shown previously to achieve very low error rates (<1%; Chia *et al.* In Review). Because we have genotyped these individuals on the Illumina 55K SNP array, we can also assess error rates by comparison to genotypes obtained from the array.

Statistical methods to identify CAL

Multiple comparisons

One of the strengths of the selection mapping approach we are proposing is the use of multiple population comparisons. The utility of multiple population comparisons has been widely shown in evolutionary analyses (Turner *et al.* 2008a, Turner *et al.* 2008b), but to our knowledge has never been applied to crops. Comparison of a single pair of populations inevitably identifies a large number of loci that are strongly differentiated (Fig. 4A). Simply comparing our sample of tropical and temperate inbred lines, for example, identifies >2500 SNPs that are strongly differentiated (F_{ST}>0.4). Many of these SNPs likely have little to do with adaptation to different climates, and instead reflect the neutral process of drift between isolated populations. By comparing two pairs of populations that have independently adapted to similar climatic conditions (Fig. 1 D-E), we immediately narrow the pool of potentially interesting loci (Fig. 4A). Inclusion of a third comparison identifying overlap with SNPs differentiated among breeding material results in a set of only 2 loci (3 total SNPs) from the >50,000 SNPs evaluated. That this is a powerful approach to identifying CAL can be seen in the genotypic comparisons in Fig. 5; the same allele at both loci is associated with tropical/lowland environs, and the lowland/tropical

allele is predominantly found in warmer environs. Because a single locus may change in multiple different ways, we will assess parallel changes at the level both of SNPs and independent genes. While undoubtedly there will be adaptive loci that are missed by this approach, one tremendous advantage of the method is that CAL identified in parallel have already been proven to have an effect in multiple genetic backgrounds and different environments, dramatically increasing their potential utility for MAS or transgenic approaches to breeding.

Identifying selection

We will rely on two complementary population genetic approaches to detect CAL under selection for adaptation to new climatic conditions. Because we know which is ancestral and derived in each of our three pairs of populations, we can make use of the composite likelihood method of Chen *et al.* (2010), which calculates the likelihood of observing extended regions of high differentiation under models of selection and drift. We have modified the method to correctly incorporate missing data, and have applied it to similar next-generation sequencing data (Hufford *et al.* In Review) to detect selection during domestication and modern breeding in maize. The second method does not look for patterns of differentiation, but instead explicitly tests for association between SNPs and environmental variables of interest (Coop et al. 2010). We have already applied variations of this method to Illumina genotyping data in both wild (Pyhäjärvi *et al.* In Prep) and cultivated (van Heerwaarden *et al.* In Prep) maize as well as pine (Eckert *et al.* 2010). This approach is similar to association mapping in having the disadvantage of only focusing on *a priori* environmental variables, but likely has greater power to detect loci that vary among individuals within populations, and allows for identification of CAL that associate with important environmental variables (e.g., precipitation) that are not starkly contrasted between highland and lowland populations. We will evaluate our CAL based on the strength of evidence for selection from both of these methodologies.

Identify CAL in breeding lines from public low-coverage sequence

To maximize the utility of the CAL we have identified in comparisons of landrace populations, we will compare our sequence data to publicly available sequences from temperate and tropical inbred lines. This serves a dual purpose: not only does it increase our power to identify likely CAL (e.g., Fig. 4), but it also allows us to identify modern breeding material with putatively adaptive alleles. Moving CAL from modern inbred breeding material into commercial varieties would be significantly faster than bringing them in from exotic germplasm. Inbred lines also allow for genetic studies that can validate and test the functional basis of our CAL. Fortunately, we note that a large number of inbred lines will be publicly available for comparison. Our resequencing work includes more than 70 inbred genomes (Chia *et al.* In Review), and we are aware of a number of other sequencing projects underway. It is important to note, however, that these sequence resources alone would likely not prove sufficient for identifying CAL (see discussion above of the advantages of multiple comparisons).

Objective 2: Validate CAL in growth chamber experiments

The well-characterized maize 282 inbred line association panel (Flint-Garcia *et al.* 2005) is an exceptional resource for independent confirmation of the "bottom-up" population genetic approach described in Objective 1. Since each of these lines has minimal residual heterozygosity, a representative genotype for each line is sufficient for conducting an association study based on phenotypes of line replicates. We have collaborated in the genotyping of >51,000 SNPs from the Illumina 55K array on this panel (Cook *et al.* In Review), and have already made these data publicly available upon request. Moreover, genotypes for ~500,000 additional loci will soon be available from large-scale, genotyping-by-sequencing efforts presently underway (see attached email from Dr. Ed Buckler). Variants from these data sets that overlap or are linked to CALs identified in Objective 1 will be tested in a targeted association analysis through a growth chamber simulation experiment of current and predicted future (2080) temperatures in the U.S. Corn Belt. While other variables (e.g., precipitation, CO_2) are likely to be affected by climate change in this region, we will focus on temperature as it is the the variable expected to be most significant for maize yield (see rationale and significance above). Also, by limiting our treatments we will gain power to detect CAL significantly associated with adaptation to changing temperature. Because this validation is essentially an exercise in association mapping, it suffers many of the limitations discussed above. Notably, CAL identified through selection mapping may not show effects in the limited number of phenotypes we can compare here. Nonetheless, observation of a phenotypic effect provides a powerful validation of our selection mapping procedures, and this objective is thus a useful demonstration of the utility of at least a subset of our CAL.

Experimental design

Many of the accessions in the 282 inbred line association panel are already in our germplasm collection at UC Davis. The remainder of this panel will be obtained from the USDA National Plant Germplasm System and collaborators. Present mean growing-season temperature for the Corn Belt will be estimated from weather records aggregated over the period 2001-2010 while future temperatures will be chosen from comparison of several climate models (e.g., Fig. 1). Phenotypic data will be recorded from five individuals per line grown under short-day conditions at present and future temperatures in multiple growth chambers available through the Plant Sciences Department and the Controlled Environmental Facility (http://cef.ucdavis.edu/cef/) at UC Davis.

Traits measured

While growth chamber space restrictions will preclude phenotypic measures at plant maturity, several traits informative for climate adaptation can be measured during the first six to eight weeks of the maize life cycle. For each maize plant we will measure: 1) Initial seed weight to account for maternal effects; 2) Days to germination; 3) Germination rate; 4) Plant height at 15, 30, and 45 days; 5) stomatal conductance (an indirect measure of photosynthetic rate); and 6) Total above-ground dry biomass. All of the necessary equipment for such phenotyping is already available either in the Ross-Ibarra laboratory or through shared departmental resources.

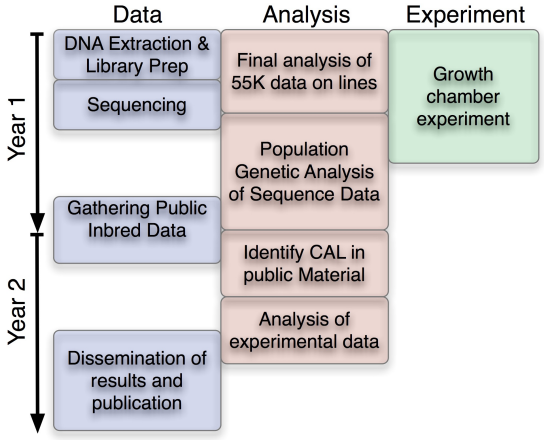

Fig. 6. Timeline of activities

	Ross-Ibarra	Hufford	Postdoc	Technician
Sequencing	X	X		X
Mapping and SNP discovery	X	X		
Population genetic analyses	X		X	
Growth chamber validation	X	X		X

Table 2. Responsibilities of project participants

Association analysis and validation of CAL

Phenotypic data from individual traits and a cumulative estimate of plant vigor will be used in a single-marker association analysis of CAL. The testing of CAL represents an hypothesis-driven approach and thus avoids multiple-test correction issues that decrease statistical power for detecting associations between genotype and phenotype. However, it is likely that a subset of CAL will lack appreciable frequencies of both alleles within the association panel, limiting our ability to infer associations for these loci. If, however, our CAL are shown to better predict phenotype under temperature treatments than random sets of SNPs, we will have demonstrated the power of using multiple natural evolutionary "experiments" to scan for adaptive polymorphisms of potential use for breeding and MAS.

Research Timeline

A timeline of the proposed research is presented in graphical format in Fig. 6, and the division of labor is shown in Table 1. Dr. Ross-Ibarra will work with Dr. Hufford, the postdoctoral scholar, and the laboratory technician throughout the project, guiding most of the activities. In the first year, Dr. Hufford will oversee the library preparation and sequencing of the 40 chosen landrace lines. Dr. Hufford will then take the lead on read mapping and SNP discovery. We will hire a full-time postdoctoral scholar to implement the population genetic analysis, starting midway through year one. After preparing the libraries for sequencing, the technician will coordinate the growth chamber experiments.

In year two we will finish the population genetic analysis of the landrace data, and the postdoc will then be in charge of comparing our results to results from publicly available genome sequences of inbred breeding material. S/he will also be responsible for the analysis of the growth chamber experiments, testing the fitness effects of the CAL identified. By the end of year two, we plan to have 1-2 manuscripts in preparation for submission. We will present our progress and findings at the maize genetics conference both years of the grant.

BROADER IMPACTS

Training

The proposed project will provide important training for two postdoctoral scholars. Dr. Hufford, currently a postdoctoral scholar in Dr. Ross-Ibarra's lab, will be applying for faculty positions in the first year of the grant. As a new PI, the project will provide valuable experience supervising a laboratory technician and undergraduate interns, in addition to the training involved in grant preparation, planning, and budgeting. Moreover, while Dr. Hufford has considerable experience working with next-generation data (Hufford *et al.* In Review; Chia *et al.* In Review, Tenaillon *et al.* 2011), this project will provide him with an opportunity to lead both read-mapping and variant discovery. The postdoctoral scholar to be employed on the project will gain considerable training in population genetic and bioinformatic analysis of next-generation data, as well as important statistical skills in the association analysis of the validation experiments in objective two. The postdoctoral scholar will also be primarily responsible for presenting the work at conferences and preparing manuscripts.

The project will also train undergraduate students in both bioinformatics and phenotyping. The laboratory has a history of training undergraduate interns through the biotechnology internship program at UC Davis. In the last 2 years, we have trained four undergraduate interns; one is now a full-time laboratory manager, another works as a programmer at a large biotech company, and a third has received a USDA National Needs graduate scholarship to work on soybean genetics; two of the four have earned authorship on publications in preparation. Potential interns on this project will train under the postdoctoral scholar or Dr. Hufford on the bioinformatic aspects of the project, or work with the laboratory technician and Dr. Ross-Ibarra on the phenotyping and validation of detected CAL. Interns will be encouraged to undertake a related independent project or independent part of the larger project and, if appropriate, will be offered authorship for their contributions to published work.

Benefits to maize breeding and genomics

The main product of our proposal will be a list of candidate agronomic loci (CAL) for adaptation to new climatic conditions in maize. This list of loci will be made available in publications and via the Internet, potentially as a track on the genome browser at the Maize Genetics and Genomics Database (www.maizegdb.org). The list will allow breeders and industry to quickly begin testing CAL discovered here as markers in MAS or in transgenic approaches. Efforts underway in the GEM (Germplasm Enhancement of Maize) project (http://www.public.iastate.edu/~usda-gem/GEM_Project/GEM_Project.htm) demonstrate the feasibility of moving alleles from tropical inbreds and landraces into temperate lines, and our preliminary analyses suggest that even strongly differentiated loci (e.g., Fig. 5) are polymorphic within at least some temperate material. These facts suggest that progress could be made relatively quickly once a set of CAL have been identified, especially given the ability to identify lines with putatively adaptive alleles via genotyping alone.

Dissemination of data and methods

We will disseminate our method and results widely, both in presentations at conferences and in peer-reviewed publications. In addition to publishing our results, we will continue our history of making data (e.g., lists of CAL and related information, the raw genotype data, source code) publicly available on our lab webpage (www.rilab.org).

Application to other breeding programs

The present proposal is focused on maize, but the analytical and technical approaches developed here should be appropriate to similar genotype-based studies in any breeding program. Our approach of using multiple comparisons of traditional breeding pools that have had many generations to adapt to different conditions should provide considerable power to identify CAL that may contribute to breeding for future climate change in a number of crops. Notably, as genotyping becomes ever cheaper, such studies are now feasible in crops not generally considered as model organisms. Because we are taking a selection-mapping approach, our methods do not require detailed pedigree information, the ability to make wide crosses with traditional varieties, or even the need for detailed genetic or physical maps. Though selection mapping is likely to be most fruitful in crops with well-developed genomic resources, rapidly expanding genomic methods (e.g., Baird *et al.* 2008) allow selection mapping even in non-model species.