Introduction 1
Main Hypotheses 1
Why Study Rare Alleles? 1
Population Genetic Theory of Rare Alleles 1
Why Study Rare Alleles in Maize? 2
Characterization of Sequencing Variants with Polymorphism Descriptors 3
Project Structure 3
A. Genome Models to Predict Functional Rare Alleles 5
A1. Pan-*Zea* GWAS 5
A2. Pan-*Zea* Population Genetics 7
A3. Data Mining to Predict Functional Alleles 9
B. Phenotype Models Incorporating Rare Allele Predictions 11
B1. Predict Overall Fitness Using PD Trained Models 12
B2.Testing of PD Hypotheses with Maize Hybrids 13
C. Evaluation of Rare Alleles in the Field 14
C1. Integration of Global *Zea* Fitness Trials 14
C2. Evaluation of Deleterious Rare Alleles in Teosinte and Landraces 17
C3. Comparative Evaluation of *Zea* Alleles 19
D. Allele Discovery and Characterization 21
D1. Whole Genome Sequencing for Rare Allele Discovery 22
D2. Genotyping to Track Alleles 24
D3. Annotate Maize Genome with Polymorphism Descriptors 26
Education and Outreach 26
References 30
Appendix 1 - Proposed Polymorphism Descriptors (PDs) 34
Appendix 2 - Table of Germplasm and Status of Genotypic and Phenotypic Data 36
Appendix 3 - Construction of a Pan *Zea* -Genome 37
Appendix 4 - SNP Calling and Identification of Rare Alleles 38
Appendix 5 - Roles 39
Appendix 6 – Timeline 41

Introduction

This project will combine the power of population genetic and molecular models with quantitative genetics to elucidate the contributions of rare versus common alleles to the phenotypic variation and evolution. While the biology of rare alleles is fundamental to our understanding of evolution and genotype-to-phenotype relationships, it has yet to be adequately explored in any system. We propose to take advantage of recent advances in high-throughput genotyping and phenotyping methodologies to identify the key biological attributes of variants[1], or "polymorphism descriptors" (PDs), that will allow us to predict the functional effects of rare alleles in *Zea*. This project will refine our understanding of natural phenotypic variation, which is critical to genetics, medicine, agriculture, and conservation.

On a practical level, the proposed research will provide tools to identify the beneficial and deleterious SNPs in maize individuals, and to estimate their overall number and distribution in populations. This information can then be used in genomic selection or future homologous recombination approaches for crop improvement. This will facilitate the use of diverse genetic resources such as landraces, and even teosinte, in elite breeding programs. The effectiveness of plant breeding will be enhanced by improving our ability to identify, predict, and select on the effects of rare variants, both deleterious and beneficial.

Main Hypotheses

(H1) Rare alleles contribute significantly to phenotypic variation in diverse maize and teosinte.
(H2) Classification with molecular and population genetic polymorphism descriptors (PDs) will enable the identification of functional rare variation.
(H3) Genetic models including the aggregate effects of rare alleles will enable more precise genotype to phenotype prediction across diverse maize germplasm.

Why Study Rare Alleles?

Recent advances in genome wide association studies (GWAS)[2–5] and genomic selection methods[6,7] have enabled the identification and efficient use of common alleles (>5% frequency). Because many of those common alleles are likely fixed in elite breeding populations, the transformative potential of genetics and breeding now lies in rare alleles (<5% frequency). Ameliorating the effects of deleterious rare alleles is a major goal of both plant breeding and medical research, while the focus of allele mining from diverse germplasm is to increase the frequency of beneficial rare alleles.

In maize, approximately 90 single nucleotide mutations occur each meiosis[8]. If even a small fraction of these mutations affect phenotypes, a sizeable proportion of phenotypic variation may be due to rare alleles. While the low frequency of any one allele limits its influence on phenotypic variance, rare alleles as a class have the potential to be an important component of heritable trait variation. Distinguishing between models of population variation due to common variants of small effect versus rare alleles of larger effects is difficult and different lines of evidence support different models, even within the same species[3,9–14]. Our efforts at genetic mapping have made considerable progress towards identifying common functional polymorphisms[4,5,9,15–17], but even some of these may result from synthetic associations due to close linkage with causative rare variants[9,18]. A better understanding of the biology of rare alleles will enhance our ability to predict phenotype from genotype and thereby accelerate breeding efforts.

Population Genetic Theory of Rare Alleles

Natural selection prevents deleterious alleles from increasing in frequency within a population, keeping such alleles at low frequencies. Rare alleles are thus enriched for deleterious, functional variants (*e.g.*, nonsynonymous mutations). Dominant deleterious alleles are more easily removed by natural selection and, at mutation-selection balance, will be very rare. Recessive alleles, on the other hand, are shielded in heterozygotes and thus will segregate at frequencies often many orders of magnitude higher.

Rare deleterious alleles contribute to **genetic load**, the difference between a population's mean fitness and its maximum potential fitness[19]. We can partition the effect of deleterious alleles on genetic load into a component that is expressed in a panmictic population, and an additional component that is expressed under inbreeding. This second component, known as **inbreeding depression (ID)**, arises from the increased frequency of homozygosity for recessive deleterious alleles in inbred populations.

Estimates of the fitness effects of new, nonsynonymous mutations in plants suggest that most are deleterious[20,21]. Strongly deleterious mutations are unlikely to persist, but weakly deleterious mutations are often found segregating at low frequencies[22] and contribute to genetic load, ID, and, likely, to heterosis through complementation[23]. Furthermore, Hill-Robertson interference among linked, selected mutations can lead to an enrichment of deleterious alleles in regions of low recombination[24–27] and around loci targeted by recent strong positive selection[28] such as likely occurred during domestication[29,30].

Unlike deleterious alleles, beneficial mutations usually fix quickly in a population. However, processes such as genetic drift, local adaptation, or recent environmental change (including anthropogenic change under domestication) can maintain species-wide polymorphism for beneficial alleles. For example, selection can efficiently act on a beneficial allele only when the product of the effective population size and the selection coefficient ($N_e s$) is much larger than one. Because of this, the effects of drift in small populations (such as those used in breeding programs) may prevent the effective utilization of potentially beneficial alleles. Genotype by environment interaction or geographic isolation can also limit the spread of beneficial alleles. In *Arabidopsis* many alleles were found to be locally beneficial but have deleterious effects on fitness in other populations[31]. These considerations suggest that some potentially beneficial alleles are likely segregating at low frequencies.

Why Study Rare Alleles in Maize?

The unique history of maize makes it a powerful model system for the study of rare alleles. For the last million years, the progenitor of maize, teosinte (*Zea mays* ssp. *parviglumis*), has consisted of large, predominantly outcrossing populations. Teosinte and maize shared a common gene pool and population history until their recent divergence during domestication[32], and both still share a common genome structure[33]. Maize landraces consist of open-pollinated populations that have experienced only a modest bottleneck in genetic diversity[34], accompanied by strong selection at hundreds of loci[30,35]. Both maize landraces and teosinte suffer substantial ID[36,37], and hence are useful for the study of deleterious rare alleles.

Intentional inbreeding and selection for inbred survival and vigor in maize were initiated only in the last 100 years and represent the result of intensive selection to purge strongly deleterious alleles[36]. However, pedigree breeding with inbred lines imposed very strong bottlenecks on diversity, so that some previously rare alleles are now quite common in modern maize. Although a small proportion of these historically rare alleles are likely favorable, most are probably not. Studying the distribution and function of these alleles across teosinte, landraces, and modern breeding lines will provide great power to dissect and predict their effects.

The quantification of the effect sizes and fitness consequences of rare alleles is one of the most challenging issues in biology, as quantitative genetic approaches have traditionally suffered from limited resolution (QTL mapping) or insufficient allelic replication (GWAS) to allow accurate estimates of the effects of rare alleles. While these obstacles are difficult to overcome in many systems (like humans), in maize we can experimentally modify the frequency of rare alleles through controlled crosses (*e.g.*, mapping populations) and breeding bottlenecks (*e.g.*, closed random mating populations). Such populations can then take advantage of the sophisticated machinery of association mapping[38] and genomic selection[39] to analyze genetic variants and predict their phenotypic consequences. These experimental

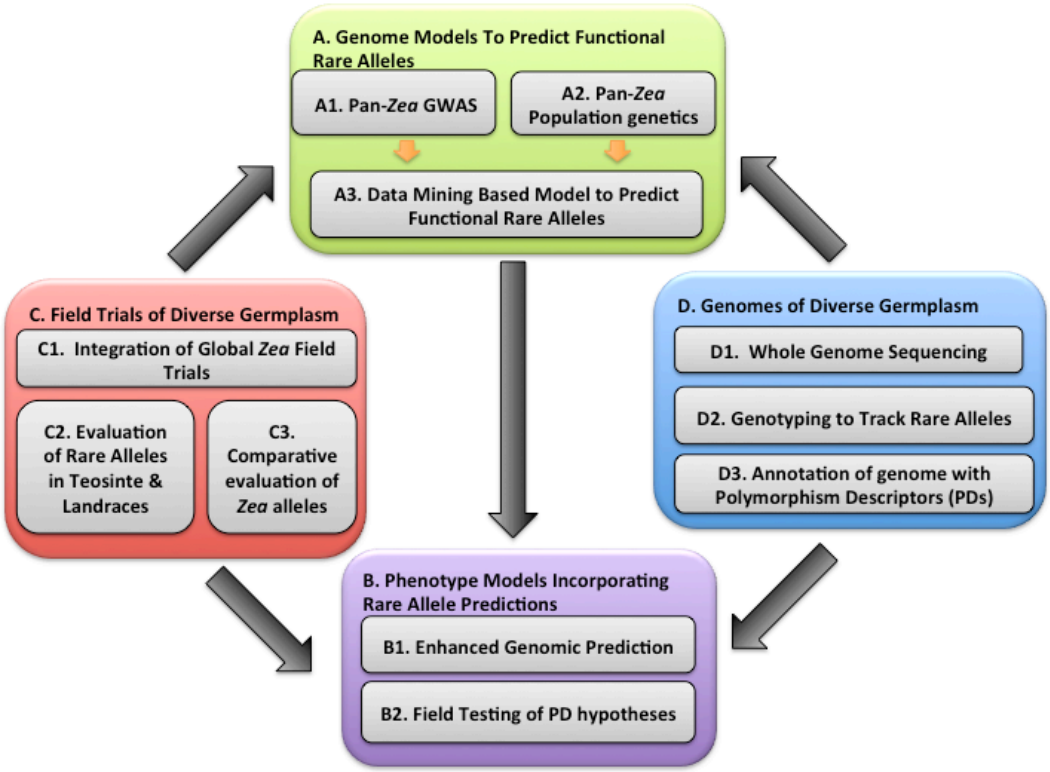
populations can be used to investigate an array of predicted effects, both fitness-related and agronomic.

Characterization of Sequencing Variants with Polymorphism Descriptors

Using field trials and quantitative genetic analyses such as GWAS to relate phenotypic values to sequence variation can be considered a forward genetics approach. Complementary to this is a reverse genetics approach involving the application of models derived from molecular biology and population genetics to make *a priori* predictions of the allelic effects of rare polymorphisms[1,40]. For example, the sequencing of large populations, combined with models to predict the effects of sequence variants, has been used to predict that hundreds of human genes carry recessive deleterious alleles, and that alleles predicted to disrupt protein functions are rare[41,42].

To better understand the phenotypic and fitness consequences of nearly all of the variants in our experimental populations, each site in the pan-*Zea* genome can be characterized with regard to a broad spectrum of biological attributes, ranging from molecular features such as chromatin status to population genetic features such as allele frequency across *Zea*. We refer to these biological attributes as "polymorphism descriptors" (PDs). Multiple PDs can be used to make testable hypotheses about the fitness and agronomic effects of sequence variants. Recent work in humans has shown that several evolutionary and functional classes are useful in predicting loss-of-function mutations[13,43] and our own work investigating unexpected heterozygosity in recombinant inbred lines suggests that local recombination rate may be informative as to the frequency of deleterious variants[25–27,44]. These and other PDs can be used to predict which polymorphisms are likely to be functional and whether their effects are beneficial or deleterious.

Project Structure



A. Genome Models to Predict Functional Rare Alleles

This section of the project directly uses hypotheses based on genomics and population genetics to identify functional alleles, with an emphasis on rare alleles. The elements of section A are:

(A1) *Pan-Zea GWAS*: Perform GWAS on extensive field trial data (detailed in section C) to estimate effects of alleles across *Zea*. Additive and dominance effects and genotype by environment interactions will be estimated for variants throughout the genome.

(A2) *Pan-Zea Population Genetics:* Evaluate patterns of selection and evolution across *Zea*. Identify regions of evolutionary constraint, estimate the fitness effects of new mutations.

(A3) *Data Mining to Predict Functional Alleles*: Use genomic and population genetic criteria to classify sequence variants according to numerous polymorphism descriptors (PDs). Determine which PDs predict effect estimates (from section A1) and allele frequency (from section D). From these results, we will create unified models that identify functional alleles based on optimal combinations of PDs.

B. Phenotype Models Incorporating Rare Allele Predictions

While basic scientific understanding of functional variation (section A) is very important, we also want to apply this knowledge to breeding. The elements of section B are:

(B1) *Enhanced Genomic Prediction*: Use the best PD hypotheses from section A3 to try to improve prediction of phenotype from genotype. Special emphasis will be placed on testing these prediction models for use with diverse germplasm (tropical breeding germplasm, landraces, and teosinte).

(B2) *Field Testing of PD hypotheses*: Test the most promising PD hypotheses (identified in sections A3 and B1) by making and evaluating new independent sets of hybrids differing for their frequencies of specific PD classes, while controlling for other genomic differences.

C. Field Trials of Diverse Germplasm

Field evaluation will provide data on fitness-related traits that will be used to estimate rare allele effects. In this section, we take advantage of our ability to artificially manipulate allele frequencies within populations so that subsets of rare alleles can be studied intensively. While these field-based studies (C1 to C3) individually address key questions, they also provide the raw data for the GWAS studies of section A1. The elements of section C are:

(C1) *Integration of Global Zea Field Trials*: Combine and curate field trial data for over 468,000 total plots representing over 89,000 maize inbreds and hybrids. Most of the field trial data sets already exist or are being completed within the next year; we will unify these massive genotypic and phenotypic data sets for combined analysis.

(C2) *Evaluation of Deleterious Rare Alleles in Teosinte and Maize Landraces*: Measure fitness of outcrossed and partly inbred teosinte and maize landrace germplasm. We expect to uncover alleles with large deleterious effects that have been purged in elite maize (thus absent from C1 experiments). Understanding these alleles will inform our PD analyses in section A3.

(C3) *Comparative Evaluation of Zea Alleles*: Directly compare the effects of alleles derived from outcrossing teosinte to those from inbred maize lines by measuring their effects in a common genetic background, the "*Zea* Synthetic."

D. Genomes of Diverse Germplasm

Although rare alleles are ubiquitous and may contribute substantially to the genetic architecture of fitness and other complex traits, we lack a robust catalog of rare alleles in *Zea*. Section D will characterize rare variants at the molecular level for all of the key germplasm evaluated in C. The elements of this section

are:

(D1) *Whole Genome Sequencing (WGS)*: The sample size of extant whole genome resequencing data is insufficient to estimate allele frequencies in teosinte and landraces, and the depth is generally insufficient for assembly of novel genes and elements that are absent from the reference genome. We will deeply sequence all founders of key experimental populations used in this grant to discover rare alleles.

(D2) *Genotyping to Track Rare Alleles*: Genotyping-by-sequencing (GBS) provides a powerful way to track rare alleles and/or haplotypes through diverse germplasm at two million sites across the genome. We have already genotyped 33,000 taxa with GBS, and this proposal will conduct GBS on another other 29,000.

(D3) *Annotation of the Genome with Polymorphism Descriptors (PDs)*: Sections A and B rely on annotation of genomic features. This section describes our approaches to develop new annotations and incorporate them with ongoing community efforts to annotate the genome.

A. Genome Models to Predict Functional Rare Alleles

Our goal is to develop a model for predicting rare allele effects based on PDs for two reasons: (1) genotyping is now cheaper and faster than phenotyping, and (2) there is a nearly infinite number of rare variants, which cannot all be field tested. *While our model will be based on hypotheses from molecular biology and population genetics, it will be trained and evaluated based on large, replicated, and powerful empirical datasets.*

Section A conducts the synthesis and analysis for the entire project with a goal of predicting functional rare alleles based on sequence context. It pulls together the GWAS results from across *Zea* (A1), which provides a top-down forward genetics analysis. Section A2 provides population genetic estimates of allele frequency, recombination, and selection. Finally, section A3 combines the results of A1, A2, and the polymorphism descriptors (D3) to evaluate which PD hypotheses are the best predictors of functional rare alleles.

A1. Pan-*Zea* GWAS

Published maize association panels have been successfully used to dissect traits controlled by relatively few large effect genes and/or with well-developed candidate genes. However, more complex traits with fewer *a priori* candidate genes have been less tractable, especially in species like maize with complex genomes and rapid linkage disequilibrium (LD) decay. Until very recently it has been too costly to genotype large enough panels at sufficient marker density to have power detect effects of rare alleles.

The largest published association studies in maize are the studies involving the 5,000 nested association mapping (NAM) lines[27]. In NAM, association mapping is conducted across a large series of bi-parental mapping families, combining the strengths of association and linkage mapping[30,45,46]. While NAM has proven to have sufficient statistical power to dissect traits as complex as flowering and plant height, the current NAM population still samples only a modest proportion of the universe of rare alleles, as it is derived from only 26 founder lines. This section (A1) will conduct a series of joint linkage and association analyses across a much larger set of 89,000 genotypes described and compiled in section C1. These experiments are geared to overcome the limitations of our previous studies:

- Inability to estimate dominance or ID in inbred lines.
- Low resolution to map alleles in regions of low recombination.
- Genotype by environment (GxE) interactions of individual SNPs have only been explored for flowering time.
- Insufficient power to test most rare alleles.

- Species-wide allele frequencies of functional alleles are unknown.

To address these limitations, this section (A1) will evaluate the relationship between allele frequency and effect size, dominance, ID. Additionally, it will provide a robust dataset for evaluation of PDs (A3) and for genome wide prediction (B1).

Research Summary

The genetic architecture of maize will be dissected for yield, flowering time, and height using an extensive array of *Zea* germplasm consisting of ~89,000 different genotypes. These traits all show heterosis and environmental interactions. When data are available, additional traits will be contrasted with these fitness traits. Using GWAS approaches, we will estimate additive and dominance effects and GxE interactions for every polymorphic variant. Several strategies will be used to unite information from field trials representing many different environments and alleles.

Goals and Research Questions

- What variants and genes control flowering, height, and fitness related traits?
- How does the distribution of dominance effects compare to additive effects?
- Are there rare variants with large effects?
- Do rare alleles contribute significantly to inbreeding depression?
- Is allele frequency associated with effect size, dominance, and/or GxE interactions?

Analyses

Two major results will be produced from these GWAS analyses: first, a list of the key variants involved in fitness related traits; and, second, effect estimates for *all* variants in the genome (although resolution will vary). These results will be used to address a series of questions regarding rare alleles.

Section C1 focuses on uniting these trials and evaluating heritability. For our GWAS analysis, we anticipate that we will be able to combine these disparate germplasm pools into, at most, five major analyses: (i) Temperate/Tropical Inbred Maize Trials (N=45,000) - NAM, Ames, PVP, Chinese Trials, and C3 and B2 from this proposal, (ii) CIMMYT SeeD landrace hybrids (N=15,000), (iii) CIMMYT Breeding Germplasm (N=5000), (iv) Teosinte inbreeding evaluations (N=10,000) – C2 from this proposal, (v) Landrace inbreeding evaluation (N=10,000) – C2 from this proposal.

From prior experience, we know that all of the following approaches will work for one or more subsets of this germplasm. For this project, we will evaluate each of the five germplasm sets outlined above individually, and use the combination that works best for the combined set of germplasm.

a) Mixed Model GWAS: Our group has been at the forefront of the development of analytical methods for structured and mixed model association mapping[47,48]. These approaches are most powerful when working with unrelated samples, and thus will be highly appropriate for the Ames germplasm and for the SeeD landrace hybrids.
b) NAM-style GWAS model selection: Over 80% of our data will come from populations derived from known parents via controlled crosses. Following Tian *et al.*[4], we can conduct association mapping on a target chromosome while using joint linkage mapping results to control for the genetic variance from the non-target chromosomes and also efficiently and accurately impute missing SNP data. This analytical approach worked well with NAM, but should be much more powerful with the vastly increased number of families available here.
c) Meta-analysis across all experiments: We anticipate that it will be difficult, perhaps impossible, to unite the full phenotypic and genotypic data set compiled in C1 in a single, global analysis. However, there will be considerable overlap between the SNPs across data sets, facilitating meta-analysis across all the experiments for each SNP.

d) Use common variants to re-scan genome for rare variants: Given the scale of these experiments, we should identify many of the common and large effect variants through analyses a-c. These variants will then be used as covariates in either a mixed model or fixed effect framework[49]. By controlling for these variants with a few degrees of freedom, this second search should have enhanced power to detect rarer variants or those with smaller effects.

Resources

A substantial part of this effort will involve dealing with the scale of the data. The association genetics community has not regularly conducted GWAS with 200M variants or more than 10,000 taxa, especially when a large proportion of the data set consists of bi-parental families. To support this analysis, extensive code optimization is needed, and the computation will need to be implemented on a massively parallel HPC cluster. New code will be developed in TASSEL[50] and the R-based GAPIT[51] to support GWAS on high-density variants. We will apply for NSF's XSEDE resource to carry out the computation for this project, and make the tools available to the research community through the iPlant Collaborative.

A2. Pan-*Zea* Population Genetics

Population genetic analyses of selection offer a means of assessing functionality that is complementary to the GWAS approaches of section A1. While A1 will allow estimation of the effect size of variants for traits related to fitness, the analyses presented here will indirectly infer function by providing estimates of long- and short-term selection on variants across the genome. *This approach has the advantage of allowing estimation of the functional importance of variants unrelated to traits studied in the environments or populations used in field evaluations.* Population genetics analyses also provide means to understand the role of recombination in patterning variation across the genome.

Research Summary

Making use of the hundreds of genomes currently available and to be generated by this project, we will use population genetic analyses to understand the effects of demography, selection, and recombination on the observed distributions of rare variants. We will characterize both individual variants and genomic regions for their levels of constraint, evidence of positive selection, and recombination rate. Finally, we will generate new recombination maps for maize and teosinte.

Goals and Research Questions

- How does population demography affect the abundance of rare variants?
- What proportion of variants are deleterious or beneficial, and how are these distributed across the genome?
- How do patterns of recombination vary among individuals, populations, and subspecies?

Analyses

Demographic modeling: Changes in population demography impact diversity, but rare alleles are especially sensitive to demographic change. Previous population genetic analyses of maize, focusing on gene sequences, identified a paucity of rare variants in maize and interpreted this as evidence of a recent population bottleneck during domestication[34,35]. In contrast, our recent whole-genome sequencing analyses[30,33] instead found a genome-wide excess of rare variants in maize (Figure 1). This is consistent with recent population expansion, and suggests that selective constraint, rather than demography, may explain the lack of rare variants in genic regions. We will further test these models, explicitly evaluating alternative demographic models for both maize and teosinte in order to more fully explain observed patterns of rare variant abundance across the genome.

Constraint: We will quantify selective constraint for all sites in the genome. In coding regions, we will utilize models that account for the physicochemical properties of substitutions at amino acids[52,53]. For other regions of the genome, multispecies alignments of genomic sequence will be used to identify

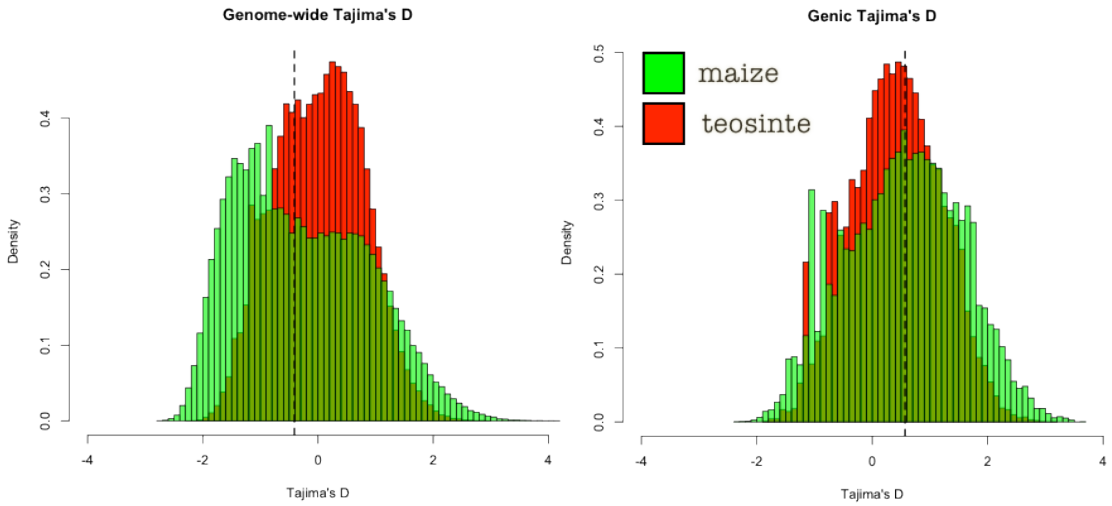


Figure 1. (Left) Genome-wide, maize shows a lower Tajima's D (a measure of allele frequency deviation from neutral expectations) than teosinte, reflecting the excess of rare variants expected in a rapidly expanding population. (Right) Within genes, however, maize shows a higher Tajima's D, likely due to purifying selection against new deleterious variants. The dotted vertical line shows the mean Tajima's D in maize.

constraint via observations of accepted and rejected mutations[54]. Finally, we will develop methods to assess the distribution of fitness effects (DFE) in local regions along the genome. The DFE has been used to infer the strength of purifying selection in numerous plant and animal genomes[55], but to date has only been applied to entire genomes. *Because the efficacy of selection is expected to correlate with local patterns of recombination, the DFE will likely vary along the genome.* We will investigate how the DFE varies along the genomes of both teosinte and maize and use it to infer the past impact of selection and current levels of selective constraint.

Positive selection: We will scan the genome for evidence of positive selection using combinations of phylogenetic (dN/dS) and population genetic (McDonald-Kreitman style tests) analyses to identify individual sites under positive selection as well as genomic regions with a high proportion of adaptive substitutions. Analysis of partial selective sweeps will identify ongoing selection, and understanding the frequency and strength of sweeps across the genome will answer questions about the role of Hill-Robertson interference in maintaining or fixing deleterious alleles.

These analyses will include investigation of local adaptation (using *e.g.*, data from C2). Work in *Arabidopsis*[31] has shown that loci deleterious in one population and environment may be adaptive in another. We will expand on our previous analyses in teosinte[56] to the full genome, and include evidence of local partial sweeps. These analyses will allow us to evaluate how genetic load varies across space, correlates with environmental variables, and is affected by population size and demographic history.

Recombination: Population genetic theory posits a central role for recombination in determining the effectiveness of natural selection, and previous work from our group[27,44] highlighted the effects of recombination on the fate of deleterious variants in maize. Here we will use new genomic data from maize and teosinte to estimate fine-scaled patterns of recombination along the genome. We will use our high-resolution diversity data to, for the first time, pinpoint gene conversion and recombination hotspots along the genome. We also expect to identify a wide range of small inversions that likely have important effects on genetic load. Finally, we will use data from C2 and C3 to build new genetic maps of both teosinte and maize.

Resources

This section will produce a large number of population genetic PDs that will be analyzed in sections A3 and B2. These include contextual PDs (*e.g.*, DFE in a window around the variant, local recombination rates) as well as PDs specific to individual sites (*e.g.*, dN/dS, GERP score, etc.). Our analyses will also identify important relationships among PDs, such as between evolutionary rate and recombination.

We will create a number of new LD- and meiosis- based maps of recombination and gene conversion in maize and teosinte, and will provide the first fine-scale genetic maps in outcrossing *Zea*. These maps will serve as important resources for mapping traits, assembling contigs in a pan-*Zea* genome, and documenting recombination variation among individuals. They will also allow identification of novel structural variants such as inversion polymorphisms[57].

A3. Data Mining to Predict Functional Alleles

Information from genomic annotations (PDs) should allow us to model the effects of variants on fitness-related phenotypes. A wide range of molecular biology and population genetic criteria can be used as PDs for every variant in the genome (section D3 and Appendix 1). In this section we will evaluate the general utility of PDs and use these results to identify the most informative set of PDs for genomic prediction (B1). Although the scale of the data sets involved is much larger than normally encountered in plant genetics, application of data mining approaches developed in other fields should enable their analysis

Goals

Our goal in this section is to understand which biological features of sequence variants (PDs) best serve as predictors of their functional effect. We will model the effects of PD classifications on allele frequency and also on allele phenotypic effects estimated through GWAS (section A1). Data mining approaches will be used to combine the wide range of PDs into a unified model for prediction of overall constraint and effects. These models will be evaluated via cross-validation. We will then harness the additional predictive power provided by informative PDs to improve prediction of whole plant phenotypes.

Analyses

Analyses will be conducted using data collected from the experiments described in sections C and D. Functional variants impact both the phenotype of the organism and the evolutionary history of the surrounding sequence. We will evaluate the predictive power of each PD using response variables that encompass both types of signal. Phenotypic effects will be captured in whole plant phenotypes (C), SNP selection in GWAS models (A1), and allele effect estimates from GWAS (A1).

The hypotheses to be tested are the polymorphism descriptors (PDs) derived in A2 and D3. Ultimately, we expect several hundred PDs to be developed, from which we will attempt to identify the most relevant. Data will be differentially weighted in the analysis by resolution and effect estimate accuracy. A schematic and hypothetical sample of the data table that we will generate is shown in Table 1.

	Response variables			Independent variables based on PDs (annotations)		
Site	Allele Freq in Zea	GWAS SNP Model Selection	Minor Allele Additive Effect	Nonsynonymous	SIFT Score	Species Conservation
1100	0.01	Yes	-0.3	Yes	0.01	90%
200,000	0.09	No	0.01	No	N/A	25%
5,000,000	0.02	Yes	-0.2	Yes	0.2	95%
5,000,001	0.00	N/A	N/A	Yes	0.3	80%
5,000,002	0.00	N/A	N/A	Yes	.001	100%
5,000,003	0.00	N/A	N/A	No	0.6	20%

Table 1. Schematic, hypothetical version of the data table to be mined to evaluate the predictive power of polymorphism descriptors (PDs). The full data table will include all >2.3 billion sites in the pan genome, all of the response variables outlined in the text, and the full set of several hundred PDs.

Estimating Effects of PDs

(i) Whole plant phenotype: A variety of association analysis models that collapse rare variants within genomic segments have been developed by human geneticists. One method that will be useful for maize data takes quantitative traits as the dependent variable, weights variants by allele frequency, and incorporates functional annotations into a score[58]. We will evaluate the effectiveness of this method using the data from the field experiments in section C.

(ii) SNP selection in GWAS analysis: We have already been conducting this style of analysis in NAM (see introduction to A). In this new work we will expand it substantially both in terms of the range of germplasm analyzed and the number of PDs examined. A standard GWAS analysis will be run (A1). Variants will be categorized as significant or non-significant using a series of alpha levels. For categorical PDs, counts of each PD class in the two groups (significant and non-significant) will be compared and tested for equality using a chi-square test. For continuously variable PDs, a t-test will be used.

After these initial screens for informative PDs, we will combine multiple PDs together through multiple regression and data mining approaches. We have successfully used these approaches in the past to call SNPs from whole genome sequence[33] and to map presence/absence variants[59]. We will use a combination of decision trees, support vector machines (SVMs), and hybrid approaches like the model tree M5' algorithm[60,61].

(iii) Minor allele effect estimates derived from GWAS: We believe that this will be a more powerful extension of (ii), where, rather than a categorical assessment of statistical significance, the response variables for each variant will instead be estimates of additive effect, dominance effect, and GxE variance. The analysis will be similar to that in (ii), except that regression and logistic regression will be used in the initial screens for informative PDs. Subsequently, multiple, informative PDs will be combined together using similar data mining approaches as in (ii).

(iv) Allele frequencies from whole genome sequencing data: By the time that this analysis is carried out, we expect that over 1000 whole genome sequences will be available from existing or future public data and from the results of D1. For each site in the genome, we will have estimates of allele frequency in each of the various germplasm pools. We will also have allele frequency estimates for many of the presence/absence variants. These allele frequencies will be used as response variables to test the predictive power of each PD. The analysis will be identical to that in (iii), except with allele frequency as the response variable.

Preliminary Results

We have already demonstrated the informative power of PDs in several cases. For example, we found a very strong relationship between residual heterozygosity and recombination rate in NAM inbreds (Figure 2), which is also paralleled by higher heterosis in low recombination regions[25,26] This observation suggests that low recombination regions will exhibit greater genetic load, and recombination rates are likely to be a useful PD in predicting functional rare alleles. We have also shown significant enrichment of GWAS associations for individual PDs4,5,9,15,16,33,62, including SNPs with lower methylation levels and copy number variants (Figure 3).

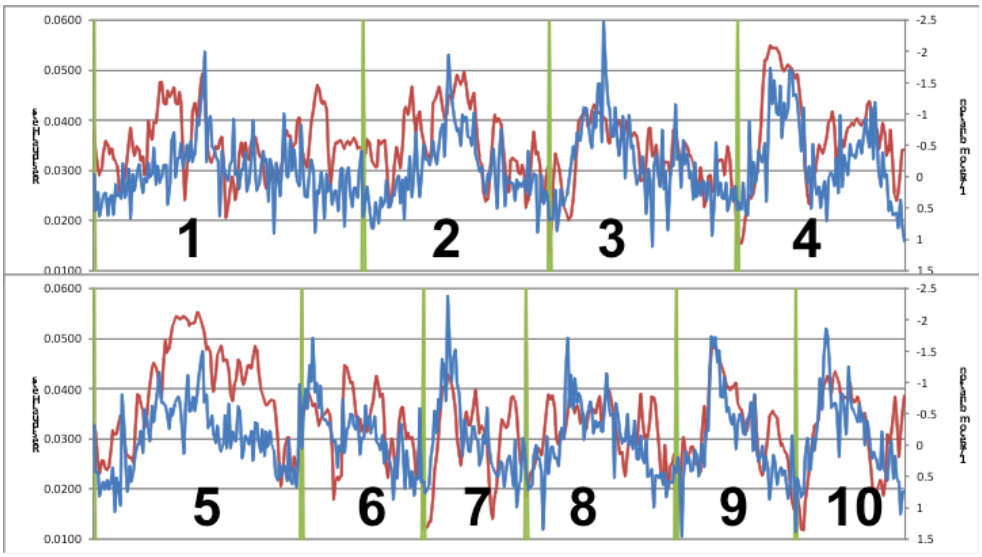
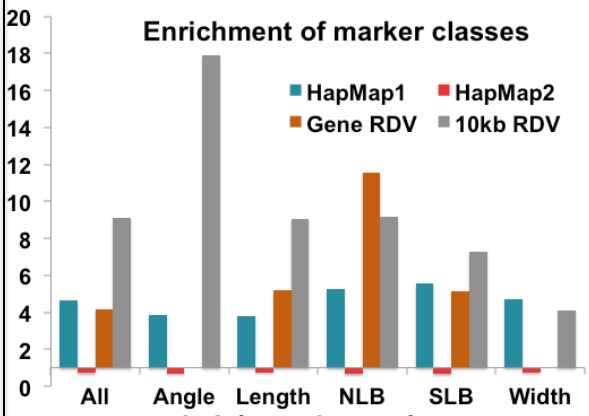


Figure 2. Residual heterozygosity (red) are correlated with the inverse of cross-over rates (blue) in maize NAM populations across 10 chromosomes (r^2=36%)[44].



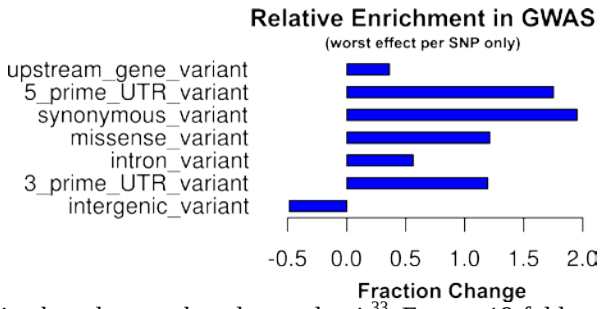


Figure 3. On the left, enrichment of GWAS association based on marker class and trait[33]. Four to 18-fold enrichment was found for read depth variants (RDV), which are indicative of copy number variation, and enrichment was seen for hypomethylated HapMap1 SNPs identified using methylation sensitive sequencing. On the right, observed patterns of NAM GWAS enrichment within genes using the same SNPs across 30 traits (unpublished results).

B. Phenotype Models Incorporating Rare Allele Predictions

While section A focuses on the basic research questions of what biological hypotheses (PDs) predict functional alleles, section B focuses on applying that basic research to predicting phenotypes as a way to accelerate planting breeding. Over the last decade, traditional breeding has begun to change as marker-based genomic selection replaces phenotypic evaluation. Accuracy of current genomic selection models are high between the training and validation datasets for the first meiosis, but then decays quickly. Can we improve model accuracy over more generations by integrating prediction of rare allele causal effects?

The aim of section B is to create models of genome wide prediction that enhance current models by incorporating more information about functional alleles and predicted effects of deleterious rare alleles. Section B1 will use the genotypic and phenotypic data from C & D combined with the best hypotheses

from A to evaluate approaches for genome wide phenotypic prediction. Section B2 will focus on validating our best PDs (A3) and testing their value in genome wide predictions of B1.

If successful, these approaches will facilitate a more rapid advancement of global breeding efforts and a more effective use of beneficial alleles in germplasm banks.

B1. Predict Overall Fitness Using PD Trained Models

Just over a decade ago, animal breeders extended marker assisted selection methods to use all available genetic markers to predict phenotype, regardless of whether or not they are significantly associated with the target traits[7]. This extended method is referred to as "genomic selection" (GS). Evaluation of GS, both by computer simulation and actual implementation, has shown that it results in improved prediction accuracy and shorter breeding cycles[39,62], yet retains genetic diversity if carefully monitored[63–65].

In general, GS methods are accurate only when the training set (the lines both genotyped and phenotyped) and prediction set (only genotyped) are fairly closely related and/or separated in time by only one to three breeding cycles. The methods do not work well for rare alleles that are absent from the training set. Hence, a primary question that will be addressed by this research is whether or not the use of additional biological information in the form of PDs can improve genome wide prediction, both for widely separated training and prediction sets, and for highly diverse and less thoroughly phenotyped training sets (*e.g.*, landraces).

Research Summary

This section will investigate the degree to which genomic selection models can be improved by incorporating the most informative polymorphism descriptors (PDs) identified in section D3 and tested in A3. This work will be carried out for two reasons. The first is to validate the collective set of hypotheses represented by the informative PDs; if these hypotheses are correct, then incorporation of these PDs as explanatory variables should improve phenotypic prediction models. Second, enhanced genomic selection models has the potential to accelerate the breeding of new, improved varieties with diverse germplasm.

Hypothesis

Models that incorporate PD information about rare variants will predict phenotypes better than the same models without that information.

Analyses

Genomic prediction models will be compared using different combinations of training and validation data sets. To make the comparisons, models will be built using training sets and then used to predict phenotypes for the validation sets. The global *Zea* trials (C1) will provide a comprehensive set of data for cross-validation of genomic prediction models. Data will be subdivided into subsets that share locations and entries. Validation will be performed both within and between these data sets. Accuracy will be used to determine whether the model is superior[66].

Rare alleles and PDs will be incorporated through two general approaches. For rare alleles that are sufficiently represented for accurate estimation of their phenotypic effects (*e.g.*, those present in the founders of our various study populations), we will use PD predictions to create separate classes of variants and use these in phenotypic prediction. An extension of ridge regression genomic selection[6] will be used that separates markers into common and rare classes and fits separate variance components for each class. We will also explore Bayesian methods that provide considerable flexibility to model prior beliefs about the data being analyzed. We will extend Bayes Cπ[67] by creating separate classes of variants defined by combinations of PD and frequency. Each class will have its own variance and value of π. The resulting analysis will be useful not only for predicting phenotypes but also for estimating effects for individual variants. Because of the large data sets in this project, computing time is likely to be an

important limiting factor. As needed, we will implement computationally efficient gBLUP analyses using relationship matrices based solely on genetic variants that are influential for the trait considered[68], including the rare variants and PD identified in section A3.

The above approaches should be most useful when the rare alleles are replicated in more than 10 genotypes. For the even rarer alleles, we will develop "genetic burden" statistics to predict the genetic load present in each individual, based both on informative PDs and on the combined PD models developed in section A3. We will then use these genetic burden statistics in a data mining model to predict phenotypes. Given the importance of dominance to load, with both additive and recessive burden will be predicted.

B2.Testing of PD Hypotheses with Maize Hybrids

Sections A3 and B1 focus on making specific predictions about fitness effects of sequence variants and their contribution to phenotypic variation based on PDs. Here we focus on hypothesis-driven field experiments to validate PD predictions and test the utility of including rare alleles and PDs in improving trait prediction.

Our completed sets of germplasm, adapted to yield trials in the U.S. and already genotyped with GBS and WGS (partially) include:

1) 2,500 Ames inbreds – Of these, 1,500 (including 150 exPVPs) have appropriate flowering times (+/- 7days of B73)
2) 5,000 NAM RILs – 3,500 with appropriate flowering
3) *Zea* synthetic DH (C3) – 1000 with appropriate flowering when available

With these core sets of adapted, diverse, and genotyped germplasm, there are over 10 million possible hybrids that can be made to address specific PD hypotheses.

Research Summary

Our objective is to validate in the field the fitness effects of specific PDs predicted in sections A3 and B1. Given that we have access to many germplasm sets and have estimates of the PDs in each line, we can custom design field trials to test specific PDs by choosing from millions of possible hybrid combinations. Two PDs or enhanced genomic selection prediction models will be tested per year in hybrid field trials across the U.S.

Approach

Using *in silico* methods, we will predict which hybrids would best represent the tails of the distribution of predicted effects for a given PD. Care will be taken to control for confounding factors such as flowering time (as mentioned above), population structure, and differences in other PDs of likely importance between the two groups to be compared. Hybrids deviating from the center of the distribution for any of these potentially confounding effects will not be considered for testing. The relatively low LD in diverse maize collections is expected to make possible the selection of hybrid groups differing in mean frequency of a single PD class but not for other PDs. Frequently, this will be tested in similar hybrids expect one group is highly homozygous for a PD class of rare variants versus another the other being highly heterozygous.

Each year we will conduct a trial of two sets of 500 hybrids (250 high crosses, 250 low crosses) specifically chosen to test two different PDs. For example, in the first year we will conduct a pilot study designed to test the effect of heterozygosity in the pericentromeric regions due to reduced recombination. These 500 hybrids will have average heterozygosity across the genome, but differ in the level of heterozygosity in pericentromeric regions (high in pericentromeric regions versus high heterozygsoity in the arms). This experiment will test the hypothesis that there is greater load due to heterozygosity in pericentromeric regions that results in differences in yield. In years two and three, we will choose the

most promising PD hypotheses (from sections A3 and B1). In years three and four, we will compare GS prediction based on common alleles versus various total load models, where the latter includes information about PDs and rare alleles (from section B1). In Year 1, the pilot trial will be conducted within the project only (MO, NY, and NC), whereas in subsequent years we will request assistance from seed companies to run the yield trials at more locations.

Analyses

For each PD, we will test whether there is a significant difference between the two contrasting PD sets of hybrids for each trait after adjusting out non-genetic effects such as block and environment effects.

C. Evaluation of Rare Alleles in the Field

To understand the functional effects of rare alleles, we will evaluate their effects in the field. Through prior efforts of this project and the USDA, international collaborations, and other ongoing efforts, section C1 will build a large dataset of field trials and genotypic data that can be used as the baseline for evaluating the effects of rare alleles. Most of this data will come from maize inbred and hybrid trials based on improved lines (*e.g.*, maize NAM), which provides powerful replication and large families. Many of the alleles will have been replicated in numerous environments, which will enable GxE quantification.

The existing field data are limited to inbred lines and their hybrids, such that the most extreme deleterious alleles that may exist in outcrossing maize populations were likely purged by selfing. To complement this limitation, we will conduct a set of experiments to evaluate rare deleterious alleles in teosinte and landraces in section C2. C2 will quantify levels of genetic load and map large effect deleterious alleles in germplasm that has never been inbred. Section C3 creates a community resource of a synthetic population of improved maize and teosinte, which allows us to compare the effects of alleles derived from teosinte to those from maize in a common genetic background and environment.

Overall, this effort will evaluate nearly 89,000 genotypes covering a range of environments and germplasm and provide a comprehensive training dataset for the effects of rare alleles. These efforts also create important links to other communities: C1 links this project to breeding germplasm efforts, C2 links to inbreeding depression experiments in outbred species (*e.g.*, humans and *Drosophila*), and C3 links to the breeding and genetics communities and provides a lasting community germplasm resource.

C1. Integration of Global *Zea* Fitness Trials

Recently, public maize research groups from around the world have been designing larger and more genetically comprehensive experimental populations to study complex traits. Changes occurring throughout the maize community will make it possible to unify these analyses. First, common large organized germplasm sets are being phenotyped by collaborative research groups across similar climatic zones. For example, maize NAM has been grown by numerous groups throughout the U.S. and China, the CIMMYT-led SeeDs Project is evaluating large samples of landrace topcrosses across numerous sites in Mexico, European scientists are collaborating on evaluating large panels of European Dent and Flint varieties, and Chinese researchers are developed a Chinese NAM platform to be evaluated across China with U.S. NAM.

Second, genotyping platforms and data have become transferable between studies. With a reference genome, the Illumina 55K array, and GBS and WGS resequencing data, it is now possible to identify genomic regions of similarity and dis-similarity among most of the key *Zea* germplasm across the world. Finally, the recognized need to maximize statistical power is fueling interest in sharing and unifying data. The large sample sizes and comprehensive sampling of maize diversity will be vital to understanding and estimating allele effects, particularly rare alleles absent in smaller scale studies. However, data sharing

alone is not enough. To make a truly collaborative and global platform for maize complex trait genetics, a substantial curation effort is required, to organize data in a universally accepted framework. Furthermore, innovative quantitative approaches will be required to combine these diverse data sets in a common analysis.

Research Summary

This project will coordinate curation and analysis of large maize genotypic and phenotypic studies that will facilitate the analysis of rare alleles globally. We will curate data for nearly 89,000 genotypes and over 460,000 trial plots of diverse germplasm (Appendix 2). We will then develop approaches for imputation of complete genotypic data for these samples. We will develop quality control checks to help identify sample mix-ups and bad phenotypic trials. Using genotypic data to estimate genomic relationships among all lines and hybrids in the analysis will permit mixed model/BLUP-based prediction of breeding values and phenotypes of all germplasm. To support this effort, we will develop the bioinformatic tools to curate, store, and share the data.

Data Collection

To have the highest QTL resolution and accuracy of effect estimation, we will unite the public field trials for flowering, height, and yield from numerous research groups around the globe (additional traits will be included when available). Our group and its members have already played a leading or collaborative role in many of these trials. This leveraging of numerous groups' efforts should be of great value to this project as, for example, we estimate that the yield trials alone cost nearly $9M (460,000 total plots x $20/plot = $9.2M) and represent millions of phenotyped plants (Appendix 2). We will remain opportunistic to incorporate other large studies as they become publicly available (*e.g.*, European efforts, NIFA-ATLAS, NSF-MAGIC, USDA-GEM, etc.).

Genotypic Data

All public, improved maize inbreds have already been genotyped by GBS, or are in the pipeline to be genotyped by the end of 2012. Given the inbred line data, computational prediction of the genotype of any hybrid made by crossing a pair of lines is trivial. WGS is currently available for a substantial portion of this germplasm through sequencing or imputation, but additional WGS is still required for unique IBD blocks and parental plants of diverse material (*e.g.*, sections C1-3). This sequence is being developed in D1.When possible, collaboration on genotyping will facilitate additional unification (*e.g.*, the Mitchell and Buckler groups are currently collaborating with EU groups to conduct GBS on their samples).

There are wide ranges of genotype imputation algorithms available from human genetics. However, we have found the nature of presence/absence variation in maize, missing data from sampling in GBS, homozygosity of inbred lines, and large shared haplotypes through maize breeding make many human genetic approaches ineffective or inefficient.

We have completed imputation algorithms for:

- Inbred segments (*i.e.*, no phasing required) using a nearest neighbor algorithm
- Bi-parental crosses using a Hidden Markov Model (HMM)
- Synthetic populations with known parental haplotypes

We will develop algorithms for:

- Imputing the presence/absence variants
- Highly heterozygous landraces (in collaboration with CIMMYT), using HMM to combine known phased haplotypes.

Since so much of this data set is derived from controlled crosses, extensive haplotypes (long LD blocks) are common, allowing for very efficient encoding. We will develop haplotype encoding approaches to reduce the data scale issues, which will facilitate sharing and analysis.

The final imputed dataset will be ~200M variable sites on 89,000 taxa.

Phenotypic Data

More than half of the germplasm listed in Appendix 2 will already have been phenotyped for flowering, height, and yield by the end 2012. Information about the field location and other experimental design and protocols will be collected. Ultimately, consistent BLUP estimates will be made for this germplasm within and across the many diverse environments. As outlined in section B2, we will continue to collect field data that compare *Zea* alleles and to test specific PD hypotheses.

Goals

Compile integrated phenotypic and genotypic datasets for all of *Zea*.
Curate the phenotypic and genotypic datasets to remove questionable data.

Analyses

Individual trials will be analyzed separately to obtain summary statistics (mean, error variance, heritability) that will be used to identify potential problematic environments. Many of the trials will have single replication data, but will include repeated checks that can be used to estimate residual variances and adjust unreplicated experimental entry values for field block effects[69]. We will develop a pipeline in R for automating these analyses (with mixed models analyses handled by calls to ASReml-R package). In some cases, cooperators may already have conducted these initial analyses and we will simply check that similar results are obtained to ensure that data have been curated accurately.

We will sequentially build comprehensive combined analysis models, trying to first join sets of genotypes that share the greatest similarity into a joint model. Realized genomic relationships will be estimated from the GBS data and will be used to model the variance-covariance matrix of all genotypes included in analysis. This is the key step for joining datasets with few or no genotypes in common: the realized genomic matrix provides information on the expected covariances of genotypes within and across different datasets. Predictions (BLUPs) for all genotypes within each site are a natural output of this type of model.

As the joint analysis encompasses more diverse environments and genotypes, we will permit the mixed model to estimate distinct genetic variance within different environments and different genetic correlations between each pair of environments. This will determine which pairs of environments result in similar or very different genotypic values. If we join data sets that have low genetic correlations between them or very distinct GxE patterns, we may encounter model convergence difficulties; we also expect the heritability for genotypes across such diverse sites to decrease. We will check the accuracies of line predictions in environments in which they were not tested. If these accuracies are poor, we will conclude that the data sets should not be joined. We will attempt to join data sets into as few combined analyses as possible. We expect to be able to unify the phenotypic data into 5 major groups as mentioned in section A1.

Resources

In order for each of our collaborating groups to get appropriate credit (*e.g.*, CAAS, CIMMYT), we have agreed that our group will generally lead the first publication for flowering and height traits, while the collaborator's group will lead the first publication on yield or other traits. After these publications, the phenotypic data will be will be made available to the public through the Panzea and MaizeGDB websites.

Preliminary Results

U.S. NAM and Ames germplasm has already been curated as above for GBS data. Heritability studies have been reported for a range of traits in U.S. NAM.

C2. Evaluation of Deleterious Rare Alleles in Teosinte and Landraces

Teosinte, landraces, and improved maize all harbor genetic load, but differ due to the combined effects of domestication, human selection, and the more recent selection under inbreeding and cross-breeding in maize. Because teosinte and landraces are predominantly outcrossing and have not been deliberately inbred, they are expected to contain more deleterious alleles with larger effects compared to improved maize. During the creation of modern maize lines via inbreeding, deleterious alleles have been partially purged and any remaining deleterious alleles will have smaller effects compared to teosinte.

The existing data for fitness related traits in maize are focused on inbreds and hybrids derived from improved maize developed over the last century. In contrast, teosinte and landraces provide the perfect opportunity to study the natural distribution of rare alleles and genetic load. This section will evaluate load and inbreeding depression in samples directly derived from outbred populations. We have identified six teosinte/landrace population pairs that are either sympatric or separated by no more than 18 km. These experiments will estimate the effects of variants (and thus PD classes) on fitness in outbred and partly inbred plants, allowing us to determine their effects on genetic load and inbreeding depression. This section will also provide a wider sample of phenotypes and genotypes for part A of the proposal than are otherwise available from improved maize.

A key aspect of this section is that it likely captures a wide range of severely deleterious rare alleles. This will aid GWAS (A1) and will broaden the range of effects to be modeled in the PD data mining models (A3). In addition, the genomics that will be performed on this material (D) will allow estimation of allele frequency (A2), definition of targets of selection (A2), and assembly of a Pan-*Zea* genome (D2).

Research Summary

We will estimate the contribution of sequence variants to both load and ID, the distribution of load/ID as a function of features of the genome (*e.g.*, recombination, gene content), and the impact of domestication on load and ID. We have sampled populations of wild teosinte and Mexican landraces and will use controlled matings to produce progenies with elevated inbreeding to expose loci that contribute to load. We will deeply sample one teosinte population and one maize landrace from central Mexico near the center of maize domestication (Experiment C2.I). We will also provide a broader assessment of load and rare alleles in teosinte and maize by sampling 10 additional populations (5 teosinte and 5 maize) from a range of environments across Mexico (Experiment C2.II).

Goals and Hypotheses

The overall goals of these experiments are to estimate levels of genetic load and ID present in teosinte and maize landraces, to measure the contribution of each individual sequence variant and each genomic region at the level of haplotype blocks to load and ID, and to test predictions of the functional effects of sequence variants on load and ID. The experiments and measurements proposed will allow us to address the following key questions:

1) How much genetic load/ID for fitness traits is carried by teosinte and maize landraces?
2) Is genetic load related to genetic diversity or local adaptation, and is it stratified geographically?
3) What specific genomic and genic features (polymorphism descriptors) are associated with differences in genetic load and ID across the genome?
4) To what extent does genetic load arise due to Hill-Robertson interference[24] and strong artificial selection unrelated to fitness?
5) How well do PDs predict load in diverse germplasm?

Importantly, this study will sample a set of large effect deleterious rare alleles for inclusion in the global nucleotide analyses in A and B.

Analyses

Experiment C2.I

For Experiment C2.I, we will sample a large natural teosinte population (Palmar Chico, Mexico)[70], and a single population of a nearby maize landrace (Tuxpeño type). For teosinte, we will grow 50 teosinte plants from this population and both intercross and self-pollinate these plants to produce seed that results from a 50:50 mix of selfing and outcrossing. This process will produce seed with half F≈0.5 and half F≈0. A sample of 2,500 plants from the above crossing scheme will be grown in each of two environments (5,000 total plants). Fitness-related traits will be scored on the plants including flowering time, plant height, above-ground biomass, and yield (estimates of total seeds produced, seed weight). The genomes of each of the 50 mother plants will be resequenced (*see* D1). The 5,000 progeny will be assayed using genotype-by-sequence (GBS) technology (*see* D2). Seed for the matching maize landrace will be produced by the same protocol and 5,000 progeny assayed in the same manner.

Experiment C2.II

Data from Experiment C2.I will provide a detailed assessment of load in a single population. To provide a broader assessment across multiple populations, Experiment C2.II will repeat the design of Experiment C2.I in samples from five additional paired landrace and teosinte populations, but sampling only 10 mother plants per population. Selection of the five teosinte populations will be based on eco-geographic diversity as well as available estimates of N_e for these populations. For each population, the 10 mother plants will be selfed and intercrossed to produce a 50:50 mix of selfing and outcrossing. In total, 50 progeny from each of the 50 mothers (10 per population) will be assayed in each of two environments for a total of 5,000 plants. Five landrace populations will be selected from locations as near as possible to the teosinte populations. Ten mothers from each of these will be selfed and intercrossed in a manner parallel to that done for teosinte. WGS of the mothers and GBS and phenotyping for the progeny will be performed as described in Experiment C2.I.

Genotyping, Parentage Analysis, and Inbreeding Coefficients

On each parent, WGS to 30X will be performed and the offspring will be genotyped via GBS (D2). Parental identification, imputation and phasing will rely on tools already built (C1) with some slight modifications for this design. Inbreeding coefficients will be calculated from phased GBS data. We will obtain an RNAseq profile of each population by sequencing the transcriptome of one seedling from each of 12 teosinte and 12 maize parents from experiment C2.I. and from each of 6 teosinte and 6 maize parents in each of the 5 populations from experiment C2.II (84 RNA samples in total). These seedling transcriptomes will assist in identification of novel expressed genes, codon usage, putatively deleterious differences in expression, and splice site variants.

Estimation of Additive Genetic Variance and Heritability for Fitness

Once we have parentage of each progeny plant identified, we can construct the pedigree relationship matrix, which will permit estimation of additive genetic variance. First, we can perform a classical half-sib design analysis[71] on the basis of paternal half-sib families (to eliminate bias due to maternal and cytoplasmic effects). Next, we can conduct a generalized analysis that estimates the additive variance using the complete pedigree relationship matrix (equivalent to the animal breeding A matrix). A factor for common maternal parents can be included in the model to account for maternal and cytoplasmic effects. Finally, we can estimate the pairwise realized genomic relationships among all individuals (equivalent to the K matrix of Yu et al., 2006[48]) and estimate the additive variance by fitting this relationship matrix.

GWAS for Fitness and Inbreeding Depression

We will conduct GWAS for genetic load and ID for at least two distinct scales of genomic variation as power is likely to vary. To map specific variants contributing to genetic load at high resolution and test hypotheses regarding specific polymorphism descriptors, we will first perform GWAS as described in A1. GWAS resolution will be variable depending on the exact genomic context of the site and effect sizes. For analyses involving single SNPs (section A), these associations will be weighted by their predicted resolution (*e.g.*, # of SNPs with LD r^2>0.4). To test hypotheses regarding the genomic spatial pattern of alleles contributing to genetic load and the influence of larger scale genomic context (such as local recombination rates) on genetic load, we combine results at a larger scale.

We will estimate both additive and dominance effects of sequence variants by constructing two genetic effect coefficient matrices for the sequence variants, one for additive, and one for dominance effects. The contribution of each variant to panmictic genetic load (L_i) and to ID load (ID_i) will be estimated as functions of their genetic effects and allele frequency. GWAS for genetic load will require fitting an effect for the inbreeding coefficient of each individual to account for differences in average inbreeding level between outcross and selfed progeny.

C3. Comparative Evaluation of *Zea* Alleles

The teosinte and landrace studies outlined in the previous section (C2) will have powerfully surveyed a wide range of these alleles in their normal genomic context. However, the fitness effects of specific variants segregating in teosinte, landraces, and improved lines may be difficult to compare, because their background genomic contexts will be quite different across the subspecies, as will whole plant morphology and physiology underlying fitness. Therefore, in this section, we will conduct an investigation into the contributions of allelic variants throughout the genome to fitness under varying levels of inbreeding for *both teosinte and maize alleles segregating in a common population*. This should provide a direct comparison of allelic effects in *Zea* taxa both with and without a history of deliberate inbreeding.

Section C2 & C3 also complement one another in allelic range versus power. Thus, although fewer rare alleles will be sampled in this experiment, they will occur at a higher frequency in these replicated experimental populations, permitting more accurate estimation of their effects on phenotypes.

Research Summary

Our objective in this section is to measure the contribution to genetic load of sequence variants and to estimate effects sizes in a population segregating for alleles both with and without a history of deliberate inbreeding. We have created an intermated "*Zea* Synthetic" population based on the 27 NAM founders (including Mo17) and 11 teosinte gametes (Figure 4). Because the NAM founders are inbred lines, their breeding histories have included strong purging selection against deleterious alleles. In contrast, the teosinte contributions to the *Zea* Synthetic originated from open-pollinated teosinte accessions, and thus have never been exposed to intentional inbreeding. Comparison of maize and teosinte allele effects simultaneously in a common reference population, and at varying levels of inbreeding in the various experiments herein, will test the combined effects of domestication and inbreeding. This will extend the inferences obtained from experiments C2.I and C2.II, where maize alleles from non-inbred maize landraces were compared to teosinte alleles.

We will measure ID and effects in the *Zea* Synthetic by evaluating paired outbred and partially inbred families derived from the synthetic. We will conduct GWAS to measure the effects of individual sequence variants and haplotype blocks (composed primarily of either maize or teosinte alleles) on fitness and ID (experiment C3.I). We will also create doubled haploid (DH) lines from the *Zea* Synthetic and determine which alleles are purged during the production of DH lines (experiment C3.II). By evaluating

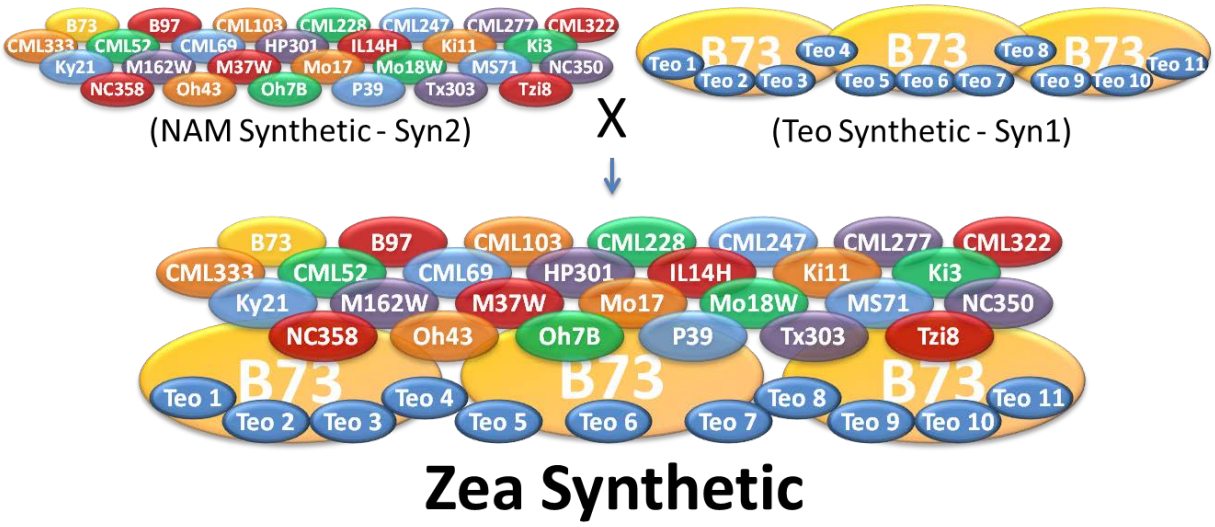


Figure 4 Development of the Zea Synthetic brings together alleles that have diverged thousands of years ago.

hybrids produced from the DH lines, we will then determine the fitness effects of those rare alleles that survive purging in the DH production process.

Approach

The *Zea* Synthetic was created by crossing two other parental synthetic populations: the NAM Synthetic (created from NAM founders) and the Teosinte Synthetic (created from F1s between B73 and 11 geographically diverse *parviglumis* accessions). The *Zea* Synthetic was random mated four generations prior to the start of this project. The expected parentage of the *Zea* Synthetic is 39.4% B73, 1.85% each NAM parent, and 12.5% teosinte contribution from 11 *parviglumis* accessions.

Experiment C3.I

We will self-pollinate 1,000 *Zea* Synthetic (S0) plants, yielding 1,000 S0:1 families. The same 1,000 S0 plants will be used as males to pollinate a second random sample of *Zea* Synthetic S0 females, yielding 1,000 full-sib (FS) families. The 1,000 S1 families (partially-inbred with F = 0.5) and the 1,000 FS families (non-inbred with F ~ 0) will be evaluated in field trials at NC, NY, and MO in each of two years. Traits related to fitness will be measured, including flowering time, plant/ear height, total seed weight per plant, and 50 kernel weight. The genomes of the NAM parents (plus Mo17) have already been sequenced at high coverage as part of our maize HapMap and HapMapV2 projects. Likewise, the teosinte genomes of the 11 founding F1 plants will be sequenced at 30X coverage (see section D1). The 2,000 parental (S0) plants will be genotyped by GBS (see section D2).

Experiment C3.II

We will create 2,000 doubled haploid (DH) lines from the *Zea* Synthetic via the service provided by AgReliant Genetics. The DH lines will be genotyped by GBS and evaluated for fitness-related traits in three locations (NY, NC, and MO) in each of two years. We will then make topcross hybrids of the DH lines and evaluate them in three locations (NY, NC, and MO) in each of two years. We will primarily collect agronomic trait data on the hybrids, including yield.

Goals and Hypotheses

1) Teosinte genome segments in the *Zea* Synthetic carry greater genetic load than maize segments.

2) Inbreeding depression is context-dependent: alleles favorable in teosinte may be disadvantageous in a maize context. This can be detected by differences in the fitness effects of teosinte-derived alleles measured in the *Zea* Synthetic context compared to the pure teosinte context (experiment C2.I.).
3) Purging during the process of doubled haploid development will further reduce the genetic load of the *Zea* Synthetic population, and will preferentially eliminate teosinte alleles. In rare cases, it is possible that maize inbred genome segments carry load that was fixed during domestication or improvement, such that teosinte alleles could be favored during the doubled haploid production process.

Analyses

Experiment C3.I

We will estimate the additive and dominant effects of SNP alleles based on the observed/imputed GBS genotypes of the parental plants and the expected genotype frequencies in their selfed and crossed families. GWAS will be conducted for S1 and FS families in a combined analysis including realized genomic inbreeding coefficient as covariate for background inbreeding depression. GWAS methods will be used to estimate the additive and dominance effects for each SNP separately, leading to estimates of their contribution to genetic load and ID.

Experiment C3.II

GBS data on the doubled haploids will be used to determine IBD from the original parents for each region of the genome. We will know the frequency of alleles for each region of the genome based on WGS of the outbred parents of the *Zea* Synthetic performed in Experiment C3.I. Comparison of allele frequencies in the outbred generation to those of the DH lines will indicate which regions of the genome underwent selection during the DH process. Further, we will be able to determine the origin of alleles that may have been purged, whether from a NAM founder or from teosinte, to address the hypothesis that purging will be stronger on teosinte alleles. Data from the field evaluations of DH lines *per se* will permit GWAS to estimate additive allelic effects for teosinte and maize alleles. Data from the field evaluations of topcross hybrids of the DH lines crossed to a common tester (in combination with our ability to identify regions of IBS between a DH line and the tester) will permit the estimation of dominance interactions with tester alleles, providing further insight into the relationship between ID (measured in C2), fitness effects in completely inbred DHs, and heterosis with tester alleles.

Preliminary Results

We have self-pollinated several hundred plants from the *Zea* Synthetic in winter 2011, and conducted preliminary evaluations of these S1s in summer 2012.

Deliverables

- 2,000 DH lines derived from the *Zea* Synthetic – deposit at Maize Genetic Stock Center.
- Seed samples of the original *Zea* Synthetic – deposit 7,000 seeds at North Central Region Plant Introduction Station.

D. Allele Discovery and Characterization

The main activities in section D of this project are next-generation, whole genome sequencing to discover rare alleles across *Zea* (section D1), genotyping by sequencing to track alleles in experimental populations (section D2), and sequence annotation to derive a comprehensive list of polymorphism descriptors (section D3). The technical challenges of capturing, assembling, and characterizing the diversity of *Zea* are substantial, as it has one of the most diverse and dynamic genomes known.

In section D1, we will perform whole genome sequencing on the founders of our experimental populations in order to discover virtually all of the alleles contained therein, with particular focus on the

rare alleles. These experimental populations are described in detail in section C. Included in section D1 is the construction of a pan-genome for *Zea*, based upon all of the available next generation sequence, in order to capture genomic segments not present in the B73 reference genome.

In section D2, we will perform genotyping by sequencing on all of the individuals in the experimental populations (*i.e.*, the offspring of the founders sequenced in section D1). This will allow us to track the inheritance of all of the variants discovered in the founders, resulting in near-complete genotype information for all of the individual plants in our experiments.

In section D3, we describe our efforts to develop polymorphism descriptors (PDs) across the pan-*Zea* genome for every nucleotide. The central hypothesis of this project, examined in sections A and B, is that data mining of this vast catalog of PDs will identify key PDs that can be incorporated into genomic selection models (section B1) and will enhance our ability to predict phenotype from genotype, even for rare alleles.

D1. Whole Genome Sequencing for Rare Allele Discovery

The HapMapV2 pipeline[33] called more SNPs with less error than any previous maize bioinformatic pipeline. We found evidence for over 150M SNPs, but because of power and paralogy issues our pipeline eliminated nearly 100M SNPs in order to keep error rates low. For example, only ~50% of the singleton SNPs were called, as depth of sequence was insufficient and/or genetic proof of position was poor. GWAS mapping of read depth variants suggested that many of these excluded variants and regions of the genome probably have functional effects. In this project, we will build upon our successful HapMapV2 pipelines using deep sequencing of key founders of mapping population, integration of existing community sequences, and a new bioinformatic fusion of physical and genetic sequence analysis.

Research Summary

We will use utilize WGS to identify rare alleles in the founders of the experimental populations described above (section C) and in diverse, underrepresented maize inbreds (section C1). We will sequence 111 teosinte parents (30X), 100 maize landrace parents (30X), and 285 diverse maize (15X). A subset of at least 9 individuals (including some teosinte inbreds) will be targeted for additional sequencing at greater depth in order to perform reference-guided or *de novo* genome assembly. *De novo* genome assembly is important given the dramatic differences in sequence content observed even among inbred maize lines[44,72,73]. These deeply sequenced and assembled individuals will be used to define a pan-*Zea* genome, expanding on the current B73 reference genome. In addition to genome sequence, we will sequence and assemble transcriptomes from a few tissues to assist in the annotation of novel genes. Sequencing will initially utilize Illumina HiSeq technology, which is currently the most cost-efficient next-generation sequencing platform. As new sequencing platforms develop, they will be evaluated and incorporated into our sequencing strategy as appropriate (reallocations will come from the diverse maize sequencing in later years).

Data Collection and Analysis

Figure 5 shows a schematic representation of our strategy for genome assembly and SNP calling, described in detail in Appendices 3 and 4. We will construct a pan-*Zea* genome so that all available *Zea* reads can be aligned both to the B73 reference genome and the pan-*Zea* genome in order to call SNPs and identify copy number and presence/absence variants. As LD between sites within a contig in the resulting SNP genotyping data set will be used to validate the integrity of each contig, pan-genome construction and SNP calling will be an iterative process. Giving the complexity of the genome, this genome will *not* be a perfect set of contigs spanning each chromosome, but rather a comprehensive catalog of Zea genome space with contigs placed in approximate positions.

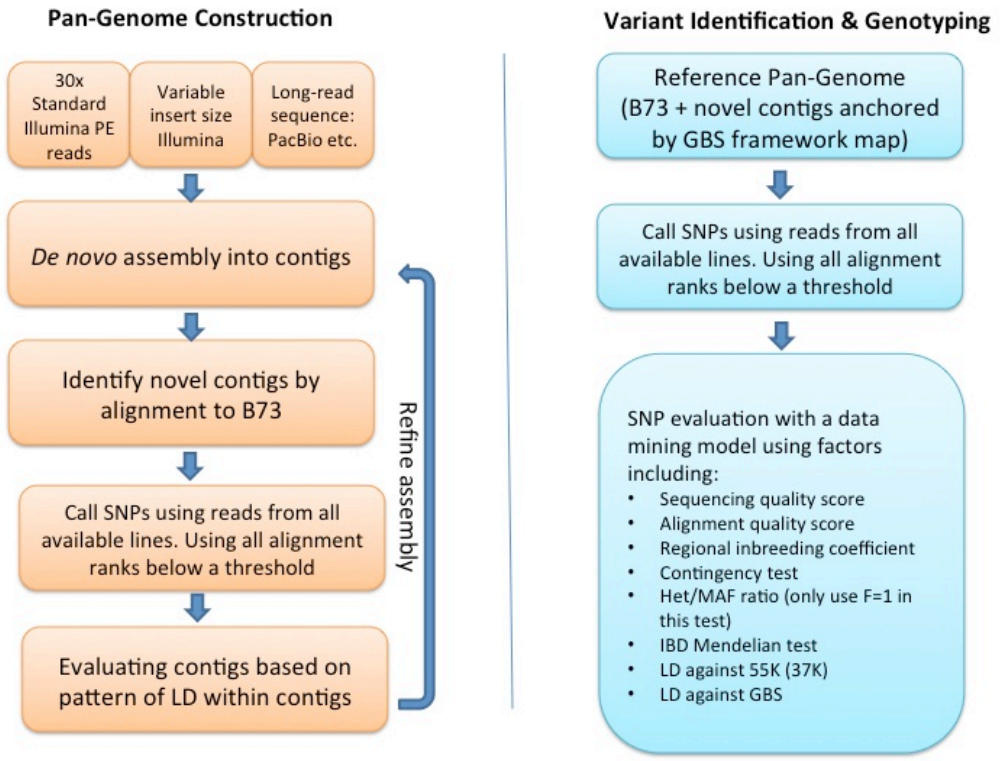


Figure 5. Schematic of analysis strategy for pan-genome assembly and SNP calling from whole genome sequence data.

Computational Implementation

Storage: File servers at UC Davis and Cornell will each host a copy of the raw sequencing data, including whole genome shotgun, RNA-seq and GBS. The raw data will also be deposited to the iPlant Data Store. The Ware group at CSHL will work with the iPlant team to develop data transferring protocols for depositing data to the iPlant file server.

Computing Resources

We will keep local copies of the data files at two sites to facilitate quick access to computer clusters at UC Davis and Cornell University and, most importantly, we will need to leverage the existing NSF Computing Infrastructure the Texas Advanced Computing Center (TACC). We will work with the iPlant team and use the TACC clusters through its XSEDE program. By working with iPlant, we can also make sure the raw data and meta data analysis pipelines are readily accessible to the research community after they are publicly released. The raw data stored at iPlant will thus represent a third copy (in addition to copies at UC Davis and Cornell).

Local clusters at Cornell, UC Davis, and CSHL will be used for prototyping the data analysis pipeline including assembly, WGS SNP calling, and GBS SNP calling. A dedicated 64-core, 0.5T RAM server with 50TB of directly attached RAID-6 storage disks will be set up at Cornell for this project to host the raw data and for development of the assembly and GBS genotyping pipelines. The Ross-Ibarra Lab at UC Davis has dedicated access to 21TB of storage on an 8-node (with 16 cores and 24G RAM per node) system that can be used for short read alignment and SNP calling. The Ware at CSHL has access to 100 compute nodes with 128GB of memory, and two "high memory" compute nodes with 1.5TB of memory.

The development nodes and 128GB compute nodes are Xeon E5-2665's running at 2.40GHz. The Sun Lab and Ross-Ibarra Lab have access to shared computing clusters at core facilities at Cornell and UC Davis, respectively, including a total of seven 0.5Tb RAM nodes that can be used for *de novo* assembly.

Deliverables

All deliverables will be released to the community initially via the NCBI short read archive (SRA) and the panzea.org website, followed by integration into large community resources such as MaizeGDB, iPlant, and Gramene.

1) Raw sequence data for all lines (to be deposited in the NCBI SRA).
2) Annual Panzea builds of sequence alignments and variant calls for all sequenced material. Upon release, each year's build will be immediately made publicly available under conditions similar to the Fort Lauderdale and Toronto agreements.
3) De novo transcriptomes for all lines sequenced.
4) Pan-*Zea* de novo (reference guided) genome assembly.

D2. Genotyping to Track Alleles

Recent advances in DNA sequencing technologies have facilitated development of cost effective and efficient strategies that allow simultaneous SNP discovery and genotyping of large numbers of individuals. These methods exploit the power of next-generation sequencing to obtain massive numbers of DNA sequences from the ends of genomic DNA fragments and DNA barcoding strategies[74] that allow pooling of up to 384 individuals in a single sequencing lane.

Several genotyping approaches based on reduction of genome complexity with restriction enzymes[75] followed by NGS have recently emerged, including the genotyping-by-sequencing (GBS) method developed in the Buckler lab[44,76]. GBS has proven to be very robust and adaptable to any species, and because of its technical simplicity is also extremely cost-effective. Briefly, the GBS protocol[76] involves reduction of genome complexity by digesting total genomic DNA from individual samples with a methylation sensitive restriction enzyme followed by ligation of adapter pairs that allow incorporation of barcode sequence and sequence complementary to sequencing primers. Our current strategy involves barcoding 384 individuals in a single lane of Illumina. With the most recent GBS analysis pipeline, the resulting reads or "tags" allowed genotyping of 2.2M SNPs across the genome for 33K maize lines.

Data Collection and Analysis

GBS will be conducted on 29,000 samples to support the tracking of alleles throughout C. For C1, 1,000 maize samples per year will be genotyped at 384-plex to support integration of community efforts. A substantial number of these will be run to clarify errors identified during curation of existing data. C2 will generate 10,000 teosinte and 10,000 landrace samples that will be genotyped at 96-plex. We will genotype these at higher density as they are highly heterozygous, and this density will facilitate anchoring of genomic contigs and detailed map construction. The *Zea* synthetic (C3) will generate 4,000 samples for GBS at 96-plex.

SNP Analysis and GBS Pipeline

The TASSEL software will be used to process the GBS sequencing data into genotype results in HapMap format. The Buckler, Sun, and Ware groups will collaborate to develop a new file exchange format and database servers for large scale genotypic data. HDF5 will be used as a replacement to the regular SQL-based database system.

We are collaborating with the iPlant team to move the GBS data analysis pipeline to the iPlant infrastructure. Once the pipeline is moved to iPlant, we will keep a copy of the data files on iPlant's iRod storage system, supplementing the storage server in Cornell Rhodes Hall. As a trial, all GBS data from

the NAM population has been copied to iRod. Software developed by the Buckler and Sun groups, including TASSEL, is deposited at SourceForge, using GIT for version control.

Data Analysis and Sequencing Validation

GBS data will be used for a variety of analyses, including parentage analysis (C2), identification of parental haplotypes (sections C2-3), quantification of relatedness or IBS among unrelated lines (C), and association mapping (A1, C). The Buckler, Sun, and Ross-Ibarra labs will work to implement these analyses (see preliminary results).

In addition to direct experimental utility, GBS data will play an important role in validating the WGS data collected in section D1. Linkage disequilibrium estimates from GBS will aid de novo assembly and comparison of GBS results (*e.g.*, IBS or heterozygosity vs. distance from reference) will enable effective validation and filtering of WGS sequence data.

Preliminary Results

To date, we have analyzed 55,000 maize DNA sample runs representing more than 32,000 lines by GBS. These lines represent a broad swath of maize germplasm including inbred lines important for U.S. maize breeding, U.S. and Chinese NAM populations, other populations including those for mapping disease resistance and nutritional quality QTL, CIMMYT landraces, DH lines, wild teosinte populations, teosinte NILs, teosinte-maize hybrids and advanced backcross populations. Figure 6 shows a minor allele frequency (MAF) distribution of alleles genotyped in the USDA Ames inbred collection. These data highlight our successful use of GBS across diverse germplasm, the apparent absence of ascertainment bias in GBS genotypes, and the utility of the approach for genotyping rare alleles.

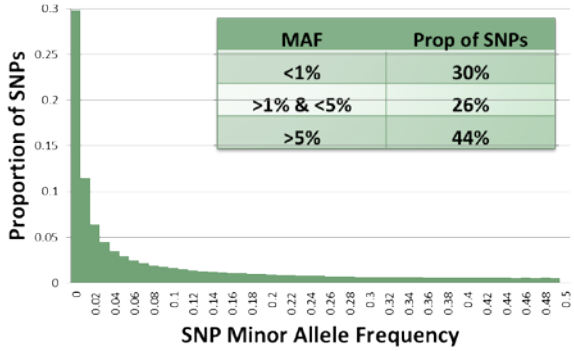


Figure 6. Distribution of minor allele frequency of SNPs genotyped using GBS technology in 2,709 inbred lines from the USDA Ames collection (Romay et al in review).

We have developed hidden Markov models (HMM) to identify parental haplotypes in mapping populations. These models take into account recombination and genotyping error to estimate parental haplotypes and impute missing data from recombinant inbred lines developed in a mapping population. Recent improvements on the imputation accuracy allow us to impute data from biparental populations with less than 0.1% error. Work is underway to extend these models to populations with multiple parents, such as the experimental populations in sections C2 and C3.

Deliverables

1) Raw sequence data for all lines, to be deposited in NCBI Sequence Read Archive.
2) Annual Panzea builds of genotype calls for all lines. This will include recalling previously released lines, as additional data allows exploitation of previously unalignable or unused sequence. Upon release, each year's build will be immediately made publicly available under conditions similar to the Fort Lauderdale and Toronto agreements.
3) A public implementation of the GBS pipeline in TASSEL on the iPlant Cloud system Atmosphere. We are working with the iPlant team to make the GBS pipeline available through the graphic user interface of the iPlant Discovery Environment (DE). A beta version of GBS for DE is available now.

D3. Annotate Maize Genome with Polymorphism Descriptors

The last few years of the genomics revolution have clearly shown that a powerful connection can be made between genomics and functional variation (*e.g.*, ENCODE). Polymorphism descriptors (PDs) are annotations of every nucleotide in the genome that either contrasts the predicted effect of the major versus minor allele or describes the genomic context of that region of the genome. This section focuses on using available datasets that the community produces over the next four to five years to annotate the genome so that the most functionally important sites and rare alleles can be identified in A3.

While many of these PDs can be developed now, we foresee that many will improve and new PDs will become available as various community projects are completed. Our group will work with maize community efforts to incorporate the most relevant PDs, and our focus will be informed by progress in efforts like the human ENCODE project. In each lab, individual graduate students and postdocs will focus on categorizing and analyzing suites of PDs that form publishable units, and the project as a whole will pull these together into a unified dataset for analysis in sections A3 and B1.

Data Collection

Polymorphism Descriptors

Our goal will be to measure, for every nucleotide in the genome, each of 100+ parameters. This will result in a matrix of perhaps 200 billion entries (2.1×10^9 bp x 100 PD) for the genome. A list of example polymorphism descriptors is shown in Appendix 1, but is far from comprehensive or complete. Throughout the course of the project we will add to and modify our list of PDs as new data become available and modeling/experimental results suggest removing or modifying existing PDs. Many of the PDs can be further split by population, *e.g.*, expression level in teosinte vs. expression level in maize, or allele frequency in various subpopulations. The data will be curated in a unified compressed HDF5 format. This format will be developed in conjunction with Gramene and MaizeGDB. Annotations will be publicly shared through Panzea, MaizeGDB, and Gramene, as each paper testing these hypotheses is accepted. We expect this approach to be model for our community and other species.

Hypothesis generation: Each lab's graduate students and postdocs will lead development of a suite of PDs. We also expect that numerous collaborations with our colleagues to create these PDs, so that a team of two students (one from this group, and one from a collaborating lab) will develop these PDs. While the creation of the PDs can be time consuming, there is also a tremendous opportunity for students to develop and test their own hypotheses of how the genome and evolution functions. Each student's paper will merge these PDs (a formalized hypothesis) with A3 & C1 analyses that relate PDs to QTL function. Each lab will likely specialize to some degree. For example, Ross-Ibarra's lab will focus on population genetic PDs, Doebley lab on transcriptional regulation, Ware's lab comparative genomics, Buckler's lab on splicing variation, etc.

Education and Outreach

Background

In collaboration with the Paleontological Research Institute's Museum of the Earth (Ithaca, NY), we developed a traveling museum exhibit on maize diversity, evolution, and genetics that has been very popular for medium size museums. The exhibit has been seen by approximately 300,000 people so far, and is on target to be viewed by 1 million visitors during the life of the project. The exhibit has been displayed at five venues to date, including the famous Corn Palace in South Dakota, and is booked through 2014 (Table 2). Numerous additional venues that have enquired about hosting it. Because of NSF support, there are no shipping costs or fees charged to hosting museums which has enabled many mid-size and smaller museums to host the exhibit.

The exhibit was featured in a number of media items, including a radio interview (Science Cabernet, WICB Ithaca) that is available as a podcast (at www.panzea.org), news articles (Ithaca Times, the University of Iowa newsletter) as well as being a cover feature in the PRI's American Paleontologist magazine.

We also completed the Teacher Friendly Guide™ to the Evolution of Maize, which is now available online (www.maize.teacherfriendlyguide.org), and is being printed out as hard copies to distribute at education-related events. The Teacher Friendly Guide™ series is hosted by PRI and written for educators (www.maize.teacherfriendlyguide.org).

In conjunction with our BREAD project (BREAD: Platform, Pipeline, and Analytical Tools for Next Generation Genotyping to Serve Breeding Efforts in Africa NSF #0965342), we have had seven GBS workshops attended by almost 200 people, at no cost to the participants. We have had participants working on a vast array of organisms, animal, bacteria in addition to crops, and they have come from all over the world, from as far away as New Zealand. Each workshop has filled up within hours of being announced and we are running a continual waiting list. Videos made of our February 2012 GBS workshop posted online have been viewed numerous times and are also available on CD (and have been sent to Nigeria upon request, where the bandwidth prohibits streaming video). We conduct an online evaluation each time and the results have been overwhelmingly positive.

Goals

Traveling Museum Exhibit

Given the popularity of the exhibit, we (together with PRI) plan to update it and make a second copy so that it can continue to travel and hit a wider audience. In addition to museums, we have been getting requests from other venue types such as libraries and botanical gardens.

Table 2. Maize Diversity Exhibit Schedule

Venue	Location	Dates
Museum of the Earth	Ithaca, NY	March – August 2011
Pentacrest Museum, Univ of Iowa	Iowa City, IA	August – November 2011*
Corn Palace	Mitchell, SD	May – August 2012
George Bush Presidential Library & Museum	College Station, TX	August – October 2012
Susquehanna River Archaeological Center of Native Indian Studies	Waverly, NY	October 2012 – January 2013
Weisman Museum of Southwest Texas	Uvalde, TX	February –May 2013
Illinois State Museum	Springfield, IL	May – August 2013
Smith College Botanic Garden	Northampton, MA	September – December 2013
Bell Museum of Natural History	Minneapolis, MN	February – June 2014

*after Iowa the exhibit was brought back to Ithaca for a quality check

With support of this grant, we will easily to be able to fill out the rest of the schedule for the next years. Overall, more than a million people will be introduced to maize diversity, agriculture, genetics, and evolution through this exhibit.

We will continue to update and promote both the exhibit and the related Teacher Friendly Guide™ to the Evolution of Maize (www.maize.teacherfriendlyguide.org) as well, by presenting it at a number of

conferences attended by teachers, such as the Association of Science-Technology Centers meeting (in years 1 and 3), and hosting a booth at the National Science Teachers Association Conference (year 3). In year 4, one person will attend the Society of the Study of Evolution meeting.

Panzea Website

Panzea (www.panzea.org) has been the public face of the project, housing many resources both for the scientific community and the general public. In light of new web-based, social media technologies such as running news updates, RSS feeds, blogs, etc., we plan to revise the Panzea website to make it more current. Furthermore, given that some of the scientific data will be migrating to MaizeGDB and Gramene, we plan to make the Panzea site more accessible and usable by the general public, teachers, museums that may be interested in the exhibit, etc.

Currently the Panzea site includes downloadable data, database searches, germplasm information and other items that will be migrating to MaizeGDB and Gramene. Therefore this is a good opportunity to shift the focus to items of more popular interest, such as where the exhibit is currently located, workshop information, updates to the Teacher Friendly Guide™, links to new videos and other educational material. We will also start a Panzea Facebook page.

Workshops and Online Video Courses

Our GBS workshops are two days long, and focus on what GBS is (both the biology and technology) and how to process the data with TASSEL, via a mix of lectures and hands-on practicals taught by 8-10 members of the Panzea project. These workshops have been attended by students and researchers from all over the world, and have included people working on species from killer whales to dogs to fungi, in addition to numerous plant species. The most recent workshop in September 2012 filled to capacity (35 participants) the day it was announced. We have scheduled the next one for February 6-7, 2013; there are already 40 people registered with 30 more on a waiting list. All the slides, tutorials and documentation for each workshop are posted on the CBSU workshop site along with the agenda, and we update with each workshop (http://cbsu.tc.cornell.edu/ww/1/Default.aspx?wid=34). We will post new videos of our workshop with each substantial topic change, as well as make CDs available upon request.

We will continue to offer workshops every few months in conjunction with CBSU. We also plan to hold a workshop in North Carolina and one in Missouri, and in Year 4 will hold one at the Plant & Animal Genome Conference. There has been interest in hosting workshops in Europe and in South America, with outside support. The subject area of the workshops will evolve with the progress of the project, so in the future will incorporate new bioinformatic approaches for predicting the effects of rare alleles. We expect the topic of the workshop to evolve over the life of the grant, in conjunction with the sequencing technologies and the science that they enable.

We will be offering video short topics online as well. These will include subjects for fellow scientists - on topics ranging from GBS (*e.g.*, preparatory information as introduction to the workshops) to DNA extractions (for which we have already received a number of requests), in addition to topics of more general interest (domestication, what is teosinte, inheritance, evolution, careers in science). We will be creating three to six videos each year; the videos will be posted on the YouTube Education site (www.youtube.com/Education).

References

1. Cooper, G. M. & Shendure, J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nature Reviews Genetics* **12**, 628–40 (2011).
2. Beló, A. *et al.* Whole genome scan detects an allelic variant of fad2 associated with increased oleic acid levels in maize. *Molecular Genetics and Genomics* **279**, 1–10 (2008).
3. Atwell, S. *et al.* Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature* **465**, 627–31 (2010).
4. Tian, F. *et al.* Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nature Genetics* **43**, 159–62 (2011).
5. Kump, K. L. *et al.* Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. *Nature Genetics* **43**, 163–8 (2011).
6. Lorenz, A. J. *et al.* Genomic selection in plant breeding: Knowledge and prospects. *Advances in Agronomy* **110**, 77–123 (2011).
7. Meuwissen, T. H., Hayes, B. J. & Goddard, M. E. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819–29 (2001).
8. Clark, R. M., Tavaré, S. & Doebley, J. Estimating a nucleotide substitution rate for maize from polymorphism at a major domestication locus. *Molecular Biology and Evolution* **22**, 2304–12 (2005).
9. Brown, P. J. *et al.* Distinct Genetic Architectures for Male and Female Inflorescence Traits of Maize. *PLoS Genetics* **7**, e1002383 (2011).
10. Gibson, G. Rare and common variants: twenty arguments. *Nature Reviews Genetics* **13**, 135–145 (2012).
11. Bansal, V., Libiger, O., Torkamani, A. & Schork, N. J. Statistical analysis strategies for association studies involving rare variants. *Nature Reviews Genetics* **11**, 773–85 (2010).
12. Mackay, T. F. C. *et al.* The Drosophila melanogaster Genetic Reference Panel. *Nature* **482**, 173–178 (2012).
13. MacArthur, D. G. *et al.* A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes. *Science* **335**, 823–828 (2012).
14. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* **42**, 565–9 (2010).
15. Buckler, E. S. *et al.* The genetic architecture of maize flowering time. *Science* **325**, 714–8 (2009).
16. Cook, J. P. *et al.* Genetic Architecture of Maize Kernel Composition in the Nested Association Mapping and Inbred Association Panels. *Plant Physiology* **158**, 824–834 (2012).
17. Harjes, C. E. *et al.* Natural genetic variation in lycopene epsilon cyclase tapped for maize biofortification. *Science* **319**, 330–3 (2008).
18. Platt, A., Vilhjálmsson, B. J. & Nordborg, M. Conditions under which genome-wide association studies will be positively misleading. *Genetics* **186**, 1045–52 (2010).
19. Charlesworth, D. & Willis, J. H. The genetics of inbreeding depression. *Nature Reviews Genetics* **10**, 783–96 (2009).
20. Tellier, A. *et al.* Fitness effects of derived deleterious mutations in four closely related wild tomato species with spatial structure. *Heredity* **107**, 189–99 (2011).
21. Gossmann, T. I. *et al.* Genome wide analyses reveal little evidence for adaptive evolution in many plant species. *Molecular Biology and Evolution* **27**, 1822–32 (2010).
22. Cao, J. *et al.* Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nature Genetics* **43**, (2011).
23. Birchler, J. A., Yao, H., Chudalayandi, S., Vaiman, D. & Veitia, R. A. Heterosis. *The Plant Cell* **22**, 2105–12 (2010).
24. Hill, W. G. & Robertson, A. The effect of linkage on limits to artificial selection. *Genetical Research* **8**, 269–94 (1966).

25. Larièpe, A. *et al.* The Genetic Basis of Heterosis: Multiparental Quantitative Trait Loci Mapping Reveals Contrasted Levels of Apparent Overdominance Among Traits of Agronomical Interest in Maize (Zea mays L.). *Genetics* **190**, 795–811 (2012).
26. Schön, C. C., Dhillon, B. S., Utz, H. F. & Melchinger, A. E. High congruency of QTL positions for heterosis of grain yield in three crosses of maize. *TAG. Theoretical and Applied Genetics.* **120**, 321–32 (2010).
27. McMullen, M. D. *et al.* Genetic properties of the maize nested association mapping population. *Science* **325**, 737–40 (2009).
28. Chun, S. & Fay, J. C. Evidence for Hitchhiking of Deleterious Mutations within the Human Genome. *PLoS Genetics* **7**, e1002240 (2011).
29. Lu, J. *et al.* The accumulation of deleterious mutations in rice genomes: a hypothesis on the cost of domestication. *TRENDS in Genetics* **22**, 126–31 (2006).
30. Hufford, M. B. *et al.* Comparative population genomics of maize domestication and improvement. *Nature Genetics* **44**, 808–11 (2012).
31. Fournier-Level, A. *et al.* A Map of Local Adaptation in Arabidopsis thaliana. *Science* **334**, 86–89 (2011).
32. Matsuoka, Y. *et al.* A single domestication for maize shown by multilocus microsatellite genotyping. *Proceedings of the National Academy of Sciences* **99**, 6080–4 (2002).
33. Chia, J.-M. *et al.* Maize HapMap2 identifies extant variation from a genome in flux. *Nature genetics* **44**, 803–7 (2012).
34. Tenaillon, M. I., U'Ren, J., Tenaillon, O. & Gaut, B. S. Selection versus demography: a multilocus investigation of the domestication process in maize. *Molecular Biology and Evolution* **21**, 1214–25 (2004).
35. Wright, S. I. *et al.* The effects of artificial selection on the maize genome. *Science* **308**, 1310–4 (2005).
36. Hallauer, A. R. & Miranda, J. B. *Quantitative Genetics in Maize Breeding*. (Iowa Sate University Press: Ames, Iowa, 1988).
37. Hufford, M. B. Genetic and ecological approaches to guide conservation of teosinte (Zea mays ssp. parviglumis), the wild progenitor of maize. **DAI/B 71-0**, 130 (2010).
38. Frazer, K. A., Murray, S. S., Schork, N. J. & Topol, E. J. Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics* **10**, 241–51 (2009).
39. Goddard, M. & Hayes, B. Genomic selection. *Journal of Animal Breeding and Genetics* **124**, 323–330 (2007).
40. Zhu, C., Li, X., Yu, J. & Yandell, B. S. Integrating Rare-Variant Testing, Function Prediction, and Gene Network in Composite Resequencing-Based Genome-Wide Association Studies (CR-GWAS). *G3: Genes|Genomes|Genetics* **1**, 233–243 (2011).
41. Lohmueller, K. E. *et al.* Proportionally more deleterious genetic variation in European than in African populations. *Nature* **451**, 994–7 (2008).
42. Tennessen, J. A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64–9 (2012).
43. Jelier, R., Semple, J. I., Garcia-Verdugo, R. & Lehner, B. Predicting phenotypic variation in yeast from individual genome sequences. *Nature Genetics* **43**, 1270–1274 (2011).
44. Gore, M. a *et al.* A first-generation haplotype map of maize. *Science* **326**, 1115–7 (2009).
45. Li, H., Bradbury, P., Ersoz, E., Buckler, E. S. & Wang, J. Joint QTL linkage mapping for multiple-cross mating design sharing one common parent. *PLoS ONE* **6**, e17573 (2011).
46. Yu, J., Holland, J. B., McMullen, M. D. & Buckler, E. S. Genetic design and statistical power of nested association mapping in maize. *Genetics* **178**, 539–51 (2008).
47. Zhang, Z. *et al.* Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics* **42**, 355–60 (2010).
48. Yu, J. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics* **38**, 203–8 (2006).

49. Listgarten, J. & Heckerman, D. Determining the Number of Non-Spurious Arcs in a Learned DAG Model : Investigation of a Bayesian and a Frequentist Approach. *UAI 2007* 251–258at <http://research.microsoft.com/pubs/70442/tr-2007-60.pdf>
50. Bradbury, P. J. *et al.* TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 2633–5 (2007).
51. Lipka, A. E. *et al.* GAPIT: genome association and prediction integrated tool. *Bioinformatics* **28**, 2397–9 (2012).
52. Ng, P. C. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Research* **31**, 3812–3814 (2003).
53. Stone, E. A. & Sidow, A. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Research* **15**, 978–86 (2005).
54. Davydov, E. V *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Computational Biology* **6**, e1001025 (2010).
55. Eyre-Walker, A. & Keightley, P. D. The distribution of fitness effects of new mutations. *Nature Reviews Genetics* **8**, 610–8 (2007).
56. Pyhäjärvi, T., Hufford, M. B., Mezmouk, S. & Ross-Ibarra, J. Complex patterns of local adaptation in teosinte. (2012).at <http://arxiv.org/abs/1208.0634>
57. Fang, Z. *et al.* Megabase-scale inversion polymorphism in the wild ancestor of maize. *Genetics* **191**, 883–94 (2012).
58. Price, A. L. *et al.* Pooled association tests for rare variants in exon-resequencing studies. *American Journal of Human Genetics* **86**, 832–8 (2010).
59. Lu, F. *et al.* Switchgrass genomic diversity, ploidy and evolution: novel insights from a network-based SNP discovery protocol. *PLoS Genetics*
60. Wang, Y. & Witten, I. H. Induction of model trees for predicting continuous classes. *Proceedings of the poster papers of the European Conference on Machine Learning, University of Economics, Faculty of Informatics and Statistics, Prague*. 128–137 (1997).
61. Quinlan, J. R. Learning with continuous classes. *Proceedings of 5th Australian Joint Conference on Artificial Intelligence, World Scientific, Singapore*, 343–348 (1992).
62. Heffner, E. L., Sorrells, M. E. & Jannink, J.-L. Genomic Selection for Crop Improvement. *Crop Science* **49**, 1 (2009).
63. Daetwyler, H. D., Villanueva, B., Bijma, P. & Woolliams, J. A. Inbreeding in genome-wide selection. *Journal of Animal Breeding and Genetics* **124**, 369–76 (2007).
64. Dekkers, J. C. M. Prediction of response to marker-assisted and genomic selection using selection index theory. *Journal of Animal Breeding and Genetics* **124**, 331–41 (2007).
65. Lillehammer, M., Meuwissen, T. H. E. & Sonesson, A. K. A comparison of dairy cattle breeding designs that use genomic selection. *Journal of Dairy Science* **94**, 493–500 (2011).
66. Heslot, N., Yang, H.-P., Sorrells, M. E. & Jannink, J.-L. Genomic Selection in Plant Breeding: A Comparison of Models. *Crop Science* **52**, 146 (2012).
67. Habier, D., Fernando, R. L., Kizilkaya, K. & Garrick, D. J. Extension of the bayesian alphabet for genomic selection. *BMC bioinformatics* **12**, 186 (2011).
68. Zhang, N. *et al.* Genetic analysis of central carbon metabolism unveils an amino acid substitution that alters maize NAD-dependent isocitrate dehydrogenase activity. *PLoS ONE* **5**, e9991 (2010).
69. Hung, H.-Y. *et al.* The relationship between parental genetic or phenotypic divergence and progeny variation in the maize nested association mapping population. *Heredity* 1–10 (2011).doi:10.1038/hdy.2011.103
70. Van Heerwaarden, J. *et al.* Fine scale genetic structure in the wild ancestor of maize (Zea mays ssp. parviglumis). *Molecular Ecology* **19**, 1162–73 (2010).
71. Lynch, M. & Walsh, B. *Genetics and Analysis of Quantitative Traits*. (Sinauer Associates: 1998).
72. Fu, H. & Dooner, H. K. Intraspecific violation of genetic colinearity and its implications in maize. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 9573–8 (2002).

73. Brunner, S., Fengler, K., Morgante, M., Tingey, S. & Rafalski, A. Evolution of DNA sequence nonhomologies among maize inbreds. *The Plant Cell* **17**, 343–60 (2005).
74. Parameswaran, P. *et al.* A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing. *Nucleic acids research* **35**, e130 (2007).
75. Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J. & Lander, E. S. High-resolution haplotype structure in the human genome. *Nature Genetics* **29**, 229–32 (2001).
76. Elshire, R. J. *et al.* A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLoS ONE* **6**, e19379 (2011).
77. Schnable, J. C., Springer, N. M. & Freeling, M. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 4069–74 (2011).

Appendix 1 - Proposed Polymorphism Descriptors (PDs)

- **Allele Frequency:** Measured via sequencing and GBS. Tripsacum or Sorghum will be used to estimate derived frequency where possible; otherwise minor allele frequency (MAF) will be used. Invariant sites will be scored as 0 or 1.
- **Sequence Quality:**
  - **Sequence quality:** This score will be a composite measure of the mapping quality of the read and the sequence quality score.
  - **SNP quality:** This score will reflect sequence depth, and confidence of genotype calls, including invariant sites.
- **Genomic Context:** A number of parameters relating to sequence context
  - **Methylation status:** (*e.g.*, Springer project: www.cbs.umn.edu/lab/springer/mev/data)
  - **GC content**
  - **Repeat density, class, and age**
  - **Gene density**
  - **Transcription factor binding sites**
  - **Conserved non-coding sites**
  - **Subgenome: Ancestral state,** which of the two subgenomes does the SNP belong to[77]
  - **Chromatin openness:** will rely predominantly on published data from several groups (*e.g.*, www.genomaize.org)
- **Recombination Rate:** These estimates of the impact of recombination will, for the same genomic region, differ in different panels or populations.
  - **Rho:** A measure of historical recombination estimated from LD
  - **Crossover rate:** Estimated from genetic map.
  - **LD:** Mean levels of LD in surrounding genomic regions
  - **Haplotype length:** Average length of shared haplotypes at a site.
- **Population Genetic Parameters:** Like recombination rates, these may vary by sample or population.
  - **Extended haplotype homozygosity:** A measure of haplotype blocks informative of the action of selection
  - **Diversity:** Estimated for each site and surrounding regions
  - **SFS:** The site frequency spectrum, estimated in a surrounding window
  - **Divergence/Differentiation:** Various estimates of differentiation and divergence, including net nucleotide divergence at a SNP, dn/ds in coding regions, and F_{ST} among populations.
  - **Geographic range of variant**
  - **Inbreeding coefficient of surrounding region**
  - **Haplotype age:** Age of surrounding haplotype is informative about whether a rare variant is simply young (and thus potentially neutral or weakly deleterious) or old (and thus very likely deleterious, as it has unable to drift to higher frequency)
  - **Constraint:** How conserved is the site or surrounding region across *Zea*, or all of Andropogonae?
- **Gene characters**
  - **Position:** Position in a gene, *e.g.*, intron, exon, 5' UTR
  - **mRNA:** splice-site, mRNA stability
  - **AA impact:** Synonymous, nonsynonymous, codon usage
  - **Physico-chemical impact of amino acid change:** Size, charge, etc. Will likely include estimates of deleterious nature of change from, *e.g.*, SIFT software
  - **miRNA: non-coding genes, position in miRNA**
  - **phosphorylation:** Data on phosphorylation from www.p3db.org project.

- **Transcription/Translation:** Estimated at the level of genes or regions, but applied to individual sites
    - **Expression level**
    - **Tissue specific expression pattern**
    - **Alternative splicing, alternative start site, antisense related Transcript**
    - **Position in biochemical pathway**
    - **Interaction in transcription co-expression network**
    - **Protein level:** Data from Briggs NSF proposal on protein level for maize genes
    - **Protein interaction:** Application of rice interaction network (ricenet) data to maize to predict number of interactions

Appendix 2 - Table of Germplasm and Status of Genotypic and Phenotypic Data

Key: WGS=Whole Genome Sequenced; GBS=Genotyping by Sequencing; F&H = year that our group will access to flowering and height data; YD = year that our group will have access to yield data; Comp. = Computational projection of genotypes (genotypes of hybrids are determined by GBS and WGS of inbreds).

Germplasm	**WGS**	**GBS**	**Distinct Genotypes**	**Locations**	**Total Plots**	**F&H**	**YD**
NAM Inbreds	HapMap V2	Done	5,000	20	100,000	12	12
NAM Hybrid U.S.	Comp.	Comp.	1,600	8	12,800	12	12
CN-NAM Inbreds	CAU Studies	Done	2,000	4	8,000	12	12
U.S. CN NAM Hybrids	Comp.	Comp.	21,000	9	126,000	13	14
CIMMYT SeeD Landraces	D1 & SeeD	Done	15,000	33	60,000	12	13
Ames Inbreds	A1	Done	3,200	4	12,800	12	12
Ames Hybrids	Comp.	Comp.	1,500	4	6,000	12	12
PVP Diallel Inbreds	In Progress	In progress	2,000	4	8,000	13	13
PVP Hybrids	Comp.	Comp.	2,000	12	24,000	14	14
CIMMYT Breeding	CIMMYT	Substantially Done	5,000	8	40,000	12	13
Teosinte (C2Exp1)	D1	D2	5,000	2	5,000	C2	C2
Landrace (C2Exp1)	D1	D2	5,000	2	5,000	C2	C2
Teosinte (C2Exp2)	D1	D2	5,000	2	5,000	C2	C2
Landrace (C2Exp2)	D1	D2	5,000	2	5,000	C2	C2
Zea Synthetic (C3Exp1)	D1	D2	2,000	6	12,000	C3	C3
Zea Synthetic DH (C3Exp2)	D1	D2	2,000	6	12,000	C3	C3
Zea Synthetic Hybrids (C3Exp2)	Comp.	Comp.	2,000	6	12,000	C3	C3
PD Hybrids (B2)	D1 & Comp.	Comp.	5,000	3	15,000	B2	B2
Total			**89,300**		**468,600**		

Appendix 3 - Construction of a Pan *Zea* -Genome

1) The WGS assembly strategy will be prototyped using the inbred Mo17 and a teosinte inbred line. These two lines will be sequenced and assembled following the AllPaths-LG approach. A paired-end library and 2-3 kb jump library will be constructed for each line and sequenced to 50x for each library. The AllPaths-LG assembler will be used for *de novo* assembly. As the technology is rapidly evolving in this area, we will explore other options including the Pac-Bio platform.
2) The Illumina paired-end sequencing data will be assembled into contigs with SOAP-DENOVO. Contigs will be created from three sets of data: High depth inbred lines each individually (>15X depth), high depth heterozygous samples each individually (>30X depth), and combined low depth sequence (~3000X cumulative depth across 800 taxa). We expect the extent of paralogy errors in the contigs to vary from moderate to high, respectively, in these datasets. Contigs generated from the combined low depth will require more rigorous validation thresholds below.
3) Several technical platforms will be explored for de novo transcriptome assembly, including Pac-Bio, and 100x2 Hiseq. *De novo* transcriptome assembly will be used to improve assembly and annotation of the gene space.
4) Contigs assembled from the different approaches will be evaluated and filtered as follows. Lines sequenced to depth (15X for inbreds, 30X for heterozygous material) will be aligned to the contigs to call SNPs. A contig will be considered valid if common SNPs (>5% frequency) all show LD with at least one other SNP on the contig. This evaluation will be used to optimize the *de novo* assembly process.
5) Contigs will be anchored using GBS markers and Reference SNPs calls. Validated contigs will be anchored to one or more genetic maps using their predicted GBS tags, which provides 30,000 taxa to map against. We will make use of the well-studied NAM, CN-NAM, CIMMYT, and IBM mapping populations as well as the half-sib mapping populations developed during this project and a teosinte x B73 F2 population currently under development. For those contigs without GBS tags, we will LD map these against the reference scored SNPs.
6) Contigs (or portions of contigs) that cannot be aligned to the reference genome will be binned based on genetic map positions, and then combined with the B73 reference to construct a pan-genome of maize. Currently, GBS mapping resolves half of the GBS tags to <10Kb resolution.
7) Gene annotation of these extra contigs will be based on RNA-seq data generated from this project and other NSF PGRP projects. The Gramene project will work on tools for visualization of the pan-genome.

Appendix 4 - SNP Calling and Identification of Rare Alleles

1) Whole genome shotgun sequencing will be performed on a total of 496 individuals as described in section D1.
2) Data from additional lines from related projects (Table A4.1) will be included as they become available.
3) Short reads from each line will be aligned with Bowtie2 to the B73 reference as well as to a pan-genome constructed from this project in order to call SNPs.
4) The SNPs will be filtered based on LD and segregation patterns in the sequenced teosinte and maize lines following pipelines developed for maize HapMapV2.
5) False-positive and error-prone SNPs will be identified through utilization of Identity by Descent (IBD) genomic segments. Known relatedness and large number of samples will help to uncover a large number of novel IBD regions.
6) A Gold Standard set of SNPs (~10K SNPs consistently scored in both Illumina 55K and GBS plus 90K IBD validated in GBS and WGS) will be identified and carefully annotated and validated. These will be used to test any new pipeline developed to ensure consistency and accuracy throughout project analyses.

Table A4.1 Additional sequence data leveraged from related projects

Germplasm	**Taxa**	**Depth**	**Source**	**Release**
Maize PVP	12	15X	Maize Diversity; J. Ross-Ibarra	2013
Maize Inbreds	55	5X	HapMap v2	2012
Teosinte Inbreds	16	5-15x	HapMap v2; Maize Diversity	2012; 2013
Landrace Inbreds	20	5X	HapMap v2	2012
Landraces	40	30X	USDA funding; J. Ross-Ibarra	2013
CIMMYT Inbreds	84	10X	BGI-CIMMYT; Buckler advising	2013
CAU Skim	278	2X	Jiao et al. Nat. Gen.	2012
CAU Skim P2	400	2X	J. Lai; CAU	2013
Various NSF projects	30	10X	*e.g.*, Vitamin A, Disease Resistance	2013

Appendix 5 - Roles

Investigator	Roles
Buckler (PI)	• Overall project management • Pan-*Zea* GWAS analyses (section A1) • Data mining to screen polymorphism descriptors (PDs) (section A3) • Genomic prediction analyses incorporating PDs (section B1) • Choose hybrids & contribute to field trials for PD tests (section B2) • Integration of global *Zea* field trials (section C1) • Comparative evaluation of *Zea* alleles field trials (section C3) • Error modeling of variant calls from whole genome sequence (section D1) • Annotation of splicing & transcript stability PDs (section D3)
Bradbury (co-PI)	• Pan-*Zea* GWAS analyses (section A1) • Data mining to screen polymorphism descriptors (section A3) • Genomic prediction analyses incorporating PDs (section B1) • Genotypic imputation (section D1)
Doebley (co-PI)	• Pan-*Zea* population genetic analyses (section A2) • Genetic load in teosinte and landraces (section C2) • Annotation of transcriptional regulation PDs (section D3)
Flint-Garcia (co-PI)	• Design and organize hybrid trials to test PD hypotheses (section B2) • Comparative evaluation of *Zea* alleles (section C3) • Annotation of breeding history polymorphism descriptors (section D3)
Holland (co-PI)	• Pan-*Zea* GWAS analyses (section A1) • Genomic prediction analyses incorporating PDs (section B1) • Field trials to test PD hypotheses (section B2) • BLUP estimation (section C) • Genetic load in teosinte and landraces (section C2) • Comparative evaluation of *Zea* alleles (section C3)
Mitchell (co-PI)	• Generation of whole genome sequence (section D1) • Collection of GBS and RNAseq data (section D2)
Ross-Ibarra (co-PI)	• Pan-*Zea* population genetic analyses (section A2) • Data mining to screen polymorphism descriptors (section A3) • Genetic load in teosinte and landraces (section C2) • Generation & analysis of whole genome sequence (section D1) • Annotation of population genetic polymorphism descriptors (section D3)
Sun (co-PI)	• Computational infrastructure & data management • Training postdocs and students to use high performance computing clusters • Pan-genome construction from whole genome sequence (section D1) • Analysis and curation of genotypic data (section D2) • Annotation of the genome with polymorphism descriptors (section D3)
Ware (co-PI)	• Pan-genome construction from whole genome sequence (section D1) • Annotation of comparative genomic polymorphism descriptors (section D3)
Fulton (co-PI, Outreach Coordinator)	• Outreach • Web site development and maintenance (static content)
Zhang (USDA-ARS funded Senior Personnel)	• Pan-*Zea* GWAS analyses (section A1) • Genomic prediction analyses incorporating PDs (section B1)

Sanchez (Collaborator)	• Identify teosinte and landrace germplasm source populations (section C2) • Phenotyping of teosinte and maize landraces (section C2)
Casstevens (Lead Program Analyst)	• Lead programmer for TASSEL and implementation of public algorithms • Computational infrastructure
Kroon (Database Curator; USDA-ARS funded Personnel)	• Curation of phenotypic data (section C) • DNA and RNA sample tracking (section D)
Glaubitz (Project Manager)	• Project management (25%) & Bioinformatics (75%) • Analysis and curation of genotypic data (sections C and D2) • Curation of phenotypic data (section C) • Error modeling of variant calls from whole genome sequence (section D1) • Annotation of the genome with polymorphism descriptors (section D3)
Miller (Administrator)	• Project management • Outreach • Curation of phenotypic data (section C)

Section	Year 1	Year 2	Year 3	Year 4	Year 5
A1: Pan-*Zea* GWAS					
Development of computational infrastructure					
Development of GWAS approaches					
GWAS on flowering and height	**Maize**		**+Teosinte & Landraces**		
GWAS on yield components		**Maize**	**+Teosinte & Landraces**		
Integrated analysis of the genetic architecture of *Zea*					
A2: Pan-*Zea* Population Genetics					
Analyses of demographics, constraint, positive selection, and recombination					
A3: Data Mining to Predict Functional Alleles					
Development of computational infrastructure					
Testing initial PDs using NAM estimates					
Testing of PDs using whole genome allele frequency estimates					
Trainee driven PD testing					
Integration across all PDs					

Section	Year 1	Year 2	Year 3	Year 4	Year 5
B1: Enhanced Genomic Prediction					
Model and computational development	**gBLUP**	**Machine Learning**	**Bayes**		
Addition of burden predictions (A2)					
Cross validation testing		**Maize**		**+Teosinte & Landraces**	
Design EGP trials (B2) and retrain from their results					
B2: Field Testing of PD Hypotheses					
Design hybrids, make seed					
Field trial PDs and enhanced genomic prediction (EGP)	**Pilot**	**2 PDs**	**2 PDs**	**1 PD, 1 EGP**	**2 EGP**

Section	Year 1	Year 2	Year 3	Year 4	Year 5
C1: Integration of Global *Zea* Fitness Trials					
Curation of community data	**NAM, Ames**	**CN-NAM**	**exPVP, SeeD**	**C2, C3**	**C2, C3**
Implementation of computational infrastructure					
Analysis pipelines	**Inbred Imputation**	**Error Detection**	**LR Imputation**	**Error Detection**	
Imputed dataset release					
BLUP, h2, and GxE					
C2: Evaluation of Deleterious Rare Alleles in Teosinte and Landraces					
Seed production	**Experiment I**	**Experiment II**			
Field trials		**Experiment I**	**Experiments I & II**	**Experiment II**	
RNAseq		**Experiment I (24)**	**Experiment II (60)**		
Analysis (RNAseq and REML)					
C3: Comparative Evaluation of *Zea*					
Generate S1, FS & DH germplasm					
Generate testcross hybrids of DH lines					
Field trials (MO, NC, NY)		**DH**	**S1,FS & DH**	**S1,FS & DH hybrids**	**DH hybrids**
Phenotypic data analysis and GWAS					

Section	Year 1	Year 2	Year 3	Year 4	Year 5
D1: Whole Genome Sequencing for Rare Allele Discovery					
Teosinte sequencing (~30x)	50	61			
Landrace sequencing (~30x)	50		50		
Inbred sequencing (276 @ ~15x; 9 @ ~60x)	4	78	4	89	110
Resequencing pipeline	Development	Implementation			
Pan-*Zea* genome pipeline	Development	Implementation			
Public release (raw data, variant calls, pan-genome)					
D2: Genotyping to Track Rare Alleles					
384-plex Genotyping-By-Sequencing (GBS) runs (C1)	1,000	1,000	1,000	1,000	1,000
96-plex GBS runs (C2 and C3)	5,000	7,000	7,000	5,000	
Cumulative genotype builds with annual release					
D3: Annotate Maize Genome with Polymorphism Descriptors					
Data structures developed					
PDs developed, evaluated, published					

Section	Year 1	Year 2	Year 3	Year 4	Year 5
Education and Outreach					
New copy of the museum exhibit constructed					
Exhibit travels every 3-5 months					
Panzea website format/structure revised					
2-4 GBS workshops					
3-6 short video courses produced					