

Project Summary:

Molecular and Functional Diversity of the Maize Genome

Edward Buckler (Cornell Univ., coPI)
John Doebley (Univ. of Wisconsin, PI)
Theresa Fulton (Cornell Univ., Senior Personnel)
Brandon Gaut (Univ. of California-Irvine, coPI)
Major Goodman (North Carolina St. Univ., coPI)
James Holland (North Carolina St. Univ., coPI)
Stephen Kresovich (Cornell Univ., coPI)
Mike McMullen (Univ. of Missouri, coPI)
Lincoln Stein (Cold Spring Harbor Lab, coPI)
Doreen Ware (Cold Spring Harbor Lab, coPI)

Scientific Objectives and Approaches

In our previous grant, we determined the genetic relationships among maize and teosinte germplasm, examined how average loci evolve, and developed methods for relating nucleotide diversity to phenotypic effects. Our objective for this new proposal is to understand how selection has shaped molecular diversity in maize and then relate molecular diversity to functional phenotypic variation. *How has selection shaped molecular diversity?* To address this, 4000 loci will be screened for selection evidence, and then 1000 loci will be studied extensively by doing SNP surveys across diverse maize and teosinte. A range of tests of selection will be used to identify genes showing positive, diversifying and purifying selection. The identified genes will be those involved in domestication, agronomic improvement, and local adaptation. *How does this molecular diversity relate to functional trait variation?* A wide range of maize and maize-teosinte linkage and association mapping populations will be created that capture a tremendous range of diversity. These populations will be genotyped for SNPs and candidate genes (many of the selected genes noted above) and phenotyped for domestication, agronomic and developmental traits. This will permit high-power and high-resolution dissection of a wide range of traits, and relate the molecular diversity to functional variation.

Broader Impacts

In 2001, maize became the number one production crop in the world, and current US maize production is almost four times wheat and rice production combined. Maize is also the most diverse known crop at the molecular and phenotypic levels. This project's diversity dissection will enable genomics and breeding to help improve maize yield and quality and to adapt maize to a wide range of biotic and abiotic conditions. This project will provide insights into the ways selection modifies a genome. Additionally, this will produce two major resources: (1) a large database of validated SNPs throughout the genome, and (2) a set of maize and teosinte mapping populations that could be used to map almost any trait. These results will be made accessible to two audiences. A strong informatics program will be used to make these results available to a wide range of researchers through public databases such as Gramene and Genbank. Additionally, we will make this research accessible to a wider range of people through outreach programs. We will focus on transmitting the excitement of science and knowledge of maize to students in rural schools and through mentoring programs in our labs.

Results under Prior NSF Support:

Plant Genome Program 0096033; Buckler, Doebley, Gaut, Goodman, Kresovich, Muse and Weir
Title: Evolutionary Genomics of Maize. Duration: 1/1/99-12/31/03. Amount: \$5,847,105

This proposal had three sections: (1) SSR diversity in maize and its genome, (2) the processes that shaped diversity in the maize genome, and (3) natural functional variation in maize genes. The project was remarkably successful and all research goals were achieved. The 36 project publications or manuscript are marked (*) in the proposal literature cited. Over all labs, we mentored 14 postdocs, 17 graduate students and 18 undergraduates.

Microsatellite (SSR) Diversity in Maize: *How is diversity distributed in the maize germplasm pool?* We addressed this question by screening 100 SSRs on a comprehensive sample of maize germplasm. All the data are in hand (genotypes for 100 SSRs on 1026 maize landraces accessions, 261 maize inbred lines, and over 217 teosinte accessions). These data are being analyzed (1) to characterize the amount and distribution of genetic diversity in maize germplasm, (2) to determine the degree of genetic similarity among different segments of the germplasm pool, and (3) to identify a core set of lines for use in mining useful diversity for agronomic traits. One paper was published this past year (Matsuoka et al. 2002b) and three others are in draft and targeted for submission in 2003. When these three papers are done, all goals on this question will have been met or exceeded. The published papers will show how genetic diversity is organized in maize inbreds, maize landraces and teosinte so that future workers will be able to intelligently sample these materials to fit the needs of their individual projects.

How is SSR diversity distributed across the maize genome? We screened 45 maize inbreds, 45 maize land races (exotics) and 45 teosintes for ~500 SSRs to make diversity maps of the maize and teosinte genomes with an SSR every 3-4 cM. We have completed this work and a manuscript on these results should be submitted in May 2003. The results enabled us to detect a mild drop in diversity near domestication quantitative trait loci (QTL) and detect some evidence for genic regions (and some individual SSRs) that experienced selection; however, overall the data argue that the maize genome is homogeneous for diversity at the 3-4 cM level.

Other results: We also did two experiments not outlined in the original proposal. First, we determined the empirical mutation rate and studied the pattern of mutation for maize SSRs (Matsuoka et al. 2002a; Vigouroux et al. 2002a). Second, we collaborated with Mike McMullen to identify SSRs in ESTs that lack variation in maize. Such ESTs were likely under selection during maize domestication and thus likely agronomically important genes. We identified 14 of 500 ESTs that were putatively targets of selection during maize domestication (Vigouroux et al. 2002b). In this work, we developed a novel test for selection that incorporates the maize domestication bottleneck.

Processes That Shape Genetic Diversity in the Genome: This section sought to measure sequence diversity in 100 loci. The questions to be addressed were: 1) What is the level and pattern of SNP diversity in maize? 2) How does diversity in maize compare to its relatives?, and 3) Can this information be integrated into a genomic context?

Levels of diversity: A survey of maize SNP diversity, based on 21 chromosome 1 loci, revealed both that maize is very diverse – roughly 11-fold more diverse than humans and that linkage disequilibrium decays rapidly over physical distance (Tenaillon et al. 2001). The survey

also demonstrated that U.S. germplasm is demonstrably less diverse than exotic germplasm. Another paper (Tenailon et al. 2002) examined diversity in the context of recombination rates along chromosome 1. In contrast to expectations based on *Drosophila*, there was no correlation between SNP diversity and recombination rates. For example, centromeric loci were not genetically depauperate. There was, however, a correlation between recombination rates and SSR diversity, suggesting that recombination contributes to SSR mutation.

Diversity among Zea taxa and domestication genes: We are analyzing a sample of 35 loci representing four *Zea* taxa. Each locus has been sequenced from roughly 70 individuals. The data are being used to infer the population and demographic histories of the taxa (target submission - Fall 2003). We have published studies that characterize diversity patterns between maize and its wild ancestor, *Z. mays* ssp. *parviglumis* (Zhang et al. 2002; Tiffin and Gaut 2002). Maize contains roughly 70% of the SNP diversity found in its ancestor at “non-domestication” genes, but far less diversity in “domestication” genes that have been subject to selection. For example, three maize domestication genes (*tb1*, *d8* and *c1*) contain, on average, only 15% SNP diversity relative to ssp. *parviglumis*. This work demonstrates that the relative levels of SNP diversity between maize and its ancestor can be used as a marker to identify domestication genes of potential agronomic importance. The approach has already been used to document selected genes in the starch biosynthesis pathway (Whitt et al. 2002).

Additional data: In collaboration with Mike McMullen and the Missouri Maize Mapping project, ≈1200 loci have been sequenced from a sample of 7 temperate and 7 tropical maize inbreds. Data processing, release and analysis are underway (target submission – Summer 2003). Another 1000 loci will be sampled from maize inbreds and wild relatives by the end of the grant period (target submission – 2004). As a result of this collaboration, we will produce 20-fold more data than initially proposed.

Other results: To better understand the structure of the maize genome, we developed a statistical method to infer chromosomal homologies from genetic map data (Gaut 2001). This method revealed that one-third of the maize genome is triplicate or higher copy.

Evaluation of Natural Maize Functional Variation: This section developed and applied candidate gene association tests for maize. In the first four years of the five-year project, all empirical cloning, sequencing, and phenotyping goals of the project have been substantially surpassed. Most importantly this research has been successful in developing candidate gene association approaches that can be applied in plants. Association approaches have had great utility in fine mapping traits in humans and *Drosophila*, but population structure has been a serious obstacle to applying these approaches to fine mapping QTL in plants. First, this project described the structure of linkage disequilibrium for diverse inbred lines, which was found to decay within 1500 bp (Remington et al. 2001). Second, populations were developed, phenotyped and genotyped so that association approaches could be applied to maize. Third, we extended statistical approaches so that quantitative traits could be analyzed, while controlling for the population structure of the mapping populations (Thornsberry et al. 2001; Buckler and Thornsberry 2002). Fourth, we developed and released 2 software packages (TASSEL and MAIZEMEISTER) to help with the analysis and phenotyping. Finally, these association approaches have helped us identify nucleotides that control flowering time, produce sweet corn, change row number, and modify starch levels (Thornsberry et al. 2001, Whitt et al. 2002, 3 more submissions in 2003). Overall, these association approaches provide a tool to map functional

variation with 5000 fold better resolution than F₂ QTL mapping. This research has provided a rapid avenue for applying functional genomic diversity to plant improvement.

Project database and other results: We developed a diversity database for maize and its relatives (<http://statgen.ncsu.edu/panzea/>). The interface has a genetic map-based tool that allows users to view genetic diversity for different loci by genetic position. A download tool allows users to select markers based on their type, chromosome position or other features, and then to select accessions based on their geographic location, source, altitude, etc. Once this is done, the user can download diversity data for the selected loci and designated germplasm. All published germplasm, SSR, sequence, and phenotype data are available in the project database. The project also developed powerful software packages for statistical analysis of genetic diversity data (POWERMARKER) and for choosing CORESETS of germplasm (statgen.ncsu.edu).

Plant Genome Program 9872655; Ed Coe PI, **McMullen coPI**. Duration: 10/1/98-9/30/03. Title: Comprehensive genetic, physical and database resources for maize. Amount: \$11,075,485

The main goal of this project was to develop an integrated genetic-physical map for maize. Dr. McMullen was responsible developing a high-resolution genetic map to serve as the anchor framework for the physical map. With G. Davis, this objective was accomplished with the IBM genetic map containing over 2200 markers released through MaizeDB. Dr. McMullen was also responsible for marker development. The initial objective was for 2000 SSR markers over the 5 years. The project developed 1340 SSRs in the first 3 years of the project and mapped most of the other public SSRs for a total resource of 2070 SSRs. These SSRs have been used to anchor BAC contigs by employing a BAC pooling strategy. Moreover, these SSRs are in extensive use throughout the maize research community. Accomplishing the SSR development in three years allowed Dr. McMullen to initiate a SNP discovery project that is beyond the goals of the original grant. We are using the SNP markers to map unigenes that detect unplaced contigs to enhance the integrated genetic/physical map. Our additional goal is to conduct SNP discovery on 3000 unigenes (2200 completed) and map 1000 (225 completed) SNP loci by the end of the project.

Introduction:

Maize is the most diverse crop species in the world. This diversity is manifested at both the molecular and phenotypic levels. Considering nucleotide polymorphism in genes, two maize lines are on average as diverged from one another as humans are from chimpanzees. The phenotypic diversity we see in maize today is the product of a long tradition of plant breeding practiced by Native Americans who first converted the wild grass, teosinte, into maize and then proceeded to adapt this crop to virtually every habitable environment in the Americas including deserts, tropical rainforest, high mountains (3500 m above sea level), and the short growing season of Canada. Most maize diversity remains undescribed, poorly understood and under utilized in modern plant improvement largely because of the difficulty of identifying useful genetic variants hidden in the background of low yielding local varieties or lines (Tanksley and McCouch 1997). Thus, there is a compelling opportunity to apply genomic tools to sift through the countless allelic variants in the maize gene pool and understand how they impact phenotypes of agronomic and evolutionary importance.

In our past project, we determined the structure of genetic diversity in the maize and teosinte gene pools, examined how genetic loci evolve, and developed methods for relating nucleotide diversity to phenotypic effects. In our new project, we want to extend this work to understand two things: (1) *Which genes have been under selection during maize domestication and improvement?* By addressing this question, we can identify specific genes that have experienced past selection and thereby improve our understanding of how evolution modulates diversity across the genome. Our group's prior research suggests that although only 1% of random maize genes are under selection (Vigouroux et al. 2002b), two-thirds of the genes involved in the agronomically important starch pathway have been modified by selection (Whitt et al. 2002). The identification of numerous genes under selection will provide insights into how evolution modifies pathways and which genes control agronomic phenotypes. (2) *Which genes and alleles control key traits?* Although finding agronomically relevant genes is important, the improvement of maize requires finding better alleles at these genes. We will develop a high-throughput, high-resolution platform for finding and evaluating allelic variation throughout the genome and the breadth of the gene pool. Combining these two questions, our overall goal is to turn the identification and evaluation of functional and evolutionarily important allelic variation into a comprehensive (genomics) activity.

Specific Aims:

The proposal has two sections. **Section 1** will examine the impact of past selection on molecular diversity in maize genes. **Section 2** will investigate a large number of candidate genes for their associated effects on agronomic or evolutionarily important phenotypes. These two sections are interrelated. If a gene is identified in Section 1 as having experienced selection, this implies that it controls an important phenotype and thus would be an appropriate candidate for functional analysis in Section 2. Similarly, genes shown to possess functional allelic diversity in Section 2 need to be assayed for molecular diversity to reveal how selection has patterned diversity across the gene pool and to identify where interesting allelic diversity occurs. Our specific aims are:

Section 1. Molecular Diversity:

- A. Perform SNP discovery for 1000 genes in diverse maize inbreds and teosinte to identify an unbiased set of SNP for molecular diversity analysis.
- B. Perform tests of selection during maize domestication and improvement on 4000 random and phenotype-candidate genes.

Section 2. Functional Diversity:

- A. Develop maize and maize-teosinte mapping populations to capture diverse alleles and dissect a wide range of quantitative traits.
- B. Conduct large-scale QTL mapping experiments in maize and maize-teosinte populations to identify, localize and characterize the genomic regions harboring QTL.
- C. Dissect domestication, agronomic, and plant development QTL using candidate gene association analyses.

In addition, we will develop a database that has user interfaces for (1) presenting our maize QTL maps in the context of the existing genetic maps for maize and other cereals and (2) viewing SNP and phenotypic diversity. These features will be integrated into the Gramene database in a way that makes it easy for researchers to perform analyses across diverse germplasm and to gain

access to the underlying biological reagents. Finally, we will work with students in rural and under resourced communities to enhance science education and to help bring underrepresented groups into plant science.

Section 1: Molecular Diversity

Background: Natural (or artificial) selection generates adaptations through its action on genetic diversity in populations. The action of selection can leave a distinctive signature at the nucleotide level. During the 1980's and 90's when nucleotide polymorphism data first became readily obtained, evolutionary biologists applied tests of selection to single genes in countless species (Otto 2000; Brookfield 2001). Over the past few years with the implementation of high-throughput methods, evolutionary biologists have begun to perform tests of selection on a genomic scale. In humans and *Drosophila*, tests of selection with SSRs have given promising results (Kauer et al. 2002; Payseur et al. 2002; Schlotterer 2002). Our group has also taken this approach in maize with similarly promising results (Vigouroux et al. 2002b). Despite these successes, SSRs have several limitations: (1) only a fraction of genes have SSRs, (2) the high mutation rate for SSRs means that they can recover rapidly from selective sweeps, and (3) an SSR in a gene may be too distant from the selected site to be a reliable reporter of past selection. Section 1 of this proposal will extend our prior work and examine the impact of selection on molecular diversity in maize on a genomic scale. First, we will conduct SNP discovery with diverse germplasm, and then perform tests of selection using multiple strategies.

Specific Aim 1A: SNP discovery in diverse germplasm. Our first goal is to perform SNP discovery with diverse germplasm as a foundation for our tests of selection and for our QTL mapping and association studies (Section 2). Our SNP discovery work will also provide a resource for the maize genetics community in general.

SNP discovery is often conducted in such a way that the utility of these powerful markers is compromised. Specifically, when discovery is performed in a narrow set of genotypes, SNPs that do not vary in that narrow set will not be discovered. This creates an *ascertainment bias* that obviates the possibility of using the SNP for projects with diverse germplasm. A biased set of SNPs derived from US maize inbreds can not be used to make inferences about selection and diversity in maize germplasm overall. A better discovery strategy would be to use teosinte and diverse tropical-temperate inbreds since maize contains a subset of the SNP diversity in teosinte and US inbreds a subset of that in more diverse maize (Gaut et al. 2000; Tenaillon et al. 2001). We have begun to create a set of SNPs for maize diversity analysis that is free of this form of ascertainment bias under our current projects (DBI0096033 and DBI9872655). We will extend this work into the proposed project with the goal of identifying 8000 new SNPs in 1000 genes.

SNP Discovery to date: By the start of the project, we will have completed SNP discovery for 1000 genes in US inbreds, 1200 genes in diverse (temperate/tropical) maize inbreds, and 1000 genes in diverse maize inbreds and teosinte partial inbreds (~3000 genes total). The reference sequences for SNP discovery are the “CornSensus” unigene set developed by the Maize Mapping Project and DuPont. Because this unigene set was synthesized from sequences of both public and private EST databases, this is the most detailed public maize unigene set available. The genomic position will be known for most candidate genes identified from the CornSensus sequences (iMAP: <http://www.maizemap.org>).

Our SNP discovery to date has been performed in three stages. In the first stage, 10-12 US inbreds were assayed for 1000 unigenes. In line with previous reports for maize (Rafalski 2001; Tenaillon et al. 2001; Ching et al. 2002), diversity was high with one SNP every 73 bp and one indel every 300 bp. In the second stage of SNP discovery, 7 temperate and 7 tropical inbreds were assayed for 1200 unigenes. These 14 lines were chosen to maximize the number of distinct SSR alleles present, assuring that we are sampling the greatest level of sequence variation possible. These data will be submitted to Genbank by the end of 2003. The third stage of SNP discovery involves sequencing an additional 1000 unigenes across the 14 tropical/temperate inbreds plus the 16 teosinte (*ssp. parviglumis*) partial inbreds. To date, we have applied this strategy to 23 unigenes. Despite the fact that the teosintes have been inbred only two generations, on average 13 of 16 individuals yielded quality direct sequence (avg. Phrap score > 50). Based on this pilot study, we are proceeding and will complete the data by the end of the current grant period. This last set of SNPs lack the ascertainment bias that affects the first set severely and the second set to some degree. This last set of SNPs provides the foundation for our large-scale study of SNP diversity and selection in maize and its relatives.

Additional SNP discovery: Since our project will be using SNP for the analysis of molecular diversity (Section 1) and QTL/association analyses (Section 2) in diverse maize germplasm, we need to expand the set of available SNPs. Specifically, we need to perform SNP discovery with a set of candidate genes for agronomic and evolutionary phenotypes. The candidate gene SNP discovery will be done in two stages. *First*, the McMullen lab will sequence a 600 bp region from 1000 candidate genes in the 14 tropical/temperate inbreds plus the 16 teosinte partial inbreds. These will be done at a rate of 250-300 per year for the first 3 to 4 years of the project. Candidate genes will be chosen based on homology to genes of known function from other plants (rice or Arabidopsis) and from genes identified by the “Inflorescence Architecture in Maize” and the “Endosperm Development in Maize” projects (see letters from Sarah Hake and Don McCarty). *Second*, the Buckler lab will sequence the same 600 bp region for 750 of the 1000 candidates from 11 additional diverse maize lines (25 total) and a second 600 bp region will be sequenced from the entire set of 25 diverse inbreds. The complete set of SNPs for the two 600 bp regions in these 25 diverse maize lines will be used to evaluate specific hypotheses as part of Section 2.

Specific Aim 1B: Tests of selection. Over the past several years, our group has used *forward screens* for selection, *i.e.* first identifying genes that were likely targets of selection and then asking whether we could detect the signature of selection (Wang et al. 1999; Whitt et al. 2002) as compared to neutral loci (Tenaillon et al. 2001). We now propose to perform genomic-scale *reverse screens* for selection in a total of 4000 genes, most of which show no prior evidence that they were under selection. We will apply several types of tests that look for either reduced diversity in maize relative to teosinte, skewed frequency spectra for SNPs, elevated divergence between maize and teosinte, and elevated levels of Linkage Disequilibrium (LD) in maize.

1. *Sampling:* Our sampling strategy for genes and germplasm is tailored to take advantage of our prior and current SNP discovery work and to allow us to test the maximum number of genes for the minimum cost. Five different sampling strategies will be employed (Table 1). The first four involve direct sequencing of genes that allows powerful tests of selection, however the high cost of sequencing requires small samples of germplasm (e. g. 16 maize and 16 teosinte). This limits statistical power and the number of specific tests that can be performed. Therefore, a fifth

strategy will use SNP genotyping to examine a comprehensive sample of maize germplasm (1000 exotic maize and 200 teosintes) with 500 random and 500 phenotype-candidate genes. This last approach will enable us to identify genes with the signature of selection by setting up specific contrasts between different portions of the maize germplasm pool.

Strategy 1. SNP discovery already completed for 1000 random genes in US inbreds revealed about 80 genes that have little or no SNP/indel variation. We will follow up on these low diversity genes by sequencing them in 16 exotic (tropical) maize and 16 teosinte lines. If these genes have low diversity in US inbreds because of some intrinsic factor (GC content, genomic location), then diversity will also be low in exotics and teosinte. However, selection during maize domestication would be revealed by high diversity in teosinte relative to US and exotic maize, and selection during US inbred development would be revealed by high diversity in teosinte and exotics relative to US inbreds. We have previously employed this strategy using SSRs and we were able to identify 14 selection-candidate genes (Vigouroux et al. 2002b).

Strategy 2. Our ongoing SNP discovery for 1200 random genes with the 14 temperate-tropical inbreds is expected to identify about 80 more low diversity genes. We will follow up on these by sequencing them in teosinte. *Strategy 3.* We are currently sequencing the 14 temperate-tropical inbreds and 16 teosintes in a set of 1000 random genes. These data will not only allow us to perform tests of selection, but as a large set of random genes, they will provide a reliable empirical distribution of diversity in maize and teosinte genes as a baseline for identifying genes that have experienced selection. *Strategy 4.* For the investigation of functional diversity (Section 2), we propose to sequence the 14 temperate-tropical inbreds and 16 teosintes for a set of 1000 phenotype-candidate genes. These genes will be of special interest for selection tests since they are more likely to control phenotypes of agronomic and evolutionary importance and thus should also show greater evidence for past selection (e.g. Whitt et al. 2002). *Strategy 5.* To provide greater statistical power and allow comparisons between different portions of the maize germplasm pool, we will assay 500 random and 500 phenotype-candidate genes by SNP genotyping in a comprehensive sample of maize germplasm. Each gene will be assayed for 3 or more SNPs in several hundred plants to test specific contrasts. For example, a sample of 200 teosinte and 200 random maize lines would provide a test of selection during domestication, while a test of 200 high elevation versus low elevation lines would allow us to identify genes involved in adaptation to altitude. We will perform a variety of such contrasts involving maize lines with different ecogeographic and climatic adaptations.

Plant Sample	Genes	Assay
10 US inbreds	1000 Random	sequencing
7 temp/7 trop inbreds	1200 Random	sequencing
7 temp & 7 trop maize inbreds/16 teosinte	1000 Random	sequencing
7 temp & 7 trop maize inbreds/16 teosinte	1000 Candidate	sequencing
1000 exotic maize & 200 teosinte	500 Random & 500 Candidate	SNP genotyping

2. Tests of selection: Selection can affect nucleotide diversity in multiple ways. Accordingly, we will apply multiple tests based on different expectations to glean as much as possible from the data. *First*, we will examine the ratio of genetic diversity between two samples (teosinte vs. maize or highland vs. lowland maize), which is reduced by directional selection. The ratio, R , will be measured either by Tajima's (1983) π or Watterson's (1975) θ . To date, it is known that most maize genes contain ~70% of the SNP diversity in teosinte (Zhang et al. 2002), but selected genes contain an average of only 15%. Our preliminary sequencing of 23 random maize genes

has identified two with $< 25\%$ of the SNP found in teosinte, although we caution that we have yet to rigorously establish the statistical threshold for selection. *Second*, we will test for skewed frequency distributions using Tajima's (1989) D for sequence data and the Ewens-Watterson test (Ewens 1972; Watterson 1978) for SNP-polymorphism data. In most genes, D is higher in maize than teosinte, reflecting accelerated drift during domestication and a commensurate loss of low frequency variants (Przerworski et al. 2000; Wall et al. 2002; Zhang et al. 2002). In contrast, selected genes often exhibit a decrease in D in maize relative to teosinte (Wang et al. 1999; Zhang et al. 2002). *Third*, the level of LD and the frequency of core haplotypes can be used as a measure of past selection (Sabeti et al. 2002; Wall et al. 2002). LD increases only gradually after a selection event (Przerworski 2001), and selection in maize, like humans (Sabeti et al. 2002), was recent enough to affect extant LD. Although we are only sampling short (~ 600 bp) regions, LD declines so rapidly in most maize genes that even short sequences may be sufficient to yield insights about elevated LD (Remington et al. 2001; Tenaillon et al. 2001). This test will be especially powerful for those candidate loci where two gene regions are sampled. *Fourth*, selection can lead to elevated divergence between two populations as measured by F_{st} . For the SNP-polymorphism data, we will test whether divergence for a gene between two samples is greater than expect under neutral models (Beaumont and Nichols 1996). We will identify selected genes based on R , D , LD and F_{st} in combination with coalescent models of domestication (see below) and other standard population genetic tests of neutrality (Tajima 1989; Hudson et al. 1987; Fay and Wu 2000).

When performing the selection tests with the SNP-polymorphism data, we will experiment with treating these data two different ways. First, each site can be analyzed individually for some tests (e.g. LD, F_{st} or Ewens-Watterson). Nevertheless, greater power may be obtained by first reconstructing the haplotype structure of the population. For this purpose, we will use a maximum likelihood method for the estimation of molecular haplotype frequencies in a diploid population (Excoffier and Slatkin 1995). The inferred haplotypes based on 3 or more SNP per gene can then be used in the tests of selection.

3. Types of selection: While the primary interest of our group is to identify genes that were under directional selection during maize domestication and improvement, the sequence data collected during that effort will also allow us to identify rapidly evolving genes that are candidates for either diversifying or balancing selection. These are an interesting class of genes, potentially including genes involved in adaptation to local environmental (abiotic) and pathogen-induced (biotic) stress (Mondragon-Palomino et al. 2002). To isolate rapidly evolving genes from maize, we will begin with two criteria. *First*, DNA sequence data gathered during our SNP discovery effort will be searched for genes at the high end of the distribution for diversity and divergence. Specifically, we will earmark genes with exceptionally high values of Tajima's π , Watterson's θ , or a high positive Tajima's D . *Second*, the "Cornensus" unigene EST set developed by the Maize Mapping Project and DuPont will be used as queries in BLASTN and BLASTX searches against Genbank's non-redundant nucleotide and protein databases. Genes having BLAST scores of $p > 10^{-10}$ will be considered possible rapidly evolving genes. Candidates identified by these criteria will be further investigated by Southern hybridization with garden blots of maize relatives and other grasses (Schmid and Tautz 1997).

4. Modeling maize domestication and diversification: The necessary context for performing tests of selection is a knowledge of the demographic and phylogenetic history of the populations in

question. Work under our past grant provides the foundation for the proposed tests of selection (Matsuoka et al. 2002b; Tenaillon et al. 2001, 2002; Vigouroux et al. 2002a, b; Zhang et al. 2002; other manuscripts in prep). However, the sequence and SNP-polymorphism data we proposed to collect will also enable us to better understand the demographic processes involved in maize domestication and diversification. This, in turn, will improve our test of selection by allowing us to better identify genes with diversity profiles beyond the range that can be explained by purely neutral models (e.g., Vigouroux et al. 2002b). Here, we describe how we will further investigate the domestication process.

Maize genes that were not selected during domestication were still subjected to a domestication bottleneck (Eyre-Walker et al. 1998; Hilton and Gaut 1998; Zhang et al. 2002). Demographic events like population bottlenecks have recently become a prominent focus in population genetics (e.g., Reich et al. 2002; Wall et al. 2002), but maize is the only case in the literature where a “known” demographic event - the domestication bottleneck - has been studied. Maize is unique because there are or will be estimates of: *i*) the time of the domestication event (9,000 years ago; Matsuoka et al. 2002b), *ii*) pre-bottleneck levels of diversity for 2000 genes (based on sequence data in teosintes), and *iii*) post-bottleneck levels of diversity for 2000 genes (based on sequence data in maize). Given these estimates, we will explore coalescent models of bottleneck events (Hudson 1990; Eyre-Walker et al. 1998). Models will incorporate recombination and will explicitly consider the variance in R among loci as one of several goodness of fit statistics. One purpose of these models will be to explore demographic effects over a wide-range of parameter values in the absence of selection. Such analyses will aid the identification of “outlier loci” with diversity distributions that cannot be easily explained without invoking selection or additional demographic events (e.g., Vigouroux et al. 2002b). We will also use coalescent models to infer the duration of the domestication bottleneck relative to population size (Hilton and Gaut 1998) and to compare expected vs. observed bottleneck effects on LD and frequency spectra (Wall et al. 2002). Finally, we will explore likelihood and Bayesian methods for using SNPs to infer demographic processes (Nielsen 2000; Wakeley et al. 2001). An advantage of these approaches is that they can incorporate any ascertainment bias created by selecting SNPs with a high diversity.

5. Follow up work: Once putatively selected genes are identified, we will use BLAST searches to determine the class of protein that they encode. We will also compare map positions with QTL identified in maize-teosinte or maize-maize crosses (Section 2). We expect to follow up in more detail on a small number of the identified selection-candidates that fall into the established research interests of individual PIs, (Doebley, domestication QTL; Buckler, kernel trait and flowering time genes; Kresovich, rapidly evolving genes). These genes will be studied using a combination of mutator knockout lines, association analysis and transformation. No funds are requested for these detailed analyses of the selection-candidates, since we expect these funds can be obtained from other sources.

6. Other analyses: The data we gather will allow several other questions to be addressed. Here are three examples. *How do mutational biases and weak selection shape the distribution of SNPs?* DNA substitution biases have been characterized in many systems. However, there have been surprisingly few orthologous comparisons of plant nuclear genes, particularly comparisons that consider multiple loci (Tiffin and Hahn 2002). As a result, there is very little information about: *i*) patterns of sequence change in plant nuclear genes, and *ii*) variation in molecular rates

among plant nuclear genes. *How is genetic diversity organized in the maize germplasm pool?* Although we have previously addressed this question using SSR data (Matsuoka et al. 2002b; Vigouroux et al. in prep a; Fukunaga et al. in prep), the SNP-polymorphism data will allow this question to be thoroughly reevaluated. *How is genetic diversity distributed in the maize genome?* We have addressed this question under our current grant with SSR and sequence data (Tenailon et al. 2001; Vigouroux et al. in prep b); however, we will now have sequence polymorphism data for 4000 loci, allowing us to address this question in far greater detail.

7. Results and products: Based on our on-going and proposed SNP discovery work, we will create a collection of 30,000 SNP in 4000 genes. The majority of these will have known map positions. These SNP will be key tools for the QTL and association analyses in Section 2, but also a key resource for the maize genetics community. All SNP will be documented in Gramene, MaizeGDB, and Genbank. Based on our multiple assays for selection, we will also produce a catalog of selection-candidate genes, perhaps numbering 400 or more. Selection-candidate genes will be tagged in Gramene as a guide for researchers with interest in these genes. We also anticipate a series of ten or more major publications describing the principal results.

Section 2: Functional Diversity

Background: The identification of genes and alleles that modify agronomic traits is the most direct approach for bridging genomics to plant breeding. Since most traits are under multigenic control, the identification of QTL is key to accessing the diversity of maize. This section will develop and apply large-scale approaches to dissect complex traits in maize and teosinte. The benefits will be the development of a diverse and unified set of genotypically and phenotypically characterized germplasm, the mapping of thousands of QTL, and an understanding of the basis of complex traits and domestication. This will facilitate the fine mapping of 100s of QTL per year, representing a substantial advance over the decade time frame required to dissect a single QTL using current methodologies and resources.

How can functional diversity be evaluated on a genome-wide scale? The first genome surveys for QTL employed linkage mapping with backcross (BC), F₂, and recombinant inbred line (RIL) mapping populations (Edwards et al. 1987; Tanksley et al. 1982; Paterson et al. 1988). Linkage mapping is still the most efficient and statistically powerful way to evaluate an entire genome. However, the resolution of this approach is often limited to 10 cM regions (~10 million bp in maize) and each population only surveys a maximum of two alleles. In contrast, association mapping relies on surveying natural variation to exploit the rich history of alleles and recombination through evolution. Association mapping in diverse maize also localizes QTL to 2000 bp and allows the evaluation of a wide range of alleles. The weakness of association approaches is that they can have less statistical power per allele. We propose to combine linkage and association mapping, thereby joining their complementary strengths. This will allow statistically powerful genomic scans for QTL. The joint analysis of linkage and association mapping of QTL in multiple, related populations will serve as an important test case for the application of these methods. If our methods prove successful in characterizing traits in maize, they potentially can be extended to any trait for any plant or crop species.

Specific Aim 2A: Develop mapping populations to capture diverse alleles. The key resource needed for high throughput QTL dissection is a variety of mapping populations that capture

diverse alleles. Six complementary sets of mapping populations will be developed. The teosinte-maize F₂ and maize (or maize-teosinte) RILs will be used for genomic scans by linkage mapping. The teosinte and maize association populations will be used for fine-mapping. The diallel and near isogenic lines (NILs) will be used in *post hoc* fashion to evaluate specific hypotheses.

1. *Teosinte-Maize Mapping Populations*: Prior QTL studies with maize and teosinte (Doebley et al. 1990; Doebley and Stec 1993) were performed more than ten years ago when both marker technologies and statistical methods for QTL analysis were in their formative years. While these studies proved informative and enabled us to clone one major domestication QTL (Doebley, Stec and Hubbard 1997) and to define genetically several others (Dorweiler et al. 1992; Doebley, Stec and Gustus 1995; Bombliet et al. in review), they lack the precision now possible with improved methodologies, especially as related to QTL of moderate or small effect. Given the importance of maize as a crop and the “teosinte-to-maize-transition” as a model for crop domestication, it is crucial to update these studies using more powerful methodologies. Our goals are (1) to better define QTL effects and map locations using larger populations and denser marker coverage, and (2) to create sets of BC₁ lines and RILs that will be available to the maize and cereal genetics communities. We will create two maize-teosinte F₂ populations of 1000 plants each from crosses of *ssp. parviglumis* and primitive maize landraces. The plants of one population will be backcrossed to B73 and the other to W22. These large populations will allow precise estimates of QTL effect, position and epistasis. They will eliminate the imprecision of QTL detection and position that occurs when using populations of a few hundred progeny (Beavis 1994). The BC₁ lines will provide 40 cross-overs within 1 cM of each QTL, facilitating map-based cloning. Precise mapping of the QTL will enhance the joint QTL and association mapping (below). We will also create 200 RILs from each maize-teosinte F₂ population to provide permanent mapping populations. We will do two generations of backcrossing to B73 or W22 prior to selfing (Doganlar et al. 2002).

Current status: The two maize-teosinte F₁ crosses were made this winter and F₂ seed will be available at the start of the project.

2. *Teosinte Association Mapping Populations*: Association analyses in teosinte will focus on *ssp. parviglumis* since it is the ancestor of maize, has the highest levels of genetic variation of any teosinte, and is geographically wide-spread. We have seed of 76 *ssp. parviglumis* populations from Mexico. These will be grown in two seasons in winter nurseries and four plants per population phenotyped and genotyped each season (608 plants total). We estimate that this depth of sampling will allow us to detect QTL with effects down to 1% of the phenotypic variance of the population (Long and Langley 1999). This sampling will also enable us to capture and assay essentially the entire range of allelic variation in *ssp. parviglumis*. Seed of *ssp. parviglumis* were obtained from NCRPIS, INIFAP, or CIMMYT and are available to others who would wish to build on the system we develop. SSR and SNP genotypes for these 76 populations are either already available or will be gathered during the project, enabling us and others to control for population structure (Thornsberry et al. 2001). Our analyses will show how functional variation is partitioned in *parviglumis* and provide a guide to the future allele-mining of these populations.

Current status: An initial set of 76 populations of *ssp. parviglumis* was grown in our 2003 winter nursery. Phenotypic data and DNA extractions are being done on the first 304 plants and

association analyses will be performed with candidate genes (*tb1*, *D8*, *id1*, *zfl1*, *zfl2* and others) for which we have already developed SNP assays (see Methods below).

3. Maize RIL Populations: To map much of the segregating variation in maize, 25 RIL mapping populations will be created. Using the SSR marker evaluations and the CORESETS Program (see Prior Support) (Liu et al. in prep), we have selected 25 diverse lines (here after referred to as the 25DL) that capture 80% of the nucleotide polymorphism in maize. In order to provide a uniform evaluation background, each line was crossed to B73 (the standard US inbred) to form 25 mapping populations. *Specifics*: Each of the 25 RIL populations will have at least 200 RILs, each descended from a unique F₂ plant; over 5000 RILs will be produced in total. Using single seed descent and low density planting, we have 88% success in advancing lines per generation. Since we started with 600 S₁ individuals per population, we should recover roughly 360 RILs at the F₆ generation. To minimize delays caused by weather, the populations will be split and advanced separately in North Carolina and Missouri during the summer, and in Florida and Puerto Rico each winter. Seed will be bulked for phenotyping and then long-term distribution will be done by the USDA Maize Stock Center (see Appendix A-1). The established intermated B73 x Mo17 population will be included in genotypic and phenotypic surveys (Lee et al. 2002). *Current Status*: One population is at S₃, three are at S₂, and the remainder is at S₁.

4. Maize Association Mapping Population: In the previous phase of this project, 102 diverse inbred lines were phenotyped in 12 field seasons, sequenced at 30 candidate genes, and genotyped for 141 SSR markers. These lines permitted 1500 bp (Remington et al. 2001) resolution for evaluation of flowering and starch quality candidate genes (Thornsberry et al. 2001). These lines have also been distributed to a wide range of researchers. The main limitation of this population was its statistical power to evaluate QTL that explained less than 8% of the variance (Long and Langley 1999). To overcome this, we propose association mapping with a larger population that will provide statistical power to detect QTL that explain as little as 2% of the variation. *Specifics*: Dr. Goodman has collected an additional 203 very diverse inbred lines from throughout the world. Together, these 305 lines represent a wide range of phenotypic and genotypic diversity and were extensively sampled for SSRs (see Prior Support) (Liu et al. in prep). Hybrids of the 305 lines with B73 will also be phenotyped to examine dominance and heterosis effects for association studies. *Current Status*: The 305 lines have been phenotyped in two environments for 41 traits.

5. Maize Half Diallel: To aid in combining QTL results from different populations all 25DL were crossed to B73. However, there will be many situations where comparisons of alleles among the 25DL are required. We will derive the base resource for these comparisons by generating a half diallel among the 25DL. The F₁s will be phenotyped and advanced to F₂s as needed to test specific hypotheses. There are four main areas where the diallel F₂s will be used: 1) to test for allelism of QTL discovered among the 25DL X B73 RIL populations; 2) to test the full array of dominance and epistatic allele combinations; 3) to test QTL effects in a heterotic background; and 4) to test combinations of favorable alleles for transgressive segregation of agronomic trait values. *Specifics*: F₁ seed will be made for all combinations of the 25DL. Phenotypes will be determined at the F₁. Seed will be distributed to coPIs within the project and other researchers as requested. Genotyping of the S₁ will be on a case-by-case basis. *Current Status*: Diallel will be generated first winter of the grant in Puerto Rico.

6. Related resources: Three other germplasm sets are being created that will complement those described above. (1) A subset of the 25DL are being used to create backcross and NIL populations with B73 as the recurrent parent. This work is partially supported by the USDA

(Jim Holland). These populations will be valuable for verifying the effects of QTL and for positional cloning of loci that cannot be efficiently mapped using joint association and RIL mapping. (2) The 25DL x B73 crosses are also being intermated for 5 generations. This will provide the opportunity to create higher resolution QTL mapping populations for the future (similar to the B73 x Mo17 population). (3) The SSR survey in our prior grant identified landrace accessions that had allelic diversity not captured in maize inbred lines. Inbreds will be derived from these landrace accessions. (4) A set of 12 to 16 teosinte inbreds will be developed.

Specific Aim 2B: QTL mapping in maize and maize-teosinte populations. The justification behind the laborious effort to develop these populations is that the resulting common set of resources can be used to evaluate essentially any trait for maize and teosinte. In this way we hope to do for trait analysis what EST projects have done for gene discovery – to transform QTL analysis into a comprehensive (genomics) activity. As a test of the efficacy of our approach, we will use the populations to dissect selected domestication, agronomic, and plant development traits. These populations will provide 2-15 cM resolution of QTL for roughly 50 traits. The phenotypes will be evaluated in up to 15 environment-replication combinations. QTL position, magnitude, and epistatic and environmental interactions will be evaluated through a variety of statistical approaches. The thousands of QTL identified will be a resource for domestication studies, breeding, understanding the genetic architecture of maize, and dissecting QTL.

Genotyping: To map QTL in these populations, a marker is needed every 10 cM, therefore 300 SNP markers will be developed that evenly span the maize genome. This will provide roughly one informative SNP marker every 5-10 cM for all the populations. Preliminary data on the SNP discovery (Section 1) and genotyping (Methods, below) are presented.

Phenotyping: Acquiring high quality, replicated phenotypic data is critical to the accurate mapping and evaluation of phenotypic effects. We plan to score the maize RILs in up to 15 environment-replication combinations, and the maize-teosinte populations in two seasons in winter nursery. Each replication will consist of 1 to 15 individuals from each line. We have developed a phenotype tracking system for planting, barcoding, and scoring individual plants called *MaizeMeister* (www.maizegenetics.net). *MaizeMeister* also integrates the data directly in a database. This system allows accurate and uniform data collection across multiple environments and multiple scorers. About four million phenotypic data points will be collected. *Traits:* A wide range of plant development, agronomic and domestication traits will be mapped in order to help understand the basic development of maize and teosinte and to manipulate maize morphology for agronomics. Our core list of traits is given in Table 2.

Table 2: Core Traits
Domestication: tillers, lateral branch number and length, barren nodes above the ear, ear structure and phyllotaxy, single vs. paired spikelets, glume shape and induration, ear number per plant, kernel weight
Agronomic: silk and tassel date, plant and ear height, lodging, yield
Plant Development: nodes above/below ear, ear number, brace root nodes, tillers, leaf length & width, leaf angle
Inflorescence Development: tassel length, spike length, branch angle, spikelet density, tassel weight, ear length, diameter, cob diameter, row number, ear mass
Kernels: kernel starch, oil, protein, moisture, length, width, depth, mass, volume, color, halo, opacity
Disease: common rust, Southern Corn Leaf Blight, European Corn Borer, <i>Fusarium</i>

A meeting of all evaluation collaborators will be held during the second year of the grant to discuss logistics and to determine which traits are to be scored in each location. While traits in Table 2 comprise the initial application of the germplasm resources (2A), it is a long-term goal

of the coPIs to apply these populations to understand the molecular genetic basis of maize responses to biotic challenges (Holland and McMullen) and to abiotic conditions (Buckler). Since the 25DL come from diverse geographic and climatic regions and therefore have different histories of pathogen-stress exposure, the 25DL populations constitute an excellent resource for dissecting a wide range of abiotic conditions and biotic challenges. Holland, Goodman, and McMullen will also conduct initial screens for response to *Fusarium* ear rot, common rust, Southern Corn Leaf Blight, and European corn borer in the maize 25DL and diallel lines. The coPIs will seek support beyond this proposal and establish collaborations to dissect many of these traits in greater depth (e.g. disease and kernel quality).

QTL Analysis: Single marker and composite interval mapping (CIM) will be used to investigate QTL (Zeng 1993; Zeng 1994). QTL by environment interactions will be evaluated using joint CIM approaches that explicitly evaluate these interactions (Basten et al. 2002; Jiang and Zeng 1995). Multiple interval mapping will also be used to search for epistatic interactions (Kao and Zeng 1997; Kao et al. 1999). One dimensional genomic search approaches will also be used to explore epistasis (Jannink and Jansen 2001).

A major weakness of traditional QTL mapping in one population is that often loci are fixed for similar alleles in any single cross. Mapping across multiple populations substantially reduces this possibility. Some recent approaches have been developed for combining QTL mapping results over multiple populations (Xu 1998; Yi and Xu 2002). These methods will be used to identify regions with substantial QTL effects and will effectively summarize the QTL effects over the populations. The dual approach of using the 25 DL populations to capture most variation in maize plus the two large maize-teosinte populations to capture domestication QTL fixed in maize but differentiated between maize and teosinte will provide a comprehensive system for QTL analysis in the maize germplasm pool.

Specific Aim 2C: Dissect QTL using candidate gene associations. The maize and teosinte association populations will be used to evaluate candidate genes and fine-map QTL associated with domestication, plant development, pathogen, and agronomic traits. These populations will provide gene level resolution of QTL. The surveys will also identify alleles and markers that could be used in breeding or positional cloning.

There will be three types of candidate genes evaluated throughout this project. The first set of candidates will be based on knowledge of trait related pathways. Maize mutagenesis, biochemical studies, and grass and Arabidopsis comparative genomics can provide a wealth of strong candidate genes for some pathways. The second set of candidate genes will be based on QTL regions identified in section 2B. The maize genome sequence of B73 should be substantially complete by the time many of the QTL are available from 2B. The genes in these QTL intervals that are most likely to affect relevant pathways will be chosen as candidate genes. The third set will come from selected genes identified under Section 1 of this project.

Trait Candidates: Thirty candidate genes involved in flowering time, kernel quality, and plant development traits were the focus of prior support. We will expand our survey to roughly 750 genes. For plant development, many of the candidate genes are available from other NSF funded projects including “Regulation of Inflorescence Architecture in Maize” for tassel and ear traits (see letters from Hake and Rocheford). These inflorescence candidate genes are also excellent candidates for the domestication genes differentiating maize and teosinte since the essence of maize domestication was a radical change in inflorescence architecture. Additionally, a large number of genes for flowering defined in Arabidopsis has already allowed our group to

identify nearly 100 maize candidate genes in this pathway. Similarly for the kernel quality pathway, molecular and biochemical research has identified 100s of candidate genes, and several NSF funded projects are continuing large-scale discovery in maize and Arabidopsis (see letter from McCarty). Finally, we have done microarray-based expression profiling of endosperm tissue across 102 diverse inbred maize lines, allowing us to identify many candidate genes that correlate with several kernel quality traits.

Genotyping: SNPs will be scored in the 25DL for the diallel and NIL populations by direct sequencing (Section 1A) as needed to evaluate specific hypotheses. SNP assays will be outsourced to a commercial provider (see Methods, below) to score the full set of maize inbreds and teosinte lines for the association studies (2C).

Phenotyping: Phenotypic analysis on the association populations will be conducted in parallel with that of the maize RIL and maize-teosinte F₂ populations (2B).

Association Analysis: Candidate gene association approaches in diverse germplasm that control for population structure were developed (Pritchard et al. 2000a; Pritchard et al. 2000b; Remington et al. 2001; Thornsberry et al. 2001) (see Prior Support). We will develop methods to study epistasis, including model-free approaches such as multifactor-dimensionality reduction (Ritchie et al. 2001). Methods similar to transmission disequilibrium tests (TDT) will be used to combine F₂ populations produced in the diallel.

Joint Linkage and Association Analysis: Joint analysis of the linkage and association populations will result in extraordinary power for QTL detection. Historically, linkage analysis and association studies were conducted as independent stages of research; with linkage analysis serving to direct where subsequent association studies should focus. Recently, however, joint linkage and association mapping has been developed and successfully applied in humans (Soubrier et al. 2002) and cattle (Farnir et al. 2002; Meuwissen et al. 2002). Using these approaches, QTL previously identified only within broad confidence intervals and with limited certainty were resolved to single narrow QTL peaks. Essentially, the association approaches identify regions that are identical by descent between the various mapping populations, and then the results from different populations are statistically combined. We will devote substantial effort to the development of statistical approaches that extend methods for combining populations and thereby provide both high resolution and greater statistical power.

Candidate Gene Validation: Once the interesting alleles have been identified for a candidate gene, then complete sequencing of the genes for representative alleles will be conducted. This will identify candidate polymorphisms that may cause the phenotypic variation. Putative polymorphisms will be evaluated in multiple populations, including NILs. Expression patterns and evidence of locus duplication or deletion will be examined. More extensive follow up molecular or biochemical experiments will generally be done in collaboration with appropriate expert researchers. We will also maintain a web list of candidate genes, their potential associations, and interesting alleles that should be evaluated.

Results and Products: (1) Mapping populations will be created that capture much of the allelic diversity available in maize and teosinte in an organized fashion. This germplasm will be an important resource for mapping throughout the maize community. (2) Numerous traits will be mapped in multiple populations and 19 environments. Thousands of QTL will be identified, which will help unravel the genetic architecture of maize and identify the genes controlling these traits. (3) Combined linkage and association tests will evaluate the effects of 1000 candidate genes. Hundreds of QTL will be resolved to the gene and nucleotide level for domestication, agronomic, and developmental traits.

Molecular Methods:

The project requires that we generate 5 million SNP data points. We are confident that this goal can be achieved based on our exploratory work.

DNA Isolations: The project will require 20,000 DNA extractions. The bulk of these will be done at Cornell's Institute for Genomic Diversity under the supervision of Kresovich. While our current method of performing Qiagen plant DNA extraction in a 96-well format is adequate, we are testing a DNA Autogen 9600 to determine if it meets our quality needs.

SNP Genotyping: Because of the highly variable nature of the maize and teosinte genomes, it is essential that we identify both an assay and high-throughput analytical platform that will perform optimally in screening SNPs. Methods developed for assaying human DNA, which is much less variable than maize, may not be adequate (BioTechniques 2002; Weiner and Hudson 2002). Therefore, we undertook a pilot study to evaluate the efficiency of converting a random set of discovered SNPs into robust assays for genotyping maize and teosinte. We supplied Qiagen Genomics (Bothell, WA) with flanking sequence data for 111 characterized SNPs previously identified from 30 different genes. Qiagen was able to develop 100 assays (90%) and they subsequently tested these on a panel of 47 DNAs representing the full geographic range of domesticated maize and teosinte. The Qiagen genotyping technology employs a 2-step PCR assay in which the second PCR is allele-specific (Kokoris et al. 2000).

These 100 SNPs by 47 DNAs were assayed in replicate for a total of 9400 assays. Assays for 7 SNPs failed entirely and another 7 produced results that could not be readily grouped into the expected two homozygous and one heterozygous class. The remaining 86 SNPs were robust across the breadth of germplasm, giving an 86% conversion rate. The assays appeared to be highly accurate based on prior sequence information. The call rate ranged from 77% to 100% per target, with an average of 96.9%. There was only one instance of discordance between replicate assays, for a concordance rate of 99.98%. Additionally, the Buckler lab has implemented an alternative platform for scoring SNP by fluorescent single-base extension with Luminex flow cytometry. This approach will be used for some studies and for problematic loci.

SNP technology is rapidly changing and the proposed SNP work will begin 2 years from now. Our pilot study proves the feasibility of SNP genotyping for diverse maize germplasm. We have or will soon initiate pilot studies with other vendors (Orchid and Illumina) and we are following corporate reorganizations (e.g. Qiagen is selling their technology platform). We will determine a final supplier when the resources are needed. Additionally, we are confident that we can produce the data inhouse if that become the most practical alternative.

Informatics:

This project informatics will develop efficient data management within the group and effective dissemination of the data and analyses to the public. Synergistic efforts will be key to the success of the bioinformatics, and Drs. Ware and Stein will coordinate these efforts.

Data Management: A workflow database (DB) system will be built for tracking the diversity data and work synergistically with the current coPI's systems. The Oracle database management system (DBMS) will manage the workflow DB and be accessible by collaborators, and serve the role of organizing and coordinating the various laboratories' activities. Laboratory information

systems (LIMs) are already set up in several labs to track the various types of data: SNP discovery pipeline (Gaut/McMullen, Section 1), DNA sample tracking (Kresovich/Buckler), and germplasm development (Buckler; MaizeMeister – Section 2). SNP genotypic data includes substantial informatics on primer design, SNP validation, and SNP calling. The resulting data is accessible to direct database input. Data will be moved from the LIMs and the contract SNP genotypes to the laboratory workflow database through a series of Perl scripts, XML files and direct database dumps. The laboratory of Dr. Stein brings considerable experience to the project with the management and display of SNP data from the human SNP consortium (snp.cshl.org).

Drs. Ware and Stein will design the workflow DB, building on ideas from Gramene, Panzea, and Cornell's Institute for Genomic Diversity (IGD) sample tracking database. The system will be developed within the first 6 months of the project. This database will be maintained at the Buckler/Kresovich labs with a mirror at CSHL. Between the IGD and Buckler USDA-ARS groups at Ithaca, a full time DB programmer and system administrator are available to maintain the workflow DB. GDPC (Genomic Diversity and Phenotype Connection; USDA-ARS funded) and its data browser will provide direct access to the diversity data.

Presentation/Analysis: Gramene (www.gramene.org) will be the main repository of the project's information. This DB will provide substantially more public exposure than the previous project's Panzea (statgen.ncsu.edu/panzea/). It will hold the part of the workflow database that is of interest to others, including such things as raw sequence, genotypic data, aligned clusters, results of motif searches, results of selection tests, genetic maps, and phenotypic data derived from the QTL linkage and association studies. The data will be updated on the public site (Gramene) on a quarterly basis and will include the integration of information from other publicly accessible data sets to provide added value to the primary data. Sequence and trace files generated through the SNP discovery will also be made available through Genbank and NCBI's trace depository. All mapping data will also be easily accessible in standard formats. The information generated by this project will be made accessible to the community via Web and FTP sites maintained at Cold Spring Harbor Laboratory. All data on the public web site will be available in a variety of bulk forms, including text files, XML files, and Oracle database dumps.

We will present the resources developed by the project as searchable Web pages that will form a browse-able front-end to the data within Gramene. These displays will be accessible via a series of Web-based queries, which can be called either from within Web query forms (which we will provide), or programmatically via specially formatted URLs (whose syntax will be published). The pages will be interactive, providing internal links to other resources in Gramene, and act as worldwide portals to other genomic resources via external links. It is anticipated over the course of the project that the implementation of display pages will mature as the data sets, analysis tools and potential links to resources become available.

Candidate Gene Annotation: These annotations will provide information on the 4000 candidate genes including; sequence, genetic and physical map position, diversity statistics, associated maize and rice sequences, associated plant anatomy terms derived from the EST libraries, mappings to maize and rice genomic sequence. Protein annotations will include mapping to maize SwissProt proteins, Pfam and Prosite mappings, Interpro to Gene Ontology (GO) mapping. These annotations will rely on adapting current Gramene displays and linking to MaizeGDB, TAIR, PlantGDB, Genbank, SwissProt, and NSF sequencing projects. These

annotations will provide key information for classifying selection effects and potential trait associations. (Yr 1-2)

SNP display. This will provide information on the germplasm and location of the SNP in relationship to gene structure (exon, intron, etc.). From the SNP pages, links will be provided to the comparative map viewer through a map location. As the data set matures, links will be made to phenotype data through the QTL analysis. Additional displays will also be derived from Dr. Stein's work on the NHGRI human haplotype map project. (Yr 1-2)

Germplasm displays: The distribution of genotypic and phenotypic data will be displayed in both geographical and phylogenetic tree displays of this project's germplasm. The page will provide external links to GRIN for more information on accessions. These geographic tools will help in dissecting the ecological distribution of the landraces (Section 1B). (Yr 2-3)

QTL display/analysis: This system will provide an integrated tool to examine QTL including: QTL positions across multiple populations, markers within intervals, sequence diversity, candidate gene associations, and joint linkage-association *P*-values. Links to germplasm, phenotypes and genotypes will be provided. Plant and trait ontologies will be used to help organize the display of QTL for related traits. Genes within QTL intervals will be searchable by gene ontology terms to facilitate the identification of candidate genes based upon function, process and cellular location. Where significant collinear regions with rice are present, links with the rice QTL and genome maps will be made. Tools will be built to facilitate running bulk QTL analyses. Additional algorithms and visualization software will be included from Dr. Stein's work on the NHGRI human haplotype map project and Dr. Buckler's USDA CRIS project. (Yr 3-5)

Comparative map displays (<http://www.gramene.org/cmap>): Gramene has extensive comparative mapping capabilities. This project will create 27 linkage maps, which will be portrayed using this system. Differences in recombination and its relationship to diversity will be explored. Statistically defined collinear regions between species will also be visualized to the user to traverse within and between species maps. (Yr 4-5)

Significance and Justification:

Sydney Brenner in his 2002 Nobel Lecture suggested that the 21st century will see a shift in biology to the study of diverse populations for the dissection of biological complexity. Our project fits precisely his prediction as it uses diverse populations to address two important issues in biology. First, our project will sample diverse populations to characterize how selection has modified both individual maize genes and the entire maize genome. This approach will allow us to identify the genes and pathways that play key roles in maize domestication, agronomics and local adaptation. Many of these genes will be responsible for resource allocation, yield, and grain quality in maize. This work will also provide insights into the basic mechanisms of evolution and population genetics. Second, our project will identify numerous alleles that control functional diversity in maize. We will identify and evaluate thousands of QTL and candidate genes for agronomic traits across a very wide range of germplasm. These genes and germplasm will provide a wealth of alleles for plant breeding and molecular dissection of pathways. By this means, we hope to help bridge the gap between genomics and applied crop improvement. This is a critical goal since in 2001 maize became the number one production crop in the world (FAO Statistics 2002).

Project Team:

While our past project was highly successful (see Prior Support), we believe this renewal is in an even stronger position to excel. We have matured as a group, learned to interact effectively (see numerous co-authored papers), and we have established many collaborations throughout the plant science community. We have added two new coPIs: McMullen, who brings proven skills in high throughput genotyping, SNP discovery, and QTL mapping, and Holland, who brings skills in quantitative genetics and maize breeding. We have also added Stein and Ware for informatics. They bring extensive expertise in data management and display and will facilitate integrating project data into Gramene, thereby making it accessible to the entire cereal genetics community. These additions along with the knowledge that the members of our past project bring in high throughput genotyping, statistical genetics, population genetics, molecular evolution, quantitative genetics, maize germplasm, and maize genetics creates a team that will accomplish the goals we have outlined in this proposal.

Project Timetable:

Time	Molecular Diversity		Functional Diversity			Informatics	
P1 2004	Prep DNA for selection study; Model maize demography	Perform test of selection	SNP Discovery, Development and Assay	Teosinte Linkage Analysis	Develop Maize RIL and Diallel Pops.; teosinte F2 and BC lines	Internal DBs, Upload tools	Basic Public Access
P2 2004							
P1 2005	Begin SNP genotyping; Model maize demography	Teosinte Association Analysis	Maize & Teosinte Association Analysis	SNP teosinte F2 populations	SNP RILs	Genomic SNP diversity displays	Display germplasm relationships
P2 2005							
P1 2006	Complete most SNP genotyping; Manuscript prep	Phenotype Maize/Teosinte Populations	Maize & Teosinte Association Analysis	Blowup maize Pops	Integrate and apply QTL mapping algorithms, QTL display	Statistical Analysis Development	Analysis and Display of Candidate Regions and Alleles
P2 2006							
P1 2007	Residual SNP genotyping	Statistical Analysis Development	Maize & Teosinte Association Analysis	Post Hoc Trait Dissection	Compare Molecular and Functional Diversity, Linkage Maps	Statistical Analysis Development	Analysis and Display of Candidate Regions and Alleles
P2 2007							
P1 2008	Final manuscript preparation	Statistical Analysis Development	Maize & Teosinte Association Analysis	Post Hoc Trait Dissection	Compare Molecular and Functional Diversity, Linkage Maps	Statistical Analysis Development	Analysis and Display of Candidate Regions and Alleles
P2 2008							

Since this is a renewal, this grant would start Jan 1, 2004. A little later than other PG grants.

Education, Training and Diversity:

The National Science Education Standards (NRC 1996) recommend an integrated approach to education that combines science, history and culture into a unified curriculum. These practices foster a deeper understanding of scientific concepts by appealing to students on different levels (whether they are more inclined toward the sciences or humanities), presenting topics as unified pictures rather than isolated facts, and making the concepts more relevant to real life.

Maize lends itself extremely well for use as an educational tool, and particularly to an integrated educational approach. Maize is one of the best examples of crop domestication and an excellent visual example of genetic inheritance. Due to its large size and fast growth, it provides a good model of the plant life cycle, and is also a good example of agricultural techniques such as intercropping. Maize can be used in education on many levels: while an elementary school child can learn about food and plants from growing maize in the school garden, a college student can use maize to study transposons. Furthermore, maize has played a central role throughout history and in the cultures of Native American peoples (Viola and Margolis 1991). Finally, Barbara McClintock's achievements make maize an excellent example of the contribution of women in science.

We plan to take advantage of these features and use maize as a tool to present scientific, historical, and cultural concepts to middle/high school students. While there are several educational web sites about maize, many students do not have ready access to a computer, or the time or inclination to read through information in a text format. This project will bring the many-faceted story of maize to the students. We will construct mobile "story boards" to be placed in the hallway display cases of selected schools. These story boards will contain pictures and real examples of a diverse array of maize and teosinte. Simple pictorials and brief text boxes will explain the domestication of maize from both historical and genetic perspectives. A pictorial description of the maize life cycle will be included, as will cultural and historical uses of maize, in particular those related to Native American and Meso-American cultures. These story boards will be made in sections that are easy to assemble and disassemble, making it possible to rotate the boards through a large number of schools. A list of relevant web resources for additional information and an email contact will be given in case the students have questions. To introduce the story board, a presentation will be arranged at each school. This presentation will also include information about careers in plant science, not only giving the students educational information but also introducing them to real scientists. In addition, seeds of various maize and teosinte types will be available to teachers who would like to plant a display garden.

Schools will be selected near each of the coPIs' home institutions, but only in either very rural areas, or in high minority/resource poor areas (e.g. New York City & Santa Ana, CA, which has a large Latino community). Schools on or near Native American reservations or in areas of high Latino populations will be especially targeted, increasing the numbers of minority, rural, and resource poor students that we will impact. These schools are unlikely to be reached by other university outreach programs, thus we will be able to have a much larger impact than by simply targeting local schools. Middle-high school principals will be contacted, and will help select the appropriate target grade(s) in each school. Story boards will be transported and assembled in the schools by one of the coPIs, graduate students or the outreach director, introduced with a presentation, and left for several weeks. In this way, we can reach a minimum of 3 schools per year, possibly reaching as many as several thousand rural students with the story of maize. Dr. Fulton will contact the schools and coordinate this part of the project.

To complement the project's outreach to the middle/high school level, we will also be employing a minimum of 9 undergraduate students and 8 graduate students per year. We will aim to include a high number of minority students in this group by contacting the minority office on each campus to announce the positions.

Literature Cited

- Basten, C., B. S. Weir and Z. B. Zeng, 2002 QTL CARTOGRAPHER, Version 1.16. Department of Statistics, North Carolina State University, Raleigh, NC.
- Beaumont, M. A. and R. A. Nichols, 1996 Evaluating loci for use in the genetic analysis of population structure. *Proc. R. Soc. London Ser. B* 263:1619-1629
- Beavis, W. D., 1994 The power and deceit of QTL experiments: lessons from comparative studies. Pp. 250-266 *in* Proceedings of the 49th Annual Corn and Sorghum Industry Research Conference, Chicago.
- BioTechniques, 2002 SNPs: Discovery of Markers for Disease. June supplement.
- Bomblies, K., R. L. Wang, B. Ambrose, R. Meeley, R. Schmidt and J. Doebley, 2003 Duplicate *FLORICAULA/LEAFY* homologs *zfl1* and *zfl2* control inflorescence architecture and flower patterning in maize. *Development* (in review).
- Brookfield, J. F. Y., 2001 Population genetics: the signature of selection. *Curr. Biol.* 11:R388-R390.
- *Buckler, E. S., J. M. Thornsberry and S. Kresovich, 2001 Molecular diversity, structure and domestication of grasses. *Genet. Res.* 77:213-218.
- *Buckler, E. S., 2001 TASSEL: Trait Association, Evolution, and Linkage Analysis software package. <http://www.maizegenetics.net/>.
- *Buckler, E. S., and J. Thornsberry, 2002 Plant molecular diversity and applications to genomics. *Curr. Opin. in Plant Biol.* 5:107-111.
- *Casa, A. M., S. E. Mitchell, O. S. Smith, J. C. Register, S. R. Wessler and S. Kresovich, 2002 Evaluation of *Hbr* (MITE) markers for assessment of genetic relationships among maize (*Zea mays* L.) inbred lines. *Theor. Appl. Genet.* 104:104-110.
- *Casa, A. M., S. E. Mitchell, Q. Zhang and S. Kresovich, Evolutionary implications of *Heartbreaker* (*Hbr*) insertional polymorphisms in maize and its wild relatives. *Theor. Appl. Genet.* (submitted).
- Ching, A, K. S. Caldwell, M. Jung, M. Dolan, O. S. Smith, S. Tingey, M. Morgante and A. Rafalski, 2002 SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. *BMC Genet.* 3:19:1-14.
- Doebley, J. and A. Stec, 1993 Inheritance of the morphological differences between maize and teosinte: comparison of results for two F₂ populations. *Genetics* 134:559-570.
- Doebley, J., A. Stec, J. Wendel and M. Edwards, 1990 Genetic and morphological analysis of a maize-teosinte F₂ population: implications for the origin of maize. *Proc. Natl. Acad. Sci. USA* 87:9888-9892.
- Doebley, J., A. Stec and C. Gustus, 1995 *Teosinte branched1* and the origin of maize: evidence for epistasis and the evolution of dominance. *Genetics* 141:333-346.
- Doebley, J., A. Stec and L. Hubbard, 1997 The evolution of apical dominance in maize. *Nature* 386:485-488.
- Doganlar, S., A. Frary, H. M. Ku and S. Tanksley, 2002 Mapping quantitative trait loci in inbred backcross lines of *Lycopersicon pimpinellifolium* (LA1589). *Genome* 45:1189-1202.
- Dorweiler, J., A. Stec, J. Kermicle and J. Doebley, 1993 *Teosinte glume architecture 1*: A genetic locus controlling a key step in maize evolution. *Science* 262:233-235.
- Edwards, M. D., C. W. Stuber and J. F. Wendel, 1987 Molecular-marker-facilitated investigations of quantitative-trait loci in maize. I. Numbers, genomic distribution and types of gene action. *Genetics* 116:113-125.

- Ewens, W. J., 1972 The sampling theory of selectively neutral alleles. *Theor. Pop. Biol.* 3:87-112.
- Excoffier, L. and M. Slatkin, 1995 Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.* 12:921-927
- Eyre-Walker, A., R. L. Gaut, H. Hilton, D. L. Feldman and B. S. Gaut, 1998 Investigation of the bottleneck leading to the domestication of maize. *Proc. Natl. Acad. Sci. USA* 95:4441-4446.
- Farnir, F., B. Grisart, W. Coppieters, J. Riquet, P. Berzi, N. Cambisano, L. Karim, M. Mni, S. Moisisio, P. Simon, D. Wagenaar, J. Vilkki and M. Georges, 2002 Simultaneous mining of linkage and linkage disequilibrium to fine map quantitative trait loci in outbred half-sib pedigrees: revisiting the location of a quantitative trait locus with major effect on milk production on bovine chromosome 14. *Genetics* 161:275-287.
- Fay, J. and C. I. Wu, 2000 Hitchhiking under positive Darwinian selection. *Genetics* 155:1405-1413.
- *Flint-Garcia, S. A., J. M. Thornsberry and E. S. Buckler, 2003 Structure of linkage disequilibrium in plants. *Ann. Rev. Plant Biol.* 54: (in press).
- *Fukunaga, K., Y. Vigouroux, Y. Matsuoka, K. Liu, E. S. Buckler, J. Hill and J. Doebley, 2003 Genetic structure and phylogeny of teosinte as revealed by multilocus microsatellite genotyping. (draft for Evolution).
- *Garcia, A. and Buckler, E. S., 2003 MaizeMeister: Integrated phenotypic and germplasm tracking software. <http://www.maizegenetics.net/>.
- *Gaut, B. S., 2001 Patterns of chromosomal duplication in maize and their implications for comparative maps of the grasses. *Genome Res.* 11:55-66.
- *Gaut, B. S., 2002 Evolutionary dynamics of grass genomes. *New Phytologist* 154:15-28.
- *Gaut, B. S., M. Le Thierry d'Ennequin, A. S. Peek, M. C. Sawkins, 2000 Maize as a model for the evolution of plant nuclear genomes. *Proc. Natl. Acad. Sci. USA* 97:7008-7015
- *Gethi, J. G., J. A. Labate, S. Kresovich, M. E. Smith and K. R. Lamkey, 2002 SSR variation in important U.S. maize inbred lines. *Crop Sci.* 42:951-957.
- *Goodman, M. M. and M. L. Carson, 2001 Reality vs. myth: Corn breeding, exotics, and genetic engineering. *Annual Corn Sorghum Research Conference Proc.* 55:149-172.
- Hilton, H., and B. S. Gaut, 1998 Speciation and domestication in maize and its wild relatives: evidence from the *Globulin-1* gene. *Genetics* 150:863-872.
- Hudson, R. R., 1990 Gene genealogies and the coalescent process. *Oxford Surv. in Evol. Biol.* 7:1-44.
- Hudson, R. R., M. Kreitman and M. Aguade, 1987 A test of neutral molecular evolution based on nucleotide data. *Genetics* 116:153-159.
- *Jaenicke, V., E. S. Buckler, B. Smith, J. Doebley and S. Pääbo, Early allelic selection in maize as revealed by Ancient DNA. (draft ready for submission to *Science*)
- Jannink, J. L., and R. Jansen, 2001 Mapping epistatic quantitative trait loci with one-dimensional genome searches. *Genetics* 157:445-54.
- Jiang, C., and Z. B. Zeng, 1995 Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics* 140:1111-27.
- Kao, C. H., and Z. B. Zeng, 1997 General formulas for obtaining the MLEs and the asymptotic variance-covariance matrix in mapping quantitative trait loci when using the EM algorithm. *Biometrics* 53:653-65.

- Kao, C. H., Z. B. Zeng and R. D. Teasdale, 1999 Multiple interval mapping for quantitative trait loci. *Genetics* 152:1203-16.
- Kauer, M., B. Zangerl, D. Dieringer, and C. Schlötterer, 2002 Chromosomal patterns of microsatellite variability contrast sharply in African and non-African populations of *Drosophila melanogaster*. *Genetics* 160:247-56.
- Kokoris, M., K. Dix, K. Moynihan, J. Mathis, B. Erwin, P. Grass, B. Hines and A. Duesterhoeft, 2000 High-throughput SNP genotyping with the Masscode system. *Mol. Diag.* 5:329-340.
- *Labate, J. A., K. R. Lamkey, S. E. Mitchell, S. Kresovich, H. Sullivan and J. S. C. Smith, Molecular, agronomic, and historical aspects of Corn Belt Dent varietal diversity. *Crop Sci.* (in press).
- Lee, M., N. Sharopova, W. D. Beavis, D. Grant, M. Katt, D. Blair and A. Hallauer, 2002 Expanding the genetic map of maize with the intermated B73 x Mo17 (IBM) population. *Plant Mol. Biol.* 48:453-61.
- *Liu, K., M. Goodman, S. Muse, J. S. Smith, E. Buckler and J. Doebley, 2003 Genetic Structure and Diversity among Maize Inbred Lines as Inferred from DNA Microsatellites. *Genetics* (in preparation).
- Long, A. D., and C. H. Langley, 1999 The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Res.* 9:720-31.
- *Matsuoka, Y., S. E. Mitchell, S. Kresovich, M. Goodman and J. Doebley, 2002a Microsatellites in *Zea* – variability, patterns of mutations, and use for evolutionary studies. *Theor. Appl. Genet.* 104:436-450.
- *Matsuoka, Y., Y. Vigouroux, M. M. Goodman, J. Sanchez G., E. Buckler and J. Doebley, 2002b A single domestication for maize shown by multilocus microsatellite genotyping. *Proc. Natl. Acad. Sci. USA* 99:8060-8064.
- Meuwissen, T. H. E., A. Karlsen, S. Lien, I. Olsaker and M. E. Goddard, 2002 Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. *Genetics* 161:373-379.
- Mondragon-Palomino M., Meyers B. C., Michelmore R. W., Gaut B. S., 2002 Patterns of positive selection in the complete NBS-LRR gene family of *Arabidopsis thaliana*. *Genome Res.* 12:1305-1315.
- Nielsen, R., 2000 Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics* 154:931-942.
- National Research Council, 1996 National science education standards. Washington, D.C.: National Academy Press. Also available at <http://www.nap.edu/html/nses/html>.
- Neuffer, M. G., E. H. Coe and S. R. Wessler, 1997 *Mutants of Maize*. Cold Spring Harbor Press, Plainview, NY.
- Otto, S. P., 2000 Detecting the form of selection from DNA sequence data. *Trends in Genetics* 16:526-529.
- Paterson, A. H., E. S. Lander, J. D. Hewitt, S. Peterson, S. E. Lincoln, S. D. Tanksley, 1988 Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. *Nature* 335:721-726.
- Payseur, B. A., A. D. Cutter and M. W. Nachman, 2002 Searching for evidence of positive selection in the human genome using patterns of microsatellite variability. *Mol. Biol. Evol.* 19:1143-53.

- Pritchard, J. K., M. Stephens and P. Donnelly, 2000a Inference of population structure using multilocus genotype data. *Genetics* 155:945-59.
- Pritchard, J. K., M. Stephens, N. A. Rosenberg and P. Donnelly, 2000b Association mapping in structured populations. *Amer. J. Hum. Genet.* 67:170-181.
- Przeworski, M., 2002 The signature of positive selection at randomly chosen loci. *Genetics* 160:1179-1189.
- Przeworski, M., R. R. Hudson and A. Di Rienzo, 2000 Adjusting the focus on human variation. *Trends Genet.* 16:296-302.
- Rafalski, A., 2001 Applications of single nucleotide polymorphisms in crop genetics. *Curr. Opin. Plant Biol.* 5:94-100.
- Reich, D. E., S. F. Schaffner, M. J. Daly, G. McVean, J. C. Mullikin *et al.*, 2002 Human genome sequence variation and the influence of gene history, mutation and recombination. *Nature Genetics* 32:135-142.
- *Remington, D. L., J. M. Thornsberry, Y. Matsuoka, L. M. Wilson, S. R. Whitt, J. Doebley, S. Kresovich, M. M. Goodman and E. S. Buckler, 2001 Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proc. Natl. Acad. Sci. USA* 98:11479-11484.
- Risch, N., and K. Merikangas, 1996 The future of genetic studies of complex human diseases. *Science* 273:1516-1517.
- Ritchie, M. D., L. W. Hahn, N. Roodi, L. R. Bailey, W. D. Dupont, F. F. Parl and J. H. Moore, 2001 Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Amer. J. Hum. Genet.* 69:138-147.
- Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Z. P. Levine, D. J. Richter *et al.*, 2002 Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419:832-837.
- *Sanchez G., J. J., M. M. Goodman and C. W. Stuber, 2000 Isozymatic diversity in the races of maize of the Americas. *Maydica.* 45:185-203.
- *Sanchez G., J. J., M. M. Goodman and C. W. Stuber, 2000 Isozymatic and morphological diversity in the races of maize of Mexico. *Econ. Bot.* 54:43-59.
- Schlötterer, C., 2002 A microsatellite-based multilocus screen for the identification of local selective sweeps. *Genetics* 160:753-63.
- Schmid, K. J., and D. Tautz, 1997 A screen for fast evolving genes from *Drosophila*. *Proc. Natl. Acad. Sci. USA* 94:9746-9750.
- Soubrier, F., S. Martin, A. Alonso, S. Visvikis, L. Tiret, F. Matsuda, G. M. Lathrop and M. Farrall, 2002 High-resolution genetic mapping of the ACE-linked QTL influencing circulating ACE activity. *Euro. J. Hum. Gen.* 10:553-561.
- Tajima, F., 1983 Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105:437-460.
- Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585-595.
- *Tallury, S. P., and M. M. Goodman, 2001 The state of the use of maize genetic diversity in the USA and sub-Saharan Africa. Pp. 159-180 in Cooper, H.D., C. Spillane and T. Hodgkin (eds.), *Broadening the Genetic Base of Crop Production*. CAB Publishing, Wallingford, UK.
- Tanksley, S. D., and S. R. McCouch, 1997 Seed banks and molecular maps: unlocking genetic potential from the wild. *Science* 277:1063-1066.

- Tanksley, S. D., H. Medinafilho and C. M. Rick, 1982 Use of naturally-occurring enzyme variation to detect and map genes-controlling quantitative traits in an interspecific backcross of tomato. *Heredity* 49:11-25.
- *Tenailon, M., M. Sawkins, A. Long, R. Gaut, J. Doebley and B. Gaut, 2001 Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea. mays* ssp. *mays*). *Proc. Natl. Acad. Sci. USA* 98:9161-9166.
- *Tenailon, M. I., M. C. Sawkins, L. K. Anderson, J. F. Doebley, S. M. Stack and B. S. Gaut, 2002 Patterns of diversity and recombination along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). *Genetics* 162:1401-1413.
- *Thornsberry, J., M. Goodman, J. Doebley, S. Kresovich, D. Nielsen and E. Buckler, 2001 *Dwarf8* polymorphisms associate with variation in flowering time. *Nat. Genetics* 28:286-289.
- *Thornsberry, J. M., J. Heer and E. S. Buckler, 2003a Quantitative variation in maize photosynthetic and respiratory rates between a broad range of inbred lines. *Curr. Genet.* (in review).
- *Thornsberry, J., M. M. Goodman, J. F. Doebley and E. S. Buckler, 2003b Analysis of maize flowering time genes using association tests. (draft for *Genetics*).
- *Tiffin, P., and B. S. Gaut, 2002 Molecular evolution of the wound induced serine protease inhibitor *wip1* in *Zea* and related genera. *Mol. Biol. Evol.* 18:2092-2101.
- Tiffin, P., and W. M. Hahn, 2002 Coding sequence divergence between two closely related plant species: *Arabidopsis thaliana* and *Brassica rapa* ssp. *pekinensis*. *J. Mol. Evol.* 54:746-753.
- *Vigouroux, Y., J. S. Jaqueth, Y. Matsuoka, O. S. Smith, W. D. Beavis, J. S. C. Smith and J. Doebley, 2002a Rate and pattern of mutation at microsatellite loci in maize. *Mol. Biol. Evol.* 19:1251-1260.
- *Vigouroux, Y., Y. Matsuoka, M. M. Goodman, J. Sanchez G. and J. Doebley, The maize gene pool: structure, diversity and ecogeography. *Genetics* (in preparation A).
- *Vigouroux, Y., M. McMullen, C. T. Hittinger, K. Houchins, L. Schulz, S. Kresovich, Y. Matsuoka and J. Doebley, 2002b Identifying genes of agronomic importance in maize by screening microsatellites for evidence of selection during domestication. *Proc. Natl. Acad. Sci. USA* 99:9650-9655.
- *Vigouroux, Y., S. Mitchell, Y. Matsuoka, M. Hamblin, J. S. C. Smith, J. Register, O. S. Smith, S. Kresovich and J. Doebley, 2003b A genetic diversity map of the maize genome using microsatellites: how domestication has shaped the diversity landscape. *Genetics* (in preparation B).
- Viola, H. J., and C. Margolis, 1991 *Seeds of change: a quincentennial commemoration.* Washington, D.C.: Smithsonian Institution Press.
- Wakeley, J., R. Nielsen, S. N. Liu-Cordero and K. Ardlie, 2001 The discovery of single-nucleotide polymorphisms-and inferences about human demographic history. *Amer. J. Hum. Genet.* 69:1332-47.
- Wall, J. D., P. Andolfatto and M. Przeworski, 2002 Testing models of selection and demography in *Drosophila simulans*. *Genetics* 162:203-216
- Wang, R.-L., A. Stec, J. Hey, L. Lukens and J. Doebley, 1999 The limits of selection during maize domestication. *Nature* 398:236-239.
- Watterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* 7:188-193.

- Watterson, G. A., 1978 The homozygosity test of neutrality. *Genetics* 88:405-417.
- Weiner, M. P., and T. J. Hudson, 2002 Introduction to SNPs: discovery of markers for disease. *BioTechniques* 32:S4-S13.
- *Whitt, S. R., and E. S. Buckler, 2002 Using natural allelic diversity to evaluate gene function. In *Plant Functional Genomics: Methods and Protocols* (Grotewald, E., ed.). Humana Press.
- *Whitt, S. R., L. M. Wilson, M. I. Tenailon, B. S. Gaut and E. S. Buckler, 2002 Genetic diversity and selection in the maize starch pathway. *Proc. Nat. Acad. Sci. USA* 99:12959-12962.
- *Wilson, L., S. Whitt, A. M. Ibanez-Carranza, M. M. Goodman, J. F. Doebley, T. R. Rocheford and E. S. Buckler, Associations for starch content and pasting properties in diverse maize inbreds. (draft for *Plant Cell*).
- Xu, S. Z., 1998 Mapping quantitative trait loci using multiple families of line crosses. *Genetics* 148:517-524.
- Yi, N. J., and S. Z. Xu, 2002 Linkage analysis of quantitative trait loci in multiple line crosses. *Genetica* 114:217-230.
- Zeng, Z.-B., 1993 Theoretical basis of separation of multiple linked gene effects on mapping quantitative trait loci. *Proc. Natl. Acad. Sci. USA* 90:10972-10976.
- Zeng, Z.-B., 1994 Precision mapping of quantitative trait loci. *Genetics* 136:1457-1468.
- *Zhang, L., A. S. Peek, D. Dunams and B. S. Gaut, 2002 Population genetics of duplicated disease-defense genes, *hm1* and *hm2*, in maize (*Z. mays* ssp. *mays* L.) and its wild ancestor (*Z. mays* ssp. *parviglumis*). *Genetics* 162:851-860.