ABSTRACT

Wilson, Larissa Mary. Association mapping of major starch biosynthesis genes in *Zea mays* ssp. *mays*. (Under the direction of Edward S. Buckler, IV and Rebecca S. Boston)

Improving maize yield by utilizing natural allelic diversity is a major objective of today's breeders, and has likely been a goal since maize domestication. Starch is the main component of maize yield, and is an important agronomic trait needed for a wide range of uses from human and animal consumption to ethanol production. The level of starch in the maize kernel is controlled by upwards of 20 different loci, and has been the focus of multiple quantitative trait loci (QTL) studies in order to find regions in the maize genome that affect both starch content and starch quality, like the amylose/amylopectin ratio. The objective of this study was to evaluate the starch biosynthesis pathway using an association mapping approach, by evaluating six starch candidate genes from a diverse set of maize germplasm: *Ae1, Bt2, Sh1, Sh2, Su1,* and *Wx1*.

The six starch candidate genes were amplified, sequenced, and aligned from 29 inbred lines and then evaluated for the level of diversity present. Estimates of π (nucleotide diversity) indicated, on average, starch genes contained 2.3- and 4.8-fold lower amounts of diversity at silent and nonsynonymous sites, respectively, than 20 randomly sampled genes from chromosome one of maize. Three of the starch loci (*Ae1, Bt2,* and *Su1*) had dramatic drops in diversity compared to *Zea mays* ssp. *parviglumis*. Furthermore, Hudson-Kreitman-Aguadé (HKA) tests for selection were significant for these same three loci. In addition,

another test for selection, Tajima's D, was significant at *Ae1*. These data suggest selection on starch genes has lowered diversity in the starch pathway.

Smaller regions throughout each gene were sampled and aligned in a larger set of 97 maize inbreds for association tests. Phenotypic measurements of kernel composition (starch, protein, oil) and viscoamylographic (viscosity, pasting) profiles of starch were used in separate principle component analyses for the association tests. Significant associations ($P \leq 0.05$) with kernel composition traits, while controlling for population structure, were found in *Sh1, Sh2,* and *Bt2*. Significant associations for starch pasting traits were found in *Sh1, Sh2,* and *Ae1*.

Possible phenotypic effects were examined between alleles with significant associations. For kernel composition traits, *Sh1*and *Sh2* showed a general genotype by environment (G × E) effect. In *Bt2,* a nonsynonymous change at residue 22 caused lower variance in oil content. For starch pasting traits, an allele in *Sh1* caused a 1% increase in pasting temperature. At *Ae1,* a nonsynonymous change at residue 58 had a 1.6% higher pasting temperature and 4.6% higher amylose content. A nonsynonymous change at residue 318 in *Sh2* caused a 6% increase in amylose. This study supports previous findings that the *Sh2* locus affects amylose content, but has offered much higher resolution than is possible with traditional linkage mapping, while examining a much broader range of alleles. Therefore, even in a moderately heritable pathway, such as the starch biosynthesis pathway, association methods can be successful in narrowing down regions of effect, most times within 1000 bp.

ASSOCIATION MAPPING OF MAJOR STARCH BIOSYNTHESIS GENES IN *Zea mays* ssp. *mays*

By

Larissa Mary Wilson

A thesis submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Master of Science

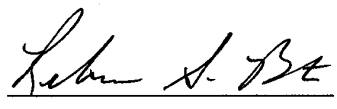
Genetics

Raleigh

2002

Approved by:

Edward S. Buckler, IV
Chair of Advisory Committee



Rebecca S. Boston
Co-chair of Advisory Committee

Michael D. Purugganan

DEDICATION

To my husband, David, and to my parents for their love, support, and understanding.

BIOGRAPHY

Larissa Mary Emenecker Wilson was born on January 18, 1972 in Waukegan, Illinois, the third child out of four girls, but lived most of her childhood in rural Bristol, Wisconsin. Science was always her favorite subject, and when in the seventh grade she dissected her first animal, the typical laboratory frog, she was hooked and knew she wanted to become a biologist, or maybe even a medical doctor.

In 1988 while she was a junior in high school, her family moved south to Smyrna, Tennessee, where she learned a deep appreciation for southern cuisine and much nicer weather. She attended the University of Tennessee at Chattanooga, where in 1998 she received a B.S. in Biology with a minor in chemistry. It was during her days as an undergraduate at UTC when the explosion of molecular biology in the popular news solidified a growing interest in pursuing a career in genetic research. A small research project with Dr. Ann Stapleton introduced her to the merits of maize as a model organism, and reiterated the desire for further education and experience in laboratory research.

After her 1999 marriage to another UTC alumnus, David Berit Wilson, the couple moved to Raleigh, North Carolina where she pursued a Master's degree in Genetics as the first graduate student of Dr. Edward Buckler, IV.

ACKNOWLEDGEMENTS

I wish to express my sincere gratitude to my advisor, Dr. Edward Buckler, IV for his mentorship and guidance throughout my graduate career. I would also like to thank the members of my committee, Dr. Rebecca Boston and Dr. Michael Purugganan for their helpful advice and direction.

The people in the Buckler lab have been a great resource of knowledge, advice, emotional support, (yes, and even fun sometimes!) and I am grateful to all who have been a part of it. In particular I'd like to thank Sherry Whitt, Dr. Jeffry Thornsberry, Brad Rauh, Dr. Sherry Flint-Garcia, Jennifer Heer, and Sandra Andaluz.

Finally, I would like to collectively thank everyone who has personally helped me to become a better scientist.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

adh1	*alcohol dehydrogenase1*
adh2	*alcohol dehydrogenase2*
ADP-glucose	adenosine diphosphate glucose
ae1	*amylose extender1*
AGPase	adenosine diphosphate glucose pyrophosphorylase
amf	*amylose free*
ANOVA	analysis of variance
bp	base pair(s)
bt2	*brittle endosperm2*
c1	*colored aleurone1*
cM	centiMorgan
CRY2	*cryptochrome2*
CSSL	chromosome segment substitution line
C-terminus(terminal)	carboxyl-terminus(terminal)
d3	*dwarf3*
d8	*dwarf8*
DBE	debranching enzyme
DE	Delaware
DNA	deoxyribonucleic acid
DPoll	days to pollen
DSilk	days to silking
EarHt	ear height
EDI	*early day-length insensitive*
FL	Florida
FRI	*FRIGIDA*
fw2.2	*fruit weight2.2*
G x E	Genotype by Environment
GBSSI	granule-bound starch synthase isoform 1
glb1	*globulin1*
Hd1	*Heading date1*
HKA	Hudson-Kreitman-Aguade
hm1	*Helminthosporium carbonum susceptibility1*
hm2	*Helminthosporium carbonum susceptibility2*
id1	*indeterminate1*
IL	Illinois
IN	Indiana
indel	insertion/deletion
kb	kilobase(s)
LD	linkage disequilibrium
Lin5	*Lycopersicon apoplastic invertase5*
MOS	malto-oligosaccharide

mRNA	messenger ribonucleic acid
NC	North Carolina
NIL	nearly-isogenic line
NIR	near infrared
nr	*nitrate reductase*
NSS	non-stiff stalk
N-terminus(terminal)	amino-terminus(terminal)
Pa.S	Pascal second
PC	principle component
PCA	principle component analysis
PCR	polymerase chain reaction
PGO	phenylglyoxal
P_i	inorganic phosphate
PlHt	total plant height
PNAS	*Proceedings of the National Academy of Sciences*
QTL	quantitative trait locus/loci
RIL	recombinant inbred line
SBEI, IIa, IIb	starch branching enzyme isoform I, IIa, or IIb
sh1	*shrunken1*
sh2	*shrunken2*
SNP	single nucleotide polymorphism
SS	stiff stalk
SSI,II,III	starch synthase isoform I,II, or III
SSR	simple sequence repeat
ST	tropical/semitropical
STA	starchless
su1	*sugary1*
sus1	*sucrose synthase1*
tb1	*teosinte branched1*
TDT	transmission/disequilibrium test
te1	*terminal ear1*
UDP-glucose	uridine diphosphate glucose
US	United States
USDA	United States Department of Agriculture
WSP	water-soluble polysaccharide
wx1	*waxy1*

CHAPTER 1

LITERATURE REVIEW

Maize Starch

Starch is the major component in the human diet, for which the Food and Agriculture Organization (FAO) and the World Health Organization (WHO) jointly recommend that 55-75% of daily food intake should come from carbohydrates (http://apps.fao.org). Cereal grains provide a major source of starch; maize, in particular, is the third largest quantity of cereal grains grown worldwide and is second in the U.S. Additionally, maize contains numerous mutants that provide a unique source of specialty starches, e.g. amylose-free waxy starches, important for many commercial applications (Hallauer, 2001), or the sugary mutants responsible for sweet corn production. Currently corn use is up 1% for starch production from the 248 million bushels used in 2000-2001(ERS-USDA, 2002). Today's food industry relies heavily on corn and its modified starches for thickening properties and low-temperature stability in freeze-thaw cycles of convenience foods. Other traditional industrial uses for starch include adhesives and paper, and the possibility of starch-based degradable plastics and super-absorbent polymers are promising (Johnson *et al.*, 2001). Ethanol production from corn produced 1.7 billion gallons in 2001, and corn use for this purpose has increased 9% for this year (ERS-USDA, 2002). Furthermore, new, modified starches that do not require additional chemical modifications are highly desirable (White, 2001). Therefore, continued increases in both yield and starch quality in maize are important goals.

Diversity In Maize

Natural allelic variation within genes can account for the phenotypic differences seen among individuals in a population. Breeders rely on the natural diversity found within crop species on selection for the improvement of a quantitative trait such as yield. Genetic diversity in maize has been studied in several different ways and the general consensus is that maize is highly variable both within and across populations. Sequencing of the *adh1* locus in maize, in *Z. mays* ssp. *parviglumis* (maize progenitor), and in *Zea luxurians*, (distant maize relative), shows that maize retains 77% of the diversity of *parviglumis* and shows more diversity than *Z. luxurians* (Eyre-Walker *et al.*, 1998). For a more general picture of maize diversity, sequencing of random genes across chromosome one of U.S. inbred maize and selected landraces show maize is highly diverse containing a single nucleotide polymorphism (SNP) every 28 base pairs (bp) and is higher than what has been observed in *Drosophila* and humans (Tenaillon *et al.*, 2001). Maize is 3-10 fold more diverse than other domesticated grasses, (reviewed in Buckler *et al.*, 2001).

Why is maize so diverse? Maize appears to have undergone an ancient tetraploidization event, in which duplicated genes allow for mutational events to further diversity in an organism. Colinearity, or shared markers in shared order and an indication of chromosomal duplication, is statistically significant for 60-82% of the maize genome, and roughly a third of the genome may even be triplicated or quadruplicated (Gaut, 2001). Additionally, maize outcrosses at high rates, facilitated by the physical separation of male and female reproductive structures.

Specific diversity studies of genes involved in key biosynthetic pathways like starch production, however, are lacking. In order for breeders to continue making progress in yield,

genetic diversity within such pathways is essential. Only a few landraces have contributed to today's elite U.S. germplasm (Goodman, 1990) and less than 1% of U.S. germplasm contains exotic sources (Hoisington *et al.*, 1999). Commercial maize's vulnerability as a result of its genetic uniformity was demonstrated by the devastating 1970 southern leaf blight epidemic caused by the fungal pathogen, *Bipolaris maydis*. Even though there are over 50,000 accessions present in germplasm collections worldwide, evaluation for useful traits in most of these accessions has not been done, making these "gene morgues" an untapped resource (Hoisington *et al.*, 1999).

In the case of tomato, introgressing alleles from wild relatives has been necessary to increase diversity in this crop (Tanksley and McCouch, 1997). By no means a trivial task, crossing a domesticated crop with a highly resistant wild species for a simple trait such as pathogen resistance can be successful (Valkoun, 2001). For a complex trait such as yield, choosing a wild relative by phenotype to introgress into the domesticated crop may not be the best approach. Suprisingly, a low-yield wild accession may have some beneficial alleles to contribute to high yield in domesticates. The key to progress in yield may be to choose wild accessions that are the most diverged from elite lines (Tanksley and McCouch, 1997). Regardless of the source of diversity of germplasm used in future maize improvement, fine-mapping of quantitative trait loci (QTL) controlling yield is a necessary requirement in order for maize breeders to perform marker assisted selection to ensure continued gains in yield.

Quantitative Trait Loci Dissection

Positional cloning of a QTL

Many important agronomic traits, like starch content, are genetically controlled by multiple loci, each with modest effects on phenotype. In order to locate regions in the maize genome that have an effect on starch levels, several QTL studies were undertaken and have estimated QTL effects on yield traits such as kernel starch, oil or protein content (Goldman *et al.*, 1993; Berke and Rocheford, 1995) and carbohydrate enzyme activity (Prioul *et al.*, 1999; Séne *et al.*, 2000). However, the desirable goal of nucleotide identification of a causal variant underlying these QTL has not yet been performed.

QTL studies alone, while providing the necessary step of identifying regions of effect, have a resolution between 3-20 cM (Falconer and Mackay, 1996). Dissection of a QTL and subsequent determination of causal molecular variants is presently an active area of interest, but may take many years. Once QTL for a particular trait are mapped using molecular markers, a map-based strategy can be used to clone genes underlying the QTL (Yano, 2001). Often, advanced backcross lines are developed, such as nearly isogenic lines (NILs) or chromosome segment substitution lines (CSSLs). Then fine-scale mapping of the QTL is performed in order to further reduce the candidate region for the QTL. So far this approach has found success in plants due to particular characteristics of the model systems, including the ability to produce large numbers of offspring, an amenability to inbreeding allowing a uniform genetic background, and the ease of transformation to verify the identified gene (Remington *et al.*, 2001). Additionally, positional cloning may require substantial time investment as the entire process from development of NILs to the identification and cloning of the gene responsible can easily take many years. This is especially problematic in long

generation plants. Examples of successful identification of genes at QTL occurred in maize, tomato, rice, and *Arabidopsis* and are discussed individually below (Doebley *et al.*, 1995; Frary *et al.*, 2000; Fridman *et al.*, 2000; Yano *et al.*, 2000; El-Din El-Assal *et al.*, 2001).

In tomato, high-resolution mapping of a major fruit weight QTL, *fw2.*2, narrowed down the region to a 150 kb area (Alpert and Tanksley, 1996). Use of the 150 kb QTL in library screenings and genetic complementation analysis located a cosmid containing open reading frames (Frary *et al.*, 2000). Cloning of *fw2.2* revealed one gene responsible for the QTL, *ORFX*, which is responsible for carpel number and is similar in sequence to a human oncogene, c-H-*ras* p21. Sequencing of the 830 nucleotide fragment containing *ORFX* in two different alleles suggests that differences may lie in the upstream regulatory region. Another yield QTL in tomato, *Brix9-2-5,* which is responsible for glucose and fructose levels, was isolated to a 484 bp region by SNP mapping of NIL recombinants (Fridman *et al.*, 2000). An exon and intron of an apoplastic invertase gene, *Lin5,* was found spanning this 484 bp region, and the QTL activity, putatively, is a result of regulation by the untranslated intron, rather than by changes in amino acid sequence. These differences in gene regulation between alleles seen in the case of both *ORFX* and *Lin5* is very similar to the earlier case demonstrated for *tb1* in maize (Wang *et al.*, 1999).

teosinte branched1 (tb1) is a domestication gene underlying a QTL controlling for inflorescence architecture and plays a key role in the difference in morphology between maize and teosinte (Doebley *et al.*, 1995). Mutant *tb1* maize plants resemble teosinte in having unregulated axillary outgrowth and at the base of the plant, prolific tillering. The *tb1* gene was isolated by the *Mutator* transposable element system and cloned (Doebley *et al.*,

1997). Sequence analysis of *tb1* indicates selection during the domestication process was limited to the promoter, rather than the coding region (Wang *et al.*, 1999).

In the dissection of a flowering time QTL called '*early day-length insensitive*' (*EDI*), NILs of *Arabidopsis* were utilized in order to positionally clone *CRY2*, which encodes the blue-light photoreceptor cryptochrome-2 (El-Din El-Assal *et al.*, 2001). *EDI* alone explains 56% of the variation in flowering time between two accessions under short day conditions (Alonso-Blanco *et al.*, 1998). Transformation of a novel allele of *CRY2,* found in a tropical population that shows the EDI phenotype, into the long-day flowering laboratory strain, Landsberg *erecta,* results in earlier flowering under short-day conditions. A single amino acid change is responsible for the light-induced down regulation of CRY2, resulting in early, short-day flowering.

In another example of photoperiod response, positional cloning of a major rice QTL for heading date, *Hd1*, was found to be a homolog of the *CONSTANS* gene from *Arabidopsis* (Yano *et al.*, 2000). *Hd1* encodes a zinc finger domain, and through expression and structural analysis, was found to be allelic to the previously identified *Se1*gene. *Hd1* mRNA levels do not change in response to day length, but are thought to regulate other genes in which day length controls transcription.

Linkage mapping and positional cloning versus association tests

Linkage analysis and the positional cloning of disease genes in humans have been highly successful approaches in Mendelian disorders like Huntington's disease (Gusella *et al.*, 1983; Duyao *et al.*, 1993), where mutations at such loci cause large effects, and

correlations between phenotype and genotype are strong. However, in more complex diseases where the effects are modest, linkage analysis has proven to be less reliable (Risch, 2000). Instead association studies have greater power to detect more modest effects (Risch and Merikangas, 1996). This approach is advantageous in human studies, because finding parent and multiple affected siblings is not necessary. Instead, data from case-control groups from a population are usually easier to obtain.

SNPs are the most common type of marker used in association analyses, and thought to be the largest contributor to phenotypic differences seen in humans. SNPs have an advantage over other markers in being stably inherited, highly abundant in most species, and adaptable for a high-throughput genotyping system, as they are relatively easy to detect (Johnson and Todd, 2000). Ideally, a high-density genome-wide SNP map would allow powerful detection of disease-causing variants. Suggestions of a SNP every 3-6 kb in humans would be required for associations, however this goal is out of reach for most laboratories (Carlson *et al.*, 2001). Instead, a candidate gene approach is the next best solution, involving testing of SNPs in coding regions of possible disease genes or the complete sequencing of one gene and then testing all possible polymorphisms (Johnson and Todd, 2000). Because of the availability of SNPs mapped in the human genome, large haplotype blocks resulting from correlations between two or more neighboring alleles (linkage disequilibrium, LD) could be used to test for association, rather than many individual SNPs. One estimate suggests that human LD in a U.S. population with north-European descent extends 60 kb (Reich *et al.*, 2001).

With the high degree of polymorphism present in maize, SNP identification should be straightforward and useful for association studies in this species (Rafalski, 2002), along

with other biallelic markers such as insertions/deletions, which are abundant within maize genes (Bhattramakki *et al.*, 2002).

Linkage disequilibrium

Having knowledge of the pattern of LD in the species under study is a prerequisite for association mapping. Whether or not an association approach is appropriate over genome-wide scans depends on the level of LD present within the organism of interest by not decaying either too fast or too slow. In a selfing species such as *Arabidopsis*, LD decays within one cM around the *FRI* locus (Nordborg *et al.*, 2002), suggesting genome-wide scans for regions of effect may be possible (Nordborg and Innan, 2002). In traditional maize landraces, LD decays quickly, as seen in 20 random loci across chromosome one (Tenaillon *et al.*, 2001). Expected r^2 values drop 65% in 100 base pairs (bp) and another 26.6% over the next 100 bp. This finding is supported within genes of diverse inbred lines (*d8*, *d3*, *id1*, *sh1*, *su1*, and *tb1*), in which LD decays on average 1.5 kb for $r^2 < 0.1$, however the rate of the decay varies among genes (Remington *et al.*, 2001). When LD between four of the loci is examined (three of the loci are linked on chromosome one), no LD is found. The case may be different for U.S. elite germplasm, where bottlenecks resulting from breeding may increase LD to levels that could be useful for SNP haplotype association analysis, rather than testing individual SNPs (Rafalski, 2002). Generally, the rapid decay of LD seen in maize may allow good resolution of a causal variant if markers are at optimal density.

Controlling for population structure

Association approaches for complex human diseases have been problematic as a result of the confounding effect of population structure. Unless population structure is controlled for, the presence of LD caused by population admixture, genetic drift or selection may result in spurious associations (Lander and Schork, 1994). This was shown in the type 2 diabetes study of two Native American tribes (Knowler *et al.*, 1988), where population admixture of Caucasian alleles resulted in the association of a particular immunoglobulin G haplotype with lower prevalence of diabetes. Once tribal members with Caucasian descent were removed from the analysis, the association became non significant. Recently, a method for using multilocus genotype data to infer population structure caused by admixture or other means has been developed (Pritchard *et al.*, 2000a). Individuals can be assigned probabilistically to a subpopulation, or if admixture is involved as reflected by genotype, can be assigned to two or more subpopulations. With a way to infer population structure, Pritchard *et al*. (2000b) designed a method to test for association in case-control studies (Pritchard *et al.*, 2000b). This method competes with the power of the family based transmission/disequilibrium test (TDT), and controls for population structure.

Plant geneticists to date have rarely used association approaches in part because of the risk of false positives caused by population structure (Pritchard, 2001). By applying the multilocus genotype method of Pritchard *et al.* (2000a), Remington *et al.* (2001) divided a group of 102 taxa of maize into three subpopulations using simple sequence repeats (SSRs) to infer the population structure. These three subpopulations distinguished Iowa Stiff Stalk Synthetics (SS), non-stiff stalks (NSS), and tropicals/semitropicals (ST). An association

study performed in maize found a suite of polymorphisms in the *Dwarf8* gene that associates with flowering time while controlling for population structure (Thornsberry *et al.*, 2001). Thornsberry *et al.* (2001) used a modified association test statistic of the Pritchard method (Pritchard *et al.*, 2000b) for use with a quantitative trait. When population structure is controlled for by using the three assigned subpopulations, many polymorphic sites lose their association but reveal, in particular, a deletion in a conserved domain of the coding region. Plant genetics so far have the benefit of using either linkage mapping or association methods in QTL dissection, but association testing has the advantage of testing multiple candidate genes simultaneously, has good resolution, and can identify a greater range of alleles.

Association studies are most useful when they point to sites or regions within a gene that make biological sense. In order to interpret results from an association study of starch genes, current knowledge of the starch pathway and the biochemistry of the enzymes in maize is useful and necessary. The following section highlights aspects of what is understood and what is still unknown about six major starch genes from maize.

Major Enzymes of Starch Metabolism In Maize

Roughly 70% of maize seed weight is due to starch. Starch for long-term storage is produced in the amyloplasts, which are colorless plastids located in maize endosperm. Starch granules are made up of two types of starch, amylose and amylopectin. Amylose is a mostly linear molecule of $\alpha(1\rightarrow4)$ glucose linkages, while amylopectin has shorter $\alpha(1\rightarrow4)$ linkage segments, attached by $\alpha(1\rightarrow6)$ branch points. Glycogenin is a self-glucosylating protein required for glycogen synthesis in animals, however, a plant counterpart for granule initiation

or a plant starch synthase primer is not known (Bligh, 1999). Formation of growth rings can be seen within the granule, caused by amorphous and semi-crystalline rings, which may be due to the diurnal variation in temperature in the light/dark cycle (Buleon *et al.*, 1997). Within the semi-crystalline ring are successive layers of crystalline and amorphous layers. The crystalline section is made up of amylopectin arranged in a hierarchical pattern, where chain lengths with a low frequency of branches form double helices, which then group together into side chain clusters. Amylose is packaged within the amorphous regions, and constitutes around 27-29% of total starch in maize.

Shrunken1: sucrose synthase

In the maize endosperm, sucrose is converted to UDP-glucose and fructose in a reversible reaction by the major isoform of sucrose synthase, the product of the *Shrunken1* (*Sh1*) gene. Mutant *sh1* kernels contain 8-10% sucrose synthase activity and accumulate 60-78% of starch levels seen in nonmutant kernels (Chourey and Nelson, 1976; Chourey *et al.*, 1998). The UDP-glucose product can enter two pathways, either cellulose biosynthesis or the glycolysis pathway, where it is converted into glucose-1-phosphate, a starch precursor (reviewed by Winter and Huber, 2000). In *sh1* mutants cell degeneration, primarily in the central region of the endosperm and occurring before starch accumulation, causes the collapsed, shrunken phenotype (Chourey *et al.*, 1998).

Both *Sh1* and the constitutive second isoform, *Sucrose synthase1* (*Sus1*) contain a conserved phosphorylation site at serine-15 (Shaw *et al.*, 1994; Huber *et al.*, 1996). This post-translational modification of the sucrose synthase enzyme is reversible and

dephosphorylation increases its hydrophobicity and allows for association with the plasma membrane, while the soluble form is phosphorylated. The phosphorylation event may be a localization signal, as gravity-induced cell elongation in maize increases the amount of plasma-associated, dephosphorylated sucrose synthase and may be involved in cell wall deposition in times of growth (Winter *et al.*, 1997). The soluble, phosphorylated sucrose synthase binds to actin in the cytosol of maize leaves, suggesting this enzyme has an association with the cytoskeleton, but the significance of this interaction has yet to be elucidated (Winter *et al.*, 1998). While *Sus1* is found in all tissues and is highly expressed, *Sh1* is more specialized as it is highly induced in anaerobic conditions and is tissue specific, working mainly in the endosperm (McCarty *et al.*, 1986).

Shrunken2 and *Brittle endosperm2*: ADP-glucose pyrophosphorylase

After the formation of glucose by sucrose synthase, ADP-glucose pyrophosphorylase (AGPase) converts glucose-1-phosphate to ADP-glucose, the substrate for starch synthases (Tsai and Nelson, 1966). AGPase is a heterotetramer encoded by the loci *Shrunken2 (sh2)* and *Brittle endosperm2 (Bt2)* for the large and small subunits, respectively (Bae *et al.*, 1990; Bhave *et al.*, 1990). The subcellular location of AGPase in maize kernels has not been resolved. Denyer *et al*. (Denyer *et al.*, 1996) found most of the AGPase activity in the cytosol when using plastid enrichment preparations, thereby controlling for cytosolic contamination. However, *in situ* detection of enzyme expression and immunolabeling of AGPase protein shows plastidial localization (Brangeon *et al.*, 1997). Four stages of maize development were examined using this approach, as opposed to just one stage done by other

studies; however, resolution may not have been clear enough to determine whether the plastidial location of AGPase was in fact inside the amyloplast or just outside, next to the cytosol. A recent study suggests that while most starch-storing organs have plastidial forms of AGPase, the major AGPase of cereal endosperms is cytosolic (Beckles *et al.*, 2001). Beckles *et al*. (2001) suggest the advantage of cytosolic AGPase activity in graminaceous endosperms allows utilization of carbon for other needs besides starch production if the sucrose supply is low. Alternatively, plastidial activity of AGPase would be committed to starch production only, regardless of the sucrose supply.

AGPase is an allosterically regulated enzyme, activated by 3-phosphoglycerate (3-PGA) and inhibited by inorganic phosphate (P_i) (Stark *et al.*, 1992). Allosteric regulation is important to the leaf enzyme, where fluctuations of photosynthetic substrates from the day/night cycle require an active enzyme in times of need. However, sensitivity to activation of the endosperm AGPase by 3-PGA varies among plant species. In rice, the endosperm AGPase can be activated over 40-fold under 3-PGA saturation to levels normally seen in leaf AGPases, while in barley the endosperm AGPase is catalytically active without 3-PGA (Sikka *et al.*, 2001) (Doan *et al.*, 1999). In maize, activation of AGPase by 3-PGA varies by line. Since some plant endosperm AGPases do not need activation by 3-PGA, allosteric regulation by 3-PGA may be an evolutionary carry over from origins in the leaf enzyme (Greene and Hannah, 1998). Allosteric regulation of AGPase is the focus of many transgenic projects in potato, barley, and rice in order to produce a highly active enzyme (allosterically insensitive) to increase yield (Stark *et al.*, 1992; Doan *et al.*, 1999; Sikka *et al.*, 2001).

The N-terminal and C-terminal region of both subunits have been studied for allosteric regulation and subunit interactions. In potato, the C-terminal regions of both the

large and small subunit of AGPase are essential for proper protein folding or assembly of the enzyme as revealed by deletions (Laughlin *et al.*, 1998). Similarly, mutation of the C-terminus in maize *sh2* causes disruption of proper protein assembly (Greene and Hannah, 1998). Little is known on how the subunit assembly of AGPase is completed, but studies in *sh2* or *bt2* null mutants suggest that SH2:SH2 or BT2:BT2 homodimers do not form.

Allosteric properties are also affected by mutations in either the N- or C-terminus of AGPase. An addition of two amino acids in the C-terminus of the SH2 protein increased seed weight while making the mutant insensitive to P_i (Giroux *et al.*, 1996). Deletions in the N-terminal region of the small subunit of potato show this area is essential for heat stability and normal kinetic properties, like inhibition by P_i, and in the large subunit, N-terminal deletions affect changes in the allosteric regulation (Laughlin *et al.*, 1998).

Waxy1: granule-bound starch synthase

ADP-glucose is either transported into the amyloplast or converted from glucose-1-phosphate by AGPase inside the organelle, and is the substrate for the starch synthases in amylose and amylopectin production. Granule-bound starch synthase I (GBSSI), the product of the *Waxy1 (Wx1)* locus in maize (Shure *et al.*, 1983), is solely responsible for amylose production. Homozygous *wx1* mutants do not accumulate amylose but can still accumulate wildtype levels of amylopectin starch, an indication that amylopectin formation does not proceed from the branching of amylose (Nelson and Rines, 1962). Furthermore a dosage effect can be seen as the number of nonmutant *Wx1* alleles increases in the endosperm (Sprague *et al.*, 1943). Amylose content and GBSSI enzyme activity increase with the

number of wildtype alleles, but the amylose increase is not linear. Amylose content is virtually the same between tetraploid potato plants with two, three, or four wildtype *amf* (the GBSSI loci) alleles (Flipse *et al.*, 1996). After a certain point, increasing the level of GBSSI activity will not increase amylose, as physical limitations are set presumably by the organized structure of amylopectin and the restricted location of the enzyme inside the granule.

Maize *Wx1* and its product have been well characterized molecularly (Nelson, 1968; Echt and Schwartz, 1981; Shure *et al.*, 1983; Klosgen *et al.*, 1986). However, there are still questions about the native primer used by GBSSI for elongation of amylose and the effect GBSSI has on amylopectin. In the monocellular alga *Chlamydomonas reinhardtii*, GBSSI produced by the *STA* locus elongates external amylopectin chains, and subsequent cleavage of the branches occur to form amylose, as supported by pulse-chase experiments (Delrue *et al.*, 1992; van de Wal *et al.*, 1998). In fact, *sta* mutants and *wx* mutants in diploid wheat do show modified amylopectin structure implying some impact of GBSSI on amylopectin, but these effects have not been seen in other higher plants (Delrue *et al.*, 1992; Fujita *et al.*, 2001). Maddelein *et al.* (1994) suggest that GBSSI is responsible for the long glucan backbone within amylopectin. Further support for the role of GBSSI in amylopectin structure is displayed in interactions with starch synthase type III (SSIII) of potato, the major isoform for amylopectin formation. SSIII antisense plants contain deep fissures and multilobed starch granules, while simultaneous GBSSI/SSIII antisense plants have normal granule morphology (Fulton *et al.*, 2002). It is likely that the increased flux of glucose through GBSSI in SSIII antisense plants caused by increases in ADP-glucose is disrupting the starch matrix. Lowering GBSSI in antisense plants relieves this disruption. Because of

this effect on amylopectin structure and its unique role in amylose synthesis, the possibility exists that *wx* mutations affect other starch synthases. However, in barley and maize, elution profiles of starch synthase activities were similar between normal and *wx* mutant starches, excluding GBSSI activity (Hylton *et al.*, 1996). This indicates GBSSI mutations do not affect the activity of other starch synthases.

Unlike *Chlamydomonas*, the model for amylose production by the cleavage of amylopectin is not supported in higher plants such as pea. An alternate model of amylose production is the elongation of linear malto-oligosaccharides (MOS) that can diffuse into the granule and provide nonreducing ends for the action of GBSSI (Denyer *et al.*, 1999). Amylopectin still plays an important role in this model, as increased levels activate GBSSI activity, perhaps by altering the conformation of the enzyme and allowing more efficient elongation of MOS (Denyer *et al.*, 1999). In *Chlamydomonas*, MOS also compete with amylopectin as a primer for amylose synthesis, but the authors remain skeptical about the *in vivo* concentration of MOS to be a significant source of nonreducing ends for GBSSI (van de Wal *et al.*, 1998). It may be that different mechanisms are at work in amylose synthesis between *Chlamydomonas* and higher plants, but the search for the native primer for GBSSI is still under scrutiny.

Amylose extender1: starch branching enzyme IIb

Within most plants there are three types of starch branching enzymes (SBEI, IIa, and IIb), and they are responsible for branching amylose and amylopectin by hydrolyzing the $\alpha(1\rightarrow4)$ bonds of linear chains and reattaching to produce $\alpha(1\rightarrow6)$ branch points. Although

precise functions for the three individual isoforms are still unclear, differences between the I and II groups lie in preferred substrate, tissue specificity, and immunological reactivity (Fisher and Boyer, 1983; Sidebottom *et al.*, 1998). Mutants for SBEI and SBEIIa so far have not been found, but in maize the SBEIIb mutation has been of interest since the observation of high amylose corn (Deatherage *et al.*, 1954). *Amylose-extender (Ae1)* is the locus for SBEIIb in maize (Fisher *et al.*, 1996; Kim *et al.*, 1998), and mutants have 33% more amylose over normal (Fergason *et al.*, 1966). *Ae1* is endosperm and embryo specific (Dang and Boyer, 1989; Kim *et al.*, 1998), and as the dosage of *ae1* mutant alleles increases, SBEIIb enzyme decreases, causing higher amounts of short chain amylose and longer chain lengths of amylopectin (Boyer *et al.*, 1980; Hedman and Boyer, 1982).

Recent work on the branching enzymes has focused on the sequence elements of the gene and conserved areas of the protein in order to elucidate the precise functions of the isoforms. Protein sequence comparisons between SBEI and SBEII indicate significant divergence of these groups of branching enzymes within the first 50 amino acids of the amino terminus and the last 50 amino acids of the carboxyl region, indicating these regions may determine some specificity of their action (Guan *et al.*, 1994). Maize SBEII contains four highly conserved regions among branching and amylolytic enzymes (Kuriki *et al.*, 1996). In particular, three amino acids located within these conserved, carboxyl terminal regions are likely to play an important role in catalysis as suggested by site-directed mutagenesis, highlighting the importance of the C-terminus in enzyme function. These regions may be sites for amylose/amylopectin binding, as arginine residues modified by phenylglyoxal (PGO, an arginine-specific reagent) inactivates SBE. In addition, incubating

SBE with amylose or amylopectin substrate protects SBE against PGO inactivation of arginine (Cao and Preiss, 1996).

In the barley *sbeIIb* gene, an element located in the second intron distinguishes *sbeIIa* from *sbeIIb* and is responsible for the endosperm-specific expression of the gene (Ahlandsberg *et al.*, 2002). This element has similarity to the B-box, an enhancer in potato that is responsible for the tissue-specific expression of patatin, a storage protein. Leaf nuclear proteins contain a repressor that binds to the intronic element suppressing transcription of the *sbeIIb* gene in those tissues. No repressive binding was seen in endosperm nuclear fractions. Although tissue-specific expression of *sbeIIa* and *Ae1* in maize is not yet known, the Ahlandsberg *et al.* (2002) study suggests investigation of either promoter or untranslated regions may be the key for future focus on this question.

Sugary1: starch debranching enzyme

Along with the action of starch branching enzymes in the production of amylopectin are the debranching enzymes (DBE), which hydrolyze the $\alpha(1\rightarrow6)$ branch linkages and allow the crystallization of amylopectin. Two types of DBE are found in maize and differ primarily in their substrate specificities: isoamylases, which hydrolyze amylopectin, glycogen, and phytoglycogen; and pullulanases, which hydrolyze pullulan, which is composed of $\alpha(1\rightarrow6)$ maltotriosyl units (Doehlert and Knutson, 1991). In maize the *Sugary1* (*Su1)* locus is the gene that codes for isoamylase activity (James *et al.*, 1995). In mutant *su1* kernels, the major carbohydrate portion that accumulates is phytoglycogen, a highly branched water-soluble polysaccharide resulting in the sweet corn phenotype (Pan and

Nelson, 1984). The *su1* mutant has pleiotropic effects on other starch biosynthetic enzymes, for example an increase is seen in AGPase, a likely result of the higher sucrose levels in these mutants. Alternatively, *su1* mutants show a reduction in the activity of the other DBE, pullulanase (Pan and Nelson, 1984; Doehlert *et al.*, 1993). Reduction of pullulanase in *sugary* mutants may occur post-translationally, since normal levels of pullulanase transcript are seen (Beatty *et al.*, 1999; Kubo *et al.*, 1999). The reduction in pullulanase in isoamylase mutants may only be characteristic of storage starches and not transient starches, as no effect on pullulanase is seen in *Chlamydomonas* or *Arabidopsis* leaves (Zeeman *et al.*, 1998; Dauvillee *et al.*, 2000).

DBEs were thought primarily to play a role in the germination of seeds in the break down of starch into useable soluble sugars for the developing seedling, along with other hydrolytic enzymes. Indeed, mutant *su1* seedlings show lower viability (Martins and da Silva, 1998; Revilla *et al.*, 2000) in support of an involvement in germination; however, the main role of DBEs is most likely in amylopectin synthesis, when transcript amounts of isoamylases are highest in the developing endosperms of maize, rice, barley, and wheat (Rahman *et al.*, 1998; Kubo *et al.*, 1999; Sun *et al.*, 1999; Genschel *et al.*, 2002).

The role of isoamylases and pullulanases in the formation of amylopectin is not known. Erlander (1958) suggested that the production of amylopectin occurs through the debranching of a phytoglycogen intermediate by DBEs (Erlander, 1958). This model has been criticized in part as debranching of phytoglycogen will not result in the higher A:B chain ratio necessary to obtain amylopectin (Marshall and Whelan, 1974). A-chains of amylopectin are the outer, unbranched chains, while B-chains are inner and branched. Although a highly branched complex, phytoglycogen has a low A:B chain ratio of 0.9:1, as

opposed to amylopectin which has a higher ratio of 1.5-2.6:1. The theory of an amylopectin precursor has recently been revived, as molecular knowledge of DBE function increases. Ball *et al*. (Ball *et al.*, 1996) have suggested that amylopectin structure depends on the discontinuous action of both SBEs and DBEs. This glucan-trimming model starts with a unstructured, preamylopectin molecule overly branched by SBEs, until the function of DBEs produce the highly structured, crystallized form of amylopectin. The linear oligosaccharides produced by the debranching activity are recycled by starch synthases to further elongate nonreducing ends until the minimum chain length is reached for SBE activity, and the cycle is repeated.

A second model for amylopectin synthesis describes a less direct role for DBEs in starch synthesis, called the water-soluble polysaccharide (WSP)-clearing model. *dbe1* mutants in *Arabidopsis* are defective in isoamylase production, but still accumulate small amounts of normal starch granules plus phytoglycogen, a result contrary to the glucan-trimming model (Zeeman *et al.*, 1998). In order to explain these results, Zeeman and colleagues propose the action of DBEs in the stroma is to degrade WSPs, preventing the SSs and SBEs at the starch granule surface from elongating both WSPs and amylopectin. In DBE mutants, elongation and branching of WSPs out compete amylopectin formation and produce the byproduct phytoglycogen.

Recent work on the molecular characterization of isoamylases in plants and bacteria may aid in resolving the controversy surrounding the role of DBEs in amylopectin synthesis. Essential residues for catalytic activity of isoamylases have been found in conserved regions of the amylase family (Abe *et al.*, 1999). Specific to the maize *su1* mutant, the cause of phytoglycogen accumulation in sweet corn has been narrowed down to a couple of likely

amino acid changes (Dinges *et al.*, 2001). Unfortunately, isoamylases from just a few organisms have been isolated, as it has been difficult to measure activity from crude extracts (Zeeman *et al.*, 1998). It is essential that more isoamylases from different species are cloned and characterized in order to elucidate the mechanism by which isoamylases contribute to amylopectin formation.

Objectives Of The Present Study

The objectives of this research were:

a) to determine the level of diversity present within starch candidate genes in a large set of maize inbred lines, thereby assessing the amount and focus of selection in genes involved in starch biosynthesis.

b) to identify the molecular basis for the quantitative starch variation seen between maize inbred lines using an association approach.

c) to produce highly reliable markers for breeding purposes.

Six major candidate genes of the maize starch metabolism pathway were evaluated in order to meet these objectives: *Ae1, Bt2, Sh1, Sh2, Su1,* and *Wx1*.

The remainder of this thesis contains two chapters and an appendix. Chapter 2 is titled: "Diversity and selection in the maize starch pathway" written by Whitt and Wilson *et al.* (2002) and originally published in *Proceedings of the National Academy of Sciences* Volume 99, pages 12959-12962. Chapter 3 is titled: "Associations with kernel composition and starch pasting properties in six major genes involved in starch biosynthesis in maize."

The appendix includes data from both *Sh1* and *su1* and describes the level of LD in selected maize genes and the suitability of association studies in maize:

Remington, DL, JM Thornsberry, Y Matsuoka, LM Wilson, SR Whitt, J Doebley, S Kresovich, MM Goodman, and ES Buckler IV. 2001. "Structure of linkage disequilibrium and phenotypic associations in the maize genome." *PNAS*. 98:11479-11484.

Literature Cited

Abe, J., C. Ushijima and S. Hizukuri (1999). "Expression of the isoamylase gene of Flavobacterium odoratum KU in Escherichia coli and identification of essential residues of the enzyme by site-directed mutagenesis." *Applied and Environmental Microbiology* **65**(9): 4163-4170.

Ahlandsberg, S., C. Sun and C. Jannson (2002). "An intronic element directs endosperm-specific expression of the *sbeIIb* gene during barley seed development." *Plant Cell Reports* **20**: 864-868.

Alonso-Blanco, C., S. El-Din El-Assal, G. Coupland and M. Koornneef (1998). "Analysis of Natural Allelic Variation at Flowering Time Loci in the Landsberg *erecta* and Cape Verde Islands Ecotypes of *Arabidopsis thaliana*." *Genetics* **149**: 749-764.

Alpert, K. B. and S. Tanksley (1996). "High-resolution mapping and isolation of a yeast artificial chromosome contig containing *fw2.2*: A major fruit weight quantitative trait locus in tomato." *Proceedings of the National Academy of Sciences of the United States of America* **93**: 15503-15507.

Bae, J. M., M. J. Giroux and L. C. Hannah (1990). "Cloning and characterization of the *Brittle-2* gene of maize." *Maydica* **35**: 317-322.

Ball, S., H. P. Guan, M. G. James, A. M. Myers, P. L. Keeling, G. Mouille, A. Buleon, P. Colonna and J. Preiss (1996). "From glycogen to amylopectin: A model for the biogenesis of the plant starch granule." *Cell* **86**: 349-352.

Beatty, M. K., A. Rahman, H. Cao, W. Woodman, M. Lee, A. M. Myers and M. G. James (1999). "Purification and Molecular Genetic Characterization of ZPU1, a Pullulanase-Type Starch-Debranching Enzyme from Maize." *Plant Physiology* **119**: 255-266.

Beckles, D. M., A. M. Smith and T. ap Rees (2001). "A cytosolic ADP-glucose pyrophosphorylase is a feature of graminaceous endosperms, but not of other starch-storing organs." *Plant Physiology* **125**(2): 818-827.

Berke, T. G. and T. Rocheford (1995). "Quantitative Trait Loci for Flowering, Plant and Ear Height, and Kernel Traits in Maize." *Crop Science* **35**: 1542-1549.

Bhattramakki, D., M. Dolan, M. Hanafey, R. Wineland, D. Vaske, J. C. Register, S. V. Tingey and A. Rafalski (2002). "Insertion-deletion polymorphisms in 3' regions of maize genes occur frequently and can be used as highly informative genetic markers." *Plant Molecular Biology* **48**: 539-547.

Bhave, M. R., S. Lawrence, C. Barton and L. C. Hannah (1990). "Identification and molecular characterization of *Shrunken-2* cDNA clones of maize." *Plant Cell* **2**: 581-588.

Bligh, H. F. J. (1999). Genetic manipulation of starch biosynthesis: Progress and potential. *Biotechnology and Genetic Engineering Reviews, Vol 16*. **16:** 177-201.

Boyer, C. D., P. A. Damewood and G. L. Matters (1980). "Effect of Gene Dosage at High Amylose Loci on the Properties of the Amylopectin Fractions of the Starches." *Starke* **32**(7): 217-222.

Brangeon, J., A. Reyss and J. L. Prioul (1997). "In situ detection of ADPglucose pyrophosphorylase expression during maize endosperm development." *Plant Physiology and Biochemistry* **35**(11): 847-858.

Buckler, E. S., J. M. Thornsberry and S. Kresovitch (2001). "Molecular diversity, structure and domestication of grasses." *Genetical Research* **77**: 213-218.

Buleon, A., D. J. Gallant, B. Bouchet, G. Mouille, C. D'Hulst, J. Kossmann and S. Ball (1997). "Starches from A to C (*Chlamydomonas reinhardtii* as a Model Microbial System to Investigate the Biosynthesis of the Plant Amylopectin Crystal)." *Plant Physiology* **115**: 949-957.

Cao, H. and J. Preiss (1996). "Evidence for essential arginine residues at the active sites of maize branching enzymes." *Journal of Protein Chemistry* **15**(3): 291-304.

Carlson, C. S., T. L. Newman and D. A. Nickerson (2001). "SNPing in the human genome." *Current Opinion in Chemical Biology* **5**: 78-85.

Chourey, P., E. W. Taliercio, S. J. Carlson and Y. L. Ruan (1998). "Genetic evidence that the two isozymes of sucrose synthase present in developing maize endosperm are critical, one for cell wall integrity and the other for starch biosynthesis." *Molecular and General Genetics* **259**: 88-96.

Chourey, P. S. and O. E. Nelson (1976). "Enzymatic Deficiency Conditioned by Shrunken 1 Mutations in Maize." *Biochemical Genetics* **14**(11-1): 1041-1055.

Dang, P. L. and C. D. Boyer (1989). "Comparison of soluble starch synthases and branching enzymes from leaves and kernels of normal and *amylose-extender* maize." *Biochemical Genetics* **27**(9/10): 521-532.

Dauvillee, D., V. Mestre, C. Colleoni, M. C. Slomianny, G. Mouille, B. Delrue, C. d'Hulst, C. Bliard, J. M. Nuzillard and S. Ball (2000). "The debranching enzyme complex missing in glycogen accumulating mutants of Chlamydomonas reinhardtii displays an isoamylase- type specificity." *Plant Science* **157**(2): 145-156.

Deatherage, W. L., M. M. Macmasters, M. L. Vineyard and R. P. Bear (1954). "A Note on Starch of High Amylose Content from Corn with High Starch Content." *Cereal Chemistry* **31**(1): 50-53.

Delrue, B., T. Fontaine, F. Routier, A. Decq, J.-M. Wieruszesi, N. Van den Koornhuyse, M. L. Maddelein, B. Fournet and S. Ball (1992). "Waxy *Chlamydomonas reinhardtii*: Monocellular algal mutants defective in amylose biosynthesis and granule-bound starch synthase activity accumulate a structurally modified amylopectin." *Journal of Bacteriology* **174**(11): 3612-3620.

Denyer, K., F. Dunlap, T. Thorbjornsen, P. L. Keeling and A. M. Smith (1996). "The major form of ADP-glucose pyrophosphorylase in maize endosperm is extra-plastidial." *Plant Physiology* **112**: 779-785.

Denyer, K., D. Waite, A. Edwards, C. Martin and A. M. Smith (1999). "Interaction with amylopectin influences the ability of granule-bound starch synthase I to elongate malto-oligosaccharides." *Biochemical Journal* **342**: 647-653.

Denyer, K., D. Waite, S. Motawia, B. L. Moller and A. M. Smith (1999). "Granule-bound starch synthase I in isolated starch granules elongates malto-oligosaccharides processively." *Biochemical Journal* **340**: 183-191.

Dinges, J. R., C. Colleoni, A. M. Myers and M. G. James (2001). "Molecular Structure of Three Mutations at the Maize *sugary1* Locus and Their Allele-Specific Phenotypic Effects." *Plant Physiology* **125**: 1406-1418.

Doan, D. N. P., H. Rudi and O. A. Olsen (1999). "The allosterically unregulated isoform of ADP-glucose pyrophosphorylase from barley endosperm is the most likely source of ADP-glucose incorporated into endosperm starch." *Plant Physiology* **121**(3): 965-975.

Doebley, J. F., 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. F., A. Stec and L. Hubbard (1997). "The evolution of apical dominance in maize." *Nature* **386**: 485-488.

Doehlert, D. C. and C. A. Knutson (1991). "2 Classes of Starch Debranching Enzymes from Developing Maize Kernels." *Journal of Plant Physiology* **138**(5): 566-572.

Doehlert, D. C., T. M. Kuo, J. A. Juvik, E. P. Beers and S. H. Duke (1993). "Characteristics of Carbohydrate-Metabolism in Sweet Corn (Sugary-1) Endosperms." *Journal of the American Society for Horticultural Science* **118**(5): 661-666.

Duyao, M., C. Ambrose, R. Myers, A. Novelletto, F. Persichetti, M. Frontali, S. Folstein, C. Ross, M. Franz, M. Abbott, J. Gray, P. Conneally, A. Young, J. Penney, Z. Hollingsworth, I. Shoulson, A. Lazzarini, A. Falek, W. Koroshetz, D. Sax, E. Bird, J. Vonsattel, E. Bonilla, J. Alvir, J. B. Conde, J. H. Cha, L. Dure, F. Gomez, M. Ramos, J. Sanchezramos, S. Snodgrass, M. Deyoung, N. Wexler, C. Moscowitz, G. Penchaszadeh, H. Macfarlane, M. Anderson, B. Jenkins, J. Srinidhi, G. Barnes, J. Gusella and M. Macdonald (1993). "Trinucleotide Repeat Length Instability and Age-of-Onset in Huntingtons-Disease." *Nature Genetics* **4**(4): 387-392.

Echt, C. S. and D. Schwartz (1981). "Evidence for the inclusion of controlling elements within the structural gene at the waxy locus in maize." *Genetics* **99**: 275-284.

El-Din El-Assal, S., C. Alonso-Blanco, A. J. M. Peeters, V. Raz and M. Koornneef (2001). "A QTL for flowering time in *Arabidopsis* reveals a novel allele of *CRY2*." *Nature Genetics* **29**: 435-440.

Erlander, S. R. (1958). "A proposed mechanism for the synthesis of starch from glycogen." *Enzymologia* **19**(5): 273-283.

ERS-USDA (2002). Feed Situation and Outlook Yearbook. Washington, DC, Economic Research Service, United States Department of Agriculture.

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." *Proceedings of the National Academy of Sciences of the United States of America* **95**: 4441-4446.

Falconer, D. S. and T. F. C. Mackay (1996). *Introduction to Quantitative Genetics*.

Fergason, V. L., J. L. Helm and M. S. Zuber (1966). "Gene Dosage Effects at Ae Locus on Amylose Content of Corn Endosperm." *Journal of Heredity* **57**(3): 90-&.

Fisher, D. K., M. Gao, K. N. Kim, C. D. Boyer and M. J. Guiltinan (1996). "Allelic analysis of the maize amylose-extender locus suggests that independent genes encode starch-branching enzymes LLa and LLb." *Plant Physiology* **110**(2): 611-619.

Fisher, M. B. and C. D. Boyer (1983). "Immunological characterization of maize starch branching enzymes." *Plant Physiology* **72**: 813-816.

Flipse, E., C. J. A. M. Keetels, E. Jacobsen and R. Visser (1996). "The dosage effect of the wildtype GBSS allele is linear for GBSS activity but not for amylose content: absence of amylose has a distinct influence on the physico-chemical properties of starch." *Theoretical and Applied Genetics* **92**: 121-127.

Frary, A., T. C. Nesbitt, A. Frary, S. Grandillo, E. van der Knaap, K. B. Cong, J. Liu, J. Meller, R. Elber, K. B. Alpert and S. Tanksley (2000). "*fw2.2*: A Quantitative Trait Locus Key to the Evolution of Tomato Fruit Size." *Science* **289**: 85-88.

Fridman, E., T. Pleban and D. Zamir (2000). "A recombination hotspot delimits a wild-species quantitative trait locus for tomato sugar content to 484 bp within an invertase gene." *Proceedings of the National Academy of Sciences of the United States of America* **97**(9): 4718-4723.

Fujita, N., H. Hasegawa and T. Taira (2001). "The isolation and characterization of a *waxy* mutant of diploid wheat (*Triticum monococcum* L.)." *Plant Science* **160**: 595-602.

Fulton, D. C., A. Edwards, E. Pilling, H. L. Robinson, B. Fahy, R. Seale, L. Kato, A. M. Donald, P. Geigenberger, C. Martin and A. M. Smith (2002). "Role of granule-bound starch synthase in determination of amylopectin structure and starch granule morphology in potato." *The Journal of Biological Chemistry* **277**(13): 10834-10841.

Gaut, B. S. (2001). "Patterns of chromosomal duplication in maize and their implications for comparative maps of the grasses." *Genome Research* **11**: 55-66.

Genschel, U., G. Abel, H. Lorz and S. Lutticke (2002). "The sugary-type isoamylase in wheat: tissue distribution and subcellular localisation." *Planta* **214**(5): 813-820.

Giroux, M. J., J. R. Shaw, G. F. Barry, B. G. Cobb, T. W. Greene, T. W. Okita and L. C. Hannah (1996). "A single gene mutation that increases maize seed weight." *Proceedings of the National Academy of Sciences of the United States of America* **93**: 5824-5829.

Goldman, I. L., T. Rocheford and J. W. Dudley (1993). "Quantitative trait loci influencing protein and starch concentration in the Illinois Long Term Selection maize strains." *Theoretical and Applied Genetics* **87**: 217-224.

Goodman, M. M. (1990). "Genetic and Germ Plasm Stocks Worth Conserving." *Journal of Heredity* **81**: 11-16.

Greene, T. W. and L. C. Hannah (1998). "Adenosine diphosphate glucose pyrophosphorylase, a rate- limiting step in starch biosynthesis." *Physiologia Plantarum* **103**(4): 574-580.

Greene, T. W. and L. C. Hannah (1998). "Maize endosperm ADP-glucose pyrophosphorylase SHRUNKEN2 and BRITTLE2 subunit interactions." *Plant Cell* **10**(8): 1295-1306.

Guan, H. P., T. Baba and J. Preiss (1994). "Expression of branching enzyme II of maize endosperm in *Escherichia coli*." *Cellular and Molecular Biology* **40**(7): 981-988.

Gusella, J. F., N. S. Wexler, P. M. Conneally, S. L. Naylor, M. A. Anderson, R. E. Tanzi, P. C. Watkins, K. Ottina, M. R. Wallace, A. Y. Sakaguchi, A. B. Young, I. Shoulson, E. Bonilla and J. B. Martin (1983). "A Polymorphic DNA Marker Genetically Linked to Huntingtons- Disease." *Nature* **306**(5940): 234-238.

Hallauer, A. R., Ed. (2001). Specialty Corns. Boca Raton, CRC Press LLC.

Hedman, K. D. and C. D. Boyer (1982). "Gene dosage at the *amylose-extender* locus of maize: Effects on the levels of starch branching enzymes." *Biochemical Genetics* **20**(5/6): 483-492.

Hoisington, D., M. Khairallah, T. Reeves, J. V. Ribaut, B. Skovmand, S. Taba and M. Warburton (1999). "Plant genetic resources: What can they contribute toward increased crop productivity?" *Proceedings of the National Academy of Sciences of the United States of America* **96**(11): 5937-5943.

Huber, S. C., J. L. Huber, P.-C. Liao, D. A. Gage, R. W. McMichael, Jr., P. Chourey, L. C. Hannah and K. Koch (1996). "Phosphorylation of serine-15 of maize leaf sucrose synthase." *Plant Physiology* **112**: 793-802.

Hylton, C. M., K. Denyer, P. L. Keeling, M.-T. Chang and A. M. Smith (1996). "The effect of *waxy* mutations on the granule-bound starch synthases of barley and maize endosperms." *Planta* **198**: 230-237.

James, M. G., D. S. Robertson and A. M. Myers (1995). "Characterization of the Maize Gene Sugary1, a Determinant of Starch Composition in Kernels." *Plant Cell* **7**(4): 417-429.

Johnson, G. C. L. and J. A. Todd (2000). "Strategies in complex disease mapping." *Current Opinion in Genetics and Development* **10**: 330-334.

Johnson, L. A., C. L. Hardy, C. P. Baumel and P. J. White (2001). "Identifying valuable corn quality traits for starch production." *Cereal Foods World* **46**(9): 417-423.

Kim, K. N., D. K. Fisher, M. Gao and M. J. Guiltinan (1998). "Molecular cloning and characterization of the amylose-extender gene encoding starch branching enzyme IIB in maize." *Plant Molecular Biology* **38**(6): 945-956.

Klosgen, R. B., A. Gierl, Z. Schwarzsommer and H. Saedler (1986). "Molecular Analysis of the Waxy Locus of Zea-Mays." *Molecular & General Genetics* **203**(2): 237-244.

Knowler, W. C., R. C. Williams, D. J. Pettitt and A. G. Steinberg (1988). "$Gm^{3;5,13,14}$ and Type 2 Diabetes Mellitus: An Association in American Indians with Genetic Admixture." *American Journal of Human Genetics* **43**: 520-526.

Kubo, A., N. Fujita, K. Harada, T. Matsuda, H. Satoh and Y. Nakamura (1999). "The starch-debranching enzymes isoamylase and pullulanase are both involved in amylopectin biosynthesis in rice endosperm." *Plant Physiology* **121**(2): 399-409.

Kuriki, T., H. P. Guan, M. Sivak and J. Preiss (1996). "Analysis of the active center of branching enzyme II from maize endosperm." *Journal of Protein Chemistry* **15**(3): 305-313.

Lander, E. S. and N. J. Schork (1994). "Genetic Dissection of Complex Traits." *Science* **265**: 2037-2048.

Laughlin, M. J., S. E. Chantler and T. W. Okita (1998). "N- and C-terminal peptide sequences are essential for enzyme assembly, allosteric, and/or catalytic properties of ADP- glucose pyrophosphorylase." *Plant Journal* **14**(2): 159-168.

Maddelein, M. L., N. Libessart, F. Bellanger, B. Delrue, C. D'Hulst, N. Van den Koornhuyse, T. Fontaine, J.-M. Wieruszesi, A. Decq and S. Ball (1994). "Toward an understanding of the biogenesis of the starch granule." *The Journal of Biological Chemistry* **269**(40): 25150-25157.

Marshall, J. J. and W. J. Whelan (1974). "Multiple branching in glycogen and amylopectin." *Archives of Biochemistry and Biophysics* **161**: 234-238.

Martins, M. and W. J. da Silva (1998). "Genic and genotypic frequencies of endosperm mutants in maize populations under natural selection." *Journal of Heredity* **89**(6): 516-524.

McCarty, D. R., J. R. Shaw and L. C. Hannah (1986). "The cloning, genetic mapping, and expression of the constitutive sucrose synthase locus of maize." *Proceedings of the National Academy of Sciences of the United States of America* **83**: 9099-9103.

Nelson, O. (1968). "The *waxy* locus in maize. II. The location of the controlling element alleles." *Genetics* **60**: 507-524.

Nelson, O. and H. W. Rines (1962). "The enzymatic deficiency in the waxy mutant of maize." *Biochemical and Biophysical Research Communications* **9**(4): 297-300.

Nordborg, M., J. O. Borevitz, J. Bergelson, C. C. Berry, J. Chory, J. Hagenblad, M. Kreitman, J. N. Maloof, T. Noyes, P. J. Oefner, E. A. Stahl and D. Weigal (2002). "The extent of linkage disequilibrium in *Arabidopsis thaliana*." *Nature Genetics* **30**: 190-193.

Nordborg, M. and H. Innan (2002). "Molecular population genetics." *Current Opinion in Plant Biology* **5**: 69-73.

Pan, D. and O. E. Nelson (1984). "A Debranching Enzyme Deficiency in Endosperms of the Sugary-1 Mutants of Maize." *Plant Physiology* **74**(2): 324-328.

Prioul, J. L., S. Pelleschi, M. Sene, C. Thevenot, M. Causse, D. de Vienne and A. Leonardi (1999). "From QTLs for enzyme activity to candidate genes in maize." *Journal of Experimental Botany* **50**(337): 1281-1288.

Pritchard, J. K. (2001). "Deconstructing maize population structure." *Nature Genetics* **28**: 203-204.

Pritchard, J. K., M. Stephens and P. Donnelly (2000a). "Inference of population structure using multilocus genotype data." *Genetics* **155**: 945-959.

Pritchard, J. K., M. Stephens, N. A. Rosenberg and P. Donnelly (2000b). "Association mapping in structured populations." *American Journal of Human Genetics* **67**: 170-181.

Rafalski, A. (2002). "Applications of single nucleotide polymorphisms in crop genetics." *Current Opinion in Plant Biology* **5**: 94-100.

Rahman, A., K. S. Wong, J. L. Jane, A. M. Myers and M. G. James (1998). "Characterization of SU1 isoamylase, a determinant of storage starch structure in maize." *Plant Physiology* **117**(2): 425-435.

Reich, D. E., M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti, D. J. Richter, T. Lavery, R. Kouyoumjian, S. F. Farhadian, R. Ward and E. S. Lander (2001). "Linkage disequilibrium in the human genome." *Nature* **411**: 199-204.

Remington, D. L., J. M. Thornsberry, Y. Matsuoka, L. M. Wilson, S. R. Whitt, J. F. Doebley, S. Kresovitch, M. M. Goodman and E. S. Buckler, IV (2001). "Structure of linkage disequilibrium and phenotypic associations in the maize genome." *Proceedings of the National Academy of Sciences of the United States of America* **98**(20): 11479-11484.

Remington, D. L., M. C. Ungerer and M. D. Purugganan (2001). "Map-based cloning of quantitative trait loci: progress and prospects." *Genetical Research* **78**(3): 213-218.

Revilla, P., R. A. Malvar, M. C. Abuin, B. Ordas, P. Soengas and A. Ordas (2000). "Genetic background effect on germination of SU1 maize and viability of the SU1 allele." *Maydica* **45**(2): 109-111.

Risch, N. (2000). "Searching for genetic determinants in the new millennium." *Nature* **405**: 847-856.

Risch, N. and K. Merikangas (1996). "The future of genetic studies of complex human diseases." *Science* **273**: 1516-1517.

Séne, M., M. Causse, C. Damerval, C. Thevenot and J. L. Prioul (2000). "Quantitative trait loci affecting amylose, amylopectin and starch content in maize recombinant inbred lines." *Plant Physiology and Biochemistry* **38**(6): 459-472.

Shaw, J. R., R. J. Ferl, J. Baier, D. Stclair, C. Carson, D. R. McCarty and L. C. Hannah (1994). "Structural Features of the Maize Sus1 Gene and Protein." *Plant Physiology* **106**(4): 1659-1665.

Shure, M., S. Wessler and N. Fedoroff (1983). "Molecular-Identification and Isolation of the Waxy Locus in Maize." *Cell* **35**(1): 225-233.

Sidebottom, C., M. Kirkland, B. Strongitharm and R. Jeffcoat (1998). "Characterization of the difference of starch branching enzyme activities in normal and low-amylopectin maize during kernel development." *Journal of Cereal Science* **27**(3): 279-287.

Sikka, V. K., S. B. Choi, I. H. Kavakli, C. Sakulsingharoj, S. Gupta, H. Ito and T. W. Okita (2001). "Subcellular compartmentation and allosteric regulation of the rice endosperm ADPglucose pyrophosphorylase." *Plant Science* **161**(3): 461-468.

Sprague, G. F., B. Brimhall and R. M. Hixon (1943). "Some effects of the waxy gene in corn on properties of the endosperm starch." *Journal of the American Society of Agronomy* **35**: 817-822.

Stark, D. M., K. P. Timmerman, G. F. Barry, J. Preiss and G. M. Kishore (1992). "Regulation of the amount of starch in plant tissues by ADP glucose pyrophosphorylase." *Science* **258**: 287-292.

Sun, C. X., P. Sathish, S. Ahlandsberg and C. Jansson (1999). "Analyses of isoamylase gene activity in wild-type barley indicate its involvement in starch synthesis." *Plant Molecular Biology* **40**(3): 431-443.

Tanksley, S. and S. R. McCouch (1997). "Seed banks and molecular maps: unlocking genetic potential from the wild." *Science* **277**: 1063-1066.

Tenaillon, M. I., M. C. Sawkins, A. D. Long, R. L. Gaut, J. F. Doebley and B. S. Gaut (2001). "Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.)." *Proceedings of the National Academy of Sciences of the United States of America* **98**(16): 9161-9166.

Thornsberry, J. M., M. M. Goodman, J. F. Doebley, S. Kresovitch, D. Nielson and E. S. Buckler, IV (2001). "*Dwarf8* polymorphisms associate with variation in flowering time." *Nature Genetics* **28**: 286-289.

Tsai, C. Y. and O. Nelson (1966). "Starch-deficient maize mutant lacking adenosine diphosphate glucose pyrophosphorylase activity." *Science* **151**: 341-343.

Valkoun, J. J. (2001). "Wheat pre-breeding using wild progenitors." *Euphytica* **119**(1-2): 17-23.

van de Wal, M., C. D'Hulst, J.-P. Vincken, A. Buleon, R. Visser and S. Ball (1998). "Amylose is synthesized *in vitro* by extension of and cleavage from amylopectin." *The Journal of Biological Chemistry* **273**(35): 22232-22240.

Wang, R.-L., A. Stec, J. Hey, L. Lukens and J. F. Doebley (1999). "The limits of selection during maize domestication." *Nature* **398**: 236-239.

White, P. J. (2001). Properties of Corn Starch. *Specialty Starches*. A. R. Hallauer. Boca Raton, CRC Press LLC**:** 33-62.

Winter, H., J. L. Huber and S. C. Huber (1997). "Membrane association of sucrose synthase: changes during the graviresponse and possible control by protein phosphorylation." *FEBS Letters* **420**: 151-155.

Winter, H., J. L. Huber and S. C. Huber (1998). "Identification of sucrose synthase as an actin-binding protein." *Febs Letters* **430**: 205-208.

Winter, H. and S. C. Huber (2000). "Regulation of sucrose metabolism in higher plants: Localization and regulation of activity of key enzymes." *Critical Reviews in Biochemistry and Molecular Biology* **35**(4): 253-289.

Yano, M. (2001). "Genetic and molecular dissection of naturally occurring variation." *Current Opinion in Plant Biology* **4**: 130-135.

Yano, M., Y. Katayose, M. Ashikari, U. Yamanouchi, L. Monna, T. Fuse, T. Baba, K. Yamamoto, Y. Umehara, Y. Nagamura and T. Sasaki (2000). "*Hd1*, a Major Photoperiod Sensitivity Quantitative Trait Locus in Rice, Is Closely Related to the Arabidopsis Flowering Time Gene *CONSTANS*." *Plant Cell* **12**: 2473-2483.

Zeeman, S. C., T. Umemoto, W. L. Lue, P. Au-Yeung, C. Martin, A. M. Smith and J. Chen (1998). "A mutant of Arabidopsis lacking a chloroplastic isoamylase accumulates both starch and phytoglycogen." *Plant Cell* **10**(10): 1699-1711.

CHAPTER 2

GENETIC DIVERSITY AND SELECTION IN THE MAIZE STARCH PATHWAY

Note: This work was funded in part by federal funds from the United States Department of Agriculture and is exempt from copyright law (Title 17, Chapter 1, Section 105, concerning public domain). "Genetic Diversity and Selection in the Maize Starch Pathway" by Whitt and Wilson, *et al.* was originally published in *Proceedings of the National Academy of Sciences*. October 1, 2002. Volume 99(20): pages 12959-12962.

Genetic Diversity And Selection In The Maize Starch Pathway

Authors: Sherry R. Whitt[*§], Larissa M. Wilson[†§], Maud I. Tenaillon[‡], Brandon S. Gaut[‡], and Edward S. Buckler IV[*†]

§ S.R.W and L.M.W. contributed equally to the work.

* USDA/ARS, Raleigh, NC 27695-7614

† Department of Genetics, North Carolina State University, Raleigh, NC 27695-7614

‡ Dept. Ecology and Evolutionary Biology, University of California, Irvine, CA 92697-2525

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. AY146786-AY146817).

Abstract

Maize is both phenotypically and genetically diverse. Sequence studies generally confirm the extensive genetic variability in modern maize is consistent with a lack of selection. For more than 6000 years, Native Americans and modern breeders have exploited the tremendous genetic diversity of maize (*Zea mays* ssp. *mays*) to create the highest yielding grain crop in the world. Nonetheless, some loci have relatively low levels of genetic variation, particularly loci that have been the target of artificial selection, like *c1* and *tb1*. However, there is limited information on how selection may affect an agronomically important pathway for any crop. These pathways may retain the signature of artificial selection and may lack genetic variation in contrast to the rest of the genome. To evaluate the impact of selection across an agronomically important pathway, we surveyed nucleotide diversity at six major genes involved in starch metabolism and found unusually low genetic diversity and strong evidence of selection. Low diversity in these critical genes suggests that a paradigm shift may be required for future maize breeding. Rather than purely relying on the diversity within maize or on transgenics, future maize breeding would perhaps benefit from the incorporation of alleles from maize's wild relatives.

Maize molecular diversity is roughly 2 to 5-fold higher than that of other domesticated grass crops (Buckler *et al.*, 2001). Tenaillon *et al* (Tenaillon *et al.*, 2001) reported that in 25 maize individuals, one nucleotide every 28 base pairs is polymorphic, and overall nucleotide diversity is almost 1.3%. That study, the largest examination of

random maize loci, found almost no evidence of selection in 21 genes along chromosome one. Maize's closest wild relative, *Z. mays* ssp. *parviglumis* (a teosinte), often has levels of nucleotide diversity that surpass 2% (Goloubinoff *et al.*, 1993; Eyre-Walker *et al.*, 1998; Hilton and Gaut, 1998; White and Doebley, 1999). The tremendous diversity of maize and teosinte has been the raw genetic material for the radical transformation of maize into the world's highest yielding grain crop.

To date only two loci have been identified as targets of selection in maize. *Teosinte branched, tb1*, is responsible for modifying apical dominance, tillering, and inflorescence position, and was vital to maize domestication (Doebley *et al.*, 1997). A nucleotide diversity survey by Wang *et al* (Wang *et al.*, 1999) showed that selection was strongly directed at *tb1*. During the improvement of maize, kernel color was modified through the anthocyanin pathway. A nucleotide diversity survey of the *c1* gene (an anthocyanin regulator) indicated the locus was also under selection (Hanson *et al.*, 1996). In both cases, surveys of nucleotide diversity were critical in linking selection pressure to specific genes in pathways. There has been little examination of selection in the maize starch metabolic pathway, which is probably the single most important pathway for grain production.

Starch production is critical to both the yield and the quality of the grain. In the maize endosperm, sucrose is converted to glucose and then into starches that normally account for 73% of the kernel's total weight (Fig. 1). Roughly three-quarters of the total starch is amylopectin, which consists of branched glucose chains that form insoluble, semi-crystalline granules. The remainder of the starch is amylose, comprised of linear

chains of glucose that adopt a helical configuration within the granule (Myers *et al.*, 2000). The chemical and structural nature of amylose and amylopectin confer specific properties related to the viscosity of starch that are important in food processing. Although amylose and amylopectin may have synergistic effects on viscosity (Jane and Chen, 1992), amylose is typically thought to affect the gelling of starch (Ott and Hester, 1965). Gelatinization is a property that controls starch firmness, due to reassociation of glucose molecules. The contribution of amylose to starch viscosity is an increase in pasting temperatures and shear stress stability (Miller *et al.*, 1973; Jane and Chen, 1992). Alternatively, amylopectin is primarily responsible for granule swelling and eventual thickening of pastes upon addition of heat (Tester and Morrison, 1990). Starch pasting modifies the ability of foods to hold fat and protein molecules that enhance flavor and texture (Johnson *et al.*, 1999). Selection for yield and better kernel quality may have contributed to the maize domestication and improvement process.

Plant genetics and biochemistry have so far identified over 20 genes involved in starch production (Nelson and Pan, 1995; Myers *et al.*, 2000). We have focused on six key genes known to play major roles in this pathway: *amylose extender1* (*ae1*), *brittle2* (*bt2*), *shrunken1* (*sh1*), *shrunken2* (*sh2*), *sugary1* (*su1*), and *waxy1* (*wx1*) (Fig. 1). *bt2*, *sh1*, and *sh2*, located upstream in the pathway, aid in the formation of glucose. High SH1 activity plays a role in better grain filling, probably by providing more glucose for ADP-glucose pyrophosphorylase (AGPase) (Chourey and Nelson, 1976; Liang *et al.*, 2001). *sh2* and *bt2* encode subunits of the AGPase enzyme, which controls the rate-limiting step in starch production and is regulated by allosteric effectors (Hannah and Nelson, 1976; Greene and Hannah, 1998). The enzymes coded by *ae1*, *su1*, and *wx1* produce the final

products of starch metabolism, amylose and amylopectin (Fisher and Boyer, 1983; James *et al.*, 1995). Mutations at the *wx1* locus eliminate amylose, and have been used in modern breeding to create high amylopectin maize (Tsai, 1974). Reduction of amylopectin has been accomplished with *ae1* and *su1* mutants (James *et al.*, 1995). We evaluated selection at these six loci by examining patterns of nucleotide diversity in maize and *Z. mays* ssp. *parviglumis*.

Materials and Methods

Sampling

To examine diversity and selection in the starch pathway, we sequenced the six loci from inbred lines of maize that represent much of the breeding diversity available. Diversity estimates were performed by sequencing 30 maize inbreds lines that included both coding and noncoding genic regions: A272, A6, B103, B14A, B37, B73, B97, CI187-2, CML254, CML258, CML333, D940Y, EP1, F2, I205, IDS28, IL101, Ki9, Ki21, Ky21, M162W, Mo17, N28Ht, NC260, NC348, Oh43, Pa91, T232, Tx601, W153R. The entire gene and 500-2000bp upstream of the translation start site were sampled from *bt2, sh1, sh2, su1*, and *wx1*. Approximately 8 kbp were sequenced from the 23 kbp *ae1* gene, spanning some of the introns and almost all of the exons. Exon 15 and the large introns 11 and 14 were not sequenced at all.

From *Z. m.* ssp. *parviglumis*, 500-3400 bp of each gene were amplified, cloned and then a single clone was sequenced. The 10 accessions sampled represent much of the range and diversity of the subspecies (USDA: PI566686, PI566688, PI566691, PI331783,

PI331785, PI331786, PI331787, Iltis & Cochrane Site 3, Beadle & Kato Site 5, and Benz 967). *Tripsacum dactyloides* (PLT457 or seeds supplied courtesy of Joseph Burns) was sequenced directly from at least 1000bp of amplification product for each gene. Since PCR errors were a concern in these heterozygous teosinte samples, high fidelity enzyme (*Pfu*) was used and statistical tests between the maize and teosinte diversity numbers were not conducted.

To clarify the origin of sweet corn, we amplified by PCR and sequenced a 1,000 bp area in the promoter, and the 1,000 bp area surrounding residue 578, from seven Mexican and South American maize accessions (AYA-32, BOV-331, BOV-344, BOV-396, CUN-465, JAL-304, NAR-494). We sequenced the entire *su1* locus in AYA-32 and JAL-304.

Statistics

Nucleotide diversity, π, is the average number of nucleotide differences per site between two sequences. π was estimated using DnaSP (Rozas and Rozas, 1999). Insertions and deletions were excluded from the estimates. Tajima's test of selection (Tajima, 1989) was also conducted using DnaSP (Rozas and Rozas, 1999).

The HKA tests were used to evaluate selection at the loci (Hudson *et al.*, 1987). We used the HKA test to compare the six starch loci with 11 neutral loci sampled previously (Tenaillon *et al.*, 2001), and applied the test at two levels within our germplasm. In the first sample we used all 30 diverse lines (set A), while in the second sample we used a narrower subset of 9 lines (set B). Testing at two levels helps determine

whether selection occurred across all breeding germplasm or only in a narrower subset of predominantly US germplasm. In the test for all breeding germplasm (set A), the complete set of 30 diverse lines in this study and all lines from Tenaillon et al. (Tenaillon *et al.*, 2001) were used. For the narrow germplasm, US lines were the predominant focus (set B). This set was designed to be equivalent to the Tenaillon et al. US germplasm (Tenaillon *et al.*, 2001). Set B excluded sweet corn, popcorn, and most maize with tropical germplasm. Six of 9 lines (B73, Mo17, W153R, Ky21, Oh43, Tx601) were identical to Tenaillon et al. (Tenaillon *et al.*, 2001), while three substitutes were made (Ki9 for Mo24W, I205 for T8, NC348 for NC258) based on the closest genetic distance between lines from SSR data (data not shown). In the narrow set, only Ki9 is from outside the US.

Association tests were conducted using the STRAT program (Pritchard *et al.*, 2000b), and population structure effects were reduced by the method of Thornsberry et al. (Thornsberry *et al.*, 2001).

Results

Although the starch loci exhibited a wide range in diversity, average diversity (π) in the starch loci was 2.3-fold lower than 20 random maize loci at silent sites (T-test; $p<0.05$) (Table 1), and 4.8 fold lower at nonsynonymous sites (Tenaillon *et al.*, 2001). We also sampled genetic diversity in Z. *mays* spp. *parviglumis*. In non-selected loci (*adh1, adh2, glb1, hm1, hm2,* and *te1*), maize has 1.3-fold lower diversity than *Z. mays* ssp. *parviglumis* (Fig. 2) (Goloubinoff *et al.*, 1993; Eyre-Walker *et al.*, 1998; Hilton and

Gaut, 1998; White and Doebley, 1999). In contrast, three of the starch loci (*su1*, *bt2*, and *ae1*) exhibited a dramatic 3 to 7-fold reduction in diversity (Fig.2), which is consistent with artificial selection since domestication. Rare divergent haplotypes at *ae1* and *su1* are responsible for most of the diversity. If these rare divergent haplotypes were excluded, then diversity at these loci, among the 30 inbreds would be almost zero.

To formally test for selection, we employed the Hudson, Kreitman, Aguade (HKA) test (Hudson *et al.*, 1987). This test uses an outgroup, in this case the wild relative *Tripsacum dactyloides*, to compare rates of divergence between species to levels of polymorphism within species. A low level of intraspecific diversity to interspecific divergence relative to other loci suggests that selection has reduced diversity. We used the HKA test to compare starch loci with 11 neutral loci sampled previously (Tenaillon *et al.*, 2001), and applied the test at two levels within our germplasm. In the first sample, we used all 30 diverse lines (set A), while in the second sample, we used a narrower subset of 9 lines (set B) that contains no sweet, popcorn, and little tropical germplasm. Testing at two levels helps determine whether selection occurred on a broad scale across all breeding germplasm or only in narrow germplasm. Over the diverse maize sample, *sh2* had significant HKA results (Table 2), but both maize and *Z. mays* ssp. *parviglumis* had low levels of diversity, indicating that selection may have occurred before the divergence of maize from *Z. m.* ssp. *parviglumis* (Figure 2). *bt2* and *su1* had highly significant HKA tests for both germplasm sets and exhibited high levels of diversity in *Z. m.* ssp. *parviglumis* (Table 1), and therefore, *bt2* and *su1* were likely targets of positive selection since the divergence of maize and *Z. m.* ssp. *parviglumis*. *ae1* had non-significant HKA results for diverse germplasm (set A), but it had low diversity for all germplasm and

significant HKA results for the narrow subset of germplasm (set B). Additionally, another test of selection, Tajima's *D*, indicated strong confirmation of selection at *ae1* (*D*=-2.29; P<0.01) (Tajima, 1989). Together this suggests selection may be ongoing at *ae1*.

Due to the close proximity of *bt2* and *su1* to the domestication locus *tga*, we were concerned about a possible hitchhiking effect. The uncloned *tga* locus is 3.3cM from *bt2* and 4 cM from *su1* (Dorweiler *et al.*, 1993). To determine whether low diversity of *bt2* and *su1* was the product of selection on the neighboring *tga* locus, we characterized diversity in *nr (nitrate reductase*) (Gowri and Campbell, 1989), the closest known and cloned gene to *tga*. *nr* diversity (π_{silent}=0.008; $\pi_{nonsyn.}$=0.003) was roughly 3-fold higher than *su1* and *bt2*. Thus a hitchhiking effect due to selection on *tga* is not responsible for the *su1* and *bt2* low diversity. Selection has been shown to impact only parts of a single gene, as in the domestication gene *tb1*(Wang *et al.*, 1999). Therefore it is not surprising that a hitchhiking effect does not extend 3cM in a species where linkage disequilibrium decays rapidly (Remington *et al.*, 2001; Tenaillon *et al.*, 2001).

Our survey of *su1* discovered a polymorphism unique to all sampled US sweet corns. This polymorphism converted tryptophan to arginine at conserved residue 578. Association tests between the *su1* polymorphism and the sweet phenotype were significant (p<0.001), even while controlling for population structure (Pritchard *et al.*, 2000b; Thornsberry *et al.*, 2001). The amino acid change was also one of two identified in the molecular and biochemical study by Dinges et al. (Dinges *et al.*, 2001). Sweet maize cultivars from central and South America did not carry the tryptophan to arginine mutation. Rather, the Mexican sample had a 1.3 kbp transposable element in exon 1 that

disrupts normal translation of the gene. Further investigation of the South American samples is needed to determine the mutation responsible for the sweet phenotype.

Discussion

Previous studies of random maize loci have shown departures from neutrality are rare (Tenaillon *et al.*, 2001), while selection was prevalent across genes in the starch production pathway. *bt2, su1* and *ae1* have average levels of diversity in *Z. mays* ssp. *parviglumis*, but low levels of diversity in maize, consistent with artificial selection during maize domestication and improvement. The significant HKA results for *bt2* and *su1* indicate that this selection probably occurred before the dispersal of maize germplasm throughout the world, while at *ae1* the HKA test with narrow germplasm (set B) and Tajima's test suggest selection is ongoing. It is striking that at least half of these starch loci exhibit strong evidence of selection, while random loci in maize show almost no proof of selection.

Why was there selection in this pathway? Given the position of *ae1, bt2 and su1* in the starch pathway, we propose that selection, both historically and currently, has been for increased yield and different amylopectin qualities. Starch (unlike protein) is often lacking in hunter-gatherer diets in the tropics and subtropics (Piperno and Pearsall, 1998), so it would be reasonable that the early cultivators of maize focused on improving their yield of starch. Native American and modern breeders have boosted yield and starch in maize several-fold over its wild relative and genes like *bt2* may have had an important role. Grain quality is also critical, as evinced in wheat for gluten levels and rice for stickiness. In maize the ratio of amylose to amylopectin, as well as amylopectin branch

chain length is important for altering starch gelatinization and pasting properties that could affect everything from porridge to tortilla texture (Jane and Chen, 1992; Jane *et al.*, 1999; Klucinec and Thompson, 2002, [‖]). Indeed, increased amylopectin improves the textural properties of tortillas, making them softer [‖]. Obviously, the relationship between amylose and amylopectin in the starch granule is a complex one. However, because of the substantial consequence to food processing, it is quite probable that people selected for quality traits very early in the process of maize improvement. If starch gelatinization had been an important target for selection, then we should have seen selection evidence at *wx1*. Instead we saw strong selection at the starch branching (*ae1*) and debranching enzymes (*su1*), which suggests that amylopectin structure and therefore pasting properties were the key targets of selection. The exact nature of this selection will not be understood until a wide-range of teosinte starch alleles are examined in maize genetic backgrounds and combined with subsistence archaeological studies. The identified genes and alleles will provide the background necessary for further dissection of this pathway.

In some varieties of maize, there has been selection for low-starch, high-sugar phenotypes, which are popular because of their sweet taste. Examples of maize races with the sweet phenotype are found throughout the Americas. The *su1* mutants give rise to the sweet corn phenotype because the mutants accumulate sucrose, and *su1* was one of the first loci described genetically (Correns, 1901). Our data clarify the origin of sweet corn, for which there were two competing hypotheses. One theory argues a single origin from a Peruvian race (Chullpi) (Mangelsdorf, 1974), while others propose independent origins

[‖] Yeggy, H., Zelaya, N., Suhendro, EL, McDonough, CM and Rooney, LW. American Association of Cereal Chemists 84th Annual Meeting, Oct. 31-Nov. 3, 1999, Seattle, abstr. 271.

from recurring mutations (Tracy, 2001). Our discovery of two independent *su1* mutations suggests that there have been at least two independent origins of sweet corn.

The reduction of diversity in starch loci is dramatic, and should motivate a paradigm shift for maize breeding. Maize has high levels of diversity at most loci (Tenaillon *et al.*, 2001), and often has 2 to 5 times the diversity of other grass crops (Fig. 3). This tremendous variation has allowed maize to respond to selection for industrial farming in the last century. However, limited diversity in starch and perhaps other critical pathways may preclude current breeding practices from reaching their full potential. Selection for yield may arguably be a constant throughout the millennia, while selective pressures on quality differ as cultural preferences change. Hence, useful variation especially for grain quality needs to be generated for these pathways. Perhaps the most efficient way to introduce potentially useful diversity into maize is to introgress or transform the abundant allelic variation present in teosinte for selected genomic regions or genes. This approach has been successful in tomato (Tanksley and McCouch, 1997), and it could provide the allelic variation necessary to further increase yield and provide a much wider range of kernel qualities.

Acknowledgements

We thank Martha James for her wonderful suggestions on the possible targets of selection. We thank Major Goodman and John Doebley for their help in choosing and acquiring germplasm. We thank Michael Purugganan and William Tracy for useful discussions. We are grateful to the College of Agriculture and Life Sciences Genome Research

Laboratory at North Carolina State University for assistance with sequencing. This research was supported by NSF grant DBI-9872631 and USDA-ARS.

Literature Cited

Buckler, E. S., J. M. Thornsberry and S. Kresovitch (2001). "Molecular diversity, structure and domestication of grasses." *Genetical Research* **77**: 213-218.

Chourey, P. S. and O. E. Nelson (1976). "Enzymatic Deficiency Conditioned by Shrunken 1 Mutations in Maize." *Biochemical Genetics* **14**(11-1): 1041-1055.

Correns, C. (1901). "Bastarde zwischen Maisrassen, mit besonderer Berucksichtigung der Xenien." *Biblioteca Bot* **53**: 1-161.

Dinges, J. R., C. Colleoni, A. M. Myers and M. G. James (2001). "Molecular Structure of Three Mutations at the Maize *sugary1* Locus and Their Allele-Specific Phenotypic Effects." *Plant Physiology* **125**: 1406-1418.

Doebley, J. F., A. Stec and L. Hubbard (1997). "The evolution of apical dominance in maize." *Nature* **386**: 485-488.

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**(5131): 233-235.

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." *Proceedings of the National Academy of Sciences of the United States of America* **95**: 4441-4446.

Fisher, M. B. and C. D. Boyer (1983). "Immunological characterization of maize starch branching enzymes." *Plant Physiology* **72**: 813-816.

Goloubinoff, P., S. Paabo and A. C. Wilson (1993). "Evolution of Maize Inferred from Sequence Diversity of an Adh2 Gene Segment from Archaeological Specimens." *Proceedings of the National Academy of Sciences of the United States of America* **90**(5): 1997-2001.

Gowri, G. and W. H. Campbell (1989). "Cdna Clones for Corn Leaf Nadh-Nitrate Reductase and Chloroplast Nad(P)+-Glyceraldehyde-3-Phosphate Dehydrogenase - Characterization of the Clones and Analysis of the Expression of the Genes in Leaves as Influenced by Nitrate in the Light and Dar." *Plant Physiology* **90**(3): 792-798.

Greene, T. W. and L. C. Hannah (1998). "Adenosine diphosphate glucose pyrophosphorylase, a rate- limiting step in starch biosynthesis." *Physiologia Plantarum* **103**(4): 574-580.

Hannah, L. C. and O. E. Nelson (1976). "Characterization of Adp-Glucose Pyrophosphorylase from Shrunken-2 and Brittle-2 Mutants of Maize." *Biochemical Genetics* **14**(7-8): 547-560.

Hanson, M. A., B. S. Gaut, A. O. Stec, S. I. Fuerstenberg, M. M. Goodman, E. H. Coe and J. F. Doebley (1996). "Evolution of anthocyanin biosynthesis in maize kernels: The role of regulatory and enzymatic loci." *Genetics* **143**(3): 1395-1407.

Hilton, H. and B. S. Gaut (1998). "Speciation and domestication in maize and its wild relatives: Evidence from the globulin-1 gene." *Genetics* **150**(2): 863-872.

Hudson, R. R., M. Kreitman and M. Aguade (1987). "A Test of Neutral Molecular Evolution Based on Nucleotide Data." *Genetics* **116**(1): 153-159.

James, M. G., D. S. Robertson and A. M. Myers (1995). "Characterization of the Maize Gene Sugary1, a Determinant of Starch Composition in Kernels." *Plant Cell* **7**(4): 417-429.

Jane, J., Y. Y. Chen, L. F. Lee, A. E. McPherson, K. S. Wong, M. Radosavljevic and T. Kasemsuwan (1999). "Effects of amylopectin branch chain length and amylose content on the gelatinization and pasting properties of starch." *Cereal Chemistry* **76**(5): 629-637.

Jane, J. L. and J. F. Chen (1992). "Effect of Amylose Molecular-Size and Amylopectin Branch Chain- Length on Paste Properties of Starch." *Cereal Chemistry* **69**(1): 60-65.

Johnson, L. A., C. P. Baumel, C. L. Hardy and P. J. White (1999). Identifying Valuable Corn Quality Traits for Starch Production. Ames, IA, Iowa State University Press.

Klucinec, J. D. and D. B. Thompson (2002). "Amylopectin nature and amylose-to-amylopectin ratio as influences on the behavior of gels of dispersed starch." *Cereal Chemistry* **79**(1): 24-35.

Liang, J. S., J. H. Zhang and X. Z. Cao (2001). "Grain sink strength may be related to the poor grain filling of indica-japonica rice (Oryza sativa) hybrids." *Physiologia Plantarum* **112**(4): 470-477.

Mangelsdorf, P. C. (1974). *Corn. Its Origin, Evolution and Improvement*. Cambridge, MA, Harvard University Press.

Miller, B. S., R. I. Derby and H. B. Trimbo (1973). "Pictorial Explanation for Increase in Viscosity of a Heated Wheat Starch-Water Suspension." *Cereal Chemistry* **50**(3): 271-280.

Myers, A. M., M. K. Morell, M. G. James and S. Ball (2000). "Recent Progress toward understanding biosynthesis of the amylopectin crystal." *Plant Physiology* **122**: 989-997.

Nelson, O. and D. Pan (1995). "Starch Synthesis in Maize Endosperms." *Annual Review of Plant Physiology and Plant Molecular Biology* **46**: 475-496.

Ott, M. and E. E. Hester (1965). "Gel Formation as Related to Concentration of Amylose and Degree of Starch Swelling." *Cereal Chemistry* **42**(5): 476-&.

Piperno, D. R. and D. M. Pearsall (1998). *The origins of agriculture in the lowland neotropics*. San Diego, Academic Press.

Pritchard, J. K., M. Stephens, N. A. Rosenberg and P. Donnelly (2000b). "Association mapping in structured populations." *American Journal of Human Genetics* **67**: 170-181.

Remington, D. L., J. M. Thornsberry, Y. Matsuoka, L. M. Wilson, S. R. Whitt, J. F. Doebley, S. Kresovitch, M. M. Goodman and E. S. Buckler, IV (2001). "Structure of linkage disequilibrium and phenotypic associations in the maize genome." *Proceedings of the National Academy of Sciences of the United States of America* **98**(20): 11479-11484.

Rozas, J. and R. Rozas (1999). "DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis." *Bioinformatics* **15**(2): 174-175.

Sokal, R. R. and F. J. Rohlf (1995). *Biometry*. New York, W.H. Freeman and Co.

Tajima, F. (1989). "Statistical-Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism." *Genetics* **123**(3): 585-595.

Tanksley, S. and S. R. McCouch (1997). "Seed banks and molecular maps: unlocking genetic potential from the wild." *Science* **277**: 1063-1066.

Tenaillon, M. I., M. C. Sawkins, A. D. Long, R. L. Gaut, J. F. Doebley and B. S. Gaut (2001). "Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.)." *Proceedings of the National Academy of Sciences of the United States of America* **98**(16): 9161-9166.

Tester, R. F. and W. R. Morrison (1990). "Swelling and Gelatinization of Cereal Starches .1. Effects of Amylopectin, Amylose, and Lipids." *Cereal Chemistry* **67**(6): 551-557.

Thornsberry, J. M., M. M. Goodman, J. F. Doebley, S. Kresovitch, D. Nielson and E. S. Buckler, IV (2001). "*Dwarf8* polymorphisms associate with variation in flowering time." *Nature Genetics* **28**: 286-289.

Tracy, W. F. (2001). *Specialty Corns*. A. R. Hallauer. New York, CRC Press LLC**:** 155-197.

Tsai, C. Y. (1974). "Function of Waxy Locus in Starch Synthesis in Maize Endosperm." *Biochemical Genetics* **11**(2): 83-96.

Wang, R.-L., A. Stec, J. Hey, L. Lukens and J. F. Doebley (1999). "The limits of selection during maize domestication." *Nature* **398**: 236-239.

White, S. E. and J. F. Doebley (1999). "The molecular evolution of terminal earl, a regulatory gene in the genus Zea." *Genetics* **153**(3): 1455-1462.

Table 1. Summary of maize nucleotide diversity.

		Diversity (π)	
Locus	Sites	Silent	Nonsyn.
ae1	6781	0.0029	0.0007
bt2	6098	0.0023	0.0010
sh1	6176	0.0121	0.0005
sh2	6754	0.0050	0.0013
su1	9378	0.0027	0.0004
wx1	2978	0.0115	0.0014
Starch Loci[1]	37,330	0.0052	0.0008
Random Loci[2]	10,908	0.0122	0.0038

1. Summary of six starch loci. Sites were summed, and π-values were averaged.

2. The random loci are the 20 loci in Tenaillon et al (Tenaillon *et al.*, 2001). The domestication gene *tb1* was excluded from these comparisons.

Table 2. HKA tests of selection.

		Diverse Germplasm Set A Lines			Narrow Germplasm Set B Lines		
Locus	Silent Sites	Ratio[1]	f(P<0.05)[2]	P_{all}[3]	Ratio[1]	f(P<0.05)[2]	P_{all}[3]
ae1	2216	0.32	0%	0.109	0.07	18%	0.023
bt2	999	0.01	100%	<0.0001	0.01	91%	<0.0001
sh1	1630	0.21	0%	0.844	0.23	0%	0.815
sh2	1485	0.08	18%	0.002	0.09	9%	0.595
su1	2345	0.04	91%	<0.0001	0.06	18%	<0.0001
wx1	417	0.13	0%	0.712	0.16	0%	0.967
Starch Loci[4]	8951	0.13	35%		0.10	23%	
Random Loci[5]	4567	0.22	2%		0.19	2%	

1. The ratio of θ within the maize sample versus average divergence between maize and *Tripsacum*.

2. The frequency of significant HKA tests between the given starch locus and the 11 neutral loci.

3. *P*-values from all 11 comparisons were combined using Fisher's method of combining probabilities from independent tests of significance to produce an overall *P*-value (Sokal and Rohlf, 1995).

4. Summary of six starch loci. Sites were summed, and other values were averaged.

5. The random loci are the 11 loci from Tenaillon *et al*. (Tenaillon *et al.*, 2001) with *Tripsacum* data. The domestication gene *tb1* was excluded from these comparisons.

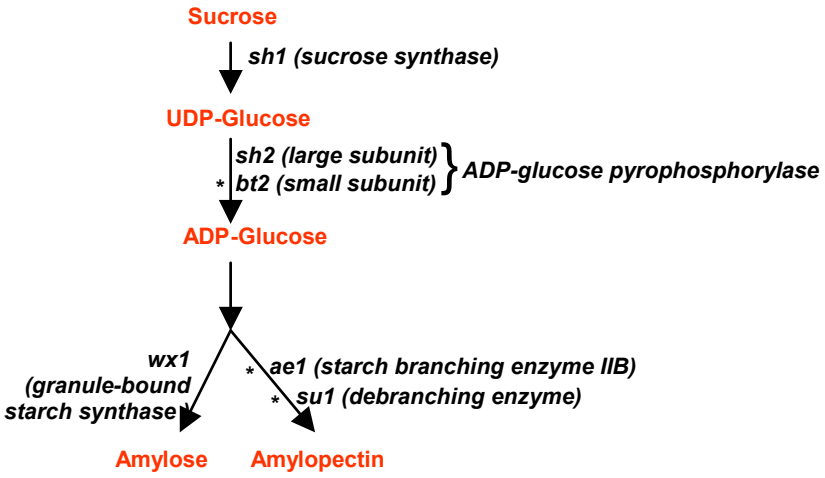


Figure 1. A simplified pathway of starch production in maize, and the position of the six sampled genes in the pathway. Loci with strong evidence of selection (either HKA or Tajima's D test) are indicated by *.

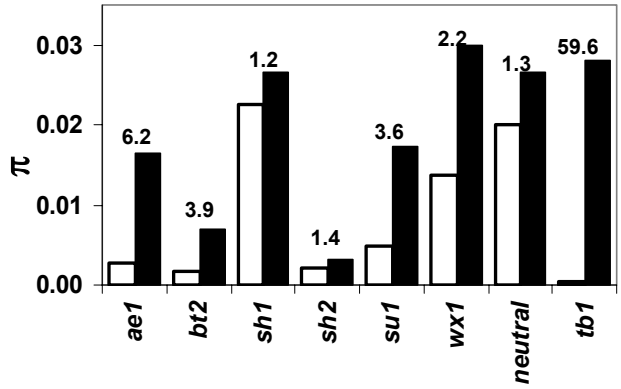


Figure 2. Comparison of silent diversity in maize (open bars) and its wild relative *Z. mays* ssp. *parviglumis* (black bars). For each locus, 500 to 2700 silent bases were sampled. The numbers above the bars indicate the fold reduction in diversity between maize and *Z. m.* ssp. *parviglumis*. Neutral refers to the average of 6 non-selected genes (*adh1, adh2, glb1, hm1, hm2,* and *te1*), and *tb1* is the only cloned domestication gene from maize (only the highly selected promoter region is shown).

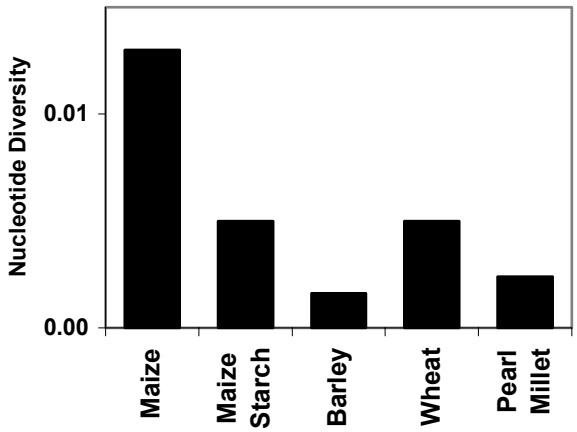


Figure 3. Comparison of nucleotide diversity in maize and various grass crops. Maize data is the average diversity at random loci from Tenaillon *et al.* (Tenaillon *et al.*, 2001); the maize starch average is from the six genes examined here. The other grasses are diversity averages from published studies (Buckler *et al.*, 2001); however, they are often based on a limited number of loci.

CHAPTER 3

ASSOCIATIONS WITH KERNEL COMPOSITION AND STARCH PASTING PROPERTIES IN SIX MAJOR GENES INVOLVED IN STARCH BIOSYNTHESIS IN MAIZE

Abstract

Starch content in the maize kernel is a complex trait controlled by many genes. The market for corn and its main product, starch, continues to grow with diverse applications, ranging from human and animal consumption, to plastics and ethanol production. For this study, six candidate genes of maize involved in kernel starch biosynthesis were evaluated for correlations with kernel composition traits and starch pasting properties using an association approach: *Amylose extender1* (*Ae1*), *Brittle endosperm2* (*Bt2*), *Shrunken1* (*Sh1*), *Shrunken2* (*Sh2*), *Sugary1* (*Su1*), and *Waxy1* (*Wx1*). Sites in *Bt2*, *Sh1*, and *Sh2* showed significant associations for kernel composition traits, while significant associations for starch pasting properties were found in *Ae1*, *Sh1,* and *Sh2*. Specifically, polymorphisms in *Sh2* were found to have an effect on amylose content, strongly supporting previous findings of linkage mapping studies implicating the *Sh2* chromosomal region in affecting amylose content. In most cases, high resolution (within 1000 base pairs) was achieved in evaluating these starch candidate genes, demonstrating the power of association mapping to detect significant variation even in a moderately heritable pathway, such as starch production in maize.

Introduction

It is clear that improved strategies in food production are necessary to meet rising demands caused by increased growth in the human population. Starch is the major component in the human diet, comprising 55-75% of daily food intake, and cereal grains provide the major source of this starch. Cereal starches are also important as food sources

for domestic animals (Pan, 2000). Maize is one of the top three cereal grains in amounts produced worldwide (http://apps.fao.org), and has the benefit of numerous starch mutants that provide a unique source of specialty starches such as the amylose-free waxy starches, important for many industrial applications (Hallauer, 2001), and the *sugary* mutants, responsible for the production of popular sweet corn varieties (Pan, 2000). Therefore improvements in both yield and starch quality are desirable goals.

Maize starch is composed of ~25% amylose, a mostly linear chain of α(1→4) linked glucose molecules, and ~75% amylopectin, a more highly branched molecule of α(1→4) linkages with α(1→6) branch points. Long-term storage of starch in maize occurs in the kernel endosperm, where the semi-crystalline nature of amylopectin allows for efficient packaging into granules. Starch is produced within amyloplasts, plastids that also serve as storage until the starch is utilized during seed germination. Amylose and amylopectin are composed of simple glucose monomers, belying the actual complexity of the organization of the starch granule. The desire to control starch amount as well as modified starch products has driven biologists, chemists, and physicists to fill remaining gaps in our current understanding of starch metabolism (Smith, 2001). For example, *in vitro* attempts at producing amylopectin have not been successful, but instead produce an animal-like glycogen product (Guan *et al.*, 1995); therefore, the exact nature of amylopectin formation is not known. To complicate matters, many starch enzymes have multiple isoforms (Fisher *et al.*, 1996; Gao *et al.*, 1996; Huang and Wang, 1998; Sidebottom *et al.*, 1998; Beckles *et al.*, 2001). Currently little is known about transcription factors associated with starch gene regulation, as geneticists and biochemists have concentrated primarily on the enzymes involved in starch production.

Research on the well-known mutants of maize has elucidated much of what is known of the pathway. Sucrose transported into the kernel is converted to UDP-glucose and fructose by the major isoform of sucrose synthase, encoded by the *Sh1* gene (Chourey and Nelson, 1976). *Sh2* and *Bt2* encode the large and small subunits, respectively, of ADP-glucose pyrophosphorylase (AGPase), which converts ADP-glucose into glucose-1-phosphate, the substrate for starch synthases (Tsai and Nelson, 1966; Bae *et al.*, 1990; Bhave *et al.*, 1990). Regarded as the rate-limiting step in starch biosynthesis, AGPase is allosterically regulated by 3-phosphoglycerate and P_i, and is the basis for much interest in controlling starch yield by modification of its allosteric effector sites (Stark *et al.*, 1992). Starch synthases then sequentially add glucose-1-phosphate molecules onto the non-reducing ends of a growing starch chain. Granule-bound starch synthase, encoded by the *Wx1* locus in maize, is solely responsible for amylose production (Shure *et al.*, 1983). Amylose-free starches caused by the *wx1* mutant in maize have long been of commercial interest (Deatherage *et al.*, 1954). Another maize mutant of interest affecting starch quality is *ae1*, which results in kernels with higher amounts of amylose than nonmutant kernels (Fisher *et al.*, 1996; Kim *et al.*, 1998). The nonmutant *Ae1* gene codes for the starch branching enzyme IIb isoform, which hydrolyzes $\alpha(1\rightarrow4)$ linkages and reattaches these chains with $\alpha(1\rightarrow6)$ branch points found in amylopectin. *Su1* encodes a debranching enzyme of the isoamylase type, and *su1* mutant kernels contain the highly branched, water-soluble phytoglycogen instead of amylopectin starch and are popular in sweet corn production (James *et al.*, 1995). In order to obtain the semi-crystalline formation of amylopectin, it may be that the correct ratio of both starch branching to debranching enzymes are important (Ball *et al.*, 1996), but

the role of isoamylase in conjunction with branching enzymes has not been resolved and differing models have been proposed (reviewed in Smith, 2001).

Like many important agronomic traits, starch content varies across maize accessions, and it is this variation that breeders utilize to improve crops. Linkage mapping has identified several regions in the maize genome that have an effect on kernel starch content and starch enzyme activity, and some of these quantitative trait loci (QTL) correspond to starch biosynthesis genes. One of the most documented genes, *Sh2,* colocalizes with a QTL effect on protein and starch levels (Goldman *et al.*, 1993), and has also been found to associate with amylose content (Prioul *et al.*, 1999; Séne *et al.*, 2000). Another positional candidate gene, *Sh1,* has an effect on starch and protein content (Berke and Rocheford, 1995). From the basis of mutational studies, *Bt2, Ae1, Su1,* and *Wx1* are functional candidate genes that are thought to have an effect on either starch content or starch quality. Mutant kernels have severely reduced starch content or contain an altered amylose/amylopectin ratio. All six genes are major players in the starch pathway and are good starting points to find functional polymorphisms that affect starch in maize. While QTL studies are suggestive, their resolution is often on the order of 10 cM, corresponding to millions of bases, so real evaluation of these genes has not occurred. Association mapping provides an avenue for high-resolution evaluation of these candidate genes.

Association mapping, although common in human genetics, has only recently been applied to plant populations. It provides high resolution, and has the potential to evaluate QTL with smaller effects than can be positionally cloned. Furthermore, advantages of association mapping over positional cloning are that several candidate genes can be tested at once, and that a wide range of alleles can be evaluated.

To date, association methods are uncommon in plants. The amount of linkage disequilibrium (LD), or the nonrandom association of alleles, present in a species is critical to whether or not association mapping would be an appropriate approach. However, two studies of LD within maize in both diverse inbreds and traditional landraces suggest that in most cases LD decays rapidly both within genes and in intergenic regions, usually within one kilobase (Remington *et al.*, 2001; Tenaillon *et al.*, 2001). Therefore, high resolution is possible in maize when using association approaches. However, one difficulty in applying association methods is that LD can be present as a result of genetic drift, selection, or population admixture. Therefore, LD can contain the confounding effect of population sub-structure, resulting in a high frequency of false positive associations, as seen in human populations (Lander and Schork, 1994). This issue too has been dealt with in maize. In the first study to successfully apply a structured association method, Thornsberry *et al*. (2001) were able to show that polymorphisms in the candidate gene, *Dwarf8*, associated with flowering time in maize. Thornsberry *et al.* (2001) applied a statistical model for case/control studies in human disease from Pritchard *et al*. (2000b) and modified the model for use with a quantitative trait (Thornsberry *et al.*, 2001). By taking into account the underlying population structure, Thornsberry, *et al.* (2001) found the risk of obtaining false positive associations was reduced.

In this study six starch candidate genes (*Ae1, Bt2, Sh1, Sh2, Su1,* and *Wx1*) were tested for associations with starch content and quality, agronomically important yield traits of maize, by employing the association method of Thornsberry *et al*. (2001). Each gene was sequenced in a diverse set of maize inbred lines in order to locate the nucleotide polymorphisms or regions associated with either starch content or starch quality traits.

Polymorphisms identified in this survey could be used in future genetic and breeding studies to manipulate maize starch content.

Materials And Methods

Plant materials

For this study, 102 inbred maize lines were used that represent most of the diversity available to breeding programs around the world. They are divided into three groups: the stiff-stalks (SS), the non-stiff stalks (NSS), and the tropical/semitropicals (ST) (Remington *et al.*, 2001) (Table 1). The plants were grown at four different field sites in the U.S. with a total of six replications: Homestead, Florida (winter 1998-1999); West Lafayette, Indiana (summer 2000); Clayton, North Carolina (summer 2001) in two replications; and Urbana, Illinois (summer 2001) in two replications.

Seed for each line was pooled from individuals of the same genotype, and total kernel starch, oil, protein, and moisture content were measured from ground, mature, dried kernels using a Dickey-John near infrared (NIR) light reflectance machine. All six replications were phenotyped by NIR. Starch was isolated to measure amylose content. The following starch pasting and viscosity profiles were determined using a controlled stress rheometer (CSR, Carri-Med CSL-100, TA Instruments, Dover, DE): starch breakdown, consistency, cool paste viscosity, hot paste viscosity, pasting temperature, peak temperature, peak time, peak viscosity, setback, and trough viscosity. For amylose content and starch pasting characteristics, only kernels from the Homestead, FL (1999) replication were phenotyped.

Amplification and sequencing

Genomic sequence data for six candidate genes for maize kernel starch were obtained from published sequences in Genbank: *Ae1* (accession AF072725), *Bt2* (accession AF334959), *Sh1* (accession X02382), *Sh2* (accession M81603), *Su1* (accession AF030882), and *Wx1* (accession XO3935). Primers for PCR were designed using PrimerSelect from DNAstar software, or from Primer3, available at www.genome.wi.mit.edu/genome_software/other/primer3.html. Gene fragments spanning most of the length of all genes, except *Ae1,* as well as upstream promoter sequence, were amplified from a subset of 29 lines (Table 1) chosen to maximize diversity utilizing SSR data. Due to the large size of *Ae1* (23 kb), amplification focused primarily on coding regions. PCR products were then either directly sequenced, or cloned into pCR-TOPO TA vector (Invitrogen, Carlsbad, CA). Sequence fragments were contiged using PHRED/PHRAP and aligned using BioLign, a custom alignment program developed by Tom Hall. For information about BioLign, visit http://www.maizegenetics.net. Alignments were edited manually using chromatographs. Alignments can be accessed at http://statgen.ncsu.edu/panzea/. Specific genomic regions sampled only in a complete set of 97 inbred lines:: for *Ae1*: exon one through exon three, exon 12 through exon 14, and exon 16 through exon 18; for *Bt2*: the promoter through intron one; for *Sh1*: the promoter through the noncoding exon one and a portion of intron one; for *Sh2*: intron eight through intron 10; for *Su1*: the promoter through exon one, and exon 13 through intron 14; for *Wx1*:

exon one through exon two, and exon eight through exon nine. Exact alignment positions corresponding to the above regions are listed in Table 2.

Statistics

To summarize the data over multiple replications, principle component analysis (PCA) was done on both the NIR and the starch pasting data using SAS software (Cary, NC) for all 97 taxa and excluding five sweet corn lines. Although sequence alignments for the six candidate loci from sampled sweet corn lines are available (Ia2132, Il14H, Il101, Il677a, and P39), these low-starch mutants were taken out of the association analyses in order to avoid false results produced by these outliers. The major principle components (PC) found for NIR and for starch pasting were then used as traits in association tests for the 97 taxa. In order to handle missing data for the PCA, missing data were imputed using the KNNimpute program (Troyanskaya *et al.*, 2001) with K set to five nearest neighbors. KNNimpute can be downloaded at http://smi-web.stanford.edu/projects/helix/pubs/impute/. Tests for association and linkage disequilibrium were performed using the software package TASSEL, available at http://www.maizegenetics.net. Interesting polymorphisms, either SNPs or insertions/deletions (indels) at a site frequency of 0.05 or greater, were found initially using TASSEL in the 29-line subset (Table 1) across the complete gene alignment. Small regions containing these polymorphisms were then PCR amplified in the complete set of 97 inbred lines and then sequenced once through the region in order to score the particular polymorphism (Table 2). Association tests were run again with the entire 97 lines of these smaller regions to determine whether the association remained significant. All association

tests were run with and without population structure included, using logistic regression or ANOVA, respectively. One thousand permutations of the data were run to account for multiple tests within a gene, and a significant association was called if the *P*-value of the best site in a region was seen in less than 5% of the permutations. The association test statistic used has been described previously (Thornsberry *et al.*, 2001).

Post hoc t- and *F*-tests were used to further dissect PC associated effects in those genes with significantly associated PC traits. *t*-tests were used to determine whether the sample means of non-PC trait values were significantly different between lines with the polymorphism and those lines without the polymorphism. The use of *F*-tests determined whether the samples had equal or unequal variances. Results of the *F*-tests indicated what type of *t*-test to use: a homoscedastic test (both samples have equal variances), or a test where both samples have unequal variances.

Results

Phenotype variation

Observed kernel composition traits and various starch pasting characteristics are listed in Table 3 and 4, respectively. Starch comprised the largest proportion of the kernel and ranged from 42-63%, with the highest starch average found in the Homestead, FL winter field season (Table 3). Starch content also showed the largest amount of variation, as suggested by the standard deviation values, while protein, oil, and moisture content varied

much less among lines. Broad-sense heritabilities for the NIR data indicated that starch content was less heritable than protein or oil content (H^2 = 0.40, 0.64, 0.57, respectively).

Principle component analysis

Phenotypic trait values for kernel composition (NIR) and starch pasting traits were analyzed separately using PCA. PCA was used in order to reduce multiple testing in the association analyses by summarizing the phenotypes over the various replications and by combining correlated traits into a single PCA index. Results for the kernel composition PCA along with interpretations for each factor are included in Table 5. Three factors explained, cumulatively, 55% of the variation in phenotypes, where factor one alone explained 34%. Table 6 shows the results of the PCA of starch pasting values for the Homestead, FL 1998-1999 field season, where the magnitude of eigenvectors indicate which pasting trait is driving the variance seen in maize viscoamylographic profiles. Four factors explained, cumulatively, 61% of the variance in pasting and viscosity.

Associations with kernel composition and starch pasting

Polymorphisms identified in complete gene alignments for the 29-line subset were tested for association with kernel composition and starch pasting traits in an initial survey for regions of interest. Areas indicating possible associations are listed in Table 2. Several of these areas of interest were sampled in the remainder of the 97 total taxa, and the results of the association tests are shown in Table 7.

Of the six starch genes sampled, four genes showed significant associations ($P \leq 0.05$) for either kernel composition or starch pasting properties: *Ae1, Sh1, Sh2,* and *Bt2*, (Table 7). A total of six significant associations controlling for population structure were identified; however, given that 42 multiple tests were conducted, it is quite possible that two at the 0.05 level are false positives. Overall, *Sh1* showed an association with a general genotype × environment (G × E) effect (kernel composition, factor three), and with amylose content and pasting temperature (starch pasting, factor four) (Table 7). *Sh2* associated with a general G × E effect (kernel composition, factor three) and with pasting temperature vs. amylose content (starch pasting, factor three). *Bt2* associated with oil vs. protein production (kernel composition, factor 2). Lastly, *Ae1* associated with pasting temperature and amylose content (starch pasting, factor four). Greater detail of these associations follows, concentrating on results found with the logistic regression analysis (population structure included).

Sh1

Two significant polymorphisms were found that associated with a general G × E effect (logistic regression; $P < 0.05$) (kernel composition, factor three) (Table 7) (Figure 1). The PCA weighted all kernel traits similarly for factor three, but in opposite directionality across replications, suggesting the G × E effect (Table 5). Both polymorphisms (sites 1195 and 1210) are SNPs in LD, and are located in intron one (Figure 1A). The best site for this specific analysis was site 1210, referred herein as allele *sh1*.1210G (*shrunken1*.site1210-Guanine) and was found in 35 out of the 97 lines surveyed.

The *sh1*.1210G allele occurred in the following lines:: NSS: A441-5, C103, CM7, D940Y, EP1, M162W, Mo17, NC258, SA24, and U267Y; SS: CMV3 and SC213; ST: A272, A6, I-29, CML258, CML281, CML287, Ki3, Ki43, Ki9, NC296, NC298, NC300, NC304, NC320, NC338, NC348, NC350, NC352, NC354, Q6199, SC55, Tzi10, and Tzi18.

Three polymorphic sites associated with amylose content and pasting temperature (starch pasting, factor four) (Table 7) all located within the promoter region of the *Sh1* gene. Two were SNPs, (sites 649 and 728), while the third was an indel in a region with a complex polymorphic sequence pattern of three major alleles. One particular SNP that was in significant LD with other polymorphisms, herein referred to as *sh1*.649C, occurred in lines that showed a higher pasting temperature over the more common allele in the 97 lines, *sh1*.649G (*t*-test; $P < 0.05$) (Figure 1B). The 30 lines that contain *sh1*.649C had on average a 1% increase in pasting temperature. Lines containing *sh1*.649C are:: NSS: A441-5, C103, EP1, H95, I137TN, Mo17, Pa91, and Va26; ST: CML10, CML258, CML281, CML287, CML5, CML61, NC296, NC298, NC304, NC320, NC338, NC348, NC352, NC354, Ki11, Ki21, Ki3, Ki43, Ki9, Tzi10, and SC55; SS: SC213.

Sh2

Two significant polymorphisms were found that associated with a general G × E effect (logistic regression; $P < 0.036$) (kernel composition, factor three) (Table 7). The first polymorphism, site 3674, is a one bp deletion in intron eight, while the other polymorphic site, 4027, is an 11 bp deletion located in intron 10 (Figure 2). Both of these sites are in significant LD with each other, and are found in nine out of the 97 lines surveyed.

Examination of the sequence alignment of the whole gene from 29 lines revealed these two polymorphisms are in LD with a suite of other polymorphisms including multiple SNPs, an eight bp insertion in the promoter, 581 bp away from the noncoding exon one, and a 67 bp deletion in intron 13 (site 4640), indicating an obvious haplotype. Therefore resolution of the causative polymorphism was not possible in this survey. The best site for this specific analysis was site 3674, and is the representative for the above-mentioned haplotype. Allele *sh2*.3674-1 containing the one bp deletion occurred in mainly NSS lines, with one ST exception. Lines with the *sh2*.3674-1 allele are: A619, B84, CI187-2, IDS28, Mo17, Mo24W, SA24, W117Ht, and Ki21 (ST).

Three sites were identified in *Sh2* that had significant associations with pasting temperature vs. amylose content (starch pasting, factor 3) (Table 7). Two of the sites were SNPs, one found in exon 10 (site 3886), and the other found in intron 10 (site 3946). In particular, site 3886 causes a nonsynonymous change (leucine to a serine) in the predicted SH2 protein sequence at amino acid position 318. Protein sequence comparisons with orthologues of SH2 showed that the leucine residue is conserved in other grasses such as sorghum, rice, and wheat, but in the dicot, tomato, a serine is found instead. Lines with the *sh2*.3886C allele contain on average 6% more amylose (*t*-test; $P < 0.007$) (Figure 2A); however, the effect on pasting temperature was insignificant. Twenty-one out of 97 lines contained the *sh2*.3886C allele, with a majority of ST origin. Lines with the *sh2*.3886C allele are:: NSS: A554, F2, F7, Gt112, Ki44, T232, W64A, and Wf9; SS: SC213; ST: A272, A6, CML281, Ki11, Ki3, Ki9, NC320, NC338, NC354, Q6199, Tx601, and Tzi8. The third polymorphism was a 47 bp deletion (site 3634) located in intron eight and was found in 13 lines out of the total 97. Deletion 3634 occurred 8-54 bp from exon nine, a likely location

of the branch site for 3' intron splice recognition. All three sites are in significant LD and contribute to the formation of a haplotype. Examination of the 29-line *Sh2* whole gene alignment showed that other polymorphisms scattered throughout the gene continued to define this haplotype. Other notable polymorphisms in this haplotype are a 31 bp deletion (site 4839), a 224 bp putative miniature transposable element insertion (site 5341), and a 1276 bp ILS-1 type transposable element (site 6214), all located within intron 13.

Bt2

Three SNPs, one located in the promoter (site 817) and two in exon one (sites 905 and 925), were found to be significantly associated with oil vs. protein production (kernel composition, factor two) (Table 7) (Figure 3). All three SNPs are in significant LD with one another, and similar to the case in *Sh2,* the whole *Bt2* gene alignment of 29 taxa revealed a distinct haplotype in the 5' end of the gene that encompassed ~1000 bp. One of these polymorphisms, allele *bt2*.925T caused a nonsynonymous change in the N-terminal region of the BT2 protein, from a proline to a leucine at amino acid 22 (P22L). While the mean oil content between lines varying for the *bt2*.925T allele was not significantly different, the variance in oil content in lines with the *bt2*.925T allele was significantly lower than lines with the more common allele (*F*-test; $P < 0.002$) (Figure 3). Nineteen lines out of 97 contained polymorphism *bt2*.925T, and were of NSS or ST origin:: NSS: A619, EP1, F2, F44, F7, ND246, Oh43, Va26, W117HT, W153R, and W182B; ST: CML247, CML254, CML287, I-29, Ki21, Q6199, Tzi10, and Tzi18.

Ae1

Four SNPs were found to be significantly associated with amylose content and pasting temperature (starch pasting, factor four) (Table 7) (Figure 4). Three out of the four SNPs were located within noncoding regions, and are in significant LD with each other. The fourth SNP (site 1509) located in exon two caused a nonsynonymous change in the predicted AE1 protein sequence from an arginine to a glycine at amino acid 58 (R58G). Twelve taxa out of the total 97 had the R58G mutation. The orthologues in rice, wheat, barley and potato all contain a glycine in the predicted amino acid sequences of AE1, while most maize lines sampled in this study contained an arginine residue. Lines with the *ae1*.1509G allele have on average 4.7% more amylose and a 1.6% higher pasting temperature (*t*-test; $P < 0.0001$ and $P < 0.02$, respectively) (Figure 5 and 6). The *ae1*.1509G allele was found in both NSS and ST lines:: NSS: A441-5, C103, EP1, F2834T, GT112, Ki44, ND246, and T8; ST: CML10, CML333, CML61, and Tzi10.

Discussion

Six major genes controlling starch content in maize were analyzed using an association approach with the intention of identifying markers that control an effect on starch content or quality. Mutational studies have already shown that *ae1, bt2, sh1, sh2, su1,* and *wx1* have major effects on either the amount of starch produced, or affect the types of starch produced in amylose or amylopectin levels. The use of these six starch candidate genes in association analyses is a starting point in the greater understanding of the interactions of

these genes in the starch metabolic pathway. Ultimately, breeders, in order to incorporate desirable alleles to meet specific starch or yield goals, could use high-resolution markers that associate with starch phenotypes.

Phenotypic traits were grouped together into two types of association analyses. General kernel composition traits such as total oil, protein, moisture, and starch content, measured by NIR spectrophotometry, were the basis of traits comprising kernel composition PCA. Specific characteristics of starch viscosity and pasting measurements comprised the second type of analysis, starch pasting PCA. *A priori* knowledge of the starch pathway would suggest that genes farther upstream, like *Sh1, Sh2,* and *Bt2*, would affect characteristics such as overall starch amounts, like the NIR data. Likewise, genes downstream in the pathway, like *Ae1, Su1,* and *Wx1*, which produce modified starches, might affect pasting properties.

Indeed 80% of the significant associations were found in the three upstream starch production genes *Sh1, Sh2,* and *Bt2* with the kernel composition PCA traits, although a direct effect on starch levels is not clear. However, use of PCA was successful in locating regions in the genes that affect other important yield components of the maize seed. *Sh1* associated with a G × E effect at two SNPs in LD with each other and located in intron one. This intron is of interest for its transcriptional enhancer properties (Vasil *et al.*, 1989; Clancy *et al.*, 1994). It is possible that the *sh1*.1210G allele may have an altered ability to affect transcription of the gene over the more common allele, *sh1*.1210A in different environments. Of the four field conditions, West Lafayette, Indiana, (2000) displayed clear signs of a stressed environment in comparison to the other environments. *Sh2* also associated significantly with a G × E effect at a one bp deletion at position 3674; however, this

particular polymorphism was in significant LD with a suite of polymorphisms located throughout the entire gene forming a limited number of haplotypes, thus limiting resolution. Therefore, the causative polymorphism may not even be located within the *Sh2* gene. Several reasons could account for the significant association with the G × E effect with regards to the SH2 subunit of the AGPase enzyme. For example, alternate alleles could perform differently in a stressed environment affecting factors such as heat lability of the enzyme or altered SH2:BT2 interactions. Mutations in the SH2 subunit have been shown to increase the stability of the SH2 subunit, thereby increasing SH2:BT2 interactions (Greene and Hannah, 1998).

Allele *bt2*.925T, a polymorphism located in exon one of *Bt2* causing a P22L mutation, associated with a decrease in variance in oil content. It appears that carbon flux through the pathways leading to the main storage products of oil and starch may be affected by mutations in *Bt2*. Genotypes with the *bt2*.925T allele result in reduced variability by cutting out the high and low extremes in oil production. A legitimate question is if this polymorphism results in less efficient enzyme subunit assembly. Mutations in the N-terminus of a peptide could affect assembly of the SH2:BT2 subunit and/or affect final AGPase stability and activity. In potato, mutations in the small subunit of AGPase affect both the enzyme's heat stability and its inhibition by P_i , an allosteric effector (Laughlin *et al.*, 1998). Because maize AGPase's major activity is cytosolic, a direct effect on oil biosynthesis is possible, as lipid biosynthesis receives the majority of carbon through pyruvate, a product of cytosolic glycolysis of sucrose (White *et al.*, 2000).

Tests for association with starch pasting PCs were significant in three genes: *Ae1,* *Sh1,* and *Sh2.* Variations in a branching enzyme gene, like *Ae1,* are likely to have an effect

on amylose content and/or pasting properties. One particular polymorphism in exon two, *ae1*.1509G caused the nonsynonymous change R58G in the predicted protein sequence. The phenotypic effects seen with this mutation were higher levels of amylose and higher pasting temperatures. The R58G mutation occurred within the identified transit peptide and is located near the cleavage site in front of the N-terminus of the mature BEII protein (Fisher *et al.*, 1993). Two possible scenarios resulting from the mutation in the transit peptide include hindered translocation of the preprotein into the amyloplast, or problems in cleavage negatively affecting activity/amounts of mature BEIIb. Dosage effects are seen when the amount of mutant *ae1* alleles are increased, where a decrease in BEIIb levels results in higher amounts of short chain amylose and longer chain lengths of amylopectin (Boyer *et al.*, 1980; Hedman and Boyer, 1982). Although an increase in amylose content was seen in lines with the *ae1*.1509G polymorphism, it is unclear what causes an increase in pasting temperature, an important indicator of stability in food processing of starches. Amylose content and pasting temperature did not significantly correlate (data not shown); however, these traits could be related to changes through physicochemical properties of the amylopectin branch lengths, thereby contributing to the stability of the starch under stress.

Two of the upstream starch genes, *Sh1* and *Sh2*, each associated significantly with pasting properties. Lines with the *sh1*.649C polymorphism resulted in higher pasting temperatures, while lines with the *sh2*.3886C polymorphism had higher amounts of amylose. It is likely that alterations in the enzyme activity of sucrose synthase and AGPase affecting sugar metabolism in a developing maize kernel would have an effect on flux in the downstream parts of the pathway. It is well known that accumulation of sucrose in plants acts as a signaling mechanism for many genes, and in particular, various endosperm mutants

of maize have been shown to affect many other enzymes involved in carbohydrate metabolism (Doehlert and Kuo, 1990). Two excellent related QTL studies by Séne *et al.* (2000) and Prioul *et al.* (1999) found an effect on amylose content near *Sh2*, using a recombinant inbred line population of a French flint (F2) and a U.S. Iodent line (Io very similar to I205) (Prioul *et al.*, 1999; Séne *et al.*, 2000). The genotypes used in the Séne (2000) study and responsible for higher amylose were sequenced in this study (Table 1). This study implicated allele *sh2*.3886C and other polymorphisms belonging to the same haplotype in higher amylose content. The *sh2*.3886C allele was also present in the line F2. The estimated effect of *sh2*.3886C on amylose content is a 1.3% increase ($R^2 = 0.10$), while a virtually identical allelic effect, (1.5%; $R^2 = 0.101$), was seen by Séne *et al.* (2000). The overall *P*-value can be estimated by using Fisher's method of combining the results of these two studies. Using the Séne reported *P*-value = 0.01 for ANOVA detection of QTL (Séne *et al.*, 2000) and this study's *P*-value = 0.04, the support that *Sh2* modifies the amylose/amylopectin ratio overall is $P < 0.004$ (df = 4). Given that this polymorphism only explains roughly 10% of the variation, this may be the first confirmed QTL with modest effect. However, QTLs that explain modest effects are the norm rather than the exception. This combining of RIL mapping populations with association studies provides a viable approach for dissecting agronomically important traits.

Although *Sh2* has a significant effect on the amylose/amylopectin ratio, a previously reported association with overall starch is probably a false positive result. Prioul *et al.* (1999) found an association with a SacI restriction site within the *Sh2* gene in a sample of 46 unrelated maize inbreds, but did not control for population structure (Prioul *et al.*, 1999). Our analysis also finds a significant association if population structure is not controlled, as

tropical, stiff stalk, and non-stiff stalk germplasm all have different mean starch levels. Although the Prioul association may reflect the G × E nature of *Sh2*, another strong possibility is that this association with starch is purely a population structure effect.

Population structure is an important issue that cannot be ignored in association studies, as it produces false positive results (Knowler *et al.*, 1988; Pritchard *et al.*, 2000a; Pritchard *et al.*, 2000b). Logistic regression analysis that incorporated estimates of population structure was used in order to assess whether polymorphisms within the six starch genes had an effect on starch content or starch pasting properties. The use of population structure enabled Thornsberry *et al.* (2001) to locate polymorphisms within the *Dwarf8* gene that associated with flowering time variation in maize. For traits that were initially responsible for establishing the population structure, such as flowering time, inclusion of these estimates in the analysis is necessary to avoid false associations. It is less clear for starch traits; however, in several instances in the present study correction for population structure was able to detect significant associations by increasing power through the use of an unlinked set of markers (Table 7). Alternatively, correcting for population structure in some instances caused some genes to lose significance; for instance, four out of the six genes tested for kernel composition, factor one (*Sh1, Sh2, Su1,* and *Wx1*) were significant at the 0.10 level in the ANOVA analysis, but lost significance when population structure was added (Table 7). Population structure is significantly related to basic kernel composition (PC1) (ANOVA, $P < 0.02$) and accounted for 10% of the variation. Since kernel quality is population sub-structuring, associations performed without a population structure correction need to be reevaluated.

No significant associations for either kernel composition or starch pasting were found within the smaller areas sampled in the 97 lines in the granule-bound starch synthase gene, *Wx1*, or within the starch debranching enzyme, *Su1*. There is evidence that further sampling of different areas within these genes could be done within the total set of maize lines and might be worthwhile (Table 2).

QTL mapping studies for starch traits have traditionally identified regions that span 10-20 cM, which may represent as many as 20 million bases. Association analysis, in contrast, identified a suite of polymorphisms within a few thousand bases in this study. This is a substantial increase in resolution. The ability for association tests to locate possible regions within starch genes that may affect a modestly heritable trait such as starch demonstrates their usefulness for other such quantitative traits of importance that indicate modest heritability. A suite of polymorphisms has been identified that were in LD with one another and have associated with either different levels of storage products in the kernel, or in the makeup of the starch itself. It is possible that the specific genes showing association with a particular phenotype have alterations in enzyme activity. This study has supplied breeders with a set of high-resolution markers for a set of six starch genes. It also provided a wealth of candidate polymorphisms that can be further studied by molecular biologists and biochemists.

Sh1 and *Sh2* exhibited the most robust associations, and it is encouraging that they are also two of the three genes in this pathway that still have substantial molecular diversity. The discovery of a lack of diversity caused by selection throughout the history of maize improvement seen in roughly half of the six genes studied indicates that more work could be done in introgressing alleles from maize's wild relatives, the teosintes (Whitt *et al.*, 2002).

Because the set of 97 maize lines were chosen to include as much diversity as possible, including a good proportion of tropical origin, introgression of tropical germplasm in an attempt to fulfill the need for diversity in the starch pathway, rather than from the teosintes, may be inadequate for starch improvement.

Acknowledgements

I would like to express my appreciation to Ana Maria Ibanez-Carranza for her work in the viscoamylographic measurements. Thanks also to Lauren McIntyre at Purdue for the planting and care the West Lafayette, IN replication, and to Torbert Rocheford for the two Urbana, IL replications. Finally, thanks to Jason Dinges and Martha James for looking into *Su1* for effects. This work was supported by a grant from the National Science Foundation **DBI-9872631**, and the United States Department of Agriculture.

Literature Cited

Bae, J. M., M. J. Giroux and L. C. Hannah (1990). "Cloning and characterization of the *Brittle-2* gene of maize." *Maydica* **35**: 317-322.

Ball, S., H. P. Guan, M. G. James, A. M. Myers, P. L. Keeling, G. Mouille, A. Buleon, P. Colonna and J. Preiss (1996). "From glycogen to amylopectin: A model for the biogenesis of the plant starch granule." *Cell* **86**: 349-352.

Beckles, D. M., A. M. Smith and T. ap Rees (2001). "A cytosolic ADP-glucose pyrophosphorylase is a feature of graminaceous endosperms, but not of other starch-storing organs." *Plant Physiology* **125**(2): 818-827.

Berke, T. G. and T. Rocheford (1995). "Quantitative Trait Loci for Flowering, Plant and Ear Height, and Kernel Traits in Maize." *Crop Science* **35**: 1542-1549.

Bhave, M. R., S. Lawrence, C. Barton and L. C. Hannah (1990). "Identification and molecular characterization of *Shrunken-2* cDNA clones of maize." *Plant Cell* **2**: 581-588.

Boyer, C. D., P. A. Damewood and G. L. Matters (1980). "Effect of Gene Dosage at High Amylose Loci on the Properties of the Amylopectin Fractions of the Starches." *Starke* **32**(7): 217-222.

Chourey, P. S. and O. E. Nelson (1976). "Enzymatic Deficiency Conditioned by Shrunken 1 Mutations in Maize." *Biochemical Genetics* **14**(11-1): 1041-1055.

Clancy, M., V. Vasil, L. C. Hannah and I. K. Vasil (1994). "Maize Shrunken-1 Intron and Exon Regions Increase Gene- Expression in Maize Protoplasts." *Plant Science* **98**(2): 151-161.

Deatherage, W. L., M. M. Macmasters, M. L. Vineyard and R. P. Bear (1954). "A Note on Starch of High Amylose Content from Corn with High Starch Content." *Cereal Chemistry* **31**(1): 50-53.

Doehlert, D. C. and T. M. Kuo (1990). "Sugar metabolism in developing kernels of starch-deficient endosperm mutants of maize." *Plant Physiology* **92**: 990-994.

Fisher, D. K., C. D. Boyer and L. C. Hannah (1993). "Starch branching enzyme II from maize endosperm." *Plant Physiology* **102**: 1045-1046.

Fisher, D. K., M. Gao, K. N. Kim, C. D. Boyer and M. J. Guiltinan (1996). "Allelic analysis of the maize amylose-extender locus suggests that independent genes encode starch-branching enzymes LLa and LLb." *Plant Physiology* **110**(2): 611-619.

Gao, M., D. K. Fisher, K. N. Kim, J. C. Shannon and M. J. Guiltinan (1996). "Evolutionary conservation and expression patterns of maize starch branching enzyme I and IIb genes suggests isoform specialization." *Plant Molecular Biology* **30**(6): 1223-1232.

Goldman, I. L., T. Rocheford and J. W. Dudley (1993). "Quantitative trait loci influencing protein and starch concentration in the Illinois Long Term Selection maize strains." *Theoretical and Applied Genetics* **87**: 217-224.

Greene, T. W. and L. C. Hannah (1998). "Enhanced stability of maize endosperm ADP-glucose pyrophosphorylase is gained through mutants that alter subunit interactions." *Proceedings of the National Academy of Sciences of the United States of America* **95**(22): 13342-13347.

Guan, H. P., T. Kuriki, M. Sivak and J. Preiss (1995). "Maize branching enzyme catalyzes synthesis of glycogen-like polysaccharide in *glgB*-deficient *Escherichia coli*." *Proceedings of the National Academy of Sciences of the United States of America* **92**: 964-967.

Hallauer, A. R., Ed. (2001). Specialty Corns. Boca Raton, CRC Press LLC.

Hedman, K. D. and C. D. Boyer (1982). "Gene dosage at the *amylose-extender* locus of maize: Effects on the levels of starch branching enzymes." *Biochemical Genetics* **20**(5/6): 483-492.

Huang, D. Y. and A. Y. Wang (1998). "Purification and characterization of sucrose synthase isozymes from etiolated rice seedlings." *Biochemistry and Molecular Biology International* **46**(1): 107-113.

James, M. G., D. S. Robertson and A. M. Myers (1995). "Characterization of the Maize Gene Sugary1, a Determinant of Starch Composition in Kernels." *Plant Cell* **7**(4): 417-429.

Kim, K. N., D. K. Fisher, M. Gao and M. J. Guiltinan (1998). "Molecular cloning and characterization of the amylose-extender gene encoding starch branching enzyme IIB in maize." *Plant Molecular Biology* **38**(6): 945-956.

Knowler, W. C., R. C. Williams, D. J. Pettitt and A. G. Steinberg (1988). "*Gm*$^{3;5,13,14}$ and Type 2 Diabetes Mellitus: An Association in American Indians with Genetic Admixture." *American Journal of Human Genetics* **43**: 520-526.

Lander, E. S. and N. J. Schork (1994). "Genetic Dissection of Complex Traits." *Science* **265**: 2037-2048.

Laughlin, M. J., S. E. Chantler and T. W. Okita (1998). "N- and C-terminal peptide sequences are essential for enzyme assembly, allosteric, and/or catalytic properties of ADP- glucose pyrophosphorylase." *Plant Journal* **14**(2): 159-168.

Pan, D. (2000). Starch synthesis in maize. Carbohydrate Reserves in Plants - Synthesis and Regulation. A. K. Gupta and N. Kaur. Amsterdam, Elsevier. **26:** 125-146.

Prioul, J. L., S. Pelleschi, M. Sene, C. Thevenot, M. Causse, D. de Vienne and A. Leonardi (1999). "From QTLs for enzyme activity to candidate genes in maize." *Journal of Experimental Botany* **50**(337): 1281-1288.

Pritchard, J. K., M. Stephens and P. Donnelly (2000a). "Inference of population structure using multilocus genotype data." *Genetics* **155**: 945-959.

Pritchard, J. K., M. Stephens, N. A. Rosenberg and P. Donnelly (2000b). "Association mapping in structured populations." *American Journal of Human Genetics* **67**: 170-181.

Remington, D. L., J. M. Thornsberry, Y. Matsuoka, L. M. Wilson, S. R. Whitt, J. F. Doebley, S. Kresovitch, M. M. Goodman and E. S. Buckler, IV (2001). "Structure of linkage disequilibrium and phenotypic associations in the maize genome." *Proceedings of the National Academy of Sciences of the United States of America* **98**(20): 11479-11484.

Séne, M., M. Causse, C. Damerval, C. Thevenot and J. L. Prioul (2000). "Quantitative trait loci affecting amylose, amylopectin and starch content in maize recombinant inbred lines." *Plant Physiology and Biochemistry* **38**(6): 459-472.

Shure, M., S. Wessler and N. Fedoroff (1983). "Molecular-Identification and Isolation of the Waxy Locus in Maize." *Cell* **35**(1): 225-233.

Sidebottom, C., M. Kirkland, B. Strongitharm and R. Jeffcoat (1998). "Characterization of the difference of starch branching enzyme activities in normal and low-amylopectin maize during kernel development." *Journal of Cereal Science* **27**(3): 279-287.

Smith, A. M. (2001). "The Biosynthesis of Starch Granules." *Biomacromolecules* **2**: 335-341.

Stark, D. M., K. P. Timmerman, G. F. Barry, J. Preiss and G. M. Kishore (1992). "Regulation of the amount of starch in plant tissues by ADP glucose pyrophosphorylase." *Science* **258**: 287-292.

Tenaillon, M. I., M. C. Sawkins, A. D. Long, R. L. Gaut, J. F. Doebley and B. S. Gaut (2001). "Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.)." *Proceedings of the National Academy of Sciences of the United States of America* **98**(16): 9161-9166.

Thornsberry, J. M., M. M. Goodman, J. F. Doebley, S. Kresovitch, D. Nielson and E. S. Buckler, IV (2001). "*Dwarf8* polymorphisms associate with variation in flowering time." *Nature Genetics* **28**: 286-289.

Troyanskaya, O., M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein and R. B. Altman (2001). "Missing value estimation methods for DNA microarrays." *Bioinformatics* **17**(6): 520-525.

Tsai, C. Y. and O. Nelson (1966). "Starch-deficient maize mutant lacking adenosine diphosphate glucose pyrophosphorylase activity." *Science* **151**: 341-343.

Vasil, V., M. Clancy, R. J. Ferl, I. K. Vasil and L. C. Hannah (1989). "Increased Gene-Expression by the 1st Intron of Maize Shrunken-1 Locus in Grass Species." *Plant Physiology* **91**(4): 1575-1579.

White, J. A., J. Todd, T. Newman, N. Focks, T. Girke and e. al. (2000). "A New Set of Arabidopsis Expressed Sequence Tags from Developing Seeds. The Metabolic Pathway from Carbohydrates to Seed Oil." *Plant Physiology* **124**: 1582-1594.

Whitt, S. R., L. M. Wilson, M. I. Tenaillon, B. S. Gaut and E. S. Buckler, IV (2002). "Genetic diversity and selection in the maize starch pathway." *PNAS* **99**(20): 12959-12962.

Table 1. Maize inbred lines surveyed.

Non Stiff Stalk					
38-11	CM7	H95	Ky21 *	NC260 *	T8
A441-5	D940Y *	H99	M162W *	ND246	Va26
A554	EP1 *	HP301	Mo17 *	Oh43 *	U267Y
A619	F2 *	I137TN	Mo24W	Oh7B	W117HT
B37 *	F2834T	I205 *	MS153	PA91 *	W153R *
B97 *	F44	IDS28 *	N28HT *	SA24	W64A
C103	F7	K55	NC250	SG18	W182B
CI187-2 *	GT112	Ki44	NC258	T232 *	WF9
Stiff Stalk					
A632	B104	B68	B84	CM174	N192
B103 *	B14A *	B73 *	CM105	CMV3	SC213
Tropical/Semi-Tropical					
A272 *	CML277	CML91	Ki43	NC320	Q6199
A6 *	CML281	I-29	M37W	NC338	SC55
CML10	CML287	Ki11	NC296	NC348 *	Tx601 *
CML247	CML333 *	Ki9 *	NC298	NC350	Tzi10
CML254 *	CML5	Ki21 *	NC300	NC352	Tzi18
CML258 *	CML61	Ki3	NC304	NC354	Tzi8
CML261					

* Twenty-nine inbred lines initially amplified across genes *Ae1, Bt2, Sh1, Sh2, Su1*, and *Wx1*.

Table 2. Significant regions from the complete gene alignments for sampling.

Gene	Kernel Composition PCA	Starch Pasting PCA
Ae1	205-337, 3088-4182, 4475-4694, **6308-6934**	205-337, **1288-1935**, **8128-9181**
Bt2	**756-1147**	n/a
Sh1	**628-1427, 3423-5600**	**628-1427**
Sh2	**3588-4096**	310-932, **3588-4096**
Su1	65-538, 5302, 10340-10869	**1011-1530, 7894-8563**
Wx1	2530-3100, **3132-3694,** 3715-4014, 4495-4597	74-916, **1385-1948**

Bold typeface denotes regions that were sampled in the full set of 97 maize lines, while regular typeface denotes regions in the initial survey of 29 lines that showed possible associations, but have not been sampled further in the larger set of taxa.

Table 3. Averages for kernel composition traits per replication.

Trait	Homestead, FL Winter 1998-1999	West Lafayette, IN Summer 2000	Clayton, NC Summer 2001 rep1	Clayton, NC Summer 2001 rep3
Kernel starch	0.626 ± 0.046	0.460 ± 0.060	0.496 ± 0.059	0.528 ± 0.048
Kernel moisture	0.088 ± 0.002	0.088 ± 0.002	0.073 ± 0.007	0.080 ± 0.006
Kernel oil	0.060 ± 0.012	0.055 ± 0.016	0.072 ± 0.014	0.068 ± 0.012
Kernel protein	0.105 ± 0.016	0.129 ± 0.018	0.122 ± 0.016	0.120 ± 0.016

Table 4. Phenotype averages for starch pasting traits and viscosity measurements.

Starch Pasting Trait	Homestead, FL Winter 1998-1999
Amylose content	0.211 ± 0.015
Pasting temperature (°C)	64.9 ± 1.3
Peak temperature (°C)	82.2 ± 3.0
Peak time (s)	216.3 ± 18.2
Peak viscosity (Pa.s) [a]	0.423 ± 0.080
Trough viscosity (Pa.s)	0.283 ± 0.051
Hot paste viscosity (Pa.s)	0.285 ± 0.050
Cool paste viscosity (Pa.s)	0.515 ± 0.096
Breakdown (Pa.s)	0.141 ± 0.049
Setback (Pa.s)	0.233 ± 0.051
Consistency (Pa.s)	0.230 ± 0.051

[a] Pa.s., pascal second, the standard unit of dynamic viscosity.

Table 5. Results of the kernel composition principle component analysis.

Principle Component Factor	Overall Eigenvector Value [b]				Proportion of Variance	Cumulative Variance	Interpretation [a]
	Oil	Protein	Starch	Moisture			
1	+ 1.21	+ 1.35	- 1.28	-/+ 0.73	0.34	0.34	Oil and protein accumulation versus starch
2	+ 1.61	- 1.37	+ 0.51	- 0.56	0.13	0.47	Oil production versus protein production
3	-/+ 0.71	-/+ 0.48	-/+ 0.90	-/+ 1.34	0.08	0.55	Genotype × Environment effects

[a] Interpretation of the trait(s) driving each principle component were determined by summing the absolute values of the eigenvectors over all environments for each phenotypic trait [b]. The greater the absolute value for a trait, the more weight that trait was given in the amount of its contribution to driving the variance seen for a particular PCA factor. A plus (+) sign indicates all eigenvectors were positive values, and a minus (-) sign indicates all eigenvectors were negative values, while both (-/+) indicate both positive and negative eigenvector values were seen among the field environments.

Table 6. Results of the starch pasting properties principle component analysis.

Starch Pasting Principle Component	Proportion of Variance	Cumulative Variance	Interpretation
PC-1	0.34	0.34	Highly correlated viscosity (v) measurements: consistency, setback, cool paste v, hot paste v, peak v, trough v.
PC-2	0.13	0.47	Pasting temperature vs. peak temperature and time
PC-3	0.08	0.55	Pasting temperature vs. amylose
PC-4	0.06	0.61	Amylose and pasting temperature

Table 7. Overall gene results of the association analyses.

	Kernel Composition PCA Factor [a]						Starch Pasting PCA Factor [b]							
Gene	1		2		3		1		2		3		4	
Ae1											**		**	**
Bt2				**										
Sh1	**	*				**	*							**
Sh2	**					**					**	**		
Su1	*													
Wx1	*			*										

*, ** = ($P \leq 0.10, 0.05$), respectively.

White columns denote ANOVA, no population structure.

Grey columns denote logistic regression analysis including estimates of population structure.

[a, b] Refer to Table 5 and 6, respectively, which detail specifics of each individual PCA factor.

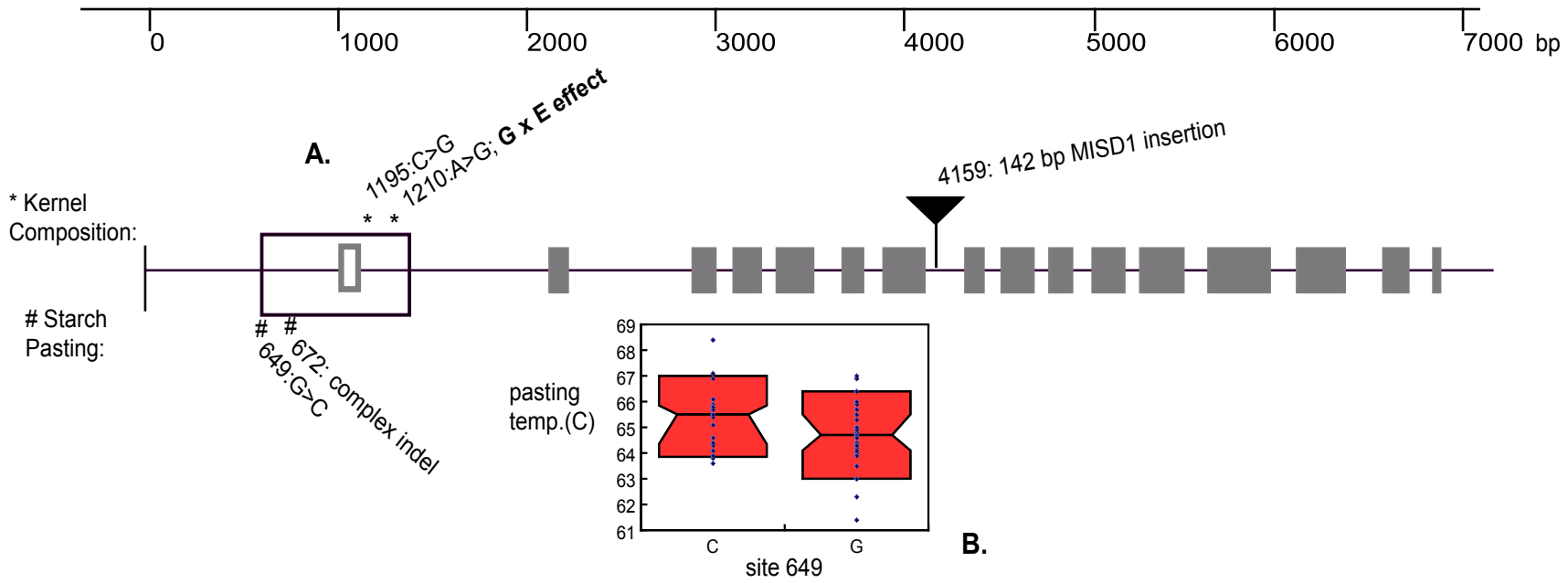


Figure 1. Genetic structure of *Sh1* sequenced from 29 maize taxa. The area outlined with a black box denotes the region sampled in the full set of 97 taxa. The gray, unfilled exon denotes *Sh1*'s noncoding exon one. Particular sites that show associations with kernel composition are highlighted (*) above the gene picture, along with a notable polymorphism found in *Sh1* (site 4159). Region A highlights the polymorphisms significant for a G x E effect (best site 1210) (logistic regression; $P < 0.05$). Sites in *Sh1* significant for starch pasting traits are highlighted (#) below the gene picture. Inset B shows the significant increase in pasting temperature seen in lines with a G>C transversion at site 649 (*t*-test; $P < 0.05$). Shaded areas within inset B denote the distribution of the data: the median for the data points is marked by the middle horizontal line; the upper and lower horizontal lines highlight the 10th and 90th percentiles, while the notch denotes the 25th and 75th percentiles. This is also the case for all insets shown for all other genes.

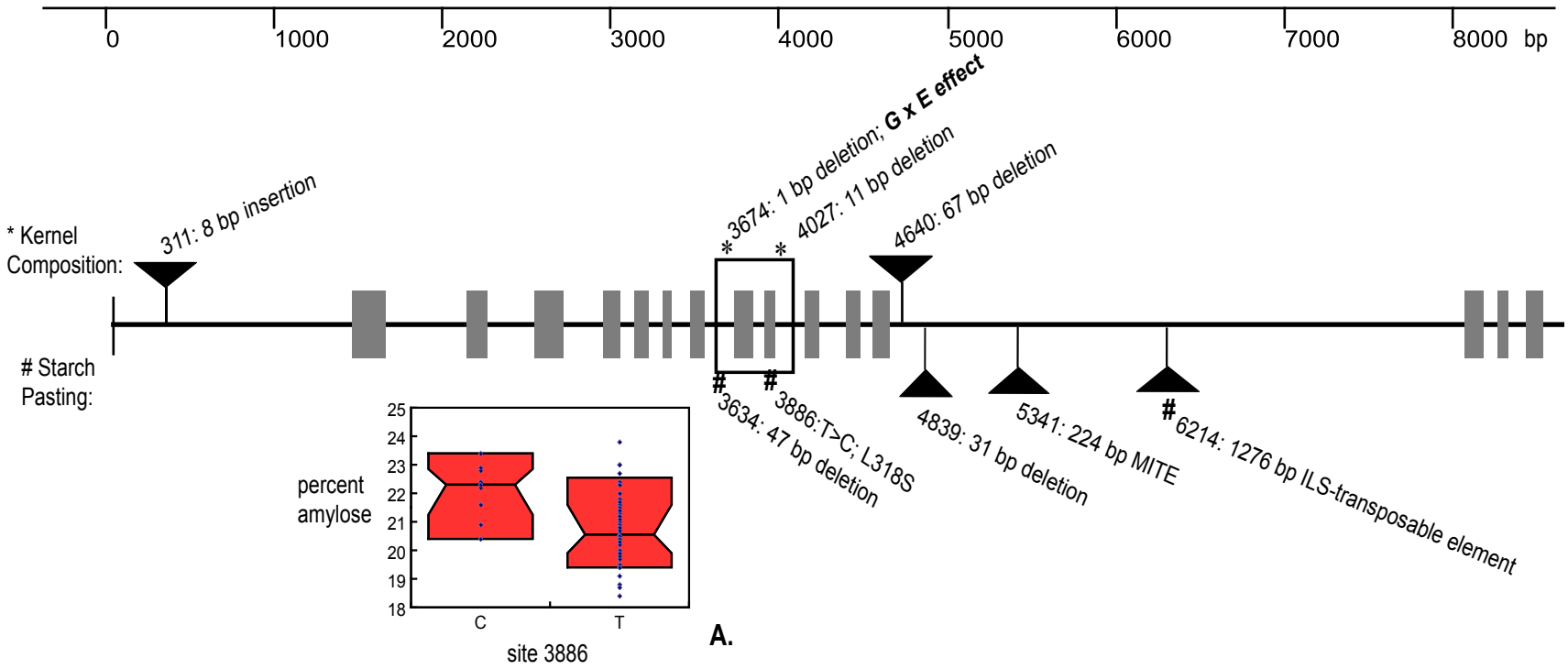


Figure 2. Genetic structure of *Sh2* sequenced from 29 maize taxa. Only coding exons are shown. Two regions in *Sh2* were scored in the total 97 taxa; the region outlined by the black box was sequenced, while the region containing the 1276 bp transposon was scored by length of PCR product. Two highlighted polymorphisms (*) were significant with kernel composition traits, showing a G x E effect (best site: 3674) (logistic regression; $P < 0.036$). Three highlighted polymorphisms (#) were significant with amylose content. Inset A shows the significant increase in amylose content seen in lines containing a transition to a cytosine at site 3886 (*t*-test; $P < 0.007$). This mutation causes a lysine to serine substitution at residue 318 in the predicted SH2 protein.

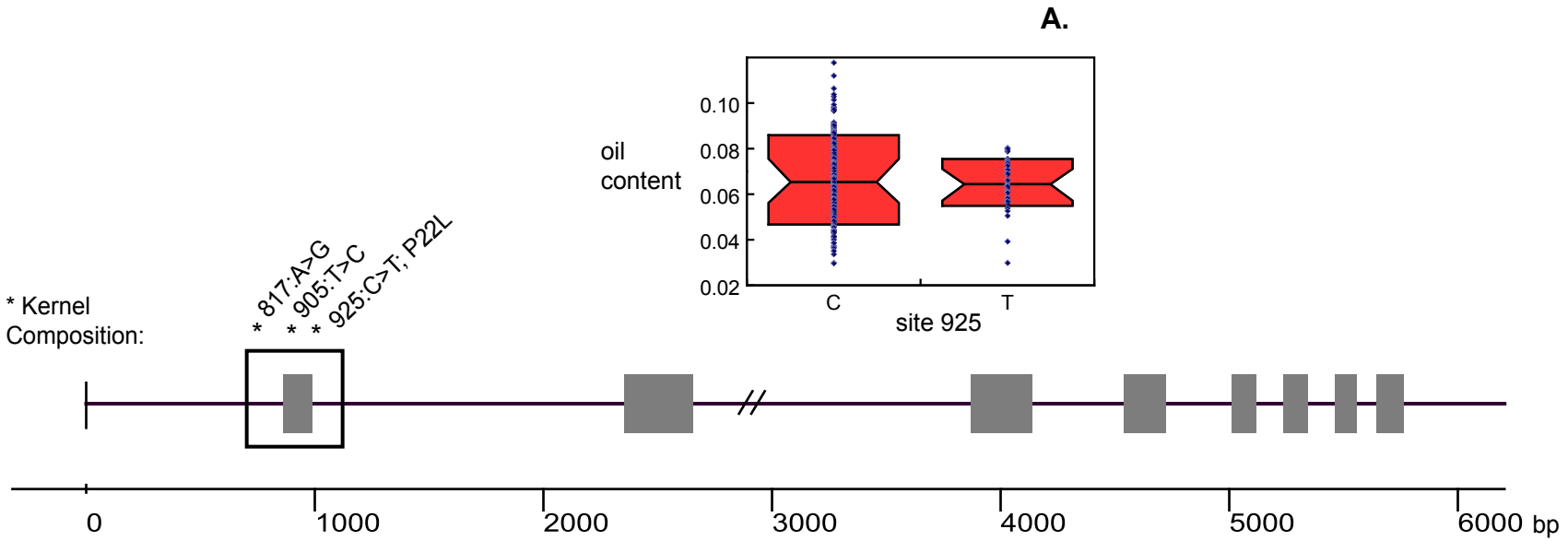


Figure 3. Genetic structure of *Bt2* sequenced from 29 maize taxa. One region was scored in the total 97 taxa, outlined by a black box. Sites significant for kernel composition are highlighted (*) above the gene picture. Inset A shows the reduced variance seen in oil content (*F*-test; $P < 0.002$) in lines containing the C>T transition mutation at site 925, which causes a proline to leucine amino acid change at residue 22 in the BT2 predicted protein. Note: the last exon of *Bt2* (exon 9) was not sequenced. Dashed lines near position 3000 (//) denote an area not sequenced due to a highly repetitive 250 bp stretch.

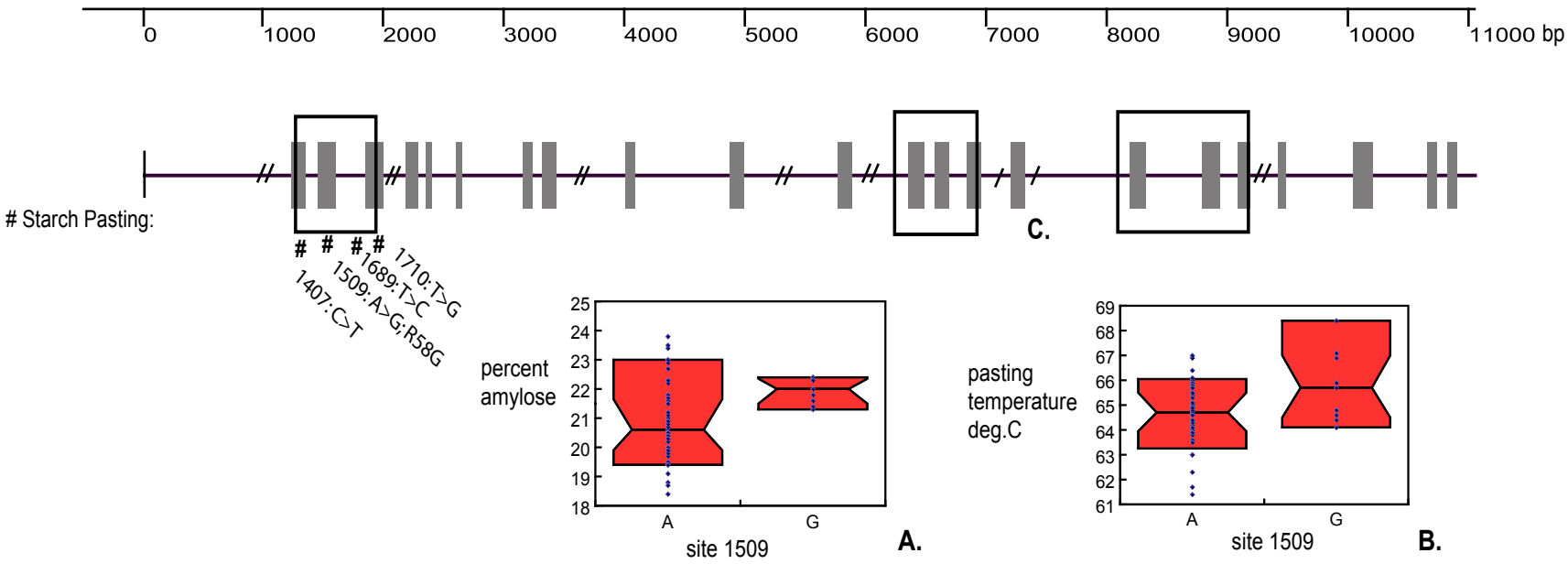


Figure 4. Genetic structure of *Ae1* sequenced from 29 maize taxa. Three areas of *Ae1* were sampled in the total 97 taxa, and are denoted by a black outline. Sites significant for starch pasting are highlighted (#) below the gene picture. Site 1509, in particular, associated with higher levels of amylose (*t*-test; $P < 0.0001$) (inset A) and higher pasting temperatures (*t*-test; $P < 0.02$) (inset B) in lines containing the A>G transition mutation. This mutation causes an arginine to glycine substitution at residue 58 in the AE1 predicted protein. Note: diagonally dashed lines (//) in the gene picture are intronic areas not sequenced in the initial 29-line subset, excepting exon 15 (4C).

APPENDIX

Note: This work was funded in part by federal funds from the United States Department of Agriculture and is exempt from copyright law (Title 17, Chapter 1, Section 105, concerning public domain). "Structure Of Linkage Disequilibrium And Phenotypic Associations In The Maize Genome" by Remington, *et al.* was originally published in *Proceedings of the National Academy of Sciences*. September 25, 2001. Volume 98(20): pages 11479-11484.

STRUCTURE OF LINKAGE DISEQUILIBRIUM AND PHENOTYPIC ASSOCIATIONS IN THE MAIZE GENOME

David L. Remington*, Jeffry M. Thornsberry*, Yoshihiro Matsuoka[†], Larissa M. Wilson*, Sherry R. Whitt*, John Doebley[†], Stephen Kresovich[‡], Major M. Goodman[§], and Edward S. Buckler IV*

Departments of *Genetics and [§]Crop Science, North Carolina State University, Raleigh, NC 27695-7614.

[†] Department of Genetics, University of Wisconsin, Madison, WI 53706.

[‡] Department of Plant Breeding, Cornell University, Ithaca, NY 14853.

Abstract

Association studies based on linkage disequilibrium (LD) can provide high resolution for identifying genes that may contribute to phenotypic variation. We report patterns of local and genome-wide LD in 102 maize inbred lines representing much of the worldwide genetic diversity used in maize breeding, and address its implications for association studies in maize. In a survey of six genes, we found that intragenic LD generally declined rapidly with distance ($r^2 < 0.1$ within 1500 bp), but rates of decline were highly variable among genes. This rapid decline probably reflects large effective population sizes in maize during its evolution and high levels of recombination within genes. A set of 47 simple sequence repeat (SSR) loci showed stronger evidence of genome-wide LD than did single-nucleotide polymorphisms (SNPs) in candidate genes. LD was greatly reduced but not eliminated by grouping lines into three empirically determined subpopulations. SSR data also supplied evidence that divergent artificial selection on flowering time may have played a role in generating population structure. Provided the effects of population structure are effectively controlled, this research suggests that association studies show great promise for identifying the genetic basis of important traits in maize with very high resolution.

Introduction

In plant genetic studies, recombinant inbred lines have been very successful in mapping quantitative trait loci (QTLs) to 10-30 cM regions (Alpert and Tanksley 1996; Stuber et al. 1999), but association studies based on linkage disequilibrium (LD) may allow

identification of the actual genes represented by QTLs. Only polymorphisms with extremely tight linkage to a locus with phenotypic effects are likely to be significantly associated with the trait in a randomly mating population, providing much finer resolution than genetic mapping. Association methods have been especially important for studying the genetic basis of human diseases, for which controlled genetic experiments are not feasible. However, these methods also have great potential for resolving individual genes responsible for QTLs (Lai et al. 1994; Laitinen et al. 1997; Slatkin 1999).

The resolution of association studies in a test sample is dependent on the structure of LD across the genome. LD, or the correlation between alleles at different sites, is generally dependent on the history of recombination between polymorphisms. However, factors such as genetic drift, selection within populations, and population admixture can also cause LD between markers and traits. (Following common practice (Falconer and Mackay 1996; Weir 1996), we refer to gametic phase disequilibrium as LD whether or not it is caused by linkage.) Since many factors affect LD, its genomic structure in particular crop plants must be empirically determined before association studies can be applied. In maize, for example, divergent selection for adaptive traits such as time of maturation in different regions may have created LD among chromosomal regions containing major genes for these traits.

Our goal in this study was to evaluate patterns of LD among 102 maize inbred lines representing the diversity of both temperate and tropical sources and address its implications for association studies in maize. Our first objective was to evaluate the rates at which LD decays within genes, using DNA sequence data from six candidate genes for important agronomic traits. Secondly, to explore the extent LD between unlinked sites, we evaluated LD between sites in different candidate genes and between 47 simple sequence repeat (SSR)

loci. Finally, we performed a number of statistical tests on the SSR LD data and SSR-trait associations in order to identify mechanisms by which selection on agronomic traits may have shaped LD in the maize genome. This evaluation of linkage disequilibrium across maize breeding lines will show that association studies could be developed for maize to map quantitative traits at very high resolution.

Materials and Methods

Plant Materials

One hundred two inbred maize lines, representing a broad cross-section of breeding germplasm from temperate and tropical regions, were used in this study. These include 53 U.S. lines, 7 European and Canadian lines, and 42 tropical/semitropical (ST) lines. Thirteen of the combined U.S.-European-Canadian lines were primarily Iowa Stiff Stalk Synthetic in origin (SS), and the remaining 47 lines were non-stiff-stalk (NSS). The lines were as follows: *ST lines*: A6, A272, A441-5, B103, CML5, CML10, CML61, CML91, CML247, CML254, CML258, CML261, CML277, CML281, CML287, CML333, D940Y, F2834T, I137TN, KUI3, KUI11, KUI21, KUI43, KUI44, KUI2007, M37W, M162W, NC296, NC298, NC300, NC304, NC338, NC348, NC350, NC352, NC354, Q6199, SC213R, Tzi8, Tzi10, Tzi18, U267Y. *SS lines*: A632, B14A, B37, B68, B73, B84, B104, CM105, CM174, MS153, N28Ht, N192, NC250. *NSS lines*: 38-11, A554, A619, B97, C103, CI187, CM7, CMV3, EP1, F2, F7, F44, Gt112, H95, H99, HP301, I29, I205, Ia2132, IDS28, Il14H, Il101, Il677a, K55, Ky21, Mo17, Mo24W, NC258, NC260, NC320, ND246, Oh43, Oh7B,

P39, Pa91, SA24, SC55, Sg18, T232, T8, Tx601, Va26, W64A, W117Ht, W153R, W182B, Wf9. Additional information on these lines is included in Table 4, published as supplemental data on the PNAS web site at www.pnas.org.

Field Data

Field tests were established at two sites, near Clayton, NC and Homestead, FL. A number of phenological and morphological traits were measured over three field seasons during 1998 and 1999 at one or both sites, for a total of five study environments. Details of test design and trait measurements have been described elsewhere (Thornsberry et al. 2001). For this report, days to pollen (DPoll) and days to silking (DSilk) were selected as measures of flowering time, and ear height (EarHt) and total plant height (PlHt) were selected as measures of plant morphology.

Candidate Gene Sequence Data

DNA sequence data were obtained from coding regions and flanking sequence of four genes: *indeterminate1* (*id1* – chromosome 1, 175.0 cM), *teosinte branched1* (*tb1* – chromosome 1, 197.6 cM), *dwarf8* (*d8* – chromosome 1, 198.5 cM), and *dwarf3* (*d3* – chromosome 9, 62.7 cM). These are considered candidate genes for variation in plant height and/or flowering time, based on mutant phenotypes and chromosomal locations near major QTLs. Sequence data was also obtained for 32 lines for two additional genes: *shrunken1* (*sh1* – chromosome 9, 36.4 cM) and *sugary1* (*su1* – chromosome 4, 60.2 cM). Gene

fragments were PCR amplified using primers designed from published sequences. Sequence data were obtained directly from PCR products or from pools of two to four clones of PCR products. Sequence chromatogram files were assembled into contigs using SeqMan (DNAStar), and consensus sequences were edited manually to resolve discrepancies. Consensus sequences for all lines were aligned using the Clustal alignment option in MegAlign (DNAStar), with further manual alignment. Polymorphisms appearing in only one or two lines were re-checked on chromatograms in order to distinguish true polymorphisms from probable polymerase or scoring errors. Well over 1.5 megabases of contiged sequence data were collected.

SSR Marker Data

Development and scoring of SSR markers has been described elsewhere by Matsuoka et al. (in review). We used data from 47 highly polymorphic loci with a mean of 6.85 alleles per locus (range 2-16 alleles). These SSRs have been found to contain frequent indels outside of repeat units and are not evolving in a stepwise manner (Matsuoka et al. in review). Map positions for all candidate genes and SSRs were based on the Pioneer Composite 1999 linkage maps obtained from the MaizeDB website (www.agron.missouri.edu/query.html).

Statistical Analyses

LD between pairs of sites in candidate genes (both SNPs and insertion-deletion polymorphisms, or indels) and in SSRs was evaluated using the software package TASSEL (available at www.statgen.ncsu.edu/~buckler/). Contiguous indel sites showing identical patterns of variation were treated as a single polymorphism. LD was estimated using standardized disequilibrium coefficients (D') per Hedrick (Hedrick 1987), and squared allele-frequency correlations (r^2) per Weir (Weir 1996) for pairs of loci. D' is affected solely by recombination and not by differences in allele frequencies between sites. r^2 is also affected by differences in allele frequencies at the two sites, and is therefore a better measure of potential allele-trait associations than D'. Only sites with a frequency of at least 0.10 for the rarer allele were included because D' and r^2 have large variances with rare alleles. The probabilities of obtaining LD estimates at least as extreme as those observed under a hypothesis of linkage equilibrium (*P*-values) were calculated using Fisher's exact test (Fisher 1935) for site pairs with two alleles each. For site pairs with more than two alleles at one or both loci, empirical *P*-values were obtained by repeatedly permuting the alleles at one of the loci as described by Weir (Weir 1996). Complete LD data for pairs of candidate gene polymorphisms and SSR loci are included in Tables 5-7, included as supplemental data on the PNAS web site.

Decay of LD with distance in base pairs (bp) between sites within the same candidate locus was evaluated by nonlinear regression (PROC NLIN in SAS software) (SAS Institute 1999). The expected value of r^2 under drift-recombination equilibrium is $E(r^2) = 1/(1 + C)$, where N is the effective population size, c is the recombination fraction between sites, and $C = 4Nc$ (Sved 1971). With a low level of mutation and an adjustment for sample

size n, the expectation becomes: $\mathrm{E}(r^2) = \left[\frac{10+C}{(2+C)(11+C)}\right]\left[1+\frac{(3+C)(12+12C+C^2)}{n(2+C)(11+C)}\right]$ (Hill and Weir 1988).

The nonlinear models based on each of these expectations contain a single coefficient, which is the least-squares estimate for $4Nc$ per bp distance between sites. Distances were weighted to adjust for indels by averaging the number of bp separating the sites across all lines for which both sites were scored. Several factors may reduce precision or create bias in the model estimates, including non-independence of linked site pairs and non-equilibrium populations (Weir and Hill 1986). Consequently, the models may not provide useful estimates of $4Nc$, but are nonetheless useful for characterizing the rate of LD decay. The distribution of D' and r^2 values for pairs of sites in different candidate loci was evaluated for *d8*, *tb1*, *id1*, and *d3*.

SSR haplotypes were used to evaluate population structure associated with the ST, NSS and SS subpopulations. Lines were also subdivided based on data from the 47 SSRs using a model-based approach with the software package STRUCTURE (Pritchard et al. 2000a). Several runs were made using various sets of initial parameter values for 2, 3, 4, and 5 subpopulations. The run producing the highest log likelihood for the observed data was obtained when the number of subpopulations was set at 3, and was used to produce a new set of model-based subpopulations, designated ST_M, NSS_M and SS_M. For analyses of structure within subpopulations, we assigned each line to the subpopulation with the largest estimated admixture contribution. Overall, individual-locus, and pairwise estimates of the correlation of alleles within subpopulations (F_{ST}) for both the origin-based and model-based

groupings were calculated using an AMOVA approach in Arlequin version 2.0 (Weir 1996; Schneider et al. 2000).

The significance of the overall matrix of pairwise LD *P*-values among all 47 SSR loci was evaluated in TASSEL by repeatedly permuting the matrix of SSR genotypes at each locus, and computing pairwise LD *P*-values for each permuted data set as described above. The numbers of site pairs with LD *P*-values less than threshold values of 0.01, 0.001, and 0.0001 were counted for the observed data and for each permuted data set, and the total *P*-value for the observed data was calculated as the proportion of permuted data sets with higher counts than the observed data.

Associations of individual SSR alleles with trait values across all five study environments were evaluated in TASSEL by simple regression. *P*-values were obtained from the *F*-value of effects of each allele on trait values. The *P*-value of the most strongly associated allele (regardless of frequency) was used as a measure of the SSR-trait association for the locus. Differences in these measures among traits were evaluated using SAS (PROC GLM). The effects of individual-locus F_{ST} values on SSR-trait associations were also evaluated using PROC GLM. Simple linear correlations between SSR allele-trait associations and the distribution of SSR LD were evaluated using SAS (PROC CORR).

Results

Linkage Disequilibrium between Candidate Locus Polymorphisms

LD between pairs of sites within the six candidate loci is summarized in Figure 1a-f. A nonlinear model of LD decay that incorporated mutation (Hill and Weir 1988) explained 9.6%-37.5% of the variance in r^2 for all loci except *su1*. The model incorporating mutation explained more of the variance in r^2 than did a recombination-drift model (Sved 1971) for *d3*, *id1*, *tb1*, and *sh1*. At *su1*, only the recombination-drift model explained more variation in r^2 than simply fitting a mean. The predicted value of r^2 declined to 0.1 or less within 1500 bp at *d3*, *id1*, *tb1*, and *sh1*. At *su1*, on the other hand, the predicted value of r^2 remained greater than 0.4 for more than 7000 bp, and *d8* showed an intermediate rate of decline. The degree of LD for sites a given distance apart was highly variable. Sites in strong LD with one another tended to occur in blocks, but pairs of sites in complete LD with each other often showed low LD with intervening sites as measured by both D' and r^2.

We also evaluated LD of inter-locus site pairs between the four loci that were scored for the entire set of 102 lines. Three contrasting levels of linkage could be evaluated: tightly linked loci (*tb1* with *d8*, which are ≈1cM apart on chromosome 1), loosely linked loci (*id1* with *tb1* and *d8*, which are ≈22cM apart), and unlinked loci (*d3* on chromosome 9 with the other 3 loci) (Table 1). Approximately 3.6% of site pairs were in significant LD at the comparisonwise 0.01 level. The pair of tightly linked loci showed by far the highest level of LD. This is due primarily to a large number of polymorphic sites within the same large insertion in the *d8* promoter, which are in LD with a cluster of sites in the 3′ untranslated region of *tb1*.

Population Structure

When we grouped the lines into the ST, NSS and SS subpopulations, the overall F_{ST} of 0.105 was highly significant, as were each of the three pairwise estimates of F_{ST} (Table 2). The pairwise comparisons show a low level of differentiation between the ST and NSS subpopulations, but the SS lines are much more highly diverged from the other two groups. The F_{ST} estimate for the three model-based subpopulations (ST_M, NSS_M and SS_M) estimated from STRUCTURE was only slightly higher at 0.121. All but 18 lines were predicted to have greater than 80% of their origin from one of the three inferred subpopulations in the highest-likelihood run. The model-based and origin-based subpopulations were in agreement for 88 of the 102 lines when each line was assigned to the subpopulation with the largest admixture proportion (see Table 4, supplemental data).

Linkage Disequilibrium between SSR Loci

LD was significant at a comparison-wise 0.01 level in nearly 10% of the SSR marker pairs when all lines were included in the analysis, or nearly 10 times the number expected by chance (Table 3). This is nearly three times the percentage of intergenic site pairs that were in LD at this level. The proportion of sites in significant LD was reduced substantially within individual model-based subdivisions. Some of this reduction could be due to reduced power to detect LD with fewer lines. To test for this possibility, we evaluated the percent of locus pairs showing significant LD in sets of 1000 randomly chosen subpopulations, with each set containing the same numbers of lines as the original subpopulations. The observed percentages of LD in the random subpopulations were substantially higher than those in the origin-based and model-based ST and NSS subpopulations (Table 3), suggesting that the

subpopulations themselves explain much of the LD. Nevertheless, each of the subpopulations still shows an excess of significant LD values. When 100 randomly permuted datasets were generated for each subpopulation, none showed more than the observed number of significant LD values at the 0.01 or the 0.001 levels. The low number of pairs with significant LD within the SS lines thus may be merely the result of limited power to detect significant deviations with such a small number of lines, but may also reflect the random-mated origin of the SS lines (Labate et al. 2000).

SSR-Phenotype Associations

We wanted to examine whether selection for maturation time in different environments may have been a factor in generating population structure and LD between unlinked genomic regions. Between 34% and 64% of SSRs showed strong associations (P<0.01) with the four traits measured. The number of SSRs with strong trait associations for the two flowering time traits (DPoll and DSilk) directly related to maturation was significantly greater (P=0.0007) than for the two morphological traits (EarHt and PlHt). Fewer SSRs showed strong trait associations when only the NSS_M lines were used in the analysis (15% to 30%). Within the NSS_M lines, the difference between SSR-flowering time and SSR-morphological trait associations was not significant (P=0.17).

Next, we evaluated whether LD between SSRs was related to the strength of SSR-trait associations, which would be expected if selection on these traits helped generate population structure. SSR-trait associations for each of the four traits were correlated weakly but highly significantly (r = 0.11 to 0.16, P<0.0001) with LD. When the same

analysis was done using only the NSS_M lines, none of the SSR-trait associations were significantly correlated with LD.

Thirdly, we investigated whether selection on flowering time loci may have directly generated SSR LD. We compared flowering time associations for SSRs near known flowering time QTLs with those for the remaining SSRs. Twenty of the 47 SSR loci are within 20 cM of estimated map positions of flowering time QTLs in eight studies summarized in MaizeDB (Abler et al. 1991; Koester et al. 1993). The mean *P*-values of SSR-flowering time associations were not significantly different for these markers than for the remaining 27 SSRs.

Finally, we examined whether the SSRs showing strong associations with flowering time also showed greater levels of differentiation between subpopulations. We separately estimated overall and pairwise F_{ST} values for the model-based subpopulations for the 21 SSR loci showing strong flowering-time associations ($P \leq 0.001$) and the remaining 26 loci. Overall F_{ST} values were consistently higher for the loci showing strong flowering time associations (0.161 vs. 0.085), as were all pairwise values among the three subpopulations. Individual-locus F_{ST} values were significant predictors of SSR associations with DPoll (R^2=0.176, F=9.64, P=0.003) and DSilk (R^2=0.149, F=7.85, P=0.008) but not with EarHt (R^2=0.034, F=1.59, P=0.215) and PlHt (R^2=0.001, F=0.05, P=0.824).

Discussion

Decay of LD with Distance between Sites

We found that LD generally decayed rapidly with distance between sites within loci, but there was substantial variation among genes. In 4 of the 6 genes sampled, predicted r^2 values declined to less than 0.1 within 2000 bp, much less than the 50 kb predicted for the same degree of LD decay in humans (Koch et al. 2000). Recent studies in humans have shown that LD typically extends 60 kb in European populations, and may extend much farther (Koch et al. 2000; Moffatt et al. 2000; Reich et al. 2001). Only at *su1* did we find evidence that LD might persist at anywhere near these distances in maize. This may be caused in part by reduced recombination rates due to the location of *su1* near the centromere of chromosome 4. Selection can also maintain elevated LD in localized regions (Huttley et al. 1999), and may provide an explanation for the persistence of LD at *su1* and to some extent at *d8*. Both loci are candidate genes for traits that have been under strong artificial selection, *d8* for flowering time variation (Thornsberry et al. 2001), and *su1* for kernel sugar and starch levels (Buckler and Whitt, in preparation). LD appeared to decay rapidly at *tb1*, as has been reported previously (Wang et al. 1999), in spite of the selective sweep at this locus during maize domestication. The relatively poor fit of the nonlinear model with *tb1* and *su*1 may be due in part to the effects of strong selective episodes on the frequency and distribution of polymorphisms. In some cases, sites separated by 1 kb or more were in complete LD, but had low D' values (indicating recombination) with intervening sites. This reflects differences in the age and genealogy of the various mutations, and possibly the effects of gene conversion and admixture.

The unlinked candidate loci had extremely low levels of LD (r^2=0.024), and it was only modestly higher in one pair of loci 1cM apart. To determine whether this slightly elevated level of LD at 1cM is due to linkage or chance, sequencing of more genes and

much longer contiguous regions will be necessary to evaluate the variability of LD decay over intermediate distances. These results are in sharp contrast with those recently reported for Dutch dairy cattle, in which LD has been found to persist over distances of many centiMorgans (Farnir et al. 2000). LD has also been reported between loci as much as 4 cM apart in European human populations (Huttley et al. 1999). The population recombination parameter *C* is dependent on both effective population size (*N*) and recombination frequency (*c*) (Hudson 1987; Hey and Wakeley 1997). High recombination frequencies have been reported for several maize genes (Xu et al. 1995; Dooner and MartinezFerez 1997; Henry and Damerval 1997; Okagaki and Weil 1997). Other studies of recombination rates and levels of polymorphism in maize have found evidence of large population sizes as well, which suggests that the domestication bottleneck was either mild or of short duration (Eyre-Walker et al. 1998; Wang et al. 1999). Our average value for *C* from six loci was 0.0080. If the overall genomic value of ~1 x 10^{-8} for *c* in maize is used, this suggests a value of approximately 2 x 10^{5} for *N*, similar to estimates from sequence diversity at the *Adh1* locus by Eyre-Walker et al. (Eyre-Walker et al. 1998). This estimate would be biased upwards, however, if the recombination rate within the studied genes were abnormally high. If a much narrower set of lines had been chosen for this study, the rate of LD decay might have been substantially lower.

Candidate-Gene Polymorphisms *vs.* SSRs as Indicators of Genome-Wide LD

The level of genome-wide LD indicated by the SSRs is much higher than that shown by the candidate genes. This could be due to chance alone, as the small set of candidate

genes may happen to share relatively little evolutionary history. It may also reflect the fact that these SSRs were initially chosen because they differentiated between a small set of U.S. inbred lines. Another possibility is that a higher percentage of SSR mutations than SNPs arose during the development of regional maize subpopulations. Maize and its wild progenitor, *Z. mays* ssp. *parviglumis*, share many of the same single-nucleotide polymorphisms at a number of loci, including *adh1* (Eyre-Walker et al. 1998), *c1* (Hanson et al. 1996), and *tb1* (Wang et al. 1999), suggesting that SNP alleles tend to predate domestication. The high level of variability in the SSRs, however, suggests a high rate of mutation to new alleles (primarily indels rather than variation in repeat number), increasing the opportunity for unique length variants to have arisen in individual races during domestication (Matsuoka et al. in review). Consequently, the SSR polymorphisms may reveal the recent development of population structure in domesticated maize much better than SNPs.

Population Structure

In spite of the genome-wide LD revealed by the SSR loci, this broad cross-section of maize breeding material shows a fairly low degree of population structure. Much of the differentiation we did detect was due to the rather divergent nature of the SS lines. The domestication and breeding history of maize may explain the low level of differentiation between the ST and NSS groups. The NSS lines are primarily Corn Belt dents, a diverse group which originated from the crossing of northern flints and southern dents and appears to consist predominantly of southern dent genetic material (Doebley et al. 1988). The SS

lines were developed from only 16 inbred Corn Belt ancestors, and their divergence from the NSS and ST lines is primarily due to genetic drift.

The degree of LD is lower within subpopulations, but it is still significantly elevated. The extent of within-subpopulation structure in domesticated maize is undoubtedly affected by the admixture origin of the Corn Belt dents, and probably by assortative mating and selection for divergent combinations of traits.

Role of Selection in Generating LD

The population structure in maize appears to reflect the effects of selection on adaptive traits such as flowering time. SSR/phenotype associations and their relationship to population structure were stronger for flowering time than for correlated height traits. This suggests that divergent selection on flowering time may have had an important role in the development of regional variation in maize germplasm. The most plausible explanation for the observed SSR-trait-F_{ST} associations is that SSRs with allelic variants that happen to distinguish subpopulations are consequently associated with differences in flowering time among subpopulations as well. SSR-trait associations among these lines are unlikely to reflect actual linkage to flowering time loci, as SSRs located near identified flowering time QTLs do not show stronger flowering time associations than other SSRs. Selection would have to generate LD over large chromosomal blocks to be detected through linkage to such a limited set of SSRs. This would probably require severe population bottlenecks generated by extremely strong selection and/or epistasis (Wiehe and Slatkin 1998; Huttley et al. 1999;

Kohn et al. 2000). In maize, however, the region affected by selective sweep at *tb1*, a major domestication locus, does not encompass the entire gene (Wang et al. 1999).

The significant relationship between SSR LD and SSR-trait associations also appeared to be an effect of population structure. These relationships disappeared entirely when the analysis was limited to the NSS_M subpopulation. Elevated levels of SSR LD and SSR-flowering time associations, however, were apparent even within subpopulations, which suggests that assigning lines to subpopulations alone may not be adequate to control for non-functional LD. The STRUCTURE analysis predicted 18 lines to be substantially admixed (<80% composition from a single population). Pritchard et al. (Pritchard et al. 2000b) have developed a methodology that uses estimated subpopulation admixture proportions, not merely subpopulation assignments, to control for population structure in disease association studies. These methods have been adapted for quantitative traits and found useful for association testing in maize (Thornsberry et al. 2001). In the future, pedigree information should also be integrated with overall population structure estimates. Such approaches will especially need to be used for traits under divergent selection such as flowering time.

Implications for Association Testing

A rapid breakdown of LD due to linkage will be favorable for association testing of candidate genes that are located near mapped QTLs and have functional relevance to trait variation. The rate of LD decay is probably too rapid to permit genome-wide association testing with SNPs as has been proposed for human populations (Reich et al. 2001).

However, a two-tiered strategy of QTL mapping followed by association testing of positional candidate genes shows substantial promise for localizing quantitative trait effects to individual genes or even subgenic regions (Thornsberry et al. 2001). The rapid LD decay in maize provides an opportunity to map quantitative trait loci with up to 5,000 fold greater resolution than current mapping with F2 or recombinant inbred populations. Statistical approaches will be needed to control for the effects of population structure, but suitable methods are now available (Thornsberry et al. 2001). Mapping QTLs to the level of individual genes will provide new insights into the molecular and biochemical basis for quantitative trait variation, and identify specific targets for crop improvement for the 21st century.

Acknowledgements

We are grateful to the College of Agriculture and Life Sciences Genome Research Laboratory at North Carolina State University for assistance with sequencing. We thank Greg Gibson, Oscar Smith and Rex Bernardo for helpful comments on the manuscript. This research was supported by a grant from the National Science Foundation (#DBI-9872631) and the U.S. Department of Agriculture, Agricultural Research Service.

Literature Cited

Abler, B. S. B., M. D. Edwards and C. W. Stuber (1991). "Isoenzymatic Identification of Quantitative Trait Loci in Crosses of Elite Maize Inbreds." *Crop Science* **31**(2): 267-274.

Alpert, K. B. and S. Tanksley (1996). "High-resolution mapping and isolation of a yeast artificial chromosome contig containing *fw2.2*: A major fruit weight quantitative trait locus in tomato." *Proceedings of the National Academy of Sciences of the United States of America* **93**: 15503-15507.

Doebley, J., J. D. Wendel, J. S. C. Smith, C. W. Stuber and M. M. Goodman (1988). "The Origin of Cornbelt Maize - the Isozyme Evidence." *Economic Botany* **42**(1): 120-131.

Dooner, H. K. and I. M. MartinezFerez (1997). "Recombination occurs uniformly within the bronze gene, a meiotic recombination hotspot in the maize genome." *Plant Cell* **9**(9): 1633-1646.

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." *Proceedings of the National Academy of Sciences of the United States of America* **95**: 4441-4446.

Falconer, D. S. and T. F. C. Mackay (1996). *Introduction to Quantitative Genetics*.

Farnir, F., W. Coppieters, J. J. Arranz, P. Berzi, N. Cambisano, B. Grisart, L. Karim, F. Marcq, L. Moreau, M. Mni, C. Nezer, P. Simon, P. Vanmanshoven, D. Wagenaar and M. Georges (2000). "Extensive genome-wide linkage disequilibrium in cattle." *Genome Research* **10**(2): 220-227.

Fisher, R. A. (1935). "The Logic of Inductive Inference." *Journal of the Royal Statistical Society* **98**: 39-54.

Hanson, M. A., B. S. Gaut, A. O. Stec, S. I. Fuerstenberg, M. M. Goodman, E. H. Coe and J. F. Doebley (1996). "Evolution of anthocyanin biosynthesis in maize kernels: The role of regulatory and enzymatic loci." *Genetics* **143**(3): 1395-1407.

Hedrick, P. W. (1987). "Gametic Disequilibrium Measures - Proceed with Caution." *Genetics* **117**(2): 331-341.

Henry, A. M. and C. Damerval (1997). "High rates of polymorphism and recombination at the Opaque-2 locus in cultivated maize." *Molecular & General Genetics* **256**(2): 147-157.

Hey, J. and J. Wakeley (1997). “A coalescent estimator of the population recombination rate.” *Genetics* **145**(3): 833-846.

Hill, W. G. and B. S. Weir (1988). “Variances and Covariances of Squared Linkage Disequilibria in Finite Populations.” *Theoretical Population Biology* **33**(1): 54-78.

Hudson, R. R. (1987). “Estimating the Recombination Parameter of a Finite Population-Model without Selection.” *Genetical Research* **50**(3): 245-250.

Huttley, G. A., M. W. Smith, M. Carrington and S. J. O'Brien (1999). “A scan for linkage disequilibrium across the human genome.” *Genetics* **152**(4): 1711-1722.

Koch, H. G., J. McClay, E. W. Loh, S. Higuchi, J. H. Zhao, P. Sham, D. Ball and I. W. Craig (2000). “Allele association studies with SSR and SNP markers at known physical distances within a 1 Mb region embracing the ALDH2 locus in the Japanese, demonstrates linkage disequilibrium extending up to 400 kb.” *Human Molecular Genetics* **9**(20): 2993-2999.

Koester, R. P., P. H. Sisco and C. W. Stuber (1993). “Identification of Quantitative Trait Loci Controlling Days to Flowering and Plant Height in 2 near-Isogenic Lines of Maize.” *Crop Science* **33**(6): 1209-1216.

Kohn, M. H., H. J. Pelz and R. K. Wayne (2000). "Natural selection mapping of the warfarin-resistance gene." *Proceedings of the National Academy of Sciences of the United States of America* **97**(14): 7911-7915.

Labate, J. A., K. R. Lamkey, M. Lee and W. Woodman (2000). "Hardy-Weinberg and linkage equilibrium estimates in the BSSS and BSCB1 random mated populations." *Maydica* **45**(3): 243-256.

Lai, C., R. F. Lyman, A. D. Long, C. H. Langley and T. F. C. Mackay (1994). "Naturally Occurring Variation in Bristle Number and DNA Polymorphisms at the *scabrous* Locus of *Drosophila melanogaster*." *Science* **266**: 1697-1702.

Laitinen, T., P. Kauppi, J. Ignatius, T. Ruotsalainen, M. J. Daly, H. Kaariainen, L. Kruglyak, H. Laitinen, A. delaChapelle, E. S. Lander, L. A. Laitinen and J. Kere (1997). "Genetic control of serum IgE levels and asthma: linkage and linkage disequilibrium studies in an isolated population." *Human Molecular Genetics* **6**(12): 2069-2076.

Moffatt, M. F., J. A. Traherne, G. R. Abecasis and W. Cookson (2000). "Single nucleotide polymorphism and linkage disequilibrium within the TCR alpha/delta locus." *Human Molecular Genetics* **9**(7): 1011-1019.

Okagaki, R. J. and C. F. Weil (1997). "Analysis of recombination sites within the maize waxy locus." *Genetics* **147**(2): 815-821.

Pritchard, J. K., M. Stephens and P. Donnelly (2000a). "Inference of population structure using multilocus genotype data." *Genetics* **155**: 945-959.

Pritchard, J. K., M. Stephens, N. A. Rosenberg and P. Donnelly (2000b). "Association mapping in structured populations." *American Journal of Human Genetics* **67**: 170-181.

Reich, D. E., M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti, D. J. Richter, T. Lavery, R. Kouyoumjian, S. F. Farhadian, R. Ward and E. S. Lander (2001). "Linkage disequilibrium in the human genome." *Nature* **411**: 199-204.

SAS Institute, I. (1999). SAS. Cary, NC.

Schneider, S., D. Roessli and L. Excoffier (2000). Geneva, Switzerland, Genetics and Biometry Laboratory, University of Geneva.

Slatkin, M. (1999). "Disequilibrium mapping of a quantitative-trait locus in an expanding population." *American Journal of Human Genetics* **64**(6): 1765-1773.

Stuber, C. W., M. Polacco and M. Lynn (1999). "Synergy of empirical breeding, marker-assisted selection, and genomics to increase crop yield potential." *Crop Science* **39**(6): 1571-1583.

Sved, J. A. (1971). *Theoretical Population Biology* **2**: 125-141.

Thornsberry, J. M., M. M. Goodman, J. F. Doebley, S. Kresovitch, D. Nielson and E. S. Buckler, IV (2001). "*Dwarf8* polymorphisms associate with variation in flowering time." *Nature Genetics* **28**: 286-289.

Wang, R.-L., A. Stec, J. Hey, L. Lukens and J. F. Doebley (1999). "The limits of selection during maize domestication." *Nature* **398**: 236-239.

Weir, B. S. (1996). *Genetic Data Analysis II*. Sunderland, MA, Sinauer.

Weir, B. S. and W. G. Hill (1986). "Nonuniform Recombination within the Human Beta-Globin Gene- Cluster." *American Journal of Human Genetics* **38**(5): 776-778.

Wiehe, T. and M. Slatkin (1998). "Epistatic selection in a multi-locus Levene model and implications for linkage disequilibrium." *Theoretical Population Biology* **53**(1): 75-84.

Xu, X. J., A. P. Hsia, L. Zhang, B. J. Nikolau and P. S. Schnable (1995). "Meiotic recombination break points resolve at high rates at the 5' end of a maize coding sequence." *Plant Cell* **7**(12): 2151-2161.

Table 1. Comparison of LD values between pairs of polymorphic sites in different genes.

		Mean ± SD			
Comparison	Degree of linkage[a]	r^2	D'	$f(P<0.01)$[b]	n_{obs}
d8 vs. tb1	tightly-linked	0.046 ± 0.059	0.486 ± 0.325	0.157	624
d8/tb1 vs. id1	loosely-linked	0.014 ± 0.021	0.237 ± 0.252	0.001	825
d8/tb1/id1 vs. d3	unlinked	0.022 ± 0.030	0.334 ± 0.309	0.018	2730
All unlinked site pairs		0.024 ± 0.001	0.338 ± 0.005	0.036	4179

[a] Tightly-linked loci are ~1cM apart; loosely-linked loci are ~22cM apart; unlinked loci are on different chromosomes.

[b] Percentage of site pairs with LD *P*-value <0.01.

Table 2. Overall and pairwise estimates of F_{ST} for 47 SSR loci, using (a) origin-based and (b) model-based population subdivisions.

	Origin-based subdivision[a]				Model-based subdivision[a]		
Subdivision[a]	ST	NSS	Overall	Subdivision[a]	ST_M	NSS_M	Overall
NSS	0.069	-	-	NSS_M	0.086	-	-
SS	0.202	0.132	-	SS_M	0.224	0.149	-
Combined	-	-	0.105	Combined	-	-	0.122

[a] ST/ST_M = tropical/semi-tropical lines. NSS/NSS_M = US/Northern non-stiff-stalk lines. SS/SS_M = US/Northern stiff-stalk lines.

Table 3. Numbers of SSR locus pairs showing LD at a *P*=0.01 level, by population subdivision.

Population subdivision	# lines	Number of locus pairs in *LD*	% of locus pairs	Expected % based on sample size[a]
All	102	105	9.7%	
Model-based subdivisions:				
ST_M	37	26	2.4%	3.0%
NSS_M	53	26	2.4%	4.6%
SS_M	12	6	0.6%	0.6%

[a] Empirically estimated % of locus pairs expected to show LD if population subdivision effect were due only to reduction in sample size, based on average percent of all locus pairs showing *LD* in a random sample containing the same number of lines.

Figure captions:

Figure 1. – Plots of squared correlations of allele frequencies (r^2) against weighted distance between polymorphic sites in six candidate genes: (a) *id1*; (b) *tb1*; (c) *d8*; (d) *d3*; (e) *sh1*; (f) *su1*. Curves show nonlinear regression of r^2 on weighted distance, using a recombination-drift model for *su1* and a mutation-recombination drift model for all other loci. Regression coefficients (b_1) and the corrected percentage of variance explained by the models (SS_M/SS_C) are shown above each plot.

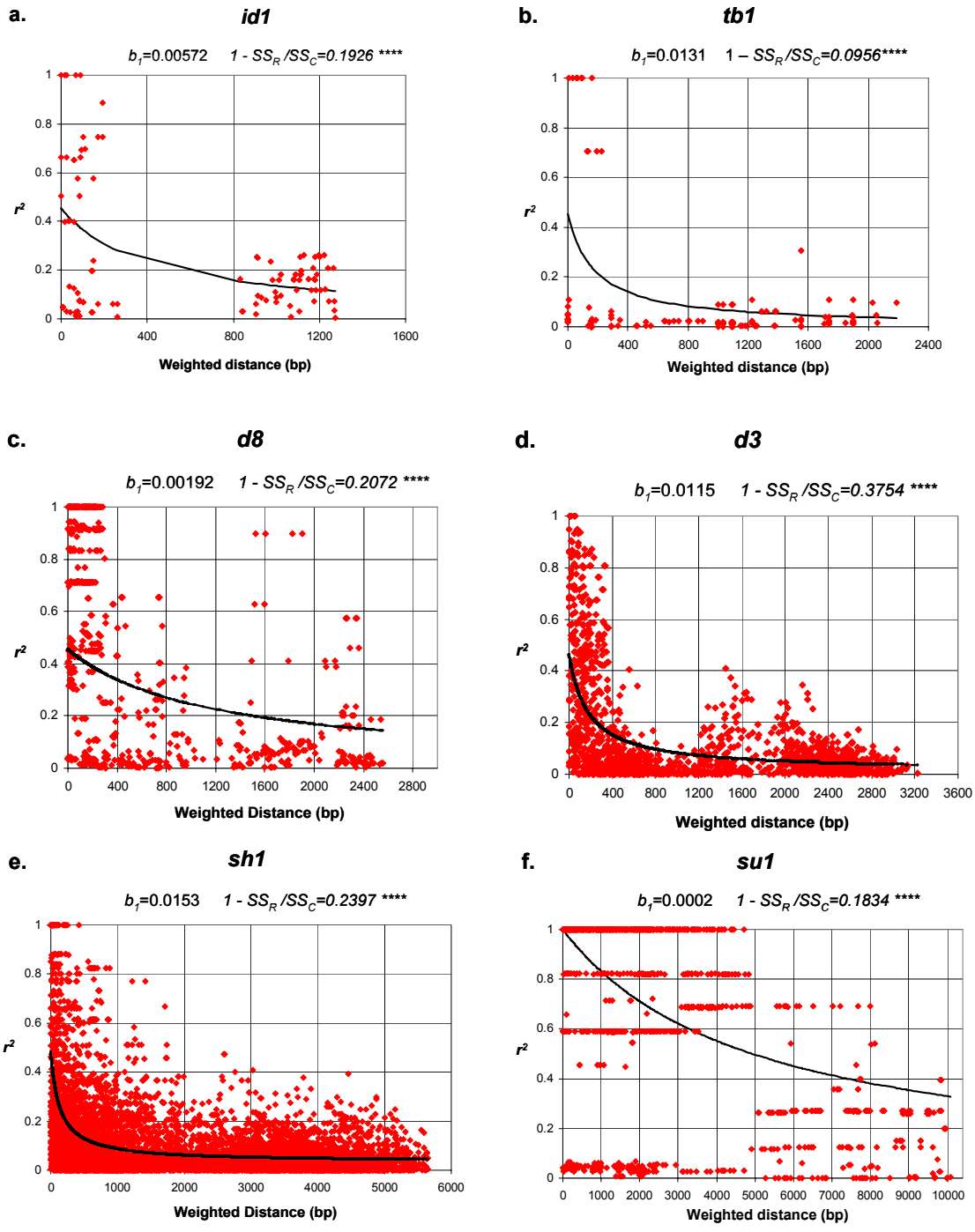

a.
id1
b1=0.00572 1 - SSR /SSC=0.1926 ****
r2
Weighted distance (bp)
b.
tb1
b1=0.0131 1 – SSR/SSC=0.0956****
r2
Weighted distance (bp)
c.
d8
b1=0.00192 1 - SSR /SSC=0.2072 ****
r2
Weighted Distance (bp)
d.
d3
b1=0.0115 1 - SSR /SSC=0.3754 ****
r2
Weighted distance (bp)
e.
sh1
b1=0.0153 1 - SSR /SSC=0.2397 ****
r2
Weighted Distance (bp)
f.
su1
b1=0.0002 1 - SSR /SSC=0.1834 ****
r2
Weighted distance (bp)

Figure 1a-f.