The Plant Journal (2005) **44**, 1054–1064 doi: 10.1111/j.1365-313X.2005.02591.x

TECHNICAL ADVANCE

Maize association population: a high-resolution platform for quantitative trait locus dissection

Sherry A. Flint-Garcia[1,2], Anne-Céline Thuillet[3], Jianming Yu[4], Gael Pressoir[4], Susan M. Romero[4], Sharon E. Mitchell[4], John Doebley[3], Stephen Kresovich[4], Major M. Goodman[5] and Edward S. Buckler[4,6,*]

[1]*US Department of Agriculture – Agricultural Research Service, Plant Genetics Research Unit, Columbia, MO 65211, USA*

[2]*Division of Plant Sciences, University of Missouri, Columbia, MO 65211, USA,*

[3]*Department of Genetics, University of Wisconsin, Madison, WI 65211, USA,*

[4]*Institute for Genomic Diversity, Cornell University, Ithaca, NY 14853, USA,*

[5]*Department of Crop Science, North Carolina State University, Raleigh, NC 27695, USA, and*

[6]*US Department of Agriculture – Agricultural Research Service, US Plant, Soil and Nutrition Research Unit, Ithaca, NY 14853, USA*

Received 5 July 2005; accepted 17 August 2005.

*For correspondence (fax +1 607 255 6249; e-mail esb33@cornell.edu).

Summary

Crop improvement and the dissection of complex genetic traits require germplasm diversity. Although this necessary phenotypic variability exists in diverse maize, most research is conducted using a small subset of inbred lines. An association population of 302 lines is now available – a valuable research tool that captures a large proportion of the alleles in cultivated maize. Provided that appropriate statistical models correcting for population structure are included, this tool can be used in association analyses to provide high-resolution evaluation of multiple alleles. This study describes the population structure of the 302 lines, and investigates the relationship between population structure and various measures of phenotypic and breeding value. On average, our estimates of population structure account for 9.3% of phenotypic variation, roughly equivalent to a major quantitative trait locus (QTL), with a high of 35%. Inclusion of population structure in association models is critical to meaningful analyses. This new association population has the potential to identify QTL with small effects, which will aid in dissecting complex traits and in planning future projects to exploit the rich diversity present in maize.

Keywords: association mapping, quantitative trait loci, diverse maize germplasm, linkage-disequilibrium mapping.

Introduction

Conventional methods for mapping quantitative trait loci (QTL) in plants include first generating a population (F_2, backcross, recombinant inbred, etc.) from a biparental cross, genotyping the individuals with genetic markers across the genome, phenotyping the individuals for the trait of interest, and then analyzing the results via linkage mapping. While linkage mapping has certainly proven useful in identifying a number of important qualitative and quantitative traits in many plant species, it is severely limited by the resolution of the mapping population. Due to small population sizes and the modest degree of recombination within the population, resolution is often in the range of 10–30 cM. Efforts to increase mapping resolution in maize (*Zea mays* ssp. *mays*) by random mating for several generations prior to inbreeding have improved resolution to a few cM [e.g. the intermated B73 × Mo17 population (IBM), Lee *et al.*, 2002], but this still corresponds to millions of bases and hundreds of genes. Linkage mapping is also limited to sampling only two alleles at a locus in any given biparental population.

Association mapping, based on linkage disequilibrium (LD), offers an alternative method for mapping QTL. Originally developed for use in mapping human disease genes

(Corder *et al.*, 1994; Kerem *et al.*, 1989), association mapping utilizes ancestral recombination events in natural populations to make marker–phenotype associations (reviewed by Buckler and Thornsberry, 2002; Rafalski, 2002). Association methods evaluate whether certain alleles within a population are found with specific phenotypes more frequently than expected.

Association mapping has several advantages over linkage mapping in traditional biparental populations: (i) currently existing populations are used versus generating a population via a biparental cross (especially relevant for long-lived species); (ii) a potentially large number of alleles per locus – as opposed to only two – can be surveyed simultaneously; and (iii) resolution can be dramatically increased (e.g. 2000 bp in diverse maize inbred lines; Remington *et al.*, 2001). Given enough statistical power, this latter improvement may allow for the identification of the causative polymorphism within a candidate gene. Current applications of association analysis include genome scans and candidate-gene testing. In a genome scan, single-nucleotide polymorphism (SNP) markers are placed across the genome at an appropriate density, while candidate-gene testing involves sequencing only the candidate gene. Success of either method depends on population size and the degree of LD in the population, genome scans being most useful in species with moderate to extensive LD (species with low LD require excessive numbers of markers to cover the genome); and candidate-gene testing being most effective for species with low LD (the LD may extend well beyond the candidate gene in species with high LD).

Association analysis has been employed only recently in plants, with initial resistance due in large part to the confounding effects of population structure and the general lack of knowledge regarding the structure of LD in many plant species (reviewed by Flint-Garcia *et al.*, 2003). The complex breeding history of many important crops and limited gene flow in most wild plants have created complex stratification within the germplasm (Sharbel *et al.*, 2000). This population stratification and an unequal distribution of alleles within a population can result in non-functional, spurious associations (Knowler *et al.*, 1988). However, the effects of population structure can be corrected for by using a large number of independent genetic markers across the genome (Pritchard and Rosenberg, 1999; Pritchard *et al.*, 2000b; Reich and Goldstein, 2001). As a result, many important plant association studies have been published to date, including flowering time in maize (Thornsberry *et al.*, 2001); growth habit and bolting in sea beet (*Beta vulgaris* ssp. *maritima*) (Hansen *et al.*, 2001); trichome density and initiation in Arabidopsis (Hauser *et al.*, 2001); kernel composition in maize (Wilson *et al.*, 2004); and carotenoid content in maize (Palaisa *et al.*, 2003).

Here we describe the latest association mapping resource available to the maize community, including accurate estimates of relatedness and population structure based on 89 simple-sequence repeat (SSR) loci. Our population of 302 maize inbred lines represents the diversity present in public-sector breeding programs around the world, and captures a large proportion of the maize germplasm pool (Liu *et al.*, 2003). We also discuss relationships between trait heritability, population structure, and power in association analysis. This association population will provide the maize community with a high-resolution platform for QTL dissection.

Results and discussion

Maize has extensive phenotypic and molecular diversity that can be exploited in breeding programs for crop improvement. To use this diversity effectively requires a thorough understanding of the phenotypic and molecular variation of these genetic resources. In addition, before initiating genetic analysis of a quantitative trait, information about the trait such as its heritability and its relationship with population structure must be known. Here we discuss these aspects of phenotypic and molecular diversity in the context of a new maize association mapping population.

Association analysis: the basic components

Detailed methods for association mapping are given by Whitt and Buckler (2003), and additional information can be found at http://www.maizegenetics.net. Here we provide only a brief introduction to the five basic steps required for association studies: germplasm choice; estimation of population structure; trait evaluation; identification of candidate polymorphisms; and statistical analysis (Figure 1).

Choice of germplasm is critical to the success of association analysis. The germplasm set should encompass as much phenotypic variation as possible, and perhaps represent the breeding pool of a crop species. Genetic or phenotypic surveys can be used to identify genotypically diverse subsets of the available germplasm in order to maximize the range of alleles sampled in the population. In some species, core sets of germplasm have already been defined and characterized, and can be used to initiate preliminary association studies. The degree of LD present in the population will determine the resolution of the analysis. The structure of LD is highly population-specific (Rafalski, 2002; Tenaillon *et al.*, 2001), and can be locus-specific due to differences in recombination and mutation rates as well as the evolutionary history of the locus (Remington *et al.*, 2001).

The presence of population structure can lead to spurious results, and must be accounted for in the statistical analysis. The underlying population stratification can be quantified by acquiring genotypic data for a large number of independent markers (50–150) for each member of the germplasm set. A more thorough discussion regarding the analysis of

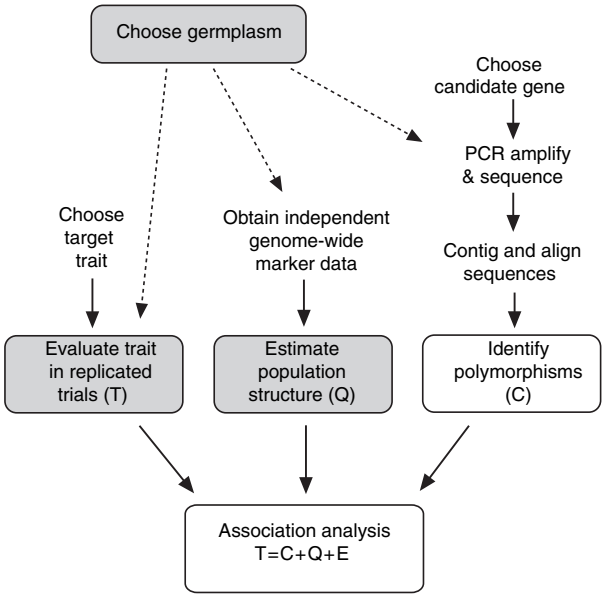

Figure 1. Flow chart illustrating the steps involved in association mapping. Steps discussed in detail are displayed in gray boxes.

population structure and generation of the 'Q matrix' follows below.

Once an adequate germplasm set has been defined and the population structure has been estimated, as is the case for this new association population, an almost endless number of experiments may be conducted regarding specific traits and candidate genes. Choice of target trait(s) should reflect the ability to measure the trait accurately (heritability). For example, many quantitative agronomic traits, such as plant-damage ratings for insect resistance, have low heritability, requiring extensive replication to obtain meaningful trait data. In those cases, and where data collection is extremely intensive, use of a highly correlated trait may be preferable. Trait measurements should balance simplicity in data collection, biological relevance and reproducibility. Typically, trait data should come from randomized plots with 10–15 plants per row, replicated within and across multiple environments, especially if estimates of genotype × environment effects are desired.

Choice of candidate genes often requires a multi-disciplinary approach. Studies involving mutagenesis, biochemical, expression-profiling and comparative genomics can be used to create a list of 'positional candidates', or candidate genes that fall within previously defined QTL intervals. A preliminary sequence for each candidate gene is needed to design overlapping primer pairs to amplify both upstream and coding regions of the gene. Primer design can be challenging in diverse germplasm, as priming in more conserved regions may amplify paralogous regions of the genome, while priming in less conserved regions may fail if there is extensive polymorphism in the primer region across the germplasm. Nucleotide sequence is determined by capillary electrophoresis, sequence quality is assessed, and sequences are contiged/joined. Alignments of contigs across the germplasm are created, and polymorphisms (indels and SNPs) are identified.

Trait data, candidate polymorphisms and population structure data are then analyzed by association analysis (Figure 1). The type III sums of squares are tested using the generalized linear model (GLM):

$$T = C + Q + \varepsilon$$

where T is the trait data, C is the genotype of the candidate polymorphism, Q is the population structure Q matrix, and ε is error. Permutation analysis, a less conservative method than the Bonferroni correction, is then used to determine significance thresholds. Significant results should be checked for phenotypic and genotypic outliers that may be driving the associations. Linkage disequilibrium should be examined to determine which polymorphisms form haplotypes within the candidate gene, and to define the resolution of the association analysis. The most straightforward way to verify a putative association is to evaluate the candidate polymorphism in an entirely different population sample with an independent population structure. Other verification methods include biochemical studies when the polymorphism causes a change in promoter activity or coding sequence, and analysis of nearly isogenic lines.

The association population

The newest association resource available to the maize community is a collection of 302 inbred lines representing the diversity present in public-sector maize-breeding programs worldwide (Table S1). This germplasm set is comprised of current breeding lines as well as historically important lines from both temperate and tropical programs, including eight popcorns and seven sweetcorn lines with genetically distinct breeding histories. A subset of this population (102 lines) has been used in previous association analyses (Remington *et al.*, 2001; Thornsberry *et al.*, 2001; Wilson *et al.*, 2004), and the population structure for 238 of the 302 lines was analyzed in an earlier study by (Liu *et al.*, 2003).

Phenotypic diversity

Here we present well replicated phenotypic data on 101 inbred lines from the original population of 102 inbred lines (Q6199 was omitted from phenotypic analysis due to lack of data and difficulty in maintaining seed stocks). During the period 1998–2002, data were collected for 60 plant, ear and kernel traits from both winter and summer seasons in up to

10 environments. Rather than discussing these traits *per se*, these data are provided in order to demonstrate the value of this population as a genomic tool.

As indicated in Table 1, there was substantial variation for nearly all of the 60 traits we measured in this diverse germplasm set. Flowering dates (days to pollen shed and silk) represented the most striking example of phenotypic variation, ranging from 50 to 90 days after planting. To date, most genetic analyses in maize have utilized only a handful of inbred lines, such as B73 and Mo17. When these two inbreds are used as inbred lines *per se*, they capture an average of only 20% of the phenotypic range seen in this diverse group of inbred lines (Table 1). Many QTL studies report transgressive segregation when examining populations derived from biparental crosses such as the IBM population. Based on the narrow range of B73 and Mo17 compared with the 101 inbreds examined here, we would expect to see even greater transgressive segregation in populations derived from more diverse germplasm. Our findings in the inbreds emphasize the importance of examining a broad germplasm set in order to identify inbreds that differ for the trait of interest.

Heritability

Broad-sense heritability (H^2) 'expresses the extent to which individuals' phenotypes are determined by the genotypes' (Falconer and Mackay, 1996), and corresponds computationally to the proportion of total phenotypic variance due to genetics. Narrow-sense heritability (h^2) is the proportion of the phenotypic variance that can be attributed specifically to additive genetic variance. Additionally, there are two ways to report heritability: mean-basis heritability, and plot-basis heritability. While their calculation varies only slightly, the interpretation of each is quite different (Holland *et al.*, 2003).

Numerous factors affect the ability to detect QTL, including population size, number of QTL and size of their effects, and error present in the phenotypic measurements (Beavis, 1994; Lande and Thompson, 1990). One way to interpret H^2 is that it indicates how much replication and what type of experimental design are required to minimize this experimental error. Thus heritability indicates the precision of genotypic mean estimates that are to be used in QTL analysis. A low heritability value can indicate several things: that a large number of genes govern the trait (biological complexity); that a significant proportion of the trait variation is due to the environment or experimental error; and/or that relative differences among genotypic values depend on the environment (genotype × environment interaction). In any case, insufficient replication may result in lower mean-basis heritability, reflecting inaccurate estimates of genetic effects in QTL experiments. A more practical interpretation is that mean-basis H^2 is the upper limit of the proportion of phenotypic variation (R^2) in the data set that can be explained in genetic studies.

For most of traits in this study there were either single-replication environments, or only one multiple-replication environment. Therefore we chose to simplify our data set by combining multiple replicates from an environment into a single mean in order to calculate heritability, causing the genotype × environment effect to be confounded with the error term. We were able to test genotype × environment effects for days to tassel and silk; plant and ear height; and kernel protein, starch, oil, and moisture, as these were the only traits with data from multiple replicates in multiple environments. The genotype × environment interaction was statistically significant ($P < 0.01$) for all but kernel starch and kernel moisture (Table 1). A consequence of significant genotype × environment effects is that different associations may be detected across environments, as is the case with most genetic analyses.

We calculated broad-sense heritability on a family mean basis (H^2_m) for 60 traits by partitioning the variance from replicated evaluations of 101 inbred lines for data combined across all environments (Table 1). H^2_m for most traits in this study was high, despite the fact that data were combined across winter and summer environments. As expected, days to pollen and silk, plant and ear height, and total number of nodes are highly heritable traits, all with H^2_m estimated as ≥0.95. Heritability for number of ears, however, was lower ($H^2_m = 0.64$), a result we also anticipated as data for this trait were collected relatively early after flowering based on the number of shoots with silks present, rather than on the number of ears with seed at harvest. Interestingly, other traits with lower heritability values ($H^2_m \leq 0.60$) included pubescence indices for the leaf blade, leaf sheath and leaf sheath margin. These lower values may be attributed to the subjective scale used during data collection. It is possible that H^2_m was low for these traits because of the biological complexity mentioned above; however, in the case of number of ears, we acknowledge that our approach to scoring this trait was unsatisfactory. Heritability for percentage vivipary was estimated to be zero because the genetic component of the phenotypic variance was insignificant relative to the genotype × environment component (error) for those environments in which we scored percentage vivipary.

The values for narrow-sense heritability on a plot basis (h^2_p) were generally lower than the H^2_m values, reflecting the effort required to obtain reliable trait measurements, as well as the varying complexity levels of the traits. Our estimates of h^2_p for 12 traits (days to pollen, days to silk, plant height, ear height, number of ears, number of kernel rows, ear diameter, cob diameter, ear length, 10-kernel length, 10-kernel mass, kernel oil) are in general agreement with the results of the large number of studies summarized by

Table 1 Descriptive statistics, family mean-basis heritability and percentage of phenotypic variation explained by population structure for 60 traits scored at up to 10 environments

	Minimum	Average	Maximum	Percentage of range captured by B73 and Mo17	Significance of G × E	H_m^{2a}	h_p^{2b}	R^{2c}
Plant traits								
Days to pollen (days)	50.2	70.1	90.8	17.9	**	0.965	0.877	35.0
Days to silk (days)	51.6	70.7	90.2	14.9	**	0.956	0.847	32.9
Plant height (cm)	86	154	211	8.8	**	0.949	0.739	8.7
Ear height (cm)	20	57	100	15.7	**	0.952	0.800	16.4
Upper leaf angle (°)	6.0	54.4	80.6	32.2	ND	0.881	0.726	20.1
Lower leaf angle (°)	31.9	55.6	76.2	16.7	ND	0.860	0.533	6.7
Leaf length (mm)	44.6	74.6	104.2	30.2	ND	0.936	0.920	17.6
Leaf width (cm)	4.4	8.5	11.6	2.6	ND	0.891	0.710	4.6
Nodes with brace roots	0.1	1.7	4.3	27.7	ND	0.773	0.593	12.2
Number of ears[d]	1.2	2.0	3.2	6.4	ND	0.639	0.195	5.0
Nodes from ear to tassel	2.9	5.4	8.4	20.2	ND	0.930	0.859	14.5
Nodes below ear	2.8	5.9	9.8	27.1	ND	0.913	0.817	22.1
Total number of nodes	7.1	13.3	18.8	24.7	ND	0.948	0.911	24.5
Stalk thickness (mm)	15.4	21.6	27.5	25.2	ND	0.811	0.541	1.1
Stalk width (mm)	15.7	21.8	28.2	25.1	ND	0.786	0.532	1.7
Tassel length (mm)	188.3	321.0	474.9	22.8	ND	0.900	0.763	17.5
Main spike length (mm)	119.0	220.5	370.4	35.4	ND	0.912	0.710	3.7
Tassel branch count	0.6	10.3	20.1	9.2	ND	0.922	0.806	21.0
Tassel angle[d] (°)	2.5	61.1	83.0	21.2	ND	0.894	0.741	4.9
Tassel mass (g)	1.3	4.5	11.6	0.5	ND	0.961	0.956	14.8
Stalk anthocyanin	0.0	1.2	7.0	29.1	ND	0.641	0.529	5.6
Brace-root anthocyanin	0.0	3.7	10.0	80.4	ND	0.709	0.554	10.2
Leaf-midrib anthocyanin	0.0	0.1	10.0	0.0	ND	ND	ND	0.6
Leaf-margin anthocyanin	0.0	1.4	10.0	0.0	ND	0.765	0.438	6.7
Anther anthocyanin	0.0	1.4	8.0	18.8	ND	ND	ND	3.2
Glume anthocyanin	0.0	1.9	8.5	5.9	ND	0.805	0.572	4.6
Glume-bar anthocyanin	0.0	1.2	10.0	0.0	ND	0.931	0.816	2.0
Leaf-blade pubescence	0.0	2.0	4.5	0.0	ND	0.604	0.591	15.5
Leaf-sheath pubescence	0.0	1.6	3.7	33.7	ND	0.269	0.327	13.9
Leaf sheath-margin pubescence	0.0	2.0	5.0	9.9	ND	0.429	0.142	3.3
Ear traits								
Number of kernel rows	10.2	13.1	17.2	97.8	ND	0.905	0.746	12.8
Ear diameter (mm)	26.4	36.4	56.0	33.4	ND	0.784	0.216	0.8
Cob diameter (mm)	18.4	23.8	31.1	77.5	ND	0.824	0.369	3.6
Ear mass (g)	20.2	55.3	103.4	8.6	ND	0.853	0.309	1.9
Cob mass (g)	4.1	13.7	30.3	39.2	ND	0.902	0.506	1.5
Total kernel mass (g)	12.1	41.5	80.9	4.6	ND	0.837	0.297	3.7
Ear length (mm)	67.4	119.6	159.6	17.0	ND	0.872	0.449	0.1
Seed-set length (mm)	33.0	102.2	167.0	15.7	ND	0.773	0.569	1.9
Total kernel volume (mL)	18.3	58.0	104.5	2.9	ND	0.833	0.286	5.1
Kernel density[e] (g L^{-1})	0.47	0.70	0.93	43.7	ND	0.451	0.127	0.2
Percentage vivipary (%)	0.0	0.5	24.8	0.8	ND	0.000	ND	1.5
Kernel traits								
10-lernel length[d] (mm)	62.5	85.0	102.8	7.5	ND	0.800	0.242	7.1
10-kernel thickness (mm)	24.8	46.5	60.1	11.7	ND	0.666	0.290	3.0
10-kernel width[d] (mm)	54.0	78.0	98.3	29.0	ND	0.840	0.308	0.9
10-kernel mass[d] (g)	1.06	2.30	3.56	14.9	ND	0.834	0.417	1.4
Kernel moisture[e] (%)	2.3	8.1	9.9	1.7	ns	0.167	ND	0.2
Kernel oil[e] (%)	2.7	6.5	15.4	2.3	**	0.811	0.596	7.9
Kernel protein[e] (%)	7.6	12.2	17.1	0.6	**	0.776	0.469	3.8
Kernel starch[e] (%)	21.7	49.4	76.7	2.0	ns	0.785	0.556	18.9
Kernel amylose[e] (%)	18.4	20.8	26.8	16.8	ND	0.738	0.469	6.4
Starch breakdown[e] (Pa·sec)	0.01	0.12	0.24	4.3	ND	0.610	0.327	8.1
Starch consistency[e] (Pa·sec)	0.08	0.24	0.33	20.0	ND	0.634	0.297	14.6
Starch cool-paste viscosity[e] (Pa·sec)	0.16	0.53	0.73	24.6	ND	0.724	0.348	17.6
Starch hot-paste viscosity[e] (Pa·sec)	0.08	0.29	0.41	27.6	ND	0.727	0.425	19.8

Table 1 Continued

	Minimum	Average	Maximum	Percentage of range captured by B73 and Mo17	Significance of G × E	H_m^{2a}	h_p^{2b}	R^{2c}
Starch-pasting temperature[e] (°C)	63.81	66.82	69.66	33.4	ND	0.666	0.451	4.5
Starch peak temperature[e] (°C)	76.67	83.30	94.03	12.1	ND	0.656	0.315	4.8
Starch peak time[e] (sec)	183.87	222.16	287.26	11.5	ND	0.679	0.318	4.8
Starch peak viscosity[e] (Pa sec)	0.08	0.41	0.63	18.5	ND	0.752	0.369	15.7
Starch setback[e] (Pa sec)	0.08	0.24	0.34	19.9	ND	0.642	0.297	13.3
Starch trough viscosity[e] (Pa sec)	0.08	0.29	0.40	26.6	ND	0.716	0.423	20.8

[a]Broad-sense heritability on family mean basis.
[b]Narrow-sense heritability on a plot mean basis.
[c]Percentage of phenotypic variation explained by population structure.
[d]Popcorn-related trait.
[e]Starch-related trait.
ND, not determined. These traits were not scored in enough environments to calculate heritability and/or to test genotype × environment effects.
$^{**}P < 0.01$; ns, not significant at $P = 0.05$.

Hallauer and Miranda Filho (1988), in which various segregating progeny were used to estimate h_p^2.

Population structure

The prerequisite for all subsequent analyses in this study was the characterization of population structure within our new set of inbred lines using the software package STRUCTURE 2.1 (Pritchard *et al.*, 2000a). We ran STRUCTURE for K (number of fixed subpopulations or clusters) ranging from 1 to 7 on (i) the entire data set (301 lines) and (ii) a subset of lines that excluded seven sweetcorn and eight popcorn lines (286 lines). Based on prior analyses using a smaller number of lines, we found consistently that models including sweetcorn and popcorn lines had lower likelihoods than models excluding these lines (Liu *et al.*, 2003). The sweetcorn and popcorn lines are so very distinct from all other lines because of the intense genetic isolation that occurred during their development as specialty maize. Despite this fact, these lines fail to create their own subpopulation(s) due to the small number of sweetcorn and popcorn lines in our germplasm collection. However, whether or not these lines were included, the model with $K = 3$ clusters showed a higher likelihood than $K = 2$, and likelihoods comparable with models $K = 4$–7. The three clusters are referred to hereafter as non-stiff-stalk (NSS), stiff-stalk (SS) and tropical and subtropical (TS) subpopulations, a naming utility based on how the majority of lines in a given cluster are regarded by maize breeders. The $K = 3$ model excluding sweetcorns and popcorns presented the highest likelihood, and was used to assign lines to the appropriate subpopulations: 113 NSS, 36 SS and 73 TS lines (Table S1). In addition, 64 lines were assigned to a fourth 'mixed' group. Results of this study were consistent with those presented by Liu *et al.* (2003). When sweetcorn and popcorn lines were included in our analyses, they clustered with the NSS lines rather than forming their own subpopulation(s).

Wright's Fst values (Wright, 1951) were used to summarize differentiation between subpopulations relative to variation within subpopulations (Table 2). While analogous Rst estimates based on the variance in microsatellite allele size can also be used for microsatellites (Slatkin, 1995), these are not appropriate when indels occur as is the case for some of the microsatellites used in this study (data not shown). Although the size homoplasy that exists at microsatellite loci may bias Fst estimates and detection of population structure, this is not usually considered to represent a significant problem, as the large amount of variability at microsatellite loci often compensates for the existence of homoplasy (Estoup *et al.*, 2002). Moreover, in our data set a potential effect of size homoplasy on population structure estimates is expected to be diluted by the large number of loci, because the probability of identity by state at a large number of loci for a given line is expected to be low.

A total of 250 lines were classified as members of the NSS, TS or mixed groups, with a maximal pairwise Fst of 0.06 ($P < 0.001$) among these three groups. This value indicates that, while consistent and significant population structure exists, it explains only a modest proportion of the marker variation in this diverse germplasm. By comparison, the sweetcorn, popcorn and SS subpopulations were more differentiated. The overall Fst value for comparing the three subpopulations (SS, NSS, TS) with the sweetcorn and popcorn subpopulations was 0.097 ($P < 0.001$; data not shown). Fst values between all groups were also significant ($P < 0.001$), supporting the existence of genetic structure. The inclusion, and subsequent clustering, of sweetcorn and popcorn lines with the NSS group did not change Fst values between this subpopulation and the others, as respective values remained the same when these lines were excluded

Table 2 Fst values for comparison of subpopulations determined by STRUCTURE analysis

Fst*	Mixed	Non-stiff stalk	Popcorn	Stiff stalk	Sweetcorn	Tropical/subtropical
Mixed		0.01	0.15	0.15	0.08	0.03
Non-stiff stalk			0.15	0.18	0.10	0.06
Popcorn				0.39	0.28	0.16
Stiff stalk					0.33	0.22
Sweetcorn						0.12
Tropical/subtropical						

*All Fst values are significant ($P < 0.001$).

(data not shown). The mixed group fell substantially closer to the TS and NSS subpopulations than it did to the SS subpopulation, which was in agreement with pedigree information (Table S1).

In the past, maize breeders have relied on qualitative pedigree information (Table S1) to estimate the genetic contribution of inbred lines and the relationships between them. Our population structure estimates based on SSR marker data allow us to quantify these relationships in the form of the *Q* matrix. This will allow breeders to make more informed decisions about potential crosses between these inbred lines.

A phylogenetic tree was constructed using the same genotypic data cited above. This tree was in good agreement with the STRUCTURE analysis, further supporting the proposed genetic structure to describe the set of inbred lines. The detection of historically meaningful phylogenetic relationships between lines (according to pedigree-based groups: SS, NSS, TS, popcorn and sweetcorn lines) also supports the fact that potential size homoplasy at microsatellites is not problematic in our data set. Consequently, the population structure detected is likely to have some evolutionary meaning, and is not significantly affected by possible homoplasy among alleles. The phylogenetic tree is presented in Figure S1.

Armed with these population-structure data, a proper association analysis can now be performed on the new set of 302 inbred lines. Several methods using independent marker data have been devised recently to correct for population structure (Pritchard *et al.*, 2000a,b; Reich and Goldstein, 2001). Pritchard's method (Pritchard *et al.*, 2000a,b), which incorporates estimates of population structure directly into the association test statistic, has been integrated into the association analysis software TASSEL (see http://www.maizegenetics.net for more information). These population-structure data can also be utilized in approaches that rely on alternative methods for dealing with relatedness (Crepieux *et al.*, 2004).

Effect of population structure on phenotype

A particular strength of this study was that estimates of population structure based on 89 SSR loci allowed us to investigate the relationship between population structure and various aspects of phenotype, such as heritability. As noted above, an understanding of the effect of population structure on the trait of interest is critical for association studies in order to prevent spurious associations. However, if population structure is found to explain too much of the variation, then structured association analyses will have little power to detect the effects of individual genes.

For each trait, we looked for any important statistical outliers related to population structure and breeding history. As sweetcorn lines are mutant outliers for kernel composition traits, they were excluded from the kernel composition population structure analyses. Likewise, popcorn lines were identified as outliers on kernel morphology traits, as well as tassel angle and number of ears, and thus were also excluded from phenotypic correlations.

Overall, population structure accounts for an average of 9.3% of the phenotypic variation across all traits in this study, even after appropriate lines were excluded from the analysis for specific traits (Table 1). Population structure is rarely the dominant factor in phenotypic variation, but its effect can be seen in nearly all traits. As a 9.3% effect is roughly equivalent to a substantial QTL detected in interval mapping, or a major QTL in association analysis, it is imperative that the effect of population structure be examined when doing association studies for any given trait. By controlling for population structure, Thornsberry *et al.* (2001) decreased the number of false-positive associations by almost fivefold for some traits in an association study of flowering time in maize. Wilson *et al.* (2004) found that a previously reported association between *shrunken2* and overall starch in the original set of 102 inbred lines was no longer evident when population structure was included in the analysis. They concluded that either the previously reported QTL was perhaps the result of the genotype × environment-sensitive nature of *shrunken2*, or that the original association may have been caused exclusively by population structure.

Structured association analysis of traits highly correlated with population structure will result in many false negatives (lack of power). For example, as flowering time is highly correlated with population structure ($R^2 = 33$–35%), functional alleles whose distribution coincide with population

structure will not be detected when association models include population structure estimates. This problem can be seen as the trait and polymorphism(s) associate very strongly when population structure is ignored, but the association disappears when structure is considered. In these cases, alternate association populations or linkage populations would be more useful for evaluating the candidate polymorphisms. In this study of a new association population, days to tassel and days to silk appear to be the only traits where this phenomenon will be common.

Heritability and population structure

Our results suggest an interesting relationship between heritability and population structure effect ($r = 0.37$). Although there was considerable scatter around the regression line, two notable outliers were apparent: days to pollen and days to silk have a higher population structure effect than other traits of similar heritability. This result is reasonable, as flowering time has probably undergone selection to form the underlying population structure of this new association population. While this population is an excellent genetic resource for basic association analyses, alternative populations may need to be designed to address specific traits where there is a very strong relationship with population structure.

Population size and power to detect associations

Simulation studies have demonstrated that sufficient power exists to detect SNP–phenotype associations for QTL that account for as little as 5% of the phenotypic variation when approximately 500 individuals are genotyped for approximately 20 SNPs within the candidate gene region (Long and Langley, 1999). This model-based study found that more power is achieved by increasing the number of individuals in the population than by increasing the SNP density within the candidate gene.

Increasing population size lessens the impact of several factors that limit the power to detect associations. Although a major advantage of association approaches is the ability to evaluate multiple alleles simultaneously (compared with only two alleles in standard linkage mapping), testing multiple alleles can pose two problems. First, the more alleles there are at any particular locus, the more allelic classes must be tested, causing a multiple-test problem. This problem can be addressed by using permutation analysis to determine an experiment-wise significance threshold (Churchill and Doerge, 1994). Second, as the number of possible alleles per locus increases, the number of individuals within each allelic class decreases, thereby decreasing overall power. However, increasing the population size necessarily increases the number of individuals with rare alleles, thus improving the power to test these rare alleles. A separate but related issue is the presence of epistasis. When alleles at different loci interact, it is necessary to include the interaction term in the model. However, with multiple alleles at a particular locus, the number of individuals in each interaction class may be very small. Increasing population size also serves to increase the number of individuals with certain allelic combinations, thereby allowing for more powerful tests of epistasis. Genes that are confirmed to play a role in the expression of a trait can also be added to the model as cofactors in order to test for epistatic effects (Szalma *et al.*, 2005).

It is critical, however, that plant geneticists should not completely abandon linkage mapping in favor of association analysis. The relative success of association analysis compared with linkage analysis is species-specific, as well as population-specific. For example, in species with low genetic diversity, linkage analysis is expected to be superior to association analysis. In this case even the best germplasm collection will not contain enough diversity to offset the loss in statistical power in association analysis. Although association analysis plays a key role in genetic analysis, it is still only one of many valuable methods. An ideal analysis combines linkage and association analysis, where the strengths of each method are used to conduct high-power, high-resolution tests (Jansen *et al.*, 2003).

Overall, this new population of 302 maize inbred lines harbors substantial phenotypic and genotypic diversity that can be exploited efficiently for maize improvement through association analysis. While population structure has a persistent effect in this diverse maize population, the effect is one of reasonable magnitude that can be controlled. To exploit fully the benefits of association analysis as a genetic/genomic tool in other plant species, a substantial effort is needed to create association populations, analyze the LD present within each population, and describe the population structure for various plant species. However, once an association population has been developed for a species, as the current population has, a community effort is needed to characterize the population phenotypically in order to maximize its potential use in crop improvement.

Experimental procedures

The association population

A set of 302 maize inbred lines was assembled to represent the diversity present in public-sector corn-breeding programs around the world. Pedigrees of lines included can be found in a number of sources, including published material by Gerdes *et al.* (1993) and in the Germplasm Resources Information Network (GRIN) database, but are summarized here in Table S1. Seed of most lines can be obtained from their original source (see http://www.panzea.org). Seed samples have also been provided to the North Central Regional Plant Introduction Station (Ames, IA, USA), and are available for distribution on request.

Phenotypic data

The original set of 102 inbred lines (a subset of the population of 302) was grown in one-row plots in a completely randomized block design at each of the following environments: 1998 Homestead, FL (one replicate); 1999 Clayton, NC (three replicates); 1999 Homestead, FL (one replicate); 2000 Clayton, NC (three replicates); 2000 West Lafayette, IN (one replicate); 2000 Homestead, FL (one replicate); 2001 Clayton, NC (three replicates); 2001 Urbana, IL (two replicates); 2002 Clayton, NC (one replicate); 2002 Homestead, FL (one replicate). Data for Q6199 were omitted from all analyses due to extensive missing data and difficulty in maintaining seed stocks. MaizeMeister, a personal digital assistant (PDA) and bar code-based phenotyping system, was used to facilitate phenotypic data collection from up to five plants per plot during both field seasons in 2002. For more information about MaizeMeister see http://www.maizegenetics.net.

Plant data collected in the field included: flowering time (days to silk and tassel); ear and plant heights; leaf angles above and below the ear zone; leaf length and width; number of nodes with brace roots; number of ears; number of nodes above and below ears; total number of nodes; stalk thickness and width; tassel and main spike length; tassel branch count and angle; tassel weight; anthocyanin indices (scale 0–10) for stalk, brace roots, leaf midrib, leaf margin, anther, glume and glume bar; and pubescence indices (scale 0–5) for leaf blade, sheath and sheath margin.

Ear and kernel data were collected from a single (uppermost) hand-harvested, self-pollinated ear per plant. Ear data included: number of kernel rows, ear diameter, cob diameter, ear mass, cob mass, total kernel mass, ear length, seed-set length, kernel volume, kernel density and percentage vivipary. Kernel data included: 10-kernel length, 10-kernel thickness, 10-kernel width, 10-kernel mass, moisture, oil, protein, starch, amylose and starch-pasting properties. The latter data set included values for starch breakdown, consistency, cool-paste viscosity, host-paste viscosity, pasting temperature, peak temperature, peak time, peak viscosity, setback and trough viscosity (methods for kernel composition and starch-pasting properties described by Wilson *et al.*, 2004).

All phenotypic data (least-squares environment means) used in our analyses are provided in Table S2. We are working towards making these data available through the Gramene (http://www.gramene.org) and Panzea (http://www.panzea.org) databases.

Molecular marker data

A subset of 238 inbred lines from the new association population was genotyped at Celera AgGen (Davis, CA, USA) (methods previously described by Liu *et al.*, 2003). The same set of SSR loci were used to genotype the remaining 63 inbred lines (no SSR data were available for NC316) and five teosinte plants at Cornell University, except that SSRs *bnlg1014*, *bnlg1189*, *bnlg1520*, *bnlg2238* and *phi116* were dropped. All SSR data are available at http://www.panzea.org.

Statistical analyses of marker data

To investigate population structure among lines, we used the software package STRUCTURE 2.1 (Pritchard *et al.*, 2000a). This software allows for the identification of different subpopulations within a sample of individuals collected from a population of unknown structure. Given a fixed number of subpopulations (*K*), this method assigns individuals to clusters (each cluster corresponding to a different subpopulation) with an associated probability. We ran STRUCTURE for *K* ranging from 1 to 7 on the entire data set (301 lines), and on a set of lines that excluded seven sweetcorn and eight popcorn lines (286 lines). Five runs were completed for each *K*. In all cases both the burn-in time and the replication number were set to 500 000. Lines with membership probabilities ≥ 0.8 were assigned to subpopulations, while lines with membership probabilities <0.8 were assigned to a mixed group. To assess further the existence of a genetic structure between identified clusters, pairwise Fst values were calculated using the software GENETIX (ver. 4.03; Belkhir *et al.*, 2001) and tested by permutation.

To construct a phylogenetic tree, we calculated the log-transformed, proportion-of-shared-alleles distance between lines. This distance is free of the stepwise assumption, enjoys low variance, and is widely used with multi-locus SSR data (Liu *et al.*, 2003; Matsuoka *et al.*, 2002). The Fitch–Margoliash least-squares algorithm implemented in the computer program PHYLIP was then applied to the distance matrix to obtain the phylogenetic tree (Felsenstein, 1993). The tree was rooted using five samples of teosinte (*Zea mays* ssp. *parviglumis*), a wild relative of maize, as the outgroup (Matsuoka *et al.*, 2002).

Statistical analyses of phenotypic data and related trait analyses

Each environment was treated as a randomized complete block design. Where there were multiple replicates per environment, the lsmeans option of PROC GLM (SAS, SAS Institute, 1999–2001) was used to compute environment means in order to simplify the data set for heritability calculations. Harmonic means were calculated for the number of replicates per line for each trait, and SAS PROC MIXED was employed to partition the variance into genotype, environment, and pooled error (genotype $\times$ environment with residual effect) (Holland *et al.*, 2003). These variance components were then used to estimate broad-sense heritability on a mean basis. We tested genotype $\times$ environment effects for days to tassel and silk, plant and ear height, and kernel protein, starch, oil, and moisture, as these were the only traits with data from multiple replicates in multiple environments. PROC GLM was used to partition the variance and test for genotype $\times$ environment effects.

Narrow-sense heritability on a plot-mean basis was calculated through a mixed model in which both marker-inferred population structure (*Q* matrix) and relatedness between inbred lines (*K* matrix) were accounted for (G.P. and co-workers, unpublished data). Essentially, with adjustment of the marker-inferred relationship among these inbred lines, the phenotypic variance of a trait was used to retrieve the additive genetic variance in a panmictic population from which these inbred lines were derived (Falconer and Mackay, 1996).

The effect of population structure was tested using SAS PROC GLM. The model statement included two of the three components of the $K = 3$ *Q* matrix (NSS, SS) with sweetcorn and popcorn lines excluded from the STRUCTURE analysis. For each trait, groups of inbred lines constituting statistical outliers related to population structure were identified and subsequently excluded from analyses involving population structure. The relationship between heritability and effect of population structure was tested using SAS PROC REG.

Acknowledgements

We would like to thank members of the Buckler laboratory, past and present, for their assistance in phenotypic data collection, and members of the Kresovich laboratory for SSR genotyping. We thank

Jim Holland with the USDA ARS in Raleigh, NC for his insights regarding heritability. We also thank Lauren McIntyre at Purdue University and Torbert Rocheford at the University of Illinois for evaluation of the West Lafayette and Urbana replications, respectively, and Natalie Stevens for technical editing of this manuscript. This project was funded by National Science Foundation DBI-9872631 and DBI-0321467, National Institutes of Health GM-58816, and USDA–ARS.

Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the US Department of Agriculture.

Supplementary Material

The following supplementary material is available for this article online:

Figure S1. Fitch–Margoliash phylogenetic tree of 301 inbred lines constructed using log-transformed proportion of allele-shared distances based on 89 SSR loci.

Table S1 The 302 inbred maize lines comprising the new association population, pedigree information, membership probabilities for subpopulations (non-stiff stalk, NSS; stiff stalk, SS; tropical/subtropical, TS; popcorn, Pop; sweetcorn, Sweet) and assigned subpopulation

Table S2 Least-squares environment means for 60 traits from up to 10 environments for a 101-line subset

This material is available as part of the online article from http://www.blackwell-synergy.com

References

Beavis, W.D. (1994) The power and deceit of QTL experiments: lessons from comparative QTL studies. In *Proceedings of the Annual Corn and Sorghum Industry Research Conference*, Vol. 49 (Wilkinson, D.B., ed.). Chicago, IL, USA: American Seed Trade Association, pp. 250–266.

Belkhir, K., Borsa, P., Chikhi, L., Raufaste, N. and Bonhomme, F. (2001) *GENETIX, Software (Windows) for Population Genetics*. Montpellier, France: Laboratoire Génome, Populations, Interactions CNRS UMR 5000, Université de Montpellier II.

Buckler, E.S. and Thornsberry, J.M. (2002) Plant molecular diversity and applications to genomics. *Curr. Opin. Plant Biol.* **5**, 107–111.

Churchill, G.A. and Doerge, R.W. (1994) Empirical threshold values for quantitative trait mapping. *Genetics*, **138**, 963–971.

Corder, E.H., Saunders, A.M., Risch, N.J. *et al.* (1994) Protective effect of apolipoprotein-E type-2 allele for late-onset Alzheimer disease. *Nat. Genet.* **7**, 180–184.

Crepieux, S., Lebreton, C., Servin, B. and Charmet, G. (2004) Quantitative trait loci (QTL) detection in multicross inbred designs: recovering QTL identical-by-descent status information from marker data. *Genetics*, **168**, 1737–1749.

Estoup, A., Jarne, P. and Cornuet, J.-M. (2002) Homoplasy and mutation models at microsatellite loci and their consequences for population genetics analysis. *Mol. Ecol.* **11**, 1591–1604.

Falconer, D.S. and Mackay, T.F. (1996) *Introduction to Quantitative Genetics*. Harlow, UK: Longman.

Felsenstein, J. (1993) *PHYLIP – Phylogeny Inference Package, Version 3.5c*. Seattle, WA, USA: Department of Genetics, University of Washington.

Flint-Garcia, S.A., Thornsberry, J.M. and Buckler, E.S. (2003) Structure of linkage disequilibrium in plants. *Ann. Rev. Plant Biol.* **54**, 357–374.

Gerdes, J.T., Behr, C.F., Coors, J.G. and Tracy, W.F. (1993) *Compilation of North American Maize Breeding Germplasm*. Madison, WI: Crop Science Society of America.

Hallauer, A.R. and Miranda Filho, J.B. (1988) *Quantitative Genetics in Maize Breeding*, 2nd edn. Ames, IA, USA: Iowa State University Press.

Hansen, M., Kraft, T., Ganestam, S., Säll, T. and Nilsson, N.-O. (2001) Linkage disequilibrium mapping of the bolting gene in sea beet using AFLP markers. *Genet. Res.* **77**, 61–66.

Hauser, M.-T., Harr, B. and Schlötterer, C. (2001) Trichome distribution in *Arabidopsis thaliana* and its close relative *Arabidopsis lyrata*: molecular analysis of the candidate gene *GLABROUS1*. *Mol. Biol. Evol.* **18**, 1754–1763.

Holland, J.B., Nyquist, W.E. and Cervantes-Martínez, C.T. (2003) Estimating and interpreting heritability for plant breeding: an update. *Plant Breed. Rev.* **22**, 9–111.

Jansen, R.C., Jannink, J.-L. and Beavis, W.D. (2003) Mapping quantitative trait loci in plant breeding populations: use of parental haplotype sharing. *Crop Sci.* **43**, 829–834.

Kerem, B.S., Rommens, J.M., Buchanan, J.A., Markiewicz, D., Cox, T.K., Chakravarti, A. and Buchwald, M. (1989) Identification of the cystic fibrosis gene: genetic analysis. *Science*, **245**, 1073–1080.

Knowler, W.C., Williams, R.C., Pettitt, D.J. and Steinberg, A.G. (1988) $Gm^{3-5,13,14}$ and type 2 diabetes mellitus: an association in American Indians with genetic admixture. *Am. J. Hum. Genet.* **43**, 520–526.

Lande, R. and Thompson, R. (1990) Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics*, **124**, 743–756.

Lee, M., Sharopova, N., Beavis, W.D., Grant, D., Katt, M., Blair, D. and Hallauer, A. (2002) Expanding the genetic map of maize with the intermated B73 × Mo17 (IBM) population. *Plant Mol. Biol.* **48**, 453–461.

Liu, K., Goodman, M.M., Muse, S.V., Smith, J.S., Buckler, E.S. and Doebley, J.F. (2003) Genetic structure and diversity among maize inbred lines as inferred from DNA microsatellites. *Genetics*, **165**, 2117–2128.

Long, A.D. and Langley, C.H. (1999) The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Res.* **9**, 720–731.

Matsuoka, Y., Vigouroux, Y., Goodman, M.M., Sanchez, G.J., Buckler, E. and Doebley, J. (2002) A single domestication for maize shown by multilocus microsatellite genotyping. *Proc. Natl Acad. Sci. USA*, **99**, 6080–6084.

Palaisa, K.A., Morgante, M., Williams, M. and Rafalski, A. (2003) Contrasting effects of selection on sequence diversity and linkage disequilibrium at two phytoene synthase loci. *Plant Cell*, **15**, 1795–1806.

Pritchard, J.K. and Rosenberg, N.A. (1999) Use of unlinked genetic markers to detect population stratification in association studies. *Am. J. Hum. Genet.* **65**, 220–228.

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

Pritchard, J.K., Stephens, M., Rosenberg, N.A. and Donnelly, P. (2000b) Association mapping in structured populations. *Am. J. Hum. Genet.* **67**, 170–181.

Rafalski, A. (2002) Applications of single nucleotide polymorphisms in crop genetics. *Curr. Opin. Plant Biol.* **5**, 94–100.

Reich, D.E. and Goldstein, D.B. (2001) Detecting association in a case–control study while correcting for population stratification. *Genet. Epidemiol.* **20**, 4–16.

Remington, D.L., Thornsberry, J.M., Matsuoka, Y., Wilson, L.M., Whitt, S.R., Doebley, J., Kresovich, S., Goodman, M.M. and

Buckler, E.S. (2001) Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proc. Natl Acad. Sci. USA*, **98**, 11479–11484.

SAS Institute (1999–2001) *SAS Proprietary Software Release 8.2*. Cary, NC, USA: SAS Institute Inc.

Sharbel, T.F., Haubold, B. and Mitchell-Olds, T. (2000) Genetic isolation by distance in *Arabidopsis thaliana*: biogeography and postglacial colonization of Europe. *Mol. Ecol.* **9**, 2109–2118.

Slatkin, M. (1995) A measure of population subdivision based on microsatellite allele frequencies. *Genetics*, **139**, 457–462.

Szalma, S.J., Buckler, E.S., Snook, M.E. and McMullen, M.D. (2005) Association analysis of flavonoid structural loci for maysin and chlorogenic acid synthesis in maize silks. *Theor. Appl. Genet.* **110**, 1324–1333.

Tenaillon, M.I., Sawkins, M.C., Long, A.D., Gaut, R.L., Doebley, J.F. and Gaut, B.S. (2001) Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). *Proc. Natl Acad. Sci. USA*, **98**, 9161–9166.

Thornsberry, J.M., Goodman, M.M., Doebley, J., Kresovich, S., Nielsen, D. and Buckler, E.S. (2001) *Dwarf8* polymorphisms associate with variation in flowering time. *Nat. Genet.* **28**, 286–289.

Whitt, S.R. and Buckler, E.S. (2003) Using natural allelic diversity to evaluate gene function. In *Plant Functional Genomics: Methods and Protocols* (Grotewald, E., ed.). Totowa, NJ, USA: Humana Press. 123–139.

Wilson, L.M., Whitt, S.R., Ibanez, A.M., Rocheford, T.R., Goodman, M.M. and Buckler, E.S.I. (2004) Dissection of maize kernel composition and starch production by candidate gene association. *Plant Cell*, **16**, 2719–2733.

Wright, S. (1951) The genetical structure of populations. *Ann. Eugenics*, **15**, 323–354.

