Dwarf8 polymorphisms associate with variation in flowering time

Jeffry M. Thornsberry[1], Major M. Goodman[2], John Doebley[3], Stephen Kresovich[4], Dahlia Nielsen[5], & Edward S. Buckler IV[1,6]

Historically, association tests have been used extensively in medical genetics[1–2], but have had virtually no application in plant genetics. One obstacle to their application is the structured populations often found in crop plants[3], which may lead to nonfunctional, spurious associations[4]. In this study, statistical methods to account for population structure[5] were extended for use with quantitative variation and applied to our evaluation of maize flowering time. Mutagenesis and quantitative trait locus (QTL) studies suggested that the maize gene *Dwarf8* might affect the quantitative variation of maize flowering time and plant height[6–8]. The wheat orthologs of this gene contributed to the increased yields seen in the 'Green Revolution' varieties[6]. We used association approaches to evaluate *Dwarf8* sequence polymorphisms from 92 maize inbred lines. Population structure was estimated using a Bayesian analysis[4] of 141 simple sequence repeat (SSR) loci. Our results indicate that a suite of polymorphisms associate with differences in flowering time, which include a deletion that may alter a key domain in the coding region. The distribution of nonsynonymous polymorphisms suggests that *Dwarf8* has been a target of selection.

We collected plant height and flowering time data from 92 inbred lines. The observed phenotypes were highly variable and highly correlated. The sequencing of *Dwarf8* from these inbred lines led to the discovery of 41 distinct haplotypes with a variety of polymorphisms (alignment at http://www.stat.ncsu.edu/~panzea/ed/d8_final.nex). The diversity of *Dwarf8* sequence was low, particularly in the coding region (Table 1). *Dwarf8* was less diverse than other published maize genes, with the exception of the domestication gene, *teosinte branched1* (refs. 9,10).

We identified several interesting polymorphisms. These included a 485-bp insertion and adjacent 117-bp deletion (position 185) in the 5′ noncoding region; this may be a miniature transposable element (MITE; ref. 11). An 18-bp deletion (position 702) was found in the promoter. We found two insertions/deletions in the coding region; the first inserted a glycine codon at the amino terminus of the putative protein (position 1964) and the second eliminated two amino acids near the SH2-like domain (position 3472), a key binding domain in this class of transcription factors[6,12].

In general, linkage disequilibrium decays rapidly across *Dwarf8* (Fig. 1), as it does for most maize genes (D. L. Remington *et al.*, manuscript submitted). We observed, however, significant disequilibrium between the deletion near the SH2-like domain (position 3472) and nearby polymorphisms in the coding region, and the 485-bp insertion in the promoter and the associated 117-bp deletion (position 185; r^2=0.31). The decay of linkage disequilibrium at such a rapid rate enhances the resolution power of association tests, allowing us to resolve within a few thousand base pairs.

The distribution of segregating polymorphisms suggests that selection has occurred at this locus. We did not observe nonsynonymous substitutions in more than 10 of the inbred lines studied (Table 2); in contrast, 39% of the silent substitutions were found in 11 or more lines. The distribution of silent versus nonsynonymous polymorphisms was significantly different (permutation of contingency table; P=0.005), which suggests that selection has prevented the proliferation of nonsynonymous substitutions. This was confirmed by the Tajima's D test of selection[13] (Table 1; P<0.001). This would normally be viewed as evidence for the slightly deleterious nature of nonsynonymous polymorphisms, but it may also be the product of recent selective pressure to modify *Dwarf8* as breeders strive to produce maize shorter in stature and earlier flowering.

Table 1 • *Dwarf8* nucleotide diversity

Region	no. of bp	no. of polymorphic sites	% Sites polymorphic	Nucleotide diversity π/bp	Nucleotide diversity θ/bp (ref.17)	Tajima's D (ref.13)
5′ Noncoding region						
Insertions/deletions		37	4.01%	0.0050	0.0080	-1.20
Point mutations		48	5.21%	0.0070	0.0100	-0.94
Total	922	85	9.22%	0.0120	0.0180	-1.10
Exon *Dwarf8*						
Synonymous	444.67	19	4.27%	0.0056	0.0085	-1.03
Nonsynonymous	1298.33	19	1.46%	0.0004	0.0029	-2.54***
Total	1747	38	2.18%	0.0018	0.0044	-1.95*

Estimates of nucleotide diversity were calculated based on average pairwise diversity (π) and upon the number of segregating sites (θ). Tajima's D is a test for selection[13]. The 5′ noncoding region has a length of 1933 aligned bp, 1011 bp of which resulted from three large insertions/deletions (indels) that were excluded from diversity analyses. Coding region indels that insert or remove amino acids were scored as nonsynonymous polymorphisms. * P<0.05; *** P<0.001.

[1]*Department of Genetics, and* [2]*Department of Crop Science, North Carolina State University, Raleigh, NC 27695, USA.* [3]*Department of Genetics, University of Wisconsin, Madison, Wisconsin 53706, USA.* [4]*Department of Plant Breeding, Cornell University, Ithaca, New York 14853, USA.* [5]*Department of Statistics, and* [6]*USDA/ARS, North Carolina State University, Raleigh, North Carolina 27695-7614, USA. Correspondence should be addressed to E.S.B (email: buckler@statgen.ncsu.edu).*

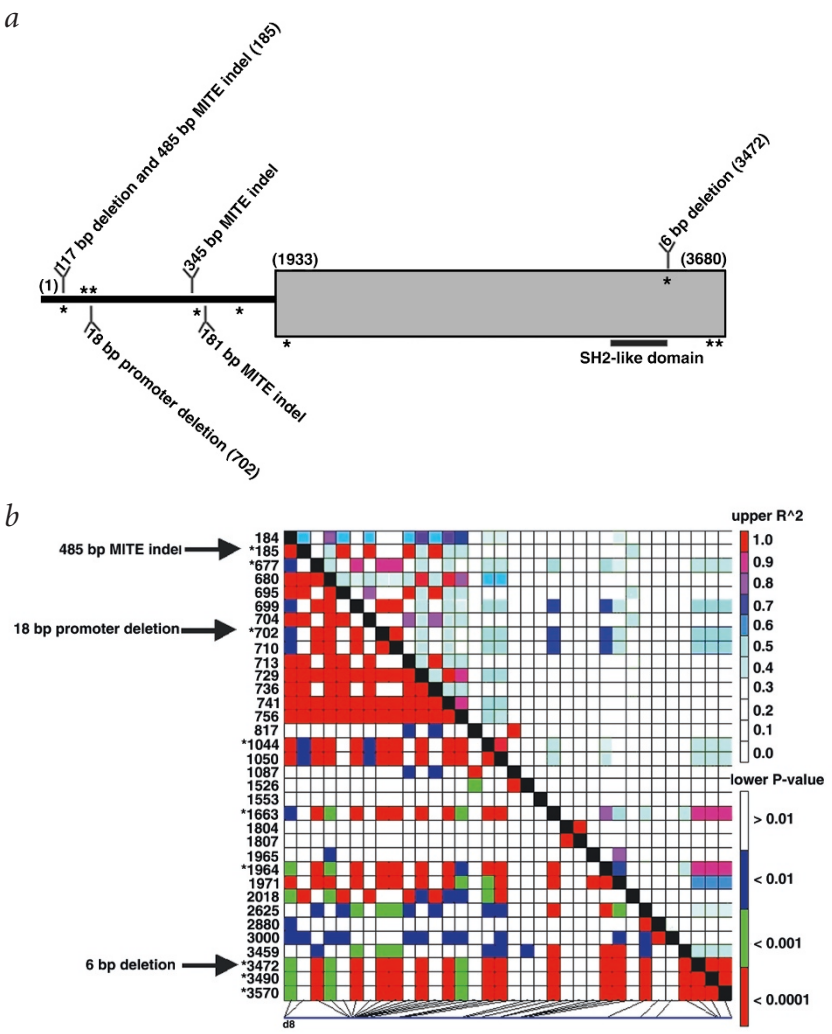

Fig. 1. Schematic diagram of *Dwarf8* gene structure and plot of linkage disequilibrium. ***a***, Major polymorphisms are highlighted, and sites significantly associated with flowering time are indicated (*). Positions relative to the sequence alignment used in this study are given in parentheses. ***b***, Linkage disequilibrium (r^2) between pairs of polymorphic sites at which the relevant alleles occurred in more than 5% of the lines was calculated (upper right), and significance was determined by a Fisher's exact test (lower left). Shading indicates the magnitude of the linkage disequilibrium and the significance level. The large block of linkage disequilibrium between aligned positions 184 and 756 is only 271 bases long if the large indel is removed.

Pritchard *et al.*[5] developed methods that control for population structure when testing for associations in a case-control study. We extended this test statistic (Λ) for quantitative traits. We calculated the likelihoods of two hypotheses using logistic regression: first, that the candidate gene distribution was associated with population structure and phenotypic variation, and second, that the candidate gene distribution was associated only with population structure. The test statistic (Λ) was the ratio of the two likelihoods. We estimated population structure by scoring 141 SSR loci, which were then used to estimate a genetic background matrix (Q) by Bayesian approaches[4]. In order to evaluate the type I error rate, all SSR alleles with a frequency greater than 0.05 were tested using the Λ statistic. When the *P* value threshold was set at 0.05, 8.1% of the SSR alleles were significant. Because of the slightly inflated type I error rate, we also re-scored the *P* values relative to the SSR *P* values. As some of the SSR are probably truly associated with the traits, the true *P* values are probably between $P(\Lambda)$ and $P_{SSR}(\Lambda)$.

For comparison, we evaluated what proportion of the SSR markers were significantly associated using logistic regression without any estimate of population structure. We observed a large difference in the proportion of SSR loci that associated by phenotype tested and by field plots. The type I error rate was, for example, 15.3% with respect to flowering time, 8.3% for plant height, and as high as 23.5% in one field. Therefore, estimates of population structure reduced the number of false positives by up to 4.7 fold.

Using this test statistic, Λ, we found significant associations between *Dwarf8* sequence polymorphisms and the traits tested (Table 3). Polymorphisms showed the most consistent association with flowering time traits. The number of days to silking was significantly associated in all fields, whereas days to pollen shed was significantly associated in four out of five field plots. Polymorphisms showed inconsistent associations with ear height and plant height.

We identified individual polymorphisms that associated with the developmental traits. Nine polymorphisms were significant at $P<0.05$ in all five field plots. These included the 117-bp deletion and 485-bp insertion caused by the putative MITE in the noncoding region (position 185), the 18-bp deletion in the noncoding region (position 702), a glycine insertion near the 5′ end of the coding region (position 1964), the 6-bp deletion just downstream of the SH2-like domain (position 3472), and other assorted point mutations (positions 677, 1044, 1663, 3490, and 3570; Fig. 1). These sites constitute a suite of polymorphisms that display significant levels of linkage disequilibrium with each other. The other nonsynonymous mutations identified, with the exception of the deletions mentioned previously, did not show any association with the traits studied.

Table 2 • Frequency of polymorphisms

	Silent polymorphisms		
Frequency of polymorphisms	5′ Noncoding region	Synonymous	Nonsynonymous
< 2%	28	10	12
2–11%	20	5	6
11–22%	28	2	0
>22%	9	2	0

Polymorphisms were divided into three classes: polymorphisms in the noncoding region, synonymous mutations, and nonsynonymous mutations in the coding region. The frequencies of the polymorphisms were determined and grouped into four categories; for example, 28 polymorphisms that occurred in the 5′ noncoding region were found in fewer than 2% of the lines tested. This was used in a permuted contingency test to examine the distribution of nonsynonymous and silent substitutions. Nonsynonymous substitutions had a very different frequency distribution from the silent substitutions (P=0.005).

Table 3 • Significance of *Dwarf8* polymorphism associations with plant phenotypes

			Entire gene		Deletion flanking SH2	
Phenotype	Field	$\ln(\Lambda_{max})$	$P(\Lambda_{max})$	$P_{SSR}(\Lambda_{max})$	Effect	Variation
Days to silking	Summer 1999, field A	9.00	0.002	0.014	-10±3 d	17%
	Summer 1999, field B	8.11	0.012	0.020	-9±3 d	12%
	Summer 1999, field C	7.08	0.008	0.017	-10±3 d	15%
	Winter 1999	7.67	0.010	0.024	-7±2 d	13%
	Winter 2000	8.56	0.014	0.041	-8±3 d	32%
Days to pollen	Summer 1999, field A	8.08	0.002	0.019	-11±3 d	19%
	Summer 1999, field B	8.03	0.012	0.027	-9±3 d	11%
	Summer 1999, field C	7.30	0.008	0.026	-11±3 d	13%
	Winter 1999	7.65	0.008	0.022	-7±2 d	10%
	Winter 2000	5.26	0.170	0.226	-9±3 d	25%
Ear height	Summer 1999, field A	4.86	0.204	0.250	-24±8 cm	14%
	Summer 1999, field B	7.95	0.096	0.133	-19±8 cm	8%
	Summer 1999, field C	4.71	0.014	0.039	-25±7 cm	17%
	Winter 1999	4.15	0.170	0.245	-11±5 cm	7%
	Winter 2000	8.32	0.014	0.027	-13±5 cm	22%
Plant height	Summer 1999, field A	4.34	0.182	0.226	-30±14 cm	8%
	Summer 1999, field B	5.33	0.078	0.097	-33±12 cm	9%
	Summer 1999, field C	5.08	0.064	0.077	-29±11 cm	9%
	Winter 1999	2.00	0.778	0.791	-19±9 cm	5%
	Winter 2000	5.38	0.124	0.173	-28±11 cm	21%

The Λ_{max} was determined for each field planting, and its experiment-wise *P* value was determined as described in the Methods. As a means of comparison, the effect of one polymorphism, the 6-bp deletion near the SH2-like domain (site 3472), was estimated in a regession model that included estimates of population structure (*Q*). The amount of variation was determined as the proportion of type III sum of squares explained by the polymorphism.

Some of these polymorphisms could have an impact upon the expression or function of *Dwarf8*. For example, it is possible that the deletion of two amino acids near the SH2-like domain (position 3472) may have an effect upon the transcription factors' activity. Additionally, insertions and deletions, such as the 485-bp MITE insertion and the deletion in the promoter (position 702), may have an effect upon the *Dwarf8* expression level. That these polymorphisms were in linkage disequilibrium with each other made it difficult for us to differentiate individual polymorphisms that are responsible for the variation in phenotype. Future studies of randomly mated maize populations should allow us to confirm the association between these polymorphisms and the traits studied.

We used regression analyses that included population structure to estimate the effect of these polymorphisms on flowering time and plant height. The deletion of two codons flanking the SH2-like domain, for example, had an estimated effect of reducing the time for flowering by 7–11 days across the multiple field plots tested (Table 3).

To serve as a comparison, the maize gene *teosinte branched1* (*tb1*) was sequenced from the same inbred lines (D. L. Remington *et al.*, manuscript submitted). Association tests did not show any association between *tb1* and flowering time or plant height, despite the proximity of *tb1* to *Dwarf8* (1 cM apart). Recombination has provided sufficient resolution to distinguish the effects of polymorphisms in each of these genes, and the statistical model seems to control the effects of population structure successfully.

Our results do not indicate a significant association between *Dwarf8* polymorphisms and plant height. The *dwarf8* mutants with severe height phenotypes identified in mutagenesis screens are the result of alterations of the DELLA domain at the N terminus of the predicted protein[6]. These mutations are dominant, gain-of-function mutations, which prevent *dwarf8* being released from its role as a negative regulator[14]. The DELLA domain is entirely conserved in all of the inbreds used in this study. Mutations of the *Arabidopsis* ortholog, *gai*, have been identified that are recessive, loss-of-function mutations, and produce polypeptides truncated upstream of the SH2-like domain[15]. The gene product no longer functions as a negative regulator, producing plants of normal height[14]. The alleles of *Dwarf8* that we have identified may be more analogous to those of this loss-of-function mutant.

In summary, the effect of structured populations has presented a serious obstacle to the use of association studies in crop plants[1]. We have demonstrated that this obstacle may be circumvented if the sampling of unlinked markers and statistical methods are used to account for population structure. We have used a method to incorporate estimates of population structure[4] in association tests[5] and extended it for use with quantitative traits such as maize flowering time. Using this method, we identified a suite of polymorphisms in the positional candidate gene *Dwarf8* that associate with differences in flowering time. Association methods that incorporate estimates of population structure in this manner should provide a powerful approach to identifying the alleles responsible for variation in a variety of quantitative traits.

Methods

Plant materials. We used a total of 92 inbred maize lines in this study to represent some of the diversity currently available; these can be subdivided into three major groups. Twelve lines fall into the stiff stalk and 45 into the non-stiff stalk group, 35 being tropical or semitropical lines, which are related to the non-stiff stalk lines (D. L. Remington *et al.*, manuscript submitted).

Field tests were conducted at two sites, one near Clayton, NC, USA (summer nursery), and the other near Homestead, FL, USA (winter nursery). The flowering time was measured in terms of the number of days to pollen shed and days to silking. Plant and ear height were measured after flowering. We collected five separate sets of measurements involving some or all lines during winter 1998 and winter 1999, and three separate plots in the summer of 1999.

Candidate gene sequence data. DNA sequence data were obtained for *Dwarf8*. We designed primers for the PCR amplification of gene fragments from published sequences in GenBank and unpublished promoter sequence (N. Harberd, personal communication). Gene fragments were PCR-amplified from each of the 92 lines. The products were cloned into pCR-TOPO2.1 vectors (Invitrogen) for sequencing. Plasmid preparations from 2–4 colonies from each PCR product were either sequenced separately or pooled prior to sequencing to minimize the contribution of polymerase errors to sequence variation. Consensus sequence contigs generated using SeqMan (DNAStar) were edited manually to resolve discrepancies. Consensus sequences for the entire set of lines were aligned using the clustal alignment option in MegAlign (DNAStar), with further

manual alignment. Chromatograms were rechecked for all singleton polymorphisms in order to distinguish true polymorphisms from probable polymerase or scoring errors. Complete *Dwarf8* sequence was obtained from each of the 92 inbred lines (alignment at http://www.stat.ncsu.edu/~panzea/ed/d8_final.nex).

Sequence analysis and statistical tests. We estimated nucleotide diversity under the infinite site model from the neutral theory as θ (ref. 16). Average pairwise diversity (π) is the average proportion of nucleotide differences between all possible pairs of sequences in the sample[17]. Tests for selection were based upon the methods of Tajima[13]. We used multiple software packages for these analyses; DNASP[18], SITES[19], and TASSEL (Buckler, software at http://statgen.ncsu.edu/buckler/bioinformatics.html). Linkage disequilibrium (r^2) between pairs of polymorphic sites with allele frequencies over 5% was calculated, and significance was determined by a Fisher's exact test.

To evaluate the associations, we followed a method that dealt with structured populations[4,5]. First, population structure was estimated using a Bayesian approach that estimated membership to various subpopulations by examining genotypic correlations at unlinked markers[4]. This approach estimated the proportion of an individual's genome that was contributed by each subpopulation, a genetic background matrix (Q). We used the set of 141 SSR loci to estimate the Q matrix (D. L. Remington *et al.*, manuscript submitted). High likelihoods of the population structure were observed when the number of subpopulations was set to three (D. L. Remington *et al.* manuscript submitted). The optimal three-subpopulation model and Q matrix was used in subsequent association analyses.

Pritchard *et al.*[5] developed their association approach for case-controls, and we have therefore modified their test statistic to deal with quantitative traits. In the null hypothesis, H_0, candidate polymorphisms are independent of phenotype, whereas in the alternative hypothesis, H_1, candidate polymorphisms are associated with the phenotype[5]. The probability of each hypothesis is compared in the following way.

$$\Lambda = \frac{\Pr_1(C;T,\hat{Q})}{\Pr_0(C;\hat{Q})}$$

C is the genotype of the candidate polymorphism for all lines, and T is the trait value for all lines. These two probabilities were estimated using logistic regression of the model (SAS), in which the response variable was the presence or absence of the candidate polymorphism, and T and Q were used as the independent variables. Q was modeled using the admixture model[4]. Significance was determined by simulation of C (ref. 5). First, the maximum value of Λ was determined over all sites. Then the phenotypic trait values were permuted relative to the haplotypes. Using this permutated data set, Λ_{max} was calculated again. The observed Λ_{max} was compared with the permuted $\Lambda_{max.}$ The proportion of random samplings with a Λ_{max} less than or equal to the observed Λ_{max} was the experiment-wise P value. Five hundred randomizations were carried out.

In order to test for a type I bias in this statistic, all 588 SSR alleles with a frequency greater than 5% were tested. This distribution of P values was also used to rescale candidate gene P values. The rescaled P value was the proportion of SSR P values that were less than or equal to the observed permuted P value.

Acknowledgments

We would like to thank T. Helentjaris for his input on this project, and N. Harberd and Plant Bioscience Ltd. for providing genomic sequence. D. Remington provided the tb1 sequence alignments and contributed helpful discussions. We would like to thank anonymous reviewers for their thoughtful comments regarding this manuscript. All sequencing was performed at the North Carolina State University Genome Research Laboratory. This work was supported by NSF (DBI-9872631) and USDA-ARS.

Received 6 March; accepted 18 May 2001.

1. Lander, E.S. & Schork, N.J. Genetic dissection of complex traits. *Science* **265**, 2037–2048 (1994).
2. Templeton, A.R. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping or DNA sequencing. V. Analysis of case/control sampling designs: Alzheimer's disease and the apoprotein E locus. *Genetics* **140**, 403–409 (1995).
3. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
4. Pritchard, J.K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
5. Pritchard, J.K., Stephens, M., Rosenberg, N.A. & Donnelly, P. Association mapping in structured populations. *Am. J. Hum. Genet.* **67**, 170–181 (2000).
6. Peng, J. *et al.* 'Green revolution' genes encode mutant gibberellin response modulators. *Nature* **400**, 256–261 (1999).
7. Koester, R., Sisco, P. & Stuber, C. Indentification of quantitative trait loci controlling days to flowering and plant height in two near-isogenic lines of maize. *Crop Sci.* **33**, 1209–1216 (1993).
8. Schon, C. *et al.* RFLP mapping in maize – quantitative trait loci affecting testcross performance of elite European flint lines. *Crop Sci.* **34**, 378–389 (1994).
9. Wang, R.-L. *et al.* The limits of selection during maize domestication. *Nature* **398**, 236–239 (1999).
10. White, S. & Doebley, J. The molecular evolution of *terminal ear1*, a regulatory gene in the genus Zea. *Genetics* **153**, 1455–1462 (1999).
11. Wessler, S.R., Bureau, T.E. & White, S.E. LTR-retrotransposons and MITEs: important players in the evolution of plant genomes. *Curr. Opin. Genet. Dev.* **5**, 814–821 (1995).
12. Koch, C.A., Anderson, D., Moran, M.F., Ellis, C. & Pawson, T. SH2 and SH3 domains: elements that control interactions of cytoplasmic signaling proteins. *Science* **252**, 668–674 (1991).
13. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989).
14. Kende, H. Hormone response mutants. A plethora of surprises. *Plant Physiol.* **125**, 81–84 (2001).
15. Peng, J. *et al.* The Arabidopsis GAI gene defines a signaling pathway that negatively regulates gibberellin responses. *Genes Dev.* **11**, 3194–3205 (1997).
16. Watterson, G. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**, 256–276 (1975).
17. Nei, M. *Molecular Evolutionary Genetics* (Columbia University Press, New York, 1987).
18. Rozas, J. & Rozas, R. DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics* **15**, 174–175 (1999).
19. Hey, J. & Wakeley, J. A coalescent estimator of the population recombination rate. *Genetics* **145**, 833–846 (1997).