Molecular Characterization of a Diverse Maize Inbred Line Collection and its Potential Utilization for Stress Tolerance Improvement

Weiwei Wen, Jose Luis Araus, Trushar Shah, Jill Cairns, George Mahuku, Marianne Bänziger, Jose Luis Torres, Ciro Sánchez, and Jianbing Yan*

ABSTRACT

A diverse collection of 359 advanced maize (*Zea mays* L.) inbred lines from the International Maize and Wheat Improvement Center (CIMMYT) and International Institute for Tropical Agriculture (IITA) breeding programs for drought, low N, soil acidity (SA), and pest and disease resistance was genotyped using 1260 single nucleotide polymorphism (SNP) markers. Model-based population partition, neighbor-joining (NJ) clustering, and principal component analysis (PCA) based on the genotypic data were employed to classify the lines into subgroups. A subgroup largely consisting of lines developed from La Posta Sequía (LPS) consistently separated from other lines when using different methods based on both SNP and SNP haplotype data. Lines related by pedigree tended to cluster together. Nine main subsets of lines were determined based on pedigree information, environmental adaptation, and breeding scheme. Analysis of molecular variance (AMOVA) revealed that variation within these subsets was much higher than that among subsets. Genetic diversity and linkage disequilibrium (LD) level were tested in the whole panel and within each subset. The potential of the panel for association mapping was tested using 999 SNP markers with minor allelic frequency (MAF) ≥ 0.05 and phenotypic data (grain yield [GY], ears per plant [EPP], and anthesis to silking interval [ASI]). Results show the panel is ideal for association mapping where type I error can be controlled using a mixed linear model ($Q + K$). Use of pedigree, heterotic group, and ecological adaptation information together with molecular characterization of this panel presents a valuable genetic resource for stress tolerance breeding in maize.

International Maize and Wheat Improvement Center (CIMMYT), Apartado Postal 6-640, 06600 Mexico, DF, Mexico; J.L. Araus, Unitat de Fisiologia Vegetal, Facultat de Biologia, Universitat de Barcelona, 08028 Barcelona, Spain; J. Yan, National Key Lab. of Crop Improvement, Huazhong Agricultural Univ., Wuhan 430070, Hubei, China; T. Shah, current address: International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru Hyderabad, Andhra Pradesh, India 502324. Received 9 Aug. 2010. *Corresponding author (yjianbing@gmail.com).

Abbreviations: AMOVA, analysis of molecular variance; ASI, anthesis to silking interval; CML, CIMMYT Maize Line; DA, days to anthesis; DS, days to silking; DTP, drought tolerant population; EPP, ears per plant; GWS, genome-wide selection; GY, grain yield; LD, linkage disequilibrium; LnP(D), natural logarithm probability of data; LPS, La Posta Sequía; MAF, minor allelic frequency; MAS, marker-assisted selection; NJ, neighbor-joining; PCA, principal component analysis; PIC, polymorphic information content; SA, soil acidity; SNP, single nucleotide polymorphism; SP, subsets selected by pedigree; SR, subsets selected randomly; Qh, *Q* value calculated based on haplotype data; Qs, *Q* value calculated based on single nucleotide polymorphism data.

It is predicted that cereal production will need to rise to over 400 million t by 2050 to meet the demands, which represents an increase in yield of 37% from current values (Tester and Langridge, 2010). However, this is a challenge, especially for smallholders and resource-poor farmers, most of whom grow maize (*Zea mays* L.) on marginal lands with low inputs and constant exposure to a variety of abiotic (drought, low N, acid soils, waterlogging, and heat) and biotic (pests, diseases, and weeds) stresses (Mugo et al., 2008). Currently, drought stress accounts for a significant percentage of annual yield losses (Bruce et al., 2002) and this is compounded by pests

Published in Crop Sci. 51 (2011).
doi: 10.2135/cropsci2010.08.0465
Published online 16 Sept. 2011.

and diseases. Given the increasing demand for maize production and limited land, maize germplasm improvement using effective breeding technologies is critical.

Broadening the genetic base by creating stress tolerant germplasm is expected to result in improved varieties that are resilient to harsh environmental conditions found in most developing countries. To develop stress tolerant germplasm, CIMMYT germplasm is tested under varying environmental conditions through on-farm regional trials (both researcher managed and farmer managed). Recurrent selection, where progeny selection is based on international testing, has resulted in improved mean performance and stability of CIMMYT's tropical maize populations (Pandey and Gardner, 1992). The destabilizing impacts of climate change, especially the increasing frequencies of drought, has resulted in CIMMYT emphasizing research to develop drought tolerant and water use efficient maize. Various numbers of cycles of selection under drought have been completed in different populations and good source drought tolerant populations have been developed (Edmeades et al., 1997, 1999). Intensive inbreeding efforts have also been underway to develop drought tolerant inbred lines and a few have been released as CIMMYT Maize Lines (CMLs). Importantly, selected drought tolerant materials also performed well under low N conditions (Bänziger et al., 1999, 2002; Monneveux et al., 2006). Secondary traits associated with stress tolerance have been emphasized for accelerating breeding gains. Bänziger et al. (2000) proposed barrenness, anthesis to silking interval (ASI), leaf senescence, and leaf rolling as secondary traits useful for improving maize yields in drought-prone environments. Selection for tolerance to low N and drought at CIMMYT has emphasized selection indices to combine data on grain yield under stressed and stress-free conditions with information of secondary traits.

In addition to drought and low N, soil acidity is an important factor accounting for genotype × environment interactions (Setimela et al., 2005). Six acid soil tolerant maize populations (SA3, SA4, SA5, SA6, SA7, and SA8, where SA represents soil acidity) have been developed by CIMMYT maize researchers at Cali, Columbia (Pandey et al., 1994, 1997; Narro et al., 1997). Products of these researches include released acid soil tolerant hybrids and open-pollinated varieties and the identification of acid soil tolerant CMLs (Mugo et al., 2008).

The ability to quickly develop maize germplasm combining tolerance to multiple stresses will be essential in the face of climate change. While conventional breeding has increased yields under both abiotic and biotic stresses, progress is slow. The availability of maize genome sequence (Schnable et al., 2009), extensive structural variation, and high levels of genomic diversity within maize (Gore et al., 2009; Springer et al., 2009) provide great potential to exploit natural maize genetic diversity for practical application within breeding programs. The use of single nucleotide polymorphisms (SNPs) as molecular markers is now widely used in many crops (Rostoks et al., 2006; Hyten et al., 2008; Muchero et al., 2009), including maize (Buckler et al., 2009; Yan et al., 2009, 2010a). To date, more than one million maize SNPs are publicly available (Gore et al., 2009). Abundant genetic diversity, rapidly developing high-throughput genotyping technology, and advanced statistical methods facilitate the application of association mapping in maize, which is a powerful tool to dissect complex agronomic traits and identify alleles that can contribute to the enhancement of a target trait (Yu et al., 2006; Harjes et al., 2008; Buckler et al., 2009; Yan et al., 2011). Advances in technology will allow marker-assisted selection (MAS) to be deployed with greater efficiency to pyramid favorable alleles.

In this study, a set of 359 advanced maize inbred lines was assembled mainly from drought, low N, soil acidity, and pest and disease resistance breeding programs at CIMMYT and IITA. We characterized this maize panel using 1260 SNP markers to investigate its genetic diversity; dissect the genetic structure, assess linkage disequilibrium (LD) level and familial relatedness, evaluate various models for association analysis based on this panel, and discuss the marker-based strategy for utilizing this panel in improvement of maize stress tolerance.

MATERIALS AND METHODS

Plant Material and Phenotyping Condition

A collection of maize germplasm was assembled representing the genetic diversity of CIMMYT and IITA's stress breeding programs (drought, disease, acid soils, low N, and entomology). An initial set of 850 advanced breeding lines was selected and evaluated in the dry season of 2006 under two different water regimes (well watered and anthesis stage drought stress) at CIMMYT's experimental station in Tlaltizapán, Mexico (18°41' N, 99°07' W, and 940 m above sea level).

Entries were planted in December 2006 in one-row plots (2.5 m rows with 0.25 m spacing between plants and 0.75 m between rows), with a final plant density of 6.67 plants m^{-2}. An α-lattice design replicated two times was used. The control plots (well watered) were irrigated through the crop cycle while in the drought stress treatment water was withheld 2 to 3 wk before anthesis. Drought stress plots received only one further irrigation 1 wk after anthesis. At physiological maturity, all the ears (here, an ear was defined as having one or more grains on it) of each plot were hand harvested, counted, air dried, and shelled and the grains were weighed. Grain yield (GY) and ears per plant (EPP) were then calculated.

In total, 246 lines from the initial 850 lines were selected based on grain yield performance under well watered and drought stress conditions. Also, 32 lines scored for tolerance to low N, 68 disease resistant lines, and a set of 13 potential testers were added to the panel. Disease resistance lines exhibited resistance to Northern corn leaf blight (*Exserohilum turcicum*), Southern corn leaf blight (*Bipolaris maydis*), Fusarium ear rots caused by *Fusarium verticillioides*, Aspergillus ear rot (*Aspergillus*

Table 1. Summary of the origin, source, and grain characteristics of maize lines within the panel.

Group identification	Breeding program	Number of lines	Main sources†	Grain color	
				White	Yellow
A	Zimbabwe	44	CML, CIMCALI, and DTPW	44	0
B	Nigeria	5	KU and P43	3	2
C	Ethiopia	2	Pool9	2	0
D	Colombia	27	SA3, SA4, SA5, SA6, SA7, and SA8	4	23
E	Mexico highland	5	A.T.Z.T.R.L.BA90	1	4
F	Mexico entomology	39	CML, MBR, ZM607, KILIMA, and P84	28	11
G	Mexico subtropical	31	CML, MBR, SPMAT, Pop 33, Pop 45, Pop 501, and Pop 502	18	13
H	Mexico tropical	41	CML, CLQ, and CL	22	19
I	Selection under drought	52	DTPW, DTPY, and LPS	41	11
J	Selection under low N	32	DTPW, DTPY, and LPS	24	8
K	Tester lines	13	CML	9	4
L	Pathology	68	CML, CLQ, CL, DTPW, DTPY, and LPS	50	18
Total		359		246	113

†A.T.Z.T.R.L., Amarillo Tardio Zona de Transition, Recombinacion de Lines; CIMCALI, CIMMYT lines from Cali, Columbia; CL, CIMMYT lines; CLQ, CIMMYT lines for grain quality; CML, CIMMYT Maize Line; DTPW, drought tolerant population white grain; DTPY, drought tolerant population yellow grain; KU, Kasetsart University in Thailand; KILIMA, a late-maturing, white grain open pollinated variety in Tanzania; LPS, La Posta Sequía; MBR, multiple borer resistant; Pool, CIMMYT gene pool; Pop, CIMMYT Population; SA, soil acidity; SPMAT, semiprolific mid altitude; ZM, population improved in Zimbabwe maize breeding program of CIMMYT.

flavus), common rust (*Puccinia sorghi*), corn stunt complex, Polysora rust (*Puccinia polysora*), or Maize streak Virus. Therefore, a final panel containing 359 lines was assembled to conduct molecular analysis. The lines were derived from 12 different breeding programs: national or regional programs in Zimbabwe (n = 44), Nigeria (n = 5), and Ethiopia (n = 2), and international (CIMMYT and IITA) breeding programs for acid soil tolerance (n = 27), highland adaptation (n = 5), insect resistance (n = 39), disease resistance (n = 68), subtropical adaptation (n = 31), tropical adaptation (n = 41), and physiology (n = 84) plus the set of 13 CIMMYT testers (Table 1). Detailed information of the 359 lines is listed in Supplemental Table S1. Information of the major original sources of these lines is described in Supplemental Table S2.

In December 2007, a subset of 253 lines (which were marked in Supplemental Table S1) from the 359 lines were planted in one-row plots (5 m rows with 0.25 m spacing between plants and 0.75 m between rows) under two different water regimes (well watered and anthesis stage drought stress) at CIMMYT's experimental station in Tlaltizapán. Two seeds per hill were sown and then thinned to one 14 d after sowing. An α-lattice design replicated two times was used and the field management was as outlined above. Days to anthesis (DA) and silking (DS) were recorded for each plot when at least 50% of the plants have reached anthesis or silking, respectively. Then ASI was calculated as DS minus DA. Grain yield and EPP were measured as mentioned above. For each genotype, ASI, GY, and EPP were averaged from the two replications and then used for association analysis.

Single Nucleotide Polymorphism Genotyping and in silico Mapping

Genotyping of the selected panel was conducted using an Illumina oligo pool assay with 1536 SNPs developed by Yan et al. (2009) and improved by Wen et al. (2011). Briefly, a total of 943 SNPs with minor allelic frequency (MAF) greater than 0.05 and good quality based on the results in 632 diverse lines (Yan et al., 2009) were combined with 593 SNPs selected from the Panzea database (Panzea, 2009; E. Buckler, personal communication, 2009) on the basis of having a designability score higher than 0.5. Designability scores of SNPs were provided by Illumina (Illumina, Inc., San Diego, CA), with a score greater than 0.5 indicating a higher probability of success in the GoldenGate assay. We used the Illumina BeadStation 500 G (Illumina, Inc., San Diego, CA) for SNP genotyping according to the protocol described by Fan et al. (2006). Each SNP was rechecked manually and rescored if any error was observed in the clustering of homozygous and heterozygous groups (Yan et al., 2010a). One thousand three hundred forty-two out of 1536 SNPs (87.4%) were successfully called with less than 20% missing data. Within the 1342 SNPs, 16 SNPs with a heterozygous rate of more than 20% were excluded from further study. Of the remaining 1326 SNPs, 62 were monomorphic in all the lines and these were excluded from further analysis.

The remaining 1264 SNP reference sequences were used to perform a BlastN search against the maize accessioned golden path version 1 for B73 (Arizona Genomics Institute, 2009). We considered the top blast hits with an e-value threshold of 10^{-18}. A total of 1260 successfully called SNPs were used to construct a unigene set by subtracting four SNPs with no hits. According to the method from Yan et al. (2009), SNPs from the same locus (a locus here was defined as a region containing SNPs located within 10 kbp of each other) were grouped into haplotypes, which were recorded as alleles; in this way, each locus could have multiple alleles, raising the information content of the markers. If the genotype of any SNP at a locus was missing in an individual, the haplotype was regarded as missing in that individual.

Genetic Structure Analysis

Population structure based on 1260 successfully called SNPs, and the unigene SNPs together with SNP haplotypes was inferred using the model-based program STRUCTURE (Pritchard et al., 2000; Falush et al., 2003). The number of subpopulations (k) was assumed to be from 1 to 12, without

admixture and with correlated allele frequencies, and the burn-in time and iterations for each run were both set to 50,000. Ten replications were used for each k. Due to the difficulties associated with finding the highest posterior probability (i.e., natural logarithm probability of data [LnP(D)]) before a large k value is examined, both LnP(D) value and Evanno's Δk (Evanno et al., 2005) were used to determine the most appropriate k value. Evanno's Δk considers the rate of change of LnP(D) as k increases and also the variance of LnP(D) among repeated runs and tends to be maximum at the true value of k. It is calculated as $\Delta k = M[|L(k-1) - 2L(k) + L(k+1)|]/S[L(k)]$, in which $L(k)$ represents the kth LnP(D), M is the mean of 10 runs, and S is their standard deviation. Membership probability in each cluster (Q value) for each line was estimated in each k. Among the 10 runs per k, the one with the highest maximum likelihood was used to assign individual genotypes to clusters. Lines with membership probabilities ≥0.60 were assigned to corresponding clusters, and lines with membership probabilities <0.60 were assigned to a mixed group.

Principal component analysis (PCA) was performed using the software NTSYSpc (Darroch and Mosimann, 1985) to visualize genetic relationships between maize lines, which was based on the Nei's genetic distances (Nei, 1972) among all the lines. A phylogenic tree was constructed using the neighbor-joining (NJ) method in the software MEGA V4.0 (Tamura et al., 2007). Relative kinship between individuals was inferred from molecular markers, approximately reflecting the identity between two given individuals over the average probability of identity between two random individuals (Yu et al., 2006). We used SPAGeDi (Hardy and Vekemans, 2002) to estimate the kinship coefficients (Loiselle et al., 1995) based on 1260 successfully called SNPs.

Pedigree information, environmental adaptation, and breeding program origin of the lines were used to compare with classification based on marker data.

Analysis of molecular variance (AMOVA) (Excoffier et al., 1992) and pairwise F-statistics were performed using Arlequin V3.11 (Excoffier et al., 2005) to investigate population differentiations among the subpopulations classified by different methods.

Diversity, Linkage Disequilibrium, and Association Analysis

Polymorphic information content (PIC), the relative value of each marker with respect to the amount of polymorphism exhibited, and gene diversity were estimated in each subset using the software PowerMarker V3.25 (Liu and Muse, 2005).

Linkage disequilibrium was estimated for all pairs of polymorphic SNPs from the same chromosome with less than 20% missing data. The parameter r^2 was calculated using the program TASSEL 2.1 (Bradbury et al., 2007). Linkage disequilibrium was computed separately for different subsets of lines and markers. Linkage disequilibrium was calculated for all 1260 SNP markers and subset of SNPs with MAF greater than 0.05 or 0.1 separately. Mean r^2 values were calculated between SNPs, which have physical distances in different ranges. Six subsets of different sample size (n = 7, 24, 31, 41, 47, and 69) were randomly selected from the 359 lines with 10 repetitions to detect the effect of sample size and population background on the extent of LD using 1260 SNPs.

Association analysis was conducted for three traits (GY, EPP, and ASI) and 999 SNPs with MAF greater than 0.05. The traits per se of 253 lines were used to test appropriate model for genome wide association analysis considering the population structure (Q), relative kinship (K) and $Q + K$ respectively (Yu et al., 2006).

RESULTS

Panel Composition and Phenotypic Performance

The initial panel was composed of 850 inbred lines. The origin of these lines is summarized in Supplemental Fig. S1a. From this set 246 lines were selected, plus an additional 113 lines from other sources, giving a total of 359 lines (Supplemental Fig. S1b) that were used for molecular analysis. The composition of this panel is summarized in Table 1. The majority of lines were developed from the drought-tolerant population (DTP), La Posta Sequía (LPS), SA, and elite CMLs. Two hundred forty-six lines had white grain while 113 lines had yellow grain (further information is described in Supplemental Tables S1 and S2).

Under drought stress, flowering synchrony was reduced (i.e., increased ASI) compared to the well-watered control (Supplemental Table S3). Grain yield under well-watered conditions was on average more than four times higher than under drought stress. Ears per plant ranged from 0.8 to 1.4 (mean = 1.1) under well watered conditions and 0.4 to 1.3 (mean = 0.9) under drought stress (Supplemental Table S3).

Performance and Quality of Single Nucleotide Polymorphism Genotyping

A total of 1260 polymorphic SNPs was used for final data analysis (Supplemental Tables S4 and S5). The 1260 SNPs were evenly distributed across the whole genome, with coverage ranging from 74 SNPs on chromosome 9 to 218 SNPs on chromosome 1. The SNP data sets were biallelic, and the loci that have rare alleles (with allelic frequency <0.05) were more common than those with intermediate frequency alleles (with allelic frequency between 0.4 and 0.5) (Fig. 1a; Table 2). The average heterozygosity of each line was 3.1%, well within expected ranges for residual heterozygosity found in inbred maize lines. In 26 lines heterozygosity was greater than 10% (Supplemental Table S6) indicating selfing is still required to reduce residual heterozygosity. Heterozygosity rate of each SNP across all the inbred lines ranged from 0 to 17.5% with an average of 3.1%.

A total of 273 SNP haplotypes that contained two or more SNPs and 492 single SNPs from the unigene analysis for all 1260 SNPs were identified, corresponding to 765 loci with an average 1.6 SNPs per locus. Among SNP haplotypes, a total of 1317 alleles was identified, ranging from 2 to 36 alleles per locus with an average of 4.8 alleles per locus. The majority of haplotype alleles were detected at low frequency within the panel; more than half had an allelic frequency of less than 0.2 (Fig. 1b).

Genetic Structure and Relative Kinship among Inbred Lines

Model-based substructure partition and PCA based on biallelic SNPs or multiallelic SNP haplotypes identified population structure within these 359 lines. In the model-based results, the value of LnP(D) increased gradually from $k = 1$ to $k = 12$ for both SNP and SNP haplotype data (Fig. 2a). At $k = 2$ and $k = 3$, Δk was much higher than other k values for SNP data. When $k = 3$, the partition of the panel based on SNP data was consistent with PCA results (Fig. 3a). Evanno's Δk peaked at $k = 2$ for the haplotype data (Fig. 2b) with good agreement to the assignment of lines to each subgroup based on PCA (Fig. 3b). A subgroup (Group 1) that largely contained lines derived from LPS was clearly and consistently separated from other groups when using both model-based and PCA methods based on both SNP or haplotype data (Fig. 3a and b). Lines classified in Group 2 based on haplotype data (Fig. 3b) were separated into two groups (Group 2 and Group 3) by biallelic SNP data (Fig. 3a).

Lines closely related by pedigree or sharing environmental adaptations or program origins were classified into corresponding subsets. A total of nine main subsets were identified based on pedigree and origin information, ranging from 7 to 69 lines per subset. The nine subsets corresponded to the germplasm of CIMCALI8843 × S9243, CML311 × MBR C3 Bc, DTPWC9, DTPYC9, LPS C7, acid soil tolerant maize population from CIMMYT (SA), Mexico subtropical, Mexico tropical, and CML (Table 3). When lines from Group 1 were excluded (Fig. 3a and b), no clear substructure was observed in the remaining lines; however, there was a tendency for lines related by pedigree to cluster together (Fig. 3c and d; Supplemental Fig. S2).

Relative kinship between lines was estimated using the 1260 informative SNP markers and it ranged from 0 to 1 with a mean of 0.0196 across all the pairwise values. Approximately 60% of the pairwise kinship estimates were around 0, indicating no relationship between these lines (Fig. 4). The mean of relative kinship within the nine subsets ranged from 0.026 of subset 7 (Mexico subtropical) to 0.838 of subset 1 (CIMCALI8843 × S9243), which was in good accordance with the known pedigree and also clearly reflected the familial relationship among the lines within each subset (Table 3). Different subpopulations were classified by different methods or criterion. The AMOVA revealed that for the model-based classification based on SNP and SNP haplotype, 6.2 and 10.6% of the molecular variation were found among the populations and 93.8 and 89.4% were found within populations, respectively. For the nine subsets based on pedigree and source, we found 11.8% of the total genetic variation was partitioned among the populations and 88.3% within populations (Table 4).

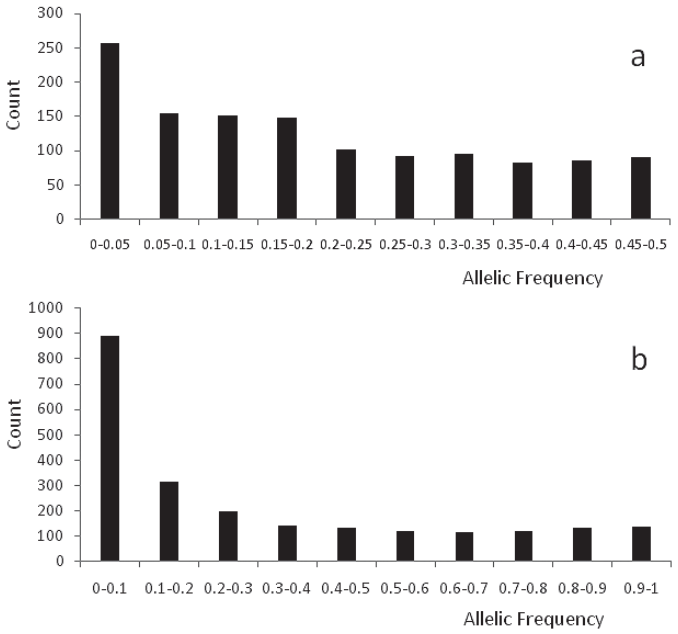

Figure 1. (a) Allelic frequency for total single nucleotide polymorphisms (SNPs). (b) Allelic frequency for SNP haplotypes.

Table 2. Summary of single nucleotide polymorphisms (SNPs) and SNP haplotypes identified within each chromosome.

		Number of loci†		Minor allelic frequency	
Chromosome	Number of SNPs	Two or more SNPs‡	One SNP§	≥0.05	≥0.1
1	218	49	76	168	141
2	159	31	65	137	110
3	141	24	71	113	99
4	130	28	52	95	77
5	147	36	61	121	104
6	76	16	34	55	45
7	112	28	34	91	77
8	116	25	40	94	84
9	74	21	22	58	52
10	87	15	37	71	58
Total	1260	273	492	1003	847

†Single nucleotide polymorphisms within 10 kbp region were combined and identified as a locus.

‡Locus contains two or more SNPs.

§Locus contains only one SNP.

Genetic Diversity of Subsets and Polymorphism between Lines

The number of polymorphic loci, the corresponding PIC values, and gene diversity within each subset are summarized in Table 3. Subset 1, originating from CIMCALI8843 × S9243, contained seven lines with only 118 polymorphic loci detected between lines that had the lowest PIC value (0.028) and gene diversity (0.033) among all subsets. Subset 8 (CIMMYT and Mexican tropical) had the highest PIC value (0.304) and gene diversity (0.259). Subset 9, comprising lines developed from CMLs, had the

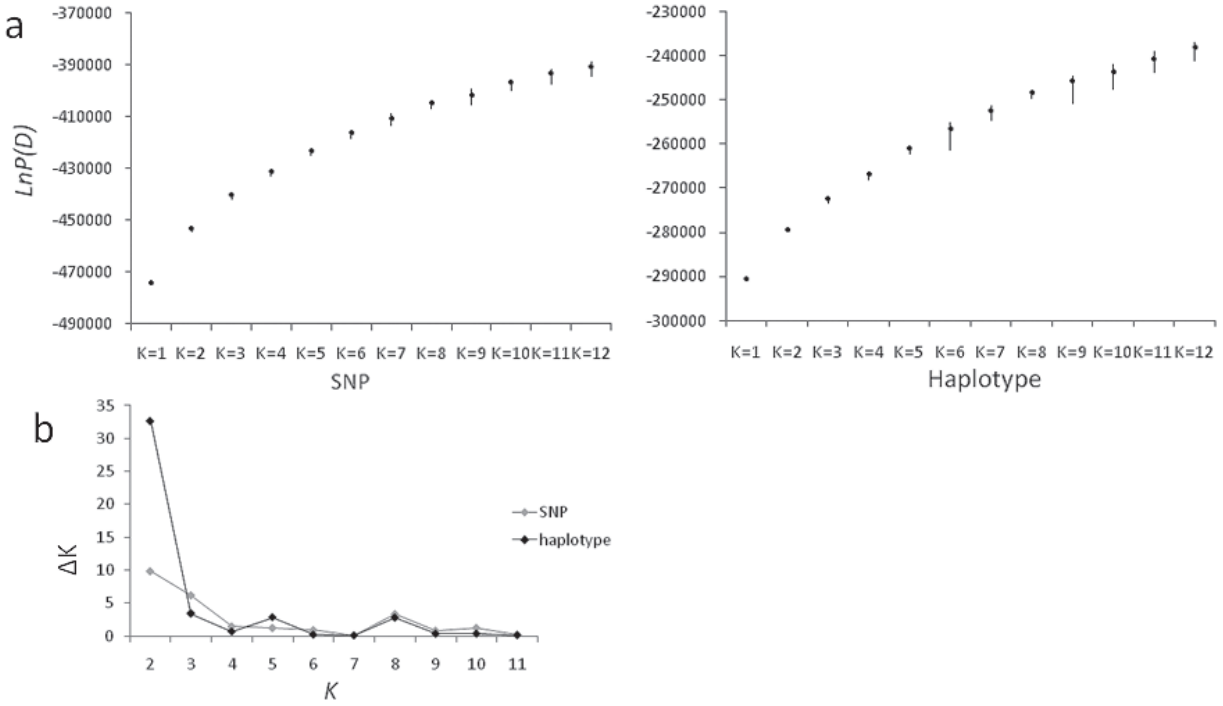

Figure 2. (a) The mean and range of the natural logarithm probability of data [LnP(D)] of 10 repeats based on STRUCTURE calculation using single nucleotide polymorphism (SNP) and haplotype data, respectively. (b) The Δ*k* of 10 repeats based on STRUCTURE calculation.

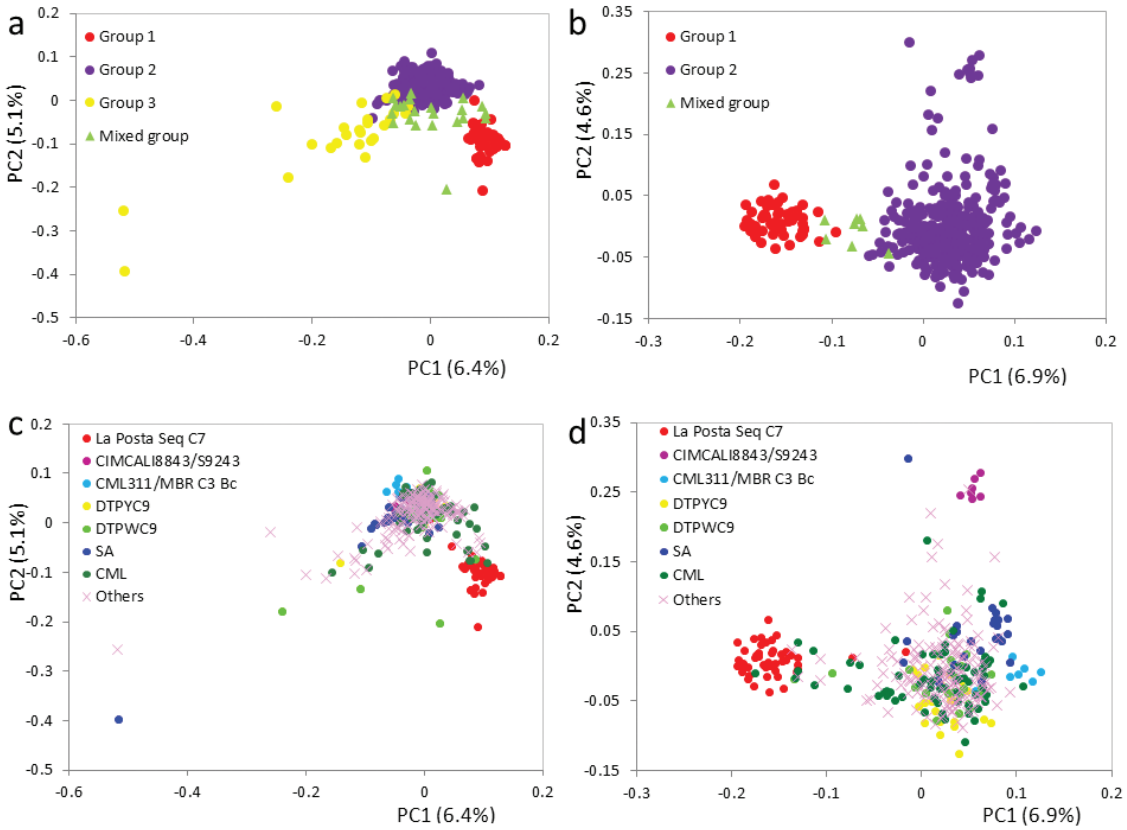

Figure 3. Principal component analysis (PCA) of 359 lines. (a) Principal component analysis based on single nucleotide polymorphism (SNP) data, groups 1, 2, and 3, and the mixed group were classified by the model-based method based on SNP data. (b) Principal component analysis based on haplotype data, groups 1 and 2, and the mixed group were classified by the model-based method based on haplotype data. (c) Principal component analysis based on SNP data; lines from different groups were colored differently. (d) Principal component analysis based on haplotype data; lines from different groups were colored differently. PC, principal component; CIMCALI, CIMMYT lines from Cali, Columbia; CML, CIMMYT Maize Line; DTPW, drought tolerant population white grain; DTPY, drought tolerant population yellow grain; MBR, multiple borer resistance; SA, soil acidity.

Table 3. Genetic diversity and mean of relative kinship within each subset identified based on pedigree and line origin information.

Subset	Pedigree or resource†	Number	Heterotic type	Number of Polymorphic loci	Gene diversity	PIC‡	Average polymorphism ratio	Mean of relative kinship
1	CIMCALI8843 × S9243	7	B	118	0.033	0.0279	0.036	0.838
2	CML311 × MBR C3 Bc	7	A, B, A and/or B	581	0.1791	0.1456	0.208	0.3235
3	DTPWC9	24	A	1083	0.2956	0.2525	0.309	0.081
4	DTPYC9	23	A and/or B	1021	0.2682	0.2259	0.282	0.1229
5	La Posta Seq C7	47	A	888	0.2307	0.1935	0.233	0.1531
6	SA	25	A, B	1064	0.2684	0.229	0.285	0.1299
7	Mexico subtropical	31	A, B, A and/or B	1141	0.2985	0.25	0.305	0.0255
8	Mexico tropical	41	A, B, A and/or B	1145	0.3044	0.2592	0.309	0.0344
9	CML	69	A, B, A and/or B	1197	0.3008	0.254	0.312	0.036

†CIMCALI, CIMMYT lines from Cali, Columbia; CML, CIMMYT Maize Line; DTPW, drought tolerant population white grain; DTPY, drought tolerant population yellow grain; MBR, multiple borer resistant; SA, soil acidity.

‡PIC, polymorphic information content.

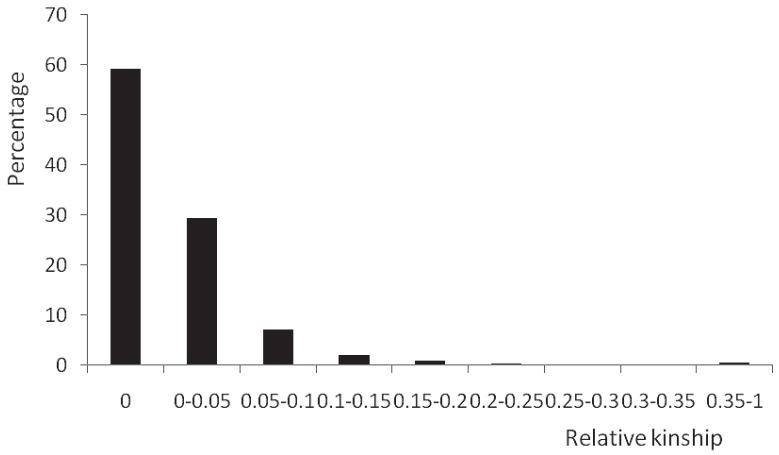

Figure 4. Distribution of pairwise relative kinship values.

greatest number of polymorphic loci. Genetic diversity of the nine subsets tended to increase as the mean of relative kinship within subsets decreased.

For any two given lines within this study, the polymorphic ratio ranged from 0.005 to 0.603 with an average of 0.157. Among the nine subsets, subset 9 had the largest average polymorphism ratio while subset 1 had the lowest, consistent with the number of polymorphic loci (Table 3). Larger genetic divergence was observed from pairwise *F*-statistics values between subset 1, 2, 5, and other subsets (Table 5). The *F*-statistics value between subset 1 and 2 was the largest while it was the smallest between subset 7 and 9 (Table 5).

Linkage Disequilibrium

Linkage disequilibrium, measured as r^2, at the whole genome level was evaluated using different sets of SNP markers (Fig. 5; Supplemental Table S7); r^2 decayed rapidly with increasing physical distance but it did not decrease with the increase of physical distance between 0.2 and 2 kbp (Fig. 5; Supplemental Table S7).The mean r^2 decreased with the reduction of MAF of SNP markers, especially within short distances (i.e., 0–100 kbp). It was less than 0.1 up to a maximum physical distance of 10 kbp estimated based on all 1260 SNPs and when estimated by SNPs with MAF greater than 0.05 and 0.1 it decreased to less than 0.1 between 10 and 100 kbp.

Table 4. Analysis of molecular variation of three populations classified by different methods.

Source of variation	df	Sum of squares	Variance components	Percent variation
Populations classified using model-based method based on biallelic single nucleotide polymorphism data				
Among populations	3	2030.2	4.9	6.2
Within populations	714	52109.8	73.4	93.8
Total	717	54140	78.3	100
Populations classified using model-based method based on multiallelic haplotype data				
Among populations	2	2130.9	8.6	10.6
Within populations	715	52009.1	73	89.4
Total	717	54140	81.6	100
Populations (the nine subsets) identified based on pedigree and line origin information				
Among populations	8	6707.5	11.5	11.75
Within populations	539	45683.9	86.2	88.25
Total	547	52391.4	97.9	

Table 5. Pairwise *F*-statistics value between each pair of the nine subsets identified based on pedigree and origin information.

Subset	Subset 1	Subset 2	Subset 3	Subset 4	Subset 5	Subset 6	Subset 7	Subset 8	Subset 9
1	0								
2	0.698	0							
3	0.434	0.242	0						
4	0.478	0.268	0.093	0					
5	0.496	0.336	0.178	0.212	0				
6	0.463	0.266	0.132	0.168	0.209	0			
7	0.360	0.177	0.066	0.095	0.119	0.083	0		
8	0.395	0.168	0.076	0.110	0.154	0.092	0.013	0	
9	0.395	0.195	0.068	0.101	0.137	0.095	0.011	0.031	0

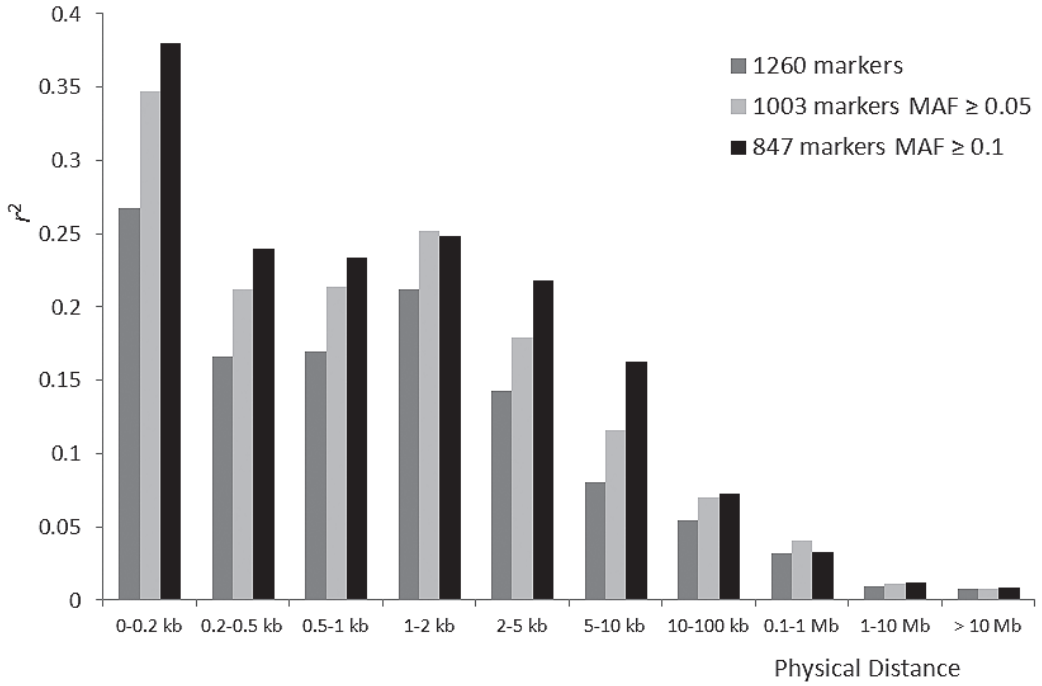

Figure 5. Mean linkage disequilibrium (LD) estimates for different physical distances measured with different marker types in 359 lines. Mean LD estimates are pooled over all chromosomes. MAF, minor allelic frequency.

Linkage disequilibrium decayed rapidly with increasing physical distance and the LD pattern varied across the chromosomes (Supplemental Fig. S3).

Linkage disequilibrium estimates generally decreased as the sample size increased in both subsets selected by pedigree (SP) and subsets selected randomly (SR) (Fig. 6). However, in SP, LD estimates for subset 5 (n = 47) were found to be larger than that of subsets with smaller sample size (i.e., subset 3, 4, 6, 7, and 8), particularly across marker intervals of 0 to 0.2, 0.2 to 0.5, 0.5 to 1, 1 to 2, and 10 to 100 kbp. Subset 5 has higher mean relative kinship and lower PIC and gene diversity than other subsets with smaller sample size (i.e., subset 3, 4, 6, 7, and 8) (Table 3). Generally, using the same sample size, LD was greater within SP than that within SR. However, for subset 2 and 5, LD was significantly higher in SP than in SR (SP estimates were 1.04 to 2.49 times higher than the mean value of 10 SR) across the marker interval of 0 to 10 Mbp.

Model Comparison and Association Panel Evaluation

Association analysis was performed using 999 SNPs with MAF ≥ 0.05 and GY, ASI, and EPP under drought stress, with Qh (i.e., Q value calculated based on haplotype data), Qs (i.e., Q value calculated based on SNP data), K, Qh + K, and Qs + K models for each trait. Quantile–quantile plots of estimated $\log_{10} p$ value indicated that the K and $Q + K$ model controlled the false positive with minimal deviation from expected values (Fig. 7). Model testing using phenotypic and genetic data classified this maize panel as a type II association sample and suitable for association analysis (Zhu and Yu, 2009).

DISCUSSION

Diverse Maize Collection Encompassing Gains in Stress Tolerance Improvement

Molecular analyses of CIMMYT maize germplasm have been performed previously (Reif et al., 2003a, b; Xia et al.,

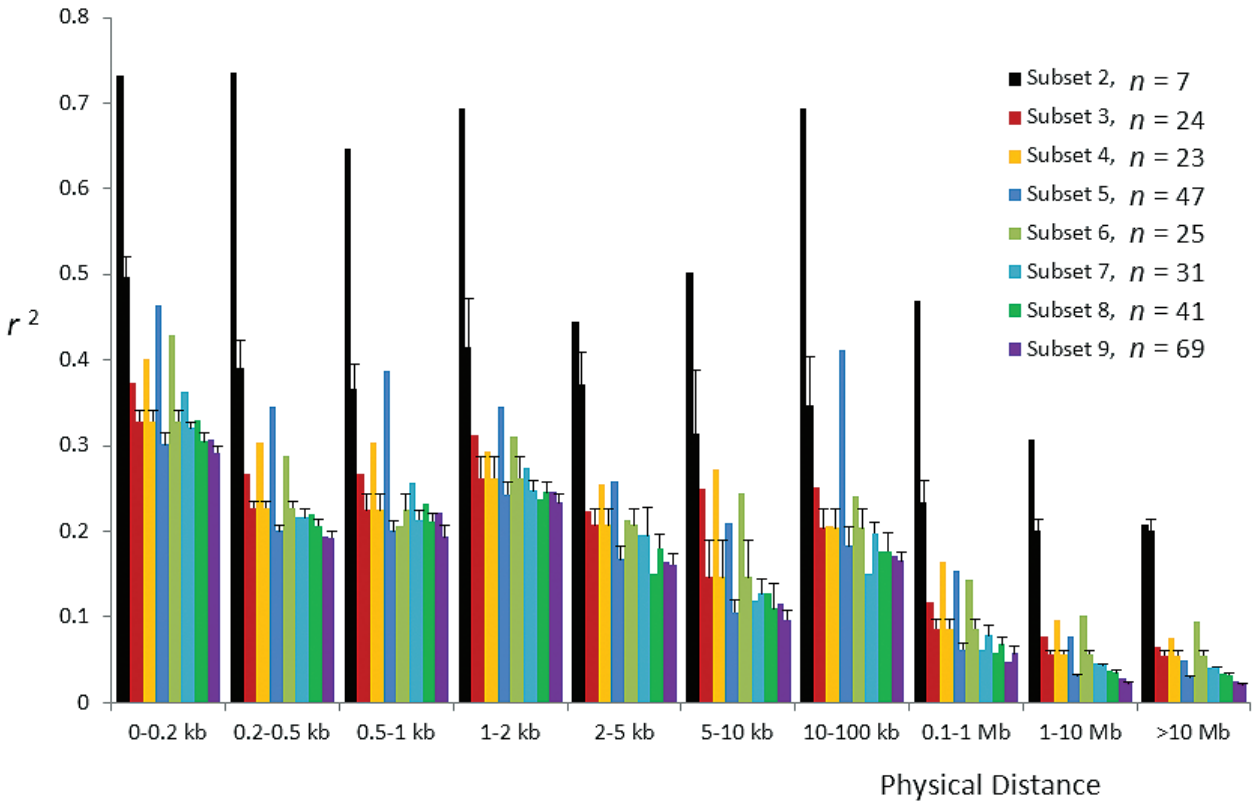

Figure 6. Mean linkage disequilibrium (LD) estimates for different physical distances pooled over all chromosomes of eight subsets selected by pedigree (SP) and subsets selected randomly (SR) with the same sample size as them. Columns with same color have the same sample size. Within the columns with the same colors, LD of the ones on the right is calculated based on SR.

2004, 2005; Warburton et al., 2002, 2005). The results of these studies highlighted the difficulties of assigning lines into genetically diverse heterotic groups due to the mixed genetic constitution of the original germplasm and may, in part, also be related to insufficient documentation of germplasm development. These reflect CIMMYT breeding history and care should be taken to properly document information in future pedigree breeding programs. The greater variation within populations relative to among populations (as revealed by AMOVA) in present results could be explained by the origin and genetic background of these CIMMYT populations. Development of heterotic groups at CIMMYT was a rather recent event that started in the mid 1980s (Vasal et al., 1999; Rief et al., 2003a, b). At CIMMYT, heterotic groups are assigned after field evaluations using testers from different heterotic groups. Two heterotic groups, A and B, have been tentatively formed, which consisted of dent and flint germplasm complexes, respectively (Vasal et al., 1999; Ortiz et al., 2010). In this study, we were unable to separate heterotic groups A and B based on partition results using either SNP or SNP haplotype data. Subgroups based on clustering analysis and PCA using marker data generally agree with pedigree information of the lines (Fig. 3; Supplemental Fig S1). This suggests variation exists within current heterotic groups. Combining current information of heterotic group and subgroups classification inferred from molecular data may be the best strategy to define heterotic groups (patterns) for future hybrid breeding practice.

Linkage Disequilibrium Pattern in Current Collection

Linkage disequilibrium is an important factor affecting the precision of association analysis and genomewide selection. The large genetic and phenotypic diversity of maize results in a rather fast decay of LD, often within several kilobase pairs (Remington et al., 2001; Tenaillon et al., 2001; Ching et al., 2002; Jung et al., 2004; Tian et al., 2009; Gore et al., 2009). Recently, using a large and diverse maize collection of 632 lines and genotyping it with 1229 SNPs, Yan et al. (2009) demonstrated that LD declines (average value of r^2 declines to 0.1) within 2 to 5 kbp; however, this distance is variable across chromosomes and not continual within a chromosome. In the present study, extent of LD varied across the genome but LD decay ($r^2 < 0.1$) was more than 10 kbp on average when using SNPs with MAF ≥ 0.05 (Fig. 5; Supplemental Table S7), which was greater than the estimation reported by Yan et al. (2009). This may be due to the level of diversity because higher average polymorphic ratio among lines was observed in the study of Yan et al. (2009) than in the present study. In this study, average r^2 between all 1260 SNP pairs fell below 0.1 in the range of 5 to 10 kbp (Fig. 5; Supplemental Table S7). In addition, our results found MAF of SNP markers and

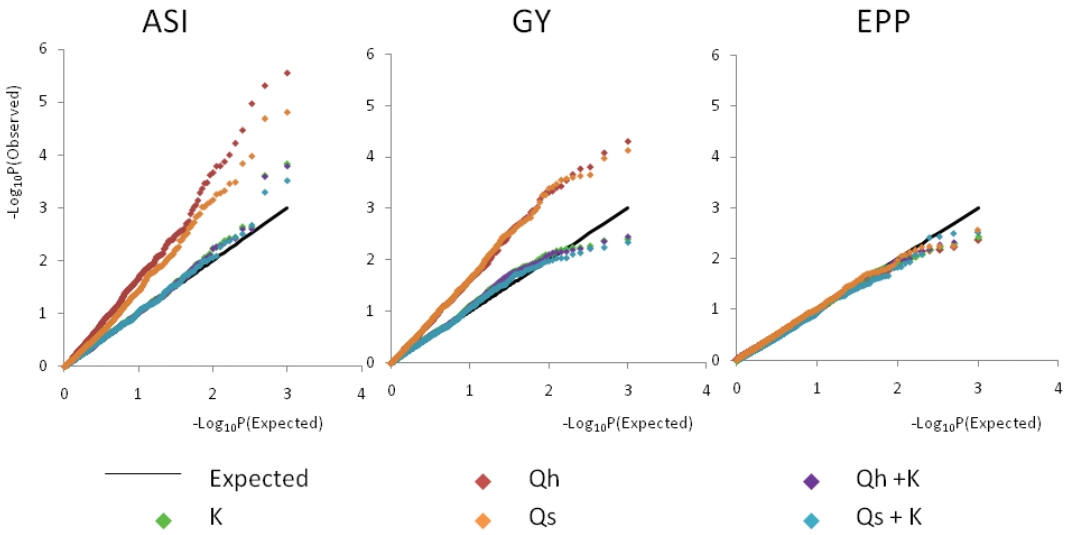

Figure 7. Quantile–quantile plots of $\log_{10} p$ value with 999 single nucleotide polymorphisms (SNPs) for three quantitative traits of anthesis to silking interval (ASI), grain yield (GY) and ear per plant (EPP) showing the adequate control of a type I error with the *K* and *Q* + *K* model. Qh, *Q* value calculated based on haplotype data; Qs, *Q* value calculated based on single nucleotide polymorphism data.

sample size of the population affected the estimation of LD in agreement with a previous study (Yan et al., 2009). Within all nine subsets classified mainly based on pedigree information, LD decay was slower (Fig. 6). Significant differences in mean r^2 were observed between the SP and SR, which have the same sample size with each other, particularly when sample size and diversity of SP were small (Fig. 6). The results suggested that diversity and kinship relation of lines within the population may also be factors affecting the LD estimation. Five genomic regions containing three contiguous SNPs with large r^2 between each other were observed in chromosomes 2, 4, 8, and 9 (Supplemental Table S8). The largest physical distance between SNP markers in the same region ranged from 19.74 to 129.12 kbp with the corresponding r^2 at these distances between 0.517 and 0.958. More than 100 LD blocks have been reported in the maize genome (Gore et al., 2009) and it is possible that LD blocks exist within these regions with large LD detected in the present study.

Strategy for Stress Tolerance Improvement based on this Maize Panel

In the present study, association mapping was performed using this maize panel, which is a powerful approach for dissection of complex agronomic traits and widely used in plants now (Atwell et al., 2010; Yan et al., 2011). After identifying genes significantly associated with target traits, the most favorable alleles can be found by resequencing in a diverse panel of germplasm, which is used for trait improvement through MAS (Harjes et al., 2008; Yan et al., 2010b). Construction and characterization of a desirable panel of maize underlies the successful application of association mapping. Complex patterns of genetic relatedness among individuals can cause spurious associations, especially for mapping traits that have been subjected to local adaptation (Zhu et al., 2008; Myles et al., 2009). In our study, spurious associations were well controlled after correcting for genetic relatedness. The *K* and *Q* + *K* models performed better than the model only involving **Q** matrix, which is consistent with previous studies (Yu et al., 2006; Zhao et al., 2007; Zhu and Yu, 2009). The data generated in this study can be combined with multilocation phenotypic data to identify the genomic regions associated with stress tolerance.

No specific genes underlying major quantitative trait loci controlling the important agronomic and economic traits have been identified in maize (Buckler et al., 2009). Tolerance to key abiotic stresses, notably drought and salinity, are complex and highly variable (Tester and Langridge, 2010). Genome-wide selection (GWS) provides an alternative strategy for maize stress-resistance improvement, simultaneously estimating loci, haplotype, and marker effects across the entire genome to calculate genomic estimated breeding values (Meuwissen et al., 2001). Genome-wide selection estimates the effect of each marker without testing it for its significance of association with the target trait. Simulation studies using GWS confirm potential breeding gains based on this strategy, and it may dramatically change the role of phenotyping especially with the rapid development of high throughput genotyping platform (Mayor and Bernardo, 2009; Bernardo, 2009; Jannink et al., 2010). In case of present study, GWS can be used in our breeding program for developing enhanced maize lines in terms of stress tolerance. Nine subsets were classified according to pedigree information, environmental adaption, and breeding scheme. Information of heterotic pattern within these nine subsets was available, which is useful for inbred line development. For example, lines from subset 5 (LPS C7) were all classified into heterotic group A (Table 3), and improved new inbred lines can be produced from the crosses among these lines. Genome-wide selection can be

utilized to select superior lines during the process combining with the phenotypic selection.

Supplemental Information Available

Supplemental material is available free of charge at http://www.crops.org/publications/cs.

Supplemental Table S1. Pedigree, heterotic group, grain color information of 359 lines.

Supplemental Table S2. Information of the main germplasm sources of the maize panel.

Supplemental Table S3. Anthesis-silking interval (ASI), grain yield (GY), and ears per plant (EPP) of a set of 253 maize lines under two water regimes, well-watered and anthesis-stage drought stress in Tlaltizapán, Mexico, in 2007.

Supplemental Table S4. Information of 1260 used single nucleotide polymorphisms (SNPs) and 276 eliminated SNPs in this study.

Supplemental Table S5. The distribution of single nucleotide polymorphisms (SNPs) for different interval of missing data percentage.

Supplemental Table S6. Information of 26 lines with heterozygosity larger than 0.1.

Supplemental Table S7. Mean linkage disequilibrium (LD) among 1260 single nucleotide polymorphisms (SNPs), 1003 SNPs (minor allelic frequency [MAF] ≥ 0.05), and 847 SNPs (MAF ≥ 0.1) over different physical distances and across 10 chromosomes.

Supplemental Table S8. Genomic regions with high level of linkage disequilibrium (LD) exist potential LD block.

Supplemental Figure S1. a). The origin of the initial panel of 850 lines. b). The origin of the final panel of 359 lines.

Supplemental Figure S2. Neighbor-joining (NJ) tree of 359 lines. Different colors represent different origins of lines.

Supplemental Figure 3. Decay of linkage disequilibrium (LD) (r^2) as a function of physical distance (Mb) between pairs of loci on individual chromosomes. All 1,260 single nucleotide polymorphisms (SNPs) were used.

Acknowledgments

This study was supported in part by the Drought-Tolerant Maize for Africa project (DTMA), funded by the Bill and Melinda Gates Foundation. We are grateful to Pedro Chepetla's team (CIMMYT Tlaltizapán Experimental Station) for their helpful contribution in field experiments.

References

Atwell, S., Y.S. Huang, B.J. Vilhjálmsson, G. Willems, M. Horton, Y. Li, et al. 2010. Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. Nature 465:627–631. doi:10.1038/nature08800

Arizona Genomics Institute. 2009. Maize genome. Available at http://www2.genome.arizona.edu/genomes/maize (verified 7 Sept. 2011). University of Arizona. Tucson, AZ.

Bänziger, M., G.O. Edmeades, D. Beck, and M. Bellon. 2000. Breeding for drought and nitrogen stress tolerance in maize: From theory to practice. CIMMYT, Mexico, D.F., Mexico.

Bänziger, M., G.O. Edmeades, and H.R. Lafitte. 1999. Selection for drought tolerance increases maize yields across a range of nitrogen levels. Crop Sci. 39:1035–1040. doi:10.2135/cropsci1999.0011183X003900040012x

Bänziger, M., G.O. Edmeades, and H.R. Lafitte. 2002. Physiological mechanisms contributing to the increased N stress tolerance of tropical maize selected for drought tolerance. Field Crops Res. 75:223–233. doi:10.1016/S0378-4290(02)00028-X

Bernardo, R. 2009. Genomewide selection for rapid introgression of exotic germplasm in maize. Crop Sci. 49:419–425. doi:10.2135/cropsci2008.08.0452

Bradbury, P.J., Z.W. Zhang, D.E. Kroon, R.M. Casstevens, Y. Ramdoss, and E.S. Buckler. 2007. TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633–2635. doi:10.1093/bioinformatics/btm308

Bruce, W.B., G.O. Edmeades, and T.C. Barker. 2002. Molecular and physiological approaches to maize improvement for drought tolerance. J. Exp. Bot. 53:13–25. doi:10.1093/jexbot/53.366.13

Buckler, E.S., J.B. Holland, P.J. Bradbury, C.B. Acharya, P.J. Brown, C. Browne, E. Ersoz, S. Flint-Garcia, A. Garcia, J.C. Glaubitz, M.M. Goodman, C. Harjes, K. Guill, D.E. Kroon, S. Larsson, N.K. Lepak, H. Li, S.E. Mitchell, G. Pressoir, J.A. Peiffer, M.O. Rosas, T.R. Rocheford, M. Cinta Romay, S. Romero, S. Salvo, H.S. Villeda, H. Sofia da Silva, Q. Sun, F. Tian, N. Upadyayula, N. Ware, H. Yates, J. Yu, Z. Zhang, S. Kresovich, and M.D. McMullen. 2009. The genetic architecture of maize flowering time. Science 325:714–718. doi:10.1126/science.1174276

Ching, A., K.S. Caldwell, M. Jung, M. Dolan, O.S. Smith, et al. 2002. SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. BMC Genet. 3:19. doi:10.1186/1471-2156-3-19

Darroch, J.N., and J.E. Mosimann. 1985. Canonical and principal components of shape. Biometrika 72:241–252. doi:10.1093/biomet/72.2.241

Edmeades, G.O., M. Bänziger, D. Beck, J. Bolaños, and A. Ortega. 1997. Development and *per se* performance of CIMMYT maize populations as drought-tolerant sources. p. 254–262. *In* G.O. Edmeades et al. (ed.) Developing drought and low-N tolerant maize. CIMMYT, Mexico D.F., Mexico.

Edmeades, G.O., J. Bolaños, S.C. Chapman, H.R. Lafitte, and M. Bänziger. 1999. Selection improves drought tolerance in tropical maize populations: I. Gains in biomass, grain yield, and harvest index. Crop Sci. 39:1306–1315. doi:10.2135/cropsci1999.3951306x

Evanno, G., S. Regnaut, and J. Goudet. 2005. Detecting the number of clusters of individuals using the software structure: A simulation study. Mol. Ecol. 14:2611–2620. doi:10.1111/j.1365-294X.2005.02553.x

Excoffier, L., G. Laval, and S. Schneider. 2005. Arlequin ver. 3.0: An integrated software package for population genetics data analysis. Evol. Bioinform. Online 1:47–50.

Excoffier, L., P.E. Smouse, and J.M. Quattro. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. Genetics 131:479–491.

Falush, D., M. Stephens, and J.K. Pritchard. 2003. Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. Genetics 164:1567–1587.

Fan, J.B., K.L. Gunderson, M. Bibikova, J.M. Yeakley, J. Chen, E. Wickham Garcia, L.L. Lebruska, M. Laurent, R. Shen,

and D. Barker. 2006. Illumina universal bead arrays. Methods Enzymol. 410:57–73. doi:10.1016/S0076-6879(06)10003-8

Gore, M.A., J.M. Chia, R.J. Elshire, Q. Sun, E.S. Ersoz, B.L. Hurwitz, J.A. Peiffer, M.D. McMullen, G.S. Grills, J. Ross-Ibarra, D.H. Ware, and E.S. Buckler. 2009. A first-generation haplotype map of maize. Science 326:1115–1117. doi:10.1126/science.1177837

Hardy, O.J., and X. Vekemans. 2002. Spagedi: A versatile computer program to analyse spatial genetic structure at the individual or population levels. Mol. Ecol. Notes 2:618–620. doi:10.1046/j.1471-8286.2002.00305.x

Harjes, C.E., T.R. Rocheford, L. Bai, T.P. Brutnell, C.B. Kandianis, S.G. Sowinski, A.E. Stapleton, R. Vallabhaneni, M. Williams, E.T. Wurtzel, J. Yan, and E.S. Buckler. 2008. Natural genetic variation in lycopene epsilon cyclase tapped for maize biofortification. Science 319:330–333. doi:10.1126/science.1150255

Hyten, D.L., Q. Song, I.Y. Choi, M.S. Yoon, J.E. Specht, L.K. Matukumalli, R.L. Nelson, R.C. Shoemaker, N.D. Young, and P.B. Cregan. 2008. High-throughput genotyping with the GoldenGate assay in the complex genome of soybean. Theor. Appl. Genet. 116:945–952. doi:10.1007/s00122-008-0726-2

Jannink, J.L., A.J. Lorenz, and H. Iwata. 2010. Genomic selection in plant breeding: From theory to practice. Brief. Funct. Genom. 9:166–177. doi:10.1093/bfgp/elq001

Jung, M., A. Ching, D. Bhattramakki, M. Dolan, S. Tingey, et al. 2004. Linkage disequilibrium and sequence diversity in a 500-kbp region around the adh1 locus in elite maize germplasm. Theor. Appl. Genet. 109:681–689. doi:10.1007/s00122-004-1695-8

Liu, K.J., and S.V. Muse. 2005. PowerMarker: An integrated analysis environment for genetic marker analysis. Bioinformatics 21:2128–2129. doi:10.1093/bioinformatics/bti282

Loiselle, B.A., V.L. Sork, J. Nason, and C. Graham. 1995. Spatial genetic structure of a tropical understory shrub, *Psychotria officinalis* (Rubiaceae). Am. J. Bot. 82:1420–1425. doi:10.2307/2445869

Mayor, P.J., and R. Bernardo. 2009. Genomewide selection and marker-assisted recurrent selection in doubled haploid versus F_2 populations. Crop Sci. 49:1719–1725. doi:10.2135/cropsci2008.10.0587

Meuwissen, T.H.E., B.J. Hayes, and M.E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829.

Monneveux, P., C. Sánchez, D. Beck, and G.O. Edmeades. 2006. Drought tolerance improvement in tropical maize source populations: Evidence of progress. Crop Sci. 46:180–191. doi:10.2135/cropsci2005.04-0034

Muchero, W., N.N. Diop, P.R. Bhat, R.D. Fenton, S. Wanamaker, M. Pottorff, S. Hearne, N. Cisse, C. Fatokun, J.D. Ehlers, P.A. Roberts, and T.J. Close. 2009. A consensus genetic map of cowpea [*Vigna unguiculata* (L) Walp.] and synteny based on EST-derived SNPs. Proc. Natl. Acad. Sci. USA 106:18159–18164. doi:10.1073/pnas.0905886106

Mugo, S., M. Bänziger, J. Araus, G. Atlin, A. Diallo, F. Kanampiu, A. Langyintuo, J. MacRobert, C. Magorokosho, G. Mahuku, D. Makumbi, L. Narro, G. Ortiz-Ferrera, and P. Zaidi. 2008. Stress tolerant maize. *In* Rodomiro Ortiz (ed.) CIMMYT 2008 science week. Program and abstracts, El Batan, Texcoco, Mexico. 3–8 Mar. 2008. Available at http://apps.cimmyt.org/english/docs/research_highlights/sciWk_bAbs.pdf (verified 7 Sept. 2011) International Maize and Wheat improvement Center (CIMMYT), Mexico City, DF, Mexico.

Myles, S., J. Peiffer, P.J. Brown, E.S. Ersoz, Z.W. Zhang, D.E. Costich, and E.S. Buckler. 2009. Association mapping: Critical considerations shift from genotyping to experimental design. Plant Cell 21:2194–2202. doi:10.1105/tpc.109.068437

Narro, L., S. Pandey, C. De León, J.C. Perez, F. Salazar, and M.P. Arias. 1997. Heterosis in acid soil-tolerant maize germplasm. p. 290–291. *In* The genetics and exploitation of heterosis in crops. An international symposium; Mexico City (Mexico). CIMMYT, Mexico, DF, Mexico.

Nei, M. 1972. Genetic distance between populations. Am. Nat. 106:283–292. doi:10.1086/282771

Ortiz, R., S. Taba, V.H.C. Tovar, M. Mezzalama, Y. Xu, J. Yan, and J.H. Crouch. 2010. Conserving and enhancing maize genetic resources as global public goods – A perspective from CIMMYT. Crop Sci. 50:1–16. doi:10.2135/cropsci2009.02.0086

Pandey, S., H. Ceballos, G. Granados, and R. Knapp. 1994. Developing maize that tolerates Aluminum toxic soils. p. 85–92. *In* Edmeades et al. (ed.) Stress tolerance breeding: Maize that resists insects, drought, low-N, and acid soils. CIMMYT, Mexico City, D.F., Mexico.

Pandey, S., C. De León, and L. Narro. 1997. The South American regional maize program. p. 61–63. *In* Maize research in 1995–96. Series: CIMMYT maize program special report. CIMMYT, Mexico City, DF, Mexico.

Pandey, S., and C.O. Gardner. 1992. Recurrent selection for population, variety, and hybrid improvement in tropical maize. Adv. Agron. 48:1–87. doi:10.1016/S0065-2113(08)60935-9

Panzea. 2009. Genetic architecture of maize and teosinte. Available at http://www.panzea.org (verified 7 Sept. 2011). Cornell University. Ithaca, NY.

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

Reif, J.C., A.E. Melchinger, X.C. Xia, M.L. Warburton, D.A. Hoisington, S.K. Vasal, G. Srinivasan, M. Bohn, and M. Frisch. 2003a. Genetic distance based on simple sequence repeats and heterosis in tropical maize populations. Crop Sci. 43:1275–1282. doi:10.2135/cropsci2003.1275

Reif, J.C., A.E. Melchinger, X.C. Xia, M.L. Warburton, D.A. Hoisington, S.K. Vasal, G. Srinivasan, M. Bohn, and M. Frisch. 2003b. Use of SSRs for establishing heterotic groups in subtropical maize. Theor. Appl. Genet. 107:947–957. doi:10.1007/s00122-003-1333-x

Remington, D.L., J.M. Thornsberry, Y. Matsuoka, L.M. Wilson, S.R. Whitt, J. Doebley, S. Kresovich, M.M. Goodman, and E.S. Buckler. 2001. Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc. Natl. Acad. Sci. USA 98:11479–11484. doi:10.1073/pnas.201394398

Rostoks, N., L. Ramsay, K. MacKenzie, L. Cardle, P.R. Bhat, M.L. Roose, J.T. Svensson, N. Stein, P.K. Varshney, D.F. Marshall, A. Graner, T.J. Close, and R. Waugh. 2006. Recent history of artificial outcrossing facilitates whole-genome association mapping in elite inbred crop varieties. Proc. Natl. Acad. Sci. USA 103:18656–18661. doi:10.1073/pnas.0606133103

Schnable, P.S., D. Ware, R.S. Fulton, J.C. Stein, F. Wei, S. Pasternak, C. Liang, J. Zhang, L. Fulton, T.A. Graves, et al. 2009. The B73 maize genome: Complexity, diversity, and dynamics. Science 326:1112–1115. doi:10.1126/science.1178534

Setimela, P., Z. Chitalu, J. Jonazi, A. Mambo, D. Hodson, and M. Bänziger. 2005. Environmental classification of maize-testing sites in the SADC region and its implication for collaborative maize breeding strategies in the subcontinent. Euphytica 145:123–132. doi:10.1007/s10681-005-0625-4

Springer, N.M., et al. 2009. Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. PLoS Genet. 5(11):E1000734. doi:10.1371/journal.pgen.1000734

Tamura, K., J. Dudley, M. Nei, and S. Kumar. 2007. MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. Mol. Biol. Evol. 24:1596–1599. doi:10.1093/molbev/msm092

Tenaillon, M.I., M.C. Sawkins, A.D. Long, R.L. Gaut, J.F. Doebley, et al. 2001. Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). Proc. Natl. Acad. Sci. USA 98:9161–9166. doi:10.1073/pnas.151244298

Tester, M., and P. Langridge. 2010. Breeding technologies to increase crop production in a changing world. Science 327:818–822. doi:10.1126/science.1183700

Tian, F., N.M. Stevens, and E.S. Buckler. 2009. Tracking footprints of maize domestication and evidence for a massive selective sweep on chromosome 10. Proc. Natl. Acad. Sci. USA 106(Suppl. 1):9979–9986. doi:10.1073/pnas.0901122106

Vasal, S.K., H. Cordova, S. Pandey, and G. Srinivasan. 1999. Tropical maize and heterosis. p. 363–373. *In* J.G. Coors and S. Pandey (ed.) The genetics and exploitation of heterosis in crops. ASA, CSSA, SSSA, Madison, WI.

Warburton, M.L., J.M. Ribaut, J. Franco, J. Crossa, P. Dubreuil, and F.J. Betrán. 2005. Genetic characterization of 218 elite CIMMYT inbred maize lines using RFLP markers. Euphytica 142:97–106. doi:10.1007/s10681-005-0817-y

Warburton, M.L., X.C. Xia, J. Crossa, J. Franco, A.E. Melchinger, M. Frisch, M. Bohn, and D.A. Hoisington. 2002. Genetic characterization of CIMMYT maize inbred lines and open pollinated populations using large scale fingerprinting methods. Crop Sci. 42:1832–1840. doi:10.2135/cropsci2002.1832

Wen, W.W., S. Taba, T. Shah, V.H.C. Tovar, and J. Yan. 2011. Detection of genetic integrity of conserved maize (*Zea mays* L.) germplasm in genebanks using SNP markers. Genet. Resour. Crop Evol. 58:189–207. doi:10.1007/s10722-010-9562-8

Xia, X.C., J.C. Reif, D.A. Hoisington, A.E. Melchinger, M. Frisch, and M.L. Warburton. 2004. Genetic diversity among CIMMYT maize inbred lines investigated with SSR markers: I. Lowland tropical maize. Crop Sci. 44:2230–2237. doi:10.2135/cropsci2004.2230

Xia, X.C., J.C. Reif, A.E. Melchinger, M. Frisch, D.A. Hoisington, D. Beck, K. Pixley, and M.L. Warburton. 2005. Genetic diversity among CIMMYT maize inbred lines investigated with SSR markers: II. Subtropical, tropical midaltitude, and highland maize inbred lines and their relationships with elite U.S. and European maize. Crop Sci. 45:2573–2582. doi:10.2135/cropsci2005.0246

Yan, J., T. Shah, M. Warburton, E.S. Buckler, M.D. McMullen, and J.H. Crouch. 2009. Genetic characterization and linkage disequilibrium estimation of a global maize collection using SNP markers. PLoS ONE 4(12):E8451. doi:10.1371/journal.pone.0008451

Yan, J., X. Yang, T. Shah, H. Héctor Sánchez, J. Li, M. Warburton, Y. Zhou, J.H. Crouch, and Y. Xu. 2010a. High-throughput SNP genotyping with the GoldenGate assay in maize. Mol. Breed. 25:441–451. doi:10.1007/s11032-009-9343-2

Yan, J.B., C.B. Kandianis, C.E. Harjes, L. Bai, E.H. Kim, X.H. Yang, D. Skinner, Z.Y. Fu, S. Mitchell, Q. Li, M.G.S. Fernandez, M. Zaharieva, R. Babu, Y. Fu, N. Palacios, J.S. Li, D. DellaPenna, T.P. Brutnell, E.S. Buckler, M.L. Warburton, and T. Rocheford. 2010b. Rare genetic variation at Zea mays crtRB1 increases β-carotene in maize grain. Nat. Genet. 42:322–327. doi:10.1038/ng.551

Yan, J.B., M. Warburton, and J. Crouch. 2011. Association mapping for enhancing maize genetic improvement. Crop Sci. 51:433–449. doi:10.2135/cropsci2010.04.0233

Yu, J.M., G. Pressoir, W.H. Briggs, I.V. Bi, M. Yamasaki, J.F. Doebley, M.D. McMullen, B.S. Gaut, J.B. Holland, S. Kresovich, and E.S. Buckler. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. 38:203–208. doi:10.1038/ng1702

Zhao, K., M.J. Aranzana, S. Kim, C. Lister, C. Shindo, C. Tang, C. Toomajian, H. Zheng, C. Dean, P. Marjoram, and M. Nordborg. 2007. An arabidopsis example of association mapping in structured samples. PLoS Genet. 3:e4 10.1371/journal.pgen.0030004. doi:10.1371/journal.pgen.0030004

Zhu, C.S., M. Gore, E.S. Buckler, and J.M. Yu. 2008. Status and prospects of association mapping in plants. Plant Gen. 1:5–20. doi:10.3835/plantgenome2008.02.0089

Zhu, C.S., and J.M. Yu. 2009. Nonmetric multidimensional scaling corrects for population structure in association mapping with different sample types. Genetics 182:875–888. doi:10.1534/genetics.108.098863