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.

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

*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.
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, Colombia (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
flavus), common rust (Puccinia sorghi), corn stunt complex, Poly-
sora 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 Zimba-
bwe (n = 44), Nigeria (n = 5), and Ethiopia (n = 2), and inter-
national (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. Informa-
tion 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 sta-
tion in Tlaltizapán. Two seeds per hill were sown and then thinned
to 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, respec-
tively. 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.

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

<table>
<thead>
<tr>
<th>Group identification</th>
<th>Breeding program</th>
<th>Number of lines</th>
<th>Main sources†</th>
<th>Grain color</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Zimbabwe</td>
<td>44</td>
<td>CML, CIMCALI, and DTPW</td>
<td>44, 0</td>
</tr>
<tr>
<td>B</td>
<td>Nigeria</td>
<td>5</td>
<td>KU and P43</td>
<td>3, 2</td>
</tr>
<tr>
<td>C</td>
<td>Ethiopia</td>
<td>2</td>
<td>Pool9</td>
<td>2, 0</td>
</tr>
<tr>
<td>D</td>
<td>Colombia</td>
<td>27</td>
<td>SA3, SA4, SA5, SA6, SA7, and SA8</td>
<td>4, 23</td>
</tr>
<tr>
<td>E</td>
<td>Mexico highland</td>
<td>5</td>
<td>A.T.Z.T.R.L.BA90</td>
<td>1, 4</td>
</tr>
<tr>
<td>F</td>
<td>Mexico entomology</td>
<td>39</td>
<td>CML, MBR, ZMG07, KILIMA, and P84</td>
<td>28, 11</td>
</tr>
<tr>
<td>G</td>
<td>Mexico subtropical</td>
<td>31</td>
<td>CML, MBR, SPMAT, Pop 33, Pop 45, Pop 501, and Pop 502</td>
<td>18, 13</td>
</tr>
<tr>
<td>H</td>
<td>Mexico tropical</td>
<td>41</td>
<td>CML, CLQ, and CL</td>
<td>22, 19</td>
</tr>
<tr>
<td>I</td>
<td>Selection under drought</td>
<td>52</td>
<td>DTPW, DTPY, and LPS</td>
<td>41, 11</td>
</tr>
<tr>
<td>J</td>
<td>Selection under low N</td>
<td>32</td>
<td>DTPW, DTPY, and LPS</td>
<td>24, 8</td>
</tr>
<tr>
<td>K</td>
<td>Tester lines</td>
<td>13</td>
<td>CML</td>
<td>9, 4</td>
</tr>
<tr>
<td>L</td>
<td>Pathology</td>
<td>68</td>
<td>CML, CLQ, CL, DTPW, DTPY, and LPS</td>
<td>50, 18</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>359</td>
<td></td>
<td>246, 113</td>
</tr>
</tbody>
</table>

†A.T.Z.T.R.L., Amarillo Tardio Zona de Transition, Recombinacion de Lines; CIMCALI, CIMMYT lines from Cali, Colombia; 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 Sequia; 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.

**Single Nucleotide Polymorphism Genotyping and in silico Mapping**

Genotyping of the selected panel was conducted using an Illu-
mina 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 com-
munication, 2009) on the basis of having a desigability 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 Gold-
enGate assay. We used the Illumina BeadStation 500 G (Illumi-
na, Inc., San Diego, CA) for SNP genotyping according
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% miss-
ing 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 contain-
ing 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 contain-
ing 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.

Genotyping of the selected panel was conducted using an Illu-
mina 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 com-
munication, 2009) on the basis of having a desigability 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 Gold-
enGate assay. We used the Illumina BeadStation 500 G (Illumi-
na, Inc., San Diego, CA) for SNP genotyping according
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% miss-
ing 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 contain-
ing 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.

Genotyping of the selected panel was conducted using an Illu-
mina 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

**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

CROP SCIENCE, VOL. 51, NOVEMBER–DECEMBER 2011  WWW.CROPS.ORG
admixtured 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 \(\ln P(D)\)) before a large k value is examined, both \(\ln P(D)\) value and Evanno’s \(\Delta k\) (Evanno et al., 2005) were used to determine the most appropriate k value. Evanno’s \(\Delta k\) considers the rate of change of \(\ln P(D)\) as k increases and also the variance of \(\ln P(D)\) 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 \(\ln P(D)\), 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 \(\geq 0.60\) 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 phylogenetic 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 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.

**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 \( k = 3 \), \( \Delta k \) 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 \( \Delta k \) 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).

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,
Figure 2. (a) The mean and range of the natural logarithm probability of data $[\text{LnP(D)}]$ of 10 repeats based on STRUCTURE calculation using single nucleotide polymorphism (SNP) and haplotype data, respectively. (b) The $\Delta K$ 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.
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 3. Genetic diversity and mean of relative kinship within each subset identified based on pedigree and line origin information.**

<table>
<thead>
<tr>
<th>Subset</th>
<th>Pedigree or resource</th>
<th>Number</th>
<th>Heterotic type</th>
<th>Number of Polymorphic loci</th>
<th>Gene diversity</th>
<th>PIC$^\ddagger$</th>
<th>Average polymorphism ratio</th>
<th>Mean of relative kinship</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CIMCALI8843 × S9243</td>
<td>7</td>
<td>B</td>
<td>118</td>
<td>0.033</td>
<td>0.0279</td>
<td>0.036</td>
<td>0.838</td>
</tr>
<tr>
<td>2</td>
<td>CML311 × MBR C3 Bc</td>
<td>7</td>
<td>A, B, A and/or B</td>
<td>581</td>
<td>0.1791</td>
<td>0.1456</td>
<td>0.208</td>
<td>0.3235</td>
</tr>
<tr>
<td>3</td>
<td>DTPWC9</td>
<td>24</td>
<td>A</td>
<td>1083</td>
<td>0.2956</td>
<td>0.2525</td>
<td>0.309</td>
<td>0.081</td>
</tr>
<tr>
<td>4</td>
<td>DTPYC9</td>
<td>23</td>
<td>A and/or B</td>
<td>1021</td>
<td>0.2862</td>
<td>0.2259</td>
<td>0.282</td>
<td>0.1299</td>
</tr>
<tr>
<td>5</td>
<td>La Posta Seq C7</td>
<td>47</td>
<td>A</td>
<td>888</td>
<td>0.2307</td>
<td>0.1935</td>
<td>0.233</td>
<td>0.1531</td>
</tr>
<tr>
<td>6</td>
<td>SA</td>
<td>25</td>
<td>A, B</td>
<td>1064</td>
<td>0.2684</td>
<td>0.229</td>
<td>0.285</td>
<td>0.1299</td>
</tr>
<tr>
<td>7</td>
<td>Mexico subtropical</td>
<td>31</td>
<td>A, B, A and/or B</td>
<td>1141</td>
<td>0.2985</td>
<td>0.25</td>
<td>0.305</td>
<td>0.0255</td>
</tr>
<tr>
<td>8</td>
<td>Mexico tropical</td>
<td>41</td>
<td>A, B, A and/or B</td>
<td>1145</td>
<td>0.3044</td>
<td>0.2592</td>
<td>0.309</td>
<td>0.0344</td>
</tr>
<tr>
<td>9</td>
<td>CML</td>
<td>69</td>
<td>A, B, A and/or B</td>
<td>1197</td>
<td>0.3008</td>
<td>0.254</td>
<td>0.312</td>
<td>0.036</td>
</tr>
</tbody>
</table>

$^\ddagger$PIC, polymorphic information content.

---

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

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

---

![Figure 4. Distribution of pairwise relative kinship values.](image-url)
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.,...
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
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 S3. 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**


