DOI 10.1007/s11032-010-9500-7

Characterization of a global germplasm collection and its potential utilization for analysis of complex quantitative traits in maize

Xiaohong Yang · Shibin Gao · Shutu Xu · Zuxin Zhang · Boddupalli M. Prasanna · Lin Li · Jiansheng Li · Jianbing Yan

Received: 2 April 2010 / Accepted: 21 August 2010

Abstract Association mapping is a powerful approach for exploring the molecular basis of phenotypic variations in plants. A maize (*Zea mays* L.) association mapping panel including 527 inbred lines with tropical, subtropical and temperate backgrounds, representing the global maize diversity, was genotyped using 1,536 single nucleotide polymorphisms (SNPs). In total, 926 SNPs with minor allele frequencies of ≥0.1 were used to estimate the pattern of genetic diversity and relatedness among individuals. The analysis revealed broad phenotypic diversity and complex genetic relatedness in the maize panel. Two different Bayesian approaches identified three specific subpopulations, which were then reconfirmed by principal component analysis (PCA) and tree-based analyses. Marker–trait associations were performed to assess the suitability of different models for false-positive correction by population structure (Q matrix/PCA) and familial kinship (K matrix) alone or in combination in this panel. The K, Q + K and PCA + K models could reduce the false positives, and the Q + K model performed slightly better for flowering time, ear height and ear diameter. Our findings suggest that this maize panel is suitable for association mapping in order to understand the relationship between genotypic and phenotypic variations for agriculturally complex quantitative traits using optimal statistical methods.

Keywords Maize · Genetic diversity · Genetic relatedness · Association mapping · Phenotypic variation

Introduction

Maize (*Zea mays* L.) is one of the most important crops in the world, serving as a source of food, feed and fuel. To address global demands due mainly to

Xiaohong Yang and Shibin Gao contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s11032-010-9500-7) contains supplementary material, which is available to authorized users.

X. Yang · S. Xu · L. Li · J. Li · J. Yan (✉)
National Maize Improvement Center of China,
Beijing Key Laboratory of Crop Genetic Improvement,
China Agricultural University, 100193 Beijing, China
e-mail: yjianbing@gmail.com

S. Gao
Maize Research Institute, Sichuan Agricultural
University, 625014 Ya'an, Sichuan, China

Z. Zhang
National Key Laboratory of Crop Improvement,
Huazhong Agricultural University,
430070 Wuhan, Hubei, China

B. M. Prasanna · J. Yan
International Maize and Wheat Improvement Center
(CIMMYT), Apdo. Postal 6-641, 06600 Mexico,
Edo Mex, Mexico

Published online: 24 September 2010

continuing population growth and energy insufficiencies, improvement of maize productivity and quality through breeding is vital (Tester and Langridge 2010). The availability of the maize genome sequence (Schnable et al. 2009; Vielle-Calzada et al. 2009) and advanced high-throughput genotyping techniques (Gupta et al. 2008; Varshney et al. 2009) will provide new insights into complex quantitative traits for maize improvement. Linkage mapping is a powerful and popular approach for identifying the genes or loci which affect the natural phenotypic variations. Generally, the resolution provided by linkage mapping is low (10–30 cM) unless huge mapping populations are used (Salvi et al. 2007; Ducrocq et al. 2009), whereas association mapping provides a complementary approach with higher resolution due to advances in rapid and cost-effective genotyping technologies and the development of statistical methods (Yu and Buckler 2006; Zhu et al. 2008; Myles et al. 2009). Since first successfully applied in maize (Thornsberry et al. 2001), a series of association mapping studies has been performed to investigate the causal variants associated with flowering time (Camus-Kulandaivelu et al. 2006; Salvi et al. 2007; Ducrocq et al. 2009; Pressoir et al. 2009), kernel starch related traits (Wilson et al. 2004), maysin synthesis (Szalma et al. 2005), forage quality related traits (Andersen et al. 2007), carotenoid content (Harjes et al. 2008; Yan et al. 2010a), kernel oil related traits (Belo et al. 2008) and kernel size (Li et al. 2010a, b), and the details have been summarized in a recent review (Yan et al. 2010b). Maize is an ideal crop for association mapping due to its great genetic diversity and rapid linkage disequilibrium (LD) decay (Yan et al. 2010b).

Successful association mapping of a species requires firstly the creation of a desirable germplasm collection that reflects genetic diversity, extent of LD decay and genetic relatedness in a population, which determine the mapping resolution and power (Zhu et al. 2008). Generally, germplasm collections need to encompass adequate genetic diversity to cover most variations for the traits of interest. Maize exhibits extensive genetic variation, so much so that the average diversity at the nucleotide level between any two maize lines is higher than that between humans and chimps (Buckler and Stevens 2005). Furthermore, the LD decays rapidly in diverse maize genotypes, to the extent of 2–5 kb in the elite inbred lines (Yan et al. 2009). These observations imply that the association mapping panel involving advanced breeding lines could provide adequate diversity and resolution for quantitative trait loci (QTL) analysis in maize. In choosing lines to construct an association mapping panel, one should consider the balance between genetic diversity and germplasm adaptation. Since maize originated from the tropical center of Mexico and then dispersed to other temperate regions worldwide, adaptation should be an important factor that must be considered for precise phenotyping in a given environment. Thus, developing an association mapping panel from elite lines chosen from breeding programs will have two obvious advantages: (1) partial avoidance of the adaptation issue with good field performance; (2) the ability to conveniently introgress the identified genes or QTL into elite genotypes used in breeding programs. Recent studies demonstrated that complex quantitative traits in maize, such as flowering time (Buckler et al. 2009) and oil content (Laurie et al. 2004), are controlled by a large number of genes/QTL with minor effects. For statistical significance, a reasonable sample size may be required to obtain enough power to identify such genes/QTL with subtle effects. However, in the reported analyses of maize association, 100–300 genotypes were usually used, which might only be adequate to identify genes with strong effects (Thornsberry et al. 2001; Wilson et al. 2004; Szalma et al. 2005; Harjes et al. 2008; Pressoir et al. 2009; Yan et al. 2010a; Li et al. 2010a, b).

Due to domestication and selection by breeders, complex patterns of genetic relatedness are common in maize association panels, generating numerous spurious associations (Yu and Buckler 2006; Zhu et al. 2008; Myles et al. 2009), which makes correcting spurious associations a major challenge for association mapping. Random molecular markers throughout the genome are generally used to estimate the genetic relatedness among individuals. Genome control (GC) (Devlin and Roeder 1999; Zheng et al. 2005) and the structured association (SA) (Pritchard et al. 2000b) are two major methods first used to control false-positive associations. The SA method, often used by the program STRUCTURE to estimate population structure (Q matrix) (Pritchard et al. 2000a; Falush et al. 2003), was further refined by incorporating the relative kinship into the mixed-liner model (MLM), thereby efficiently reducing spurious associations when genetic relatedness among individuals is complex (Yu et al. 2006; Kang et al. 2008; Stich et al. 2008; Zhu and Yu 2009). Recently, it was suggested that principal component analysis (PCA) is a fast and effective way to infer population structure (Patterson et al. 2006; Price et al. 2006), and it showed some promise in replacing the Q matrix in the mixed model (Zhao et al. 2007; Zhu and Yu 2009). The development of statistical methods makes association mapping appealing for exploring the genetic architecture of quantitative traits, but additional research is required to improve the statistical methods for association mapping, especially for genome-wide association studies (GWAS).

We have assembled a global germplasm collection with more than 1,000 maize elite inbreds representing the major temperate and tropical/subtropical breeding programs of China, CIMMYT and Germplasm Enhancement of Maize (GEM). Some of the lines in the collection have been described in previous studies (Yang et al. 2010). All the lines were genotyped using the 1,536-SNP GoldenGate assay and phenotyped in Beijing and Sanya, China from 2005 to 2008 (Yang et al. 2010), and Ya'an, China in 2009. Finally, based on the genetic diversity information provided by single nucleotide polymorphism (SNP) markers and adaptation data obtained from the field experiments, 527 lines were chosen for the present study. Our objectives were: (1) to estimate the phenotypic and genetic diversity of the elite maize inbred collection; (2) to examine the population structure and familial relatedness of the elite maize inbreds collection; and (3) to evaluate the power and the appropriate statistical models of this panel for association analysis.

Materials and methods

Plant germplasm, phenotyping and genotyping

A set of 527 global diverse lines, representative of tropical, subtropical and temperate germplasm, was collected to construct a large association panel in maize. This collection included 54 lines from the GEM project, 235 lines from the CIMMYT maize breeding programs and 238 lines from China. The latter contained a small association panel of 143 elite lines of the parents of commercial hybrids widely used in China, lines derived from Chinese landraces, high-oil lines and high provitamin A lines (Yang et al. 2010). Pedigree details are summarized in Electronic Supplementary Material Table S1.

All 527 lines (excepted the 54 GEM lines) were divided into two groups (temperate and tropical/subtropical) based on pedigree information and planted in one-row plots in an incompletely randomized block design within the group with two replicates at the agronomy farm of Sichuan Agricultural University, Ya'an, Sichuan in the summer of 2009. Five plants and five self-pollinated mature ears per line were used to score plant and ear traits, respectively. The measured traits included the following: days to pollen shedding, days to silk, plant height, ear height, leaf width, leaf length, tassel length, number of tassel branches, ear length, ear diameter, cob diameter and number of kernel rows.

Leaf tissue samples for the whole association panel were obtained from the bulk of at least six individuals for each line. DNA was extracted by a modified CTAB procedure according to Murray and Thompson (1980). All 527 lines were genotyped using GoldenGate assays (Illumina, San Diego, CA, USA) containing 1,536 SNPs (Yan et al. 2010c). The SNP genotyping was performed on an Illumina BeadStation 500G at Cornell University Life Sciences Core Laboratories Center using the protocol supported by Illumina Company (Fan et al. 2006). The details of the SNP genotyping procedure and allele scoring have been described in previous studies (Yan et al. 2010c). Of the 1,536 SNPs genotyped, 1,067 SNPs having missing data less than 20% and of good quality were used for subsequent analysis, among which 926 SNPs with minor allelic frequencies (MAFs) $\geq$0.1 were used for genotypic data analyses.

Genotypic data analyses

The number of alleles, MAFs, gene diversity, observed heterozygosity and polymorphic information content (PIC) were calculated using Powermarker version 3.25 (Liu and Muse 2005). The statistical significance of differences in all estimators, except the number of alleles, was assessed across loci using the Wilcoxon paired test.

Two Bayesian Markov chain Monte Carlo programs, STRUCTURE (Pritchard et al. 2000a; Falush et al. 2003) and INSTRUCT (Gao et al. 2007), were used to infer population structure and to assign genotypes to subpopulations. In both analyses, three independent simulations having 150,000 MCMC (Markov chain Monte Carlo) replications and

100,000 burn-ins were performed with the number of subpopulations (k) ranging from 1 to 15. For STRUCTURE, the ancestry model allowed for population mixture and correlated allele frequencies; the k value was determined by the log likelihood of the data (LnP(D)) in the STRUCTURE output and an ad hoc statistic Δk based on the second-order rate of change in LnP(D) between successive k (Evanno et al. 2005). INSTRUCT runs allowed inference of population structure and selfing rates at individual levels. Optimal k was inferred using the log likelihood of the data and deviance information criterion (DIC). Results of replicate simulations from both programs were integrated by using the CLUMPP software (Jakobsson and Rosenberg 2007). The correlation coefficients of membership probabilities estimated from STRUCTURE and INSTRUCT were calculated for each k using PROC CORR in SAS 8.02 (SAS Institute 1999). To compare the results from STRUCTURE/INSTRUCT with the pedigree knowledge, lines with membership probabilities $\geq$0.60 were assigned to corresponding clusters; lines with membership probabilities <0.60 were assigned to a mixed group. Structure results of individual assignments to corresponding groups were graphically displayed using the DISTRUCT software (Rosenberg 2004).

An analysis of molecular variance (AMOVA) (Excoffier et al. 1992) and F-statistics (F_{st}) across all subpopulations and between pairwise subpopulations were performed using Arlequin V3.11 (Excoffier et al. 2005) to investigate population differentiations among the subpopulations clustered by STRUCTURE. Additionally, Nei's genetic distances (Nei 1972) among these given subpopulations and individuals were calculated using Powermarker version 3.25 (Liu and Muse 2005). The Nei's genetic distance (Nei 1972) among individuals was then used to construct a neighbor-joining (NJ) phylogenetic tree with 1000 runs of bootstrapping using Powermarker version 3.25 (Liu and Muse 2005). Furthermore, the Nei's genetic matrices created were double-centered, and used to obtain eigenvectors by the modules DCENTER and EIGEN implemented in NTSYSpc 2.1 (Rohlf 2000). Finally, the relative kinship coefficients were calculated using the SPAGeDi software package (Hardy and Vekemans 2002). All negative values between individuals were set to 0 (Yu et al. 2006).

Phenotypic data analyses

Descriptive statistical analyses were carried out using SAS 8.02 (SAS Institute 1999). The trait means of all lines were used in subsequent analyses. The Shannon–Weaver indices (Poole 1974), measuring genetic diversity in categorical data, were calculated to investigate the phenotypic diversity in this maize panel. The details have been described in a previous study (Yang et al. 2010). Briefly, the phenotypic values were subdivided into ten classes with an interval of 0.5 SD using the means and SD of each trait in the maize panel; the number and frequency of phenotypic classes were used to calculate the Shannon–Weaver indices as defined by Poole (1974). The effects of population structure on all traits were tested using PROC GLM in SAS 8.02 (SAS Institute 1999). When the population structure was estimated by STRUCTURE, the model statement included two of the three components of the $k = 3$ Q matrix from the STRUCTURE analysis, while the top 10 axes of variations from NTSYSpc analysis were included for PCA. The principal components with the top axis numbers ranging from 1 to 9 were additionally used to evaluate the effects of population structure on flowering time (represented by days to pollen shedding), ear height and ear diameter.

Evaluation of the maize association panel

The statistical power of this maize panel for identifying genetic factors associated with quantitative traits was evaluated using various population size and genetic effects without considering population structure and relative kinship. We assumed that the population size was 100, 200, 300, 400, 500, 600, 700, 800, 900 and 1,000, and that the genetic effect (the ratio of explained total phenotypic variance) was 0.01, 0.02, 0.03, 0.04 and 0.05. All evaluations were performed using the Genetic Power Calculator (GPC, Purcell et al. 2003). The LD coefficient, r^2, was assumed to be 0.8 as GWAS for numerous small genetic effects required high LD in maize ($r^2 > 0.8$) (Gore et al. 2009). The power of detecting the causal genetic factors was given at a type I error rate of 0.05.

To assess the effect of genetic relatedness on association mapping on various quantitative traits and to identify the perfect model to correct for genetic relatedness in this maize panel, data for three traits, including flowering time (represented by days to pollen shedding), ear height and ear diameter, were collected to conduct marker–trait associations using 1067 SNPs. These SNPs were not expected to have significant effects on related traits as they were randomly selected and had low marker coverage across the whole genome. Six models were used to evaluate the effects of population structure (Q, PC) and relative kinship (K) on three traits for marker–trait associations: (1) the GLM model, without controlling for population structure and relative kinship; (2) the Q model, controlling for Q; (3) the PCA model, controlling for PC; (4) the K model, controlling for K; (5) the Q + K model, controlling for both Q and K; and (6) the PCA + K model, controlling for both PC and K. The GLM, Q and PCA models were performed using a general linear model (GLM) in TASSEL V2.1; the K, Q + K and PCA + K models were performed using mixed linear model (MLM) in TASSEL V2.1 (Yu et al. 2006; Zhang et al. 2009). The quantile–quantile plots of estimated –$\log_{10}$(p) were displayed using the observed P values from marker–trait associations, and deviations from the expectation demonstrated that the statistical analysis may cause spurious associations. Before carrying out model comparisons, we first determined the optimal dimension for PCA and PCA + K models by testing the PCA models with various numbers of dimensions using these three traits and the 1067 SNPs measured above. The top 10 axes of variations from the NTSYSpc analysis were used.

Results

Phenotypic variations of measured quantitative traits

Extensive phenotypic variations were observed for all the measured quantitative traits in this maize panel, as shown by the descriptive statistics in Table 1. The number of tassel branches, which varied from 1.7 to 27.7 with an average of 10.8, had the highest maximum change of 16.3-fold, while days to pollen shedding, which varied from 60.5 to 97.5 days with an average of 79.6 days, had the lowest change (1.6-fold). The Shannon–Weaver index (H') across all traits further confirmed that this panel encompassed abundant phenotypic diversity. An average of 2.06 (±0.01) for H' was investigated with a range from 2.04 (number of kernel rows) to 2.08 (plant height).

Summary of SNPs

An even distribution of MAFs was observed (Fig. S1) with 50 continued classes from 0.01 to 0.50 with a similar number of SNPs in each MAF class. Only

Table 1 Phenotypic variations for 12 traits and the effects of population structure on each trait in this maize panel

Traits	Min ± SD	Max ± SD	Mean ± SD	H'[a]	R^2_{Q}[b]	R^2_{PCA}[c]
Days to pollen (days)	60.5 ± 0.7	97.5 ± 3.5	79.6 ± 6.1	2.06	37.3	40.0
Days to silk (days)	60.5 ± 0.7	102.0 ± 1.4	82.3 ± 6.6	2.06	30.3	34.6
Plant height (cm)	105.2 ± 0.9	235.2 ± 17.3	172.6 ± 25.9	2.08	7.9	22.6
Ear height (cm)	14.9 ± 4.4	125.4 ± 1.4	65.8 ± 19.0	2.07	15.9	24.1
Leaf width (cm)	5.5 ± 0.7	12.4 ± 0.0	9.0 ± 1.2	2.07	0.6	5.3
Leaf length (cm)	49.7 ± 0.7	109.4 ± 0.3	76.0 ± 10.1	2.07	19.3	26.3
Tassel length (cm)	16.1 ± 2.2	50.3 ± 6.2	29.4 ± 4.7	2.06	17.1	21.7
Number of tassel branches	1.7 ± 0.4	27.7 ± 5.0	10.8 ± 4.8	2.05	5.7	12.4
Ear length (cm)	5.0 ± 0.0	18.2 ± 0.0	11.0 ± 2.1	2.07	1.1	9.4
Ear diameter (cm)	1.9 ± 0.6	4.8 ± 0.0	3.3 ± 0.5	2.06	6.9	16.1
Cob diameter (cm)	1.0 ± 0.0	3.1 ± 0.1	2.1 ± 0.3	2.06	1.9	9.9
Number of kernel rows	8.0 ± 0.0	19.7 ± 0.5	12.7 ± 1.8	2.04	3.7	12.6

[a] Shannon–Weaver index

[b] Percentage of phenotypic variation explained by population structure estimated by STRUCTURE

[c] Percentage of phenotypic variation explained by population structure estimated by PCA

13.3% (142/1067) of the SNPs had a MAF of less than 0.1. Most of the SNPs were mapped in silico and/or genetically in maize chromosomes, and were well distributed in the 10 chromosomes (Yan et al. 2010c). For any two given lines, the polymorphism ratio varied from 0.3 to 76.6%, with an average of 40.3%. The highest level of polymorphism occurred between the lines CIMBL153 and GEMS53, and the lowest occurred between CY72 and 4F1. The average polymorphism ratio for any given line to the other 526 lines ranged from 34.9% for CML474 to 72.0% for GEMS53. The 926 SNPs with MAFs of ≥0.1 were used to estimate the genetic diversity of the maize panel (Fig. 1). In total, 1,852 alleles were detected with an expected average of 2.0 alleles per loci. The MAFs of 926 SNPs averaged 0.30, and about 50% of the SNPs had an MAF greater than 0.30. For all these SNPs, heterozygosity, gene diversity and PIC varied from 0.01 to 0.21, 0.18 to 0.50 and 0.16 to 0.38, with an average of 0.04, 0.39 and 0.31, respectively.

Population structure and genetic clustering

To examine the relatedness among 527 lines, the data for 926 SNPs were first analyzed using STRUCTURE. The LnP(D) value for each given k (the number of subpopulations based on the model) increased with the increase in k but did not show evidence of a maximum (Fig. 2a). The second-order likelihood, Δk, was then calculated: $k = 2$ showed a much higher likelihood than $k = 3$–15 among all runs of the program (Fig. 2a). Furthermore, the Δk also decreased sharply when k increased from 3 to 5, and that at $k = 3$ was significantly higher than at $k = 4$. Accordingly, $k = 2$ and $k = 3$ were considered as the two best possible numbers of subpopulations. This was further supported by INSTRUCT analysis as there appeared a inflection point at $k = 2$, 3 for the log P(D) and DIC (Fig. 2b). The membership probabilities for all lines from INSTRUCT were highly correlated with those from STRUCTURE ($R^2 > 0.99$, all $P < 0.001$) when k ranged from 2 to 4, indicating ancestry assignments to corresponding subpopulations were similar between the two methods. Consequently, only the STRUCTURE results are shown. The individual assignments by varying the presumed number of subpopulations suggested k to be 3 (Fig. 2c), similar to the known pedigree and germplasm. The first level of clustering ($k = 2$) reflects the primary division of a subpopulation from all lines, representative of B73 and termed SS. At $k = 3$, another subpopulation, consisting of mostly tropical or subtropical lines, separated following SS and was termed TST. The third subpopulation, termed NSS, contained most temperate lines except the lines within the SS subpopulation. From the NSS subpopulation, smaller subpopulations were further separated when k increased to 4, and consequently, more lines indicated mixed ancestry. In summary, this

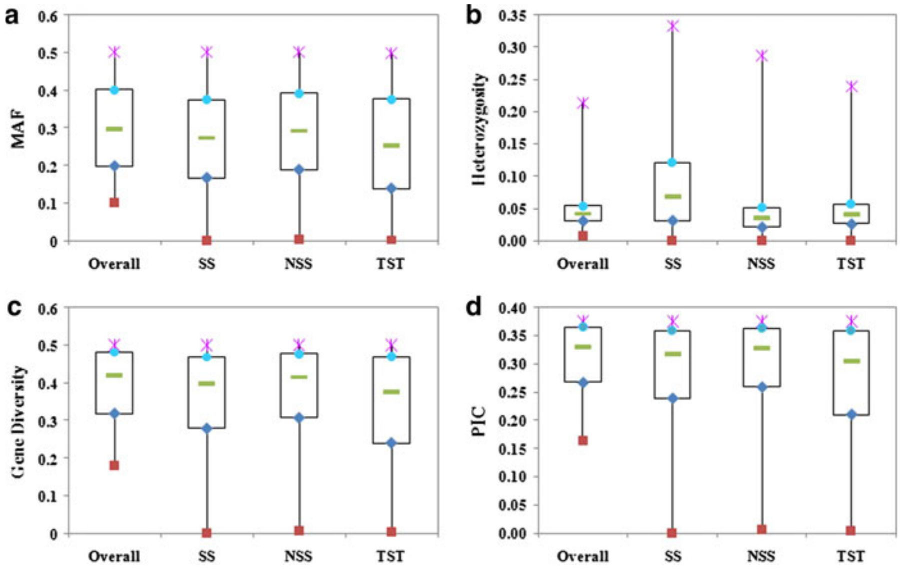

Fig. 1 *Box* and *Whisker box* of summary statistics for 926 SNPs in all inbreds and each subpopulation in 527 lines. **a** MAF; **b** heterozygosity; **c** gene diversity; **d** polymorphic information content (PIC)

maize panel was clustered into three clear subpopulations with 33 SS lines, 143 NSS lines and 232 TST lines; the remaining 119 lines were thus classified into a mixed subpopulation as they had membership probabilities lower than 0.60 in any given subpopulation (Table S2).

Furthermore, the NJ phylogenetic tree based on Nei's genetic distances displayed a similar pattern of relationships among the 527 lines estimated by STRUCTURE, with minor difference (Fig. 3a). The tree had three clear clades with the lines within mixed subpopulation distributing across the whole tree. Except for the lines from the mixed subpopulation, the smallest clade of 58 lines corresponded to SS with 27 lines from SS, 3 lines from NSS, and 18 lines from TST; the largest clade of 214 lines corresponded to TST with all lines from TST; the remaining clade of 147 lines corresponded to NSS with 140 lines from NSS and 7 lines from SS. No clearly distinct clusters were further identified within these three clades.

Similarly, PCA based on Nei's genetic distances presented a picture with all lines separating into SS, NSS and TST subpopulations, with the mixed subpopulation being in the middle of these three defined subpopulations (Fig. 3b). The top two principal components clearly separated these subpopulations. The first principal component (PC1) accounted for 18.2% of the genetic variation in this maize panel and reflected the differentiation between SS and NSS or TST, whereas the second (PC2) accounted for 6.9% of the genetic variation and reflected the differentiation between NSS and TST. It appeared that SS was relatively distant from NSS and TST, while NSS and TST were close to each other in this collection.

Population divergence and genetic diversity

The results for detecting the three subpopulations by using different statistical methods such as STRUCTURE/INSTRUCT, NJ tree-based, and PCA were quite consistent. F_{st} values across the three subpopulations averaged 0.11 ($P < 0.001$), which was confirmed by AMOVA analysis, and we found that only 10.7% ($P < 0.001$) of the total genetic variation was partitioned among subpopulations and 89.3% ($P < 0.001$) within subpopulations. The pairwise F_{st} between the three subpopulations was 0.25 ($P < 0.001$) for SS versus NSS, 0.31 ($P < 0.001$) for SS versus TST and 0.09 ($P < 0.001$) for NSS versus TST. This result demonstrates that there was much higher differentiation between SS and NSS or TST, while it was significantly lower between NSS and TST. A similar pattern of differentiation among subpopulations was supported by Nei's minimum genetic distance (Table S3).

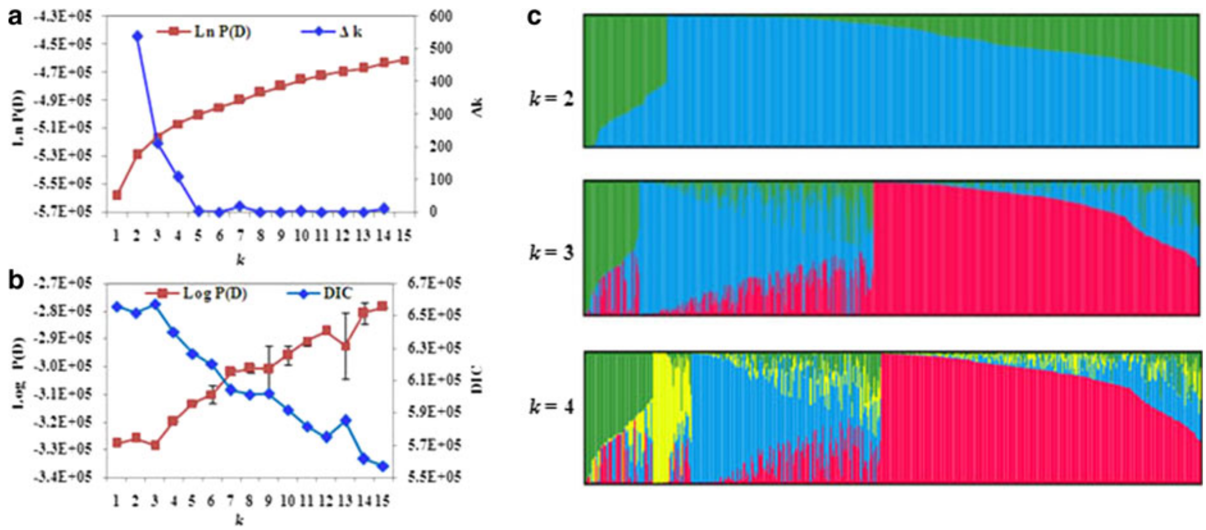

Fig. 2 Analysis of the population structure of 527 maize inbred lines. **a** Estimated LnP(D) and Δk over three repeats of STRUCTURE analysis; **b** Estimated Log P(D) and DIC over three repeats of INSTRUCT analysis; **c** Population structure assessed by STRUCTURE. Each individual is represented by a vertical bar, partitioned into colored segments with the length of each segment representing the proportion of the individual's genome from $k = 2$, 3 and 4 groups. For all classes, a given group is represented: *Green*, SS; *Blue*, NSS; *Red*, TST; *Yellow* , a small group separated from NSS. (Color figure online)

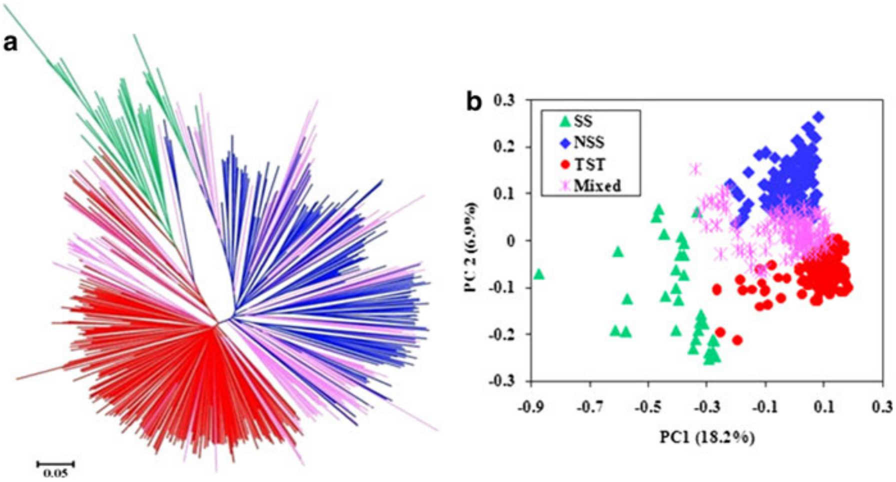

Fig. 3 Neighbor-joining phylogenetic tree based on Nei's genetic distance (**a**) and PCA plot (**b**) for 527 maize inbred lines. For NJ-tree, *SS* green, *NSS* blue, *TST* red, *Mixed* pink. (Color figure online)

MAF, heterozygosity, gene diversity and PIC for 926 SNPs were also estimated within inferred subpopulations and compared (Fig. 1). Compared to the entire panel, all three subpopulations including SS, NSS and TST had significantly lower MAFs ($z = -15.2$ to -4.2, all $P < 0.01$) (Fig. 1a). Among subpopulations, MAFs within SS (0.27) and TST (0.26) were similar but lower than within NSS (0.29). A similar picture was also obtained for genetic diversity within subpopulations estimated by gene diversity and PIC (Fig. 1c, d). For heterozygosity of SNPs, SS had the highest level of heterozygosity among all subpopulations, followed by TST and NSS in that order (Fig. 1b).

Relative kinship

The distribution of kinship coefficients between 0 and 0.50 (Fig. S2) represents 99.7% of the data. A total of 56.1% of kinship coefficients were 0, suggesting that there was no relatedness between these pairs of lines. A significant fraction (38.0%) indicated weak similarity, with kinship coefficients varying from 0 (excluding 0) to 0.10. Only 5.6% showed various degrees of relatedness, with kinship coefficients ranging from 0.10 (excluding 0.10) to 0.50. For the remaining 0.3% of the data, the kinship coefficients varied from 0.50 to 1.26 with an average of 0.68. This pattern of genetic relatedness demonstrated that few lines showed strong similarities, and most lines were weakly or modestly related in this complex maize panel.

Effects of sample size on association power

The statistical power for detecting the significant variants in an association panel was assessed with various sample sizes and effect sizes (Fig. S3). When the sample size of an association panel was less than 100, less than 33% of the significant variants were captured even when they had moderate effects (effect = 0.05). As expected, a significant increase of power was observed when the sample size or genetic effects increased. When the sample size reached 500, the association panel was large enough to capture most of the significant variants (>78%) accounting for over 3% of the phenotypic variation. For this panel with 527 individuals, over 81% of the significant variants explaining ≥3% of phenotypic variations were captured, 62% for 2%, and 35% for 1%.

Correction of spurious associations

Various levels were observed for effects of population structure on phenotype in this maize panel (Table 1). For all measured traits, the percentage of phenotypic variations explained by the Q matrix from STRUCTURE analysis (R^2_{Q}) averaged 12.1%, with a range from 0.6 (leaf width) to 37.3% (days to pollen shedding). Half of the traits were influenced weakly by population structure as the R^2_{Q} values were lower than 10%. The phenotypic variations explained by the top 10 axes of variations from PCA analysis (R^2_{PCA}) showed similar difference levels among various traits for effects of population structure. However, the R^2_{PCA} values were higher than the R^2_{Q} values, especially for plant height with the values increasing from 7.9% for R^2_{Q} to 22.6% for R^2_{PCA}. Subsequently, three traits, representing three levels of the effects of population structure on phenotype, were used to determine the optimal dimension of PCA for associations, and to evaluate the performance of various statistical models in controlling spurious associations. Population structure, estimated by STRUCTURE and PCA, accounted for 37.3% (40.0%) of phenotypic variations for flowering time (represented by days to pollen shedding), 15.9% (24.1%) for ear height, and 6.9% (16.1%) for ear diameter in this maize panel.

The *P* value distributions seen in Fig. 4 show that association analysis with principal components resulted in the reduction of false positives for all traits. The portion of corrected false-positive associations increased significantly when the axes of variation increased from 1 to 4 (flowering time and ear height) or 3 (ear diameter), but were virtually identical when the axes were greater. With the axes of variations varying from 1 to 4, the number of significant SNPs identified at $P < 0.001$ reduced from 157 to 25 for flowering time, 79–11 for ear height, and 34 to around 20 for ear diameter. Furthermore, the significant SNPs identified at $P < 0.001$ were almost the same when axes of variations varied between 4 and 10 (Table S4). According to these tests, the false-positive corrections performed the best using the variations with number of axes greater than 4 when only using principal components to account for genetic structure. This was further supported by the percentage of phenotypic variation explained by the axes of variations (Fig. S4). As the association analysis using principal components was sensitive to the number of axes of variations (when the number is small) as well as traits, we finally choose the top 10 axes of variations to account for genetic structure.

With the optimal number of dimensions for principal components, we evaluated the performance of six statistical models for controlling false positives in this maize panel (Fig. 5). For all three traits, any model controlling population structure or relative kinship performed significantly better than the GLM model. For flowering time, the PCA + K and Q + K models were a little better than the K model; but for the other two traits, the three models were similar with the Q + K model having the greatest success in reducing the type I errors. Without considering the relative kinship, the PCA model showed better type I control than the Q model. Comparing Fig. 5a, c, it is obvious that the type I error control was sensitive to traits, consistent with the different percentage of phenotypic variations explained by the population structure. For flowering time, the number of significant SNPs identified at $P < 0.001$ was reduced from 304 (GLM) to 2 (Q + K or PCA + K); 153 (GLM) to 1 (Q + K) for

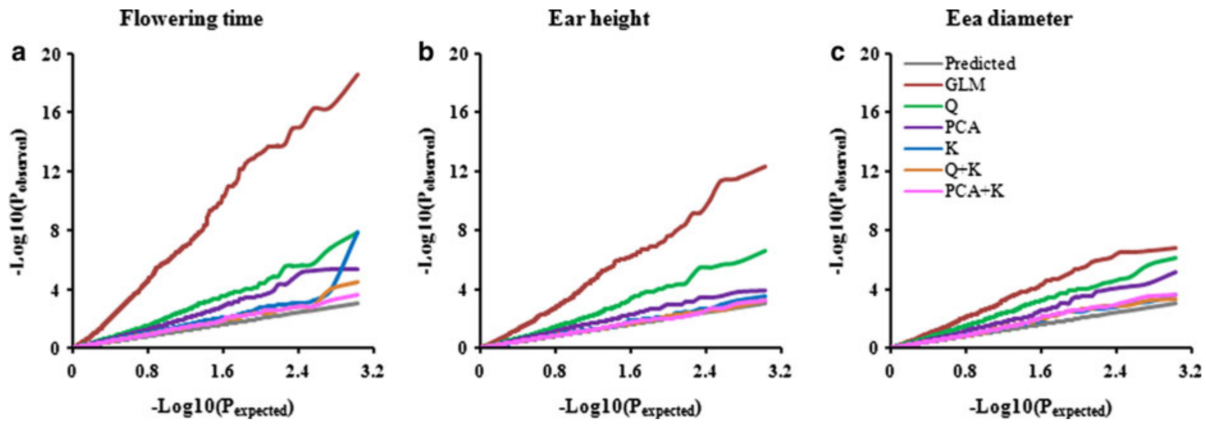

Fig. 4 *Quantile–quantile plots* of estimated $-\log_{10}(p)$ from association analysis using the PCA model with various dimensions of three traits: **a** flowering time; **b** ear height; **c** ear diameter. The *black line* is the expected line under the null distribution. Under the assumption that there are few true marker associations, the observed *P* values are expected to nearly follow the expected *P* values. The deviations from the expectation demonstrate that the statistical analysis may cause spurious associations. (Color figure online)

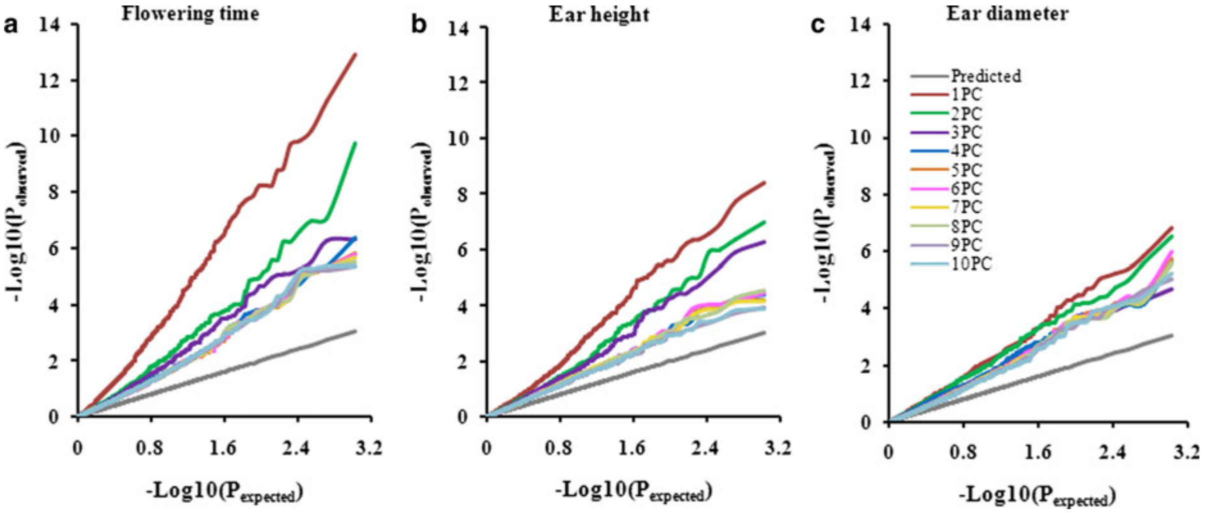

Fig. 5 Quantile–quantile plots of estimated $-\log_{10}(p)$ from association analysis using six methods in three traits: **a** flowering time; **b** ear height; **c** ear diameter. The *black line* is the expected line under the null distribution. Under the assumption that there are few true marker associations, the observed *P* values are expected to nearly follow the expected *P* values. The deviations from expectation demonstrate that the statistical analysis may cause spurious associations. (Color figure online)

ear height; and 76 (GLM) to 2 (Q + K) for ear diameter (Table S5).

Discussion

Genetic diversity and population structure

SNPs are biallelic and less informative compared to multiallelic simple sequence repeats (SSRs). Nevertheless, the high-throughput and cost-effective SNP genotyping system can achieve a large number of markers (i.e. thousands to millions) that may overcome the disadvantages of the SNPs. The utility of SSRs and SNPs in assessment of population structure was compared using 89 SSRs and 847 SNPs, showing that SSRs performed better at clustering individuals into populations than SNPs, but that the population structure assessed by both marker systems was consistent (Hamblin et al. 2007). More recently, SNP markers have been used successfully for maize genetic diversity and population structure analysis with fairly consistent results (Lu et al. 2009; Yan et al. 2009; Inghelandt et al. 2010; Yang et al. 2010). Inghelandt et al. (2010) and Yu et al. (2009) further pointed out that between 7 and 11 times more SNPs than SSRs should be used for analyzing population structure and genetic diversity, while 10 times more should be used for estimating relative kinship. Therefore, the genetic diversity and relatedness presented here should be similar to those assessed by about 100 SSRs with 10 SSRs per chromosome on average.

The 527 maize lines had a higher gene diversity value (0.39) when compared to the gene diversity of around 0.32 across three sets of maize inbred lines: a set of 259 lines from the USA representative of global diversity (Hamblin et al. 2007), a set of 770 global maize lines from six countries (Lu et al. 2009) and a set of 1537 elite lines representing European and North-American diversity (Inghelandt et al. 2010). The comparably abundant genetic diversity in this panel was primarily due to the broad range of the germplasm, as all maize collections were estimated using similar numbers of SNP markers (except the European and North-American collection which were estimated using many more markers than the others).

Our study identified three separate subpopulations within the global maize germplasm. This result was confirmed by STRUCTURE, INSTRUCT, PCA and phylogenetic tree-based analyses. Despite the wide application of the STRUCTURE program for identifying population structure, spurious inference of population structure often occurred for partially selfed populations because of the algorithm assumptions (Falush et al. 2003). The panel of inbred maize lines did not conform to the assumption of Hardy–Weinberg equilibrium, and low genetic divergence occurred among some individuals because of similar pedigree. Based on STRUCTURE, Gao et al. (2007) developed another approach, INSTRUCT, to assign individuals within partially selfing populations into more appropriate subpopulations without the assumption of Hardy–Weinberg equilibrium. However, the inferred population structure was consistent for both methods, although individual assignments tended to show low correlation when *k* was set above 7 (data not shown). Although consistent subpopulations were inferred by STRUCTURE, INSTRUCT, PCA and tree-based analyses, the assessment of individual relationships was different: membership probabilities or values are given by STRUCTURE/INSTRUCT or PCA, whereas tree-based analysis only classifies individuals to a fixed position on a tree. This demonstrates that a phylogenetic tree does not represent well the relationships among individuals with complex genetic relatedness, although it may generate similar results. This phenomenon also explained why individuals within mixed subpopulations inferred by STRUCTURE/INSTRUCT clustered into one of the three clades.

Historically, two maize groups, temperate and tropical/subtropical, were formed during the distribution of maize from the tropical center of origin in Mexico to northern climates. The population subdivision in this panel, like that by Liu et al. (2003) and Flint-Garcia et al. (2005), supported maize adaptation and separated temperate and tropical/subtropical lines into independent subpopulations, SS + NSS and TST, as described by Yan et al. (2009). However, investigations by Vigouroux et al. (2008) and Camus-Kulandaivelu et al. (2006) suggested that the Northern Flint subpopulation played a unique role in the adaptation of maize to temperate climates and was the first to split from the maize population. The difference is probably caused by few Northern Flint lines in this association panel. Further, the temperate population in this panel was divided into two subpopulations, SS and NSS. The pairwise F_{st} of 0.25 ($P < 0.001$) also indicated the existence of significant population differentiation between the SS and NSS subpopulations. Additionally, a few small subpopulations separated from NSS with the increase of *k* presumed in STRUCTURE analysis, consistent with previous studies in which three to seven subpopulations still existed apart from the SS subpopulation, namely the Reid subpopulation in the studies of Wang et al. (2008) and Yang et al. (2010) and the BSSS subpopulation in Xie et al. (2008) and Lu et al. (2009). Contrary to NSS, SS and TST did not show any separation except for a few lines clustering with the mixed subpopulation. This may be due to the limited lines used for SS and the CIMMYT maize breeding history for TST, as discussed previously (Dhliwayo et al. 2009).

The fact that maize originated from tropical environments suggests that there should be much more diversity in the tropical lines, which is well supported by previous estimation using molecular markers (Liu et al. 2003; Yan et al. 2009). Although most of the lines were of the tropical/subtropical group in this study, the genetic diversity of this group was not the richest. A similar finding was also reported in a previous study using the same set of SNP markers in 770 maize inbred lines representing tropical, subtropical and temperate maize germplasm (Lu et al. 2009). This may be partially due to the fact that the SNPs used in this study were originally developed from 27 founder lines between a common temperate line B73 and 26 other diverse (temperate and tropical) lines (McMullen et al. 2009), potentially causing bias estimation of the diversity, especially within the tropical and subtropical germplasm (Hamblin et al. 2007; Lu et al. 2009; Yan et al. 2010c). However, Lu et al. (2009) also demonstrated that these types of markers had no or very few effects on the inferences of population structure. With the development of next-generation sequencing techniques, the inexpensive cost of sequencing a genome (Varshney et al. 2009) may allow further detailed studies to achieve the final solution to this problem.

Power and statistical models

When using an association panel to uncover a variant for quantitative traits of interest, a primary consideration should be the power of this panel, namely the probability of detecting the causal variant. Studies of power evaluations have suggested that population size is one of the most fundamental decisions when identifying associations (Long and Langley 1999; Spencer et al. 2009). It is reasonable that as the population size increases, the probability of identifying presumed causal alleles present in a panel will consequently increase. Long and Langley (1999) found that a panel with 500 individuals was sufficient to detect the presence of causal variants even with small effects, consistent with our results. Our panel with 527 individuals achieved over 81% of the variants explaining ≥3% of phenotypic variations based on the assumption of an LD coefficient of 0.8. This indicated that the population size of our panel is suitable for most quantitative traits with modest effects.

Marker density is another determinant for increasing the power of association analysis (Mackay et al. 2009), especially for GWAS. It is often related to the LD pattern of a maize association panel at the genome-wide level. The number of SNPs genotyped in this panel is too small to precisely estimate LD between linked SNPs. Nevertheless, the LD pattern in a maize panel of 632 diverse breeding lines was roughly estimated within 2–5 kb at the genome-wide level when the LD coefficient was set to 0.1 ($r^2 = 0.1$) using 1536 SNPs from 582 genes (Yan et al. 2009). We supposed that a similar LD pattern occurred in our association panel based on a similar genetic diversity. Therefore, 240,000–480,000 markers will be required to perform GWAS when the LD coefficient is set to 0.1. The cutoff of $r^2 = 0.1$ may be too low to achieve enough power, and we would need to add more markers to increase the LD level to $r^2 = 0.8$. However, r^2 did not increase significantly when the LD distance decreased tenfold to between 0.2 and 0.5 kb in the study by Yan et al. (2009). This result implied that the power may not be increased significantly even if the number of markers is increased tenfold to between 2.4 and 4.8 million. Myles et al. (2009) roughly estimated that more than 10 million markers may ideally be needed to perform GWAS in maize, and this would be a significant challenge. Since less than 10% of the maize genome encodes over 32,000 genes (Schnable et al. 2009), an alternative should be to develop markers only from gene-rich regions using the exome resequencing strategy (Ng et al. 2009) similar to that developed and applied in human studies with the rapid development of next-generation sequencing technology. This approach may allow us to use the most informative markers for GWAS in maize to achieve the highest power (Yan et al. 2010b).

An ideal association panel is a population with uniform genetic background, which will not significantly influence the expression of traits. Most cultivated plants, such as maize, which have experienced domestication and breeding, show complex patterns of genetic relatedness among individuals. In such a situation, association analysis often generates a large number of false-positive associations, especially for the traits associated with adaptation (Yu and Buckler 2006; Zhu et al. 2008; Myles et al. 2009). As expected, the integration of genetic relatedness into statistical methods greatly reduced spurious associations in our study, as well as in others (Yu et al. 2006; Zhu and Yu 2009; Stich et al. 2008; Yang et al. 2010). The Q model performed well for correcting false-positive associations, although it did not completely control the population structure. This is because the STRUCTURE program divides the maize panel into a few discrete populations, and the Q matrix only gives a rough dissection of population differentiation. Consequently, it was suggested that the PCA model dissects the phenotypic variation from population structure along continuous axes (Patterson et al. 2006; Price et al. 2006). Indeed, the PCA model performed better compared to the Q model; however, a few residual false-positive associations still existed.

It was reported that association analysis using the PCA model was not sensitive to the dimension numbers of principal components (Patterson et al. 2006; Price et al. 2006). In the present study, we found that the type I error control was sensitive to the number of PCs when the axes ranged from 1 to 4, but the results were consistent when they were ≥4. The dimension numbers of principal components for type I error control were also trait-dependent (Fig. 4; Table S4). This is not surprising, as the effects of population structure vary among complex quantitative traits in maize (Flint-Garcia et al. 2005; Yang et al. 2010). The PCA models with the first top component reduced nearly half of the false-positive associations for all traits. All these results suggest that it is appropriate to evaluate the correlation between PCA models and the dimension number of principal components for various traits in a given newly constructed association panel.

The K model and two mixed models (Q + K and PCA + K) performed well for all traits in this maize model. It seems that a K matrix incorporated into the K model was sufficient to minimize false-positive associations, consistent with other model simulations and comparisons (Yu et al. 2006; Zhu and Yu 2009; Stich et al. 2008; Yang et al. 2010). However, both the Q + K and PCA + K models performed slightly better than the K model for all three traits. As the estimation of the Q matrix is computationally intense, the PCA + K model may be an ideal choice for association analysis, especially for most large, genome-wide data sets.

Potential utilization of this maize panel and further perspectives

An adequate understanding of genetic variation, pattern of complex genetic relatedness and the performance of statistical methods in the maize panel allowed us to uncover the actual variants affecting quantitative traits. A subset of this panel with 155 lines has been successfully applied using a candidate-gene association approach in the validation of several genes controlling relatively simple traits, such as *crtRB1* (Yan et al. 2010a), *ZmGW2* (Li et al. 2010a), and *ZmGS3* (Li et al. 2010b). The detected variants can be converted into functional markers (Andersen and Lübberstedt 2003) and then used for maize improvement by marker-assisted selection. For example, the action of two genes, *lcyE* and *crtRB1*, encoding the key enzyme of the carotenoid pathway, was confirmed using association analysis and the functional sequence polymorphisms were used to develop functional markers related to β-carotenoid content (Harjes et al. 2008; Yan et al. 2010a). Using these functional markers, the favorable alleles from these two genes were jointly introgressed through marker-assisted selection into adapted elite tropical breeding lines (provitamin A ranging 8–10 μg g^{-1}) in HarvestPlus/CIMMYT breeding programs which target developing countries (Yan et al. 2010a). All the high provitamin A lines selected via marker-assisted selection in these programs are lines from, or derived from, one of the association mapping panels used to identify the favorable alleles. Thus, the natural variation of quantitative traits hidden in the maize panel will greatly assist targeted efforts to improve traits of interest.

Although an association mapping panel with about 500 genotypes can help to capture most of the variations that are $\geq$3%, it may still not be large enough to obtain sufficient power for some important complex quantitative traits in maize controlled by a great number of genes with small effects. For example, only a few QTL with effects greater than 3% in the approximately 50 QTL identified affected flowering time in the Nested Association Mapping (NAM) population (Buckler et al. 2009). Therefore, sample size should be one of the most important factors to be carefully considered in future maize GWAS, especially for the complex traits. In GWAS for human diseases, genes or SNPs with effects <0.5% were also identified when combining large numbers of individuals from different groups. For example, more than 50 variants (each of which can only explain 0.3–0.5% of variation) affecting human height were identified when about 63,000 individuals were combined (Visscher 2008). It would be very difficult for any single researcher or institution to handle such a large phenotyping maize trial. However, many maize association mapping panels with different sizes and genetic backgrounds have already been developed and phenotyped for the same or similar traits (Flint-Garcia et al. 2005; Yang et al. 2010) and could be genotyped using common and high-density markers (i.e. a commercial maize 50 K array) by different maize researchers worldwide. It will be extremely useful to exploit the genetic architecture of complex traits by combining the information from all the possible panels and traits using appropriate statistical methods in the future.

Acknowledgments We thank Drs. Pixley Kevin, Palacios Natalia and Taba Suketoshi from CIMMYT for providing partial materials used in this study. We gratefully acknowledge the editor Dr. Roberto Tuberosa and two anonymous reviewers for their valuable suggestions. This research was supported by the National Hi-Tech Research and Development Program of China (2006AA100103, 2006AA10Z183, 2006AA10A107), the NSFC-CIMMYT Project (30821140352), and the Genetically Modified Varieties Breeding Science and Technology Projects of China.

References

Andersen JR, Lübberstedt T (2003) Functional markers in plants. Trends Plant Sci 8:554–560

Andersen JR, Zein I, Wenzel G, Krützfeldt B, Eder J, Ouzunova M, Lübberstedt T (2007) High levels of linkage disequilibrium and associations with forage quality at a *Phenylalanine Ammonia-Lyase* locus in European maize (*Zea mays* L.) inbreds. Theor Appl Genet 114:307–319

Belo A, Zheng P, Luck S, Shen B, Meyer DJ, Li B, Tingey S, Rafalski A (2008) Whole genome scan detects an allelic variant of *fad2* associated with increased oleic acid levels in maize. Mol Genet Genomics 279:1–10

Buckler ES, Stevens NM (2005) Maize Origins, Domestication, and selection. In: Motley TJ, Zerega N, Cross H (eds) Darwin's Harvest. Columbia University Press, New York, pp 67–90

Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, Browne C, Ersoz E, Flint-Garcia S, Garcia A, Glaubitz JC, Goodman MM, Harjes C, Guill K, Kroon DE, Larsson S, Lepak NK, Li HH, Mitchell SE, Pressoir G, Peiffer JA, Rosas MO, Rocheford TR, Romay MC, Romero S, Salvo S, Villeda HS, da Silva HS, Sun Q, Tian F, Upadyayula N, Ware D, Yates H, Yu JM, Zhang ZW, Kresovich S, McMullen MD (2009) The genetic architecture of maize flowering time. Science 325:714–718

Camus-Kulandaivelu L, Veyrieras JB, Madur D, Combes V, Fourmann M, Barraud S, Dubreuil P, Gouesnard B, Manicacci D, Charcosset A (2006) Maize adaptation to temperate climate: relationship between population structure and polymorphism in the Dwarf8 gene. Genetics 172:2449–2463

Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55:997–1004

Dhliwayo T, Pixley K, Menkir A, Warburton M (2009) Combining ability, genetic distances, and heterosis among elite CIMMYT and IITA tropical maize inbred lines. Crop Sci 49:1201–1210

Ducrocq S, Giauffret C, Madur D, Combes V, Dumas F, Jouanne S, Coubriche D, Jamin P, Moreau L, Charcosset A (2009) Fine mapping and haplotype structure analysis of a major flowering time quantitative trait locus on maize chromosome 10. Genetics 183:1555–1563

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol 14:2611–2620

Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131:479–491

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

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164: 1567–1587

Fan JB, Gunderson KL, Bibikova M, Yeakley JM, Chen J, Garcia EW, Lebruska LL, Laurent M, Shen R, Barker D (2006) Illumina universal bead arrays. Meth Enzymol 410:57–73

Flint-Garcia SA, Thuillet AC, Yu JM, Pressoir G, Romero SM, Mitchell SE, Doebley J, Kresovich S, Goodman MM, Buckler ES (2005) Maize association population: a high-resolution platform for quantitative trait locus dissection. Plant J 44:1054–1064

Gao H, Williamson S, Bustamante CD (2007) A Markov Chain Monte Carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype data. Genetics 176:1635–1651

Gore MA, Chia JM, Elshire R, Sun Q, Ersoz E, Hurwitz BL, Peiffer JA, McMullen MD, Grills GS, Ross-Ibarra J, Ware DH, Buckler ES (2009) A first-generation haplotype map of maize. Science 326:1115–1117

Gupta PK, Rustgi S, Mir RR (2008) Array-based high-throughput DNA markers for crop improvement. Heredity 101:5–18

Hamblin MT, Warburton ML, Buckler ES (2007) Empirical comparison of simple sequence repeats and single nucleotide polymorphisms in assessment of maize diversity and relatedness. PLoS ONE 12:e1367

Hardy OJ, Vekemans X (2002) Spagedi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. Mol Ecol Notes 2:618–620

Harjes CE, Rocheford TR, Bai L, Brutnell TP, Kandianis CB, Sowinski SG, Stapleton AE, Vallabhaneni R, Williams M, Wurtzel ET, Yan JB, Buckler ES (2008) Natural genetic variation in lycopene epsilon cyclase tapped for maize biofortification. Science 319:330–333

Inghelandt DV, Melchinger AE, Lebreton CL, Stich B (2010) Population structure and genetic diversity in a commercial maize breeding program assessed with SSR and SNP markers. Theor Appl Genet 120:1289–1299

Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 21:1801–1806

Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E (2008) Efficient control of population structure in model organism association mapping. Genetics 178:1709–1723

Laurie CC, Chasalow SD, LeDeaux JR, McCarroll R, Bush D, Hauge B, Lai CQ, Clark D, Rocheford TR, Dudley JW (2004) The genetic architecture of response to long-term artificial selection for oil concentration in the maize kernel. Genetics 168:2141–2155

Li Q, Li L, Yang XH, Warburton ML, Bai GH, Dai JR, Li JS, Yan JB (2010a) Relationship, evolutionary fate and function of two maize orthologous genes of rice *GW2* associated with kernel size and weight. BMC Plant Biol 10:143

Li Q, Yang XH, Bai GH, Warburton ML, Mahuku G, Gore M, Dai JR, Li JS, Yan JB (2010b) Characterization of a putative *GS3* ortholog involved in maize kernel development. Theor Appl Genet 120:753–763

Liu KJ, Muse SV (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. Bioinformatics 21:2128–2129

Liu KJ, Goodman M, Muse S, Smith JS, Buckler ES, Doebley J (2003) Genetic structure and diversity among maize inbred lines as inferred from DNA microsatellites. Genetics 165:2117–2128

Long AD, Langley CH (1999) The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. Genome Res 9:720–731

Lu YL, Yan JB, Guimarães GT, Taba S, Hao ZF, Gao SB, Chen SJ, Li JS, Zhang SH, Vivek BS, Magorokosho C, Mugo S, Makumbi D, Parentoni SN, Shah T, Rong TZ, Crouch JH, Xu YB (2009) Molecular characterization of global maize breeding germplasm based on genome-wide single nucleotide polymorphisms. Theor Appl Genet 120:93–115

Mackay TFC, Stone EA, Ayroles JF (2009) The genetics of quantitative traits: challenges and prospects. Nat Rev Genet 10:565–577

McMullen MD, Kresovich S, Villeda HS, Bradbury P, Li HH, Sun Q, Flint-Garcia S, Thornsberry J, Acharya C, Bottoms C, Brown P, Browne C, Eller M, Guill K, Harjes C, Kroon D, Lepak N, Mitchell SE, Peterson B, Pressoir G, Romero S, Rosas MO, Salvo S, Yates H, Hanson M, Jones E, Smith S, Glaubitz JC, Goodman M, Ware D, Holland JB, Buckler ES (2009) Genetic properties of the maize nested association mapping population. Science 325:737–740

Murray MG, Thompson WF (1980) Rapid isolation of high molecular weight plant DNA. Nucleic Acids Res 8: 4321–4325

Myles S, Peiffer J, Brown PJ, Ersoz ES, Zhang ZW, Costich DE, Buckler ED (2009) Association mapping: critical considerations shift from genotyping to experimental design. Plant Cell 21:2194–2202

Nei M (1972) Genetic distance between populations. Am Nat 106:283–292

Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, Bamshad M, Nickerson DA, Shendure J (2009) Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461:272–276

Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLoS Genet 2:e190

Poole RW (1974) An introduction to quantitative ecology. McGraw-Hill, NY, USA, p 532

Pressoir G, Brown PJ, Zhu WY, Upadyayula N, Rocheford T, Buckler ES, Kresovich S (2009) Natural variation in maize architecture is mediated by allelic differences at the PINOID co-ortholog *barren inflorescence2*. Plant J 58:618–628

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904–909

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

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000b) Association mapping in structured populations. Am J Hum Genet 67:170–181

Purcell S, Cherny SS, Sham PC (2003) Genetic power calculator: design of linkage and association genetic mapping studies of complex traits. Bioinformatics 19:149–150

Rohlf FJ (2000) NTSYS-pc. Numerical taxonomy and multivariate analysis system, Version 2.1. Exeter Software, New York

Rosenberg NA (2004) DISTRUCT: a program for the graphical display of population structure. Mol Ecol Notes 4:137–138

Salvi S, Sponza G, Morgante M, Tomes D, Niu XM§, Fengler KA, Meeley R, Ananiev EV, Svitashev S, Bruggemann E, Li BL, Hainey CF, Radovic S, Zaina G, Rafalski JA, Tingey SV, Miao GH, Phillips RL, Tuberosa R (2007) Conserved non-coding genomic sequences associated with a flowering-time quantitative trait locus in maize. Proc Natl Acad Sci USA 104:11376–11381

Schnable PS, Ware D, Fulton RS, Stein JC, Wei FS, Pasternak S, Liang CZ, Zhang JW, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C, Strong C, Delehaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du FY, Kim K, Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B et al (2009) The B73 maize genome: complexity, diversity, and dynamics. Science 326:1112–1115

Spencer CCA, Su Z, Donnelly P, Marchini J (2009) Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. PLoS Genetics 5:e1000477

Stich B, Möhring J, Piepho HP, Heckenberger M, Buckler ES, Melchinger AE (2008) Comparison of mixed-model approaches for association mapping. Genetics 178: 1745–1754

Szalma SJ, Buckler ES, Snook ME, McMullen MD (2005) Association analysis of candidate genes for maysin and chlorogenic acid accumulation in maize silks. Theor Appl Genet 110:1324–1333

Tester M, Langridge P (2010) Breeding technologies to increase crop production in a changing world. Science 327:818–822

Thornsberry JM, GoodmanM M, Doebley J, Kresovich S, Nielsen D, Buckler ES (2001) *Dwarf8* polymorphisms associated with variation in flowering time. Nat Genet 28:286–289

Varshney RK, Nayak SN, May GD, Jackson SA (2009) Next-generation sequencing technologies and their implications for crop genetics and breeding. Trends Biotechnol 27:522–530

Vielle-Calzada JP, de la Vega OM, Hernández-Guzmán G, Ibarra-Laclette E, Alvarez-Mejía C, Vega-Arreguín JC, Jiménez-Moraila B, Fernández-Cortés A, Corona-Armenta G, Herrera-Estrella L, Herrera-Estrella A (2009) The palomero genome suggests metal effects on domestication. Science 326:1078

Vigouroux Y, Glaubitz JC, Matsuoka Y, Goodman MM, Sánchez GJ, Doebley J (2008) Population structure and genetic diversity of new world maize races assessed by dna microsatellites. Am J Bot 95:1240–1253

Visscher PM (2008) Sizing up human height variation. Nat Genet 40:489–490

Wang RH, Yu YT, Zhao JR, Shi YS, Song YC, Wang TY, Li Y (2008) Population structure and linkage disequilibrium of a mini core set of maize inbred lines in China. Theor Appl Genet 117:1141–1153

Wilson LM, Wllitt SR, lbáñes AM, Rocheford TR, Goodman MM, Buckler ES (2004) Dissection of maize kernel composition and starch production by candidate associations. Plant Cell 16:2719–2733

Xie CX, Warburton M, Li MS, Li XH, Xiao MJ, Hao ZF, Zhao Q, Zhang SH (2008) An analysis of population structure and linkage disequilibrium using multilocus data in 187 maize inbred lines. Mol Breed 21:407–418

Yan JB, Shah T, Warburton ML, Buckler ES, McMullen MD, Crouch J (2009) Genetic characterization and linkage disequilibrium estimation of a global maize collection using SNP markers. PloS ONE 4:e8451

Yan JB, Kandianis CB, Harjes CE, Bai L, Kim EH, Yang XH, Skinner D, Fu ZY, Mitchell S, Li Q, Fernandez MGS, Zaharieva M, Babu R, Fu Y, Palacios N, Li JS, DellaPenna D, Brutnell BucklerES, Warburton ML, Rocheford T (2010a) Rare genetic variation at *Zea mays crtRB1* increases β-carotene in maize grain. Nat Genet 42:322–327

Yan JB, Warburton M, Crouch J (2010b) Association mapping for enhancing maize genetic improvement. Crop Sci (in press)

Yan JB, Yang XH, Hector S, Sánchez H, Li JS, Warburton M, Zhou Y, Crouch JH, Xu YB (2010c) High-throughput SNP genotyping with the GoldenGate assay in maize. Mol Breed 25:441–451

Yang XH, Yan JB, Shah T, Warburton ML, Li Q, Li L, Gao YF, Chai YC, Fu ZY, Zhou Y, Xu ST, Bai GH, Meng YJ, Zheng YP, Li JS (2010) Genetic analysis and characterization of a new maize association mapping panel for quantitative trait loci dissection. Theor Appl Genet 121:417–431

Yu JM, Buckler ES (2006) Genetic association mapping and genome organization of maize. Curr Opin Biotechnol 17:1–6

Yu JM, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Holland JB, Kresovich S, Buckler ES (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38:203–208

Yu JM, Zhang ZW, Zhu CS, Tabanao DA, Pressoir G, Tuinstra MR, Kresovich S, Todhunter RJ, Buckler ES (2009) Simulation appraisal of the adequacy of number of background markers for relationship estimation in association mapping. Plant Genome 2:63–77

Zhang ZW, Buckler ED, Casstevens TM, Bradbury PJ (2009) Software engineering the mixed model for genome-wide association studies on large samples. Brief Bioinform 10:664–675

Zhao K, Aranzana MJ, Kim S, Lister C, Shindo C, Tang C, Toomajian C, Zheng H, Dean C, Marjoram P, Nordborg M (2007) An arabidopsis example of association mapping in structured samples. PLoS Genet 3:e4

Zheng G, Freidlin B, Li ZH, Gastwirth JL (2005) Genomic control for association studies under various genetic models. Biometrics 61:186–192

Zhu CS, Yu JM (2009) Nonmetric multidimensional scaling corrects for population structure in association mapping with different sample types. Genetics 182:875–888

Zhu CS, Gore M, Buckler ES, Yu JM (2008) Status and prospects of association mapping in plants. Plant Genome 1:5–20