Theor Appl Genet (2010) 121:417–431
DOI 10.1007/s00122-010-1320-y

ORIGINAL PAPER

Genetic analysis and characterization of a new maize association mapping panel for quantitative trait loci dissection

Xiaohong Yang · Jianbing Yan · Trushar Shah · Marilyn L. Warburton · Qing Li · Lin Li · Yufeng Gao · Yuchao Chai · Zhiyuan Fu · Yi Zhou · Shutu Xu · Guanghong Bai · Yijiang Meng · Yanping Zheng · Jiansheng Li

Received: 7 September 2009 / Accepted: 5 March 2010 / Published online: 27 March 2010

Abstract Association mapping based on the linkage disequilibrium provides a promising tool to identify genes responsible for quantitative variations underlying complex traits. Presented here is a maize association mapping panel consisting of 155 inbred lines with mainly temperate germplasm, which was phenotyped for 34 traits and genotyped using 82 SSRs and 1,536 SNPs. Abundant phenotypic and genetic diversities were observed within the panel based on the phenotypic and genotypic analysis. A model-based analysis using 82 SSRs assigned all inbred lines to two groups with eight subgroups. The relative kinship matrix was calculated using 884 SNPs with minor allele frequency $\geq$20% indicating that no or weak relationships were identified for most individual pairs. Three traits (total tocopherol content in maize kernel, plant height and kernel length) and 1,414 SNPs with missing data $<$20% were used to evaluate the performance of four models for association mapping analysis. For all traits, the model controlling relative kinship (K) performed better than the model controlling population structure (Q), and similarly to the model controlling both population structure and relative kinship ($Q + K$) in this panel. Our results suggest this maize panel can be used for association mapping analysis targeting multiple agronomic and quality traits with optimal association model.

Communicated by J. Yu.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-010-1320-y) contains supplementary material, which is available to authorized users.

X. Yang · J. Yan (✉) · Q. Li · L. Li · Y. Gao · Y. Chai · Z. Fu · Y. Zhou · S. Xu · G. Bai · Y. Meng · Y. Zheng · J. Li (✉)
National Maize Improvement Center of China, Beijing Key Laboratory of Crop Genetic Improvement, China Agricultural University, Yuanmingyuan West Road, Haidian, Beijing 100193, China
e-mail: yjianbing@gmail.com

J. Li
e-mail: lijiansheng@cau.edu.cn

J. Yan · T. Shah
International Maize and Wheat Improvement Center (CIMMYT), Apdo. Postal 6-641, 06600 Mexico D.F., Mexico

M. L. Warburton
USDA-ARS Corn Host Plant Resistance Research Unit, Box 9555, Columbia, MS 39762, USA

G. Bai
Agronomy College, Xinjiang Agricultural University, Urumqi 830052, Xinjiang, China

T. Shah
International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru, Hyderabad 502 324, Andhra Pradesh, India

Introduction

Association mapping based on linkage disequilibrium (LD), both for genome-wide and candidate-gene approaches, provides a powerful tool for dissecting quantitative traits in plants (Yu and Buckler 2006; Buckler and Gore 2007; Zhu et al. 2008). When compared with linkage analysis, association mapping has a number of advantages that include shorter research time, higher mapping resolutions and investigation of a greater number of alleles (Yu and Buckler 2006). Several association panels in maize, wheat and sorghum have been constructed for performing association mapping studies (Flint-Garcia et al. 2005; Maccaferri et al. 2006; Casa et al. 2008). As an outbred species, maize

encompasses a high level of phenotypic and molecular diversity. When considering nucleotide diversity, any two random maize lines differ from one another in 1.4% of total DNA, similar to the divergence observed between humans and chimpanzees (Buckler and Stevens 2005). LD decays very rapidly in maize landraces (within 1 kb) and diverse maize inbred lines (within 1–5 kb) (Tenaillon et al. 2001; Remington et al. 2001; Yan et al. 2009). The abundant diversity and rapid LD decay make maize as an ideal crop for association mapping. The maize panel of 302 inbred lines assembled for association mapping by Flint-Garcia et al. (2005) consisted of both current breeding lines and historically important lines from temperate and tropical regions, representing a large fraction of the global genetic diversity in maize breeding. The genetic diversity and population structure were evaluated using 94 SSRs (Liu et al. 2003), and the panel or part of the panel has been used for association analysis for flowering time (Remington et al. 2001; Thornsberry et al. 2001; Pressoir et al. 2009), kernel compositions and starch pasting properties (Wilson et al. 2004), maysin synthesis (Szalma et al. 2005) and carotenoid content (Harjes et al. 2008; Yan et al. 2010b). Other association panels representative of American and European diversity have also been used in maize association analysis (Andersen et al. 2005, 2008; Camus-Kulandaivelu et al. 2006; Salvi et al. 2007; Belo et al. 2008), and a large-scale maize QTL/association mapping population (nested association mapping, or NAM), comprised 5,000 recombinant inbred lines derived from the common parent, B73, crossed to each of 25 diverse founder lines, was constructed to dissect the genetic basis of quantitative traits with great power (Yu et al. 2008; McMullen et al. 2009; Buckler et al. 2009).

The resources mentioned above are very useful for the maize community. The choice of germplasm for association mapping, collected from elite inbred lines, diverse inbred lines or landraces, is a key issue for the success of association analysis (Flint-Garcia et al. 2003; Breseghello and Sorrells 2006; Yu and Buckler. 2006; Zhu et al. 2008). An ideal association panel should harbor as much genetic diversity as possible, which is often used to resolve complex trait variation to a single gene or nucleotide. However, the genetic diversity should be balanced with genetic homogeneity of phenotypic traits, to ensure equal adaptation of all lines in multiple environments for phenotypic data collection. Because most of the maize in China is planted in temperate region, it is difficult to observe all interesting traits in the association panels representing of American and European diversity (Flint-Garcia et al. 2005; Salvi et al. 2007), as they consist of numerous landrace and tropical lines unadapted to growing conditions in these regions. Therefore, it was necessary to construct an appropriate association panel adapted to local environments in China.

The presence of population structure within an association mapping population can be an obstacle to the application of association mapping as it often generates spurious genotype–phenotype associations (Yu and Buckler 2006; Zhu et al. 2008; Myles et al. 2009). The non-reproducibility of associations between polymorphisms in the *dwarf8* gene and flowering time variation in three association panels validated the importance of population structure estimation prior to association analyses (Thornsberry et al. 2001; Andersen et al. 2005; Camus-Kulandaivelu et al. 2006). To account for population structure in association analysis, two major statistical methods, genome control (Devlin and Roeder 1999; Zheng et al. 2005) and structure association (SA) (Pritchard et al. 2000b), were applied in early studies, both of which used random makers spaced throughout the genome, but incorporated them into statistical analysis in different approaches. Recently, Yu et al. (2006) developed a mixed-linear model (MLM) approach to perform association analysis. The MLM approach, accounting for both population structure (*Q*) and relative kinship (K), can be performed with the TASSEL software package (Bradbury et al. 2007). It is one of the most common methods of association analysis in plants, and has been successfully applied in maize (Yu et al. 2006), wheat (Breseghello and Sorrells 2006), sorghum (Murray et al. 2009), Arabidopsis (Zhao et al. 2007) and potato (Malosetti et al. 2007).

A thorough understanding of genetic diversity, population structure and familial relatedness in a given panel is necessary for successful association studies. Currently, SSRs and SNPs are two main types of molecular markers used to evaluate genetic diversity, population structure and familial kinship of association panels. SSR markers have played a predominant role in the estimation of genetic diversity and population structure in numerous plant species because of the properties of multiple alleles, reproducibility, cost-effectiveness and selective neutrality (Smith et al. 1997). As SNP markers are less informative than SSRs due to their biallelic nature (Rosenberg et al. 2003; Liu et al. 2005), the number of SNPs must be increased to obtain the same information (Hamblin et al. 2007). This is easily possible due to the wide genomic distribution, cost-effective and high-throughput detection systems of SNP markers; these properties have made SNPs a popular choice in linkage analysis and association mapping.

In the current study, a maize association mapping panel developed for studies to unravel the genetic basis of quantitative traits useful to Chinese and similar temperate growing environments is presented. This mapping panel has been characterized with a total of 82 SSR markers, 1,536 SNP markers, and 34 phenotypic traits, to: (1) investigate the genetic and phenotypic diversity; (2) estimate the levels of population structure and assess familial

relatedness; (3) evaluate the effect of population structure on phenotype; and (4) evaluate this panel for association analysis.

Materials and methods

The association panel

A set of 155 diverse lines, mostly selected from temperate germplasm, was assembled to construct an association mapping panel in maize. The panel contained 91 inbred parents of the commercial hybrids most used widely in China (Teng et al. 2004), 35 high-oil lines developed from the major high-oil breeding populations in the world (Song and Chen 2004), 25 inbred lines derived from Chinese landraces, and four high provitamin A lines introduced from the University of Illinois in the United States. Pedigree details are summarized in Supplementary Table 1.

Phenotypic data collection

The association panel was planted in a randomized complete block design with two replications in the following six environments: Changping Agronomy Farm, China Agricultural University, Beijing, in 2005, 2007 and 2008; Shangzhuang Agronomy Farm, China Agricultural University, Beijing, in 2006 and 2008; and Winter Nursery, Sanya Agronomy Farm, China Agricultural University, Hainan Island, in 2007. Each line was grown in a single 3 m row, and rows were 0.67 m apart, with a planting density of 45,000 plants/ha. More than six plants in each row were self-pollinated in all environments except Changping Agronomy Farm in 2008, in which environment all lines were open-pollinated. Pollinated ears were harvested at maturity and allowed to air dry to reach about 13% seed moisture content.

A total of 34 traits were measured in different environments ranging from one to four environments per trait (Table 4). Only nine plants in the middle of each row were used to score plant traits. Traits measured included flowering traits: days to pollen (days from planting to male flowering for 50% of the plants in each row) and days to silk (days from planting to silk emergence for 50% of the plants in each row); plant architecture traits: number of tassel branches, number of plant nodes, number of leaves above the ear, and number of nodes above the ear; ear traits: ear length, ear diameter, cob diameter, number of rows per ear, and cob mass; kernel traits (for which equal amounts of grain from each harvested ear were bulked): 100-kernel weight, kernel length, kernel width, and kernel thickness; oil-related traits: palmitic, stearic, oleic, linoleic, linolenic acid concentration and oil content; carotenoid-related traits: lutein, zeaxanthin, β-cryptoxanthin, α-carotene, β-carotene, total carotenoids, and provitamin A; and tocopherol-related traits: delta-tocopherol, gamma-tocopherol, alpha-tocopherol and total tocopherol. Methods of scoring the oil, carotenoid and tocopherol-related traits were described in previous studies (Yang et al. 2010; Chander et al. 2008a, b).

SSR and SNP genotyping

DNA was extracted by a modified CTAB procedure from bulked leaves from at least six individuals for each line according to Murry and Thompson (1980). A set of 82 SSRs evenly distributed throughout the maize genome was used (Supplementary Table 2). SSRs were amplified via PCR with fluorescently labeled primers in a 20 μl reaction volume containing 40 ng genomic DNA, 10× PCR reaction buffer, 1.5 mM $MgCl_2$, 0.2 mM of each dNTP, 0.2 μM of each 5′-labeled forward and unlabeled reverse primer, and 0.5 unit of *Taq* polymerase (Tiangen, China). Thermal cycling conditions were 95°C for 5 min; 35 cycles of 95°C for 45 s, optimal annealing temperature for 45 s, and 72°C for 1 min; followed by a final extension of 10 min at 72°C. PCR products were size separated on a 3730XL DNA Sequencer equipped with GENESCAN software (ABI, US). Fragment size was recorded by the software GeneMarker V1.6 (SoftGenetics, State College, PA) and manually re-checked.

The details of SNP genotyping were described in a previous study (Yan et al. 2010a). Briefly, a GoldenGate assay (Illumina, San Diego, CA) containing 1,536 SNPs was applied to genotype 155 lines. The SNP genotyping was performed on Illumina BeadStation 500 G (Illumina, San Diego, CA) at Cornell University Life Sciences Core Laboratories Center following the protocols described by Fan et al. (2006). The specifications of the 1,536 SNPs can be found in the studies by Yan et al. (2010a). One thousand four hundred and fourteen SNPs with missing data <20% were used in the subsequent analysis, among which 884 SNPs with minor allelic frequencies greater than 20% were used to evaluate the kinship of this panel.

Genotypic data analyses

Powermarker version 3.25 (Liu and Muse 2005) was used to calculate allele number, gene diversity, group-specific alleles, line-specific alleles, polymorphic information content (PIC) and Nei's genetic distance (Nei 1972). Considering the effects of sample size on estimating genetic diversity, allelic richness was further estimated by a rarefaction method implemented in HP-RARE software (Kalinowski 2005). The significance of different statistics including gene diversity, PIC and allelic richness was assessed using Wilcoxon's paired test across loci.

To investigate population differentiation, an analysis of molecular variance (AMOVA) (Excoffier et al. 1992) was performed and pairwise F statistics (F_{st}) among populations was calculated using Arlequin V3.11 (Excoffier et al. 2005). These analyses mentioned above were performed using 82 SSRs.

The model-based program STRUCTURE 2.2 (Pritchard et al. 2000a; Falush et al. 2003) was used to infer population structure using 82 SSRs. Five independent runs were performed setting the number of populations (k) from 1 to 10, burn in time and MCMC (Markov Chain Monte Carlo) replication number both to 500,000, and a model for admixture and correlated allele frequencies. The k value was determined by LnP(D) in STRUCTURE output and an ad hoc statistic Δk based on the rate of change in LnP(D) between successive k (Evanno et al. 2005). The results of replicate runs from STRUCTURE were integrated by CLUMPP software (Jakobsson and Rosenberg 2007). Lines with membership probabilities $\geq$0.75 were assigned to corresponding clusters; lines with membership probabilities $<$0.75 were assigned to a mixed group. Structure results of individual assignments to corresponding groups were graphically displayed using the DISTRUCT software package (Rosenberg 2004). Groups were further subdivided into subgroups using a similar methodology. The runs most consistent with breeder's knowledge about pedigree were used to assign lines into clusters. Finally, 884 SNPs with minor allele frequencies over 20% were used to calculate the kinship matrix (K) using the SPAGeDi software package (Hardy and Vekemans 2002). All negative values between individuals were set to 0 (Yu et al. 2006).

Statistical analyses of phenotypic data

All statistical analyses of the phenotypic data were carried out using SAS 8.02 (SAS Institute 1999). The trait means across environments and the effects of genotype $\times$ environment (G $\times$ E) on traits measured in greater than one environment were evaluated by PROC GLM using the LSMEANS option. Variance components of genotype, environment, and pooled error containing G $\times$ E with residual effect were estimated using PROC MIXED. These variance components were then used to estimate the broad-sense heritability on a family mean base (Holland et al. 2003). The effects of population structure on all traits were tested based on the means across environments for each trait using PROC GLM. The model statement included one of the two components of the $k = 2$ Q matrix from STRUCTURE analysis.

To further evaluate the phenotypic diversity in this panel, the means of all lines across environments were used to calculate Shannon–Weaver index for each trait. The index was defined by Poole (1974) as

$$H' = \sum_{i=1}^{n} \text{pi lnpi}$$

where n is the number of phenotypic class, and pi is the frequency of the phenotypic classes. The means (M) and standard deviation (SD) in this maize panel were used to subdivide the phenotypic values (x_i) into ten classes ranging from class 1 ($x_i < M - 2\text{SD}$) to class 10 ($x_i > M + 2\text{SD}$), the class interval being 0.5SD (Pecetti et al. 1992).

Power simulations and model comparisons

To assess the power of association mapping achieved by this panel, the genetic power calculator (GPC, Purcell et al. 2003) was performed to simulate power for various population size ($N = 50$, 100, 150, 155, 200) and genetic effect (effect = 0.01, 0.05, 0.10, 0.15, 0.20). Three traits including total tocopherol content, plant height and kernel length, affected by Q and K at different levels, were selected to perform marker-trait associations. Four models were used to evaluate the effects of population structure (Q) and relative kinship (K) on three selected traits for marker-trait associations: the GLM model, similar to simple ANOVA analysis without considering Q and K; the Q model, considering Q; the K model, considering K; the $Q + K$ model, considering both Q and K. The GLM model and the Q model were performed using general linear model (GLM) in TASSEL V2.1; the K model and the $K + Q$ model were performed using MLM in TASSEL V2.1 (Yu et al. 2006; Bradbury et al. 2007). The quantile–quantile plots of estimated $-\log10$ (p) were displayed using the observed p values from SNP-trait associations and the expected P values assuming that no associations happened between markers and any trait.

Results

The 82 SSRs used to measure genetic diversity were polymorphic across all 155 individuals in this maize panel, and a total of 675 alleles were detected with an average of 8.23 alleles per locus (Table 1). Of this total, 203 alleles (30.07%) were restricted to one of the two groups and 142 alleles (21.04%) occurred in a single inbred. Gene diversity, PIC and allelic richness over the whole panel were 0.65, 0.61 and 6.93, respectively (Table 1).

Population structure and relative kinship

The population structure in the panel containing 155 maize inbred lines was calculated using 82 SSRs and a model-based approach of STRUCTURE. Fifty datasets were

Table 1 Diversity related summary statistics for all inbred lines, groups and subgroups, which were classified using STRUCTURE analysis

Items	Overall	P1							P2					Mixed
		P1 overall	Reid	Lancaster	Zi330	ByGy	RySy	P1 mixed	P2 overall	Tang SPT	Tem-tropic I	Landrace	P2 mixed	
Sample size	155	79	12	9	10	12	19	17	44	10	4	24	6	32
Alleles	675	491	195	215	207	224	341	346	495	200	142	423	219	484
Alleles per locus	8.23	5.99	2.38	2.62	2.52	2.73	4.16	4.22	6.04	2.44	1.73	5.16	2.67	5.90
Gene diversity	0.65	0.62	0.37	0.42	0.43	0.43	0.58	0.58	0.60	0.32	0.29	0.60	0.47	0.65
PIC	0.61	0.56	0.32	0.37	0.37	0.38	0.53	0.53	0.55	0.29	0.23	0.55	0.41	0.61
Allelic richness	6.93	5.69	2.38	2.62	2.52	2.73	4.16	4.22	6.03	2.44	1.73	5.16	2.67	5.90
Group-specific alleles	203	69	3	5	6	5	19	13	77	6	2	49	8	57
Group-specific alleles/line	1.31	0.87	0.25	0.56	0.60	0.42	1.00	0.76	1.75	0.60	0.50	2.04	1.33	1.78
Group-specific alleles (%)	30.07	14.05	1.54	2.32	2.90	2.23	5.57	3.76	15.56	3.00	1.41	11.58	3.65	11.78
Line-specific alleles	142	38	1	4	5	2	14	12	53	6	2	38	7	51
Line-specific alleles (%)	21.04	7.74	0.51	1.86	2.42	0.89	4.11	3.47	10.71	3.00	1.41	8.98	3.20	10.54

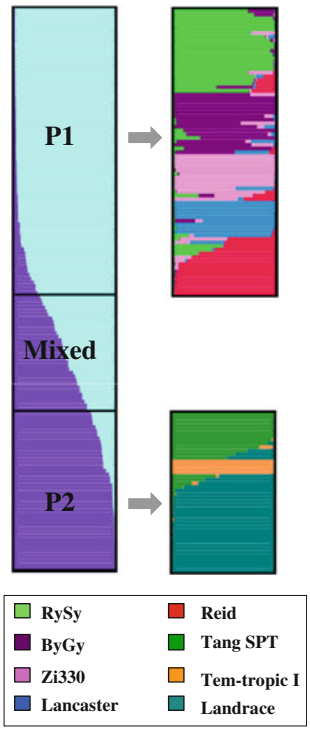

Fig. 1 Model-based cluster membership for 155 lines in two groups and eight subgroups

obtained by setting the number of possible clusters (k) from 1 to 10 with five replications each. The LnP(D) value for each given k increased with the increase of k and the most significant change was observed when k was increased from 1 to 2 (Supplementary Fig. 1a). In addition, there was a sharp peak of Δk at $k = 2$ (Supplementary Fig. 1b). Both parameters suggested the number of clusters was set to 2 (Fig. 1; Table 2, Supplementary Table 3). The first group, called P1, included 79 lines, most of which were high-oil lines and commercial inbred lines with non-flint grain texture. The second group, called P2, contained 44 lines, most of which were related to Chinese landraces with flint grain texture. The remaining 32 lines had membership probabilities lower than 0.75 in any given group and were thus classified into a mixed group.

The two main groups were further subdivided into eight subgroups as suggested by the STRUCTURE analysis. The LnP(D) and Δk values suggested $k = 2$ was the number of subgroups for P1 and P2 groups (Supplementary Fig. 1c, d), but the differentiations at $k = 2$ were not fully consistent with pedigree knowledge. Thus, the pedigree information was used to guide the subdivision of P1 and P2 groups combining with the cluster membership. The P1 group was classified into 5 subgroups and a mixed subgroups including 17 lines. Subgroups were named Reid, containing 12 lines, which were representative of B73; Lancaster, containing 9 lines, which were representative of

Table 2 List of the 155 lines by their model-based groupings

Groups	Subgroups	Number	Inbreds
P1	Reid	12	B73, Ye478, U8112, Zheng32, Hu803, C8605, Tie7922, Ye8001, 832, Ye488, 812, Xun971
	Lancaster	9	Ji846, ZaC546, Mo17, Hai1134, Mo113, 4F1, HTH-17, Ji842, CY72
	Zi330	10	HuangC, Zi330, Zong3, Shen5003, Zheng653, Zong31, LK11, Si446, BEM, A619
	ByGy	12	By804, By815, By843, By4944, Gy220, Gy386, By807, By809, By813, By4960, By855, Gy462
	RySy	19	Ry684, Ry713, Ry732, Sy1090, Sy998, Sy1052, Sy1128, Ry729, Gy1032, Ry697, Sy999, Sy1032, Sy1035, Sy1077, Ye107, 7884-4Ht, K10, Chang3, Nan21-3
	P1-mixed	17	Gy923, Gy1007, By4839, Gy237, Gy246, Gy798, Ry737, Sy1039, Zheng58, Dan340, J4112, Yu374, K14, chuan48-2, K22, 8902, Si434
P2	Tang SPT	10	HZS, Si444, HYS, TYS, H21, Xi502, 5237, WH413, Lx9801, BS16
	Tem-tropic I	4	Qi319, P178, Shen137, Dan599,
	Landrace	24	Tian77, Hai014, SW1611, 5311, S37, Jiao51, TX5, WMR, MN, BNBG, NMJT, QTHHSBTS, 04K5702, NBG, YSBN, BGY, 04K5672, BXZLN, BR2, DSB, D047, B11, SW92E114-15-1, SC55
	P2-mixed	6	Chang7-2, Ji853, 3H-2, 04K5686, HSBN, 303WX
Mixed		32	Sy3073, Ye515, Yan414, Ji53, K12, Dong237, Ji63, Yu87-1, S22, Ye52106, Zheng22, Dong46, BT1, DH02, Dan9046, Hai268, Wu109, Lv28, P138, Qi205, Q1261, 81162, Dan598, Cheng698, E28, H8123, 647, BZN, Hua83-2, HB, CI7, DE3

Mo17; Zi330, containing 10 lines, most of which were related to inbred line Zi330; ByGy, containing 12 lines, which were derived from BHO (Beijing high-oil population) or AIHO (an population developed from IHO C80 × Alexho C23); RySy, containing 19 lines, most of which were derived from RYD or Syn.D.O, both populations introduced from the University of Illinois. The P2 group was classified into three subgroups with a mixed subgroup of six lines; groups included Tang SPT, containing 10 lines, most of which were related to inbred line HZS; Tem-tropic I, containing 4 lines, which were derived from American hybrids; and Landrace, containing 23 local inbred lines derived from Chinese landraces and one high provitamin A line SC55 (Fig. 1; Table 2; Supplementary Table 4).

Relative kinship estimates based on the SNP data showed that 61.9% of the pairwise kinship estimates were equal to 0, and the remaining estimates ranged from 0.05 to 0.5, with a continuously decreasing number of pairs falling in higher estimate categories (Fig. 2). The kinship analysis indicated most lines had no or weak relationship with the other lines in this maize panel, which agreed with various sources of the collected lines.

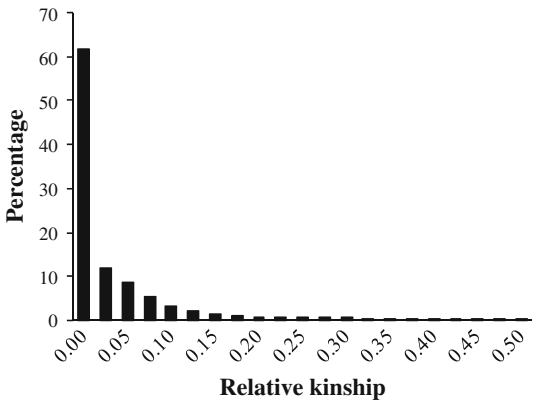

Fig. 2 Distribution of pairwise relative kinship estimates between 155 maize inbred lines. Values are from SPAGeDi estimates using 884 SNPs. For simplicity, only percentages of relative kinship estimates ranging from 0 to 0.50 are shown

Population differences

Comparing P1 and P2, AMOVA results indicated that only 6.1% ($P < 0.001$) of the total genetic variation was partitioned among groups, 92.6% ($P < 0.001$) within groups and 1.3% ($P < 0.001$) within lines. Analyses with subgroups revealed that 17.4% ($P < 0.001$) of the variation occurred among subgroups and 76.3% ($P < 0.001$) within subgroups. Pairwise F_{st} showed low levels of differentiation between P1 and P2 ($F_{st} = 0.10$, $P < 0.001$), but high levels of differentiation between subgroups with F_{st} ranging from 0.13 (RySy with Landrace, $P < 0.001$) to 0.51 (Tang SPT with Tem-tropic I, $P < 0.001$) (Table 3). Among subgroups in the P1 group, the differentiation between Reid and ByGy was the highest ($F_{st} = 0.37$, $P < 0.001$) while the differentiation between RySy and

Table 3 Genetic distances as measured by Nei's minimum distance (top diagonal) and pairwise F_{st} comparisons (bottom diagonal) between maize inbred groups

Groups	Subgroups	P1					P2		
		Reid	Lancaster	Zi330	ByGy	RySy	TangSPT	Tem-tropic I	Landrace
P1	Reid		0.24	0.23	0.26	0.18	0.31	0.36	0.24
	Lancaster	0.34**		0.22	0.22	0.17	0.31	0.34	0.18
	Zi330	0.34**	0.31**		0.21	0.16	0.25	0.33	0.17
	ByGy	0.37**	0.31**	0.31**		0.16	0.29	0.36	0.19
	RySy	0.24**	0.22**	0.21**	0.21**		0.22	0.28	0.10
P2	TangSPT	0.44**	0.42**	0.38**	0.41**	0.29**		0.38	0.17
	Tem-tropic I	0.47**	0.42**	0.43**	0.43**	0.30**	0.51**		0.23
	Landrace	0.29**	0.22**	0.21**	0.23**	0.13**	0.23**	0.25**	

** Significant at $P < 0.01$

Zi330 or ByGy was the lowest ($F_{st} = 0.21$, $P < 0.001$). Among subgroups in the P2 group, the differentiation between Tang SPT and Tem-tropic I was the highest ($F_{st} = 0.51$, $P < 0.001$) while the difference between Landrace and TangSPT was the lowest ($F_{st} = 0.23$, $P < 0.001$). A similar pattern of differentiation among subgroups was observed using Nei's minimum distance, which averaged 0.24 and ranged from 0.10 to 0.38 (Table 3).

Allelic diversity within groups and subgroups

Total diversity within each group was similar, but varied widely within each subgroup (Table 1). The total number of alleles in P1 was 491, with an average of 5.99 alleles per locus. Within this group, subgroups RySy and Reid contained the highest (average of 4.16) and lowest (average of 2.38) number of alleles per locus, respectively. When compared with P1, P2 had fewer lines and slightly lower allelic richness ($z = -2.36$, $P = 0.0182$), but encompassed a similar level of gene diversity ($z = -1.87$, $P = 0.0615$) and PIC ($z = -0.52$, $P = 0.6041$). In total, 495 alleles were detected within P2, and the number of alleles in each P2 subgroup ranged from 142 (Tem-tropic I, 1.73 alleles per locus) to 423 (Landrace, 5.16 alleles per locus). Group-specific alleles, line-specific alleles, gene diversity, PIC and allelic richness showed a similar trend in subgroups (Table 1).

Phenotypic diversity

A total of 86 data sets related to 34 traits were collected for lines in this association panel; each trait was measured in one to four environments, and a broad range of variation was observed for each trait measured (Table 4). As seen by the maximum change fold, tocopherol-related traits had the richest phenotypic diversity, followed by carotenoid-related traits, oil-related traits, ear traits, plant traits and kernel traits, in that order. Total tocopherol content, total carotenoid content and oil content varied from 11.81 to 174.63, 2.38 to 21.14 and 3.15 to 12.61 μg g^{-1}, with an average of 55.93 (±32.09), 10.32 (±4.07) and 5.92 (±2.56) μg g^{-1}, respectively. Alpha-tocopherol content displayed the most striking phenotypic diversity with 145.2 fold maximum change that varied from 0.44 to 63.87 μg g^{-1} with an average of 12.45 (±11.75) μg g^{-1}, while days to silk was the least diverse trait with only a 1.4-fold maximum change that ranged from 69 to 96 days with an average of 77.2 (±4.50) days. The Shannon–Weaver index (H') across all traits averaged 1.80 (±0.20) with a range from 1.44 (alpha-carotene) to 2.06 (plant height). The plant traits had the highest H' values with an average of 2.03 (±0.07). The H' values were similar for oil-related traits (1.67 ± 0.19), carotenoid-related traits (1.67 ± 0.12) and tocopherol-related traits (1.66 ± 0.18), which were slightly lower than that of ear traits (1.88 ± 0.06) and kernel traits (1.83 ± 0.04).

Twenty-seven of 34 traits were measured in multiple environments and replicates, on which G × E effects can be tested. The analysis of variance (ANOVA) revealed that G × E interactions was significant for all 27 traits ($P < 0.01$) except number of plant nodes, linolenic acid concentration, and the content of tocopherol components (Table 4). The broad-sense heritability of all measured traits based on a family mean was relatively high, ranging from 65.3% for 100-kernel weight to 97.8% for linoleic acid concentration (Table 4).

Effect of population structure on phenotype

As illustrated in Table 4, only a small part of the phenotypic variation for any trait was due to the presence of groups within the panel, with an average of 6.4% across all traits. For 26 of the measured traits, the percentage of phenotypic variation explained by population structure did

Table 4 Descriptive statistics, Shannon–Weaver index, family mean-basis heritability and percentage of phenotypic variation explained by population structure for 34 traits scored in up to six environments each

Traits	Min	Max	Mean ± SD	H^a	Env no.[b]	G × E	H^{2c}_m	R^{2d}
Plant traits								
Days to pollen (days)	41.0	77.0	69.1 ± 4.3	1.91	3	**	69.3	1.8
Days to silk (days)	69.0	96.0	77.2 ± 4.5	1.94	1	ND	89.7	0.6
Plant height (cm)	127.6	257.0	186.2 ± 26.6	2.06	4	**	93.7	5.9
Ear height (cm)	20.7	106.6	70.1 ± 15.9	2.05	4	**	94.3	0.0
Number of tassel branch	2.9	23.0	10.1 ± 4.2	2.03	2	**	91.7	0.1
Number of plant node	9.5	16.9	12.9 ± 1.2	2.05	2	ns	90.5	0.7
Number of leaves above ear	4.7	8.4	6.3 ± 0.7	2.14	1	ND	85.2	3.3
Number of nodes above ear	5.1	8.1	6.4 ± 0.6	2.08	2	**	87.4	5.1
Ear traits								
Ear length (cm)	6.9	14.2	10.4 ± 1.5	1.90	1	ND	82.7	0.1
Ear diameter (cm)	3.1	4.9	3.9 ± 0.3	1.97	1	ND	91.5	5.2
Cob diameter (cm)	2.1	3.6	2.7 ± 0.3	1.87	1	ND	93.3	18.9
Number of kernel rows	10.0	23.0	13.9 ± 2.1	1.87	1	ND	93.2	0.0
Cob mass (g)	4.6	20.4	9.3 ± 3.0	1.79	1	ND	88.2	3.6
Kernel traits								
100-kernel weight (g)	13.4	35.4	21.8 ± 3.9	1.78	2	**	65.3	0.5
Kernel length (mm)	7.3	11.1	9.1 ± 0.7	1.88	2	**	72.1	13.3
Kernel width (mm)	6.2	10.2	8.1 ± 0.8	1.82	2	**	89.1	7.1
Kernel thickness (mm)	3.6	5.8	4.8 ± 0.5	1.82	2	**	80.6	2.3
Oil-related traits								
Palmitic acid (μg g^{-1})	0.44	1.63	0.88 ± 0.27	1.74	3	**	96.6	8.7
Stearic acid (μg g^{-1})	0.05	0.36	0.14 ± 0.07	1.64	3	**	94.5	9.1
Oleic acid (μg g^{-1})	0.60	4.77	1.85 ± 1.13	1.54	3	**	97.6	11.0
Linoleic acid (μg g^{-1})	1.57	5.86	2.94 ± 1.13	1.58	3	**	97.8	23.0
Linolenic acid (μg g^{-1})	0.03	0.10	0.06 ± 0.01	2.01	3	ns	73	19.5
Oil content (μg g^{-1})	3.15	12.61	5.92 ± 2.56	1.48	3	**	97.7	16.3
Carotenoid-related traits								
Lutein (μg g^{-1})	0.76	15.68	5.39 ± 2.73	1.75	4	**	92.9	2.8
Zeaxanthin (μg g^{-1})	0.88	8.43	3.41 ± 2.03	1.72	4	**	94.8	2.5
Beta-cryptoxanthin (μg g^{-1})	0.09	2.87	0.73 ± 0.57	1.60	4	**	88.4	9.9
Alfa-carotene (μg g^{-1})	0.01	0.43	0.11 ± 0.10	1.44	4	**	72.6	0.7
Beta-carotene (μg g^{-1})	0.13	1.98	0.68 ± 0.37	1.68	4	**	75.2	1.6
Total carotenoid (μg g^{-1})	2.38	21.14	10.32 ± 4.07	1.81	4	**	91.2	0.0
Pro-vitamin A (μg g^{-1})	0.22	3.43	1.10 ± 0.60	1.69	4	**	79.8	4.7
Tocopherol-related traits								
Delta-tocopherol (μg g^{-1})	1.74	7.61	2.39 ± 0.81	1.46	2	ns	94.4	5.9
Gamma-tocopherol (μg g^{-1})	7.33	122.51	41.18 ± 24.92	1.84	2	ns	95.2	14.9
Alfa-tocopherol (μg g^{-1})	0.44	63.87	12.45 ± 11.75	1.57	2	ns	77.8	5.1
Total tocopherol (μg g^{-1})	11.81	174.63	55.93 ± 32.09	1.78	2	ns	92.3	14.8

ns nonsignificant at $P < 0.05$, *ND* not determined (these traits were not scored in enough environments to test genotype × environment effects)

[a] Shannon–Weaver index

[b] The number of environments in which each trait was measured

[c] Broad-sense heritability on family mean basis

[d] Percentage of phenotypic variation explained by population structure

** Significant at $P < 0.01$

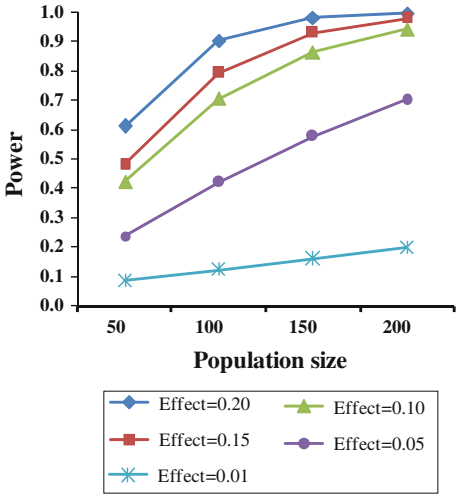

Fig. 3 Power simulations of association panels with various population size for the genetic factors with different explained phenotypic variation

not exceed 10%. For ear height, number of kernel rows and beta-carotene, almost no effect due to population structure was observed. Population structure accounted for the highest percentage (23%) of phenotypic variation for linoleic acid concentration. The greatest influence of population structure was observed for oil-related traits, with a range of explained phenotypic variation from 8.7 to 23.0%, while plant traits were least affected with a range from 0 to 5.9%.

Evaluation of association panel

Figure 3 shows the power of association panels with various numbers of individuals for the genetic factors accounting for different phenotypic variations. Greater power to detect marker-trait associations was observed as the population size increased and as the genetic factor explained more of the phenotypic variations. When the genetic factors accounted for <10% of phenotypic variations, the power to detect an association was relatively low and increased sharply with the increase in population size in an association panel with ≤200 individuals. In contrast, the power was quite high when the population size reached over 150. For this panel with 155 individuals, over 87.6% of the genetic factors explaining ≥10% of phenotypic variations were captured, 59.2% for 5%, and 16.5% for 1%.

Combined with 1,414 SNPs over the whole genome and three typical traits, total tocopherol content in maize kernel, plant height and kernel length, associations were performed to evaluate the effects of Q and K for controlling false associations. For any trait, the P value from the GLM model greatly deviated from the expected P value, followed by the Q model, while the P values from the K model and the $K + Q$ model were close to the expected P value (Fig. 4). However, the degree of the effects of Q and K was different for different traits. At $P < 0.01$, the Q model performed well for plant height and kernel length with significant SNPs reducing 4.89 and 4.90%, respectively; the correction of false positive was well conducted by the K model for all three traits, especially for total tocopherol content with significant SNPs reducing 19.28% (Table 5). Using the $K + Q$ model, 1.15,1.15, 1.66% SNPs were significantly associated with total tocopherol content in maize kernel, plant height and kernel length at $P < 0.01$, and 0.29, 0.29, 0.43% at $P < 0.001$, respectively.

Discussion

Genetic diversity in the maize panel

A suitable association mapping panel should encompass as much phenotypic and molecular diversity as can be reliably measured in a common environment (Flint-Garcia et al. 2005). An average of 8.23 alleles per locus over 82 SSRs was observed in this association panel containing 155 inbred lines. The value was significantly lower than that of 21.7 over 94 SSR loci in 260 US inbreds (Liu et al. 2003), slightly lower than 9.4 over 145 SSR loci in a mini core set with 95 temperate inbreds mostly developed in China (Wang et al. 2008), but exceeded most reported values in other diversity studies of maize inbred lines (Supplementary Table 5; Taramino and Tingey 1996; Senior et al. 1998; Lu and Bernardo 2001; Matsuoka et al. 2002; Labate et al. 2003; Clerc et al. 2005; Xia et al. 2005; Reif et al. 2005, 2006; Yu et al. 2007; Xie et al. 2008). The difference in SSR allelic richness can be explained by the number of maize lines analyzed and the choice of maize germplasm, the number of SSR loci and the SSR repeat type (Vigouroux et al. 2002; Liu et al. 2003). A higher number of lines in the sample leads to a more diverse range of germplasm simply by sampling, and a larger number of loci (and in particular, the use of dinucleotide repeat SSRs rather than tri- or higher) will lead to a higher number of alleles and thus a higher apparent level of genetic diversity. The genetic diversity across all lines in this association panel was 0.65, which was higher or equal to all reported values except those of Taramino and Tingey (1996), and Liu et al. (2003) (Supplementary Table 5).

One excellent association mapping panel is reported by Flint-Garcia et al. (2005), and has been used for association studies of several traits in maize to date (Remington et al. 2001; Thornsberry et al. 2001; Wilson et al. 2004; Szalma et al. 2005; Harjes et al. 2008; Pressoir et al. 2009; Yan et al. 2010b). This panel has been shown to possess abundant genetic diversity (Liu et al. 2003) and will be

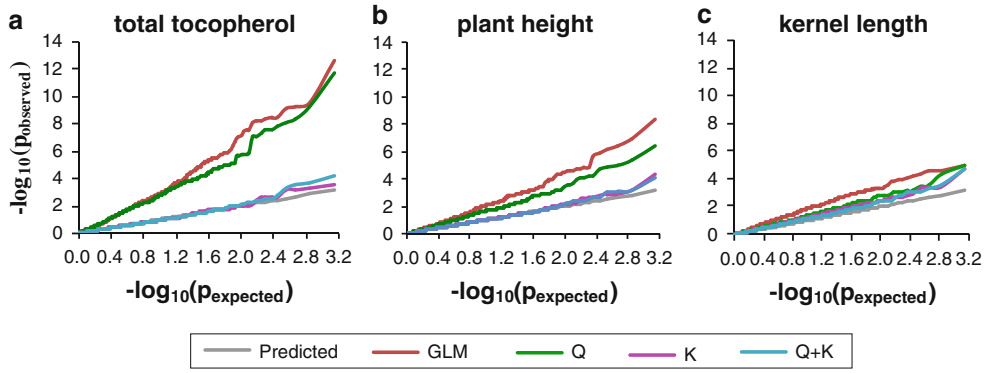

Fig. 4 Quantile–quantile plots of estimated −log10 (p) from association analysis using four methods in three traits: **a** total tocopherol, **b** plant height, **c** kernel length. The *black line* is the expected line under the null distribution. *Red line* represents the observed *P* values using GLM; *green line* represents the observed *P* values using GLM with *Q*; *pink line* represents the observed *P* values using MLM model with *K*; *blue line* represents the observed *P* values using MLM model with *Q* and *K*

Table 5 The percentage of significant SNPs associated with three traits using four statistic models

Traits	GLM	*Q*	*K*	*Q* + *K*
Total tocopherol content	20.29/9.64	19.21/8.42	1.01/0.29	1.15/0.29
Plant height	10.49/3.45	5.6/1.44	1.22/0.14	1.15/0.29
Kernel length	8.07/2.02	3.17/0.65	2.24/0.29	1.66/0.43

The value before the slash shows the percentage of significant SNPs at $P < 0.01$ and the value after the slash shows the percentage of significant SNPs at $P < 0.001$

referred to as P_{US} here. To compare the genetic diversity between P_{US} and our panel (referred to as P_{CAU} in this study), 47 SSR loci scored in both panels were reanalyzed jointly, using six inbreds common to both panels (B73, Mo17, SC55, A619, DE3 and CI7) as controls. A total of 805 alleles were detected with an average of 17.1 alleles per locus over the 47 loci across all 415 inbred lines. The allelic richness in P_{CAU} was lower than that in P_{US} ($z = -5.36$, $P < 0.001$), but similar to that in the temperate lines from the P_{US} panel analyzed separately from the tropical ($z = -0.03$, $P = 0.9747$) (Supplementary Table 6). Of 716 alleles detected in the P_{US} panel and 544 alleles in the temperate subset of the P_{US} panel, the P_{CAU} panel captured 428 (59.8%) and 376 (69.1%) alleles, respectively. Thus, P_{CAU} contained about 60% of the genetic diversity of present in P_{US}, which was created to capture a large proportion of the alleles in all cultivated maize (Flint-Garcia et al. 2005). As most lines from P_{CAU} were popular inbred lines for hybrid production in China (Teng et al. 2004), the results mentioned above indicate that the genetic diversity of maize breeding resources in China is somewhat narrow and could be broadened by introducing foreign germplasm, especially tropical germplasm, which contains many 'unique' alleles not present in Chinese maize breeding resources. However, 50 unique line-specific alleles were detected in P_{CAU}, which is more than were seen in the temperate subset of P_{US} (43). This indicates variation contained in Chinese maize germplasm that is not presented in the diverse P_{US} and may be due to the origins of the Chinese germplasm, sampling effects when creating both panels, or due to the effect of artificial selection or genetic drift. These line-specific alleles offer a source of genetic variation for further crop improvement and for dissecting the genetic basis of quantitative traits in maize.

Consequences of genetic relatedness in the maize panel

Detailed knowledge of genetic relatedness among individuals in an association panel is a key factor to avoid spurious associations. Population structure (*Q* matrix), estimated using STRUCTURE (Pritchard et al. 2000a; Falush et al. 2003) and expressed as membership probabilities, is one way to correct spurious associations due to genetic relatedness. It is often difficult to estimate the true number of population (*k*). Generally, *k* is taken to be the value with the highest estimated LnP(D) value returned by STRUCTURE (Pritchard et al. 2000a). However, in real situations, few data sets confirm precisely to the STRUCTURE model, and the LnP(D) value keeps increasing when *k* reaches the true number. Evanno et al. (2005) suggested an ad hoc method, Δk, the second-order rate of change of the likelihood function with respect to *k* performed well in indicating the real number. In this study, the Δk values indicated splitting the maize panel into two groups was the most biologically meaningful population structure because of the rapid rate of change in LnP(D) values between successive *k*. These two groups were consistent with groups known to have diverged in the very distant past (the flint and non-flint groups). However, the mode of Δk at the true *k* was absent for further subdividing inbreds in each of

the two groups, P1 and P2, into subgroups using STRUCTURE. This may be due to the reduced sample size and weak genetic differentiation within groups (Evanno et al. 2005; Waples and Gaggiotti 2006). Therefore, the pedigree information was used for subgroup subdivision. The P1 group was subdivided into five subgroups: Reid, Lancaster, Zi330, ByGy and RySy. Subgroups Reid, Lancaster and Zi330 reflect known heterotic groups (Teng et al. 2004) while subgroups ByGy and RySy split high-oil lines, which agreed that most lines in ByGy were close to Lancaster heterotic group and that most lines in RySy were close to Reid heterotic group (Song Tongming, personal communications). In P1, pairwise F_{st} values indicate that RySy was genetically more similar to the other subgroups; that Reid was strongly differentiated from subgroups Lancaster, Zi330 and ByGy; and that subgroups Lancaster, Zi330 and ByGy were more similar than other subgroups (excluding RySy). Based on the SSR markers, lines from P2 were organized into two known heterotic groups (Teng et al. 2004), Tang SPT and Tem-tropic I, and one subgroup consisting of landraces. Pairwise F_{st} values indicated that Tang SPT and Tem-tropic I were highly unrelated, and that the subgroup Landrace were equally diverged from both. Six of the eight subgroups were well agreed with the previous studies on Chinese maize inbred lines that separated Chinese lines into six groups (Xie et al. 2008) or four groups (Wang et al. 2008). A few lines in some groups were not consistent with pedigree information perfectly, which may be due to the marker density, power, lines purity and other unknown reasons.

The effect of population structure on interesting traits, determined by the percentage of trait variation explained by population structure, is one indication of the power of an association mapping panel to detect the effects of individual genes using structured association analysis. Population structure based on the two groups in this association panel accounted for an average of 6.4% of the phenotypic variation across all 34 traits, indicating that population structure is only a minor factor contributing to phenotypic variation in this panel. However, the effects varied depending on the trait, and there was a marked influence on population structure affecting cob diameter, linoleic acid, linolenic acid, oil content, gamma-tocopherol and total tocopherol (R^2 ranged from 14.9 to 23.0% for these traits). On the other hand, virtually no or weak correlations were observed between population structure and ear height, number of tassel branch, ear length, number of kernel rows and total carotenoid. The significant responses to population structure for measured traits may be explained by a selection effect, especially for oil-related traits and tocopherol-related traits, as high-oil maize is developed by intense artificial selection and tocopherols are highly correlated to oil content (Lambert 2001).

As expected, some traits were differently affected by population structure in different association panels. This was one of the underlying motivations for the selection of a suitable association panel and showed the necessity of validating trait-associated genes in multiple panels. The adaption trait, flowering time, was greatly affected by population structure in American and European association panels ($R^2 > 32\%$) (Flint-Garcia et al. 2005; Camus-Kulandaivelu et al. 2006), but very weakly affected in this association panel ($R^2 < 2\%$). The extreme difference was caused by the germplasm in each association panel, as the American and European association panels contained both temperate and tropical germplasm while the maize panel reported here mainly contained temperate germplasm. Therefore, this maize panel is more suitable for performing association analysis between genotypes and flowering time. On the other hand, because population structure in different association panels accounted for a low percentage of phenotypic variation for cob mass, 100-kernel weight and kernel thickness (Flint-Garcia et al. 2005; Table 4), these traits are weakly correlated with population structure and could be investigated equally well in different association panels.

As mentioned above, population structure plays an important role in association analysis. However, spurious associations cannot be controlled completely by population structure as the Q matrix only gives a rough dissection of population differentiation. Therefore, Yu et al. (2006) suggested incorporating the pairwise kinship (K matrix) into a mixed model to correct for relatedness in association mapping. The K matrix is generally superior to association models using only Q matrix (Yu et al. 2006; Myles et al. 2009). In this panel, SNP-trait associations were performed for three traits using simple variance analysis, the Q model, the K model, and the $Q + K$ model. We found that the K model performed better than the Q model, but similarly to the $K + Q$ model. In addition, simulations performed by Zhu and Yu (2009) indicated that nMDS (nonmetric multidimensional scaling) combining the K matrix were better for correcting false associations than simple regression analysis, especially for samples with complex genetic relatedness.

Application and weakness of the maize panel

The maize panel presented here exhibits considerable natural variation for most traits, especially for levels and compositions of carotenoid, tocopherol and oil in the kernels. All lines in the panel display good adaptation to most growing conditions in China, which will allow field experiments to be conducted for multiple phenotypic traits by different research groups throughout China. Unlike complex quantitative traits, such as yield, the

heritability of kernel carotenoid, tocopherol and oil are extremely high. A few major QTL for these traits were identified to determine their genetic basis (Yang et al. 2010; Chander et al. 2008a, b). Furthermore, the metabolic pathways for these three traits are relatively simple and most of the genes encoding key enzymes are well known in *Arabidopsis* (Thelen and Ohlrogge 2002; Beisson et al. 2003; DellaPenna and Pogson 2006; DellaPenna and Last 2006; Matthews and Wurtzel 2007). All these characteristics, together with the knowledge of population structure and relative kinship presented here, will allow this panel to be used in mining favorable alleles of genes by pathway-driven association mapping. For example, two major genes affecting provitamin A content in maize kernel [*lycopene epsilon cyclase* (*lcyE*); *β-carotene hydroxylase 1* (*crtRB1*)] were also validated by association mapping using this panel (Yan et al. 2010b). Currently, several candidate genes involved in carotenoid, tocopherol and oil metabolism are being characterized to mine favorable alleles related to corresponding traits. In addition, association mapping of some complex quantitative traits, such as plant morphology traits, ear traits and kernel traits, can be done in this panel, where they were only weakly affected by population structure (Table 4). This has been confirmed in a successful association study of *ZmGS3* and several kernel traits (Li et al. 2010). Although the LD decay of this panel was not estimated at the genome-wide level, the results from several candidate genes showed that the LD decay is very fast in this panel ranging from a few hundreds to thousands base pairs (Li et al. 2010; Zhiyuan Fu et al. unpublished data).

However, there are two main disadvantages of this panel used for association mapping. First, the sample size of this panel was small and it had relatively low power of association mapping. Simulation studies by genetic power calculate (Purcell et al. 2003) demonstrated that this panel consisting of 155 diversity lines could only capture 59.2% of the quantitative genes explaining 5% of phenotypic variation and 87.6% for 10%. It shows that this panel is suitable for the traits controlled by major QTL and needs to be extended for further investigating the genetic basis of interesting traits controlled by genes with moderate or even minor effects. Secondly, structured association analysis using this panel would increase false negatives and reduce the power for traits strongly correlated with population structure, such as oil-related traits and tocopherol-related traits. Therefore, we considered combining association analysis with linkage analysis to identify the true genetic variants for these traits as the covariance between genotypes and phenotypes can be broken up by generating controlled cross (Myles et al. 2009). In addition, increasing the population size, combined careful genotype selection for population structure estimate, may render this panel useful in identifying the genetic factors associated with the traits, which is highly correlated with the population structure.

The genetic architecture of complex quantitative traits is generally studied with the final objective of improving crop performance. Functional markers are developed and applied in molecular breeding programs after favorable alleles are identified by linkage analysis or association mapping (Andersen and Lübberstedt 2003). Six functional markers derived from *lcyE* (Harjes et al. 2008) and *crtRB1* (Yan et al. 2010b) are being used to improve the levels of provitamin A in maize breeding at several institutes including this laboratory with excellent results (data not shown). All high pro-vitamin 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. Breseghello and Sorrells (2006) pointed out that including individuals in association panels from current breeding programs offers the advantage of easy incorporation into future breeding programs. Many (91) of the lines in this maize panel are derived from advanced breeding programs in China, and will certainly prove useful in future breeding programs as well, especially as more information on favorable alleles contained in each line for a wide range of agronomic and morphological traits are identified.

Acknowledgments Helpful comments on the early manuscript from Drs. Michael Gore and Maruthi Prasanna Boddupalli are appreciated. The authors gratefully thank the editor Dr. Jianming Yu and three anonymous reviewers for their valuable suggestions. This research was supported by National Hi-Tech Research and Development Program of China (2006AA100103, 2006AA10Z183), National Basic Research and Development Program of China (2007CB10900).

References

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

Andersen JR, Schrag T, Melchinger AE, Zein I, Lübberstedt T (2005) Validation of *Dwarf8* polymorphisms associated with flowering time in elite European inbred lines of maize (*Zea mays* L.). Theor Appl Genet 111:206–217

Andersen JR, Zein I, Wenzel G, Darnhofer B, Eder J, Ouzunova M, Lübberstedt T (2008) Characterization of phenylpropanoid pathway genes within European maize (*Zea mays* L.) inbreds. BMC Plant Bio 8:2

Beisson F, Koo AJK, Ruuska S, Schwender J, Pollard M, Thelen JJ, Paddock T, Salas JJ, Savage L, Milcamps A, Mhaske VB, Cho Y, Ohlrogge JB (2003) Arabidopsis genes involved in acyl lipid metabolism 2003. A census of the candidates, a study of the distribution of expressed sequence tags in organs, and a web-based database. Plant Physiol 132:681–697

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

Bradbury PJ, Zhang ZW, Kroon DE, Casstevens RM, Ramdoss Y, Buckler ES (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23(19):2633–2635

Breseghello F, Sorrells ME (2006) Association analysis as a strategy for improvement of quantitative traits in plants. Crop Sci 46:1323–1330

Buckler ES, Gore M (2007) An Arabidopsis haplotype map takes root. Nat Genet 39:1056–1057

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

Casa AM, Pressoir G, Brown PJ, Mitchell SE, Rooney WL, Tuinstra MR, Franks CD, Kresovich S (2008) Community resources and strategies for association mapping in Sorghum. Crop Sci 48:30–40

Chander S, Guo YQ, Yang XH, Zhang J, Lu XQ, Yan JB, Song TM, Rocheford TR, Li JS (2008a) Using molecular markers to identify two major loci controlling carotenoid contents in maize grain. Theor Appl Genet 116:223–233

Chander S, Guo YQ, Yang XH, Yan JB, Zhang YR, Song TM, Li JS (2008b) Genetic dissection of tocopherol content and composition in maize grain using quantitative trait loci analysis and the candidate gene approach. Mol Breed 22:353–365

Clerc VL, Bazante F, Baril C, Guiard J, Zhang D (2005) Assessing temporal changes in genetic diversity of maize varieties using microsatellite markers. Theor Appl Genet 110:294–302

DellaPenna D, Last RL (2006) Progress in the dissection and manipulation of plant vitamin E biosynthesis. Physiol Plant 126:356–368

DellaPenna D, Pogson BJ (2006) Vitamin synthesis in plants: tocopherols and carotenoids. Annu Rev Plant Biol 57:711–738

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

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. Methods Enzymol 410:57–73

Flint-Garcia SA, Thornsberry JM, Buckler ES (2003) Structure of linkage disequilibrium in plants. Annu Rev Plant Biol 54:357–374

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

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

Holland JB, Nyquist WE, Cervantes-Martínez CT (2003) Estimating and interpreting heritability for plant breeding: an update. Plant Breed Rev 22:9–111

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

Kalinowski ST (2005) HP-RARE 1.0: a computer program for performing rarefaction on measures of allelic richness. Mol Ecol Notes 5:187–189

Labate JA, Lamkey KR, Mitchell SE, Kresovich S, Sullivan H, Smith JSC (2003) Molecular and historical aspects of corn belt dent diversity. Crop Sci 43:80–91

Lambert RJ (2001) High-oil corn hybrids. In: Hallau AR (ed) Special corn. E. CRC Press Inc, Boca Raton, pp 131–153

Li Q, Yang XH, Bai GH, Warburton ML, Mahuku G, Gore M, Dai JR, Li JS, Yan JB (2010) Cloning and 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

Liu N, Chen L, Wang S, Oh C, Zhao H (2005) Comparison of single nucleotide polymorphisms and microsatellites in inference of population structure. BMC Genet 6(Suppl 1):S26

Lu H, Bernardo R (2001) Molecular diversity among current and historical maize inbreds. Theor Appl Genet 103:613–617

Maccaferri M, Sanguineti MC, Natoli V, Ortega JLA, Salem MB, Bort J, Chenenaoui C, Ambrogio DE, Moral LGD, Montis AD, El-Ahmed A, Maalouf F, Machlab H, Moragues M, Motawaj J, Nachit M, Nserallah N, Ouabbou H, Royo C, Tuberosa R (2006) A panel of elite accessions of durum wheat (*Triticum durum* Desf.) suitable for association mapping studies. Plant Genet Resour 4:79–85

Malosetti M, van der Linden CG, Vosman B, van Eeuwijk FA (2007) A mixed-model approach to association mapping using pedigree information with an illustration of resistance to phytophthora infestans in potato. Genetics 175:879–889

Matsuoka Y, Mitchell SE, Kresovich S, Goodman M, Doebley J (2002) Microsatellites in Zea—variability, patterns of mutations, and use for evolutionary studies. Theor Appl Genet 104:436–450

Matthews PD, Wurtzel ET (2007) In: Socaciu C (ed) Biotechnology of food colorant production in food colorants: chemical and functional properties. CRC Press, Boca Raton, pp 347–398

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 SC, Rooney WL, Hamblin MT, Mitchell SE, Kresovich S (2009) Sweet sorghum genetic diversity and association mapping for brix and height. Plant Genome 2:48–62

Murry 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 (www.plantcell.org/cgi/doi/10.1105/tpc.109.068437)

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

Pecetti L, Annicchiarico P, Damania AB (1992) Biodiversity in a germplasm collection of durum wheat. Euphytica 60:229–238

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

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

Reif JC, Hamrit S, Heckenberger M, Schipprack W, Maurer HP, Bohn M, Melchinger AE (2005) Genetic structure and diversity of European flint maize populations determined with SSR analyses of individuals and bulks. Theor Appl Genet 111:906–913

Reif JC, Warburton ML, Xia XC, Hoisington DA, Crossa J, Taba S, Muminovic J, Bohn M, Frisch M, Melchinger AE (2006) Grouping of accessions of Mexican races of maize revisited with SSR markers. Theor Appl Genet 113:177–185

Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doebley J, Kresovich S, Goodman MM, Buckler ES (2001) Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc Natl Acad Sci USA 98:11479–11484

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

Rosenberg NA, Li LM, Ward R, Pritchard JK (2003) Informativeness of genetic markers for inference of ancestry. Am J Hum Genet 73:1402–1422

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

Senior ML, Murphy JP, Goodman MM, Stuber CW (1998) Utility of SSRs for determining genetic similarities an relationships in maize using an agarose gel system. Crop Sci 38:1088–1098

Smith JSC, Chin ECL, Shu H, Smith S, Wall SJ, Senior ML, Mitchell SE, Kresovich S, Ziegle J (1997) An evaluation of the utility of SSR loci as molecular markers in maize (*Zea mays* L.): comparisons with data from RFLPS and pedigree. Theor Appl Genet 95:163–173

Song TM, Chen SJ (2004) Long term selection for oil concentration in five maize populations. Maydica 49:9–14

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

Taramino G, Tingey S (1996) Simple sequence repeats for germplasm analysis and mapping in maize. Genome 39:277–287

Tenaillon MI, Sawkins MC, Long AD, Gaut RL, Doebley JF, Gaut BS (2001) Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). Proc Natl Acad Sci USA 98:9161–9916

Teng WT, Can JS, Chen YH, Liu XH, Jing XQ, Zhang FJ, Li JS (2004) Analysis of maize heterotic groups and patterns during past decade in China. Sci Agric Sin 37:1804–1811

Thelen JJ, Ohlrogge JB (2002) Metabolic engineering of fatty acid biosynthesis in plants. Metab Eng 4:12–21

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

Vigouroux Y, Jaqueth JS, Matsuoka Y, Smith OS, Beavis WD, Smith JSC, Doebley J (2002) Rate and pattern of mutation at microsatellite loci in maize. Mol Biol Evol 19:1251–1260

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

Waples RS, Gaggiotti O (2006) What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. Mol Ecol 15:1419–1439

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

Xia XC, Reif JC, Melchinger AE, Frisch M, Hoisington DA, Beck D, Pixley K, Warburton ML (2005) Genetic Diversity among CIMMYT maize inbred lines investigated with SSR markers: II. subtropical, tropical midaltitude, and highland maize inbred lines and their relationships with elite US and European maize. Crop Sci 45:2573–2582

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 M, Buckler ES, McMullen MD, Crouch J (2009) Genetic characterization of a global maize collection using SNP markers. PLoS ONE 4:e8451

Yan JB, Yang XH, Hector S, Shah T, Li JS, Warburton M, Zhou Y, Jonathan C, Xu YB (2010a) High-throughput SNP genotyping with the GoldenGate assay in maize. Mol Breed 25:441–451

Yan JB, Kandianis CB, Harjes CE, Bai L, Kim E, 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 T, Buckler ES, Warburton ML, Rocheford T (2010b) Rare genetic variation at *zea mays crtRB1* increases β-carotene in maize grain. Nat Genet. doi:10.1038/ng.551

Yang XH, Guo YQ, Yan JB, Zhang J, Song TM, Rocheford T, Li JS (2010) Major and minor QTL and epistasis contribute to fatty acid composition and oil content in high-oil maize. Theor Appl Genet 120:665–678

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 YT, Wang RH, Shi YS, Song YC, Wang TY, Li Y (2007) Genetic diversity and structure of the core collection for maize lines in China. Maydica 52:181–194

Yu J, Holland JB, McMullen MD, Buckler ES (2008) Genetic design and statistical power of nested association mapping in maize. Genetics 178:539–551

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