Mol Breeding
DOI 10.1007/s11032-011-9696-1

The strategy and potential utilization of temperate germplasm for tropical germplasm improvement: a case study of maize (*Zea mays* L.)

Weiwei Wen · Tingting Guo · Victor H. Chavez Tovar · Huihui Li · Jianbing Yan · Suketoshi Taba

Received: 21 January 2011 / Accepted: 20 December 2011

Abstract The organization of maize (*Zea mays* L.) germplasm into genetically divergent heterotic groups is the foundation of a successful hybrid maize breeding program. In this study, 94 CIMMYT maize lines (CMLs) and 54 United States germplasm enhancement of maize (GEM) lines were assembled and characterized using 1,266 single nucleotide polymorphisms (SNPs) with high quality. Based on principal component analysis (PCA), the GEM lines and CMLs were clearly separated. In the GEM lines, there were two groups classified by PCA corresponding to the heterotic groups "stiff stalk" and "non-stiff stalk". CMLs did not form obvious subgroups by PCA. The allelic frequency of each SNP differed in GEM lines and CMLs. In total, 3.6% alleles (46/1,266) of CMLs are absent in GEM lines and 4.4% alleles (56/1,266) of GEM lines are absent in CMLs. The performance of F1 plants ($n = 654$) produced by crossing between different groups based on pedigree information was evaluated at the breeding nurseries of two CIMMYT stations. Genomic estimated phenotypic values of plant height and days to anthesis for a testing set of 45 F1 crosses were predicted based on the training data of 600 F1 crosses using a best linear unbiased prediction method. The prediction accuracy benefitted from the adoption of the markers associated with quantitative trait loci for both traits; however, it does not necessarily increase with an increase in marker density. It is suggested that genomic selection combined with association analysis could improve prediction efficiency and reduce cost. For hybrid maize breeding in the tropics, incorporating GEM lines which have unique alleles and clear heterotic

Weiwei Wen and Tingting Guo contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s11032-011-9696-1) contains supplementary material, which is available to authorized users.

W. Wen (✉) · V. H. C. Tovar · S. Taba (✉)
International Maize and Wheat Improvement Center (CIMMYT), Apartado Postal 6-640, 06600 Mexico, D.F., Mexico
e-mail: wenweiwei1982@gmail.com

S. Taba
e-mail: s.taba@cgiar.org

W. Wen · J. Yan
National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan 430070, China

T. Guo
National Maize Improvement Center, China Agricultural University, Beijing 100193, China

H. Li
Institute of Crop Science, The National Key Facility for Crop Gene Resources and Genetic Improvement, Chinese Academy of Agricultural Sciences, No. 12 Zhongguancun South Street, Beijing 100081, China

Published online: 07 January 2012

patterns into tropically adapted lines could be beneficial for enhancing heterosis in grain yields.

Keywords Heterotic groups · Association analysis · Genomic selection · SNP

Abbreviations

AF	Agua Fria
ANOVA	Analysis of variance
BLUP	Best linear unbiased prediction
CIMMYT	International Maize and Wheat Improvement Center
CML	CIMMYT maize line
crtRB1	β-Carotene hydroxylase
DA	Days to anthesis
GCA	General combining ability
GEM	Germplasm enhancement of maize
GWAS	Genome-wide association study
ISU	Iowa State University
LAMP	Latin American Maize Project
MAF	Minor allele frequency
MAS	Marker-assisted selection
NCU	North Carolina University
NSS	Non-stiff stalk
OPA	Oligo pool assay
PCA	Principal component analysis
SS	Stiff stalk
TL	Tlaltizapán

Introduction

Maize (*Zea mays* L.) is one of the most important staple food crops across the world as well as being a main feed and energy crop for global livestock production and the emerging biofuel industry.

The International Maize and Wheat Improvement Center (CIMMYT) has developed and released CIMMYT maize lines (CMLs) since 1984. The CMLs were initially developed from 35 broad-based populations and pools with mixed germplasm origins. They are carefully selected with good general combining ability (GCA) and a significant number of value-adding traits such as drought tolerance, nitrogen use efficiency, acid soil tolerance, and resistance to disease and insect pests (http://www.cimmyt.org/ru/component/content/article/459-international-maize-improvement-network-imin/434-cimmyt-maize-inbred-lines-cml). In many instances they are used as parental lines for the hybrids in one or more maize mega-environments. To date, the total number of CMLs held in trust in the CIMMYT genebank is 530 (S. Taba, unpublished data) and more lines will be included from the CIMMYT maize breeding program. The development of elite maize lines and populations as global public goods will continue at CIMMYT, using new genetic variations of maize genetic resources.

The germplasm enhancement of maize (GEM) project is developing enhanced lines through the introduction and incorporation of novel and useful germplasm gathered from around the globe (http://www.public.iastate.edu/~usda-gem/). It has used some of the elite germplasm of the Latin American Maize Project (LAMP) identified as a source of new genetic diversity for widening the genetic base of United States maize hybrids. The LAMP project involved the cooperative efforts of 12 countries (Argentina, Bolivia, Brazil, Colombia, Chile, Guatemala, Mexico, Paraguay, Peru, United States, Uruguay, and Venezuela) and evaluated 12,000 accessions (Salhuana et al. 1991). The GEM project has developed breeding crosses between selected germplasm accessions from the LAMP and other exotic germplasm sources, such as proprietary lines of United States hybrid companies who are members of the GEM project. GEM breeding crosses are grouped into "stiff stalk" (SS) and "non-stiff stalk" (NSS) heterotic patterns (Salhuana and Sevilla 1995; Salhuana et al. 1998). Enhanced breeding lines used in this study by GEM contained 50 or 75% temperate elite germplasm and 25 or 50% exotic tropical germplasm in the respective heterotic patterns of SS and NSS. They were released from the GEM project to the North Central Plant Introduction Station, Ames, Iowa, USA.

Based on previous studies using both simple sequence repeat (SSR) and single nucleotide polymorphism (SNP) data, use of tropical and subtropical germplasm for temperate maize breeding was suggested for broadening the genetic base of commercial hybrid breeding (Liu et al. 2003; Yan et al. 2009; Ortiz et al. 2010). On the other hand, use of temperate maize germplasm in tropical maize breeding has been difficult due to its lack of adaptation and lack of good documentation (Goodman 1999). Some favorable alleles may be unique to temperate germplasm. For example, the most favorable allele of the gene encoding β-carotene hydroxylase (crtRB1), associated with β-carotene concentration and conversion in

maize kernels, was only detected in temperate germplasm (Yan et al. 2010a). Experience of tropical maize breeders on temperate germplasm often shows that heterosis in grain yields is enhanced when crossing with tropical germplasm. There should be favorably unique alleles or genomic regions in temperate maize germplasm that can be useful in a tropical maize improvement. However, intolerance to some insects and diseases in the tropics and poor grain quality and adaptation of temperate germplasm suggest that a long-term breeding program is required (S. Taba, unpublished data). Successful hybrid performance prediction can substantially increase breeding efficiency, which is of great interest to breeders. Molecular markers have been predicted and conceived as an efficient tool to reshape maize breeding programs and facilitate rapid gains from selection. In maize, several methods have been investigated for predicting hybrid performance using molecular markers (Maenhout et al. 2009; Reif et al. 2003; Schrag et al. 2007, 2009). Genomic selection to predict genetic values became possible with the development of high-throughput genotyping platforms and the availability of thousands of genome-wide molecular markers (Bernardo and Yu 2007; Piepho 2009; Crossa et al. 2010; Jannink et al. 2010). High correlation between true and genomic estimated breeding values in several simulation studies supports selection based on molecular markers alone (Heffner et al. 2009); however, more studies using real data should complement their utility in crop breeding. In addition, handling and selecting the rapidly increasing number of markers poses a challenge for genomic selection. For example, in the case of a limited budget, genotyping all mapped markers for a small number of individuals may be less efficient than genotyping a restricted set of well-chosen markers on a wider set of individuals (Maenhout et al. 2010).

In this study we characterized a total of 148 GEM lines and CMLs using 1,266 SNPs and evaluated the agronomic performance of 654 F1s from crosses between GEM and CML lines. The objectives of this study were: (1) to show the potential use of GEM lines for enhancement of CMLs in tropical maize hybrid breeding on the basis of genetic differences between them; (2) to construct a simple model for predicting hybrid performance using information on phenotypic values and molecular markers; and (3) to compare the efficiency of hybrid performance prediction by using random markers and selective markers with association analysis.

Materials and methods

Plant materials and phenotyping

Publicly released GEM enhanced lines ($n = 54$) obtained from the maize genebank at Iowa State University (ISU), USA (http://www.public.iastate.edu/~usda-gem/) and 94 CIMMYT lines released by the CIMMYT maize program were planted in a breeding nursery at Tlaltizapán station during the summer planting cycle (2007B). Twenty-nine of the 54 GEM lines adapted for the southern USA were registered in *Crop Science*, 2006 (Balint-Kurti et al. 2006; Carson et al. 2006). The 148 inbred lines used in this study are listed in Electronic Supplementary Material Table S1. Of the 54 GEM lines, 35 belonged to the SS heterotic pattern and 19 to the NSS heterotic pattern. Of the CMLs, 48 belonged to heterotic pattern A (dent grain type) and 38 belonged to heterotic pattern B (flint grain type) of the CIMMYT maize heterotic groups. There were eight CMLs considered to be both A and B patterns (A/B) (Table S1). To obtain F1 crosses among GEM × CML or CML × GEM, CML A lines and GEM SS lines were inter-mated plant to obtain as many different crosses as possible and CML B lines were also inter-mated with GEM NSS lines in the same manner. CML A/B lines were used for inter-mating with both GEM SS and NSS lines. A total of 654 F1 ears were obtained and planted in breeding nurseries in the following planting season (2008A) at CIMMYT stations in Agua Fria (AF, 20°27′00″N; 97°38′24″W, 100 m above sea level) in the state of Puebla and Tlaltizapán (TL, 18°40′48″N; 99°07′48″W, 940 m above sea level) in the state of Morelos, Mexico. Five plants of each F1 were planted, spaced at 0.25 m between plants and 0.75 m between rows, and selfed at flowering to produce F2 progeny. Phenotypic data were taken from five plants of each cross at both stations. A few F1 ears were saved for advancing inbreeding. Days to anthesis (DA) and plant height (PH) were measured on each F1 cross at the two locations mentioned above. PH was recorded for each plant as the distance between the ground surface and the top of the tassel. DA was recorded for each plot when at least 50% of the plants had reached anthesis.

The phenotypic value of each parental line was calculated as the mean of all of its F1 progeny from at least two cross combinations. Hence, the GCA was used as a surrogate for phenotypic trait of each inbred line.

Analysis of the phenotypic data was carried out using SAS 9.0 (SAS Institute 2002). For F1 crosses, the analysis of variance (ANOVA) across two environments (two locations) was performed and variance components of genotype, environment, G × E and residual effect were estimated. Broad-sense heritability (H^2) was calculated according to Knapp et al. (1985) as: $H^2 = \sigma_g^2/(\sigma_g^2 + \sigma_{ge}^2/n + \sigma^2/nb)$; where σ_g^2 is the genetic variance, σ_{ge}^2 is the genotype × environment interaction variance, σ^2 is the error variance, n is the number of environments, and b is the number of replications in each experiment.

SNP genotyping and in-silico mapping

We used an Illumina oligo pool assay (OPA) with 1,536 SNPs for genotyping 148 lines. The SNP genotyping details and the detailed information on each SNP can be found in the previous study (Yan et al. 2010b). SNP genotyping and data calling were performed using the Illumina BeadStation 500 G (Illumina, Inc., San Diego, CA, USA) according to the protocol described by Fan et al. (2006). The genotypes of F1 crosses were inferred from the genotypes of their parental lines. If one or both parents were heterozygous, the genotype of their F1 cross was recorded as missing.

Genetic structure analysis

We performed principal component analysis (PCA) to visualize genetic relationships among the 148 inbred lines and 654 F1s by using the software NTSYSpc (Darroch and Mosimann 1985). PCA for the 148 lines used 1,266 SNPs with good quality (i.e. polymorphic SNPs with less than 10% missing values and a heterozygosity less than 15%), and PCA for the 654 F1s used 872 SNPs with missing values less than 20%.

Relative kinship analysis

A relative kinship matrix was calculated by adopting the software package SPAGeDi (Hardy and Vekemans 2002). Familial relatedness between each two lines from the 148 inbred lines and between two F1 combinations were calculated based on the molecular data. To infer the kinship matrix, 1,266 and 384 SNPs were used for the inbred lines and F1 crosses, respectively, with less than 10% missing values.

Association analysis

Association analysis was performed on 148 inbred lines as well as 654 F1s based on the PCA, K and PCA + K models integrated in TASSEL 3.0 (Zhang et al. 2009), respectively. In the case of the 148 inbred lines, association was detected between 1,184 SNPs having a minor allele frequency (MAF) of greater than 0.05 and means of two traits (DA and PH) evaluated in two locations (TL and AF, CIMMYT stations). In the analysis of 654 F1s, association between 756 SNPs having minor genotype frequency of greater than 0.05 and the same traits mentioned above was detected.

Predicting hybrid performance

The mixed linear model used to estimate additive and dominant effects of SNPs is expressed as:

$$y = X\beta + Z_1 a + Z_2 d + e$$

where y is an $N_{F1} \times 1$ vector of phenotypic values in the training set; N_{F1} is the number of F_1 crosses; β is fixed effects (population mean); X is an $N_{F1} \times 1$ vector with all elements 1; Z_1 is an $N_{F1} \times N_M$ design matrix with elements equal to 1, −1 and 0 if the marker type is MM, mm and Mm, respectively; N_M is the total number of markers; a is an $N_M \times 1$ vector of additive effects of markers; Z_2 is an $N_{F1} \times N_M$ design matrix with elements equal to −1 or 2 if the marker type is homozygous or heterozygous, respectively; d is an $N_M \times 1$ vector of dominant effects of markers; and e is an $N_{F1} \times 1$ vector of residual effects.

In the mixed model, a, d and e are assumed to be random and follow normal distributions $N(0, V_A I)$, $N(0, V_D I)$ and $N(0, V_e I)$, respectively, where V_A is the additive variance, V_D is the dominant variance, and V_e is the error variance. Estimates of genetic variance (V_G) and residual variance (V_e) are obtained from an analysis of variance phenotypic means across locations. V_A and V_D are determined from V_G by $V_A + V_D = V_G$ and $V_A/V_D = 2.5$. Since it is not possible to partition V_A and V_D from V_G based on this data, the ratio 2.5 is empirically designated considering the proportion of V_A and V_D for the trait of interest.

The variance of additive and dominant at each of the N_M maker loci are assumed to be V_A/N_M and V_D/N_M, respectively. The effects of a and d are obtained by solving the mixed-model equations (Henderson 1984), with β as fixed effect, and a and d as random effects.The F_1 hybrid performance is predicted by the following:

$$\hat{y} = \hat{b}_0 + \sum_{i}^{n} \left(x_i \hat{a}_i + z_i \hat{d}_i \right) + e$$

where $\hat{b}_0$ is the estimate of overall mean; n is the number of SNPs; $\hat{a}_i$ and $\hat{d}_i$ are the estimates of additive and dominant effects of the ith SNP, respectively; and x_i and z_i are indicators with elements of 1 and 0 if the locus is homozygous, 0 and 1 if the locus is heterozygous.

In this study we used PH and DA as the target traits for predicting hybrid performance. The actual phenotypic data from two locations were averaged and 9 F1 crosses were excluded because of missing data at one location. Thus, prediction was performed based on information from a total of 645 F1s. A subset, composed of 600 randomly chosen F1 crosses, was considered as the training set, and the remaining 45 crosses were regarded as the testing set. In order to get better estimates of prediction effectiveness, we sampled training and testing sets ten times. The performance of PH and DA of the testing set was predicted on the basis of data from the training set, using the entire set of genome-wide markers (i.e. 1,266 SNPs) as well as different subsets of markers selected by different criteria. Regression of observed on predicted phenotypes of the testing set, which could provide estimates of the slope and the model fit (R^2), was informative for evaluating prediction accuracy.

The significance of each marker among the 1,184 and 756 SNPs was represented by the P value from the association analysis on inbreds and F1s, respectively. According to association results on inbreds performed based on the PCA + K model, markers used for prediction were selected as follows. All markers were ranked according to their P value, from the lowest to the highest. Subsets of SNP markers with sizes of 25, 50, 100, 250, 500 and 1,184 used for prediction were selected, following their ranking of P values. In addition, subsets of markers with the same sizes as mentioned above were randomly selected from the entire set of 1,266 SNPs, to be used for prediction. When referring to the results of association analysis on F1s using the PCA + K model, the same procedure for marker selection was used, but the sizes of the marker subsets were 25, 50, 100, 250, 500 and 756.

Results

Phenotypic data of F1 crosses between GEM and CML

Pedigree and heterotic groups of the CML and GEM lines are summarized in Table S1. F1 crosses of GEM × CML or CML × GEM grew well in CIMMYT stations. The 654 F1 crosses and their parent combinations are listed in Table S2. Table S3 summarizes means and ranges of DA and PH, for both F1s and 148 inbred lines at AF and TL stations in 2008. On average, F1 populations had longer flowering days (DA) at TL than at AF. However, plants grew relatively taller (PH) at AF than at TL. The broad-sense heritabilities of DA and PH were 84.0 and 65.9%, respectively (Table S3).

SNP genotyping

A total of 1,330 SNPs (87.4%) were successfully called with less than 20% missing data. Within the 1,330 SNPs, 34 were mono-polymorphic, 11 had more than 15% heterozygous data and 19 had more than 10% missing data in the 148 inbred lines. These SNPs were excluded from further analysis. Therefore, a total of 1,266 SNPs were used for the final data analysis for the 148 inbred lines. CMLs ($n = 94$) showed a range of heterozygosity from 0 to 12.8%, with an average of 1.1%. This is a normal level of residual heterozygosity in inbred lines of maize. On the other hand, seven lines out of 54 GEM lines had more than 20% heterozygosity. These need further selfing to reduce the residual heterozygosity.

SNP distribution and allelic frequency in GEM and CML

Among the 1,266 SNPs, 20 were mapped to contigs with unknown location and the remainder were evenly distributed across the whole genome, ranging from 77 SNPs on chromosome 7 to 217 SNPs on chromosome 1 (Table 1). The allelic frequency of each SNP varied in

Table 1 Distribution of the 1,266 SNPs used in the study across the maize genome

Chromosome	No. of SNPs
1	217
2	132
3	140
4	134
5	154
6	98
7	77
8	115
9	96
10	83
Unknown	20
Total	1,266

different germplasm groups, as shown in Fig. 1. A total of 3.6% alleles (46/1,266) were unique to CMLs and 4.4% alleles (56/1,266) were unique to GEM lines. The distribution of allelic frequency difference between GEM lines and CMLs is shown in Figure S1. Nearly half of the SNPs showed an allelic frequency difference of more than 0.2. Those SNPs ($n = 126$) with more than 0.4 allelic frequency difference were distributed on the whole genome. The physical position of these SNPs and their allelic frequency difference are summarized in Table S4.

Genetic structure and kinship relation

Principal component analysis classified three clear subgroups corresponding to GEM SS and NSS heterotic patterns, and CMLs, among the 148 inbred lines. Based on the genotypic classification, nine GEM SS lines were located between the main group of GEM SS lines and NSS lines, indicating they were of mixed origin. Within CMLs, there were no obvious subgroups identified by PCA (Fig. 2). The F1 populations were separated into two groups based on the first two principal components (Fig. 3).

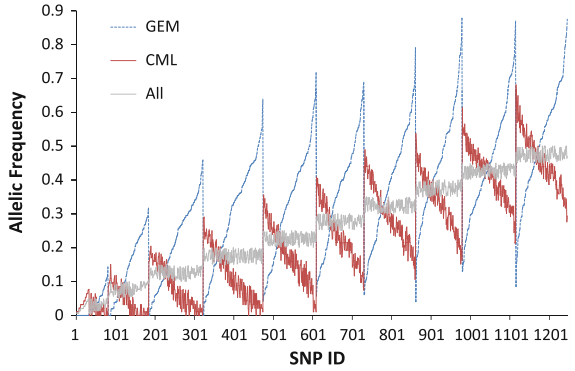


Fig. 1 Allelic frequency of each SNP of 1,266 bi-allelic SNPs used in this study varied in different germplasm groups: shown in *blue* in the group of GEM *lines*, shown in *red* in the group of CML *lines*, and shown in *grey* in all 148 inbred *lines*. (Color figure online)

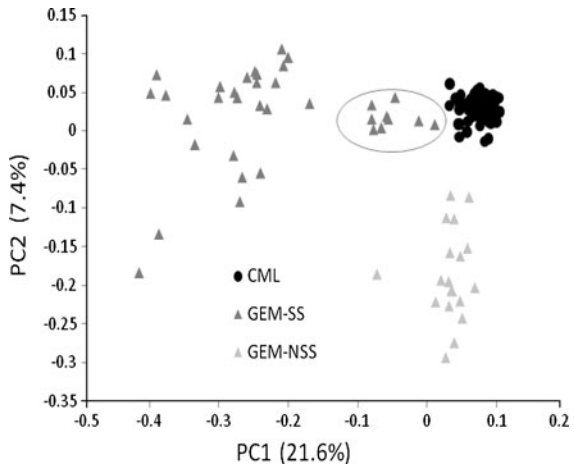


Fig. 2 Principal component analysis (PCA) of 148 inbred *lines* based on 1,266 SNPs. *CML* CIMMYT maize line, *GEM-SS* germplasm enhancement of maize-stiff stalk, *GEM-NSS* germplasm enhancement of maize-non-stiff stalk. *Lines* in the *circle* may have a mixed origin

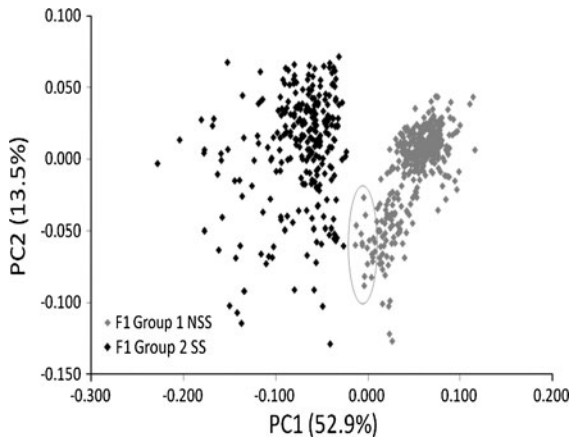


Fig. 3 Principal component analysis (PCA) of 654 F1 crosses based on 872 SNPs. *Lines* in Group 1 NSS (non-stiff stalk) are from the crosses CIMMYT maize line (CML) B or A/B × germplasm enhancement of maize (GEM) NSS *lines*. *Lines* in Group 2 SS (stiff stalk) are from the crosses CML A or A/B × GEM SS *lines*. *Lines* in the *circle* are from the GEM parental *lines* which have a mixed origin

The two large F1 clusters constructed by the GEM genotypes and CMLs were expected to form clearly separate genetically different groups or heterotic groups that were influenced by SS and NSS heterotic patterns of GEM lines. The SS lines × CML A and A/B lines formed Group 2 of the F1 cluster and the NSS lines × CML B and A/B lines formed Group 1 of the F1 cluster. A small part of the F1 crosses located between Group 1 and Group 2, and their GEM parents, were the nine lines mentioned above, which may have a mixed origin (Fig. 3).

The distribution of pair-wise kinship relation between any two lines from both inbred lines (148) and F1s (654) showed complex familial relatedness among these lines (Fig. S2). In both cases, more than half of the pairs had a kinship value of 0. The majority of the pair-wise kinship values in the case of the inbred lines were less than 0.2. However, they were less than 0.3 for the F1s.

Association analysis

Associations were conducted to evaluate the performance of PCA, K, and PCA + K models for controlling false positives in using both inbreds and F1s. In Figure S3, quantile–quantile plots of $-\log_{10}(P)$ value for results of association for the two traits based on both inbreds and F1s indicated that the models can effectively control Type I error in both cases when K was taken into account. When performing association analysis for both DA and PH on inbreds, all the three models were similar, and the PCA + K model showed a little superiority to the other two in reducing the Type I error (Fig. S3a). However, compared to the PCA model, the K and PCA + K model can greatly reduce the Type I error for both traits when performing association analysis on F1s (Fig. S3b).

Prediction of hybrid performance for PH and DA

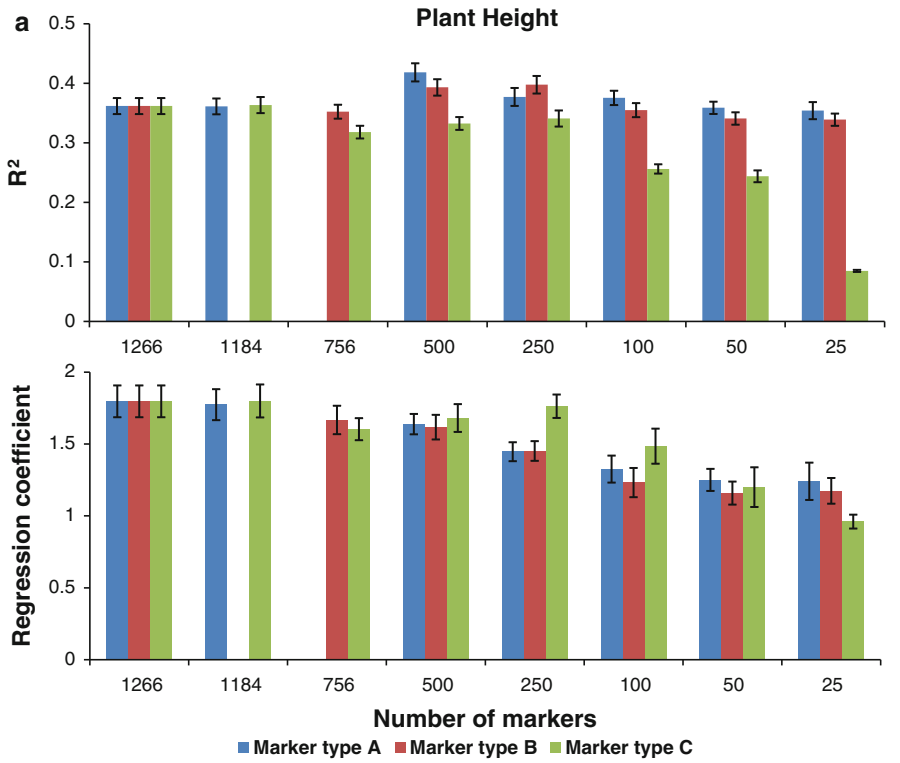
For both PH and DA, the correlations between predicted values based on different sets of SNP markers and observed values are shown in Fig. 4. Prediction accuracy was investigated as the correlation (R^2) between the predicted and the true phenotypic values. For PH, R^2 ranged from 0.354 to 0.418, when using the most significant makers based on results of association analysis on inbreds (marker type A) (Fig. 4a). On the other hand, when the most significant markers based on the results of association analysis on F1s (marker type B) were used, R^2 ranged from 0.339 to 0.398 (Fig. 4a). R^2 reached a peak when using the 500 most significant markers in the case of marker type A, and using the 250 most significant markers in the case of marker type B (Fig. 4a). The regression coefficients were 1.639 and 1.452 for the peaks in marker type A and marker type B, respectively, which indicated that the predicted values of F1 hybrids were slightly over-estimated. In the case of randomly selected markers (marker type C), R^2 basically increased with marker density, and ranged from 0.085 to 0.364. Unexpectedly, when the number of selected markers became smaller (e.g. $n \leq 500$), using marker type A or B outperformed using the same number of randomly selected markers, but the difference between them decreased with an increase in the number of markers (Fig. 4a).

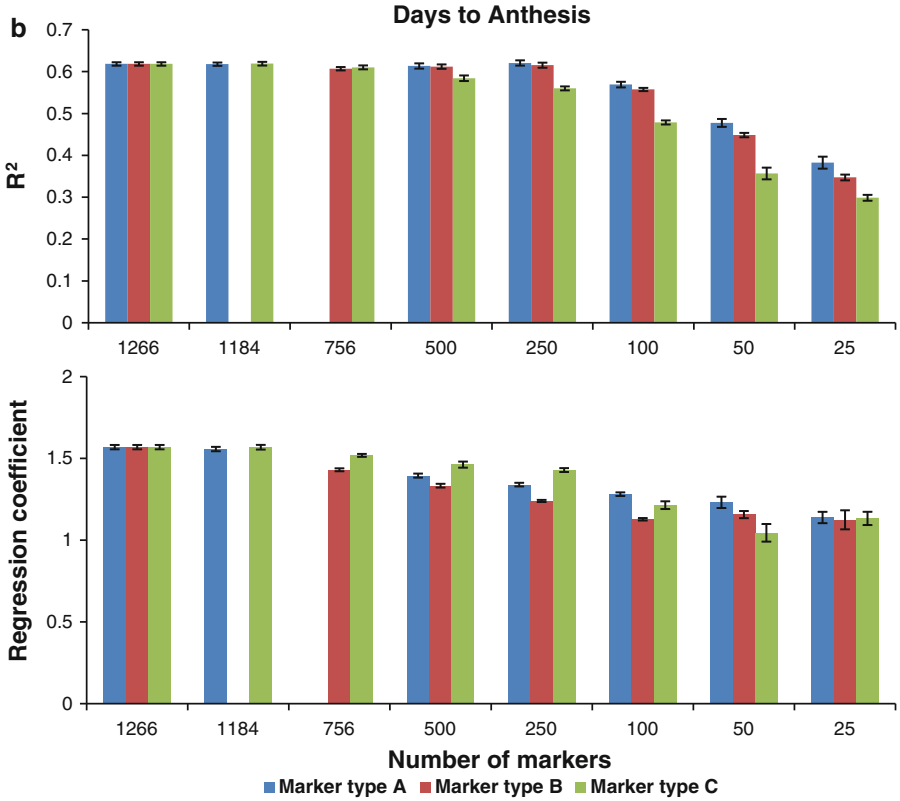
Similar results were obtained for the hybrid performance prediction of DA (Fig. 4b). In both PH and DA, using fewer, but trait-associated, markers showed a relatively higher prediction accuracy than employing the entire set of genome-wide markers ($n = 1{,}266$). In this study, applying the results of association analysis on inbreds and on F1s in genomic selection to predict hybrid performance produced similar prediction accuracy in both PH and DA.

Discussion

Implications of utilizing temperate lines in tropical maize breeding program

Recently, a large number of SNP markers have become available in maize for genome-wide fingerprinting, and they have been successfully used in maize genetic diversity and genetic structure analysis with reasonable results (Wen et al. 2011; Yan et al. 2009; Yang et al. 2010). In tropical maize breeding, research and development of the germplasm that belongs to different heterotic groups and/or patterns is fundamental for breeding high-yielding maize hybrids. CIMMYT maize gene pools and populations have been traditionally divided into flint and dent heterotic patterns, largely by phenotypic selection (Taba and Chávez 2007; Ortiz et al. 2010). CML lines are grouped by heterotic pattern A and B, or A/B based on the hybrid performance with testers


a
Plant Height
R²
0.5
0.4
0.3
0.2
0.1
0
1266
1184
756
500
250
100
50
25
Regression coefficient
2
1.5
1
0.5
0
Number of markers
Marker type A
Marker type B
Marker type C


b
Days to Anthesis
R²
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1266
1184
756
500
250
100
50
25
Regression coefficient
2
1.5
1
0.5
0
Number of markers
Marker type A
Marker type B
Marker type C

◀ **Fig. 4** Regression of observed on predicted single-cross performance of **a** plant height and **b** days to anthesis on different types of markers. The *upper part* shows the model fit (R^2) and the *lower part* shows the estimate of the slope (regression coefficient). *Marker type A*: selection of markers referring to results of association analysis in inbreds, and starting from the marker with highest significance. *Marker type B*: selection of markers referring to results of association analysis in F1s, and starting from the marker with highest significance. *Marker type C*: randomly selected markers from the 1,266 SNPs

(http://www.cimmyt.org/ru/component/content/article/459-international-maize-improvement-network-imin/434-cimmyt-maize-inbred-lines-cml). In this study, SNP markers were used to identify the genetic structure for both GEM lines and CMLs. Based on the SNP data, the heterotic patterns of GEM lines are clearly shown to be associated with SS and NSS patterns bred by the United States GEM (Salhuana and Sevilla 1995; Salhuana et al. 1998).

However, based on the SNP data, the divergence shown between the heterotic groups (i.e., A and B type) within CMLs is not as large as that of GEM lines which contain 50 or 75% temperate germplasm by pedigree. Although the difference within temperate lines may be overestimated compared to that within tropical lines due to ascertainment bias arising from the SNPs used in this study, similar results on the heterotic patterns of CMLs were reported based on SSR data (Xia et al. 2005), indicating that obvious heterotic groups are not apparent because CIMMYT lines have been developed from gene pools and populations with a wide germplasm base. The development of heterotic groups at CIMMYT was rather recent, after inbreeding in these gene pools and populations for line development from the mid-1980s (Vasal et al. 1999; Reif et al. 2003). With the availability of the maize genome and the advances in genotyping by sequencing technology, we may obtain larger numbers of SNPs with good quality for characterizing maize lines, making it possible to control ascertainment bias (Elshire et al. 2011).

Exotic germplasm can provide new desirable alleles for line and population improvement. Large allelic frequency differences observed between GEM and CML, together with the unique alleles harbored within both germplasms, imply a mutual improvement between the two sets of germplasm. The obviously larger genetic divergence between the SS and NSS heterotic groups of the GEM lines, as compared to the A and B heterotic groups of CMLs, demonstrated the accumulations of different genes or alleles in opposite heterotic groups of GEM lines. In order to enhance heterosis in grain yields of tropical maize, it is suggested that wider resources for introgression of exotic germplasm are needed to increase the genetic distances between opposite heterotic lines and populations (Ron Parra and Hallauer 1997; Reif et al. 2003). According to the current results of molecular characterization and trial evaluation, GEM lines showed great potential to be utilized in tropical maize improvement. However, they need to acquire good adaptation in tropical maize production environments as they contain 50 or 75% elite temperate germplasm and only 25 or 50% tropical germplasm. The strategy of developing F1s between lines of GEM SS and CML heterotic type A, and also between lines of GEM NSS and CML heterotic type B, was initially employed for introgression of useful alleles from GEM to CML. F1 crosses were fairly well adapted at CIMMYT stations (AF and TL) in general, producing reasonable ears; whereas the performance of some GEM lines was poor in the breeding nursery. At the AF station, plants were affected by late infection of tropical leaf rust. The plants stood well at both stations, showing good stalk strength and root development. In general, minimum ear rot was observed. Variable F1s were selected and followed by second and third selfing-selection seasons in both TL and AF.

A new population for association analysis and Type I error control using different models

Crosses between exotic germplasm and elite local germplasm are fundamental to a crop enhancement program. Thus, a population of F1 crosses can be formed. Developed from limited parental lines, F1 crosses provide us with novel material for association analysis. Larger population size compared to their parental lines as well as the easy accessibility of their phenotypic information underlies the utility of F1s for association analysis, which is especially applicable and helpful in a breeding program, as presented here.

Several methods have been proposed for association analysis in populations with a complex genetic structure. In many plant species, it is shown that a mixed-model approach outperforms the much discussed methods developed in the context of human genetics (Zhu et al. 2008). A recent study indicated

that the K model and two mixed models (Q + K and PCA + K) performed well for all traits in the maize panel of 527 maize inbred lines, and both Q + K and PCA + K models performed slightly better than the K model for all three traits (Yang et al. 2011). The performance of these models in the present study was consistent with these previous reports. It is suggested that the K matrix is generally superior to the model using only Q for association analysis, since the Q matrix only provides a rough dissection of the population differentiation (Yu et al. 2006; Myles et al. 2009). In our study, the K model greatly outperformed using only PCA when conducting association analysis on F1s, and it was slightly better than the latter when conducting association analysis on inbreds. This makes sense because the K matrix captures the relatedness between each pair of individuals within the panel while PCA or Q takes only a few axes of variation into account (Myles et al. 2009). In terms of Type I error control, the difference in the performance of PCA between association analysis on inbreds and F1s may due to the genetic background of the samples. The 148 inbred lines are samples with both population structure and familial relationship. However, the heterozygous F1 population also has population subdivision and the familial relatedness among the individuals within it is more complex. The PCA model which contains 10 dimensions of principal components in this study may be enough to capture the differentiation within the 148 inbred lines and correct for its genetic relatedness.

The potential and strategy of genomic selection in maize breeding

The accurate prediction of maize single crosses among heterotic groups could facilitate hybrid breeding. With the rapidly increasing number of available molecular markers, prediction based on genome-wide markers has become a new trend for identifying superior single crosses. By applying best linear unbiased prediction (BLUP), Bayesian analyses and other statistical methods, selection on genetic values predicted from the whole genome markers could substantially increase the rate of genetic gain in animals and plants (Meuwissen et al. 2001; Crossa et al. 2010; Hayes et al. 2009; Jannink et al. 2010). In this study, a strong correlation between the predicted and the true phenotypic values was identified, when using the entire set of genome-wide markers. The high prediction accuracy indicates the potential and efficiency of genomic selection, based on our model, for predicting hybrid performance.

However, prediction with the higher marker density has not performed best for the traits (PH and DA) in this study. The prediction accuracy benefitted from adoption of the markers that were associated with quantitative trait loci for both traits. Other than using the entire set of genome-wide markers, the selection of markers with the highest significance, and accounting for only 40 and 20% of the total marker number for PH and DA, respectively, provided the highest prediction accuracy. The results demonstrated that using fewer, but trait-associated, markers may be more effective than exploiting the entire set of genome-wide markers as they produce less noise in estimating the genetic values. There may also be a trade-off between marker size and marker quality (i.e., the effect of each marker on the specific traits), which affects the prediction accuracy of genomic selection models for hybrid performance. From the present results, it is suggested that genomic prediction combined with association analysis could improve prediction efficiency and reduce costs.

On the other hand, using a limited number of markers is insufficient to detect significantly associated markers with some important traits of maize. It poses a question about using genome-wide association study (GWAS) for identifying and empirically validating a set of significant markers for genomic selection. Myles et al. (2009) suggested that 10–15 million markers may be necessary for performing GWAS in diverse maize varieties. Even when adopting markers developed from the expressed portion of the genome (i.e., assuming 50,000 genes in the maize genome and 10–20 markers developed within the expressed regions of each gene), 500,000–1,000,000 well-chosen markers is considered sufficient (Yan et al. 2011), which currently may not be achievable in most research laboratories. Moreover, most quantitative traits like flowering time (Buckler et al. 2009) and drought tolerance (Messmer et al. 2009) are controlled by many SNPs of small effect. The same situation was identified in the case of human height, where almost 88% of the variation due to SNPs has been undetected in published GWAS because the effects of the SNPs are too small to be statistically significant under overly stringent significance tests (Yang et al. 2010). This

conclusion was confirmed by the present results, to some extent, where selecting a certain number of markers, according to their significance, performed best in the prediction. In this study, only about 1,000 SNPs were used, but high prediction efficiency was observed, which may be due to the relatively simpler population background as well as the even genomic distribution of those SNP markers. Present results imply that profiling and detecting SNPs or genes previously underlying the target traits may be more informative for genomic selection and enhancing predictive power. For a given trait, using a fixed set of markers or genes for prediction may be more efficient. GWAS, as well as the transcriptional data suggested by Frisch et al. (2010), and metabolite profiles as predictors reported by Steinfath et al. (2010), can be complementary to genomic selection, and these items should be utilized together if possible.

Acknowledgments This work was supported by the government of Japan for germplasm enhancement of the CIMMYT maize genebank. We acknowledge both the U.S. Germplasm Enhancement of Maize (GEM) and the International Maize and Wheat Improvement Center (CIMMYT) breeders for the lines used in this study. They are provided from North Carolina University (NCU), Iowa State University (ISU) and the CIMMYT maize genebank.

References

Balint-Kurti PJ, Blanco M, Millard M, Duvick S, Holland J, Clements M, Holley R, Carson ML, Goodman MM (2006) Registration of 20 GEM maize breeding germplasm lines adapted to the southern USA. Crop Sci 46:996–998

Bernardo R, Yu J (2007) Prospects for genome-wide selection for quantitative traits in maize. Crop Sci 47:1082–1090

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 H, Mitchell SE, Pressoir G, Peiffer JA, Rosas MO, Rocheford TR, Cinta Romay M, Romero S, Salvo S, Villeda HS, Sofia da Silva H, Sun Q, Tian F, Upadyayula N, Ware N, Yates H, Yu J, Zhang Z, Kresovich S, McMullen MD (2009) The genetic architecture of maize flowering time. Science 325:714–718

Carson ML, Balint-Kurti PJ, Blanco M, Millard M, Duvick S, Holley R, Hudyncia J, Goodman MM (2006) Registration of nine high-yielding tropical by temperate maize germplasm lines adapted for the southern USA. Crop Sci 46:1825–1826

Crossa J, Campos G, Pérez P, Gianola D, Burguen J, Araus JL, Makumbi D, Singh R, Dreisigacker S, Yan J, Arief V, Bänziger M, Braun HJ (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186:1–12

Darroch JN, Mosimann JE (1985) Canonical and principal components of shape. Biometrika 72:241–252

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLos One 6(5):e19379

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

Frisch M, Thiemann A, Fu J, Schrag TA, Scholten S, Melchinger AE (2010) Transcriptome-based distance measures for grouping of germplasm and prediction of hybrid performance in maize. Theor Appl Genet 120:441–450

Goodman MM (1999) Broadening the genetic diversity in maize breeding by use of exotic germplasm. In: Coors JG, Pandey S (eds) The genetics and exploitation of heterosis in crops. ASA, CSSA and SSSA, Madison, pp 139–148

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

Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009) Invited review: genomic selection in dairy cattle: progress and challenges. J Dairy Sci 92:433–443

Heffner EL, Sorrels MR, Jannink JL (2009) Genomic selection for crop improvement. Crop Sci 49:1–12

Henderson CR (1984) Applications of linear models in animal breeding. University of Guelph, Guelph

Jannink JL, Lorenz AJ, Iwata H (2010) Genomic selection in plant breeding: from theory to practice. Brief Funct Genomics 9(2):166–177

Knapp SJ, Stroup WW, Ross WM (1985) Exact confidence intervals for heritability on a progeny mean basis. Crop Sci 25:192–194

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

Maenhout S, De Baets B, Haesaert G (2009) Prediction of maize single-cross hybrid performance: support vector machine regression versus best linear prediction. Theor Appl Genet 120(2):415–427

Maenhout S, De Baets B, Haesaert G (2010) Graph-based data selection for the construction of genomic prediction models. Genetics 185(4):1463–1475. doi:10.1534/genetics.110.116426

Messmer R, Fracheboud Y, Bänziger M, Vargas M, Stamp P, Ribaut JM (2009) Drought stress and tropical maize: QTL-by-environment interactions and stability of QTLs across environments for yield components and secondary traits. Theor Appl Genet 119(5):913–930

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829

Myles S, Peiffer J, Brown PJ, Ersoz ES, Zhang ZW, Costich DE, Buckler ES (2009) Association mapping: critical considerations shift from genotyping to experimental design. Plant Cell 21(8):2194–2202

Ortiz R, Taba S, Tovar VHC, Mezzalama M, Xu Y, Yan J, Crouch JH (2010) Conserving and enhancing maize genetic resources as global public goods—a perspective from CIMMYT. Crop Sci 50:1–16

Piepho HP (2009) Ridge regression and extensions for genome-wide selection in maize. Crop Sci 49:1165–1176

Reif JC, Melchinger AE, Xia XC, Warburton ML, Hoisington DA, Vasal SK, Srinivasan G, Bohn M, Frisch M (2003) Genetic distance based on simple sequence repeats and heterosis in tropical maize populations. Crop Sci 43:1275–1282

Ron Parra J, Hallauer AR (1997) Utilization of exotic maize germplasm. Plant Breed Rev 14:165–187

Salhuana W, Sevilla R (eds) (1995) Latin American Maize Project (LAMP), stage 4 results from homologous areas 1 and 5 (Catalog and CD-ROM). National Seed Storage Laboratory, Fort Collins

Salhuana W, Jones Q, Sevilla R (1991) The Latin American Maize Project: model for rescue and use of irreplaceable germplasm. Diversity 7:40–42

Salhuana W, Pollak LM, Ferrer M, Paratori O, Vivo G (1998) Agronomic evaluation of maize accessions from Argentina, Chile, The United States, and Uruguay. Crop Sci 38:866–872

SAS Institute (2002) Statistical analysis software for windows, 9.0. SAS Institute Inc., North Carolina, USA

Schrag TA, Maurer HP, Melchinger AE, Piepho HP, Peleman J, Frisch M (2007) Prediction of single-cross hybrid performance in maize using haplotype blocks associated with QTL for grain yield. Theor Appl Genet 114:1345–1355

Schrag TA, Mohring J, Maurer HP, Dhillon BS, Melchinger AE, Piepho HP, Sørensen AP, Frisch M (2009) Molecular marker-based prediction of hybrid performance in maize using unbalanced data from multiple experiments with factorial crosses. Theor Appl Genet 118:741–751

Steinfath M, Gärtner T, Lisec J, Meyer RC, Altmann T, Willmitzer L, Selbig J (2010) Prediction of hybrid biomass in *Arabidopsis thaliana* by selected parental SNP and metabolic markers. Theor Appl Genet 120:239–247

Taba S, Chávez VH (2007) Enhancement of core accessions and the broad based gene pools for tropical maize improvement. Centro Internacional de Mejoramiento de Maíz y Trigo, México, D.F

Vasal SK, Cordova H, Pandey S, Srinivasan G (1999) Tropical maize and heterosis. In: Coors JG, Pandey S (eds) The genetics and exploitation of heterosis in crops. ASA, CSSA and SSSA, Madison, pp 363–373

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

Xia 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 U.S. and European maize. Crop Sci 45:2573–2582

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

Yan JB, Warburton M, Crouch J (2011) Association mapping for enhancing maize genetic improvement. Crop Sci 51:1–17

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, Goddard ME, Visscher PM (2010) Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42:565–569

Yang XH, Gao SB, Xu ST, Zhang ZX, Prasanna BM, Li L, Li JS, Yan JB (2011) Characterization of a global germplasm collection and its potential utilization for analysis of complex quantitative traits in maize. Mol Breed 28:511–526. doi:10.1007/s11032-010-9500-7

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

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

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