Theor Appl Genet (2011) 122:1149–1160
DOI 10.1007/s00122-010-1519-y

ORIGINAL PAPER

Effect of population structure corrections on the results of association mapping tests in complex maize diversity panels

Sofiane Mezmouk · Pierre Dubreuil · Mickaël Bosio · Laurent Décousset · Alain Charcosset · Sébastien Praud · Brigitte Mangin

Received: 17 June 2010 / Accepted: 11 December 2010 / Published online: 11 January 2011

Abstract Association mapping of sequence polymorphisms underlying the phenotypic variability of quantitative agronomical traits is now a widely used method in plant genetics. However, due to the common presence of a complex genetic structure within the plant diversity panels, spurious associations are expected to be highly frequent. Several methods have thus been suggested to control for panel structure. They mainly rely on ad hoc criteria for selecting the number of ancestral groups; which is often not evident for the complex panels that are commonly used in maize. It was thus necessary to evaluate the effect of the selected structure models on the association mapping results. A real maize data set (342 maize inbred lines and 12,000 SNPs) was used for this study. The panel structure was estimated using both Bayesian and dimensional reduction methods, considering an increasing number of ancestral groups. Effect on association tests depends in particular on the number of ancestral groups and on the trait analyzed. The results also show that using a high number of ancestral groups leads to an over-corrected model in which all causal loci vanish. Finally the results of all models tested were combined in a meta-analysis approach. In this way, robust associations were highlighted for each analyzed trait.

Communicated by E. Carbonell.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-010-1519-y) contains supplementary material, which is available to authorized users.

S. Mezmouk · P. Dubreuil · M. Bosio · L. Décousset · S. Praud
BIOGEMMA, Genetics and Genomics in Cereals, rue des Frères Lumière, 63028 Clermont-Ferrand Cedex 2, France

S. Mezmouk · B. Mangin (✉)
BIA Unit, INRA, Chemin de Borde-Rouge, BP 52627, 31326 Castanet-Tolosan Cedex, France
e-mail: Brigitte.Mangin@toulouse.inra.fr

A. Charcosset
INRA, UMR de Génétique Végétale, Ferme du Moulon, 91190 Gif-sur-Yvette, France

Introduction

Association mapping aims at linking phenotypic variation to common sequence polymorphisms in collections of unrelated individuals. This approach, initially developed in human genetics (see for reviews Khoury et al. 2009; Hirschhorn et al. 2002), was introduced with success in various plant species (see for reviews Zhu et al. 2008; Ersoz et al. 2007) and is now widely used in plant genetics. Compared to linkage mapping, it offers several advantages such as: (1) the saving of time and money by using existing populations instead of creating cross-controlled populations, (2) analysis of more than two alleles per locus on average (depending on the panel diversity), and (3) high expected resolution owing to a short extent of linkage disequilibrium (Flint-Garcia et al. 2005). However, besides these assets, there are still some disadvantages such as (1) the presence of rare alleles at some loci, (2) the need for a high marker density and (3) the need for efficient control of panel population structure and/or relatedness between individuals (Thornsberry et al. 2001; Yu et al. 2006).

The short extent of linkage disequilibrium in the diversity panels requires a high marker density to increase the chances of detecting a causal polymorphism or polymorphisms that are tightly linked to the former. With the advent of plant genome sequencing projects (http://www.ncbi.nlm.nih.gov/genomes/leuks.cgi) and with the development of new medium- and high-throughput genotyping

and sequencing techniques (Lin et al. 2008; Wang et al. 2009; Shendure and Ji 2008), this inconvenience should be solved in the near future for both model plant species and those that have economic importance.

The plant panels comprise samples of mixed and/or admixed individuals from different genetic origins. The presence of several genetic origins within the panels, in different and unknown proportions, induces linkage disequilibrium between unlinked loci (Ersoz et al. 2007); consequently, this may increase the rate of false positives that are statistically associated to the analyzed trait without actually being involved in its phenotypic variation. In synthetic association populations, such as the multi-parent advanced generation inter-cross (MAGIC) populations (Cavanagh et al. 2008) and the maize nested association mapping (NAM) population (Yu et al. 2008), the structure is well characterized. They allow thus both high mapping resolution and better control of population structure. Conversely, within the classical association panels, the real structure of the material is not totally known; it is thus inferred with various statistical methods using molecular markers. In the association mapping test, the structure control methods can be divided into three types of approach: (1) genomic control, (2) fixed model and (3) unified mixed model.

The genomic control method uses random markers to evaluate the global structure effects on *P* values; the latter are then adjusted to account for the statistical inflation caused by the structure (Devlin and Roeder 1999). The fixed model approaches use molecular markers to estimate the panel structure; these estimates are integrated in the association mapping tests as covariate fixed effects. Likewise, mixed model approaches use both fixed and random effects to control the panel structure (Yu et al. 2006; Zhao et al. 2007; Stich et al. 2008).

To compute the fixed structure effects, Bayesian methods (Pritchard et al. 2000; Corander et al., 2003) and Principal Components Analyses (Price et al. 2006; Patterson et al. 2006; Zhu and Yu 2009) are widely used. The Bayesian model-based clustering methods assume Hardy–Weinberg and linkage equilibrium between the loci within the subpopulations. Starting with uniform priors, information about the origins of the individuals (in the case of mixture) or about the origin of proportions of individual genomes (in the case of admixture) is inferred. Approximations of posterior distributions are obtained using Markov chain Monte Carlo (MCMC) methods.

Principal component analysis (PCA) is one of the most widely used techniques for dimensional reduction and data summary. It enables the identification of key components of the structure within the data without resorting to a model (McVean 2009). Since the markers used to estimate the panel structures are chosen so as to be physically distant and selectively neutral, the linkage disequilibrium between them is due principally to the panel stratification; the panel structure is thus the main information that is summarized by the first components.

With the above methods, the number of ancestral groups used to account for the panel structure is set by the user. Prior information and several independent approaches, relying mostly on ad hoc criteria, are used to help select the suitable number of groups. However, this choice is quite difficult to make for complex panels, because the results given by the different criteria are often inconsistent and prior information may not help to select one of them. The structure model might thus be mistaken and lead to an under- or over-estimation of the true panel structure. The under- and over-structured models may either increase the rate of false positives or false negatives. The problem is then to determine the extent of this effect and to find a way to select true positives in a real data set if no obvious structure model is revealed by the different criteria.

The aim of this study was to investigate the effect of the number of covariates used to control the structure effects on the results of association tests by using a maize diversity panel that was known to have a very complex internal structure (Camus-Kulandaivelu et al. 2006; Camus-Kulandaivelu et al. 2007). The structure estimates were computed using both Bayesian model-based and dimensional reduction methods. The results of the association mapping tests with these structure models were compared for several group numbers.

To select the true positives, the association mapping results with the entire tested models were summarized in a meta-analysis approach, which highlighted the loci showing the most robust associations.

Materials and methods

Plant material and phenotypic data

A total of 342 maize inbred lines from the diversity panel previously described by Camus-Kulandaivelu et al. (2006) were analyzed (see Table 1 on supplemental data). This panel is representative of European and American maize diversity and covers a wide range of flowering times. Within the 342 inbred lines, 139 were directly obtained by selfing from traditional open pollinated varieties. The inbred lines used in the study have non-missing phenotypic data and less than 10% of heterozygous or missing genotypes.

The analyses focused on two phenotypic traits with different correlations to population structure: male flowering time (MFT) and thousand-kernel weight (TKW). Compared to MFT, TKW is less correlated to the structure.

The field trials and the phenotypic data analyses have been described by Camus-Kulandaivelu et al. (2006) for MFT and by Manicacci et al. (2009) for TKW. The phenotypic variation explained by the genotypes extended from 808.2 up to 1,556 growing day degrees (GDD) for MFT and from 69 up to 264 g for TWK.

Genotypic data

A set of 12,000 proprietary single nucleotide polymorphisms (SNPs), from 3,704 gene coding sequences, were genotyped using an Illumina Infinium BeadArray technology (Steemers and Gunderson 2007). Among these loci, 9,945 were polymorphic in the panel with minor allele frequencies higher than 3%.

Within the gene coding sequences, 4,058 haplotypic markers were reconstituted by concatenating SNPs that are physically located in a 1,500 base pairs (bp) window, yielding multiallelic markers. The concatenations were performed within a 1,500-bp window, as linkage disequilibrium decline very quickly beyond this distance (data not shown). SNPs that were alone in a given genic region were, however, kept for the analyses.

Statistical analyses

The association mapping tests were carried out at the SNP level using a two-step linear model. First, a covariate matrix S was introduced to deal with the phenotypic variability due to the panel structure:

$$Y = \mathbf{1}\mu + S\beta + e \quad (1)$$

where Y is a vector of phenotypic data; $\mathbf{1}$ is an identity matrix; μ is the trait mean; S is the matrix of structure covariates; β is a vector of the panel structure effects and e is a vector of the residual effects.

Second, the SNP effects were tested on the adjusted phenotypes using the following model:

$$\hat{e} = \mathbf{1}\mu' + M\theta + e' \quad (2)$$

where $\hat{e}$ is the vector of the phenotypes corrected from the panel structure; $\mathbf{1}$ is an identity matrix; μ' is the adjusted phenotype mean; M is a tested locus; θ is a vector of locus effects and e' is a vector of model residuals.

When no structure control was applied, the SNP effects were directly tested on the phenotypes.

The multiplicity problems were resolved by controlling the false discovery rate (Benjamini and Hochberg 1995) at 10%. The R package qvalue (Storey 2002) was used.

To compute the structure estimates, a set of 641 haplotypic markers were selected among the 4,058 available ones; these selected loci had less than 5% missing data and were genetically mapped at intervals of more than 1 cM, allowing a good coverage of the IBM maize genetic map (Falque et al. 2005). Among these 641 loci, 233 were biallelic (SNPs that are alone in a given genic region); the remaining 408 loci were multiallelic with an allele number varying from 3 to 38; the average allele number was 5 per locus. A total of 2,485 alleles were identified. Each inbred line was analyzed as a haploid individual.

In all, 98 different models (S matrices) were used to control the panel structure in the association mapping tests. They were computed by the STRUCTURE Software (Pritchard et al. 2000) for the most likely and the second most likely outputs, principal component and multiple correspondence analyses (Table 1).

The admixture model of STRUCTURE.2.2 software (Pritchard et al. 2000) was run on the assumption that allele frequencies are correlated among subgroups (Falush et al. 2003). This assumption is in agreement with the results of Matsuoka et al. (2002), which suggest that cultivated maize traces back from a single domestication event. Twenty independent repetitions of a number of groups varying from 1 (no structure within the panel) to 20 were performed with a 50,000 burn-in period followed by 100,000 iterations.

PCA is a model-free method that is widely used to describe population structures. It is generally carried out on biallelic markers for which numerical values are attributed (0 and 2 for the homozygous and 1 for the heterozygous

Table 1 Summary of the tested structure models

Structure methods	Corresponding structure models			
	Name of the S matrices	Number of structure groups[a]	Degrees of freedom[b]	Number of resulting models
STRUCTURE2.2 software (most likely model)	Q_1	2–20	1–19	19
STRUCTURE2.2 software (second most likely model)	Q_2	2–20	1–19	19
Principal component analysis	P	2–31	1–30	30
Multiple correspondence analysis	M	2–31	1–30	30

[a] The number of ancestral groups in the structure models

[b] The number of degrees of freedom absorbed by the S matrix in the association mapping tests

loci) (Price et al. 2006; Patterson et al. 2006). Since this study's loci were multiallelic, the PCA was adapted to the data set as follows.

Starting with the table of genotypes $X_{(N,L)}$ (N being the number of individuals and L the number of loci), a disjunctive table $A_{(N,M)}$ (a 0–1 binary table in which each column indicates whether an allele is absent or present in a given genotype) was generated with N individuals and M alleles. Table A entries were centered and standardized by subtracting the column means p_{al} ($p_{\mathrm{al}} = \frac{1}{N}\sum_{i=1}^{N} A_{\mathrm{ial}}$) and then dividing by the column standard deviation ($\sqrt{p_{\mathrm{al}}(1-p_{\mathrm{al}})}$); the missing values were set to 0. From each locus, one allele was suppressed and the PCA was performed on the resulting table A' using the R package Ade4 (Chessel et al. 2004). The alleles were normalized in this way because each one approximates a binomial distribution with a variance $Np_{\mathrm{al}}(1-p_{\mathrm{al}})$.

Given that the study's genotypic data were qualitative in nature, a multiple correspondence analysis (MCA) was also performed. Similarly to PCA, MCA is a dimensional reduction and data summary technique applied to categorical variables (Tenenhaus and Young 1985). It is an extension of the bivariate simple correspondence analysis to more than two variables. MCA was performed on table $X_{(N,L)}$ using the R package Ade4 (Chessel et al. 2004).

To help select the "appropriate" number of groups to control the panel structure, statistical test and several ad hoc criteria were proposed (Evanno et al. 2005; Pritchard et al. 2007; Camus-Kulandaivelu et al. 2007; Patterson et al. 2006). The best structure models in accordance with these criteria were selected and their association results were compared with the results of the neighboring models.

Results

Structure estimates

The STRUCTURE2.2 admixture model was run for a group number (K) varying from 1 to 20 with 20 repetitions for each group. To select suitable K, the criterion suggested by Pritchard et al. (2007) was applied to choose the smallest K after having reached a plateau of the "Ln $P(D)$" values ("Ln $P(D)$" being the log-likelihood of the STRUCTURE model estimates). As shown in Fig. 1, no such plateau was clearly reached in this study's panel. Evanno et al. (2005) proposed a more formal criterion leading to a more salient break in the slope of the distribution of the Ln $P(D)$ values. Figure 2 shows the results of *delta*(K), which is the mean of the second order rate of change of the Ln $P(D)$ values of a given K divided by the Ln $P(D)$ standard deviation; the curve shows an upper *delta*(K) value at 2 groups (1 structure covariate), followed by a local upper value at 16 groups (15 structure covariates). Similarly, the reliability of STRUCTURE software outputs was analyzed by calculating the distance described by Camus-Kulandaivelu et al. (2007). The neighbor joining tree of these distances indicates that until $K = 5$ groups (4 covariates), STRUCTURE software outputs are quite stable with the outputs of each group number being grouped together; only the three less likely outputs at $K = 4$ and the two less likely outputs at $K = 5$ were not clustered with the others from the same group numbers. For a group number higher than 5, the outputs were not clearly grouped according to K (Fig. 3). Hence, if the model choice relies on STRUCTURE software output reliability, then five groups (4 covariates) would be used, which corresponds to the number of groups commonly used for the panel studies (Camus-Kulandaivelu et al. 2006; Ducrocq et al. 2008).

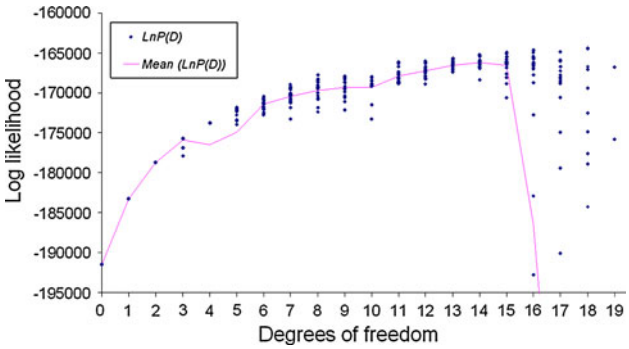

Fig. 1 Log-likelihood of the STRUCTURE software outputs as a function of the number of degrees of freedom absorbed by the structure model (*filled diamond*), and the mean of the log-likelihood of 20 STRUCTURE independent runs as a function of the number of degrees of freedom absorbed by the structure model (*solid line*). Outlier STRUCTURE runs with very low Log-likelihood values (until −2,356,000) induced a high mean decrease at 4 degrees of freedom (5 groups) and at more than 16 degrees of freedom (17 groups)

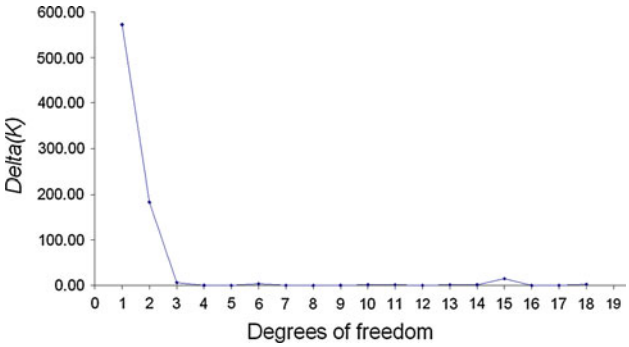

Fig. 2 Values of the *delta*(K) criterion as a function of the number of degrees of freedom absorbed by the structure model

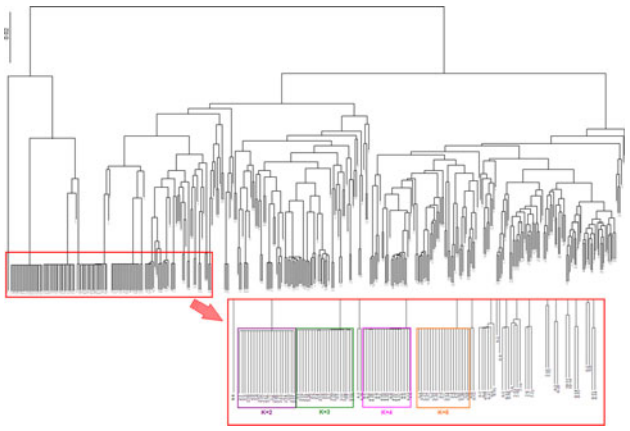

Fig. 3 Neighbor joining tree of the STRUCTURE software outputs. The pairwise Euclidian distances were computed from the predicted allele frequencies from among the 381 STRUCTURE software outputs (no genetic structure model and 20 runs of a number of ancestral groups from 2 to 20). The surrounded part of the tree was zoomed; it corresponds to the outputs that are grouped in accordance with the number of ancestral groups (K)

Principal component and multiple correspondence analyses were also performed. The number of significant principal components was set using the Tracy–Widom statistic (Tracy and Widom 1994) without correcting for linkage disequilibrium. The first four PCA and six MCA components were significant at the 5% level. These 4 PCA and 6 MCA first components explained, respectively, 8.64 and 9.80% of the whole variability described by the 641 haplotypic markers used. These percentages are quite large since 71 PCA components and 74 MCA components are required to explain 50% of the total panel variability.

A graphical criterion was then applied for the purpose of selecting all the covariates until a convex pattern occurred in the curve of eigenvalues with respect to their rank (the "elbow" criterion). This criterion clearly indicated four covariates for both PCA and MCA. However, secondary inflection points were observed on the curve (results not shown).

For the model reduction methods, this would thus mean selecting four PCA components and either four or six MCA components to control the panel structure. Nonetheless, the aim of this study was to compare the association results of these models with the remaining ones.

The above test and criteria do not take the phenotypic information into account when selecting a structure model. Zhu and Yu (2009) used the phenotypic data in a two-step model selection criterion. Similarly, the Bayesian information criteria (BIC) for the fit of the structure models to the phenotypes (Schwarz 1978) were used as a criterion to select the best one. When analyzing MFT (Fig. 4a), the minimum BIC value was reached at 22 covariates for both P and M models, at 16 covariates (17 groups) for the Q_1 models and at 14 covariates (15 groups) for Q_2 models. For TWK (Fig. 4b), the lowest BIC value was reached at four covariates for both P and M models, at seven covariates for Q_1 model and at ten covariates for Q_2 models. These BIC values are quite similar between P and M models, on the one hand, and between Q_1 and Q_2 models, on the other. When confounding all the methods, the M model with 22 and 4 covariates were the best fitting ones for, respectively, MFT and TWK. Figure 4 clearly shows that the BIC values depend on the analyzed trait and on the method used to estimate the panel structure.

Depending on the structure method and on the criteria used to choose a given number of structure groups, the selected models were different. In this study, the following models were selected: the Q_1 models with 1, 4, 7 or 16 covariates, the Q_2 models with 10 or 14 covariates, the P models with 4 or 22 covariates or the M models with 4, 6 and 22 covariates. This was why this study investigated the effects of choosing such a model on the results of association mapping tests.

Association mapping tests

The association mapping tests were carried out for the 9,945 available SNPs, using each of the models described above to control population structure. For each model, the significant loci were selected after having controlled the FDR at 10%.

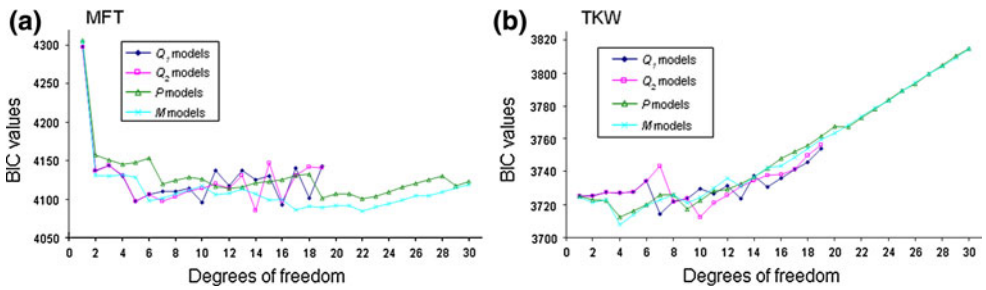

Fig. 4 BIC values of **a** male flowering time (MFT) and **b** thousand-kernel weight (TKW) as a function of the number of degrees of freedom absorbed by the structure model. The *blue* and *red lines* represent the most likely (Q_1) and second most likely (Q_2) STRUCTURE software models, respectively; the *green line* represents PCA (P) models and the *light-blue line* represents the MCA (M) models

Without taking into account the panel structure, 6,409 SNPs were significantly associated with MFT (~65% of the total number of tested SNPs) and 3,500 SNPs were significantly associated with TWK (~35% of the total number of tested SNPs). By introducing only one structure covariate (2 structure groups), the number of significant SNPs was reduced by 30% for MFT and by 90% for TWK. As shown in Fig. 5, the number of significant SNPs continued to decrease by increasing the number of structure covariates in the models until all loci were declared non-significant. This decrease in the number of significant SNPs is not a simple subtraction from the significant loci with a lower covariate number, since loci that are not significant with an n covariates model can become significant with an $n + 1$ covariates model, in spite of the fact that the total number of significant loci decreases in general. Similarly, in some cases, the number of significant loci can increase by adding covariates in the structure model; this is clearly shown in Fig. 5b with Q_1 and Q_2 models, where the number of significant loci with six covariates increases compared to the association results with five structure covariates. This observation is explained by the fit of the Q models on TWK, which is better with five covariates when compared with six covariates for both traits (see the BIC values in Fig. 4).

To evaluate the repeatability of the association results for a given number of structure groups, the proportion of significant loci, which are common between the models, was calculated. Figure 6 shows, for each number of structure covariates, the number of SNPs that were detected with at least one model (Q_1, Q_2, P or M). Within these loci, the percentage of those that were common to all the models is indicated. As expected, the number of significant loci decreases in inverse proportion to the degrees of freedom of structure covariates. Similarly, the percentage of common loci between the methods decreases as the number of structure covariates increases. For MFT, a small increase in the percentage of common loci was observed with eight and nine covariates, but it continued to decrease just after the nine covariates model (see Fig. 6a). These observations are quite different for TWK, because no significant loci were observed at more than seven covariates (Fig. 6b). Common loci were only observed for less than four structure covariates.

The proportion of common loci for the models compared two by two was then calculated, for every number of structure covariates. When analyzing MFT, Q_1 and Q_2 had almost 100% of common loci for all the structure models with six covariates at most. This is not surprising because these models are very similar according to the distance described by Camus-Kulandaivelu et al. (2007). With more than seven covariates, the percentage of common loci fluctuates depending on the similarity between Q_1 and Q_2 matrices. The percentage of common loci in the remaining

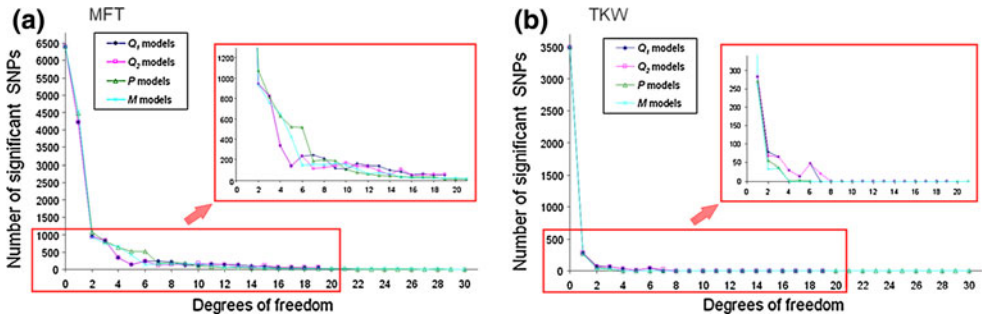

Fig. 5 Number of SNPs significantly associated with **a** male flowering time (MFT) and **b** thousand-kernel weight (TKW) as a function of the number of degrees of freedom absorbed by the structure model. The *blue* and *red lines* represent the most likely (Q_1) and second most likely (Q_2) STRUCTURE software models, respectively; the *green line* represents PCA (P) models and the *light-blue line* represents the MCA (M) models. The surrounded parts correspond to a zoom of the curves

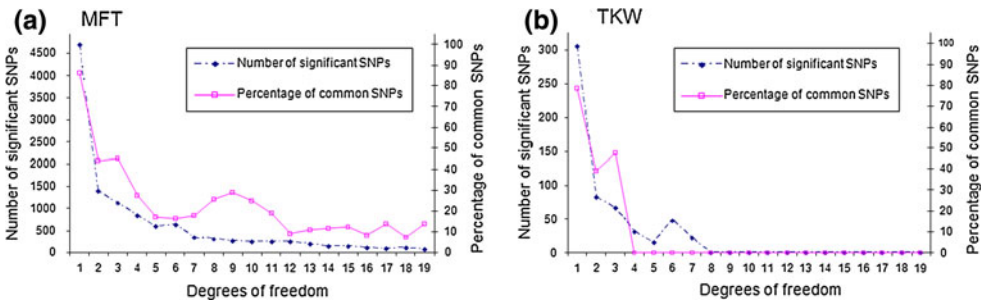

Fig. 6 Total number of significant SNPs over the four structure models tested (*solid line curve*) and percentage of SNPs common to all of them (*dotted line curve*) as a function of the number of degrees of freedom absorbed by the structure model

comparisons decreases as the number of structure groups in the compared models increases. With a high number of structure covariates (more than 24 covariates), no significant loci for MFT were common for the *P* and *M* models (see Table 2 on supplemental data). If the percentage of common loci is used as a criterion of model similarity, it appears that *P* and *M* models, on the one hand, and the STRUCTURE software models (Q_1 and Q_2), on the other, are the most similar ones. Comparable results were obtained for TWK in spite of no significant loci being detected when using structure models with more than seven covariates.

The association results of all the models tested were then compared. The more the structure models diverged according to the group numbers, the less they carried common significant loci.

Finally, it is interesting to note that 4,544 SNPs were significantly associated with MFT with at least one of the structure models that would have been selected with the different criteria (BIC values, log-likelihood plateau and *delta*(*K*) criteria for STRUCTURE outputs; "elbow" criterion and Tracy–Widom statistic for PCA and MCA). Within these SNPs, only three were common to all of them. For TWK, no significant loci were detected with the models that would have been selected with the BIC values and no significant associations were detected with the *P* and *M* models, which would have been selected with the different criteria.

Association meta-analysis

Instead of selecting a given structure model relying on a particular criteria, the entire association mapping results were used in a model summing approach. This involved first eliminating, for each trait, the models for which no significant loci were observed. A total of 3 and 77 models were thus eliminated for MFT and TKW, respectively. For the remaining models, 5,296 and 317 SNPs were significantly associated with MFT and TWK, respectively. They account for, respectively, 53.25 and 3.18% of the tested SNPs. These numbers of significant SNPs are too high and suggest that the results will include false positives. To select the robust loci, the number of models for which a given SNP was significantly associated with the analyzed trait was simply counted and ordered, starting with the SNPs that were detected with the highest number of models. The loci that were found to be significant with at least a threshold of model numbers were retained. This threshold was set according to the aimed robustness. In this case for example, 67 and 35 were selected for MFT and TWK, respectively, for at least 50% of the retained models and 21 SNPs were selected for both traits for at least 75% of the retained models (Table 2).

Table 2 Frequency table of the number of models in which a given SNP is significant for male flowering time (MFT) and thousand-kernel weight (TKW)

Class intervals	Number of significant SNPs	
	MFT	TKW
90–98 models	–	–
86–89 models	2	–
81–85 models	1	–
76–80 models	8	–
71–75 models	3	–
66–70 models	7	–
61–65 models	9	–
56–60 models	4	–
51–55 models	14	–
46–50 models	18	–
41–45 models	16	–
36–40 models	23	–
31–35 models	31	–
26–30 models	34	–
21–25 models	45	–
16–20 models	135	21
11–15 models	286	14
06–10 models	603	48
01–05 models	4,055	234

The class interval is every five models, except for the extreme classes

For the meta-analysis approaches, Fisher's inverse chi-square (ICS) method is widely used. It combines the results of independent tests by summing the logarithm of the p-values (Hedges and Olkin 1985). The ICS values in this study were calculated for the retained structure models, but were not tested because the models were not independent. The SNPs were ordered in accordance with the ICS statistic or in accordance with the model summing approach used. The ranks of the selected SNPs with at least 50% of the retained models were linearly correlated between the ICS method and the model summing approach (the square of the Pearson coefficient of correlation was $R^2 = 0.69$ and 0.74 for MFT and TWK, respectively).

Discussion

There are numerous methods for estimating the structure of a diversity panel. For each one, several approaches, relying mostly on ad hoc criteria, were proposed to help select the suitable number of ancestral groups to use in the association mapping tests as covariate fixed effects. However, because of the complexity of the panel stratification, the "true" group number was generally not known and these ad hoc criteria did not enable an obvious choice of structure

model. The model used in such cases may not be suitable. The goal was therefore to evaluate the effect of the number of structure groups on the results of association mapping tests. Different methods were also compared (STRUCTURE software, PCA and MCA).

Structure estimates

To carry out the structure estimates, a set of haplotypic markers built up from the concatenation of SNPs that are physically mapped in the same genic regions were used. These haplotypic markers should be more informative than the SNPs in spite of the fact that Hamblin et al. (2007) found only a small improvement in the measurement of genetic distances when converting SNPs to haplotypic markers.

The major difficulty when a high number of markers is used to estimate the panel structures is the presence of redundant markers (due to physical linkage disequilibrium), which introduces a bias (Price et al. 2006). Our study distinguishes two scales of possible redundancy: within a genic region—because the amplicons do not carry the same number of genotyped SNPs—and within the whole genome. We avoided intra-genic redundancy by concatenating the SNPs and utilizing haplotypic markers. These multiallelic markers carry a high number of alleles if they result from a high number of SNPs. They will thus have more weight on the structure estimates, but will not bias them. Within each chromosome, marker redundancy was avoided by selecting them so that they are genetically mapped at a distance of more than 1 cM from the others. Assuming that recombination hotspots within biparental populations and diversity panels are similar, this 1-cM distance would avoid the marker redundancy in all parts of the genome. In addition, the extent of linkage disequilibrium (LD) in this panel is very low, and the loci mapped at intervals of 1 cM will be in linkage disequilibrium only due to the panel structure. This low extent of linkage disequilibrium was also observed in similar maize panels (Remington et al. 2001; Tenaillon et al. 2001).

The Bayesian admixture model implemented in STRUCTURE software (Pritchard et al. 2000) was first run with 20 repetitions of a group number (K) varying from 1 to 20. To select the "suitable" K, the plateau criterion proposed by Pritchard et al. (2007) was applied. Such a plateau of the log-likelihood values with respect to K was not clearly reached in the study. Similar observations have been reported in numerous studies dealing with different plant species (Lia et al. 2009; Lu et al. 2009; Abdurakhmonov et al. 2009; Wang et al. 2008; Camus-Kulandaivelu et al. 2007; Heuertz et al. 2004). In the present material, this could be explained by the presence of several levels of stratification since the panel represents the whole genetic diversity of the temperate maize material and includes some related materials. Therefore, adding new groups to the structure model yields a continuous improvement in the description of the panel structure (Camus-Kulandaivelu et al. 2006).

The *delta*(K) criterion suggested by Evanno et al. (2005) gave the highest value at two groups (1 structure covariate). This method is known to give rise to the first structure level (Lia et al. 2009), which appears to principally discriminate, in the present study panel, the European flint and their American Northern flint progenitors from the American dent lines. Nevertheless, we are firmly of the opinion that using only one covariate in the association model would not fully control the stratification.

The reliability criterion introduced by Camus-Kulandaivelu et al. (2007) was also applied. This relies on a distance between the STRUCTURE software outputs, which evaluates the similarity of the Q matrices. The neighbor joining tree of these distances indicates a consistent output clustering with respect to K up to five groups (4 structure covariates). In the whole panel from which the subset of 342 lines was taken, Camus-Kulandaivelu et al. (2007) showed a consistent clustering in accordance with the group number at only $K = 2$ and $K = 3$; the remaining group numbers did not show any clear patent of gathering in accordance with K. We thus improved the reliability of the structure outputs in the present analysis. This improvement is probably essentially due to the high number of loci used; furthermore, the present loci were different, more iterations were used (100,000 instead of 50,000 in the study of Camus-Kulandaivelu et al. 2007) and the panel size was reduced by discarding 33 lines that had an unexpected high frequency of heterozygote loci.

The panel structure was also estimated by principal component and multiple correspondence analyses. When projecting the panel lines onto the first axes, the plots of PCA and MCA were quite similar (graphics not shown). This is not really surprising as the two methods are diagonalization of two particular binary tables A and A'. The only differences between the two methods are (1) the number of alleles used to perform the analysis and (2) the column weights. MCA was performed using all the alleles generated by the 641 loci and each allele weight corresponded to its frequency in the panel. This method thus gives more weight to the rare alleles. In the present PCA, one allele per locus was suppressed because the arrangement of a given allele can be inferred from the arrangement of the other ones for the same locus. This avoided the dependency within each locus. The allele weights in the present PCA correspond to their standard deviation thus giving more weight to the common alleles.

The Tracy–Widom statistic allowed for selection of four PCA and six MCA components. However, the results of this statistic were ambiguous in the case of MCA because of the dependency between the alleles of the same locus. With the "elbow" criterion, four PCA and MCA components were retained. This criterion aims to select all the components before the first inflexion point in the curve of eigenvalues according to their rank. However, the principal inflexion points were not very prominent and secondary ones were observed on the curves. This observation is in agreement with the log-likelihoods of the STRUCTURE software outputs, which suggest that the more the structure covariates introduced in the model, the better is the panel structure described.

The goodness of fit of the model to the phenotypic data was integrated in the selection of structure model by calculating the BIC. The BIC values were calculated for all of the models tested with both MFT and TWK. The model in which the best BIC value is reached depends greatly on the method of structure estimate and on the analyzed trait: the more the trait is correlated to the panel structure, the higher is the number of covariates required to control the structure.

Association mapping tests

The criteria used to select a structure model for the different structure methods did not converge to the same number of groups. In addition to the structure method, the selected model depends on the criteria used to select it and on the analyzed trait. This variation in the model choice is principally caused by the complexity of the study's panel structure, which leads to no perfect matching between any model and the panel structure. The association tests were thus carried out with all the computed models.

A two-step association test was carried out. The phenotypes were first corrected for the panel structure and the SNP effects in the adjusted phenotypes were then tested. This two-step model is asymptotically similar to the one-step one. In mixed association models, which integrate both fixed and random structure effects, Stich et al. (2008) did not observe a large increase of the type one error in their two-step model in comparison with the one-step tests. Furthermore, the present study tested a high number of SNPs and the two-step model was the more practical and less time consuming, since the phenotypes were only adjusted once. The two-step model may also be computationally more stable if the effects of the tested loci are collinear to those of the structure covariates.

The main concern was to investigate the impact of the number of structure groups on the association mapping results of a real data set. Also, given that the panel structure had been estimated with different methods, the association results with the different estimates were then compared.

The result shows a significant effect of the structure models on the association mapping tests. This effect varies depending on the analyzed traits. The more a trait is correlated to the panel structure, the higher is the number of structure covariates required to control the false positives and the higher is the number of covariates required to reach an over-structuring model in which all the associated loci vanish. Likewise, the structure models that would be selected with the criteria used gave quite different association results and only a very small proportion of the associations were common to all of them. This was principally due to the false positives and to the false negatives.

Two kinds of false positives should be distinguished: (1) false positives that are due to the panel structure and (2) statistically false positives due to the high number of tests carried out. The former were controlled by introducing structure covariates in the association model and the latter were limited by controlling the false discovery rate at 10%.

Meta-analysis of association mapping results

It was quite difficult to select a structure model with the panel studied; furthermore, the association results depend strongly on the model chosen. Due to the complex panel structure, we are of the firm opinion that it will be quite unlikely to find one model, which would fully describe it; instead, several models will be close to the "true" one. Consequently, instead of trying to identify the best model according to a given criterion, the aim was to make use of the results for all of the models tested. This approach highlights the causal loci that are not correlated with these first levels of panel structure. It is quite easy to justify for dimensional reduction methods (PCA and MCA), because a structure model with n covariates can be seen as the model with $(n - 1)$ covariates for which a supplementary structure dimension is added. Therefore, a given locus is significant and stays significant until it correlates with a structure covariate introduced into the model. Two kinds of correlations can be observed: (1) a collinearity between a structure dimension and a false positive or (2) a correlation between a causal locus and a structure covariate (false negative). The rank of the structure covariates to which a given loci is correlated depends on the loci and on the correlation of the analyzed trait to the panel structure. However, most of the first SNPs to become non-significant are false positives, correlated with the first covariates that describe the panel structure. Among the SNPs correlated with the first covariates, causal loci may be observed but it seems very difficult to differentiate them from the false positives if no *prior* information is available (Ducrocq et al. 2008). In addition, the true positives that are

highlighted by the meta-analysis approach would be less affected by the methods used to compute the structure estimates and they would be generally correlated to structure covariates that have high ranks (over-structuring dimensions).

The repeatability of the association results for all of the models helps when selecting true positives, because the higher the rank of the structure dimension to which a given locus is correlated, the higher is its repeatability and the higher the likelihood of it being a true positive. Thus, by counting the number of models for which a given locus is significantly associated with the analyzed trait and by setting a minimal threshold of model number for which it should be detected, the robust loci will be selected. The user sets this threshold according to the desired statistical power.

The only inconvenience of this summing approach is that it requires testing all the structure models and is thus time consuming. However, testing all the SNP associations with the entire structure models took less time compared to Bayesian estimates of the panel structure and, by using the two-step model, the analysis time was reduced.

By combining in this way the association results of several structure models, robust SNPs were picked up. The most often-repeated SNP, for association tests with MFT, was detected with 89 models. It is located in the MADS-box genes ZMM14, which is specifically expressed in the upper floret of maize ear spikelets. It could be involved in specifying the identity of the upper floret and may have an early function in the spikelet meristem (Cacharrón et al. 1999). The MADS-box genes encode a family of transcription factors, which are known to affect the flower organs. In maize, several genes belonging to this family were shown to be involved in the development of maize sexual organs (Danilevskaya et al. 2008; Heuer et al. 2000). Two other SNPs from ZMM14 were detected with 73 and 68 models; they were mapped at, respectively, 109 and 261 bp from the first one and were in linkage disequilibrium with it ($D' \approx 0.98$ and $R^2 \approx 0.70$).

Similarly, for the association tests with TWK, an interesting SNP was detected with 19 of the 20 retained models. It is mapped in the gene of *Granule-bound starch synthase-I* (*GBSS-I*), which is also known as *waxy1* (Nelson and Rines 1962). *GBSS-I* is involved in the synthesis of amylose, which accounts for 25% of the total starch contained in the endosperm (Hannah 2007). Starch accounts for 73% of the kernel's total weight, and the genes involved in starch synthesis are critical to grain yield and quality (Buckler and Stevens 2005). Two other SNPs from *GBSS-I* were detected with 13 models; they were located at 400 and 414 bp from the first one and were in linkage disequilibrium with it ($D' \approx 0.98$ and $R^2 \approx 0.34$).

It is also interesting to note that the result of this summing-model approach were in accordance with the results of the ICS statistic in spite of the fact that the loci ranks were not totally conserved in the two approaches. The observed inversions in the SNP ranks are not surprising because the ICS statistic favored the loci having the lowest *P* values.

Fixed and mixed structure effects in the association mapping models

Several association mapping studies appealing for different structure models were published. Yu et al. (2006) introduced a unified mixed association model accounting for both fixed and random structure effects. The fixed effects were estimated with the Bayesian STRUCTURE software, while the random effects were approximates of the identity by descent between two individuals (Loiselle et al. 1995; Ritland 1996). They showed that integrating both effects generally improved the model fit to the phenotypic variation of the analyzed maize quantitative traits. Similarly, Zhao et al. (2007) applied the above approach within an *Arabidopsis thaliana* diversity panel. In addition, PCA and a matrix based on shared alleles (identity by state) were used for the fixed and the random structure effects, respectively. They yielded similar conclusions to Yu et al. (2006) about the mixed models; these models were the best in terms of reducing the false positive rate and maintaining statistical power. Using a kinship matrix estimated by REML, it was also shown that mixed models are appropriate for association mapping in a winter wheat panel (Stich et al. 2008) and in rapeseed, potato, sugar beet, maize and *Arabidopsis thaliana* panels (Stich and Melchinger 2009).

All these studies found that mixed association models were suitable. However, the present study focused on the fixed structure covariates due to their main effect in the analyzed panel. This panel represents a large maize diversity with several heterotic groups and using only a relatedness matrix would not be enough to correctly account for the different origins of the inbred lines. Moreover, with no complete pedigree information, selecting an appropriate matrix to model the genetic covariance between the inbred lines is not an easy task. The commonly used kinship estimates based on molecular markers (Loiselle et al. 1995; Ritland 1996) were described in a population genetic context; their initial assumptions are not met in panels of inbred lines (Maenhout et al. 2009). However, Loiselle et al. (1995) kinship estimator was calculated in the analyzed material. The negative values reached up to −0.20. After replacing theses negative values by 0, the matrix was not positive semi-definite; it thus required further statistical transformation to the closest positive

semi-definite matrix to avoid the different difficulties reported by Maenhout et al.(2009). The consequences of such transformation in the genetic covariance modeling are not well known. Similarly, the use of a matrix based on the identity by state does not guarantee a common ancestral origin of the inbred lines.

Selecting an appropriate kinship estimator is important when using a mixed model approach for association mapping tests, which is beyond the scope of the present study. However, association mapping tests were carried out with different mixed models. Preliminary results indicated an influence of both the fixed and random structure effects on the association mapping tests. By changing the matrix used to model the genetic covariance and by changing the number of covariates used to control the fixed structure effects, the association results changed (data not shown). Therefore, with an appropriate modeling of the genetic covariance, combining the association results, using different fixed structure effects, will improve the pinpointing of robust loci.

Acknowledgments We thank the associate editor and three anonymous reviewers for their comments on the manuscript. This work was conducted with the financial support of the French "Association Nationale de la Recherche et de la Technologie" (ANRT).

References

Abdurakhmonov IY, Saha S, Jenkins JN, Buriev ZT, Shermatov SE, Scheffler BE, Pepper AE, Yu JZ, Kohel RJ, Abdukarimov A (2009) Linkage disequilibrium based association mapping of fiber quality traits in *G. hirsutum* L. variety germplasm. Genetica 136:401–417

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc 57:289–300

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

Cacharrón J, Saedler H, Theißen G (1999) Expression of MADS box genes ZMM8 and ZMM14 during inflorescence development of *Zea mays* discriminates between the upper and the lower floret of each spikelet. Dev Genes Evol 209:411–420

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

Camus-Kulandaivelu L, Veyrieras JB, Gouesnard B, Charcosset A, Manicacci D (2007) Evaluating the reliability of structure outputs in case of relatedness between individuals. Crop Sci 47:887–890

Cavanagh C, Morell M, Mackay I, Powell W (2008) From mutations to MAGIC: resources for gene discovery, validation and delivery in crop plants. Curr Opin Plant Biol 11:215–221

Chessel D, Dufour AB, Thioulouse J (2004) The ade4 package-I—one-table methods. R News 4:5–10

Corander J, Waldmann P, Sillanpää MJ (2003) Bayesian analysis of genetic differentiation between populations. Genetics 163: 367–374

Danilevskaya ON, Meng X, Selinger DA, Deschamps S, Hermon P, Vansant G, Gupta R, Ananiev EV, Muszynski MG (2008) Involvement of the MADS-Box Gene ZMM4 in floral induction and inflorescence development in maize. Plant Physiol 147:2054–2069

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

Ducrocq S, Madur D, Veyrieras JB, Camus-Kulandaivelu L, Kloiber-Maitz M, Presterl T, Ouzunova M, Domenica M, Charcosset A (2008) Key impact of Vgt1 on flowering time adaptation in maize: evidence from association mapping and ecogeographical information. Genetics 178:2433–2437

Ersoz ES, Yu J, Buckler ES (2007) Applications of linkage disequilibrium and association mapping in crop plants. Genomics-assisted crop improvement. Springer, The Netherlands

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

Falque M, Décousset L, Dervins D, Jacob AM, Joets J, Martinant JP, Raffoux X, Ribière N, Ridel C, Samson D, Charcosset A, Murigneux A (2005) Linkage mapping of 1454 new maize candidate gene loci. Genetics 170:1957–1966

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

Flint-Garcia SA, Thuillet AC, Yu J, Pressoir G, Romero SM, Sharon EM, 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 2(12):e1367

Hannah LC (2007) Starch formation in the cereal endosperm. Endosperm: developmental and molecular biology, vol 8. Springer, Berlin/Heidelberg, pp 179–193

Hedge VL, Olkin I (1985) Statistical methods for meta-analysis. Academic Press, London

Heuer S, Lörz H, Dresselhaus T (2000) The MADS box gene ZmMADS 2 is specifically expressed in maize pollen and during maize pollen tube growth. Sex Plant Reprod 13:21–27

Heuertz M, Hausman JF, Hardy OJ, Vendramin GG, Frascaria-Lacoste N, Vekemans X (2004) Nuclear microsatellites reveal contrasting patterns of genetic structure between Western and Southeastern European populations of the common ash (*Fraxinus excelsior* L.). Evolution 58(5):976–988

Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K (2002) A comprehensive review of genetic association studies. Genet Med 4(2):45–61

Khoury MJ, Bertram L, Boffetta P, Butterworth AS, Chanock SJ, Dolan SM, Fortier I, Garcia-Closas M, Gwinn M, Higgins JPT, Janssens ACJW, Ostell J, Owen RP, Pagon RA, Rebbeck TR, Rothman N, Bernstein JL, Burton PR, Campbell H, Chockalingam A, Furberg H, Little J, O'Brien TR, Seminara D, Vineis P, Winn DM, Yu W, Ioannidis JPA (2009) Genome-wide association studies, field synopses, and the development of the knowledge base on genetic variation and human diseases. Am J Epidemiol 170(3):269–279

Lia VV, Poggio L, Confalonieri VA (2009) Microsatellite variation in maize landraces from Northwestern Argentina: genetic diversity, population structure and racial affiliations. Theor Appl Genet 119:1053–1067

Lin CH, Yeakley JM, McDaniel TK, Shen R (2008) Medium- to high-throughput SNP genotyping using VeraCode Microbeads in DNA and RNA profiling in human blood. Methods Protoc 496:129–142

Loiselle BA, Sork VL, Nason J, Graham G (1995) Spatial genetic structure of a tropical understory shrub, *Psychotria officinalis* (Rubiaceae). Am J Bot 82:1420–1425

Lu Y, Yan J, Guimarães CT, Taba S, Hao Z, Gao S, Chen S, Li J, Zhang S, Vivek BS, Magorokosho C, Mugo S, Makumbi D, Parentoni SN, Shah T, Rong T, Crouch JH, Xu Y (2009) Molecular characterization of global maize breeding germplasm based on genome-wide single nucleotide polymorphisms. Theor Appl Genet 120:93–115

Maenhout S, De Baets B, Haesaert G (2009) Marker-based estimation of the coefficient of coancestry in hybrid breeding programmes. Theor Appl Genet 118:1181–1192

Manicacci D, Camus-Kulandaivelu L, Fourmann M, Arar C, Barrault S, Rousselet A, Feminias N, Consoli L, Francès L, Méchin V, Murigneux A, Prioul J, Charcosset A, Damerval C (2009) Epistatic interactions between Opaque2 transcriptional activator and its target gene CyPPDK1 control kernel trait variation in maize. Plant Physiol 150:506–520

Matsuoka Y, Vigouroux Y, Goodman MM, Sanchez JG, Buckler E, Doebley J (2002) A single domestication for maize shown by multilocus microsatellite genotyping. PNAS 99(9):6080–6084

McVean G (2009) A genealogical interpretation of principal components analysis. PLoS Genet 5(10):e1000686

Nelson OE, Rines HW (1962) The enzymatic deficiency in the waxy mutant of maize. Biochem Biophys Res Commun 9:297–300

Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PloS Genet 2(12):e190

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38(8):904–909

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

Pritchard JK, Wen X, Falush D (2007) Documentation for structure software: version 2.2. http://pritch.bsd.uchicago.edu/software

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. PNAS 98(20):11479–11484

Ritland K (1996) Estimators for pairwise relatedness and inbreeding coefficients. Genet Res 67:175–186

Schwarz G (1978) Estimating dimension of a model. Ann Stat 6:461–464

Shendure J, Ji H (2008) Next-generation DNA sequencing. Nat Biotechnol 26:1135–1145

Steemers FJ, Gunderson KL (2007) Whole genome genotyping technologies on the BeadArrayTM platform. Biotechnol J 2:41–49

Stich B, Melchinger AE (2009) Comparison of mixed-model approaches for association mapping in rapeseed, potato, sugar beet, maize, and Arabidopsis. BMC Genomics 10:94

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

Storey JD (2002) A direct approach to false discovery rates. J R Stat Soc Ser B 64:479–498

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.). PNAS 98(16):9161–9166

Tenenhaus M, Young F (1985) An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. Psychometrika 50:91–119

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

Tracy CA, Widom H (1994) Level spacing distributions and the airy kernel. Commun Math Phys 159:151–174

Wang R, Yu Y, Zhao J, Shi Y, Song Y, Wang T, 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

Wang J, Lin M, Crenshaw A, Hutchinson A, Hicks B, Yeager M, Berndt S, Huang WY, Hayes RB, Chanock SJ, Jones RC, Ramakrishnan R (2009) High-throughput single nucleotide polymorphism genotyping using nanofluidic dynamic arrays. BMC Genomics 10:561

Yu J, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, 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 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(1):e4

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

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