Theor Appl Genet (2010) 121:1171–1185
DOI 10.1007/s00122-010-1380-z

ORIGINAL PAPER

Variability of grain quality in sorghum: association with polymorphism in *Sh2*, *Bt2*, *SssI*, *Ae1*, *Wx* and *O2*

L. F. de Alencar Figueiredo · B. Sine · J. Chantereau · C. Mestres · G. Fliedel · J.-F. Rami · J.-C. Glaszmann · M. Deu · B. Courtois

Received: 16 October 2009 / Accepted: 3 June 2010 / Published online: 22 June 2010

Abstract To ensure food security in Africa and Asia, developing sorghum varieties with grain quality that matches consumer demand is a major breeding objective that requires a better understanding of the genetic control of grain quality traits. The objective of this targeted association study was to assess whether the polymorphism detected in six genes involved in synthesis pathways of starch (*Sh2*, *Bt2*, *SssI*, *Ae1*, and *Wx*) or grain storage proteins (*O2*) could explain the phenotypic variability of six grain quality traits [amylose content (AM), protein content (PR), lipid content (LI), hardness (HD), endosperm texture (ET), peak gelatinization temperature (PGT)], two yield component traits [thousand grain weight (TGW) and number of grains per panicle (NBG)], and yield itself (YLD). We used a core collection of 195 accessions which had been previously phenotyped and for which polymorphic sites had been identified in sequenced segments of the six genes. The associations between gene polymorphism and phenotypic traits were analyzed with Tassel. The percentages of admixture of each accession, estimated using 60 RFLP probes, were used as cofactors in the analyses, decreasing the proportion of false-positive tests (70%) due to population structure. The significant associations observed matched generally well the role of the enzymes encoded by the genes known to determine starch amount or type. *Sh2*, *Bt2*, *Ae1*, and *Wx* were associated with TGW. *SssI* and *Ae1* were associated with PGT, a trait influenced by amylopectin amount. *Sh2* was associated with AM while *Wx* was not, possibly because of the absence of waxy accessions in our collection. *O2* and *Wx* were associated with HD and ET. No association was found between *O2* and PR. These results were consistent with QTL or association data in sorghum and in orthologous zones of maize. This study represents the first targeted association mapping study for grain quality in sorghum and paves the way for marker-aided selection.

Communicated by J. Yu.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-010-1380-z) contains supplementary material, which is available to authorized users.

L. F. de Alencar Figueiredo · J.-F. Rami · J.-C. Glaszmann · M. Deu · B. Courtois (✉)
Cirad, UMR DAP, TA A96/3, Avenue Agropolis,
34398 Montpellier, France
e-mail: brigitte.courtois@cirad.fr

L. F. de Alencar Figueiredo
Dept. Biologia Celular, Instituto de Ciências Biológicas,
Universidade de Brasília, Brasília 70910-900, Brazil

B. Sine
CERAAS, Thiès, Senegal

J. Chantereau
Cirad, UPR AIVA, 34398 Montpellier, France

C. Mestres · G. Fliedel
Cirad, UPR QualiSud, 34398 Montpellier, France

Introduction

In Africa and Asia, cultivated sorghum (*Sorghum bicolor* ssp. *bicolor*) is mostly used for human consumption and plays an important role in the food security of populations in arid and semi-arid zones (Belton and Taylor 2004). Sorghum is used as whole grain or as flour for a broad range of dishes and beverages (Fliedel et al. 1996). The development of varieties with grain quality to match local consumer demand has become an important target for sorghum breeding programs (Chantereau et al. 1997).

Grain quality in sorghum destined for the food market is partly determined by starch properties, notably amylose content (AM) and peak gelatinization temperature (PGT), while grain filling and, consequently, thousand grain weight (TGW) are affected by the accumulation of starch. Other important traits for a large number of preparations are endosperm texture (ET), kernel hardness (HD), and, to a lesser extent, protein (PR) and lipid (LI) contents (Fliedel 1994; Aboubacar and Hamaker 1999; Yetneberk et al. 2004).

The development of DNA markers and statistical methods to detect markers linked to quantitative trait loci (QTLs) in mapping populations has provided insights into the genetics of quality traits in sorghum. Rami et al. (1998) and Rami (1999) positioned QTLs and protein quantity loci (PQLs) for grain quality and yield component traits on two sorghum populations (Supplementary Fig. 1). Murray et al. (2008), focusing on the use of sorghum sugars for biofuel production, also detected QTLs for grain composition that, for some of them, co-localized with the QTLs identified by Rami et al. (1998).

To date, however, QTL detection studies for grain quality traits in sorghum have been limited to these three studies. Moreover, with commonly used mapping population sizes, the precision in localization of the QTLs is limited, with confidence intervals generally in the order of 10–30 cM. Both better resolution and complementary information can be obtained through association mapping in populations in which linkage disequilibrium (LD) due to physical proximity between loci has been broken down by recombination and which segregate for a much larger number of alleles (Flint-Garcia et al. 2003). In situations where LD spans only a few hundred base pairs, association mapping not only enables identification of the gene responsible for the phenotypic variability, but also of the functional mutation within the gene. The first studies in crops were conducted in maize, revealing the unexpected importance of *Dwarf8* polymorphism in the determination of maize flowering time in various populations (Thornsberry et al. 2001; Andersen et al. 2005; Camus-Kulandaivelu et al. 2006). Additional studies have been conducted using different approaches in crops as diverse as barley (Kraakman et al. 2004), potato (Gebhardt et al. 2004), *Lollium perenne* (Skot et al. 2005), wheat (Breseghello and Sorrells 2006), and rice (Agrama et al. 2007; Tian et al. 2009). In sorghum, association studies between dwarfing genes and plant height (Brown et al. 2008), and between enzymes of the sucrose pathways and plant height and brix (Murray et al. 2009) have recently been conducted, but none targeted grain quality.

In association studies, false positives can be observed if the association panel is structured. The structure is due to factors linked to species history such as bottlenecks, selection, drift, and admixture (Flint-Garcia et al. 2003). Several statistical methods have been developed to eliminate or decrease the effect of population structure (Devlin and Roeder 1999; Pritchard et al. 2000a; Price et al. 2006; Yu et al. 2006; Gao et al. 2007).

In crops, the association approach has been used in two ways: a genome-wide approach that enables positioning of all QTLs controlling a trait but requires a very large genotyping effort when LD decays fast (e.g. Breseghello and Sorrells 2006 for grain quality traits in wheat), and a targeted gene approach that focuses on specific areas where preliminary analyses enabled identification of segments carrying QTLs supported by co-localization of candidate genes (e.g. Wilson et al. 2004 for grain quality traits in maize). In the targeted approach, a few markers, generally single nucleotide polymorphisms (SNPs) or insertions or deletions (Indels), can be used to assess the association between gene polymorphism and the trait(s) of interest. The use of marker haplotype classes can further improve the detection power of the method and may provide a predictive value in uncharacterized germplasm (Buntjer et al. 2005).

Sorghum bicolor ssp. *bicolor* is a dominantly autogamous species (6–30% outcrossing) that was domesticated 5,000 (Doggett 1988) to 8,000 years ago (Wendorf et al. 1992). Hamblin et al. (2005) demonstrated that, in sorghum, LD could extend up to 100 kb but had largely decayed by 15 kb, meaning that targeted association mapping is possible in this species.

Several genes are known to be involved in the genetic control of grain quality in cereals, notably in maize (Wilson et al. 2004). Genes such as *Shrunken 2* (*Sh2*), *Brittle 2* (*Bt2*), *Soluble starch synthase I* (*SssI) Amylose extender 1* (*Ae1*), and *Waxy* (*Wx*), represent three enzyme classes out of four that are responsible for starch synthesis in the grain of higher plants. *Sh2* encodes the large subunit of ADP-glucose phosphorylase (AGPase) while *Bt2* encodes the small subunit of AGPase, both being involved in starch synthesis (Schultz and Juvik 2004). *Ae1* codes for a branching enzyme that, together with starch synthases such as the one coded by *SssI* and debranching enzymes, plays an essential role in the biosynthesis of amylopectin (Myers et al. 2000; Nishi et al. 2001). *Wx* is responsible for amylose synthesis in the grain endosperm. The respective positions of these genes in the starch biosynthesis pathway, shown in Fig. 1, are well known (Schultz and Juvik 2004). The *Opaque2* (*O2*) gene is a transcriptional factor involved in the regulation of protein storage (Pirovano et al. 1994). The way *O2* acts to determine zein content and kernel hardness in maize is well established (Motto et al. 1996).

Sorghum genes homologous to the six genes *O2*, *Sh2*, *Ae1*, *Bt2*, *SssI*, and *Wx* have been localized on sorghum chromosomes 2, 3, 4, 7, 10, and 10, respectively, through

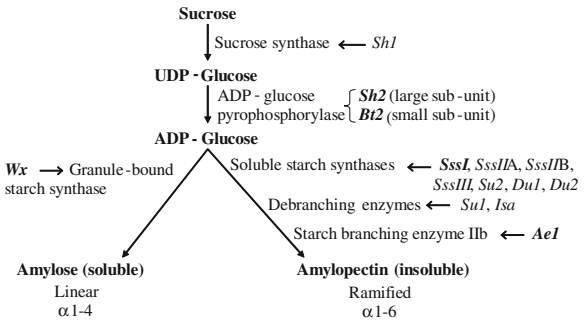

Fig. 1 Starch metabolic pathway. Genes studied are in *bold*. *Sh* shrunken, *Bt* brittle, *Sss* soluble starch synthase, *Su* sugary, *Du* dull, *Isa* isoamylase-type, *Ae* amylose extender, *Wx* waxy

direct genetic mapping for *O2*, *Wx*, and *Ae1*, and by Blast query of the gene sequences on the sorghum genome for *Sh2*, *Bt2* and *SssI* (Supplementary Fig. 1). In sorghum, some of these genes co-localize with the QTLs or PQLs detected by Rami et al. (1998), Rami (1999) and Rami (unpublished results): *O2* with QTLs for AM, ET, HD, albumins and prolamins; *Sh2* with QTLs for TGW, albumins and prolamins; *Bt2* with QTLs for ET, HD, PR, albumins, NBG and TGW; *Wx* and *SssI* with QTLs for PR and albumins. *Ae1* was located not far from QTLs for AM, PR and ET. These genes are reasonable candidates for a targeted association approach.

With the goal of conducting an association study, we recently sequenced a total of 19 fragments of these six genes (*Sh2*, *Bt2*, *SssI*, *Ae1*, *Wx* and *O2*) in a core collection of cultivated sorghum (de Alencar Figueiredo et al. 2008) shown to be representative of the species diversity (Deu et al. 2006). This core collection was also phenotypically characterized for the main food grain quality traits (AM, PR, LI, HD, ET, and PGT; de Alencar Figueiredo et al., in preparation) and for yield and yield components (Sine 2003).

Thus, the objective of the present study was to determine whether the polymorphism in genes involved in the starch (*Sh2*, *Bt2*, *SssI*, *Ae1*, and *Wx*) and grain storage protein (*O2*) synthesis pathways explained the variability of grain quality traits and grain-related yield components observed in a core collection of cultivated sorghum.

Materials and methods

Material

The core collection used in this study was designed to represent the world's cultivated sorghum landraces, with sampling based on races according to the classification proposed by Harlan and de Wet (1972), latitude of origin, response to day length and production system. The sampling methodology is detailed in Deu et al. (2006). The set used for quality evaluation was composed of 195 sorghum accessions originating from 39 countries and belonging to the 5 basic and 10 intermediate races. They were maintained as inbred samples and had a very low level of heterozygosity.

Methods

Quality trait measurement

The samples used in the grain quality analyses were harvested from an irrigated trial conducted in the 2002–2003 dry season at CERAAS (Centre d'Etude Régional pour l'Amélioration de l'Adaptation à la Sécheresse) in Senegal. A dry season trial was chosen to ensure the most homogeneous conditions during grain filling, enhancing the expression of genetic differences in grain quality among varieties. Varieties were sown in an augmented design with 10 blocks and 27 plots per block, with 6 varieties used as controls in each block. Local varieties were added to complete the experimental design. Data on thousand grain weight (TGW), number of grains per panicle (NBG) and grain yield per plant (YLD) were also collected during this trial (Sine 2003).

Detailed explanations concerning the measurement of the quality traits can be found in de Alencar Figueiredo et al. (2006). Briefly, amylose content (AM), endosperm texture (ET) and peak gelatinization temperature (PGT) were evaluated using standard biochemical methods. Near-infrared reflectance spectroscopy was used to predict the other grain quality traits: protein content (PR), lipid content (LI), and kernel hardness (HD), based on the prediction equations developed and validated by de Alencar Figueiredo et al. (2006).

For all traits, values per accession were adjusted according to the experimental design.

Population structure, percentages of admixture and kinship coefficients

All accessions had been characterized previously by Deu et al. (2006), using 74 RFLP probes distributed throughout the genome. In their study, a similarity matrix was computed using a Nei and Li index. A neighbor-joining tree was then built to determine the aggregation of the accessions into clusters using DARwin software v5 (Perrier and Jacquemoud-Collet 2006), and 10 clusters were identified.

In our study, a subset of 60 of the 74 probes separated by more than 10 cM was established after eliminating those that were too tightly linked, carried null alleles, revealed duplicated loci or revealed no polymorphism at the 95% threshold. This new dataset was analyzed using Structure

(Pritchard et al. 2000a), a software that uses a Bayesian approach to simultaneously determine k (the number of subpopulations in a collection), and estimate for each accession the proportion of its genome, also called percentage of admixture, that originates from each subpopulation. Each simulation included 50,000 burn-in and 500,000 iterations. We ran 15 simulations per k value. We chose the options of an admixture model, since intermediates between the basic races were present in our collection, with correlated frequencies on haploid data. To determine the k value, we plotted the mean estimate across runs of the log posterior probability of the data for a given k, Pr($X|k$), called L(k) and chose the k value from which the distribution of L(k) plateaus or continues to increase but much more slowly. Because this point is known to be difficult to determine, we also used Δk, an ad hoc quantity proposed by Evanno et al. (2005) related to the second order rates of change of the likelihood function with respect to k. These authors demonstrated that Δk was a good predictor of the real number of subpopulations based on simulations. Δk is supposed to show a clear peak at the true value of k. The percentages of admixture of each accession (Q matrix) given by the software were further used as cofactors in the variance analyses. For trait analyses per subpopulation, notably mean comparisons of subpopulations, an accession was assigned to a subpopulation when it showed more than 80% membership in this subpopulation. Accessions with a higher degree of admixture (denoted nc) were not assigned to any subpopulation.

The kinship coefficient approach proposed by Yu et al. (2006) allows taking possible family relatedness into account and can help removing additional false positives. A number of markers higher than for estimation of population structure and corresponding to roughly 2,000 alleles for a population similar to ours is normally necessary to compute robust estimates of kinship coefficients (Yu et al. 2009). We were in a sub-optimal situation with only 60 RFLP markers genotyped and 160 alleles. In addition, our sample mostly comprised landraces for which the kinship component might be much less important than for breeding lines. We nevertheless computed these coefficients (K matrix) with the software Tassel (Bradbury et al. 2007) and used the two matrices (Q + K) in the variance analyses for tentative model comparisons.

Molecular polymorphism

The position of the sequenced segments in the six genes is indicated in Supplementary Fig. 2 adapted from the paper of de Alencar Figueiredo et al. (2008). Two segments were sequenced for *Sh2*, *Bt2*, *Ae1*, and *SssI*, seven segments covering most of the gene and a large part of its promoter for *O2*, and four segments for *Wx* (de Alencar Figueiredo et al. 2008). For three of the genes, the sequenced segment corresponded to one of the functional domains of the protein. For *SssI*, segment A partially overlapped the starch synthase catalytic domain. For *Wx*, segments A and B also corresponded to a starch synthase catalytic domain while, for *O2,* segment F partly overlapped the bZip domain. The gene segments were sequenced in all accessions of the core collection and the data deposited in GenBank under the following accession numbers: EU388245–EU388607 for *Sh2*, EU388985–EU389363 for *Bt2*, EU388608–EU388984 for *SssI*, EU387881–EU388244 for *Ae1*, EU387138–EU387880 for *Wx*, and EU389364–EU390699 for *O2*. SNPs and Indels and resulting haplotypes were determined as indicated in de Alencar Figueiredo et al. (2008). Polymorphic sites including alleles with a frequency below 1% in the set of accessions were not included in the data presented in Supplementary Tables 1–6 because they could result from PCR or sequencing errors. Polymorphic sites were named "s" followed by the site starting position in bp in the GenBank sequences minus 1, because of the polymorphic site numbering system of the Tassel package (Bradbury, personal comm.). The letters following the site name indicated whether mutations in the exons were synonymous (s) or non-synonymous (ns). The correspondence between the site number and the position on the pseudomolecules v1.0 (http://www.phytozome.org/sorghum) as well as the site flanking sequences (35 bp each side) are given in Supplementary Table 7. The heterozygosity rate was very low (0.8%) so the rare heterozygotes encountered were coded as missing data. In addition, rare SNPs or haplotypes, i.e. grouping five or fewer accessions, were discarded from the association analyses, because the statistical power to test for association would be insufficient with such low-frequency polymorphism.

Statistical analyses

Normality of trait distribution was assessed with a Kolmogorov–Smirnov test. Means between clusters were compared using Newman and Keuls tests. The percentage of variation of each trait explained by the structure was computed through multiple linear regression of the phenotypes on the percentages of admixture using R (Ihaka and Gentleman 1996). The tests of associations between molecular polymorphism (individual polymorphic sites, segment haplotypes and gene haplotypes) and phenotypes were computed using the software package Tassel (Bradbury et al. 2007). Three models were used: a simple General Linear Model (GLM) (model 1), a GLM model using the percentages of admixture of each accession (Q matrix) as cofactors to take population structure into account (model 2), and a Mixed Linear Model (MLM)

(model 3) using both the percentages of admixture and the kinship coefficients as cofactors (Q and K matrices). The tests were run with 1,000 permutations allowing the determination for each polymorphic site, segment, or haplotype of the site-wise P value, which is the probability of a greater F value under the null hypothesis that the polymorphic site, segment, or haplotype is independent of phenotype. The adjusted P value (called "p-adj_Marker" in Tassel), which is the site-wise P value adjusted for multiple tests and which takes dependence between hypotheses due to linkage disequilibrium into account (Bradbury et al. 2007) was also computed for each site. Because each gene was treated separately, we used an additional Bonferroni correction to correct for the number of genes studied (six), and the adjusted P value threshold at which an association was said to be significant was set to 0.01 (rounding up from 0.0083). The permutation method was not available for model 3 and a threshold of $P < 0.01$ was chosen in this last case.

Results

Structure of the population

The criteria used to define the number of subpopulations in the core collection, which are the position of a break point in the L(k) curve and a peak in the Δk distribution, supported values of $k = 2$ and $k = 6$ (Fig. 2). For both k values, most accessions were assigned by Structure to a subpopulation. The results of the assignments were projected on the NJ tree established by Deu et al. (2006), showing the very good congruence of the results obtained with the two methods (Fig. 3). Differences were mainly due to merging by Structure of clusters individualized on the NJ tree. With $k = 2$, it was possible to distinguish between sorghums from south (subpopulations 4 and 5 on Fig. 3), and north of the equator (subpopulations 1, 2, 3 and 6 on Fig. 3). With $k = 6$, a finer sub-grouping corresponding to a combined racial and geographical organization was obtained. Separate analyses conducted within the two subpopulations detected by Structure for $k = 2$ led to a sub-structure of two and four subpopulations, respectively, confirming the relevance of $k = 6$ (data not shown). For further analyses, $k = 6$ was chosen as being more effective in accounting for population structure in the analyses.

The percentage of variation of each phenotypic trait explained by population structure ranged from 3.7% (AM) to 36.9% (ET), with most regressions being very highly significant (Table 1). The traits most affected by population structure were, in decreasing order, ET, TGW, PR, HD, YLD and LI, all with a proportion of variance explained by population structure above 20%.

Phenotypic data

All traits were normally distributed ($P < 0.01$). The highest range of variation among accessions was obtained for grain texture (ET and HD) and yield components (Table 1). The smallest variations were obtained for PGT and AM.

We compared the means of the 146 accessions assigned to the 6 subpopulations for the different traits, excluding the admixed accessions (Table 2). We observed significant differences between subpopulations for all traits except NBG. Among traits showing significant differentiation among subpopulations, PGT and AM displayed the weakest separation. These two traits were also among those with the smallest percentage of variance explained by the structure.

Molecular polymorphism

Between two and seven segments were sequenced per gene (Tables S1–S6). Depending on the segment, between 165 and 194 good quality sequences were recovered from the original total of 195 sequences (de Alencar Figueiredo

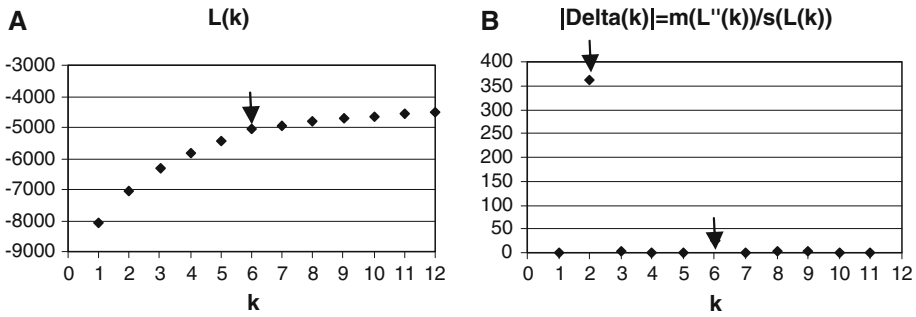

Fig. 2 Evolution of the criteria that enabled detection of the true number of subpopulations (k) in the collection using 60 RFLP markers and 205 accessions for k values varying from 1 to 12. **a** L(k) = mean of the likelihood distribution LnP(D) over 15 runs for each k value. **b** Delta(k) = m|L″(k)|/s(L(k)). The *black arrows* indicate the slope breakdown in the L(k) curve (**a**) and the peak points of Δk (**b**)

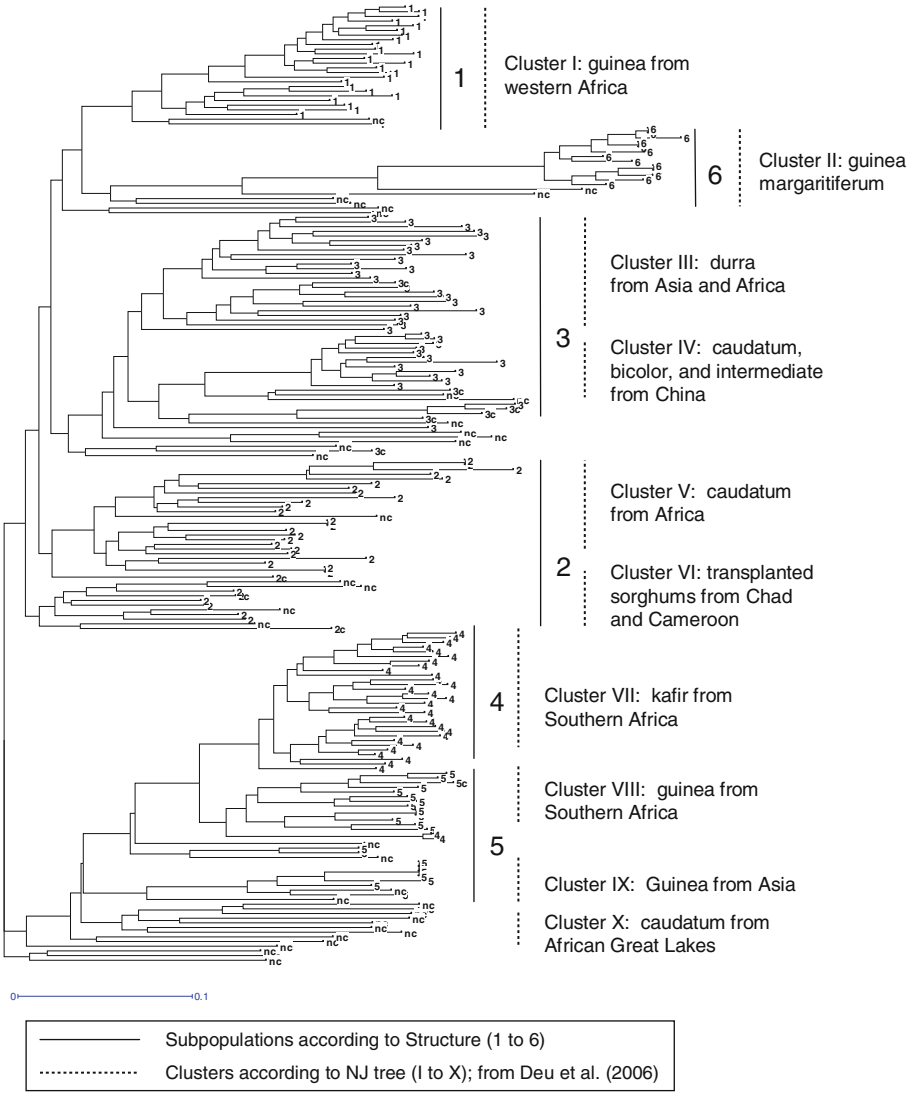

Fig. 3 Clustering of sorghum accessions. Projection of assignments based on Structure results ($k = 6$ and 80% membership in each cluster; *nc* non-classified accessions with no membership above the 80% threshold) on the NJ tree based on 74 RFLP probes and the Nei and Li similarity index from Deu et al. (2006)

et al. 2008). Because sequences of all segments of a gene were not available for all accessions, "gene haplotypes" resulting from the concatenation of all sequenced segments were only obtained for 123–183 accessions, depending on the gene. In this paper, to maximize the association detection power, we chose to keep the different segments of a gene separate and first to test the polymorphic sites individually, then organized as segment haplotypes, and finally as gene haplotypes. These three levels of the analysis of diversity are presented in the supplementary material (Supplementary Tables 1a–6h). The organization of the nucleotide diversity of the gene segments and haplotypes is extensively described and analyzed in de Alencar Figueiredo et al. (2008). Briefly, polymorphic sites, including SNP and Indels, varied from 1 to 17 depending on the segment (Table 3). For all segments, the polymorphism was organized in a limited number of segment haplotypes (2 to 5), indicating a strong LD within segments and a high level of redundancy in the information obtained from the polymorphic sites. In many segments, some minor variants, which probably appeared more recently, often broke the LD between polymorphic sites (de Alencar Figueiredo et al. 2008). Clustering of these variants with the dominant haplotypes generally allowed discrimination of two, or less frequently three, different groups. The number of gene haplotypes varied from three to nine, indicating the occasional occurrence of recombination (Table 3).

Association tests

Structure and relatedness effects

We first ran association tests using two models, a simple GLM model (model 1) and a model using the percentages of admixture (Q matrix) obtained for $k = 6$ as cofactors

Table 1 Statistics for grain quality and yield component traits and percentage of variation of these traits explained by population structure ($K = 6$) through multiple linear regression

Trait	Mean	Min	Max	SD	CV (%)	% Var
ET	2.83	1.02	5.02	1.04	36.8	36.9***
HD	15.31	9.37	25.85	3.01	19.7	24.1***
PR	14.2	9.6	18.1	1.86	13.1	34.8***
LI	3.89	2.47	5.79	0.61	15.7	21.4***
AM	20.4	14.8	24.3	1.54	7.5	3.7 ns
PGT	73.50	69.90	76.80	1.27	1.7	12.6***
TGW	25.90	7.32	54.16	8.42	32.5	35.1***
NBG	1,689	134	3,371	580	34.3	6.6*
YLD	42.31	5.56	94.02	16.30	38.5	22.4***

ET Endosperm texture, *HD* hardness, *PR* protein content (%), *LI* lipid content (%), *AM* amylose content (%), *PGT* peak gelatinization temperature (°C), *TGW* thousand grain weight (g), *NBG* number of grains per panicle, *YLD* grain yield per plant (g), $N = 192$ accessions with data for all traits, *% var* percentage of variation explained by population structure for $K = 6$, *ns* non significant

* Significant at the 5% level; *** significant at the 0.5% level

(model 2). Using model 2, considerably fewer segments and sites showed a significant association than using model 1. Altogether, with model 2, we found 83 (7%), 15 (9%), and 8 (15%) significant associations ($P < 0.01$) between one trait and polymorphic sites, segment haplotypes, and gene haplotypes, respectively (Table 3), which represent a decrease of 64, 75 and 70% from model 1.

The importance of the differences in the number of significant tests between models 1 and 2 showed that the structure effect was strong. However, the percentage of variation of each trait explained by the structure (Table 1) was not a very good predictor of this effect, since PGT (only 12.6% of the phenotypic variability explained by the structure) and, to a lesser extent, AM (3.7%) also showed large differences in the number of significant tests between model 1 and model 2.

We also compared the results between model 2 (Q matrix alone) and model 3 (Q + K matrix). The majority of associations remained significant (Table 3). With model 3, we found 48 (4%), 10 (6%), and 5 (9%) significant associations ($P < 0.01$) between one trait and polymorphic sites, segment haplotypes, and gene haplotypes, respectively (Table 3), which represent a decrease of 42, 33 and 37% from model 2. These changes in significance concerned mostly markers that were close to the significance threshold in model 2. Most of them would have remained significant at $P < 0.05$. We assumed that most of the tests that were significant only with model 1 were false positives due to population structure. It is possibly the case also for the tests that were significant for model 2 but not for model 3, but given that we are not in an optimal situation to estimate relatedness, from now on, we consequently mainly discuss model 2. The level of probability and the percentage of variance explained by the polymorphic sites, segment or gene haplotypes that were significant with model 2 are shown in Table 4, with the polymorphisms that were also significant for model 3 indicated in bold.

Sh2

Segment A (from exon 1 to exon 2) of *Sh2* was strongly associated with AM and TGW. All polymorphic sites found in segment A (except one) showed associations with both traits. This result is not surprising given that these seven SNPs (s87ns, s264, s281, s480, s525, s570, and s636) showed complete LD (Table 4). The SNPs differentiated haplotypes A1 and A2 from haplotype A3, which is mainly composed of accessions from subpopulation 6 (guinea margaritiferum) and a few admixed bicolor accessions. In addition, s253 (Indel), which characterized the minor haplotype A2 (not specific to any accession type) derived from the major haplotype A1, was also significant for TGW. For segment B (3′ end), one SNP (s240s) was significantly linked with AM and TGW, as well as with ET

Table 2 Mean comparison between subpopulations for the accessions assigned to a subpopulation (>80% membership in the subpopulation)

Sub pop	ET	HD	PR	LI	AM	PGT	TGW	NBG	YLD
1	2.10 c	13.6 c	14.8 b	3.98 ab	21.3 a	73.5 b	28.4 b	1,510 a	42.1 bc
2	3.87 a	17.5 a	12.5 c	3.47 c	20.1 b	73.0 b	32.5 a	1,611 a	51.7 a
3	3.04 b	15.6 b	14.5 b	4.28 a	20.5 ab	73.2 b	27.4 b	1,740 a	46.7 ab
4	3.00 b	15.1 b	13.0 c	3.97 ab	20.9 ab	73.0 b	24.3 bc	1,727 a	41.3 bc
5	1.63 c	12.5 cd	15.3 b	3.71 bc	20.1 b	73.5 b	21.4 c	1,821 a	35.3 c
6	1.68 c	12.2 d	16.4 a	4.06 ab	20.4 ab	74.9 a	15.8 d	1,526 a	24.1 d
N	146	146	146	146	146	146	146	144[a]	144[a]

ET Endosperm texture, *HD* hardness, *PR* protein content, *LI* lipid content, *AM* amylose content, *PGT* peak gelatinization temperature, *TGW* thousand grain weight, *NBG* number of grains per panicle, *YLD* grain yield per plant, *N* number of accessions

[a] Two samples with missing field data

Table 3 Synthesis of the significant associations for each gene or gene segment × trait combination using correction for population structure and adjustment for multiple tests

Gene segment	P	No	ET		HD		PR		LI		AM		PGT		TGW		NBG		YLD	
			Q	Q + K	Q	Q + K	Q	Q + K	Q	Q + K	Q	Q + K	Q	Q + K	Q	Q + K	Q	Q + K	Q	Q + K
Sh2 A	8	184	–	–	–	–	–	–	–	–	7H	7H	–	–	8H	1H	–	–	–	–
Sh2 B	11	169	1	–	–	–	–	–	–	–	1	1	1	–	1	–	–	–	–	–
Sh2	19	155	–	–	–	–	–	–	–	–	H	H	H	–	H	H	–	–	–	–
Bt2 A	4	177	–	–	–	–	2	2	–	–	–	–	2	2	2	–	–	–	–	–
Bt2 B	3	191	–	–	–	–	1H	H	1H	–	–	–	–	–	1H	1H	–	–	H	–
Bt2	7	162	–	–	–	–	–	–	–	–	–	–	–		H	H	–	–	–	–
SssI A	5	185	–	–	–	–	–	–	1	–	–	–	4H	–	–	–	–	–	–	–
SssI B	4	191	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
SssI	9	179	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Ae1 A	17	185	–	–	–	–	–	–	–	–	–	–	6	6	–		–	–	–	–
Ae1 B	11	174	–	–	–	–	–	–	1	–	–	–	9	8H	8H	1H	–	–	1	–
Ae1	28	158	–	–	–	–	–	–	–	–	–	–	–	–	H	–	–	–	–	–
Wx A	14	165	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Wx B	6	191	–	–	–	–	–	–	–	–	–	–	–	–	5H	5H	–	–	–	–
Wx C	1	187	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Wx D	7	189	4H	–	2H	–	–	–	–	–	–	–	–	–	4H	4H	–	–	–	–
Wx	28	123	–	–	–	–	–	–	–	–	–	–	–	–	–	H	–	–	–	–
O2 A	12	186	–	–	2H	2H	–	–	–		–	–	–	–	2	2	–	–	–	–
O2 B	5	193	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
O2 C	6	190	–	–	–	–	–	–	–	–	–	–	–	–	–	–	1	–	–	–
O2 D	5	189	1H	1H	2H	2H	–	–	–	–	1	–	1H	1	1	–	–	–	–	–
O2 E	8	186	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
O2 F	3	183	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
O2 G	1	194	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
O2	40	183	H	H	H		–	–	–	–	–	–	H	–	–	–	–	–	–	–
Total P	131		6	1	6	4	3	2	3	0	9	8	23	17	32	14	1	0	1	0

Polymorphism or haplotypes with less than 5 sequences were not included in the association test. The number of significant polymorphic sites is indicated, followed by H when the haplotype is also significant

P number of polymorphism, *No* number of sequences, *Q* model 2 with percentage of admixture used as cofactors in the analyses, *Q + K* model 3 with both percentages of admixture and coefficients of kinship used as cofactors in the analyses

and PGT (Table 4). This SNP, differentiated B1, the most frequent haplotype, from haplotypes B2 and B3 which did not show specific characteristics (Supplementary Table 1b). In addition, the segment A haplotype and the gene haplotype were also significantly associated with variation in both AM and TGW (Table 4). No association was found with the remaining traits.

Bt2

In segment A (from exon 4 to exon 5), two SNPs (s530 and s578) were associated with PR, PGT and TGW (Table 4). These SNPs, in complete LD, differentiated haplotypes A1, A2 and A3 from A4, itself composed mostly of accessions from subpopulation 6 (guinea margaritiferum), plus a few from subpopulation 3 and admixed bicolor (Supplementary Table 2a). In segment B (from exon 9 to exon 10; Figure S2), associations were significant between s178 and PR, and s145 and TGW and LI (Table 4). These two SNPs, in weak LD, are not involved in the differentiation of the same populations, since the s178 variant was a minor allele mostly found in subpopulation 6 (guinea margaritiferum accessions), and the s145 variant was a minor allele found in subpopulations 2 and 3 (caudatum and intermediate accessions from Africa and China). Significant segment haplotype associations were only found in segment B. These associations concerned PR, LI, TGW, and YLD (Table 4). The gene haplotype was significantly associated with TGW. No significant association was found between *Bt2* and ET, HD, AM, or NBG.

Table 4 Detailed results of the significant association tests using correction for population structure (model 2)

Trait	Gene or gene segment	Locus	*df* error	*F*	*P**	R^2
ET	*Sh2* B	s240s	159	6.94	0.0092	0.027
ET	*Wx* D	s121, s445, s471, s562	179	8.32	0.0044	0.028
ET	*Wx* D	H	177	5.67	0.0010	0.054
ET	***O2* D**	**s328**	180	**18.28**	**0.0000**	**0.059**
ET	***O2* D**	**H**	178	**8.90**	**0.0000**	**0.084**
ET	***O2***	**H**	178	**4.43**	**0.0001**	**0.094**
HD	*Wx* D	s535, s539	179	9.19	0.0028	0.036
HD	*Wx* D	H	177	5.79	0.0008	0.066
HD	***O2* A**	**s85, s258**	**175**	**9.46**	**0.0024**	**0.037**
HD	***O2* A**	**H**	**173**	**4.12**	**0.0075**	**0.048**
HD	***O2* D**	**s39**	**180**	**9.89**	**0.0019**	**0.040**
HD	***O2* D**	**s328**	**180**	**7.65**	**0.0063**	**0.031**
HD	***O2* D**	**H**	**178**	**7.03**	**0.0002**	**0.081**
HD	*O2*	H	178	3.88	0.0006	0.100
PR	***Bt2* A**	**s530, s578**	**168**	**7.80**	**0.0058**	**0.028**
PR	*Bt2* B	s178	181	7.92	0.0054	0.027
PR	***Bt2* B**	**H**	**179**	**4.92**	**0.0026**	**0.050**
LI	*Bt2* B	s145	181	7.98	0.0053	0.033
LI	*Bt2* B	H	179	4.48	0.0047	0.055
LI	*SssI* A	s859	175	7.93	0.0054	0.033
LI	*Ae1* B	s705	164	7.17	0.0082	0.032
AM	***Sh2* A**	**s87, s264, s281, s480, s525, s570, s636**	**174**	**11.27**	**0.0010**	**0.059**
AM	***Sh2* A**	**H**	**173**	**5.98**	**0.0031**	**0.062**
AM	***Sh2* B**	**s240s**	**159**	**8.90**	**0.0033**	**0.052**
AM	***Sh2***	**H**	**146**	**4.88**	**0.0010**	**0.114**
AM	***O2* D**	**s328**	**180**	**6.80**	**0.0099**	**0.035**
PGT	*Sh2* B	s240s	159	9.22	0.0028	0.047
PGT	*Sh2*	H	146	4.48	0.0019	0.097
PGT	***Bt2* A**	**s530, s578**	**168**	**7.51**	**0.0068**	**0.037**
PGT	*SssI* A	s57, s491, s509, s610	175	7.24	0.0078	0.035
PGT	*SssI* A	H	174	5.11	0.0070	0.048
PGT	***Ae1* A**	**s244s, s370, s393**	**174**	**12.12**	**0.0006**	**0.058**
PGT	***Ae1* A**	**s241s, s246ns, s343**	**174**	**9.66**	**0.0022**	**0.047**
PGT	***Ae1* B**	**s187, s189, s202, s222, s224, s239,** s244	**164**	**10.71**	**0.0013**	**0.053**
PGT	***Ae1* B**	**s401**	**163**	**5.35**	**0.0056**	**0.053**
PGT	***Ae1* B**	**s471**	**164**	**6.89**	**0.0095**	**0.035**
PGT	***O2* D**	**s328**	**180**	**15.82**	**0.0001**	**0.071**
PGT	*O2* D	H	178	5.24	0.0017	0.071
PGT	*O2*	H	178	4.49	0.0001	0.131
TGW	*Sh2* A	s87, s264, s281, s480, s525, s570, s636	172	6.81	0.0098	0.025
TGW	***Sh2* A**	**s253**	**172**	**12.60**	**0.0005**	**0.045**
TGW	***Sh2* A**	**H**	**171**	**9.45**	**0.0001**	**0.066**
TGW	*Sh2* B	s240s	157	7.03	0.0088	0.028
TGW	***Sh2***	**H**	**144**	**4.93**	**0.0009**	**0.081**
TGW	*Bt2* A	s530, s578	167	9.89	0.0020	0.036
TGW	***Bt2* B**	**s145**	**179**	**9.24**	**0.0027**	**0.032**
TGW	***Bt2* B**	**H**	**177**	**6.65**	**0.0003**	**0.066**
TGW	***Bt2***	**H**	**159**	**6.20**	**0.0000**	**0.122**

Table 4 continued

Trait	Gene or gene segment	Locus	*df* error	*F*	*P**	R^2
TGW	*Ae1* B	s187, s189, s202, s222, s224, s239, s244,	162	8.59	0.0039	0.033
TGW	***Ae1* B**	**s705**	**162**	**8.21**	**0.0047**	**0.033**
TGW	***Ae1* B**	**H**	**159**	**5.29**	**0.0005**	**0.076**
TGW	*Ae1*	H	147	3.50	0.0017	0.091
TGW	***Wx* B**	**s110, s187s, s191s, s263, s267**	**179**	**13.48**	**0.0003**	**0.046**
TGW	***Wx* B**	**H**	**178**	**16.09**	**0.0000**	**0.100**
TGW	***Wx* D**	**s121, s445, s471, s562**	**177**	**16.12**	**0.0001**	**0.052**
TGW	***Wx* D**	**H**	**175**	**6.51**	**0.0003**	**0.063**
TGW	***O2* A**	**s85, s258**	**173**	**8.66**	**0.0037**	**0.031**
TGW	*O2* D	s39	178	7.66	0.0062	0.026
NBG	*O2 C*	s440ns	176	7.94	0.0054	0.040
YLD	*Bt2* B	H	174	4.69	0.0035	0.057
YLD	*Ae1* B	s705	160	8.22	0.0047	0.038

Tests that gave exactly the same results (F, P and R^2) are listed in the same row. In bold, markers or haplotypes also significant for model 3 ($P < 0.01$)

* Site-wise or haplotype-wise P value adjusted for multiple tests and which takes dependence between hypotheses due to linkage disequilibrium into account

SssI

In segment A (from exon 2 to exon 7), four SNPs in complete LD (s57, s491, s509 and s610s) were associated with PGT (Table 4). These SNPs differentiated haplotype A1 from haplotypes A2 and A3 that included accessions from subpopulations 6 and 3 and a few admixed accessions (Supplementary Table 3a). In addition, an isolated SNP in segment A (s859) characteristic of haplotype A3 (accessions from subpopulation 3 mostly) was associated with LI. The segment A haplotype was also significantly associated with PGT. No significant association was found with ET, HD, PR, AM, TGW, NBG, or YLD. Segment B (3′ end) did not show significant association with any trait.

Ae1

In segment A (from exon 4 to exon 8), two sets of three SNPs were associated with PGT. These three SNPs, s241, s246 and s343, differentiated haplotype A1 (mixture of accessions from populations 2, 3 and 4) from the four other haplotypes, and s244, s370 and s393 differentiated haplotypes A1 and A5 (mixture of accessions of all populations except the guineas of subpopulation 5) from haplotypes A2, A3 and A4 (Supplementary Table 4a). In segment B (from exon 19 to exon 22), nine polymorphic sites (s187, s189, s202, s222, s224, s239, s244, s401, and s471) were also associated with PGT (Table 4). The first seven polymorphic sites were in complete LD and differentiated B1, B2 and B3 from B4 and B5 (all subpopulations except 5, Supplementary Table 4b). These seven polymorphic sites plus s705 were also significantly associated with TGW. They differentiated haplotype B2 (mostly durra and caudatum from subpopulations 2 and 3) from the others. In addition, s705 in segment B was associated with LI and YLD. The segment B haplotype and the gene haplotype were both associated with TGW. No association was found with the remaining traits.

Wx

For *Wx*, four segments were sequenced but only two showed significant associations with ET (segment D; from exon 14 to 3′UTR), HD (segment D) and TGW (segments B, from exon 7 to exon 9, and D). Four polymorphic sites in segment D in complete LD (s121, s445, s471, and s562) were found to be significantly linked with ET and TGW. They separated haplotypes D1, D2 and D3 from D4 (subset of guinea accessions from subpopulation 5). Two other SNPs (s535 and s539) themselves in full LD in segment D (Supplementary Table 5d) were associated with HD. They discriminated haplotypes D1, D3 and D4 from D2 composed only of guineas (subset of subpopulation 1). The segment D haplotype was associated with ET, HD, and TGW. In addition, five polymorphic sites in segment B (s110, s187s, s191s, s263, and s267) were found to be significantly linked with TGW. They separated B1 from B2 and B3 (accessions from subpopulations 3, 4, and 5). The segment B haplotype was also associated with TGW. No association was found with PR, LI, AM, PGT, NBG or YLD.

O2

Seven segments of *O2* covering almost entirely the promoter (96%) and a large part of the whole gene (87%) were sequenced. Most polymorphic sites, in strong LD, were not associated with any trait but a few specific polymorphisms representing local ruptures of LD were linked with ET, HD, AM, PGT, NBG, or TGW. One Indel (s85) and one SNP (s258) in LD in segment A of the promoter (Supplementary Table 6a) were significantly associated with HD and TGW. These are the two polymorphic sites that differentiate the minor haplotype A3 (a subset of accessions from East Africa from subpopulation 5) from the most common haplotype (A1). One SNP (s328) in segment D in intron 1 was significantly associated with ET, HD, and AM, separating haplotype D2 (subset of guinea accessions from subpopulation 1) from the others (Supplementary Table 6d). The polymorphic sites of both segments were in LD at the gene level (Supplementary Table 6h). The segment A haplotype was significantly associated with HD, and the segment D haplotype as well as the gene haplotype, were significantly associated with ET, HD, and PGT.

Discussion

To improve our understanding of the genetics of grain quality in sorghum, we analyzed in a structured core collection of sorghum the associations that exist between polymorphic sites within five genes involved in starch biosynthesis (*Sh2*, *Bt2*, *SssI*, *Ae1*, and *Wx*) and one gene controlling grain storage proteins (*O2*), and key traits determining grain quality.

Population structure

Sorghum displays a racial and geographical structure (Ollitrault 1987; Deu et al. 1994) and this structure was detected, as expected, in the core collection using both classical multivariate (Deu et al. 2006) and Bayesian analyses. The patterns obtained with the two methods were very similar, but a finer resolution was obtained with the multivariate approach (ten clusters versus six subpopulations with the Bayesian approach) for which additional markers were used. Population structure is the primary obstacle to successful association studies in any organism (Buckler and Thornsberry 2002). Among the available methods to control population structure in association studies, we chose that proposed by Pritchard et al. (2000b), which uses the percentages of admixture of each accession as covariates in the variance analysis. The accuracy of the estimation of the percentages of admixtures is consequently important for the results of variance analysis. The congruence of patterns obtained with Bayesian and multivariate analyses suggests that the estimates of these admixture proportions are reasonably reliable.

The percentages of admixture explained up to 37% of the trait variability, so an effect of population structure on the phenotypic trait significance was expected. In our study, the proportion of false-positive tests due to population structure reached 64, 75 and 70% for polymorphic sites, segment haplotypes, and gene haplotypes, respectively. These rates may seem high but they are comparable with the 80% of false-positive reduction observed by Thornsberry et al. (2001) in the first published association study conducted with inbred maize lines.

With a method that better accounts for kinship relationships such as that described by Yu et al. (2006), it might be possible to remove more of the structure effect, as shown by the tentative comparisons between models 2 and 3. This point was demonstrated by Brown et al. (2008) in sorghum and Cockram et al. (2008) in barley. However, to get robust estimates of kinship, many more markers than our current set of RFLPs are required (Yu et al. 2009). A set of 1,200 additional DArT markers are presently being genotyped on the population (Bouchet et al., unpublished) and should soon permit both a correct estimation of kinship coefficients and a whole genome analysis.

Trait–gene associations

For all genes, several associations remained significant with model 2 after elimination of the false positives. These associations generally matched well with what is presently known of the function of the enzymes for which the genes code. They were well supported by the results of QTL studies both in sorghum and maize, and those of the association study of Wilson et al. (2004) conducted with a panel of diverse maize inbred lines. Although the sequenced segments of the sorghum and maize association studies did not overlap, the strong LD in the studied genes of both species made comparisons between associations relevant at the gene level.

Except *SssI*, all the genes coding for enzymes of the starch production pathway tested in this study (Fig. 1) were found associated with TGW. This reflects their complementary role in determining starch amount and grain filling. Evidence derived from natural mutant analyses summarized by Wilson et al. (2004) and Manicacci et al. (2007) demonstrated that *Bt2* and *Sh2* encoded the two subunits of AGPase, an enzyme generally regarded as the rate-limiting step in starch biosynthesis, and directly affected seed weight through starch content in maize kernels. QTL evidences from sorghum or maize also support the likelihood of an association with *Bt2* and *Sh2*. QTLs for TGW on chromosome 7 in sorghum co-localized with *Bt2*

(Rami et al. 1998) as did QTLs for TGW and starch content on the homologous region of chromosome 8 in maize (Séne et al. 2001). QTLs for TGW co-localized with *Sh2* in maize (Doebley et al. 1994).

The two genes that code for enzymes specifically involved in amylopectin production (*Ae1* and *SssI*) were found associated with PGT. PGT is a starch thermal property associated with the time and energy demanded for cooking which is influenced notably by the richness in amylopectin. Studies in maize also showed a significant association between *Ae1* and pasting temperature, which is affected both by gelatinization temperature and by AM (Wilson et al. 2004). QTLs for the ratio of amylose to amylopectin, AM, and starch quantity are co-localized with *Ae1* on maize chromosome 5 (Goldman et al. 1993; Séne et al. 2000, 2001) while in the orthologous area of chromosome 4 in sorghum, a QTL for AM was detected in the vicinity of the gene (Rami et al. 1998). We localized *SssI* on sorghum chromosome 10 in an area that corresponds to chromosome 9 in maize. In maize, QTLs for the ratio of amylose to amylopectin are co-localized with *SssI* (Séne et al. 2000).

Among the three genes possibly involved in amylose production, only *Sh2* was found strongly associated with AM. The effect of *Sh2* on AM may be indirect in the sense that AGPase converts UDP-glucose into glucose-1-phosphate, which is the substrate for activity of the granule-bound starch synthase encoded by *Wx* and responsible for amylose production. Associations between *Sh2* and amylose content, pasting temperatures and starch viscosity were also found in maize (Wilson et al. 2004) and QTL for AM (Séne et al. 2000) co-localized with *Sh2*.

Wx itself, however, was not associated to AM in our study while it is a gene that is well known to determine amylose content in grain in cereals (James et al. 2003) as well as several viscosity parameters (Larkin et al. 2003). Several explanations are possible for the absence of associations with *Wx*. First, our core collection did not include any accessions without amylose (waxy), which are much rarer in sorghum than in rice (Mestres, personal communication). Consequently, the non-synonymous SNP in exon 8 of *Wx* that was proposed as a strong candidate for the causal mutation underlying the waxy phenotype in the accession concerned (Hamblin et al. 2007; McIntyre et al. 2008) was monomorphic in our sample. Secondly, since AM is a quantitative trait, *Wx* may not be the only gene involved in determining amylose content, as observed in rice (Ayres et al. 1997; Olsen and Purugganan 2002), nor the most important. In sorghum, two QTLs for AM were indeed detected on different chromosomes to that carrying *Wx* (Rami et al. 1998).

Associations with ET and HD, two very strongly correlated traits, were observed for *O2* and *Wx*. *O2* is known to have a strong influence on grain texture in maize (Lopes and Larkins 1991). Rami et al. (1998) reported co-localization of QTLs for AM, HD, ET with *O2* on chromosome 2.

Knowledge of the function of *O2* made this an a priori logical candidate gene to also explain PR variability. *O2* is a transcription factor that is assumed to play a central role in storage protein synthesis in cereals and may control allocation and balance between starch and protein content (Maddaloni et al. 1996; Henry and Damerval 1997). *O2*, however, showed no association with PR in this study. Rami et al. (1998) reported co-localization of QTLs for albumins, and kafirins with *O2* on chromosome 2, but not with total protein content. Dosage of kafirins and albumins in the protein fraction may help determine whether *O2* is associated with a narrower protein group rather than directly with PR.

Lastly, the specific association of *Bt2* with PR and LI was also found in maize (Wilson et al. 2004).

Some of the observed associations represent good examples of the interest of association studies. For example, an association between *Sh2* and TGW was detected while no QTL for TGW was found close to *Sh2* in sorghum (Rami et al. 1998) because the parents of the mapping population, which were included in our genotyped set, were monomorphic for all SNPs or Indels detected in the two segments of *Sh2*. The large allelic diversity of association panels represents a clear advantage on genetic mapping studies provided these alleles have a high enough frequency to be tested with sufficient power.

Appropriateness of a structured core collection for association studies

Positive associations were found for several gene × trait combinations showing that the association mapping approach was successful. But is a structured core collection the most appropriate population for association mapping? We observed that many suspected false positives disappeared when population structure and familial relatedness were taken into account. However, the remaining significant variation seems to still be linked, to some extent, to population structure. This is shown by the frequency of differentiated haplotypes composed mainly of guinea margaritiferum with a few admixed bicolor accessions, as highlighted for *Bt2* and *Sh2*. To improve population composition, one possible solution would be to extract a subsample of accessions from which the most redundant units would be removed, as conducted by Breseghello and Sorrells (2006) in wheat and/or LD minimized, using sampling tools specifically developed for such purposes (Perrier and Jacquemoud-Collet 2006). The risk associated with the latter strategy would be a potential loss in

statistical power to detect associations, due to the reduced population size.

Another possible way to limit the population structure problem would be to run association mapping in one of the subpopulations defined by Structure, showing very little structuring by construction, provided that the subpopulation contained enough phenotypic variability. This choice would also enable exploration of allelic diversity, which is often unexploited using classical linkage mapping, which mainly focuses on inter-subpopulation mapping populations because of improved polymorphism rates. Nevertheless, in this case, a sample larger than the one used in this study appears to be necessary.

In association mapping in strongly structured populations, another issue is the organization of phenotypic variability. In a situation where most phenotypic variability is distributed between subpopulations, association mapping is not an efficient method to identify the polymorphism responsible for such variation because the functional polymorphism will be included among the false positives linked to population structure. In such a case, QTL analysis in classical mapping populations derived from inter-subpopulation crosses may be more appropriate to detect phenotype–genotype associations. Association mapping should work better with phenotypic variability distributed more evenly between and within subpopulations. Consequently, it may be useful to initially determine the extent and structure of phenotypic variability in the target study population to help select the most appropriate mapping panel.

Complementary approaches

Selection tests that compare the level of nucleotide diversity of a gene with that expected under the hypothesis of neutral evolution are sometimes used as an indirect way to confirm that a candidate gene controls an agronomically useful phenotype (Vigouroux et al. 2002; Vasemägi and Primmer 2005). It is assumed that if a gene has been the target of selection before and/or since domestication, it must be of functional importance. Such tests performed on sorghum grain quality genes gave inconclusive results. Hamblin et al. (2007), using HKA's tests on a sample of 23 accessions, did not find any trace of selection at starch pathway loci including *Sh2*, *Bt2*, *Ae1*, *SssI* and *Wx*. De Alencar Figueiredo et al. (2008), using Tajima's tests on a sub-sample of 53 accessions from our collection, found signatures of selection for *Ae1*, *Wx* (positive D) and *Sh2* (negative D), whereas tests were not significant for *Bt2*, *SssI* and *O2*. Demographic processes such as population expansion or a bottleneck, recombination or population subdivision can all lead to false positives in such tests (Nielsen 2005); conversely, small sample size, as in the two studies, can limit the power of these tests and bias the results through the absence of rare haplotypes. Adding these limitations to the fact that HKA and Tajima's tests are different in nature, it is therefore difficult to draw firm conclusions regarding the action of selection through these results.

Purifying selection can affect association studies in another way. It tends to create or maintain monomorphic gene segments that play an important role in the conservation of gene function. Consequently, even if the gene is functionally important, it may not be detected in QTL or association mapping, rendering it useless for its original purpose in potential crop improvement.

What about functional polymorphism?

We selectively sequenced the 5′ and the 3′ ends of genes, with the exception of *Wx* which was sequenced almost entirely, and *O2* for which the promoter was also included. On the one hand, sequencing of the whole gene is not required for genes that are characterized by a strong haplotype structure. On the other hand, because of the high level of intra-genic LD, in analyzing the association with grain quality traits it is not possible to resolve down to the polymorphic site level. Since we did not study the LD around the six genes, it is not even possible to definitely conclude that the genes studied are involved in determining grain quality rather than other genes located nearby and in partial or full LD with our targets. However, this risk appears to be limited since, on average, LD in sorghum is said to have largely decayed after 15 kb (Hamblin et al. 2005) and the results are very well supported by those on orthologous genes in maize (Wilson et al. 2004).

Additional work is needed in other backgrounds showing more recombination and less structure such as randomly mated populations, to assess LD around the target genes to confirm the role of these genes and, eventually, draw inferences on the nature of the polymorphic site responsible for the phenotypic differences. Still, this association study, which is the first one to target grain quality traits in sorghum, enabled us to confirm and/or rule out the potential effect of the polymorphic sites tested in the six genes. This study will help sorghum breeders to speed up the development of varieties for human consumption with improved grain quality, ultimately resulting in wider acceptance of these improved varieties by farmers, who are also the main consumers.

Acknowledgments The authors gratefully acknowledge the financial support from the GABI-Génoplante project "Bridging genomics and genetic diversity: associations between gene polymorphism and trait variation in cereals" for the sequencing of *O2*, and from the CNPq and from the Universidade Católica de Brasília through a grant to L.F. de A.F.

References

Aboubacar A, Hamaker BR (1999) Physicochemical properties of flours that relate to sorghum couscous quality. Cereal Chem 76:308–313

Agrama HA, Eizenga GC, Yan W (2007) Association mapping of yield and its components in rice cultivars. Mol Breed 19:341–356

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

Ayres NM, McClung AM, Larkin PD, Bligh HFJ, Jones CA, Park WD (1997) Microsatellites and a single-nucleotide polymorphism differentiate apparent amylose classes in an extended rice pedigree of US rice germplasm. Theor Appl Genet 94:773–781

Belton PS, Taylor JRN (2004) Sorghum and millets: protein sources for Africa. Trends Food Sci Technol 15:94–98

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

Breseghello F, Sorrells ME (2006) Association mapping of kernel size and milling quality in wheat. Genetics 172:1165–1177

Brown PJ, Rooney WL, Franks C, Kresovich S (2008) Efficient mapping of plant height quantitative trait loci in a sorghum association population with introgressed dwarfism genes. Genetics 180:629–637

Buckler ES, Thornsberry JM (2002) Plant molecular diversity and applications to genomics. Curr Opin Plant Biol 5:107–111

Buntjer JB, Sorensen AP, Peleman JD (2005) Haplotype diversity: the link between statistical and biological association. Trends Plant Sci 10:466–471

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

Chantereau J, Trouche G, Luce C, Deu M, Hamon P (1997) Le Sorgho. In: Charrier A, Jacquot M, Hamon S, Nicolas D (eds) L'amélioration des plantes tropicales. Cirad and Orstom, Montpellier, pp 565–590

Cockram J, White J, Leigh FJ, Lea VJ, Chiapparino E, Laurie DA, Mackay IJ, Powell W, O'Sullivan DM (2008) Association mapping of partitioning loci in barley. BMC Genetics 9:1–14

de Alencar Figueiredo LF, Davrieux F, Fliedel G, Rami J-F, Chantereau J, Deu M, Courtois B, Mestres C (2006) Development of NIRS equations based on a core collection to predict quality traits in sorghum grain. J Agric Food Chem 54:8501–8509

de Alencar Figueiredo LF, Calatayud C, Dupuis C, Billot C, Rami JF, Brunel D, Perrier X, Courtois B, Deu M, Glaszmann J-C (2008) Phylogeographic evidence of crop neo-diversity in sorghum. Genetics 179:997–1008

Deu M, Gonzàlez-de-Leòn D, Glaszmann J-C, Dégremont I, Chantereau J, Lanaud C, Hamon P (1994) RFLP diversity in cultivated sorghum in relation to racial differentiation. Theor Appl Genet 88:838–844

Deu M, Rattunde F, Chantereau J (2006) A global view of genetic diversity in cultivated sorghums using a core collection. Genome 49:168–180

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

Doebley J, Bacigalupo A, Stec A (1994) Inheritance of kernel weight in two maize–teosinte hybrid populations: implications for crop evolution. J Hered 85:191–195

Doggett H (1988) Sorghum, 2nd edn. Longman, New York (512 pp)

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

Fliedel G (1994) Evaluation de la qualité du sorgho pour la fabrication du tô. Agriculture et développement 34:12–21

Fliedel G, Marti A, Thiebaut S (1996) Caractérisation et valorisation du sorgho. Cirad Montpellier, 404 pp

Flint-Garcia SA, Thornsberry JM, Buckler ES (2003) Structure of linkage disequilibrium in plants. Annu Rev Plant Biol 53:357–374

Gao H, Williamson S, Bustamante C (2007) An MCMC approach for joint inference of population structure and inbreeding rates from multi-locus genotype data. Genetics 176:1635–1651

Gebhardt C, Ballvora A, Walkemeier B, Oberhagemann P, Schüler K (2004) Assessing genetic potential in germplasm collections of crop plants by marker–trait association: a case study for potatoes with quantitative variation of resistance to late blight and maturity type. Mol Breed 13:93–102

Goldman LL, Rocheford TR, Dudley JW (1993) Quantitative trait loci influencing protein and starch concentration in Illinois long term selection maize strains. Theor Appl Genet 87:217–224

Hamblin MT, Salas Fernandez MG, Casa AM, Mitchell SE, Paterson AH, Kresovich S (2005) Equilibrium processes cannot explain high levels of short- and medium-range linkage disequilibrium in the domesticated grass *Sorghum bicolor*. Genetics 171:1247–1256

Hamblin MT, Salas Fernandez MG, Tuinstra MR, Rooney WL, Kresovich S (2007) Sequence variation at candidate loci in the starch metabolism pathway in sorghum: prospects for linkage disequilibrium mapping. Plant Genome 2:125–134

Harlan JR, de Wet JMJ (1972) A simplified classification of cultivated sorghum. Crop Sci 12:172–176

Henry AM, Damerval C (1997) High rates of polymorphism and recombination at the *Opaque-2* locus in cultivated maize. Mol Gen Genet 256:147–157

Ihaka R, Gentleman R (1996) R: a language for data analysis and graphics. J Comput Graph Stat 5:299–314

James MG, Denyer K, Myers AM (2003) Starch synthesis in the cereal endosperm. Curr Opin Plant Biol 6:215–222

Kraakman ATW, Niks RE, Van den Berg PMMM, Stam P, VanEuuwijk FA (2004) Linkage disequilibrium mapping of yield and yield stability in modern spring barley cultivars. Genetics 168:435–446

Larkin PD, McClung AM, Ayres NM, Park WD (2003) The effect of the *Waxy* locus (Granule Bound Starch Synthase) on pasting curve characteristics in specialty rices. Euphytica 131:243–253

Lopes MA, Larkins BA (1991) Gamma-zein content is related to endosperm modification in quality protein maize. Crop Sci 31:1655–1662

Maddaloni M, Donini G, Balconi C, Rizzi E, Gallusci P, Forlani F, Lohmer S, Thompson R, Salamini F, Motto M (1996) The transcriptional activator *Opaque-2* controls the expression of a cytosolic form of pyruvate orthophosphate dikinase-1 in maize endosperms. Mol Gen Genet 250:647–654

Manicacci D, Falque M, Le Guillou S, Piegu B, Henry A-M, Le Guilloux M, Damerval C, De Vienne D (2007) Maize *Sh2* gene is constrained by natural selection but escaped domestication. J Evol Biol 20:503–516

McIntyre CL, Drenth J, Gonzalez N, Henzell RG, Jordan DR (2008) Molecular characterization of the waxy locus in sorghum. Genome 51:524–533

Motto M, Hartings H, Maddaloni M, Lohmer S, Salamini F, Thompson R (1996) Genetic manipulation of protein quality in maize grains. Field Crop Res 45:37–48

Murray SC, Sharma A, Rooney WL, Klein PE, Mullet JE, Mitchell SE, Kresovich S (2008) Genetic improvement of sorghum as a biofuel feedstock: I. QTL for stem sugar and grain non-structural carbohydrates. Crop Sci 48:2165–2179

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

Myers AM, Morell MK, James MG, Ball SG (2000) Recent progress toward understanding biosynthesis of the amylopectin crystal. Plant Physiol 122:989–998

Nielsen R (2005) Molecular signatures of natural selection. Annu Rev Genet 39:197–218

Nishi A, Nakamura Y, Tanaka N, Satoh H (2001) Biochemical and genetic analysis of the effects of amylose-extender mutation in rice endosperm. Plant Physiol 127:459–472

Ollitrault P (1987) Evaluation génétique des sorghos cultivés (*Sorghum bicolor* L. Moench) par l'analyse conjointe des diversités enzymatique et morphophysiologique. Relations avec les sorghos sauvages. PhD Thesis, Université Paris XI Orsay, 187 pp

Olsen KM, Purugganan MD (2002) Molecular evidence on the origin and evolution of glutinous rice. Genetics 162:941–950

Perrier X, Jacquemoud-Collet JP (2006) DARwin software. Available at http://darwin.cirad.fr/darwin

Pirovano L, Lanzini S, Hartings H, Lazzaroni N, Rossi V, Joshi R, Thompson RD, Salamini F, Motto M (1994) Structural and functional analysis of an *Opaque-2*-related gene from sorghum. Plant Mol Biol 24:515–523

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

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

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

Rami JF (1999) Etude des facteurs génétiques impliqués dans la qualité technologique du grain chez le maïs et le sorgho. PhD thesis, Orsay University, 96 pp

Rami JF, Dufour P, Trouche G, Fliedel G, Mestres C, Davrieux F, Blanchard P, Hamon P (1998) Quantitative trait loci for grain quality, productivity, morphological and agronomical traits in sorghum (*Sorghum bicolor* L. Moench). Theor Appl Genet 97:605–616

Schultz JA, Juvik JA (2004) Current models for starch synthesis and the *sugary enhancer1* (*se1*) mutation in *Zea mays*. Plant Physiol Biochem 42:457–464

Séne M, Causse M, Damerval C, Thévenot C, Prioul J-L (2000) Quantitative trait loci affecting amylose, amylopectin and starch content in maize recombinant inbred lines. Plant Physiol Biochem 38:459–472

Séne M, Thévenot C, Hoffmann D, Bénétrix F, Causse M, Prioul J-L (2001) QTLs for grain dry milling properties, composition and vitreousness in maize recombinant inbred lines. Theor Appl Genet 102:591–599

Sine B (2003) Evaluation d'une core collection de sorgho en conditions de stress hydrique pré-floral. Master University Cheikh Anta Diop, Dakar, Sénégal, 67 pp

Skot L, Humphreys MO, Armstead I, Heywood S, Skot KP, Sanderson R, Thomas ID, Chorlton KH, Sackville Hamilton NR (2005) An association mapping approach to identify flowering time genes in natural populations of *Lolium perenne*. Mol Breed 15:233–245

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

Tian Z, Qian Q, Liu Q, Yan M, Liu X, Yan C, Liu G, Gao Z, Tang S, Zeng D, Wang Y, Yu J, Gu M, Li J (2009) Allelic diversities in rice starch biosynthesis lead to a diverse array of rice eating and cooking qualities. Proc Natl Acad Sci USA 106:21760–21765

Vasemägi A, Primmer RC (2005) Challenges for identifying functionally important genetic variation: the promise of combining complementary research strategies. Mol Ecol 14:3623–3642

Vigouroux Y, McMullen M, Hittinger CT, Houchins K, Schulz L, Kresovich S, Matsuoka Y, Doebley J (2002) Identifying genes of agronomic importance in maize by screening microsatellites for evidence of selection during domestication. Proc Natl Acad Sci USA 99:9650–9655

Wendorf F, Close AE, Schild R, Wasylikowa K, Housley RA, Harlan JR, Krolik H (1992) Saharan exploitation of plants 8000 years BP. Nature 359:721–724

Wilson LM, Whitt SR, Ibanez AM, Rocheford TR, Goodman MM (2004) Dissection of maize kernel composition and starch production by candidate gene association. Plant Cell 16:2719–2733

Yetneberk S, de Kock HL, Rooney LW, Taylor JRN (2004) Effects of sorghum cultivar on injera quality. Cereal Chem 81:314–321

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