Breeding Science 59: 217–227 (2009)

Genome-wide association mapping of quantitative traits in sorghum (*Sorghum bicolor* (L.) Moench) by using multiple models

Tariq Shehzad[1)], Hiroyoshi Iwata[2)] and Kazutoshi Okuno*[1)]

[1)] *Lab. of Plant Genetics and Breeding Science, Graduate School of Life and Environmental Sciences, University of Tsukuba*, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8572, Japan

[2)] *Data Mining and Grid Research Team, National Agricultural Research Center*, 3-1-1 Kannondai, Tsukuba, Ibaraki 305-8666, Japan

Association or linkage disequilibrium mapping is alternative to identify QTLs in plants. In this study we used SSR based sorghum diversity research set (SDRS) of 107 sorghum accessions. The representative set included a geographically diverse collection of accessions from 27 countries in Asia and Africa. For association analysis 98 sorghum SSR markers were selected from three previously published linkage maps. Phenotypic data was recorded for 26 morphological traits. Different association models were used to identify QTLs controlling major agronomic traits including both single QTL approaches as well as a multiple QTL approach. All models revealed loci having different strength of association with morphological traits. A total of 14 common significant SSR loci were identified by three different models of association analysis namely, single-QTL models with the effects of population structure, single-QTL models with the effects of population structure and familial relatedness, and multiple-QTL model with the effects of population structure. These loci were associated with 12 different morphological traits including days to heading, days to flowering, culm length, number of tillers, number of panicles and panicle length. Comparing results from different models may be an efficient way to detect reliable associations in the genome-wide association studies.

Key Words: core collection, SSR markers, morphological traits, linkage disequilibrium.

Introduction

Sorghum bicolor L. (Moench) is one of the most important cereal crops in the world. The small genome of sorghum and representativeness of tropical grasses because of "C4" photosynthesis makes it an attractive model for better understanding of the structure, function, and evolution of cereal genomes (Paterson 2008). Recently, the idea of utilization of association mapping in crop plants is gaining more attention than conventional linkage mapping. Sorghum is well suited to association mapping methodologies because of its medium-range patterns of linkage disequilibrium (Hamblin *et al.* 2005) and its self-pollinating mating system. Association mapping (a.k.a. linkage disequilibrium (LD) mapping) is a way to detect causal genes by exploiting LD which is non-random association of alleles at two or more loci. It exploits both historical recombination and genetic diversity for high resolution mapping.

Association mapping can be classified into two main categories (Chengsong *et al.* 2008). First one is candidate-gene association mapping. Here candidate genes are selected based on prior information from different ways e.g. mutational analysis, biochemical pathway or linkage analysis. It is trait-specific and low cost, but there is a chance to miss other unknown loci. Another is the genome-wide association mapping where genome-wide marker polymorphisms are used to study casual genetic variations. Although a large number of markers are necessary for detecting association with complex morphological traits in general, it does not require any prior information about candidate genes and there are chances to detect unknown loci. As an alternative to traditional linkage analysis, association mapping offers three advantages, (i) increased mapping resolution, (ii) reduced research time, and (iii) greater allele number (Yu and Buckler 2006). Since its introduction to plants (Thornsberry *et al.* 2001), association mapping has continued to gain favorability in genetic research because of advances in high throughput genomic technologies, interests in identifying novel and superior alleles, and improvements in statistical methods.

Pattern of LD is dependent on the occurrence of a new mutation that is associated with the variants on the chromosome on which it arises. Since recombination breaks the association, the rate of recombination is a key parameter in the process of LD decay. The pattern of LD is also affected by demographic factors, like population size, selection, migration, and founder effects. Therefore, the analysis of LD pattern is necessary to understand the feasibility and resolution of mapping based on LD (i.e., association mapping).

Complex breeding histories of many important crops

Communicated by M. Yano
Received January 17, 2009. Accepted May 16, 2009.
*Corresponding author (e-mail: okusan@sakura.cc.tsukuba.ac.jp)

have created the complex population structure in germplasm that hinders the application of association mapping in crop species (Flint-Garcia *et al.* 2003). The presence of population structure and unequal distribution of alleles within subpopulations can cause LD between unlinked genes and result in nonfunctional, spurious associations between a phenotype and unlinked candidate gene (Knowler *et al.* 1988, Lander and Schork 1994). To control false-positive and false-negative rate, Yu *et al.* (2006) proposed a mixed-linear-model method in which effects caused by population structure estimated by the model-based approach (Pritchard *et al.* 2000b) and background polygenic effects are included as independent variables. This method has been applied in various association mapping studies (e.g., Zhao *et al.* 2007, Stich *et al.* 2008). Iwata *et al.* (2007, 2009) combined a Bayesian multiple-QTL mapping approach with association mapping model including the effects of population structure, and demonstrated its power and precision of QTL detection in the genome-wide association study of rice germplasm.

In this study, we conducted genome-wide association studies in sorghum with multiple association mapping methods. We also analyzed the LD pattern in our germplasm collection to understand the feasibility and resolution of the association mapping study. In the association mapping study, we used two different types of approaches: the single QTL approaches proposed by Yu *et al.* (2006) and the multiple QTL approach proposed by Iwata *et al.* (2007). We compared the results obtained from different methods and tried to detect SSR loci that have strong association with morphological traits.

Materials and Methods

Plant materials and genomic DNA isolation

An SSR based sorghum diversity research set (SDRS) of 107 accessions (landraces) were selected from our previous study (Shehzad *et al.* 2009) which is preserved in National Institute of Agrobiological Sciences (NIAS), Genebank, Japan. The SDRS was developed from a geographically diverse base population of 320 sorghum landraces by the assessment of 38 SSR markers which were randomly selected from all linkage groups of sorghum. The representative set includes accessions from 27 geographically diverse countries selected from Asian and African regions. In SDRS, 25 accessions are from East Asia (Japan; 11, Korea; 7, Taiwan; 1, China; 6), two are from Southeast Asia (Cambodia; 1, Myanmar; 1), 26 from South Asia (India; 8, Pakistan; 13, Afghanistan; 2, Bangladesh; 1, Nepal; 2) and two accessions are from Southwest Asia (Iran; 1, Israel; 1). While the remaining 52 accessions are from African origin including Chad; 2, Congo; 1, Lesotho; 3, Morocco; 5, South Africa; 7, Central Africa; 1, Sudan; 11, Nigeria; 4, Algeria; 1, Uganda; 4, Ethiopia; 5, Kenya; 3, Zimbabwe; 3 and Tanzania; 2.

Leaves from 40 days old seedlings were cut and then subjected to vacuum freeze drying method for dehydration. Genomic DNA was extracted from leaf tissues using the CTAB method described by Murray and Thompson (1980) with some modification. Extraction buffer was composed of 2% CTAB, 50 mM Tris-HCl pH 8.0, 10 mM EDTA, 0.7 M NaCl, 0.1% SDS, 0.1 mg/ml Proteinase K, 2% insoluble PVP and 2% 2-mercaptoethanol. Chloroform extraction was performed to remove cellular debris and proteins by using chloroform-isoamyl alcohol (24 : 1 v/v), DNA was precipitated by the addition of 2-isopropanol and the precipitate was washed twice in 70%/90% ethanol. The final precipitate was dissolved in 50 µl of 1/10 TE solution containing RNase A, incubated at 42°C overnight, and stored at 4°C. The DNA concentration was measured by NanoDrop ND-1000 (Thermo scientific) spectrophotometer and diluted to a working concentration of 5 ng/µl.

Selection of SSR markers

Microsatellite primers were selected from published linkage maps of sorghum as revealed by Bhattramakki *et al.* (2000), Kong *et al.* (2000) and Taramino *et al.* (1997). All SSRs were screened by using eight diverse accessions and finally, a total of 98 markers were selected based on clear polymorphic banding patterns. The list of 98 sorghum microsatellite markers with linkage group (LG), sequence information, size range and other information are given in electronic supplementary material (ESM 1). Some of the SSRs used in this study have homology to known genes as previously described in Bhattramakki *et al.* (2000). Such as *Xtxp212* (LG-D) and *Xtxp34* (LG-C) have a high degree of homology to an expressed sequence tag derived from a gene coding a putative protein in *Arabidopsis thaliana* (L.), *Xtxp92* (LG-E) has homology to heat-shock-like protein gene in *Picea glauca* and *Xtxp100* (*Kaf*) in LG B has homology to *S. bicolor* kafirin gene/gene cluster. Two other SSR loci, *Xtxp38* (*lg*) in LG C and *Xtxp273* (Pbbf) in LG-H have a high degree of homology to other well-characterized genes. Similarly three markers are derived from gene loci; *Cba* (carbonic anhydrase) and *PepC* (phosphoenolpyruvate carboxylase) in LGs C and G, respectively, and *Kaf2* (Kafirin2) in LG-J (Bhattramakki *et al.* 2000). Summary statistics, including number of alleles, allele richness and gene diversity for all 98 SSR markers were calculated by using FSTAT software ver. 2.9.3 (available from http://www.unil.ch/izea/softwares/fstat.html), updated from Goudet (1995).

PCR conditions and electrophoresis

PCR amplification of sorghum SSRs were performed in 10 µl reaction mixture containing 10 ng DNA template, 10 × PCR buffer (Mg^{2+} concentration: 20 mM), 2 mM dNTPs, 25 ng of each primer and 0.02 U of Blend Taq Plus polymerase (Toyobo Co., LTD., Japan) enzyme in either Eppendorf Master cycler or Applied Biosystems 9700/2700 PCR system and Applied Biosystems 2720 thermal cycler. Annealing temperature was determined for all primers by using Eppendorf Master Cycler ep. gradient S. Thermal cycler protocol was set as denaturation at 98°C for 3 min, 30 cycles of 98°C (10 s), 60°C (30 s), and 72°C (30 s), followed by 7 min

at 72°C and then cooling at 4°C. PCR products were run on 10% polyacrylamide gel (10 cm in size) with constant supply of 200 V power, 500 mA current for 70 min to 120 min depending upon the size of PCR product. 10 × TBE buffer was used in making the gel while 1 × Tris Glycine Buffer was subjected to the tank. Gel was stained in ethidium bromide solution and photograph was taken by using Kodak Digital Science EDAS 290 ver. 3.6 with Kodak ID Image analysis software ver. 3.5. Different bands of the same SSR primer were grouped according to their respective sizes by comparing with 50 bp DNA size marker ladder and genotyping was done visually according to the format of different softwares used.

Phenotypic evaluations

The data obtained from the characterization of sorghum core set in Shehzad *et al.* (2009) were used here in association mapping. The selected set of 107 sorghum accessions were sown at NIAS field during growing season of year 2007. The statistical design used for field evaluation was Randomized Complete Block Design (RCBD) with two replications. From NIAS Genebank sorghum descriptors, 26 major traits were selected for phenotypic evaluation (ESM 2). Among them, continuous type of data was recorded for 12 phenotypic traits including, days to heading (DH), days to flowering (DF), days to maturity (DM), culm diameter (CD), grain weight per panicle (GWP), 100 grain weight (100GW), culm length (CL), number of tillers (NT), number of panicles (NP), panicle length (PL), leaf length (LL) and leaf width (LW). Similarly, panicle shape (PS), panicle type (PT), coleoptile's color (CC), quantity of lipid white powder on stem and leaves (LWP), color of midrib (CM), neck length of panicle (PNL), awn presence (AP), glume color (GC), growth in early stage (GES), endosperm type (ET), aphid resistance (AR), number of regenerated tillers (NRT), regrowth (RG) and resistance to insecticides (RI) were traits with categorical data.

Population structure and kinship matrix

The population structure among the 107 accessions by using the genotype data of SSR markers was estimated by using a program Structure ver. 2.2 (Pritchard *et al.* 2000a). The analysis was conducted on 49 markers that were selected, so that distances between adjacent markers were more than 10 cM in order to avoid using closely linked markers. The population structure was inferred with Bayesian clustering analyses with the admixture models in which the number of populations (J) ranged from 2 to 9. Markov chain Monte Carlo (MCMC) sampling was repeated 1×10^5 times after 1×10^4 cycles of a burn-in period. The optimal number of populations was determined on the basis of estimated logarithmic posterior probability of the Bayesian clustering. The analysis was repeated twice for each number of J. The posterior probability of $J=3$ was the largest among other values of J (ESM 3). Thus, we chose $J=3$ and obtained estimates for the proportion of accession i's genome that originated from population j, q_{ij}. The **Q** matrix, whose (i, j)-th element was q_{ij}, was further incorporated into the model of association mapping for both single and multiple QTL approaches. A kinship matrix, **K**, was calculated as allele sharing rates of the 89 SSR markers as suggested by Zhao *et al.* (2007), and used in the single-QTL approach. In the calculation of the kinship matrix, nine markers that had missing data for more than half of the accessions were eliminated.

Linkage Disequilibrium (LD)

LD between markers were estimated by D' and r^2, where D' is the standardized disequilibrium coefficient that is used for determining whether recombination or homoplasy has occurred between a pair of alleles; r^2 represents the correlation between alleles at two loci, and is informative for evaluating the resolution of association approaches. Statistical software TASSEL (Trait Analysis by aSSociation, Evolution and Linkage) ver.2.0.1 was used for this purpose (Bradbury *et al.* 2007). A weighted average of D' or r^2 was calculated between the two loci (Farnir *et al.* 2000) for all possible combinations of alleles, and then weighting them according to the allele's frequency. To test the significance of the LD, we also obtained P-values that were determined by permutation test to calculate the proportion of permuted gamete distributions, which were less probable than the observed gamete distribution under the null hypothesis of independence (Weir 1996). LD between the multi-allelic loci was measured by following Lewontin (1964) as:

$$D' = \sum_{i=1}^{u} \sum_{j=1}^{v} p_i q_j |D'_{ij}|$$

where u and v are the respective number of alleles at the two marker loci, pi and qj are the frequencies of marker allele i at locus A and marker allele j at locus B, and $|D'_{ij}|$ is the absolute value of normalized LD measure computed as:

$$D'_{ij} = \frac{D_{ij}}{D_{\max}} \text{ where;}$$

$$D_{\max} = \min [p(A_i)(1 - p(B_j)), (1 - p(A_i))p(B_j)] \quad if\ D > 0;$$

$$D_{\max} = \min [p(A_i)p(B_j), (1 - p(A_i))(1 - p(B_j))] \quad if\ D < 0$$

whereas, r^2 was calculated as described by Hill and Robertson (1968):

$$r^2 = \sum_{i=1}^{u} \sum_{j=1}^{v} p(A_i)p(B_j)r_{ij}^2$$

$$\text{where } r_{ij}^2 = \frac{D_{ij}^2}{p(A_i)(1 - p(A_i))p(B_j)(1 - p(B_j))}$$

Statistical models for Association analysis

a) Single QTL approach

Two different types of models, i.e., general linear model (GLM) and mixed linear model (MLM), were used for the single QTL method using TASSEL ver. 2.0.1 software. In GLM, we used two different models: (i) the model with no control for population structure and relatedness (hereafter, it

is called as naive model) and (ii) the model with population structure (hereafter, it is called as Q model) (Yu *et al.* 2006). In the second model, we used **Q** matrix estimated by the Structure analysis for controlling the effect caused by population structure. In MLM, we used two models: (i) the model that accounted for familial relatedness between accessions (hereafter, it is called as K model) and (ii) the model that takes into account both the population structure and the familial relationship (hereafter, it is called as Q+K model). The MLM approach was shown to be superior to more conventional linear models in association analyses (Yu *et al.* 2006).

b) Multiple QTL approach

A Bayesian model proposed by Iwata *et al.* (2009) was used in the multiple-QTL association mapping. The model is similar to one proposed by Iwata *et al.* (2007), but can be also applied to multi-allelic marker data, like SSR. In this study, the prior for the number of QTL (N_Q) was defined differently from Iwata *et al.* (2009) as $p(N_Q) = {}_K\mathrm{C}_{N_Q}(\lambda/K)^{N_Q}(1-\lambda/K)^{K-N_Q}$, where λ is the expected number of QTLs included in the model and K is the number of markers. This prior was expected to be conservative than the previous prior (Iwata *et al*. 2009), because we observed that it suppressed a greater number of significant markers than the previous prior. The model included the **Q** matrix as independent variables for controlling the effects caused by population structure. All the parameters in the model were estimated by MCMC sampling as described in Iwata *et al.* (2007, 2009). The hyper-parameters of prior distributions of parameters were set as $\upsilon_\beta = 4$, $s_\beta^2 = 0.04$, $\upsilon_e = -2$, $s_e^2 = 0$, and $\lambda = 10$. MCMC cycles were repeated 6×10^4 and the first 1×10^4 cycles (burn-in) were not used for estimating parameter values. In the MCMC cycles, missing genotypes were sampled based on allele frequencies observed in non-missing genotypes. Sampling of parameters in the model was carried out every ten cycles to reduce serial correlation so that the total number of samples we retained was 5×10^3. In the model, each marker position k (k = 1, 2, ..., K) has its own indicator variable γ_k, where the value one ($\gamma_k = 1$) corresponds to the case in which the marker is included in the model as a QTL representative, and the value zero ($\gamma_k = 0$) implies exclusion. We used the posterior average of γ_k for determining significant markers. That is, the markers that had γ_k larger than the specific threshold, 0.5, were regarded as significant. This threshold corresponds to the "moderate" threshold in Iwata *et al.* (2007), and was expected to have smaller false negative rate as well as larger false positive rate than a stricter threshold like 0.9 (Iwata *et al*. 2007, 2009). Despite the problem of large false positive rate, we employed the moderate threshold because of the small sample size used in this study. Under the strict threshold 0.9, we may miss a number of true associations between QTLs and markers because of low statistical power caused by the small sample size. In this study, we tried to detect reliable associations by comparing significant markers between different methods. In other words, we assumed that associations detected with both single-and multiple-QTL models were reliable.

Results

Patterns of diversity in SDRS

Results obtained from previous study (Shehzad *et al.* 2009) showed a wide range of phenotypic diversity in SDRS for all the traits studied. The accessions showed a highly significant difference for the morphological traits with continuous type numerical data including DH, DF, DM, CD, GWP, 100GW, CL, NT, NP, PL, LL and LW when tested for analysis of variance (ANOVA). Correlations among the traits showed that DH, DF and DM were highly significantly correlated with most of the traits. However, NT was only highly significantly correlated with CD and the remaining combinations were not-significant. Similarly for 14 morphological traits with categorical data, the core set (107 accessions) was distributed with different frequencies according to the ranks of sorghum descriptors list. The accessions could be divided into five geographic regions such as; East Asia, Southeast Asia, South Asia, Southwest Asia and Africa. According to Pearson's chi-square test, the geographic regions were found to be independent for six traits including ET, RI, CC, PS, PT and AP while for other traits (GES, CM, AR, LWP, NRT, RG, PNL and GC) they were positively associated with each other.

In the assessments of the 98 SSR markers, a total of 470 alleles were observed and the alleles could specifically classify 107 sorghum accessions of diverse origin. The number of alleles observed in each locus ranged from two to 10 with the average of 4.79. In the Structure analysis, we found $J = 3$. The accessions in SDRS were assigned to one of the three populations in which they have the highest probability of membership estimated in the Structure analysis when $J = 3$. The first population contained 33 sorghum accessions mostly from Africa origin except for four accessions from Asia. The second population contained 36 accessions, including 23 from Africa and 13 from Asia. The third population contained 38 accessions form East and South Asia. The three groups were classified according to their geographic origins and are similar to the three clusters observed in our previous report (Shehzad *et al.* 2009). The summary statistics of the 98 SSR markers over the whole accessions and three estimated populations are given in ESM 4.

Linkage disequilibrium (LD) plot

A short to medium range of LD (i.e., LD across chromosomes) was observed in this germplasm. The triangle plot for pairwise LD between marker sites in a hypothetical genome fragment, where pairwise LD values of polymorphic sites were plotted on both the X- and Y-axis; above the diagonal displays r^2 values and below the diagonal displays the corresponding P-values from rapid 1000 shuffle permutation test (Fig. 1). Each cell represents the relationship between two markers with the color codes indicating the significance of LD. Maximum number of SSR markers with highly significant LD ($P < 0.0001$) were situated in LGs A and B (marker index 1–41). On the other hand, a considerable degree of LD between markers closely locating

on the chromosomes was not obvious. Being a predominantly self-pollinating species, sorghum is expected to show higher levels of LD than outcrossing species like maize, reported by Tenaillon *et al*. (2001) and Remington *et al*. (2001). Fig. 1 shows that LD generally decayed rapidly with distance between sites within loci, but there was substantial variation among loci.

Association by GLM model

The association analysis by using the model without population structure and kinship detected a large number of markers suggesting associations between genotypes and phenotypes. This model had no control for the heterogeneity of genetic background (*i.e.*, population structure and familial relatedness among accessions) and thought to be affected largely by false positives. A total of 76 significant loci were identified to be associated with 20 morphological traits. On the other hand, in Q model, only 15 SSR loci had strong association at $P < 0.001$ (Table 1). Among them *Xtxp316* in LG-A and *Xtxp104* in LG-I were associated with GES. Locus *Xtxp23* in LG-J had association with DH. The same locus (*Xtxp23*) was moderately associated ($P < 0.01$) with DF and DM. Similarly *Xtxp88* (LG-A), *Xtxp24* (LG-D) and *Xtxp278* (LG-E) were found to be associated with presence of lipid white powder on stem and leaves (LWP). Two SSR loci; *Xtxp76* and *Xtxp88* in LG-A showed strong association with CL. A single locus *Xtxp100* (LG-B) and *Xtxp38* (LG-C) were identified as significant for LW and PS, respectively. Multiple number of SSR loci had strong association with PL by using this model, namely, *Xtxp335*, *Xtxp302* and *SbAGF06* in LG-A, *Xtxp7* in LG-B and *Xtxp10* in LG-E.

Association by MLM model

In K model, 35 SSR loci had strong association ($P < 0.001$) with 14 morphological traits. The number of significant markers was second-largest after naive model. Similarly a total of 33 markers were detected as significantly associated with 16 traits in Q + K method. The *P*-value for associations between SSR markers and morphological traits in Q + K model are shown in Fig. 2 (i). A single locus; *Xtxp316* (LG-A), *Xtxp14* (LG-J), *Xtxp18* (LG-H), *Xtxp100* (LG-B) and *Xtxp38* (LG-C) were associated with GES, AR,

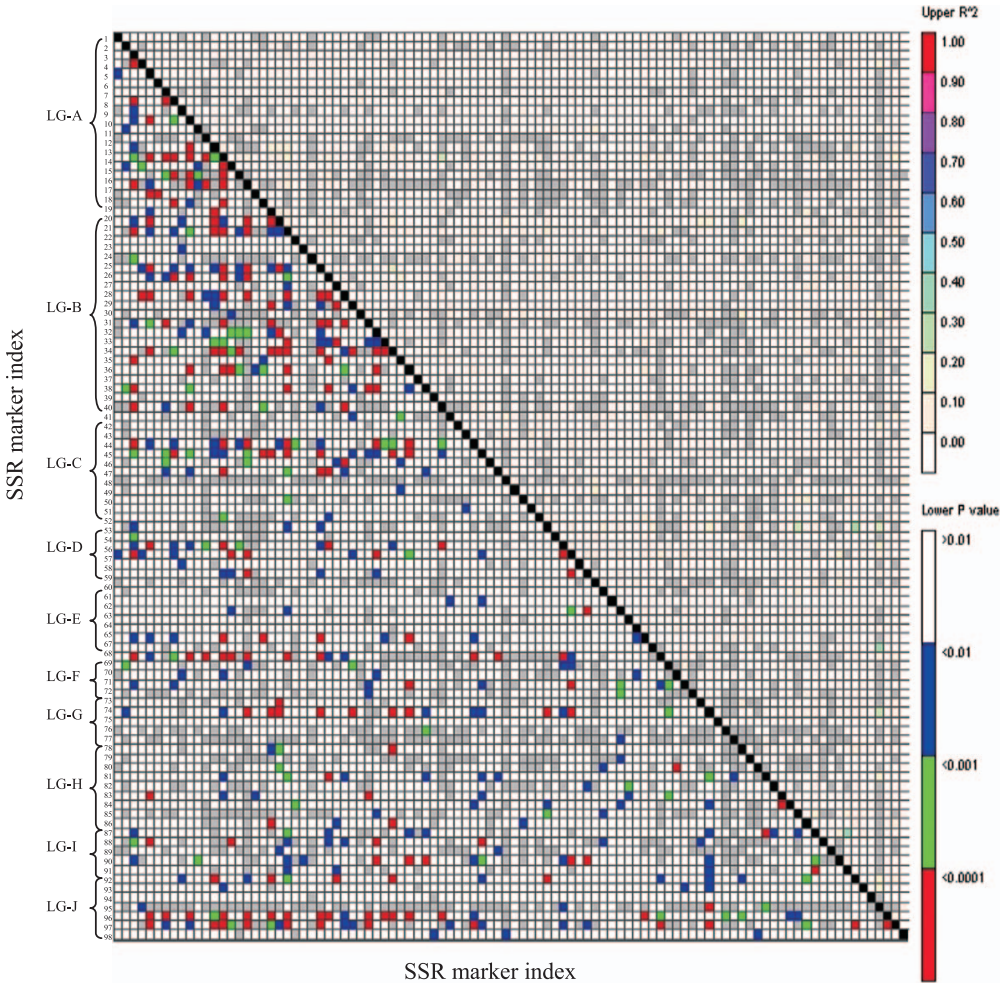

Fig. 1. LD plot generated by 98 SSR markers. Each cell represents the relationship between two markers with the color codes for the presence of significant LD. Colored bar code for the significance threshold levels.

Table 1. Significant SSR loci associated with different traits as identified by single-QTL (Q, Q+K) models and multiple-QTL model with **Q** matrix

Trait	Model	LG	Location[a]	Marker[b]	$-\text{Log}_{10}(P)^{c}/\gamma^{d}$
ET	Q+K	A	5.0	Xtxp316	3.068
		B	148.7	Xtxp315	4.477
GES	Q	A	5.0	**Xtxp316**	3.000
		I	47.7	Xtxp104	3.000
	Q+K	A	5.0	**Xtxp316**	3.682
	Multiple-QTL	A	5.0	**Xtxp316**	0.646
CM	Q+K	B	171.5	Xtxp296	3.033
		C	92.1	**Xtxp31**	3.590
		D	135.0	Xtxp21	3.305
		H	104.1	Xtxp105	4.315
	Multiple-QTL	C	92.1	**Xtxp31**	0.743
DH	Q	J	69.9	**Xtxp23**	3.000
	Q+K	E	42.5	Xtxp312	3.407
		J	69.9	**Xtxp23**	3.880
	Multiple-QTL	J	69.9	**Xtxp23**	0.698
DF	Q+K	J	69.9	**Xtxp23**	3.816
	Multiple-QTL	J	69.9	**Xtxp23**	0.705
AR	Q+K	J	48.7	Xtxp14	4.570
LWP	Q	A	117.4	**Xtxp88**	3.000
		D	95.6	Xtxp24	3.000
		E	73.2	**Xtxp278**	3.000
	Q+K	A	117.4	Xtxp88	3.195
		D	95.6	Xtxp24	4.313
		E	73.2	**Xtxp278**	4.349
	Multiple-QTL	A	117.4	**Xtxp88**	0.698
		E	73.2	**Xtxp278**	0.826
CL	Q	A	60.0	**Xtxp75**	3.000
		A	117.4	**Xtxp88**	3.000
	Q+K	A	60.0	**Xtxp75**	3.428
		A	117.4	**Xtxp88**	3.154
		C	8.0	Xtxp69	3.479
	Multiple-QTL	A	60.0	**Xtxp75**	0.500
PS	Q	C	30.2	**Xtxp38**	3.000
	Q+K	C	30.2	**Xtxp38**	3.190
AP	Q+K	F	86.6	**Xtxp67**	3.509
	Multiple-QTL	F	86.6	**Xtxp67**	0.637
NP	Q+K	B	38.2	Xtxp211	3.051
		B	41.2	Xtxp84	4.871
		B	160.4	**Xtxp100**	10.923
	Multiple-QTL	A	90.1	Xtxp37	0.544
		B	160.4	**Xtxp100**	1.000
		G	53.6	Xtxp331	0.667
		J	69.9	Xtxp23	0.510
100GW	Q+K	J	69.9	Xtxp23	3.515
PL	Q	A	86.7	**Xtxp335**	3.000
		A	174.0	**Xtxp302**	3.000
		A	76.5	**SbAGF06**	3.000
		B	163.7	**Xtxp7**	3.000
		F	63.5	**Xtxp10**	3.000
	Q+K	A	60.0	Xtxp75	3.473
		A	86.7	**Xtxp335**	5.387
		A	174.0	**Xtxp302**	4.281
		A	76.5	**SbAGF06**	6.377
		B	163.7	**Xtxp7**	3.407
		F	63.5	**Xtxp10**	6.316
	Multiple-QTL	A	86.7	**Xtxp335**	0.859
		A	127.8	SbAGB02	0.994
		B	163.7	**Xtxp7**	0.551
LL	Q+K	H	91.0	Xtxp18	3.553
NT	Q+K	B	41.2	Xtxp84	3.748
		B	160.4	**Xtxp100**	9.436
		E	0.0	SbAGE03	3.324
	Multiple-QTL	A	90.1	Xtxp37	0.963
		B	160.4	**Xtxp100**	0.968
		B	196.8	Xtxp8	0.564
		G	53.6	Xtxp331	0.965
LW	Q	B	160.4	**Xtxp100**	3.000
	Q+K	B	160.4	**Xtxp100**	3.597

[a] Location of SSRs on linkage map as described in (Bhattramakki *et al.* 2000, Kong *et al.* 2000 and Taramino *et al.* 1997).
[b] Markers with bold face shows significant association with traits as revealed by any two of Q, Q+K and multiple-QTL models.
[c] –Log10 of *P*-values determined for Q and Q+K model with 3.0 as threshold value for strong association.
[d] Multiple-QTL model; mean of posterior distribution of γ_k with 0.5 as threshold value.

LL, LW and PS, respectively. Locus *Xtxp23* (LG-J) was found to be strongly associated with DF as well as 100 GW. *Xtxp67* in LG-F was found to be strongly associated with AP. Presence or absence of awn is supposed to be controlled by a single recessive gene in wheat and other cereal crops (Tsunewaki 1983). This model identified two SSR loci for DH (*Xtxp312*; LG-E, *Xtxp23*; LG-J), two loci for ET (*Xtxp316*; LG-A, *Xtxp315*; LG-B), three loci for LWP (*Xtxp88*; LG-A, *Xtxp24*; LG-D, *Xtxp278*; LG-E), three loci for CL (*Xtxp75*; LG-A, *Xtxp88*; LG-A, *Xtxp69*; LG-C) and four loci for CM (*Xtxp296*; LG-B, *Xtxp31*; LG-C, *Xtxp21*; LG-D, *Xtxp105*; LG-H). The maximum number of significant SSR loci was observed for PL, including *Xtxp75*, *Xtxp335*, *Xtxp302* and *SbAGF06* in LG-A, *Xtxp7* in LG-B and *Xtxp10* in LG-F.

Multiple-QTL model

The strength of association between SSR loci and traits detected by the multiple-QTL model is shown as the posterior mean of γ_k (Fig. 2. ii). In this study, SSR loci with posterior mean of γ_k above 0.5 were considered as significant. A total of 19 SSRs, associated with 10 morphological traits were significant (Table 1). Among them single QTL *Xtxp340* (LG-A) was found to be associated with GES, *Xtxp31* (LG-C) with CM, *Xtxp75* (LG-A) with CL and *Xtxp23* (LG-J) with both DH as well as DF. One of the quantitatively inherited trait i.e., awn presence (AP) was found to be associated with a locus named *Xtxp67* in LG-F. In this model of multiple QTL association analysis, two SSR loci had significant association with LWP (*Xtxp88*; LG-A, *Xtxp278*; LG-E), three SSR loci for PL (*Xtxp335* and *SbAGB02*; LG-A, *Xtxp7*;

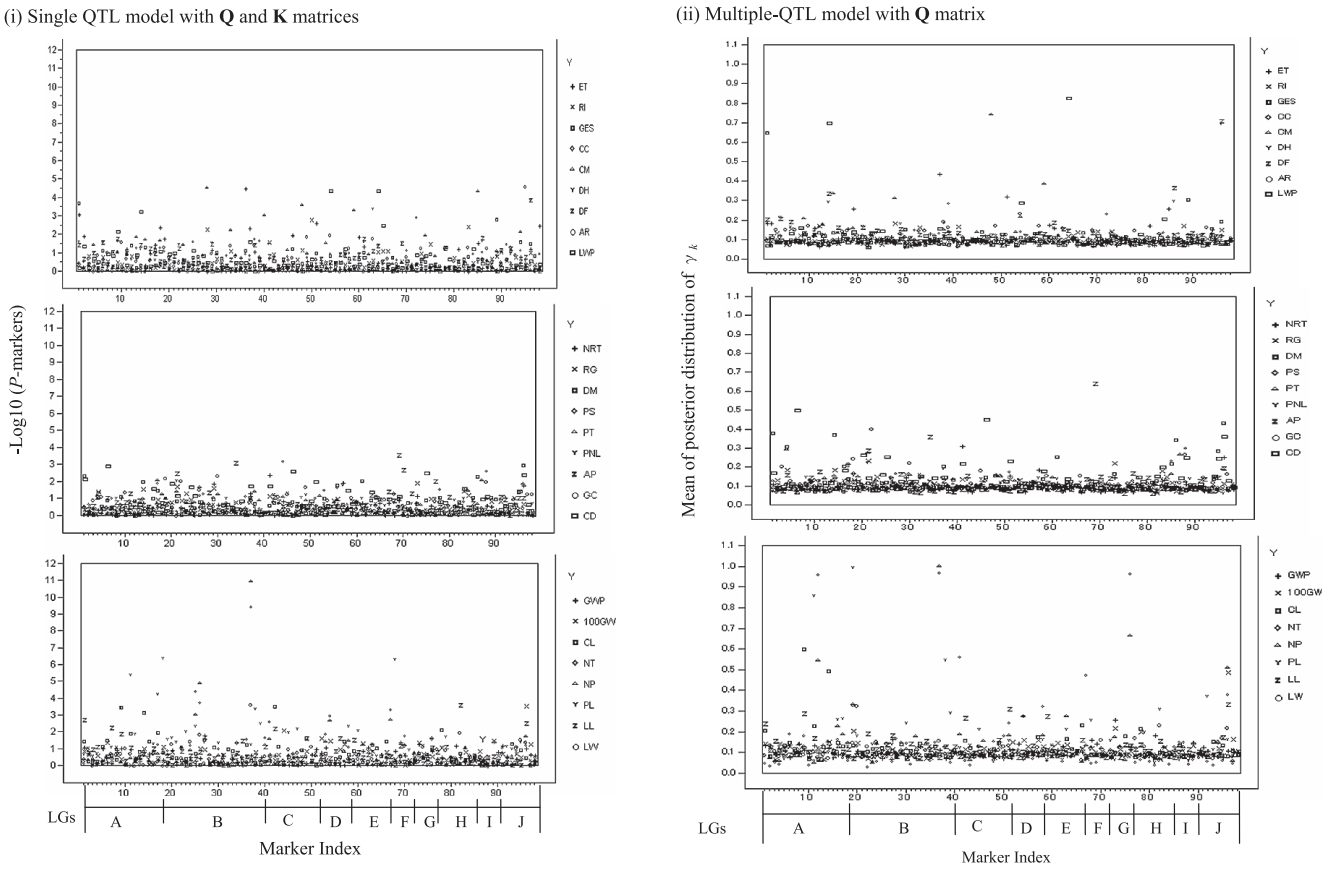

Fig. 2. Comparison between single-QTL model with Q and K matrices and multiple-QTL model with **Q** matrix. i) $-\text{Log}_{10}$ (*P*-marker) values of 98 SSR markers determined for 26 morphological traits. ii) Mean of posterior distributions of γ_k of 98 SSR markers, estimated for 26 traits.

LG-B), four loci for NT (*Xtxp37*; LG-A, *Xtxp100* and *Xtxp8*; LG-B, *Xtxp331*; LG-G) and four loci were significant for NP (*Xtxp37*; LG-A, *Xtxp100*; LG-b, *Xtxp331*; LG-G, *Xtxp23*; LG-J). The strength of association of the loci with traits depends upon the mean values of posterior distribution of γ_k and β_k. With threshold of 0.9, three QTLs had association with NT while one locus with PL. *Xtxp100* (Kaf) showed a maximum value of posterior mean distribution i.e. $\gamma_k = 1$ for NP, which shows that the marker had been always included in the model during MCMC sampling process as a QTL.

Comparison among different statistical models

To evaluate the possibility of false positives in association models, we plotted observed *P*-values against expected *P*-values as described by Stich *et al.* (2008) (Fig. 3). As a result, naive model showed the highest deviation from the $y = x$ line than K, Q and Q + K models, indicating the method might detect a larger number of false positives than the others. The K model was better than naive but still showed the deviation from the $y = x$ line in comparison with Q and Q + K models. Casa *et al.* (2008) also showed the performance of models in sorghum that account for both population structure and kinship better than those controlled only for Q or K while Zhao *et al.* (2007) reported the same result in *Arabidopsis.* From Fig. 3, the Q and Q + K models showed the smallest possibility of false negatives among single-QTL models.

The results obtained from naive and K models detected large number of markers associated with different morphological traits. The results, however, might be come from a large number of false positives, and was not comparable to ones obtained from the other methods. The Q and Q + K models and multiple-QTL model gave comparable results. After taking consensus among methods, a total of 14 loci were identified either by any two or all of the three models that have strong association with 12 morphological traits (Table 1). Among them, a single SSR locus *Xtxp67* in LG-F was identified by Q + K model and multiple-QTL model as significantly associated with AP. For CM, locus *Xtxp201* (LG-B) was identified by Q + K model and multiple-QTL model. All the three models of association analysis identified *Xtxp23* (LG-J) as associated with DH while the same locus had association with DF in Q + K model and multiple-QTL model. One of the morphological traits, growth in early stage (GES) had strong association with *Xtxp316* in LG-A, as revealed by both Q + K model and multiple-QTL model. For LWP, two SSR loci (*Xtxp88*; LG-A, *Xtxp278*; LG-E) were found significant in all three of the models while locus *Xtxp24* (LG-D) was identified only by Q and Q + K models. Similarly, for CL one locus *Xtxp75* (LG-A) was associated

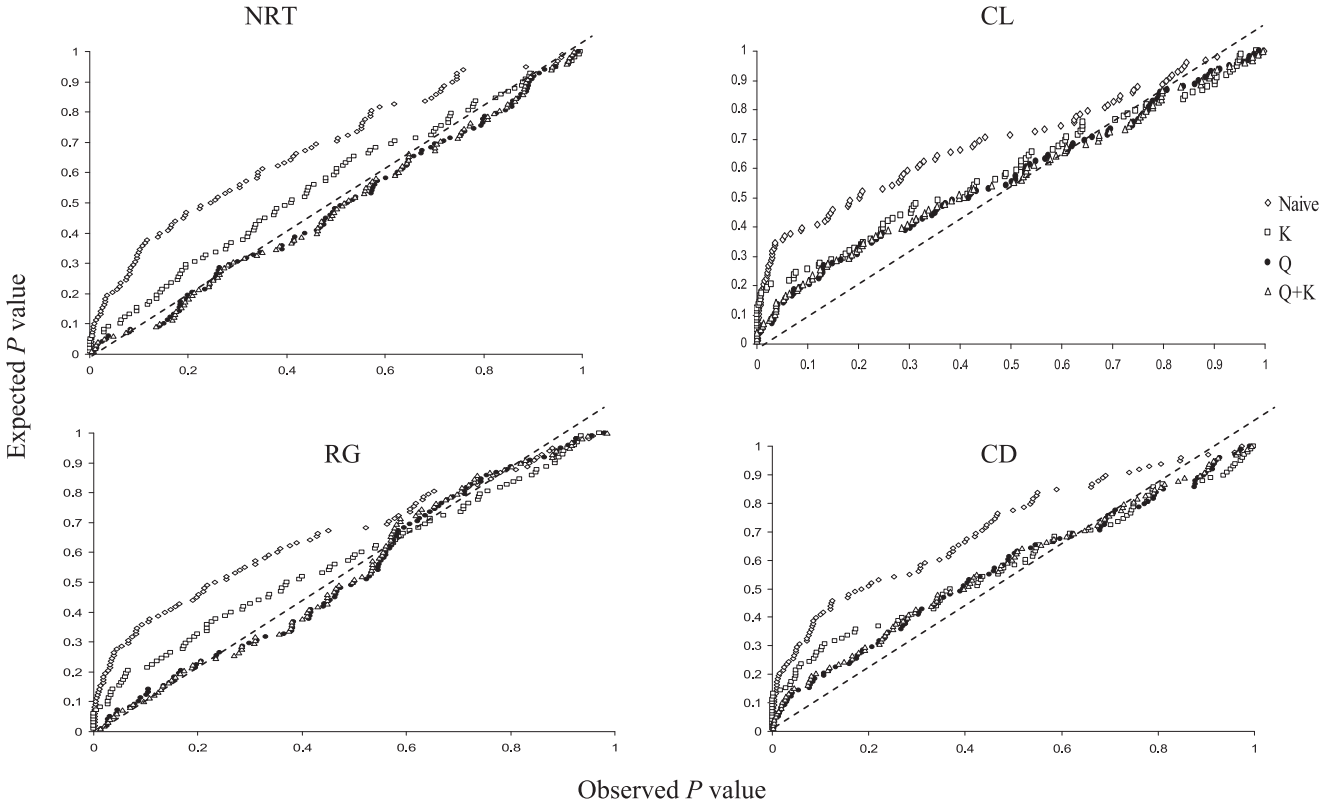

Fig. 3. Comparison among association models for the control of false positives in four traits; Observed *P*-values vs. Expected *P*-values.

in Q and Q + K models and multiple-QTL model, while *Xtxp88* (LG-A) had association in Q and Q + K models. A single locus *Xtxp100* (LG-B) was found to be strongly associated with two morphological traits (NT; NP) in Q + K model and multiple-QTL model. Panicle length (PL) was controlled by two common SSRs (*Xtxp335*; LG-A, *Xtxp7*; LG-B) in three models of association while three loci (*Xtxp302* and *SbAGF06*; LG-A, *Xtxp10*; LG-F) identified by Q and Q + K models.

Discussion

Association mapping is a powerful tool for fine mapping of quantitative traits and is dependent on the structure of linkage disequilibrium of alleles at different loci (Flint-Garcia *et al.* 2003). Association analysis is strongly affected by both false positives as well as false negatives. In this study we have used different models for association mapping to control both "false positives" and "false negatives". Some of the significant markers showed same level of association in all the models, while in some cases same markers identified with different level of significance by different models.

Choice of germplasm is one of the key factors determining the resolution of association mapping in plants. To detect more alleles, germplasm selected should include maximum diversity of the genepool with more extensive recombination in the history to allow a high level of resolution. The representative set of accessions used here retained more than 90% genetic diversity of its base population assessed by sorghum SSR markers (Shehzad *et al.* 2009). This type of germplasm is considered an ideal material for association mapping (Whitt and Buckler 2003).

The success of association mapping depends upon the possibility of detecting LD between marker alleles and alleles affecting the expression of phenotypic traits (Stich *et al.* 2005). SSR markers has more affinity towards genome-wide association mapping than either amplified fragment length polymorphism (AFLP) markers (Stich *et al.* 2006) or single-nucleotide polymorphisms (SNPs) (Remington *et al.* 2001). In this study, we found a wide-range of LD, which ranged over chromosomes, whereas a short-range of LD between markers closely locating on the same chromosome was not obvious. A wide range of LD might be caused by population structure, and might be responsible for a large number of false positives when the association mapping models did not take into account the population structure (i.e., in the naive and K models). On the other hand, a short range of LD is caused mainly by the physical linkage on the chromosome. Low LD in a short range may indicate that marker density in this study is not enough for detecting QTLs in a genome-wide manner. Thus, many QTLs might be missed because of the low density of markers used in the study, although some markers still captured the signal of QTL even in this density. For the genome-wide association study with sorghum germplasms, we should use much larger number of markers in the future.

The naive and K models, which did not control the effects caused by population structure, detected a larger number of significant associations between markers and traits. These models showed large discrepancy of the observed *P*-values from the expected *P*-values, indicating these models were affected by a larger number of spurious associations in comparison with the other models. For example, in two qualitative traits (single gene control), endosperm type (ET) and awn presence (AP), a large number of loci were detected as significant by these two models, whereas only a single locus was detected by Q + K and multiple-QTL models. When the population structure was taken into account in a single-QTL model (i.e., Q and Q + K models), a smaller discrepancy from the uniform distribution of *P*-values was observed. Moreover, the results obtained from multiple QTL model were more similar to Q + K than Q models. This may indicate the familial relatedness (i.e., kinship) should also be taken into account in the model for association mapping. The results are in accordance with the findings of Casa *et al.* (2008) in sorghum, indicating the population structure should be taken into account in the association mapping in sorghum.

For both NT and NP, *Xtxp100* (Kaf) was identified to be highly significant ($\gamma_k = 1$) by multiple-QTL model. This marker is also identified by the Q + K model with strong association for these two traits, whereas Q model identified it as weakly associated with NT and as moderately associated with NP. Similarly, the locus *Xtxp23* had association with three correlated traits i.e., DH, DF and DM, as identified by all these three models. In multiple-QTL model, this locus was found to present at the similar posterior mean of γ_k for DH and DF. These correspondences among results from different methods indicate the reliability of the detected associations.

Association mapping is also useful for identification of genes controlling qualitative traits. In this study, one locus named *Xtxp67* (LG-F) was mapped by the Q + K model and multiple-QTL model as associated with AP (a qualitative trait) in sorghum. Here, however, we treated as qualitative traits as quantitative traits by scoring the qualitative variation in an ordinal way. More sophisticated methods, however, will improve the power and precision of the association mapping of qualitative traits (Iwata *et al.* 2009). The comparison between single-QTL and multiple-QTL approaches of association analysis is reported for the first time in our study. Association mapping is a new tool for identifying complex traits in plant species. We tried to detect reliable associations by comparing significant markers among different methods and assumed that associations detected by both single-and multiple-QTL models were more reliable.

In sorghum, some reports have been published regarding linkage mapping and identification of QTLs responsible for some important traits including yield, maturity, photoperiod sensitivity, resistance/tolerance to biotic and abiotic stresses etc. Some of the previously identified QTLs by using linkage mapping in sorghum for different traits are in accordance to the findings of association studies reported here. Hart *et al.* (2001) reported several QTLs identified by using linkage mapping for different morphological traits. The positions of two QTLs responsible for height of main culm in LG-A (50 and 90 cM) and two QTLs for panicle length in LG-F (100 and 104 cM) are in accordance with the findings reported in this study such as, two QTLs in LG-A *Xtxp75* (60 cM) and *Xtxp88* (117.4 cM) were associated with CL whereas, *Xtxp10* (63.5) in LG-F had significant association with PL. The QTLs reported for number of basal tillers with heads per plant and number of basal tillers per basal tillered plant reported in Hart *et al.* (2001) are inconsistence with the results of association mapping, similarly a single gene for awn presence/absence was mapped in LG-C, which is in contrary to our findings (i.e., LG-F). Chantereau *et al.* (2001) identified three QTLs on LGs C, F and H controlling photoperiod response in sorghum which is contradictory to the finding of association analysis in this study (i.e., LG-J). There are several possible causes for the discordance between this study and previous QTL mapping studies. One of them is that this study may not have detected all the existing major QTLs because of the small number of markers and accessions used in the study. Another cause is that a major QTL detected in bi-parental-cross QTL mapping may not have large effect in the phenotypic variation of a germplasm collection and may be difficult to be detected with association mapping.

Feltus *et al.* (2006) have aligned genetic maps obtained from two different sorghum populations and detected 61 new QTLs from 17 traits. Among them single QTL for awn length was found at a position of 66.5–92.4 cM in LG-F with four co-localized RFLP markers including *Xucm58.2* (86.3 cM) which is almost at same position of the locus identified in this study i.e., *Xtxp67* (86.6 cM) associated with AP. Similarly, two of the QTLs for culm length at the positions of 31.0–77.0 cM and 127.1 cM in LG-A, two QTLs for days to flowering in LG-J at position of 90.0–132.4 cM and 102.6–132.4 cM, one for leaf width at 100.80–134.7 cM in LG-B and a single QTL for seed weight in LG-A at the position of 75.7–113 cM were identified (Feltus *et al.* 2006). Comparing with the results of association mapping in sorghum reported here, QTLs identified for these traits falls in the same positions on chromosomes which shows the accuracy of the models and its reliability. For example, *Xtxp75* (60.0 cM) and *Xtxp88* (117.4 cM) in LG-A associated with CL, *Xtxp23* (69.9 cM) in LG-J for DF, *Xtxp100* (160.4 cM) for LW while for 100GW two loci *Xtxp75* (60.0 cM) and *Xtxp88* (117.4 cM) were found to be significantly associated.

There have been only a few reports on linkage disequilibrium and association mapping in sorghum. For example, Hamblin *et al.* (2005, 2007) reported the patterns and prospects of LD in sorghum while Casa *et al.* (2008) compared different models of association mapping in a panel of 377 sorghum accessions. Recently, Brown *et al.* (2008) utilized the same panel of sorghum accessions from the study of Casa *et al.* (2008) in association mapping to characterize the phenotypic effects of the *dw3* (dwarfing gene) mutation.

The SDRS used in this study showed a wide range of genetic as well as phenotypic diversity and is suitable for association mapping. Although for the genome-wide association study a huge number of molecular markers are necessary, our study can serve as initial effort for the association mapping studies in sorghum. We employed different models for association mapping in sorghum. By comparing multiple models for association mapping, we might be able to find reliable associations between markers and traits.

Literature Cited

Bhattramakki,D., J.Dong, K.A.Chhabra and G.E.Hart (2000) An integrated SSR and RFLP linkage map of *Sorghum bicolor* (L.) Moench. Genome 43: 988–1002.

Bradbury,P.J., Z.Zhang, D.E.Kroon, T.M.Casstevens, Y.Ramdoss and E.S.Buckler (2007) TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics 23: 2633–2635.

Brown,P.J., W.L.Rooney, C.Franks and S.Kresovich (2008) Efficient mapping of plant height quantitative trait loci in a sorghum association population with introgressed dwarfing genes. Genetics 180: 629–637.

Casa,A.M., G.Pressoir, P.J.Brown, S.E.Mitchell, W.L.Rooney, M.R. Tuinstra, C.D.Franks and S.Kresovich (2008) Community resources and strategies for association mapping in sorghum. Crop Sci. 48: 30–40.

Chantereau,J., G.Trouche, J.F.Rami, M.Deu, C.Barro and L.Grivet (2001) RFLP mapping of QTLs for photoperiod response in tropical sorghum. Euphytica 120: 183–194.

Chengsong,Z., G.Michael, E.S.Buckler and J.Yu (2008) Status and Prospects of Association Mapping in Plants (Review and Interpretation). The Plant Genome doi: 10.3835/plantgenome2008.02.0089

Farnir,F., W.Coppieters, J.J.Arranz, P.Berzi, N.Cambisano, B.Grisart, L.Karim, F.Marcq, M.Mni, C.Nezer, P.Simln, P.Vanmanshoven, D.Wagenaar and M.Georges (2000) Extensive genome-wide linkage disequilibrium in cattle. Genome Res.10: 220–227.

Feltus,F.A., G.E.Hart, K.F.Schertz, A.M.Casa, S.Kresovich, S. Abraham, P.E.Klein, P.J.Brown and A.H.Paterson (2006) Alignment of genetic maps and QTLs between inter- and intra-specific sorghum populations. Theor. Appl. Genet. 112: 1295–1305.

Flint-Garcia,S.A., J.M.Thornsberry and E.S.Buckler (2003) Structure of linkage disequilibrium in plants. Annu. Rev. Plant Biol. 54: 357–374.

Goudet,J. (1995) FSTAT (vers. 1.2): a computer program to calculate F-statistics. J. Hered. 86: 485–486.

Hamblin,M.T., G.Maria, S.Fernandez, A.M.Casa, S.E.Mitchell, A.H. Paterson and S.Kresovich (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,M.T., M.G.S.Fernandez, S.E.Mitchell, M.R.Tuinstra, W.L. Rooney and S.Kresovich (2007) Sequence variation at candidate loci in the starch metabolism pathway in sorghum: prospects for linkage disequilibrium mapping. Crop Sci. 47: 125–134.

Hart,G.E., K.F.Schertz, Y.Peng and N.H.Syed (2001) Genetic mapping of Sorghum bicolor (L.) Moench QTLs that control variation in tillering and other morphological characters. Theor. Appl. Genet. 103: 1232–1242.

Hill,W.G. and A.Robertson (1968) Linkage disequilibrium in finite populations. Theor. Appl. Genet. 38: 226–231.

Iwata,H., Y.Uga, Y.Yoshioka, K.Ebana and T.Hayashi (2007) Bayesian association mapping of multiple quantitative trait loci and its application to the analysis of genetic variation among *Oryza sativa* L. germplasms. Theor. Appl. Genet. 114: 1437–1449.

Iwata,H., K.Ebana, S.Fukuoka, J.L.Jannink and T.Hayashi (2009) Bayesian multilocus association mapping on ordinal and censored traits and its application to the analysis of genetic variation among Oryza sativa L. germplasms. Theor. Appl. Genet. doi: 10.1007/s00122-008-0945-6.

Knowler,W.C., R.C.Williams, D.J.Pettitt and A.G.Steinberg (1988) GM3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. Am. J. Hum. Genet. 43: 520–526.

Kong,L., J.Dong and G.E.Hart (2000) Characteristics, linkage-map positions, and allelic differentiation of *Sorghum bicolor* (L.) Moench DNA simple-sequence repeats (SSRs). Theor. Appl. Genet. 101: 438–448.

Lander,E.S. and N.Schork (1994) Genetic dissection of complex traits. Science 265: 2037–2048.

Lewontin,R.C. (1964) The Interaction of Selection and Linkage. I. General considerations; heterotic models. Genetics 49: 49–67.

Murray,M. and W.F.Thompson (1980) Rapid isolation of high molecular weight plant DNA. Nucleic Acids Res. 8: 4321–4325.

Paterson,A.H. (2008) Genomics of sorghum. Int. J. Pl. Gen. doi: 10.1155/2008/362451.

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

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

Remington,D.L., J.M.Thornsberry, Y.Mastsuoka, L.M.Wilson, S.R. Whitt, J.Doebley, S.Kresovich, M.M.Goodman and E.S.Buckler (2001) Structure of linkage disequilibrium and phenotypic associations in maize genome. Proc. Natl. Acad. Sci. 98: 11479–11484.

Shehzad,T., H.Okuizumi, M.Kawase and K.Okuno (2009) Development of SSR based sorghum (*Sorghum bicolor* (L.) Moench) diversity research set and its evaluation by morphological traits. Genet. Res. Crop Evol. doi: 10.1007/s10722-008-9403-1.

Stich,B., E.Albrecht, M.M.Frisch, H.P.Maurer, M.Heckenberger and J.C.Reif (2005) Linkage disequilibrium in European elite maize germplasm investigated with SSRs. Theor. Appl. Genet. 111: 723–730.

Stich,B., H.P.Maurer, A.E.Melchinger, M.Frisch, M.Heckenberger, J.R.van der Voort, J.Peleman, A.P.Sørensen and J.C.Reif (2006) Comparison of linkage disequilibrium in elite European maize inbred lines using AFLP and SSR markers. Mol. Breed. 17: 217–226.

Stich,B., J.M.Möhring, H.P.Piepho, M.Heckenberger, E.S.Buckler and A.E.Melchinger (2008) Comparison of mixed-model approaches for association mapping. Genetics 178: 1745–1754.

Taramino,G., R.Tarchini, S.Ferrario, M.Lee and M.E.Pe (1997) Characterization and mapping of simple sequence repeats (SSRs) in *Sorghum bicolor*. Theor. Appl. Genet. 95: 66–72.

Tenaillon,M.I., M.C.Sawkins, A.D.Long, R.L.Gaut, J.F.Doeble and B.S.Gaut (2001) Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.) Proc. Natl. Acad. Sci. 98: 9161–9166.

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

Tsunewaki,K. (1983) Gene transfer between three groups of wheat. II.

Differential transmission rates of deleterious genes in 6× and 4× wheats. Jpn. J. Genet. 58: 219–229.

Weir, B.S. (1996) Disequilibrium. *In*: Genetic data analysis II: methods for discrete population genetic data. Sinaur Associates, Sunderland, MA, pp. 91–139.

Whitt, S.R. and E.S. Buckler (2003) Using natural allelic diversity to evaluate gene function. Methods Mol. Biol. 236: 123–139.

Yu, J. and E.S. Buckler (2006) Genetic association mapping and genome organization of maize. Curr. Opin. Biotechnol. 17: 155–160.

Yu, J., G. Pressoir, W.H. Briggs, I. Vroh Bi, M. Yamasaki, J.F. Doebley, M.D. McMullen, B.S. Gaut, D.M. Nielsen, J.B. Holland, S. Kresovich and E.S. Buckler (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. 38: 203–208.

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