Theor Appl Genet (2012) 124:447–458
DOI 10.1007/s00122-011-1719-0

ORIGINAL PAPER

Identification of single nucleotide polymorphisms and haplotypes associated with yield and yield components in soybean (*Glycine max*) landraces across multiple environments

**Derong Hao · Hao Cheng · Zhitong Yin ·
Shiyou Cui · Dan Zhang · Hui Wang · Deyue Yu**

Received: 1 April 2011 / Accepted: 27 September 2011 / Published online: 14 October 2011

Abstract Genome-wide association analysis is a powerful approach to identify the causal genetic polymorphisms underlying complex traits. In this study, we evaluated a population of 191 soybean landraces in five environments to detect molecular markers associated with soybean yield and its components using 1,536 single-nucleotide polymorphisms (SNPs) and 209 haplotypes. The analysis revealed that abundant phenotypic and genetic diversity existed in the studied population. This soybean population could be divided into two subpopulations and no or weak relatedness was detected between pair-wise landraces. The level of intra-chromosomal linkage disequilibrium was about 500 kb. Genome-wide association analysis based on the unified mixed model identified 19 SNPs and 5 haplotypes associated with soybean yield and yield components in three or more environments. Nine markers were found co-associated with two or more traits. Many markers were located in or close to previously reported quantitative trait loci mapped by linkage analysis. The SNPs and haplotypes identified in this study will help to further understand the genetic basis of soybean yield and its components, and may facilitate future high-yield breeding by marker-assisted selection in soybean.

Introduction

Seed yield of soybean (*Glycine max* [L.] Merr.) is controlled in a complex manner by quantitative trait loci (QTLs), and environmental variations can trigger and modify the actions of related genes (Li et al. 2005). In soybean, it is becoming more difficult to improve yield using traditional methods. However, the development of genomics has provided alternative tools to improve breeding efficiency in plant breeding programs. Molecular markers linked to the causal genes and/or QTLs can be used for marker-assisted selection (MAS) (Xu and Crouch 2008).

Over the past 20 years, a large number of QTLs associated with soybean seed yield and other important agronomic traits have been reported (Chung et al. 2003; Guzman et al. 2007; Hoeck et al. 2003; Kassem et al. 2006; Keim et al. 1990; Li et al. 2008a; Mansur et al. 1993, 1996; Mar 1996; Maughan et al. 1996; Mian et al. 1996; Orf et al. 1999a; Palomeque et al. 2010; Smalley et al. 2004; Vieira et al. 2006; Wang et al. 2004; Zhang et al. 2004). However, approximately 85% of the previously reported QTLs could not be confirmed in subsequent studies, and few have actually been applied in breeding programs (Kassem et al. 2006). This is because most QTLs were population-specific, and the genetic variation detected in the specific bi-parental population might not be shared in other genetic

Communicated by J. Yan.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-011-1719-0) contains supplementary material, which is available to authorized users.

D. Hao · H. Cheng · D. Zhang · H. Wang · D. Yu (✉)
National Center for Soybean Improvement,
National Key Laboratory of Crop Genetics and Germplasm
Enhancement, Nanjing Agricultural University,
210095 Nanjing, China
e-mail: dyyu@njau.edu.cn

D. Hao · S. Cui
Jiangsu Yanjiang Institute of Agricultural Sciences,
226541 Nantong, China

Z. Yin
Jiangsu Provincial Key Laboratory of Crop Genetics
and Physiology, Yangzhou University,
225009 Yangzhou, China

populations (Wang et al. 2008; Xu and Crouch 2008). In addition, the limited recombination in most populations used for linkage mapping makes it difficult to precisely map QTLs, which severely limited their use in MAS (Cardon and Bell 2001; Gupta et al. 2005).

With the potential to exploit all recombination events that occurred in the evolutionary history of a specific germplasm, genome-wide association analysis based on linkage disequilibrium (LD) has become a powerful approach for dissection of complex agronomic traits and identification of causal variation with modest effects for target traits in crops (Atwell et al. 2010; Yan et al. 2009, 2010; Yu and Buckler 2006; Chan et al. 2011). The key constraint for the successful use of association analysis in plants is the population structure and genetic relatedness, which can result in spurious marker-trait associations that may make it difficult to distinguish loci that truly affect the target traits (Ersoz et al. 2007; Gupta et al. 2005; (Chan et al. 2011). Several statistical strategies have been developed to account for the population structure and relatedness. One powerful strategy is the unified mixed model approach (MLM), which accounts for multiple levels of relatedness simultaneously, and can improve control of both typeIand typeII error rates (Yu et al. 2006).

Recent advances in genome sequencing and single nucleotide polymorphism (SNP) genotyping have increased the applicability of association analysis for QTL mapping in crops (Morgante and Salamini 2003; Rafalski 2010). Genome-wide association analyses with SNP markers have been conducted for several important traits in many plant species, including *Arabidopsis thaliana* (Atwell et al. 2010; Sulpice et al. 2009), maize (Beló et al. 2008; Lai et al. 2010; Lu et al. 2010; Yang et al. 2010), and rice (Huang et al. 2010). In soybean, several genetic bottleneck events and two major duplication events of genome have resulted in lower sequence diversity in cultivated soybean, that limited the amount of SNPs (Blanc and Wolfe 2004; Hyten et al. 2006; Schlueter et al. 2004; Schmutz et al. 2010). SNP genotyping via the GoldenGate assay has resulted in the identification of 8,000 SNPs in soybean, and the construction of the fourth version of the soybean integrated linkage map (Choi et al. 2007; Hyten et al. 2010). These SNP resources will facilitate high-throughput genotyping in genome-wide association analysis for dissection of complex agronomic traits (Hyten et al. 2010). Identification of the SNPs associated with yield-related traits would facilitate combining the desirable genes in soybean breeding programs.

Furthermore, haplotype association is likely to be more powerful in the presence of LD (Garner and Slatkin 2003). Using haplotypes for QTL mapping could compensate for the bi-allelic limitation of SNPs, and substantially improve the efficiency of QTL mapping (Lu et al. 2010; Yan et al. 2011). And, haplotype-traits association analyses are helpful for precise mapping of important genomic regions and location of favor alleles or haplotypes for breeding (Barrero et al. 2011).

The aim of this study was to identify SNPs and haplotypes underlying soybean yield and yield components in different environments. In addition, the population structure and genetic relatedness were analyzed by SNPs. Our results suggest that genome-wide association analysis in soybean landraces using SNPs and haplotypes is an alternative mapping approach for identifying QTLs underlying soybean yield and yield components.

Materials and methods

Plant materials and phenotypic data collection

A population of 191 soybean landraces from different geographic origins and with phenotypic variations was selected to construct the association mapping panel. The trials were performed in 2009 and 2010 at three different locations along the Yangtze River Basin, as follows: Jiangpu Experimental Station of Nanjing Agricultural University (32°12′N 118°37′48″E), Nanjing, in 2009 (designated as environment E1) and 2010 (designated as environment E2); Experimental Farm of Jiangsu Yanjiang Institute of Agricultural Sciences (31°58′48″N 120°53′24″E), Nantong, in 2009 (designated as environment E3) and 2010 (designated as environment E4); and Experimental Farm of Agricultural College of Yangzhou University (32°23′24″N 119°25′12″E), Yangzhou, in 2010 (designated as environment E5). A randomized complete-block design was used for all trials. For environment E1, all landraces were planted with two replications. For other environments (E2, E3, E4, and E5), all landraces were planted with three replications. In all five environments, each landrace was planted in three rows per plot, each row 200 cm-long and with 50 cm row spacing. Four traits were evaluated in all environments, including number of pods per plant (PN), number of seeds per plant (SN), 100-seed weight (SW) (g), and seed yield (SY) (g/m^2). The middle row in each plot was harvested to measure these four traits after maturity. SY and SW were adjusted to 13% moisture content in all environments.

SNP genotyping

Genomic DNA samples were extracted from the leaf of soybean seedlings using the CTAB method (Murray and Thompson 1980). All these 191 soybean landraces were genotyped with 1,536 SNP chips via the GoldenGate assay (Illumina, San Diego, CA, USA) (Shen et al. 2005). To design the GoldenGate assay, a total of 2,435 random SNPs evenly covering the genome were selected from the Soybean SNP database (http://bfgl.anri.barc.usda.gov/soybean/) (Choi et al. 2007). SNP genotyping was performed on the Illumina Beadlab system at the National Engineering Center for Biochip (Shanghai, China). Based on their quality scores, the best 1,536 SNPs were chosen for the GoldenGate assay and 1,142 SNPs with minor allele frequency (MAF) of ≥10% in the present population were used for subsequent analyses.

Population genotypic data analysis

Genetic diversity characteristics in the 191 soybean landraces, including MAF, gene diversity, heterozygosity and polymorphism information content (PIC), were evaluated using the software Powermarker 3.25 (Liu and Muse 2005). This software was also used to construct a Neighbor-joining tree based on Nei's genetic distance matrix (Nei et al. 1983; Liu and Muse 2005). Haplotypes were constructed with an accelerated EM algorithm using the software Haploview 4.2 (Barrett et al. 2005), and the haplotype blocks were defined with the default algorithm of 95% confidence intervals as described (Gabriel et al. 2002), all haploytpes (including the rare haplotypes) were used for further analyses in this study. The linkage disequilibrium parameter (r^2) for estimating the degree of LD between pair-wise SNPs (1,142 of 1,536 SNPs with MAF ≥10%) was calculated using the software TASSEL 2.1 (Bradbury et al. 2007) with 1,000 permutations. The decay distance of LD was determined with a threshold of $r^2 = 0.1$ as described (Hyten et al. 2007; Malysheva-Otto et al. 2006).

The Bayesian model-based program STRUCTURE 2.2 (Pritchard et al. 2000) was used to infer the population structure using the 1,142 SNPs mentioned above. The length of burn-in period and the number of Markov Chain Monte Carlo replications after burn-in were all assigned at 100,000 with an admixture and allele frequencies correlated model. Five independent iterations of running were performed with the hypothetic number of subpopulation (k) ranging from 1 to 10. The correct estimation of k was provided by joining the log probability of data [LnP(D)] from the STRUCTURE output and an ad hoc statistic Δk (Evanno et al. 2005), which was based on the rate of change in the log probability of data between successive k values. Based on the correct k, each soybean landrace was assigned into a subpopulation for which the membership value (Q value) was >0.5 (Pritchard et al. 2000; Breseghello and Sorrells 2006), and the population structure matrix (Q) was generated for further analyses.

Analysis of molecular variance (AMOVA) and population pair-wise F statistics (F_{ST}) for the inferred subpopulations were performed to investigate the population difference using Arlequin 3.01 (Excoffier et al. 2005). The software SPAGeDi (Hardy and Vekemans 2002) was used to calculate the pair-wise relatedness coefficients (K, kinship matrix) to estimate the genetic relatedness among individuals with the negative value of kinship set as zero.

Phenotypic data analysis

Statistical analysis of all phenotypic data across five environments was conducted using the software SAS 8.0 (SAS Institute 1999). Analysis of variance (ANOVA) of all phenotypic data based on means of traits of each landrace across five environments was conducted using PROC GLM. Decomposition of variance components (genotype, year, location, block, and the interactions among these factors) was evaluated using PROC VARCOMP. The broad-sense heritability (h^2) (Holland et al. 2003) of each trait was estimated using the variance components. The effects of population structure on the phenotypic variation were estimated based on the mean values for each trait using PROC GLM, the model statement included one of the two components of the Q matrix ($k = 2$) (Yang et al. 2010). Correlation coefficients between soybean yield and yield components were calculated with PROC CORR.

Genome-wide association analysis

To account for the population structure and genetic relatedness, various statistical models were evaluated: the GLM model without considering Q and K; the Q model with considering Q; the K model with considering K; the MLM model with considering Q and K. Genome-wide association analyses based on these models were conducted with the software TASSEL 2.1 (Bradbury et al. 2007; Yu et al. 2006). Markers were defined as being significantly associated with traits on the basis of their significant association threshold ($-\text{Log}P \geq 2.00$, $P \leq 0.01$).

Results

Genetic diversity

Among the total 1,536 SNPs in 191 soybean landraces, 394 (25.7%, data not shown) showed minor alleles frequency (MAF) of less than 10%, and were, therefore, excluded from further analyses. The remaining 1,142 SNPs with MAF greater than 10% were used to determine genetic diversity and for further analyses. The average MAF value of the 1,142 SNPs was 29.1 (range 10.2–50.0). The gene diversity, heterozygosity and PIC of the 1,142 SNPs averaged 0.391, 0.013 and 0.313, with ranges of 0.071–0.615, 0–0.245 and 0.069–0.537, respectively (Fig. 1). The low rate of heterozygous loci was observed in the present popu-

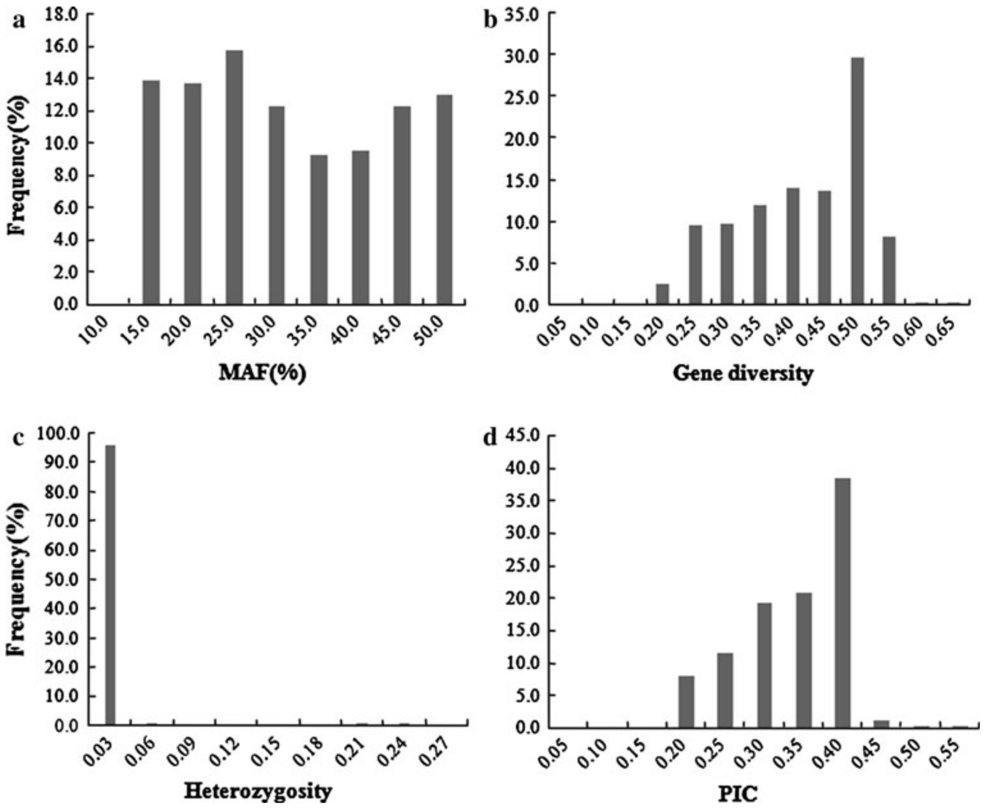

Fig.1 Distribution of genetic diversity of 1,142 SNPs across 191 landraces. **a** MAF, **b** gene diversity, **c** heterozygosity, **d** PIC

Table 1 Summary of SNPs and alleles in each haplotype

Number of SNPs per haplotype	Number of haplotypes	Number of haplotypes with different alleles per haplotype						
		2	3	4	5	6	7	8
2	163	57	101	5	0	0	0	0
3	35	4	17	12	1	0	0	1
4	7	1	2	1	3	0	0	0
5	3	0	2	1	0	0	0	0
6	1	0	0	0	1	0	0	0
Total	209	62	122	19	5	0	0	1

lation, with the average heterozygosity rate of each landrace of 0.012. 209 haplotypes were identified from 1,142 SNPs among 191 landraces. These haplotypes consisted of 599 alleles, with 2–8 alleles per haplotype (Table 1). There were 2–6 SNPs per haplotype. Haplotypes consisting of two SNPs were most common (78% of all haplotypes).

Population structure and genetic relatedness

The evaluation of the population structure of the 191 soybean landraces indicated that the distribution of the LnP(D) value corresponding to each hypothetical k did not show any peaks, but there was an increase in LnP(D) value with increasing k value (data not shown). The ad hoc quantity (Δk) showed a much higher likelihood at $k = 2$ than at $k = 3$–10 (Fig. 2a), suggesting that the population could be clustered into two major subpopulations (Fig. 2b). The population pair-wise F_{ST} was 0.24 ($P < 0.001$) between the two subpopulations, which revealed high level of difference. The result of AMOVA indicated that 23.8% of the total genetic variation was among subpopulations, whereas 76.2% was within subpopulations (Table 2).

The information of neighbor-joining tree (Fig. 2c) was consistent with the results from STRUCTURE. For the subpopulation P1 (including 161 landraces), most landraces were from South China with relatively late maturity. The landraces in subpopulation P2 (including 30 landraces) were mainly from North China with relatively early maturity. The corresponding Q-matrix at $k = 2$ was used for the following genome-wide association analysis.

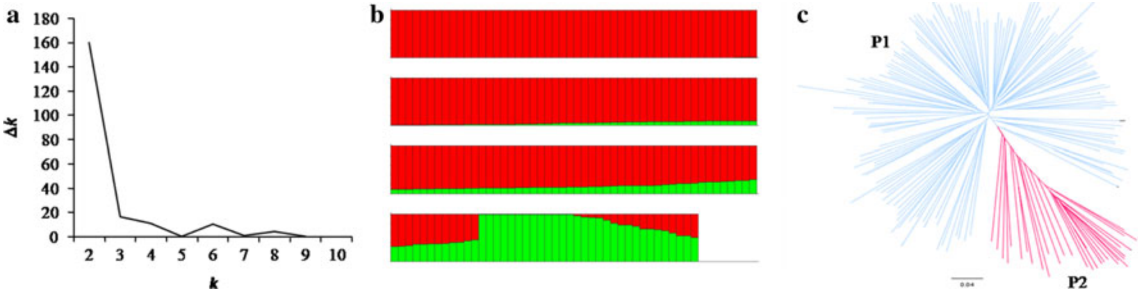

Fig. 2 Divergence of 191 soybean landraces. **a** Δk value over five iterations of running with putative k ranging from 1 to 10, **b** model-based population structure for all 191 soybean landraces, **c** neighbor-joining tree based on Nei's genetic distance matrix

Table 2 Analysis of molecular variance (AMOVA) and F_{ST} for two subpopulations of soybean landraces inferred from STRUCTURE

Source of variation	*df*	Sum of squares	Variance components	Percentage variation	*P* value
Among subpopulaions	1	6,739.00	64.58	23.8	<0.001
Within subpopulations	380	78,579.54	206.79	76.2	<0.001
Total	381	85,318.54	271.37		

Population pair-wise F_{ST} : 0.24 ($P < 0.001$)

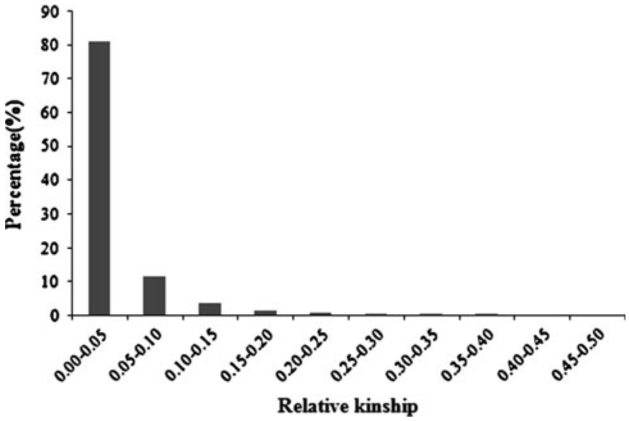

Fig. 3 Distribution of pair-wise kinship coefficients among 191 soybean landraces. Kinship coefficients between landraces were calculated using 1,142 SNPs

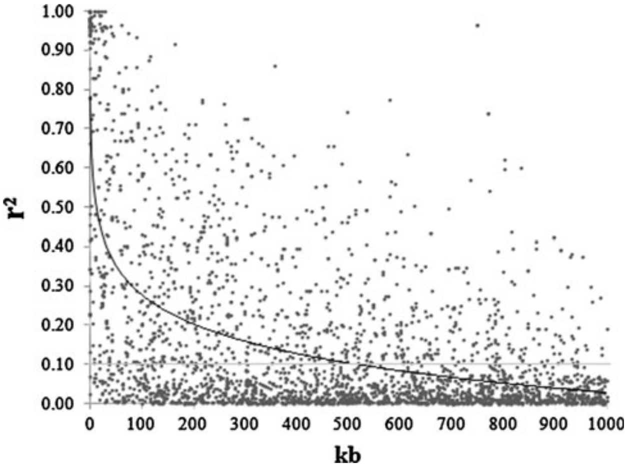

Fig. 4 Scatter plot of r^2 against genetic distance within a distance of 1,000 kb in the population of 191 soybean landraces

Genetic relatedness analysis indicated that landraces in the present population were distantly related (Fig. 3). For the kinship coefficient values, more than 80% were less than 0.05, 11.6% had a range of 0.05–0.10, and the remaining 7.4% showed various degrees of genetic relatedness. This result suggested that there was no or weak relatedness between pair-wise soybean landraces. Based on the result of the relatedness analysis, a K matrix was constructed for association analysis.

Linkage disequilibrium across whole genome

The r^2 values between pair-wise markers physically linked within a distance of 1,000 kb were plotted against distance (Fig. 4). The scatter plots of the r^2 versus.distance showed that there was a high LD in this population. The average decay of intra-chromosomal LD declined to the threshold of $r^2 = 0.1$ at around 500 kb, and the r^2 value on average was 0.16 across whole genome.

Phenotypic variation analysis

Each trait varied widely among different environments (Table 3). For all traits, the maximum value was approximately ten times the minimum value. The values of PN, SN, SW, and SY across all five environments showed ranges of 26.48–136.85, 39.17–272.44, 4.65–38.65, and

Table 3 Descriptive statistics, ANOVA, and broad-sense heritability (*h2*) of soybean yield and its components across five different environments

Trait	Environment	Mean	SD	Min.	Max.	G[a]	G × E[b]	h^{2c} (%)	R^{2d}
PN	E1	38.84	16.40	10.33	105.33	**	**	66.0	1.71
	E2	55.42	18.37	22.89	115.00				
	E3	81.20	26.34	18.00	216.33				
	E4	91.63	32.23	30.80	201.57				
	E5	88.84	28.98	35.25	174.64				
	Mean	71.83	19.71	26.48	136.85				
SN	E1	68.32	29.97	16.67	194.00	**	**	60.0	2.61
	E2	94.65	34.03	29.00	201.89				
	E3	142.25	51.53	28.33	450.17				
	E4	183.26	64.46	61.60	403.13				
	E5	136.92	52.47	38.00	350.00				
	Mean	126.14	37.20	39.17	272.44				
SW	E1	14.23	6.30	5.46	41.10	**	**	93.0	1.49
	E2	13.24	5.89	4.29	39.88				
	E3	13.08	5.70	3.93	41.31				
	E4	14.63	6.03	4.47	36.27				
	E5	13.56	5.44	4.56	34.68				
	Mean	13.81	5.55	4.65	38.65				
SY	E1	127.41	45.33	33.18	268.70	**	**	67.0	8.95
	E2	170.25	64.52	47.13	385.52				
	E3	246.83	104.81	39.95	690.65				
	E4	291.49	100.55	93.10	704.90				
	E5	261.56	119.20	43.74	667.01				
	Mean	222.52	72.27	69.21	421.28				

PN number of pods per plant, *SN* number of seeds per plant, *SW* 100-seed weight, *SY* seed yield

**Significant at $P < 0.01$

[a] Genotype across different environments

[b] Genotype × environment

[c] Broad-sense heritability

[d] Percentage of phenotypic variation explained by population structure

69.21–421.28, with average values of 71.83, 126.14, 13.81, and 222.52, respectively. The h^2 of the four traits had a range of 60.0–93.0% in this population (Table 3). The highest h^2 value was for SW (93.0%), indicating that SW was less affected by environmental factors than the other three traits.

The ANOVA showed that the genotype (G) and the interactions between genotype and all environmental factors (G × E, including genotype × year, genotype × location, and genotype × year × location) were all significant ($P < 0.01$) for yield and yield components (Table 3). The effects of population structure on soybean yield and yield components were various in present population, with *R2* average of 3.69%, ranging from 1.49% (SW) to 8.95% (SY). (Table 3).

Phenotypic correlation analysis showed that there were significant positive correlations between soybean yield and yield components (Table 4). The phenotypic correlation coefficient of SY with PN, SN, and SW was 0.27, 0.27, and 0.62, respectively. There was also a significant positive correlation between PN and SN ($r = 0.97$). Among the three yield components, there was a significant negative correlation between SW and PN ($r = -0.45$), and between SW and SN ($r = 0.49$).

Genome wide association analysis

As shown in quantile–quantile plots (Supplementary Fig. 1–Supplementary Fig. 4), the MLM model (Q + K) and the K model were significantly better than the GLM and the Q model for soybean yield and yield components in reducing the effect of population structure and the genetic relatedness. Of which, the MLM model performed a little better than the K model. So, we conducted the genome-wide

Table 4 Phenotypic correlations among yield and its components based on means of traits in all soybean landraces across five environments

Traits	PN	SN	SW
SN	0.97***		
SW	−0.45***	−0.49***	
SY	0.27***	0.27***	0.62***

PN number of pods per plant, *SN*: number of seeds per plant, *SW* 100-seed weight, *SY* seed yield

***Significant at $P < 0.001$

association analysis for soybean yield and yield components with the MLM model (Q + K) to correct for population structure and genetic relatedness, using 1,142 SNPs and 209 haplotypes.

We identified 50 SNPs associated with PN, 72 SNPs associated with SN, 40 SNPs associated with SW, and 46 SNPs associated with SY (Supplementary Fig. 5). Most SNPs were detected only in a specific environment, and only a small number of SNPs were identified in three or more environments (Tables 5, 6). For PN, eight SNPs were identified in three or more environments; seven in three environments, one in four environments, and none in all five environments. For SN, eight SNPs were identified in three or more environments; seven in three environments, one (BARC-042035-08159) in four environments, and none in all five environments. For SW, nine SNPs were identified in three or more environments; three in three environments, two in four environments, and four (BARC-040407-07733, BARC-040075-07652, BARC-029185-06106 and BARC-028709-05992) in all five environments. For SY, only one SNP (BARC-028709-05992) was identified in three environments, and none were detected in more environments.

Haplotypes (18, 22, 9, and 8) were identified for PN, SN, SW and SY, respectively (Supplementary Fig. 6). Similar to the result with SNPs, only some of the haplotypes were detected in three or more environments (Tables 5, 7). For PN, only one haplotype (hp40) was identified in three environments, and none in more environments. For SN, two haplotypes (hp27 and hp40) were identified in three environments, and none in more environments. For SW, one haplotype (hp138) was detected in three environments, one haplotype (hp115) in four environments, and one (hp40) in all five environments. For SY, two haplotypes (hp27 and hp183) were detected in three environments.

Nine loci were co-associated with two or more different traits (Tables 6, 7). BARC-043191-08550 and BARC-030807-06945 on chromosome 1, BARC-015003-01948 on chromosome 10 and BARC-042035-08159 on chromosome 13 were associated with PN and SN. BARC-028709-05992 on chromosome 3 was associated with SW and SY. The haplotype of hp40 on chromosome 11, consisting of the two closely linked markers BARC-040407-07733 and BARC-040075-07652, was associated with, SW, SN, and PN.

Discussion

Population genetic diversity and linkage disequilibrium in soybean

SNPs are widely distributed throughout the whole genome, and thus, can be used to obtain accurate sequence information via high-throughput sequencing technology, and to compensate for its bi-allelic shortcoming (Lu et al. 2009; Yan et al. 2010; Wen et al. 2011), which have been widely used for population genetic research in many plant species, including soybean (Hyten et al. 2007). Yu et al. (2009) and Van Inghelandt et al. (2010) suggested that about ten times more SNPs than SSRs used for analysis of population structure and genetic diversity could achieve similar accuracy to

Table 5 Summary of SNPs and haplotypes significantly associated with PN, SN, SW, and SY in three or more environments

Traits	Marker	Number of markers detected in different environments			
		Total	3E	4E	5E
PN	SNP	8	7	1	0
	Haplotype	1	1	0	0
SN	SNP	8	7	1	0
	Haplotype	2	2	0	0
SW	SNP	9	3	2	4
	Haplotype	3	1	1	1
SY	SNP	1	1	0	0
	Haplotype	2	2	0	0

3E, *4E*, and *5E* Number of individual SNP and haplotype detected across three, four, or five different environments, respectively, *PN* number of pods per plant, *SN*: number of seeds per plant, *SW* 100-seed weight, *SY* seed yield

Table 6 SNPs with significant association signals ($-\mathrm{Log}P \geq 2.00$, $P \leq 0.01$) for soybean yield and yield components detected in three or more environments

Traits	Marker	Chr.	Position	SNP	$-\mathrm{Log}P^{a}$				
					E1	E2	E3	E4	E5
PN	BARC-043191-08550	1	43926412	G/A	3.43	ns[b]	2.89	2.85	ns
	BARC-030807-06945	1	53063797	A/T	2.62	2.03	2.22	2.62	ns
	BARC-015003-01948	10	44376675	G/A	Ns	3.07	ns	2.15	4.67
	BARC-024093-04723	11	38631595	G/A	Ns	2.7	ns	2	2.33
	BARC-024093-04724	11	38631705	G/A	Ns	3.28	ns	2	3.02
	BARC-042035-08159	13	43467832	C/A	3.12	2.29	3	ns	ns
	BARC-041815-08101	14	45478975	G/A	2	3.52	2.34	ns	ns
	BARC-021603-04153	18	61162404	A/G	2.57	2.28	ns	3.19	ns
SN	BARC-043191-08550	1	43926412	G/A	2.15	ns	2.19	2.74	ns
	BARC-030807-06945	1	53063797	A/T	2.28	ns	2.42	2.27	ns
	BARC-039653-07533	2	51233483	T/A	2.8	2.1	2.15	ns	ns
	BARC-019805-04378	2	51243486	G/C	2.8	2.09	2.14	ns	ns
	BARC-015003-01948	10	44376675	G/A	Ns	3.91	ns	2.28	3.58
	BARC-040407-07733	11	30159557	C/A	2.85	2.44	ns	2.37	ns
	BARC-040075-07652	11	30159839	C/A	2.85	2.44	ns	2.37	ns
	BARC-042035-08159	13	43467832	C/A	2.74	2.6	3.82	ns	2.82
SW	BARC-029185-06106	3	40131490	A/C	4.76	4.24	2.74	4.8	5
	BARC-016485-02069	3	40585266	G/A	Ns	ns	2.23	2.01	2.1
	BARC-028709-05992	3	40654334	G/A	3.39	2.28	2.28	3.84	2.96
	BARC-018515-02927	9	44398393	A/C	Ns	3.43	ns	3.34	2.27
	BARC-040407-07733	11	30159557	C/A	3.77	3.34	3.15	3.95	3
	BARC-040075-07652	11	30159839	C/A	3.77	3.34	3.15	3.95	3
	BARC-030931-06978	13	1699386	G/A	2.17	2.15	ns	2	2.49
	BARC-018985-03048	13	1766981	A/G	2.72	3.05	ns	3.28	3.86
	BARC-021827-04218	19	48091800	G/A	2.12	ns	ns	2	2.1
SY	BARC-028709-05992	3	40654334	G/A	3.93	2.47	ns	3.07	ns

Chr. chromosome number of soybean, *PN* number of pods per plant, *SN* number of seeds per plant, *SW* 100-seed weight, *SY* seed yield

[a] Significant at $-\log(P) \geq 2.00$ ($P \leq 0.01$)

[b] Marker was not detected at significant level in corresponding environment

SSRs. In this study, we used 1,142 SNPs to estimate the genetic diversity and population structure in 191 soybean landraces. This landrace population showed relatively high genetic diversity with mean genetic diversity coefficient of 0.39, that was a little higher than that of 0.35 over 554 SNPs for 303 cultivated soybeans (*G. max*) and wild soybeans (*G. soja*) by Li et al. (2010). Such rich genetic diversity in this population might result from the diverse range of landraces, and the large number of SNPs used for evaluation. Li et al. (2010) also suggested that SNPs had lower genetic diversity than SSRs, but analysis of population structure based on the same SNP dataset gave similar results, especially for cultivated soybeans. In present population, few lines showed high heterozygosity for some SNP loci (Fig. 1), which may be due to the natural outcrossing during the propagation of few germplasms.

The population genetic structure of soybean landraces has been documented previously (Li et al. 2008b, 2010). In the present study, the 191 soybean landraces were classified into two subpopulations with significant divergence ($F_{ST} = 0.24$). To account for population structure in association analysis, the optimal model of MLM (Yu et al. 2006) was applied in our study, which greatly reduced false positives, as shown in quantile–quantile plots (Supplementary Fig. 1–Supplementary Fig. 4). The MLM has been successfully applied to account for population structure in several crops (Yu et al. 2006; Breseghello and Sorrells 2006; Zhao et al. 2007).

Linkage disequilibrium is the basis of association analysis for detecting genetic polymorphisms associated with important quantitative traits (Flint-Garcia et al. 2003). In this study, the decay distance of LD was about 500 kb in

Table 7 Haplotypes with significant association signals ($-\text{Log}P \geq 2.00$, $P \leq 0.01$) for soybean yield and yield components detected in three or more environments

Traits	Haplotype	SNP	Chr.	Position	−LogP[a]				
					E1	E2	E3	E4	E5
PN	hp40	BARC-040407-07733	11	30159557	2.96	2.32	ns[b]	2.8	ns
		BARC-040075-07652	11	30159839					
SN	hp27	BARC-014837-01682	8	34981007	2.8	ns	2	ns	2.28
		BARC-014847-01910	8	34980901					
	hp40	BARC-040407-07733	11	30159557	3.39	3.09	ns	2.66	ns
		BARC-040075-07652	11	30159839					
SW	hp40	BARC-040407-07733	11	30159557	4.13	3.84	3.65	4.64	3.62
		BARC-040075-07652	11	30159839					
	hp138	BARC-020149-04485	12	413234	2.28	2.07	ns	2	ns
		BARC-013545-01156	12	483243					
		BARC-044181-08640	12	554095					
	hp115	BARC-030931-06978	13	1699385	2.12	2.34	ns	2.47	2.96
		BARC-018985-03048	13	1766980					
SY	hp27	BARC-014837-01682	8	34981007	2.92	ns	2.85	ns	2.11
		BARC-014847-01910	8	34980901					
	hp183	BARC-029825-06443	7	2472109	ns	3.69	2.7	ns	2.05
		BARC-029831-06446	7	2472109					
		BARC-029831-06445	7	2471948					
		BARC-029825-06442	7	2471948					
		BARC-025961-05189	7	2477192					

Chr. Chromosome number of soybean, *PN* number of pods per plant, *SN* number of seeds per plant, *SW* 100-seed weight, *SY* seed yield

[a] Significant at $-\log(P) \geq 2.00$ ($P \leq 0.01$)

[b] Marker was not detected at significant level in corresponding environment

191 soybean landraces, which is consistent with those obtained by Hyten et al. (2007), in which LD extended from 90 to 574 kb in soybean landraces population. This high level of LD in soybean was also reported in other studies (Jun et al. 2008; Wang et al. 2008; Zhu et al. 2003), which far exceeds that of other crops, such as rice (about 150 kb) (Mather et al. 2007) and maize (about 10 kb) (Yan et al. 2009). The high level of LD in the soybean genome suggests that the mapping resolution gained from LD is likely to be limited, and that marker-assisted breeding will be less challenging than map-based cloning (Lam et al. 2010). So the high level of LD in present population would promise that the identified SNPs and haplotypes would facilitate soybean high-yield breeding by marker-assisted selection.

Several key SNPs and haplotypes are associated with yield and yield components in soybean

Using the optimal model of MLM, we identified 19 SNPs and 5 haplotypes associated with soybean yield and yield components in three or more environments. Some of these SNPs and haplotypes were located in or near regions where QTLs for yield and yield-related traits have been mapped by linkage analysis (Csanádi et al. 2001; Funatsuki et al. 2005; Guzman et al. 2007; Hoeck et al. 2003; Hyten et al. 2004; Kabelka et al. 2004; Lee et al. 2001; Mian et al. 1996; Orf et al. 1999b; Specht et al. 2001; Wang et al. 2004; Yuan et al. 2002). BARC-040407-07733 and BARC-040075-07652 on chromosome 11 (stably associated with SW) were located in the same region of a QTL for SW (Specht et al. 2001). BARC-021827-04218 on chromosome 19 (associated with SW) and hp183 on chromosome 7 (associated with SY) were respectively located near Dt1 and Satt150, where the main QTLs for seed weight and yield have been reported in several studies (Cui et al. 2008; Csanádi et al. 2001; Hoeck et al. 2003; Mian et al. 1996; Palomeque et al. 2009). These results indicated that some causal gene/genes might exist in these genome regions, these associated markers may be useful for aggregation of causal genes of interest to improve soybean yield. In this study, several SNPs and haplotypes significantly associated with soybean yield and its components have never been reported, such as BARC-029185-06106, BARC-

028709-05992, and BARC-016485-02069 on chromosome 3 and BARC-030931-06978, BARC-018985-03408, and hp115 on chromosome 13. These new loci are attractive candidate regions for further understanding the genetic basis of soybean yield and yield components.

Nine markers were co-associated with two or more traits in present study, which coincided with significant phenotypic correlations among the studied traits, as reported elsewhere (Hyten et al. 2004; Kabelka et al. 2004; Palomeque et al. 2010). The genome regions where multi-traits were co-associated indicated pleiotropy of single causal gene or tight linkage of multiple causal genes. In soybean MAS schemes, MAS of a co-associated genetic locus could simultaneously improve multi-associated target traits.

Only four SNPs and one haplotype were stably detected for SW with highest broad-sense heritability ($h^2 = 93.0\%$) in all five environments, which was due to that agronomic traits are the result of the combined actions of multiple genes and environmental factors, and gene expression varies in different environments (Mansur et al. 1993). The inheritance of quantitative traits classically involves multiple genes with small effect that are sensitive to environmental changes (Xing and Zhang 2010). Only the traits with high heritability could be mapped stably. The resulting stably associated markers in this study should be well useful for breeding with broad adaptability to different environments. Using those markers detected in the specific environment, breeders could identify the best landraces that are specifically adapted to local environments.

In this study, we identified five haplotypes associated with yield and yield components. Among which, one haplotype of hp40, containing two linked SNPs, may be from a possible causal gene (Glyma11g29460) encoding Cinnamoyl CoA reductase (CCR) (Supplementary Fig. 7). The CCR catalyses the first step of the lignin specific biosynthetic pathway (Lauvergeat et al. 2001). Lignin is mainly deposited in the walls of certain specialized cells (e.g. tracheary elements), relevant to the transport of water and nutrients within xylem tissue by modifying the permeability of the cell wall (Ma 2007). That are important in the plant developments, including seed development. However, other haplotypes were not involving some possible genes. This may be due to the higher LD (~500 kb) in the studied population, which meant that there was a low probability of these associated-markers from the causal genes themselves. Instead, the causal genes might exist within the genomic regions in LD where associated-markers located. Choice of more diverse germplasm with lower LD in and around the gene of interest, and use of more markers (especially of gene-based markers, including markers from key genes of metabolic networks), will directly detect the causal genes or get closer to the gene of interest (Chan et al. 2011; Yan et al. 2011). In addition haplotyping of critical genome regions will be a preferred method of gene discovery in association mapping (Barrero et al. 2011). Therefore, further studies will be conducted using a larger population size with more diverse genetic background, to validate the associated-markers identified in this study, and to mine some rarely functional alleles. In addition, a larger number of SNPs derived from putative yield-related candidate genes will be chosen to improve the scanning power and the accuracy of detection, and to capture the haplotype blocks underlying the soybean yield and its components.

Acknowledgments This work was supported by the National Basic Research Program of China (973 Program) (2010CB125906, 2009CB118400) and the National Natural Science Foundation of China (30800692, 31000718). Two anonymous reviewers are thanked for their highly valuable and very helpful comments.

References

Atwell S, Huang Y, Vilhjálmsson B, Willems G, Horton M, Li Y, Meng D, Platt A, Tarone A, Hu T (2010) Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. Nature 465:627–631

Barrero RA, Bellgard M, Zhang X (2011) Diverse approaches to achieving grain yield in wheat. Funct Integr Genomics 11:37–48

Barrett J, Fry B, Maller J, Daly M (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21:263–265

Beló A, Zheng P, Luck S, Shen B, Meyer D, Li B, Tingey S, Rafalski A (2008) Whole genome scan detects an allelic variant of fad2 associated with increased oleic acid levels in maize. Mol Genet Genomics 279:1–10

Blanc G, Wolfe KH (2004) Widespread pale polyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell 16:1667–1678

Bradbury P, Zhang Z, Kroon D, Casstevens T, Ramdoss Y, Buckler E (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633

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

Cardon L, Bell J (2001) Association study designs for complex diseases. Nat Rev Genet 2:91–99

Chan EKF, Rowe HC, Corwin JA, Joseph B, Kliebenstein DJ (2011) Combining genome-wide association mapping and transcriptional networks to identify novel genes controlling glucosinolates in *Arabidopsis thaliana*. PLoS Biol 9:e1001125

Choi I, Hyten D, Matukumalli L, Song Q, Chaky J, Quigley C, Chase K, Lark K, Reiter R, Yoon M (2007) A soybean transcript map: gene distribution, haplotype and single-nucleotide polymorphism analysis. Genetics 176:685–696

Chung J, Babka HL, Graef GL, Staswick PE, Lee DJ, Cregan PB, Shoemaker RC, Specht JE (2003) The seed protein, oil, and yield QTL on soybean linkage group I. Crop Sci 43:1053–1067

Csanádi G, Vollmann J, Stift G, Lelley T (2001) Seed quality QTLs identified in a molecular map of early maturing soybean. Theor Appl Genet 103:912–919

Cui S, He X, Fu S, Meng Q, Gai J, Yu D (2008) Genetic dissection of the relationship of apparent biological yield and apparent harvest index with seed yield and yield related traits in soybean. Aust J Agric Res 59:86–93

Excoffier L, Laval G, Schneider S (2005) Arlequin ver. 3.0: an integrated software package for population genetics data analysis. Evol Bioinform Online 1:47–50

Ersoz E, Yu J, Buckler E (2007) Applications of linkage disequilibrium and association mapping in crop plants. In: Genomics-assisted crop improvement, vol :97, p 119

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

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

Funatsuki H, Kawaguchi K, Matsuba S, Sato Y, Ishimoto M (2005) Mapping of QTL associated with chilling tolerance during reproductive growth in soybean. Theor Appl Genet 111:851–861

Gabriel S, Schaffner S, Nguyen H, Moore J, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M (2002) The structure of haplotype blocks in the human genome. Science 296:2225–2229

Garner C, Slatkin M (2003) On selecting markers for association studies: patterns of linkage disequilibrium between two and three diallelic loci. Genet Epidemiol 24:57–67

Gupta P, Rustgi S, Kulwal P (2005) Linkage disequilibrium and association studies in higher plants: present status and future prospects. Plant Mol Biol 57:461–485

Guzman P, Neece B, Martin DJS, LeRoy S, Grau A, Hughes C, Nelson T (2007) QTL associated with yield in three backcross-derived populations of soybean. Crop Sci 47:111–122

Hardy O, Vekemans X (2002) SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. Mol Ecol Notes 2:618–620

Hoeck JA, Fehr WR, Shoemaker RC, Welke GA, Johnson SL, Cianzio SR (2003) Molecular marker analysis of seed size in soybean. Crop Sci 43:68–74

Holland J, Nyquist W, Cervantes-Martinez C (2003) Estimating and interpreting heritability for plant breeding: an update. Plant Breed Rev 22:9–112

Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y, Li C, Zhu C, Lu T, Zhang Z, Li M, Fan D, Guo Y, Wang A, Wang L, Deng L, Li W, Lu Y, Weng Q, Liu K, Huang T, Zhou T, Jing Y, Li W, Lin Z, Buckler ES, Qian Q, Zhang QF, Li J, Han B (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. Nat Genet 42:961–967

Hyten D, Song Q, Zhu Y, Choi I, Nelson R, Costa J, Specht J, Shoemaker R, Cregan P (2006) Impacts of genetic bottlenecks on soybean genome diversity. Proc Natl Acad Sci USA 103:16666

Hyten D, Choi I, Song Q, Specht J, Carter JT, Shoemaker R, Hwang E, Matukumalli L, Cregan P (2010) A high density integrated genetic linkage map of soybean and the development of a 1536 universal soy linkage panel for quantitative trait locus mapping. Crop Sci 50:960–968

Hyten DL, Pantalone VR, Sams CE, Saxton AM, Landau-Ellis D, Stefaniak TR, Schmidt ME (2004) Seed quality QTL in a prominent soybean population. Theor Appl Genet 109:552–561

Hyten DL, Choi IY, Song Q, Shoemaker RC, Nelson RL, Costa JM, Specht JE, Cregan PB (2007) Highly variable patterns of linkage disequilibrium in multiple soybean populations. Genetics 175:1937–1944

Jun TH, Van K, Kim M, Lee SH, Walker D (2008) Association analysis using SSR markers to find QTL for seed protein content in soybean. Euphytica 162:179–191

Kabelka E, Diers B, Fehr W, LeRoy A, Baianu I, You T, Neece D, Nelson R (2004) Putative alleles for increased yield from soybean plant introductions. Crop Sci 44:784–791

Kassem M, Shultz J, Meksem K, Cho Y, Wood A, Iqbal M, Lightfoot D (2006) An updated ‘Essex’ by ‘Forrest’ linkage map and first composite interval map of QTL underlying six soybean traits. Theor Appl Genet 113:1015–1026

Keim P, Diers BW, Olson TC, Shoemaker RC (1990) RFLP mapping in soybean: association between marker loci and variation in quantitative traits. Genetics 126:735–742

Lai J, Li R, Xu X, Jin W, Xu M, Zhao H, Xiang Z, Song W, Ying K, Zhang M (2010) Genome-wide patterns of genetic variation among elite maize inbred lines. Nat Genet 42:1027–1030

Lam HM, Xu X, Liu X, Chen W, Yang G, Wong FL, Li MW, He W, Qin N, Wang B, Li J, Jian M, Wang J, Shao G, Wang J, Sun SM, Zhang G (2010) Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nat Genet 42(12):1053–1059

Lauvergeat V, Lacomme C, Lacombe E, Lasserre E, Roby D, Grima-Pettenati J (2001) Two cinnamoyl-CoA reductase (CCR) genes from *Arabidopsis thaliana* are differentially expressed during development and in response to infection with pathogenic bacteria. Phytochemistry 57:1187–1195

Lee S, Park K, Lee H, Park E, Boerma H (2001) Genetic mapping of QTLs conditioning soybean sprout yield and quality. Theor Appl Genet 103:702–709

Li D, Pfeiffer TW, Cornelius PL (2008a) Soybean QTL for yield and yield components associated with Glycine soja alleles. Crop Sci 48:571–581

Li J, Huang X, Heinrichs F, Ganal M, Röder M (2005) Analysis of QTLs for yield, yield components, and malting quality in a BC3-DH population of spring barley. Theor Appl Genet 110:356–363

Li Y, Li W, Zhang C, Yang L, Chang R, Gaut B, Qiu L (2010) Genetic diversity in domesticated soybean (*Glycine max*) and its wild progenitor (*Glycine soja*) for simple sequence repeat and single nucleotide polymorphism loci. New Phytol 188:242–253

Li Y, Guan R, Liu Z, Ma Y, Wang L, Li L, Lin F, Luan W, Chen P, Yan Z (2008b) Genetic structure and diversity of cultivated soybean (*Glycine max* (L.) Merr.) landraces in China. Theor Appl Genet 117:857–871

Liu K, Muse S (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. Bioinformatics 21:2128

Lu Y, Yan J, Guimares C, Taba S, Hao Z, Gao S, Chen S, Li J, Zhang S, Vivek B (2009) Molecular characterization of global maize breeding germplasm based on genome-wide single nucleotide polymorphisms. Theor Appl Genet 120:93–115

Lu Y, Zhang S, Shah T, Xie C, Hao Z, Li X, Farkhari M, Ribaut J, Cao M, Rong T (2010) Joint linkage–linkage disequilibrium mapping is a powerful approach to detecting quantitative trait loci underlying drought tolerance in maize. Proc Natl Acad Sci USA 107(45):19585–19590

Ma QH (2007) Characterization of a cinnamoyl-CoA reductase that is associated with stem development in wheat. J Exp Bot 58:2011–2021

Malysheva-Otto L, Ganal M, Röder M (2006) Analysis of molecular diversity, population structure and linkage disequilibrium in a worldwide survey of cultivated barley germplasm (*Hordeum vulgare* L.). BMC Genet 7:6

Mansur L, Lark K, Kross H, Oliveira A (1993) Interval mapping of quantitative trait loci for reproductive, morphological, and seed traits of soybean (*Glycine max* L.). Theor Appl Genet 86:907–913

Mansur LM, Orf JH, Chase K, Jarvik T, Cregan PB, Lark KG (1996) Genetic mapping of agronomic traits using recombinant inbred lines of soybean. Crop Sci 36:1327–1336

Mar L (1996) Molecular markers association associated with soybean plant height, lodging, and maturity across locations. Crop Sci 36(3):728–734

Mather K, Caicedo A, Polato N, Olsen K, McCouch S, Purugganan M (2007) The extent of linkage disequilibrium in rice (*Oryza sativa* L.). Genetics 177:2223–2232

Maughan PJ, Maroof MAS, Buss GR (1996) Molecular-marker analysis of seed-weight: genomic locations, gene action, and evidence for orthologous evolution among three legume species. Theor Appl Genet 93:574–579

Mian MAR, Bailey MA, Tamulonis JP, Shipe ER, Carter TE, Parrott WA, Ashley DA, Hussey RS, Boerma HR (1996) Molecular markers associated with seed weight in two soybean populations. Theor Appl Genet 93:1011–1016

Morgante M, Salamini F (2003) From plant genomics to breeding practice. Curr Opin Biotechnol 14:214–219

Murray M, Thompson W (1980) Rapid isolation of high molecular weight plant DNA. Nucl Acids Res 8:4321–4326

Nei M, Tajima F, Tateno Y (1983) Accuracy of estimated phylogenetic trees from molecular data. J Mol Evol 19:153–170

Orf JH, Chase K, Adler FR, Mansur LM, Lark KG (1999a) Genetics of soybean agronomic traits: II. Interactions between yield quantitative trait loci in soybean. Crop Sci 39:1652–1657

Orf JH, Chase K, Jarvik T, Mansur LM, Cregan PB, Adler FR, Lark KG (1999b) Genetics of soybean agronomic traits: I. Comparison of three related recombinant inbred populations. Crop Sci 39:1642–1651

Palomeque L, Li-Jun L, Li W, Hedges B, Cober E, Rajcan I (2009) QTL in mega-environments: I. Universal and specific seed yield QTL detected in a population derived from a cross of high-yielding adapted: a high-yielding exotic soybean lines. Theor Appl Genet 119:417–427

Palomeque L, Liu L, Li W, Hedges B, Cober E, Smid M, Lukens L, Rajcan I (2010) Validation of mega-environment universal and specific QTL associated with seed yield and agronomic traits in soybeans. Theor Appl Genet 120:997–1003

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

Rafalski J (2010) Association genetics in crop improvement. Curr Opin Plant Biol 13:174–180

Schlueter J, Dixon P, Granger C, Grant D, Clark L, Doyle J, Shoemaker R (2004) Mining EST databases to resolve evolutionary events in major crop species. Genome 47:868–876

Schmutz J, Cannon S, Schlueter J, Ma J, Mitros T, Nelson W, Hyten D, Song Q, Thelen J, Cheng J (2010) Genome sequence of the palaeopolyploid soybean. Nature 463:178–183

Shen R, Fan J, Campbell D, Chang W, Chen J, Doucet D, Yeakley J, Bibikova M, Wickham Garcia E, McBride C (2005) High-throughput SNP genotyping on universal bead arrays. Mutat Res Fundam Mol Mech Mutagen 573:70–82

Smalley MD, Fehr WR, Cianzio SR, Han F, Sebastian SA, Streit LG (2004) Quantitative trait loci for soybean seed yield in elite and plant introduction germplasm. Crop Sci 44:436–442

Specht JE, Chase K, Macrander M, Graef GL, Chung J, Markwell JP, Germann M, Orf JH, Lark KG (2001) Soybean response to water: a QTL analysis of drought tolerance. Crop Sci 41:493–509

Sulpice R, Pyl E, Ishihara H, Trenkamp S, Steinfath M, Witucka-Wall H, Gibon Y, Usadel B, Poree F, Piques M (2009) Starch as a major integrator in the regulation of plant growth. Proc Natl Acad Sci USA 106:10348

Van Inghelandt D, Melchinger A, Lebreton C, Stich B (2010) Population structure and genetic diversity in a commercial maize breeding program assessed with SSR and SNP markers. Theor Appl Genet 120:1289–1299

Vieira AJD, DAd Oliveira, Soares TCB, Schuster I, Piovesan ND, Martínez CA, Barros EGD, Moreira MA (2006) Use of the QTL approach to the study of soybean trait relationships in two populations of recombinant inbred lines at the F7 and F8 generations. Brazil J Plant Physiol 18:281–290

Wang D, Graef GL, Procopiuk AM, Diers BW (2004) Identification of putative QTL that underlie yield in interspecific soybean backcross populations. Theor Appl Genet 108:458–467

Wang J, McClean P, Lee R, Goos R, Helms T (2008) Association mapping of iron deficiency chlorosis loci in soybean (*Glycine max* L. Merr.) advanced breeding lines. Theor Appl Genet 116:777–787

Wen W, Taba S, Shah T, Chavez Tovar VH, Yan J (2011) Detection of genetic integrity of conserved maize (*Zea mays* L.) germplasm in genebanks using SNP markers. Genet Res Crop Evol 58:189–207

Xing Y, Zhang Q (2010) Genetic and molecular bases of rice yield. Annu Rev Plant Biol 61:421–442

Xu Y, Crouch J (2008) Marker-assisted selection in plant breeding: from publications to practice. Crop Sci 48:391–407

Yan J, Shah T, Warburton M, Buckler E, McMullen M, Crouch J (2009) Genetic characterization and linkage disequilibrium estimation of a global maize collection using SNP markers. PloS One 4:e8451

Yan J, Warburton M, Crouch J (2011) Association mapping for enhancing maize (*Zea mays* L.) genetic improvement. Crop Sci 51:433

Yan J, Yang X, Shah T, Sánchez-Villeda H, Li J, Warburton M, Zhou Y, Crouch JH, Xu Y (2010) High-throughput SNP genotyping with the GoldenGate assay in maize. Mol Breed 25:441–451

Yang X, Yan J, Shah T, Warburton M, Li Q, Li L, Gao Y, Chai Y, Fu Z, Zhou Y (2010) Genetic analysis and characterization of a new maize association mapping panel for quantitative trait loci dissection. Theor Appl Genet 121:417–431

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

Yu J, Pressoir G, Briggs WH, Vroh Bi I, 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

Yuan J, Njiti VN, Meksem K, Iqbal MJ, Triwitayakorn K, Kassem MA, Davis GT, Schmidt ME, Lightfoot DA (2002) Quantitative trait loci in two soybean recombinant inbred line populations segregating for yield and disease resistance. Crop Sci 42:271–277

Zhang WK, Wang YJ, Luo GZ, Zhang JS, He CY, Wu XL, Gai JY, Chen SY (2004) QTL mapping of ten agronomic traits on the soybean (*Glycine max* L. Merr.) genetic map and their association with EST markers. Theor Appl Genet 108:1131–1139

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

Zhu YL, Song QJ, Hyten DL, Van Tassell CP, Matukumalli LK, Grimm DR, Hyatt SM, Fickus EW, Young ND, Cregan PB (2003) Single-nucleotide polymorphisms in soybean. Genetics 163:1123–1134