Theor Appl Genet (2008) 116:777–787
DOI 10.1007/s00122-008-0710-x

ORIGINAL PAPER

Association mapping of iron deficiency chlorosis loci in soybean (*Glycine max* L. Merr.) advanced breeding lines

Ju Wang · Phillip E. McClean · Rian Lee · R. Jay Goos · Ted Helms

Received: 2 May 2007 / Accepted: 8 January 2008 / Published online: 22 February 2008

Abstract Association mapping is an alternative to mapping in a biparental population. A key to successful association mapping is to avoid spurious associations by controlling for population structure. Confirming the marker/trait association in an independent population is necessary for the implementation of the marker in other genetic studies. Two independent soybean populations consisting of advanced breeding lines representing the diversity within maturity groups 00, 0, and I were screened in multi-site, replicated field trials to discover molecular markers associated with iron deficiency chlorosis (IDC), a major yield-limiting factor in soybean. Lines with extreme phenotypes were initially screened to identify simple sequence repeat (SSR) markers putatively associated with the IDC. Marker data collected from all lines were used to control for population structure and kinship relationships. Single factor analysis of variance (SFA) and mixed linear model (MLM) analyses were used to discover marker/trait associations. The MLM analyses, which include population structure, kinship or both factors, reduced the number of markers significantly associated with IDC by 50% compared with SFA. With the MLM approach, three markers were found to be associated with IDC in the first population. Two of these markers, Satt114 and Satt239, were also found to be associated with IDC in the second confirmation population. For both populations, those lines with the tolerance allele at both these two marker loci had significantly lower IDC scores than lines with one or no tolerant alleles.

Communicated by C. Hackett.

J. Wang · P. E. McClean (✉) · R. Lee · T. Helms
Department of Plant Sciences, Loftsgard Hall 270B,
North Dakota State University, Fargo, ND 58105-5051, USA
e-mail: phillip.mcclean@ndsu.edu

J. Wang · P. E. McClean · T. Helms
Genomics and Bioinformatics Program, North Dakota State University, Fargo, ND 58105, USA

R. J. Goos
Department of Soil Science, North Dakota State University, Fargo, ND 58105, USA

Introduction

Association mapping (AM) attempts to use the variation in a population to uncover a significant association between a trait and a gene or molecular marker. In plants, AM techniques were used to describe genetic variants associated with flowering time in maize (Thornsberry et al. 2001), disease resistance in rice (Garris et al. 2003), potato (Gebhardt et al. 2004; Simko et al. 2004) and corn (Szalma et al. 2005), yield traits in barley (Kraakman et al. 2004), ecological adaptation in cultivated African rice (Semon et al. 2005), and milling quality and kernel size in wheat (Breseghello and Sorrells 2006). Each of these studies utilized a different statistical approach.

Association mapping relies upon the variation and extent of linkage disequilibrium within the population under study. In contrast to traditional linkage analysis that is limited by the variation in the two parents of the segregating population, AM procedures can effectively compare a greater portion of the variation within a species. Plant geneticists typically have access to several different types of populations for AM. First, large plant introduction collections maintained by units such as the National Plant Germplasm System in the United States Department of Agriculture (http://www.ars-grin.gov/npgs/index.html)

represent much of the diversity in a species and can be considered a random sample of the variation available for analysis. Another population is the collection of cultivars released over the many years of crop improvement. This type of population, or a subset of it, represents a smaller set of the available diversity. Associated with this population would be advanced breeding populations under development and evaluation for future release. Collections of these lines have great potential for applied association mapping experiments because they are typically evaluated in local or regional trials for local adaptation or response to biotic or abiotic stresses. But to utilize any of these populations for AM, it is important to consider the effect of population structure and/or kinship because any association may partially be caused by population admixture, and the admixture can result in spurious marker/trait associations. This may indeed be the case for populations drawn from large collections or from released cultivars. Therefore it is important to apply appropriate statistical methods that account for population structure or kinship. The method developed by Pritchard et al. (2000b) incorporates the estimates of population structure, which are estimated by Bayesian clustering methods (Pritchard et al. 2000a; Falush et al. 2003), into a test statistic for case-control study. First members of the population are assigned to subpopulations in a manner that maximizes Hardy-Weinberg and linkage equilibrium within the subpopulations (Pritchard et al. 2000a). Subpopulations membership is then considered during the process to discover marker/trait associations (Pritchard et al. 2000b). Although the method was originally designed for binary traits, it was later modified by Thornsberry et al. (2001) and extended to the discovery of molecular variants associated with quantitative traits in structured crop populations. Populations of advanced breeding lines from different breeding programs that are tested in local trials pose another problem. Although seemingly unrelated by pedigree, these lines may have common ancestry in their not too distant past. Yu et al. (2006) realized that family relatedness as well as population structure could be a source of Type I and II errors and developed a unified mixed model method to discover marker/trait associations in corn.

Another issue for AM is repeatability, an issue of particular concern when the goal is to discover marker/trait associations that have broad application. This is a concern noted by both Cardon and Bell (2001) and Gambaro et al. (2000) for humans where few associations discovered in one population were confirmed in a second population. The wealth of populations available to plant geneticists, though, especially those available in regional trials, offers the opportunity to confirm associations in a second independent population.

Here we describe an experiment designed to detect associations between molecular markers and iron deficiency chlorosis (IDC) in soybean. IDC occurs in the interveinal tissue of young leaves when iron is unavailable to the plant. This is a common problem in soybean production fields on calcareous soils in the north central states of USA (Hansen et al. 2004). There are multiple steps in the uptake and transfer of iron from the soil to the leaf (Clemens et al. 2002). First, iron must be available in the ferrous form in the soil. If it is available, it must be taken up by the root system where it then enters the xylem stream. Once it has traveled via the xylem to the leaf, it must be unloaded into the cell. Finally, the iron needs to be transferred to the correct cellular location. If any step is not functioning, iron deficiency may result.

As this description suggests, the genetic control of IDC in soybean is quantitative, and the phenotype can range from tolerance to full susceptibility where severe leaf yellowing can significantly reduce yield. Molecular markers linked with this trait were mapped in soybean F_2 and recombinant inbred populations (Lin et al. 2000; Charlson et al. 2003). Those markers though had limited utility for marker-assisted selection because the polymorphisms were not shared among the two different populations. Recently, though, Charlson et al. (2005) described a population specific SSR marker that accounted for 11% of the variation for IDC.

The goal of this research was to apply AM approaches to first identify SSR markers associated with IDC tolerance in a base population of advanced breeding lines and to then determine the utility of those markers in a second independent breeding population. The populations used in this project are composed of modern advanced breeding lines developed by public and private breeding programs for the north central states in the USA. Pedigree information was not available for the lines. Therefore there is a potential for false marker/trait associations because of population structure or family relatedness. Statistical procedures that account for population structure and family relatedness were employed to minimize false positives and maximize power.

Materials and methods

Plant material and IDC rating

Two entirely unique soybean populations were analyzed. First, 139 soybean lines, supplied by major public and private breeding programs, were evaluated in the field in 2002. Plants were grown at three sites near Argusville, Ayr, and Galesburg, ND. The sites ranged in pH from 8.1 to 8.5, and salinity (EC) ranged from 0.2 to 0.5 mmho/cm. Thirty seeds were planted in 1.5 m rows on 0.76 m centers. The experimental design was a randomized complete block with four replications.

Three IDC ratings were made for each location, but only two ratings were available for the Ayr location. The rating was made at the two to three trifoliolate stage, at the five to six trifoliolate stage, and 2 weeks after the five to six trifoliolate stage. The last rating was not made at Ayr. A 1–5 scale was used in this project, where 1 no chlorosis and plants were normal and green; 2 a slight yellowing of the upper leaves and there was no differentiation in color between the leaf veins and interveinal areas; 3 interveinal chlorosis (veins green and interveinal areas chlorotic) was observed in the upper leaves, but no obvious stunting of growth or death of leaf tissue (necrosis) was evident; 4 interveinal chlorosis of the upper leaves with some apparent stunting of growth or necrosis of plant tissue; and 5 severe chlorosis with stunted growth and the youngest leaves and growing point necrotic. This rating system is essentially identical to that used by Lin et al. (1997). The second population ($n = 115$), again consisting of advanced breeding lines from public and private programs, was evaluated in 2003. Plants were grown at three locations near Amenia, Arthur, and Galesburg, ND. At these sites, pH ranged from 7.8 to 8.0, and salinity (EC) ranged from 0.6 to 1.8 mmho/cm. Again, thirty seeds were planted in 1.5 m rows on 0.76 m centers. The experimental design and IDC rating for this population were the same as for the year 2002 population. Check lines known to exhibit IDC tolerance, susceptibility or an intermediate phenotype were included in both field trials.

DNA isolation and SSR fragment amplification

For DNA extraction, the lines were planted in the greenhouse. After 2–3 weeks growth, young leaves were harvested and stored at -80ºC prior to DNA extraction. DNA was isolated by the procedure of Brady et al. (1998). From the 2002 population, 20 lines with IDC rating $\leq$2.2 and 20 lines with IDC rating $\geq$3.4 were screened with 84 SSR markers (Table 1). These lines represented the phenotypic extremes of the population. The SSR markers are evenly distributed across the soybean linkage groups and are known to be polymorphic among many genotypes (Cregan, personal communication). A marker was considered polymorphic between the two phenotypic classes if >85% of the individuals in one phenotypic class contained one of the alleles at that marker locus. Marker sequence information found in SOYBASE (http://soybase.agron.iastate.edu/) were used for primer design. Four SSR markers polymorphic among the extreme lines were selected and used to analyze all lines from the 2002 population. The SSR marker fragments were amplified in a 10 µl reaction mixture consisting of 1 µl 10x PCR buffer, 1 µl containing 1.25 mM of each dNTP, 1 µl forward and reverse primers (20 µM), 20 ng DNA, 0.3 µl Taq DNA polymerase (5 U/µl), and sterile deionized H_2O to volume. Amplification conditions were 95°C for 2 min, 35 cycles of 95°C for 45 s; 53°C for 45 s; 72°C for 1 min; and a 10 min extension at 72°C. PCR products were size separated on 4% SFR agarose (Amresco) gel or 10% polyacrylamide gels. Laser detection techniques were also used to score all lines within the two populations for markers SSRs Satt020, Satt199, and Satt239. A Beckman CEQ2000XL detector was used for this purpose.

Table 1 Distribution of soybean SSR markers used for iron deficiency chlorosis mapping

Linkage group	Marker (genetic location)
A1	Satt276 (17.2), Satt300 (30.9), Satt385 (64.7), Satt236 (93.2)
A2	Satt177 (36.8), Satt187 (54.9), Satt424 (60.6), Satt329 (110.9), Satt409 (145.6)
B1	Satt426 (28.3), Satt197 (46.4), Satt415 (82.9), Satt453 (123.9)
B2	Satt577 (6.1), Satt168 (55.2), Satt020 (72.1), Satt070 (72.8), Satt534 (87.6), Satt063 (93.5)
C1	Satt565 (0.0), Satt194 (26.4), Satt294 (78.7), Satt180 (127.8)
C2	Satt281 (40.3), Satt307 (121.3), Satt357 (151.9)
D1a	Satt184 (17.5), Satt179 (56.2), Satt147 (108.9)
D1b	Satt157 (37.1), Satt141 (72.9), Satt172 (100.9), Satt271 (137.1)
D2	Satt002 (47.7), Satt226 (85.2), Satt186 (105.5)
E	Satt411 (12.9), Satt268 (44.3), Satt231 (70.2)
F	Satt146 (1.92), Satt114 (63.7), Satt510 (71.4), Satt554 (111.9)
G	Satt038 (1.84), Satt324 (33.3), Satt199 (62.2), Satt012 (66.6), Satt191 (96.6)
H	Satt353 (8.5), Satt192 (44.0), Satt541 (53.4), Satt253 (67.2), Satt434 (105.7)
I	Satt419 (21.9), Satt239 (36.9), Satt354 (46.2), Satt292 (82.8), Satt440 (112.7)
J	Satt249 (11.7), Satt414 (37.4), Satt431 (78.6)
K	Satt242 (14.4), Satt441 (46.2), Satt196 (104.8), Satt588 (117)
L	Satt143 (30.2), Satt156 (56.1), Satt373 (107.2
M	Satt590 (7.8), Satt175 (67.0), Satt308 (130.1)
N	Satt009 (28.5), Satt485 (38.1), Satt387 (53.2), Satt339 (75.9), Satt022 (102.1)
O	Satt358 (5.4), Satt259 (39.8), Satt173 (58.4), Satt592 (100.4), Satt243 (119.5)

Map location obtained from soybase (http://soybase.ncgr.org/cgi-bin/ace/generic/search/soybase) on 9/21/2005 are in parentheses

Underlined SSR markers were used for population structure analysis

Statistical analysis

IDC phenotypic data were analyzed separately for the 2002 and 2003 populations. ANOVAs were performed using

SAS (SAS Institute, Cary, NC) to determine IDC score differences between lines and locations, and to assess the allelic effects of two SSR makers. Least significant differences (LSDs) were calculated to test for differences among the four genotypic groups defined by markers Satt114 and Satt239. The means were considered different if the F-test was significant at $P < 0.05$. Broad sense heritability, on an entry-mean basis, was calculated using the following formula: $H^2 = MS_L/(MS_L + MS_E/rt + MS_{LE}/t)$, where MS_L is lines mean square error, MS_{LE} is the line x location mean square, MS_E is the error mean square, r is the number of replications, and t is the number of locations (Fehr 1987). Powermarker (Liu and Muse 2005; http://statgen.ncsu.edu/powermarker/) was used to measure population genetic parameters using the SSR data. Three AM analyses were carried out with the TASSEL 2.0 software (http://www.maizegenetics.net/bioinformatics/tasselindex.htm). First, a single factor analysis of variance (SFA) that did not consider population structure was performed using each marker as the independent variable and comparing the mean performance of each allelic class. This was performed using the general linear model (GLM) function in TASSEL. Next population structure was included in a GLM. Population structure consisted of a Q matrix that describes the percent subpopulation parentage for each line in the analysis. These percentages were derived using the model-based approach described by Pritchard et al. (2000a) and Falush et al. (2003) and implemented in the software STRUCTURE (Pritchard et al. 2000a; http://pritch.bsd.uchicago.edu). The SSR genotype data from 24 SSR markers (20 random SSR markers and the four IDC polymorphic SSR markers; Table 1) was used in the analysis. Since we knew a priori that some of the SSR markers were linked, the linkage ancestry model with correlated allele frequencies was used. Given that soybean is a highly selfing species, the haploid phase setting was used for data analysis. We set k (the number of subpopulations) from 1 to 10 and performed 10 runs for each k value. For each run, a burn in of 10,000 iterations was followed by an additional 20,000 iterations. To choose the best k value, we used the Wilcoxon two sample test as described by Rosenberg et al. (2001). Since ln Pr($X|K$), the natural log of the posterior probabilities from ten runs did not differ between $k = 5$ and $k = 6$, the percentage parentage for the first run of $k = 5$ was used in the Q matrix. This number of markers seemed sufficient because it was shown that 15–20 unlinked SSRs accurately represented human population stratification (Pritchard and Rosenberg 1999) and a similar number of markers properly clustered chicken breeds (Rosenberg et al. 2001). Two mixed linear model (MLM) analyses were performed (Yu et al. 2006). One method used a kinship (K) matrix, and a second used a kinship matrix and the population structure Q matrix. The K matrix was also based on the data for the 24 SSR and consisted of pairwise kinship coefficients for all pairs of lines in each population. The SPAGeDi software (Hardy and Vekemans 2002) was used to calculate kinships coefficients described by Loiselle et al. (1995) and Ritland (1996). As recommended by Yu et al. (2006), all negative kinship values were set to zero. The Loiselle et al. and Ritland matrices was used for the MLM analyses.

Results

Phenotypic analysis of IDC two independent soybean populations

IDC in soybean is a complex trait controlled by both genetic and environmental factors requiring extensive phenotypic scoring. Therefore, each line was rated two or three times. Because the correlations between ratings ranged from 0.61 to 0.92 for the 2002 population and from 0.71 to 0.93 for the 2003 population, we used the average rating for data analysis. The distribution of IDC scores for the 2002 population (Fig. 1a) and 2003 populations (Fig. 1b) were tested for normality using the Kolmogorov-Smirnov (KS) test at a significance level of $P \leq 0.05$. Since the KS significance value for both populations was above the P value threshold ($P_{2002} = 0.070$; $P_{2003} = 0.069$), the phenotypic data was considered to be normally distributed. The IDC scores for the 2002 population ranged from 1.3 to 4.0 with an average of 2.69, while for the 2003 population, the mean was 2.68, and the range ran from 1.8 to 4.3. The mean IDC scores in 2002 were: 1.3, 2.0, 2.2, and 2.3 for the tolerant checks lines A11, KG20, Traill and Council, respectively; 2.9 and 3.0 for the intermediate check lines Barnes and Glacier, respectively; and 3.6 for the susceptible check Stine 480. The mean IDC scores in 2003 were: 1.4, 1.8, 2.2, and 2.2 for the tolerant checks lines A11, S200-2070, Traill and Council, respectively; 3.1 for the intermediate check line Glacier, respectively; and 3.3 for the susceptible check Stine 480.

As expected from lines representing multiple breeding programs, the analysis of variance (Table 2) showed significant variation among lines for both populations. In addition, there was a significant location effect as well as a significant line by location effect. This is indicative of the variation for chlorosis inducing ability that can be observed among locations within a production region such as eastern North Dakota. The analysis of variance results were also used to measure broad sense heritability on an entry mean basis. Those values for the 2002 and 2003 populations were of similar magnitude, 0.96 and 0.95, respectively. These values suggest a low level of error in determining the phenotypic value for each population.

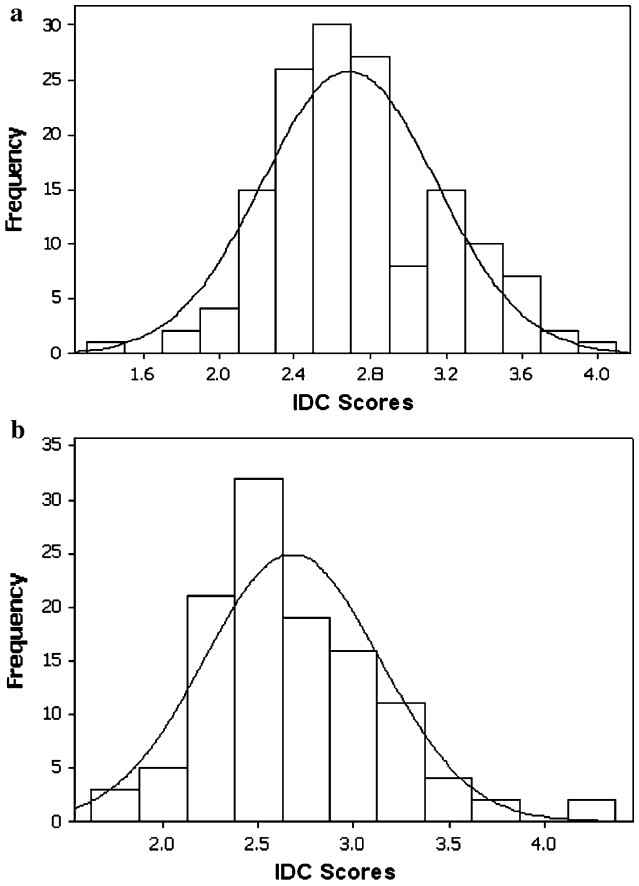

Fig. 1 Comparison of distribution of IDC scores for individual soybean lines in the 2002 (**a**) and 2003 (**b**) populations with normal distribution

Table 2 Analysis of variance mean squares values for iron deficiency chlorosis ratings of two soybean populations grown at three locations

Source of variation	Population			
	2002		2003	
	df	MS	df	MS
Location	2	10.10***	2	244.17***
Line	138	2.18***	114	2.60***
Location x line	276	0.23***	228	0.40***
Replication/location	9	19.69***	9	6.75***
Error	1242	0.15	1026	0.19

*** $P \leq 0.001$

Initial marker discovery

Selective genotyping (Lander and Botstein 1989) was used as an initial screen to find putative marker/trait associations. DNA from individual lines from the 2002 population with the 20 most tolerant and 20 susceptible IDC scores were screened with 84 highly polymorphic and multiple allelic SSR markers. Four markers (Satt020, Satt114, Satt199, and Satt239) were found to be polymorphic between the two subgroups. In general, nearly all lines of the tolerant subgroup contained a single fragment whereas the susceptible pool contained one to several fragments of a different size. All lines in the 2002 and 2003 populations were then scored with these four markers. The mean IDC score for each allele in each population is presented in Table 3. Single factor analyses (SFA) were performed to determine the association between these four markers and the IDC phenotype. Those results are found in Tables 4 and 5. Of these four markers, only Satt199 was not significantly associated with the trait in the 2002 population. Satt020, Satt114, and Satt239 were significantly associated with IDC in all three locations and with the overall mean. For the overall mean scores, these three markers accounted for 19.9, 27.8, and 24.2% of the variation. In addition, two other loci, Satt424 and Satt308, not detected with the bulk screening procedure, were significant in at least two locations. Next, we used the 2003 confirmation population to further measure the association of these markers with IDC. For this population, only Satt114 and Satt239 were significantly associated with traits in each location and with the overall mean. Again, for the overall

Table 3 Mean IDC scores for alleles of SSR loci initially identified by selective genotyping to be potentially affecting IDC phenotype and then analyzed by association mapping techniques

Locus	Allele fragment size (nt)	Year	
		2002 mean*	2003 mean*
Satt114	97	2.96b	3.50b
	100	2.87b	2.73a
	106	2.97b	2.83ab
	109	2.47a	2.54a
Satt239	171	2.50b	2.30a
	177	2.53b	2.70ab
	180	2.10a	2.73ab
	183	2.26ab	2.73ab
	186	2.45b	2.46ab
	189	2.51b	2.60ab
	192	2.89c	2.80b
	195	2.81c	2.79b
Satt020	102	2.84b	2.79 ns
	114	2.56a	2.61 ns
Satt199	158	2.69 ns	2.62 ns
	200	2.64 ns	2.84 ns

* Means followed by different letters were significantly different by the protected LSD test at $P \leq 0.05$. The mean square error was derived from the Q + K MLM model for each marker locus

ns The Q + K MLM model did not find significant differences for these markers

Table 4 Significance of tests for association between soybean SSR markers and iron deficiency chlorosis ratings for the 2002 soybean population using four statistical approaches

Marker	Linkage group	Genetic location	Location															
			Argusville				Ayr				Galesburg				Overall average			
			SFA	Q GLM	K MLM	Q + K MLM	SFA	Q GLM	K MLM	Q + K MLM	SFA	Q GLM	K MLM	Q + K MLM	SFA	Q GLM	K MLM	Q + K MLM
Satt177	A2	36.8	*	ns	*	ns	ns	ns	nd	ns	*	ns	*	ns	*	ns	ns	ns
Satt424	A2	60.6	**	***	**	***	ns	ns	ns	ns	*	*	*	*	*	*	*	*
Satt409	A2	145.6	**	*	*	*	**	ns	nd	ns	*	ns	*	ns	**	*	**	ns
Satt020	B2	72.1	***	*	***	*	***	*	nd	**	***	***	***	***	***	**	***	**
Satt271	D1b	137.1	ns	ns	ns	ns	*	*	*	**	ns	ns	ns	ns	ns	*	ns	*
Satt411	E	12.9	ns	ns	ns	ns	ns	ns	nd	ns	*	ns	*	ns	*	ns	ns	ns
Satt231	E	70.2	**	ns	**	*	*	*	*	*	*	*	*	ns	*	*	**	*
Satt114	F	63.7	***	***	***	***	***	**	***	**	***	***	nd	***	***	***	***	***
Satt199	G	62.2	ns	ns	ns	ns	ns	ns	nd	ns	**	ns	*	ns	ns	ns	ns	ns
Satt191	G	96.6	*	ns	*	ns	ns	ns	nd	ns	ns	ns	ns	ns	ns	ns	ns	ns
Satt239	I	39.6	***	*	***	**	***	ns	***	ns	***	ns	nd	ns	***	*	***	*
Satt414	J	37.4	*	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
Satt308	M	130.1	***	ns	***	ns	**	ns	**	ns	***	*	***	*	***	ns	***	ns
Satt009	N	28.5	ns	ns	ns	ns	**	ns	**	ns	**	ns	**	ns	*	ns	*	ns
Satt339	N	75.9	ns	ns	*	*	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
Satt173	O	5.4	ns	*	ns	*	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns

SFA single factor analysis of variance, *Q GLM* general linear model using the Q population structure matrix, *K MLM* mixed linear model using the K kinship matrix, *Q + K MLM* mixed linear model using the Q population structure matrix and the K kinship matrix

* $P \leq 0.05$, ** $P \leq 0.01$, *** $P \leq 0.001$, *ns* $P > 0.05$, *nd* not determined, software did not converge onto a *P* value

Table 5 Significance of tests for association between soybean SSR markers and iron deficiency chlorosis ratings for the 2003 soybean population using four statistical approaches

Marker	Linkage group	Genetic location	Location															
			Amenia				Arthur				Galesburg				Overall average			
			SFA	Q GLM	K MLM	Q + K MLM	SFA	Q GLM	K MLM	Q + K MLM	SFA	Q GLM	K MLM	Q + K MLM	SFA	Q GLM	K MLM	Q + K MLM
Satt424	A2	60.6	**	ns	ns	ns	ns	ns	ns	ns	**	*	**	*	*	ns	*	ns
Satt020	B2	72.1	ns	ns	ns	ns	ns	ns	ns	ns	**	ns	**	ns	ns	ns	ns	ns
Satt307	C2	121.3	*	ns	*	ns	*	*	*	*	ns	ns	ns	ns	*	ns	*	ns
Satt357	C2	151.9	*	ns	*	ns	ns	ns	ns	ns	*	ns	*	ns	ns	ns	ns	ns
Satt271	D1b	137.1	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	*	ns	ns
Satt411	E	12.9	*	ns	*	*	*	ns	*	*	*	*	nd	ns	*	*	*	**
Satt231	E	70.2	ns	*	ns	ns	ns	ns	ns	ns	*	*	*	ns	ns	ns	ns	ns
Satt114	F	63.7	***	*	***	**	*	*	*	ns	***	**	***	**	***	**	***	**
Satt239	I	39.6	***	*	***	***	**	ns	**	ns	***	*	***	*	***	**	**	**
Satt249	J	11.7	ns	ns	ns	*	ns	ns	ns	ns	ns	**	ns	**	ns	ns	ns	ns

SFA single factor analysis of variance, *Q GLM* general linear model using the Q population structure matrix, *K MLM* mixed linear model using the K kinship matrix, *Q + K MLM* mixed linear model using the Q population structure matrix and the K kinship matrix

* $P \leq 0.05$, ** $P \leq 0.01$, *** $P \leq 0.001$, *ns* $P > 0.05$, *nd* not determined, software did not converge onto a *P* value

mean scores, these two markers accounted for 10.0% and 18.3% of the variation, respectively. Satt020 was only associated with mean scores from the Galesburg site, Satt424 was associated with IDC in two locations, while Satt199 and Satt308 were not associated with the IDC ratings.

SSR marker polymorphism and population structure and kinship analyses

Although these results are promising, it is important to recognize that these associations may be spurious results because of population structure or kinship relationships among the lines in the populations. An additional 20 SSR markers (Table 1) were scored to provide population structure and kinship data for the lines. Collectively, these loci map to 15 of the 20 soybean linkage groups. The number of SSR alleles varied from 2 to 8, with an average of 3.68 and 3.30 markers, respectively, for the 2002 and 2003 populations. The Wilcoxon two-sample test *P*-value that compared the average number of alleles per marker among the 2002 and 2003 populations was 0.374. The major allele frequency for the 2002 population ranged from 0.41 to 0.93 with an average of 0.61. In the 2003 population, the average major allele frequency was 0.58 and ranged from 0.39 to 0.93. The Wilcoxon two-sample test was not significant ($P = 0.339$) when comparing the major allele frequency of the two populations. Gene diversity for the 2002 population ranged from 0.055 to 0.683 with an average of 0.495, and for the 2003 population the range was 0.340 to 0.721 with an average of 0.485. As determined by the Wilcoxon two-sample test, gene diversity was not significantly different between the two populations ($P = 0.191$).

Two analyses were performed to estimate the number of subpopulations. First, a phylogenetic tree was generated using the UPGMA (unweighted pair-group methods using arithmetic average) algorithm with a genetic chord distance matrix (Cavalli-Sforza and Edward 1967). The tree suggested two populations consisting of a total of five clades (data not shown).

Population structure was also estimated using the model-based approach as implemented in the software program STRUCTURE (Pritchard et al. 2000a). This method attempts to determine the number of subpopulations (k) that consist of loci that are in Hardy-Weinberg and linkage equilibrium. Multiple runs are performed for various k values, and posterior probabilities are determined for each run of a given k value. For this analysis, probabilities were determined for 10 runs with the number of subpopulations ranging from one to ten. To determine the appropriate k value, we used the procedure of Rosenberg et al. (2001) that utilizes the Wilcoxon two-sample test to determine which number of subpopulations best characterized the population as a whole by sequentially comparing the posterior probabilities from all runs of a given k, with those probabilities from the run with a k value one larger ($k = 1$ vs. $k = 2$; $k = 2$ vs. $k = 3$ etc.). The smaller k value in the first non-significant Wilcoxson two-sample test was considered to be the estimate of the subpopulation number. For both the 2002 and 2003 populations, it was determined that each consisted of five subpopulations. The F_{ST} statistic for the 2002 population was 0.206 and 0.201 for the 2003 population. For each year, a Q matrix was developed. This $n \times 5$ matrix (n number of lines in each year's population) contained the percentage ancestry from each of the five subpopulations for each line in the entire population.

Pairwise kinship coefficients were calculated using the procedures of Loiselle et al. (1995) and Ritland (1996). These two procedures gave similar kinship coefficients. The distribution of the Loiselle et al. (1995) based coefficients is displayed in Fig. 2. For both populations, more than 50% of the values were less than 0.05, whereas about 30% of the values ranged from 0.05 to 0.25. These results suggest that a subset of the advanced breeding lines we evaluated have a low to a moderate level of relatedness. A nxn K matrix was developed for each year's population using both the Loiselle et al. and Ritland methods.

SSR marker/IDC phenotype associations

Associations between 24 SSR markers (the four identified by prescreening and the 20 used for population structure and kinship analysis) and IDC rating were next determined by GLM and MLM methods. Since IDC ratings varied among locations (Table 2), these associations were determined for each location as well as using the mean rating over all three locations. Tables 4 and 5 present the significance

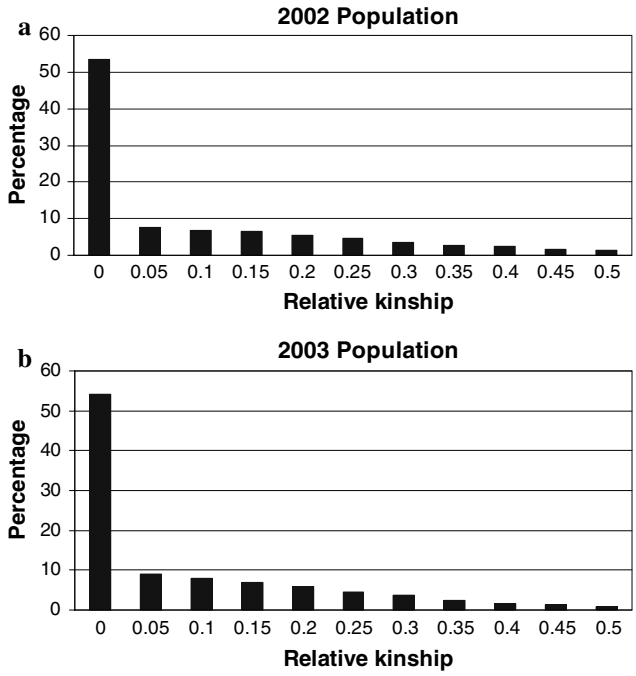

Fig. 2 The distributions of pairwise Loiselle et al. (1995) kinship coefficients for the 2002 (**a**) and 2003 (**b**) soybean populations. Values greater than 0.5 are not shown and account for only 3.7 and 1.7% of the 2002 and 2003 population distributions, respectively

levels for all markers for each of the analyses. Using SFA with the 2002 population data, we observed that 10 SSR markers were significantly associated with IDC rating when the data was averaged over the locations. Fourteen of the markers were significant in at least one location. Five markers were significant for data averaged over all locations using SFA, and eight were significant in at least one location for the 2003 population. Four of the five SSR markers significant for 2003 population were also significant for the 2002 population.

Soybean genotypes are typically classified into thirteen maturity groups that range from 000 to X. In the north central US region, where our trial was conducted, maturity groups 00, 0 and I are predominant. Because our trial was restricted to a small subset of maturity groups, our lines may represent only a minor part of the already limited diversity found in US soybean germplasm (Kisha et al. 1998; Zhu et al. 2003). Therefore, it is a distinct possibility that spurious associations, which may have occurred because of unrecognized population structure or because of kinship between individual lines, may produce a confounding effect which in turn will lead to false positive marker/trait associations. To account for this possibility, we applied the GLM and MLM procedures described by Yu et al. (2006). These procedures require a population structure matrix (Q) or a population structure (Q) and kinship (K) matrix. The Q matrix was derived from the STRUCTURE output using the model-based approach, and the K matrix was developed using the procedures of Loiselle et al. (1995) and Ritland (1996). As mentioned above, genotypes within both populations exhibited a low level of relatedness. For all of the marker trait analysis reported here, both kinship matrices gave identical significance levels for all marker/trait/environment associations. Therefore, we are only reporting the data based on the Loiselle et al. (1995) coefficients.

Seven significant SSR/trait associations were detected with the Q GLM model using 2002 data averaged over the three locations. Four associations were discovered with the 2003 population. The maximum number of associations at a single location was seven, and the range was two to six. Markers Satt114 and Satt239 were the only two confirmed in both populations. The K MLM model defined eight SSR/trait associations from the 2002 population when data averaged overall sites was considered. A total of 13 associations were discovered in at least one location. A similar analysis with the 2003 population discovered five associations overall environments, and eight associations were discovered in at least one location. Only markers Satt424, Satt114, and Satt239 were confirmed in both populations.

Since Yu et al. (2006) showed with maize a better goodness-of-fit with the Q + K than the K MLM model, we also searched for significant associations using that analytical approach. For the 2002 data, six markers were discovered to be associated with IDC ratings averaged over all environments with this method. Only two of these, markers Satt114 and Satt239, were confirmed in the 2003 population.

To determine which of the Q GLM, K MLM, and Q + K MLM was the best fit for the data, the Bayesian Information Criterion (BIC) value was calculated for the IDC data for each model for each year. For both years, the BIC value for the K MLM model (2002 = −263.5; 2003 = −196.7) was less than the Q GLM (2002 = −245.1; 2003 = −187.9) and Q + K MLM (2002 = −245.1; 2003 = −196.2) models. These results suggest that of the three models, the K MLM model provided the best goodness-of-fit.

Among the markers tested, only Satt114 and Satt239 showed a consistent association with IDC ratings among all procedures that considered population structure or relatedness. These were also the only two markers found to be associated with the trait in our confirmation population. To determine the effect of the various allelic combinations at these two loci on IDC rating, the four possible genotypic classes were compared for the two populations (Table 6). For each population, the mean IDC rating of the lines containing the Satt114 and Satt239 tolerance alleles were significantly lower than the other three genotypic classes.

Discussion

Molecular markers associated with a quantitative trait in plants are traditionally identified by developing a population based on a biparental cross, scoring the population phenotypically and with a group of polymorphic markers, and applying one of several statistical approaches to discover significant marker/trait associations. This was the method employed by Lin et al. (1997) and Charlson et al. (2003) to discover several QTL associated with IDC in

Table 6 Iron deficiency chlorosis mean scores for the four Satt114 and Satt239 genotypic classes

Genotype		2002		2003	
Satt114	Satt239	No.	IDC mean*	No.	IDC mean
T[a]	T	30	2.22a	29	2.39a
T	S[a]	44	2.66b	27	2.70b
S	T	15	2.61b	16	2.70b
S	S	49	3.00c	42	2.85b

* Means followed by different letters were significantly different by the protected LSD test at $P \leq 0.05$

[a] *T* tolerance allele: Satt114 = 109 nt band, Satt239 = not 192 or 195 nt band, *S* susceptible allele: Satt114 = not 109 nt band, Satt239 = 192 or 195 nt band

soybean. Collectively, those two analyses utilized three populations. For one population, a major QTL was discovered on linkage group N that accounted for >70% of the phenotypic variation based on a similar visual rating system used in the analyses here (Lin et al. 1997; Charlson et al. 2003). QTL were also discovered with two other populations, but they did not account for nearly the same magnitude of variation. Furthermore, these QTL were population specific. Given those results, we searched for additional markers associated with IDC using AM procedures and verified those markers by using a second population. Since our populations represent the breadth of breeding materials for the north central US, it is thought that these markers might have broad applicability, at least for this region.

AM is an alternative QTL discovery method that relies upon linkage disequilibrium between a marker and a locus that affects a phenotypic trait. The major difference between the traditional and association mapping methods of QTL discovery is the amount of recombination that is analyzed. Whereas a bi-parental mapping population undergoes limited recombination, association mapping populations represent the cumulative recombination history of all of the lines in the study. This increases the possibility of finding significant marker/trait associations over a short genetic distance.

The two populations studied here consist of advanced breeding lines provided by over 30 private and public breeding programs. As a group, North American soybean germplasm has limited diversity (Kisha et al. 1998; Zhu et al. 2003), and the germplasm analyzed here only represents maturity groups 00, 0, and early I. Because of the manner in which private companies develop or obtain germplasm, any two lines may even share the same pedigree. Therefore it is important to consider the relationship among members of the two populations and account for population structure and kinship. To address these issues, we first measured the relatedness of the lines both within and between the two populations. These lines showed a low degree of relatedness using the kinship estimator of Loiselle et al. (1995). Only one line from the 2003 population contained the same molecular genotype as a line in the 2002 population, and this line was eliminated from the study. This suggests the second population is an independent sample of the genotypic variation found within these maturity groups and can serve as a confirmation population.

Single factor analysis of variation, a traditional QTL statistical method, identified 10 loci associated with IDC rating averaged over all locations for the 2002 population. The Q GLM model that utilized population structure identified only seven marker/trait associations. Six of the seven were discovered to be associated with IDC using SFA. The K MLM model, which accounts for kinship alone, discovered eight associations. This number was reduced to six when population structure was also included in the Q + K MLM model. This reduction in significant associations by the Q + K MLM method is generally consistent with results in maize (Yu et al. 2006). Five markers were significantly associated with IDC with the three procedures that considered kinship and/or population structure for this population.

The selective genotyping step (Lander and Botstein 1989) that evaluated lines with extreme phenotypic values identified four makers, Satt020, Satt114, Satt199, and Satt239, associated with IDC. Of these, markers Satt114 and Satt239 were found to be associated with IDC by the three AM methods that controlled population structure and kinship. The selective genotyping step utilized four times as many markers as the diversity analysis. Since the diversity analysis only used a subset of possible markers, and analysis with those markers discovered only one additional significant marker, the selective genotyping step appears to have been worth the expense.

One of the challenges for any QTL analysis is to demonstrate that a marker discovered to be associated with a trait in one population is also associated with that trait in a second population. For humans, many associations between a marker and a disease discovered with AM techniques are not substantiated in a second study (Cardon and Bell 2001; Gambaro et al. 2000). To determine if the IDC marker/trait associations were useful across populations, we tested the second confirmation population. Two of the markers discovered in the 2002 population by using methods that control for population kinship and structure were also associated with IDC in the 2003 population. If instead, the 2003 population was considered to be the discovery population and the 2002 population the confirmation population, the same two markers, Satt114 and Satt239, were also consistently associated with the IDC. And again, it should be noted these two markers were initially identified through the prescreening process, further emphasizing the value of that step.

An important question is whether traditional QTL and AM analyses can uncover the same loci affecting a trait. With AM, a locus must have an effect in multiple lines to be detected whereas a single locus may exhibit a major effect in a population from a bi-parental cross if other factors are not segregating. Therefore it is possible that some loci detected in a bi-parental cross may go undetected using AM procedures. As an example, Lin et al. (1997) discovered a QTL on linkage group N that had a major effect in one of the two populations they analyzed. Our screening of five markers over 73 cM of linkage group N did not detect a locus with an effect on IDC expression. If the QTL from the single bi-parental population of Lin et al. (1997) is present in our AM population, it may be represented at too low a frequency to be detected by the AM procedure.

Lin et al. (1997) identified eight soybean linkage groups that contain QTL associated with IDC. Of the two markers

we discovered to be associated with the IDC in the two populations, one (Satt114) identified a new factor on linkage group F. The second diagnostic marker, Satt239, is located 7 cM from the border of the confidence intervals for several IDC QTL discovered by Lin et al. (1997) on linkage group I. Given that these confidence intervals were broad (60–70 cM) and that Satt239 maps just outside range of markers used in the study, it is feasible that this factor defines the same linkage group I QTL. At the same time though, Satt354, located 2 cM from the border and within the confidence interval, and 9 cM from Satt239, was not associated with IDC in our study. Given the greater number of recombination events sampled in the AM population, it may be that the peak of the Lin et al. (1997) QTL lies nearer Satt239. This example highlights differences that may be observed between bi-parental QTL and AM procedures.

Local or regional variety trials are commonly managed by public plant breeding, plant pathology, and soil science programs. These trials evaluate advanced breeding lines and typically collect yield, agronomic, or other specific performance data such as disease or soil type response. With the advent of high throughput genotyping techniques, it is becoming more feasible to collect molecular marker diversity data that can be used to estimate population structure or kinship. That data could then be coupled with performance data, and the AM techniques described here could be used to identify QTL. Since these trials are repeated annually with new advanced lines, the value of any QTL could be tested in subsequent populations.

For AM to be an effective method of QTL discovery, it is important to consider how many and what type of markers should be used. Population structure has been estimated with from the 20 (Rosenberg et al. 2001) to nearly 100 (Semon et al. 2005) SSR markers. For kinship estimates based on identity by state calculations, as few as 15 were effective to detect relationships among closely related individuals, but about 100 were needed for distantly related individuals (Wilkening et al. 2006). Yet, using an identity by descent calculation (Queller and Goodnight 1989), as few as 20 markers were sufficient for both closely and distantly related individuals (Blouin et al. 1996), a number similar to that calculated by Staub et al. (2000). As for marker type, research has shown that many fewer SSR markers are needed for accurate kinship estimates than single nucleotide polymorphism markers (Blouin 2003). This is important given that for most species SSR marker development is more mature than SNP procedures.

Therefore, an effective and practical approach to AM would include the development of the diversity estimates necessary for population structure and kinship estimates using a medium density suite of SSR markers. To accelerate the process, this step can begin during the growing season by sampling field grown plants and be completed prior to phenotypic data collection. Since screening the extreme lines in the distribution was an efficient method to discover potential associated markers in our experiment, adding that step once the phenotypic data is collected may also accelerate the discovery of significant associations.

In summary, we took advantage of annual field trials of soybean to apply association mapping techniques that consider population structure and kinship to discover loci associated with IDC in soybean. These markers were confirmed with a second population of different advanced breeding lines. This demonstrated that a subset of the originally discovered markers was also associated with IDC in that population. The discovery that several of the markers mapped near previously discovered IDC QTL further substantiate this approach as a valuable experimental method that has potential broad applications for crop genetics and breeding.

Acknowledgments We thank the North Dakota Soybean Council for supporting this research and providing financial assistance to Ju Wang. Special thanks to Edward Buckler IV, Jianming Yu, and Zhiwu Zhang for helpful discussions regarding the mixed linear models analysis. Also, we thank Shahryar Kianian and Marvin Fawley for reviewing the manuscript prior to publication.

References

Blouin MS (2003) DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. Trends Ecol Evol 18:503–511

Blouin MS, Parsons M, LaCaille V, Lotz S (1996) Use of microsatellite loci to classify individuals by relatedness. Mol Ecol 5:393–401

Brady L, Bassett MJ, McClean PE (1998) Molecular markers associated with *T* and *Z*, two genes controlling partly colored seed coat patterns in common bean. Crop Sci 38:1073–1075

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

Cardon RL, Bell JI (2001) Association study designs for complex disease. Nat Rev Genet 2:91–99

Cavalli-Sforza LL, Edward AWF (1967) Phylogenetic analysis: models and estimation procedure. Evolution 32:550–570

Charlson DV, Cianzio SR , Shoemaker RC (2003) Associating SSR markers with soybean resistance to iron deficiency chlorosis. J Plant Nutr 26:2267–2276

Charlson DV, Bailey TB, Cianzio SR, Shoemaker RC (2005) Molecular marker Satt481 is associated with iron-deficiency chlorosis resistance in a soybean breeding population. Crop Sci 45:2394–2399

Clemens S, Palmgren MG, Kramer U (2002) A long way ahead: understanding and engineering plant metal accumulation. Trends Plant Sci 7:309–315

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164:1567–1587

Fehr WR (1987) Principles of cultivar development, Volume 1: Theory and technique. McGraw Hill Inc, New York

Gambaro G, Anglani F, D'Angelo A (2000) Association studies of genetic polymorphisms and complex disease. Lancet 355:308–111

Garris A J, McCouch SR, Kresovich S (2003) Population structure and its effect on haplotype diversity and linkage disequilibrium surrounding the *xa5* lacus of rice (*Oryza sativa* L.). Genetics 165:759–769

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

Hansen NC, Jolley VD, Naeve SL, Goos RJ (2004) Iron deficiency of soybean in the North Central U.S. and associated soil properties. Soil Sci Plant Nutr 50:983–987

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

Kisha TJ, Diers BW, Hoyt JM, Sneller CH (1998) Genetic diversity among soybean plant introductions and North American germplasm. Crop Sci 38:1669–1680

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

Lander ES, Botstein D (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185–199

Lin S, Cianzio S, Shoemaker R (1997) Mapping genetics loci for iron deficiency chlorosis in soybean. Mol Breed 3:219–229

Lin S, Cianzio S, Shoemaker R (2000) Molecular characterization of iron deficiency chlorosis in soybean. J Plant Nutr 23:1929–1939

Liu K, Muse M (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. Bioinformatics 21:2128–2129

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

Pritchard JK, Rosenberg NA (1999) Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 65:220–228

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

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

Queller DC, Goodnight KF (1989) Estimating relatedness using genetic markers. Evolution 43:258–275

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

Rosenberg NA, Burke T, Elo K, Feldman MW, Freidlin PJ, Groenen MAM, Hillel J, Mäki-Tanila A, Tixier-Boichard M, Vignal A, Wimmers K, Weigend S (2001) Empirical evaluation of genetic clustering methods using multilocus genotypes from 20 chicken breeds. Genetics 159:699–713

Semon M, Nielsen R, Jones MP, McCouch SR (2005) The population structure of African cultivated rice *Oryza glaberrima* (Steud.): evidence for elevated levels of linkage disequilibrium caused by admixture with *O. sativa* and ecological adaptation. Genetics 169:1639–47

Simko I, Costanzo S, Haynes KG, Christ BJ, Jones RW (2004) Linkage disequilibrium mapping of a *Verticillium dahliae* resistance quantitative trait locus in tetraploid potato (*Solanum tuberosum*) through a candidate gene approach. Theor Appl Genet 108:217–224

Staub JE, Danin-Poleg, Fazio G, Horejsi T, Reis N, Katzir (2000) Comparative analysis of cultivated melon groups (*Cucumis melo* L.) using random amplified polymorphic DNA and simple sequence repeat markers. Euphytica 115:225–241

Szalma S J, Buckler IV ES, Snook ME, McMullen MD (2005) Association analysis of candidate genes for maysin and chlorogenic acid accumulation in maize silks. Theor Appl Genet 110:1324–1333

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

Wilkening S, Chen B, Hemminki K, Forst A (2006) STR markers for kinship analysis. Hum Biol 78:1–8

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

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