Theor Appl Genet (2015) 128:73–91
DOI 10.1007/s00122-014-2413-9

ORIGINAL PAPER

Genome-wide association study (GWAS) of carbon isotope ratio ($\delta^{13}C$) in diverse soybean [*Glycine max* (L.) Merr.] genotypes

Arun Prabhu Dhanapal · Jeffery D. Ray · Shardendu K. Singh · Valerio Hoyos-Villegas · James R. Smith · Larry C. Purcell · C. Andy King · Perry B. Cregan · Qijian Song · Felix B. Fritschi

Received: 14 March 2014 / Accepted: 11 October 2014 / Published online: 4 November 2014

Abstract

Key message **Using genome-wide association studies, 39 SNP markers likely tagging 21 different loci for carbon isotope ratio ($\delta^{13}C$) were identified in soybean.**

Abstract Water deficit stress is a major factor limiting soybean [*Glycine max* (L.) Merr.] yield. Soybean genotypes with improved water use efficiency (WUE) may be used to develop cultivars with increased yield under drought. A collection of 373 diverse soybean genotypes was grown in four environments (2 years and two locations) and characterized for carbon isotope ratio ($\delta^{13}C$) as a surrogate measure of WUE. Population structure was assessed based on 12,347 single nucleotide polymorphisms (SNPs), and genome-wide association studies (GWAS) were conducted to identify SNPs associated with $\delta^{13}C$. Across all four environments, $\delta^{13}C$ ranged from a minimum of −30.55 ‰ to a maximum of −27.74 ‰. Although $\delta^{13}C$ values were significantly different between the two locations in both years, results were consistent among genotypes across years and locations. Diversity analysis indicated that eight subpopulations could contain all individuals and revealed that within-subpopulation diversity, rather than among-subpopulation diversity, explained most (80 %) of the diversity among the 373 genotypes. A total of 39 SNPs that showed a significant association with $\delta^{13}C$ in at least two environments or for the average across all environments were identified by GWAS. Fifteen of these SNPs were located within a gene. The 39 SNPs likely tagged 21 different loci and demonstrated that markers for $\delta^{13}C$ can be identified in soybean using GWAS. Further research is necessary to confirm the marker associations identified and to evaluate their usefulness for selecting genotypes with increased WUE.

Keywords Carbon isotope ratio · Genome wide association studies (GWAS) · SNPs · Water use efficiency (WUE)

Mention of a trademark or proprietary product does not constitute a guarantee or warranty of the product by the US Department of Agriculture and does not imply approval or the exclusion of other products that may also be suitable.

Communicated by Jianbing Yan.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-014-2413-9) contains supplementary material, which is available to authorized users.

A. P. Dhanapal · F. B. Fritschi (✉)
Division of Plant Sciences, University of Missouri, Columbia, MO 65211, USA
e-mail: fritschif@missouri.edu

J. D. Ray · J. R. Smith
Crop Genetics Research Unit, USDA-ARS, 141 Experimental Station Road, Stoneville, MS 38776, USA

S. K. Singh
Crop Systems and Global Change Lab, USDA-ARS, Beltsville, MD, USA

V. Hoyos-Villegas
Department of Plant, Soil and Microbial Sciences, Michigan State University, East Lansing, MI 48824, USA

L. C. Purcell · C. Andy King
Department of Crop, Soil, and Environmental Sciences, University of Arkansas, Fayetteville, AR 72704, USA

P. B. Cregan · Q. Song
USDA-ARS, Soybean Genomics and Improvement Lab, Beltsville, MD 20705, USA

Introduction

Water deficit stress is one of the major factors limiting crop production and productivity in many regions, and food security in the twenty-first century depends, in part, on improved drought-tolerant cultivars with high yield stability (Tuberosa 2013; Tuberosa et al. 2002). Various abiotic stresses including drought, salinity and extreme temperature affect crop growth and development at different stages. Among these, drought may be the most daunting challenge faced by breeders (Tuberosa 2013; Tuberosa and Salvi 2006). Drought tolerance is complex and driven by diverse drought-adaptive mechanisms that are controlled by a number of genes and environmental factors (Blum 2005; Pinto et al. 2010; Reynolds and Tuberosa 2008). Soybean is one of the most important sources of plant protein and oil worldwide and soybean crops face many challenges posed by environmental stresses.

Over millions of years, plants have evolved mechanisms to tolerate or escape water deficits. These mechanisms range from morphological modifications to physiological adaptations such as water use efficiency (WUE) (Baum et al. 2007; Taji et al. 2004). Water use efficiency can be defined in different ways: in crop production it is commonly defined as the ratio of grain yield to water used during crop growth and often called agronomic WUE (Angus and van Herwarden 2001; Gilbert et al. 2011; Passioura 1977, 2004); at the plant level, WUE can be defined as the amount of biomass produced per unit water transpired, and at the leaf level, it generally refers to photosynthetic carbon gain per unit of water transpired and is generally termed intrinsic WUE (Angus and van Herwarden 2001; Gilbert et al. 2011; Passioura 1977, 2004). Under water-limited condition, grain yield at the crop production level can be expressed as a function of the amount of water used (WU), WUE, and harvest index (HI) (grain yield = WU × WUE × HI; (Passioura 1977; Salekdeh et al. 2009). While the target for crop improvement ultimately is increased agronomic WUE, the focus of this article is at the leaf level and the definition of WUE used is in this context.

Although attempts to improve crop yields by selection for greater WUE can have significant limitations and do not always prove successful, it is widely recognized that improved WUE can enhance yield in certain environments (Condon et al. 2004; Gilbert et al. 2011; Sinclair 2012). Selection for increased WUE has played an important role in improving performance of wheat yield under late-season drought conditions (Condon et al. 2004). However, the use of the WUE trait in breeding programs has largely been limited because of the difficulty associated with measuring actual WUE in large populations. Nonetheless, a promising screening method for WUE was developed in the 1980s based on plant tissue carbon isotope composition (Farquhar et al. 1982; O'Leary 1981). Approximately, 1.1 % of the carbon in the biosphere naturally occurs in the form of the stable isotope ^{13}C and the remaining 98.9 % is ^{12}C (Condon et al. 2002; Farquhar et al. 1989; O'Leary 1981). However, the molar abundance ratio of ^{13}C/^{12}C in plant tissues is usually less than that in atmospheric CO_2 because of discrimination against the 'heavier' ^{13}C during photosynthesis (Farquhar et al. 1989; Farquhar and Richards 1984). The magnitude of this discrimination varies with photosynthetic type (C_3 or C_4), environment and genotype.

Extensive studies on C_3 species have been reported and have confirmed the relationship between carbon isotope composition, whether measured as carbon isotope ratio (δ^{13}C) or carbon isotope discrimination (Δ^{13}C), and WUE (Condon et al. 1990; Ehleringer et al. 1991; Ismail and Hall 1992; Rebetzke et al. 2002). Farquhar and Richards (1984) proposed the use of Δ^{13}C as an expression of the ^{13}C/^{12}C ratio in the plant tissue relative to the ^{13}C/^{12}C ratio in the air. While both δ^{13}C and Δ^{13}C are related to WUE, the correlation between δ^{13}C and WUE is positive, while the correlation between Δ^{13}C and WUE is negative. Because of the correlation with WUE, carbon isotope composition has been analyzed in tissues from a wide range of plant species to assess WUE (Brüggemann et al. 2011; Wingate et al. 2010). The association between WUE and carbon isotope composition is due to a common relationship of the ratio of CO_2 inside and outside of the leaf. That is, as the ratio of internal to external CO_2 decreases, both WUE and carbon isotopic composition increase. Carbon isotope composition has been used widely as an indirect method for the selection of genotypes with improved WUE and productivity in some environments (Cattivelli et al. 2008; Condon et al. 2004), and there has been interest in improving crop performance through direct selection for carbon isotope composition (Araus et al. 2002; Rebetzke et al. 2002). Further, in a growing number of studies it has been used successfully to investigate the role of WUE in drought adaptation (Ahmed et al. 2013; Chen et al. 2012).

Drought tolerance is a complex, quantitative trait. Selection efficiency of drought-tolerance traits could be enhanced with a better understanding of its genetic control (Chen et al. 2011). Quantitative trait locus (QTL) mapping and analysis provides unprecedented opportunities to identify and locate chromosomal regions controlling adaptive traits such as δ^{13}C during plant growth in water-limited conditions. Associations of carbon isotope composition with leaf characteristics and other physiological traits have been reported for several plant species (Condon et al. 1990; Geber and Dawson 1997; Johnson 1993; Saranga et al. 1999). QTLs for δ^{13}C or Δ^{13}C have been reported for Arabidopsis (*Arabidopsis thaliana*) (Juenger et al. 2005), rice (*Oryza indica* and *Oryza indica*) (Laza et al. 2006; Takai et al. 2006), soybean (Specht et al. 2001), cotton (*Gossypium* spp) (Saranga et al. 2004) and barley (*Hordeum vulgare* L.) (Teulat et al. 2002). Sufficient genotypic variation, stability across environments and high broad-sense heritability (H^2) in carbon isotope composition indicate that it may be a promising surrogate for WUE that can be applied in breeding programs for legumes such as soybean as well as in cereal crops (Condon and Richards 1992; Rebetzke et al. 2008; Specht et al. 2001).

Among the different classes of molecular markers currently available, single nucleotide polymorphisms (SNPs) have proven to be the marker of choice for a variety of applications, particularly in breeding. Genome-wide association study (GWAS) is a powerful approach to identify the positions of genetic factors underlying complex traits (Riedelsheimer et al. 2012; Zhao et al. 2011). In a recent study, GWAS was performed in soybean to identify quantitative trait loci (QTLs) controlling seed protein and oil concentration in 298 germplasm accessions exhibiting a wide range of seed protein and oil content (Hwang et al. 2014). As compared to QTL analysis, GWAS can provide relatively higher resolution in terms of defining the genomic position of a gene or QTL, because the level of linkage disequilibrium (LD) is much lower in naturally occurring populations such as human populations or germplasm collections than in biparental populations which are generally used in QTL analyses (Abdurakhmonov and Abdukarimov 2008). Recent advances in genome sequencing and SNP genotyping facilitate association analysis for identification of genomic regions of importance for crop improvement (Rafalski 2010). In soybean, LD has been shown to be more extensive in the heterochromatic than euchromatic regions (Hwang et al. 2014) and it may differ between populations of soybean ancestors (*Glycine soja* Seib. et Zucc.), landraces and cultivars (Hyten et al. 2007). In cultivated soybean, Hwang et al. (2014) reported that the mean LD was 0.2 (r^2) within about 360 Kbp in euchromatic regions whereas it was about 9,600 Kbp at 0.2 in heterochromatic regions. In a study on soybean GWAS analysis, Hwang et al. (2014) were able to successfully map most of the previously reported seed protein and oil QTLs to narrower genomic regions than originally reported.

Investigations of the genetics underlying WUE in soybean have been limited. Previous studies were based on biparental populations and employed restriction fragment length polymorphism (RFLP) and simple sequence repeat (SSR) markers, respectively (Mian et al. 1998, 1996; Specht et al. 2001). WUE of ~36 day old greenhouse-grown plants was determined gravimetrically and identified four independent QTLs in one population (Mian et al. 1996) and two independent QTLs in a second population (Mian et al. 1998), with one of the markers apparently linked to the same QTL in the two populations. In another study, δ^{13}C was determined on juvenile trifoliates from a population of 236 recombinant inbred lines derived from parents that did not differ in δ^{13}C (Specht et al. 2001). Using δ^{13}C data from one irrigated and one non-irrigated field environment in the same year, they only identified QTLs for δ^{13}C that either coincided with maturity or determinacy QTLs, or were not associated with WUE. Thus, despite the importance of drought tolerance, our knowledge of the genetics underlying soybean WUE is minimal at best. Genome-wide association studies, coupled with δ^{13}C, provide opportunities for rapid identification of novel SNP-based markers associated with WUE and facilitate the selection of promising parental genotypes for germplasm improvement and further genetic studies. The primary objective of this study was to use GWAS to identify SNPs associated with δ^{13}C which ultimately may be used to improve WUE in soybean.

Methods

Field experiments and management

Field experiments were conducted in 2009 and 2010 at the Bradford Research and Extension Center (BREC) in Columbia, MO USA (38°53′N, 92°12′ W) and the Rice Research Experiment Station near Stuttgart, AR (34°30′N, 91°33′W). At Columbia, plants were grown on a Mexico silt loam (fine, smectitic, mesic Aeric Vertic Epiaqualf) and at Stuttgart on a Crowley silt loam (fine, montmorillonitic, thermic Typic Albaqualfs). Fields were tilled prior to sowing which occurred on 23 May 2009 and 27 May 2010 in Columbia and 2 June 2009 and 10 June 2010 in Stuttgart. Seeds were planted at 2.5 cm depth at a density of 25 seeds m^{-2}. In Columbia, plots were 4.87 m long and four rows wide with 0.76 m row spacing. At Stuttgart, single-row plots 6.1 m in length and with 0.76 m between rows were sown. The experiments in Columbia were conducted under rainfed conditions, while plots at Stuttgart were furrow irrigated as needed. Preplant applications of P and K were conducted based on results from soil test analyses and corresponding recommendations from the University of Missouri (Columbia) and the University of Arkansas (Stuttgart). Weed control in Columbia was conducted by applying the pre-emergence herbicide sulfentrazone at a rate of 0.3 kg ai ha^{-1} and a post-emergence herbicide sethoxydim at a rate of 2.6 kg ai ha^{-1}. Lambda-cyhalothrin at a rate of 0.23 kg ai ha^{-1} was applied to control insects. Prior to emergence, imazaquin and metolachlor were applied at Stuttgart at rates of 0.14 and 2.24 kg ai ha^{-1}, respectively. For post-emergence weeds control, fomesafen and clethodim were applied at rates of 1.46 and 0.73 kg ai ha^{-1}, respectively.

Experimental design

A total of 385 soybean (*Glycine max* (L.) Merr.) genotypes (376 genotypes from Soybean Germplasm Collection, USDA-ARS, and 9 other genotypes) within maturity group

(MG) IV were planted in a randomized complete block design with three replications at both locations and in both years. The two growing seasons at each location were considered as four environments.

The MG IV genotypes evaluated were selected from the USDA-ARS Germplasm collection. Selection was based on GRIN (Germplasm Resources Information Network, www.ars-grin.gov) data with genotypes falling into one of two groups. The first group (Group A, 182 genotypes) consisted of the highest yielding (>2.5 Mg ha^{-1}) MG IV genotypes with good agronomic traits (height, lodging, shattering, etc.) without regard to any consideration of genetic diversity. In the second group (Group B, 191 genotypes), good agronomic traits were maintained, but the yield threshold was lowered (<2.5 Mg ha^{-1}) and consideration was given to country and province of origin in an attempt to maximize diversity. Accessions were included in the two groups based on GRIN data and recommendations from the soybean germplasm collection curator, Dr. Randall Nelson.

Biomass sampling and δ^{13}C analysis

Shoots of five plants from each plot were randomly harvested at the soil surface at 53 days after planting (DAP) at Columbia in both years, and 50 and 61 DAP at Stuttgart in 2009 and 2010, respectively. Aboveground biomass samples were harvested when all genotypes were at beginning bloom to full bloom (R1 to R2) (Fehr et al. 1971). The samples were dried in an oven at 60 °C until completely dry and then ground to pass a 2 mm screen using a Wiley Mill (Thomas Model 4 Wiley® Mill, Thomas Scientific, NJ USA). After mixing, a subsample of about 1/4th of each sample was ground again using a UDY Cyclone sample mill (MODEL 3010-014, UDY Corporation, CO USA). After thorough mixing, a subsample of about 0.2 g was then transferred to a 15 ml tube (part # 2252-PC-30; SPEX CertiPrep, Inc., NJ USA) and a 9.52 mm diameter stainless ball (440C Stainless Steel Ball, Tolerance/Grade: 100, Abbott Ball Company, Inc., CT USA) was placed inside the tube along with the sample for grinding. Each sample was ground for 9 min at 1,200 rpm using a Geno/Grinder equipped with a large clamp assembly and a 15 ml tube foam holder (SPEX CertiPrep, Inc., NJ USA). Thereafter, about 3 mg of powdered sample was carefully packed in tin capsules and arranged in 96-well plates (Costech Analytical Technologies Inc., CA USA) according to the procedure described by the University of California, Davis Stable Isotope Facility for stable isotope analysis (UC Davis Stable Isotope Facility, CA USA).

The δ^{13}C isotope analysis was conducted using an elemental analyzer interfaced to a continuous flow isotope ratio mass spectrometer. The final δ^{13}C values were expressed relative to the international standard V-PDB (Vienna PeeDee Belemnite). For more information refer to the Stable Isotope facility website, http://stableisotopefacility.ucdavis.edu/13cand15n.html.

SNP genotyping and LD analysis

Genotypic data of ~23,000 SNPs for a select group of 385 soybean genotypes from the application of the SoySNP50 K iSelect SNP Beadchip were obtained (Song et al. 2013). From the original set of 385 genotypes evaluated in four environments (2 years and two locations), phenotypic data for δ^{13}C were available for 373 genotypes. Based on previous work (Pasam et al. 2012), a minimum minor allele frequency (MAF) of ≥5 % was employed. Of the ~23,000 polymorphic SNPs, 12,347 had an MAF ≥5 % across the 373 genotypes and these were evaluated in this study for associations with δ^{13}C.

Linkage disequilibrium (LD) was calculated using 12,347 SNPs with minor allele frequency ≥5 % covering the 20 chromosomes. Calculation of pairwise LD (r^2) among SNPs and identification of haplotype blocks were based on SNPs within 1 Mb windows using the Haploview software (Barrett et al. 2005).

ANOVA, BLUP and heritability

The experimental design was a randomized complete block with three replications at two locations (Columbia and Stuttgart) in two consecutive years (2009 and 2010). Analysis of variance (ANOVA) was calculated by PROC ANOVA (SAS-Institute-Inc 2004). The four environments were designated CO-09, CO-10, ST-09 and ST-10, corresponding to Columbia 2009, Columbia 2010, Stuttgart 2009 and Stuttgart 2010. To analyze the G × E interaction for δ^{13}C from 373 soybean genotypes, the 2 years and two locations were treated as four environments, and analysis of variance was performed using PROC MIXED procedure ($\alpha = 0.05$) of SAS 9.3 using the model as suggested by Bondari (2003).

$$Y_{\mathrm{ijk}} = \mu + G_{\mathrm{i}} + E_{\mathrm{j}} + GE_{\mathrm{ij}} + B_{\mathrm{jk}} + \varepsilon_{\mathrm{ijk}} \quad (1)$$

where μ is the mean, G_{i} is the effect of the ith genotype, E_{j} is the effect of jth environment, GE_{ij} is the interaction of the ith genotype with the jth environment, B_{jk} is the effect of the kth replication within the jth environment and ε_{ijk} is the error. Genotype was considered as a fixed effect and replication nested within environment was used as a random effect.

To minimize the effects of environmental variation, the best linear unbiased predictions (BLUPs) were used for genome-wide association analysis (Kump et al. 2011). BLUP values were derived for each environment independently and also across all environments and were used as phenotypic values for genome-wide association analyses of data from each environment and across all environments. The BLUPs and variance components for $\delta^{13}C$ per genotype were obtained using PROC MIXED procedure of SAS 9.3 (SAS-Institute-Inc 2004). For BLUP determinations for individual environments, all factors were considered as random effects (Littell et al. 1996). To derive across-environment BLUP values, environment was considered as fixed effect and all other factors as random (Piepho et al. 2008; Edae et al. 2014). The broad-sense heritability (Holland et al. 2003) for $\delta^{13}C$ was derived using the variance components obtained from above PROC MIXED procedure of SAS 9.3 (SAS-Institute-Inc 2004) as described (Piepho and Möhring 2007).

Genetic diversity analysis and AMOVA

Summary statistics for the marker data such as minor allele frequency, heterozygosity, gene diversity and polymorphism information content (PIC) were calculated by Power Marker software V 3.25 (Liu and Muse 2005). The PIC value described by Bostein et al. (Botstein et al. 1980) was used to refer the relative value of each marker with respect to the amount of polymorphism exhibited. PIC value was estimated by the following formula:

$$\mathrm{PIC_i} = 1 - \sum_{j=1}^{n} P_{\mathrm{ij}}^2 - \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} 2P_{\mathrm{ij}}^2 P_{\mathrm{ik}}^2 \tag{2}$$

In this formula, P_{ij} and P_{ik} are the frequencies of *j*th and *k*th alleles for marker *i*, respectively. The heterozygosity value indicates the proportion of heterozygous loci detected in single soybean genotype. The gene diversity is defined as the probability that two alleles randomly chosen from the test sample are different. The heterozygosity and gene diversity were calculated to quantify the genetic variation in soybean genotypes evaluated. The common biased estimator of the gene diversity for marker *i* can be obtained using the above equation by dropping the last item as previously described (Chen et al. 2011; Lu et al. 2009). Allele frequency was calculated for characterizing the differentiation and geographic patterns of genetic diversity in the sampled genotypes. Analysis of molecular variance (AMOVA) was calculated by GeneAlEx 6.41 (Peakall et al. 2006) with 1,000 permutations.

Population structure and clustering

The population structure was inferred using the Bayesian model-based software program STRUCTURE 2.2 (Pritchard et al. 2000) using the 12,347 SNPs. The burn-in iteration was 100,000, followed by 100,000 Markov chain Monte Carlo (MCMC) replications after burn-in with an admixture and allele frequencies correlated model. The population structure analysis was performed with five independent iterations with the hypothetical number of subpopulations (*k*) ranging from 1 to 10. By plotting the estimated likelihood value of data [LnP(D)] from the STRUCTURE output and an ad hoc statistic Δ*k,* the correct value of *k* was determined (Evanno et al. 2005). Further analysis was based on the rate of change in the log probability of data between successive *k* values which best describes the population structure based on maximizing log probability or the value at which LnP(D) reaches a plateau. All soybean accessions were assigned to a subpopulation based on the correct *k* ($k = 8$), for which the membership value (*Q* value) was >0.5 (Breseghello and Sorrells 2006), and the population structure matrix (Q) was generated for further association analyses. The kinship matrix (K) was calculated by a built-in function of TASSEL 3.0 software (Bradbury et al. 2007; Buckler et al. 2009) using 12,347 SNPs to obtain the pairwise relatedness without any missing values. The kinship matrix was first generated using the TASSEL cladogram function to calculate a distance matrix. Each element d_{ij} of the distance matrix is equal to the proportion of the SNPs which are different between taxon i and taxon j. The distance matrix is converted to a similarity matrix by subtracting all values from 2 and then scaling, so that the minimum value in the matrix is 0 and the maximum value is 2.

Clustering of genotypes was done with the cladogram function in TASSEL 3.0 (Bradbury et al. 2007; Buckler et al. 2009) to produce a neighbor-joining (NJ) relationship using parsimony substitution models and an unweighted pair group method with arithmetic mean (UPGMA) Newick file. The output Newick file was used as input in TreeDyn 198.3 software (Chevenet et al. 2006) to obtain the final tree.

Genome-wide association analysis

To account for the population structure and genetic relatedness, two statistical models were tested: i) general linear model (GLM) with Q-matrix, and ii) MLM with Q-matrix and K-matrix. The Q- and K-matrices were used as corrections for population structure and/or genetic relatedness (Dhanapal and Crisosto 2013; Pasam et al. 2012; Yang et al. 2010; Yu et al. 2006). Genome-wide association analyses based on these models were conducted with the software TASSEL 3.0 (Bradbury et al. 2007; Buckler et al. 2009). Markers were defined as being significantly associated with $\delta^{13}C$ on the basis of their significant association threshold ($-\mathrm{Log10}\ P \geq 3.00$; $P \leq 0.001$) for GLM + Q and MLM Q + K ($-\mathrm{Log10}\ P \geq 2.00$; $P \leq 0.01$) (Hao et al. 2012; Yang et al. 2010). The *P* values obtained from both GLM + Q and MLM

(Q + K) were used as an input file for a script written with small modifications in R software (R Development Core team 2013) to generate Manhattan plots.

A permutation testing approach was employed to establish the marker trait significance associations using the GLM + Q model. For the GLM + Q model, 10,000 permutation runs were performed using TASSEL software (Bradbury et al. 2007; Bush and Moore 2012; Anderson and Ter Braak 2003). The associations were regarded as significant when adjusted p values were <0.05. Further, a probability value ≤0.001 (−log10 P value ≥3) was used for selecting significant marker trait associations that fit the $p < 0.05$ criterion. To help avoid false positives, population structure (Q) was used in both models and for the MLM (Q + K) model, the kinship matrix (K) was also engaged. Both models were employed in the analysis of each environment. For an SNP to be considered a candidate, it had to exhibit a significant association in both models as well as in at least two environments, which also served to reduce the number of false positives. For both models and analyses by environment as well as over all environments, multiple testing was performed to assess the significance of marker trait associations using QVALUE R 3.1.0 employing the smoother method (Storey and Tibshirani 2003), an extension of the false discovery rate (FDR) method (Benjamini and Hochberg 1995). Markers with qFDR < 0.01 were considered to be significant.

Results

δ^{13}C descriptive statistics

Measurements of δ^{13}C were obtained on 373 MG IV soybean genotypes over 2 years (2009 and 2010) at two locations (Columbia, MO, and Stuttgart, AR). For analysis, each location in each year was considered a different environment. Figure 1 shows the 7-day running averages for solar radiation, maximum and minimum temperature and the daily rainfall for each environment. Conditions during the growing period between planting and sampling (denoted by horizontal lines in the top panel of the figure) were different in each environment. In general, it was warmer (both minimum and maximum temperatures) in Stuttgart than in Columbia and warmer in 2010 than in 2009 for both locations. Most of the time, daily solar radiation was generally greater in Stuttgart compared to Columbia, and especially for Stuttgart daily radiation was greater in 2010 compared to 2009. For the most part, Columbia received more rainfall than Stuttgart, except at the end of the sampling period in 2009. At Stuttgart, it was considerably wetter at the end of the growing period in 2009 than in 2010.

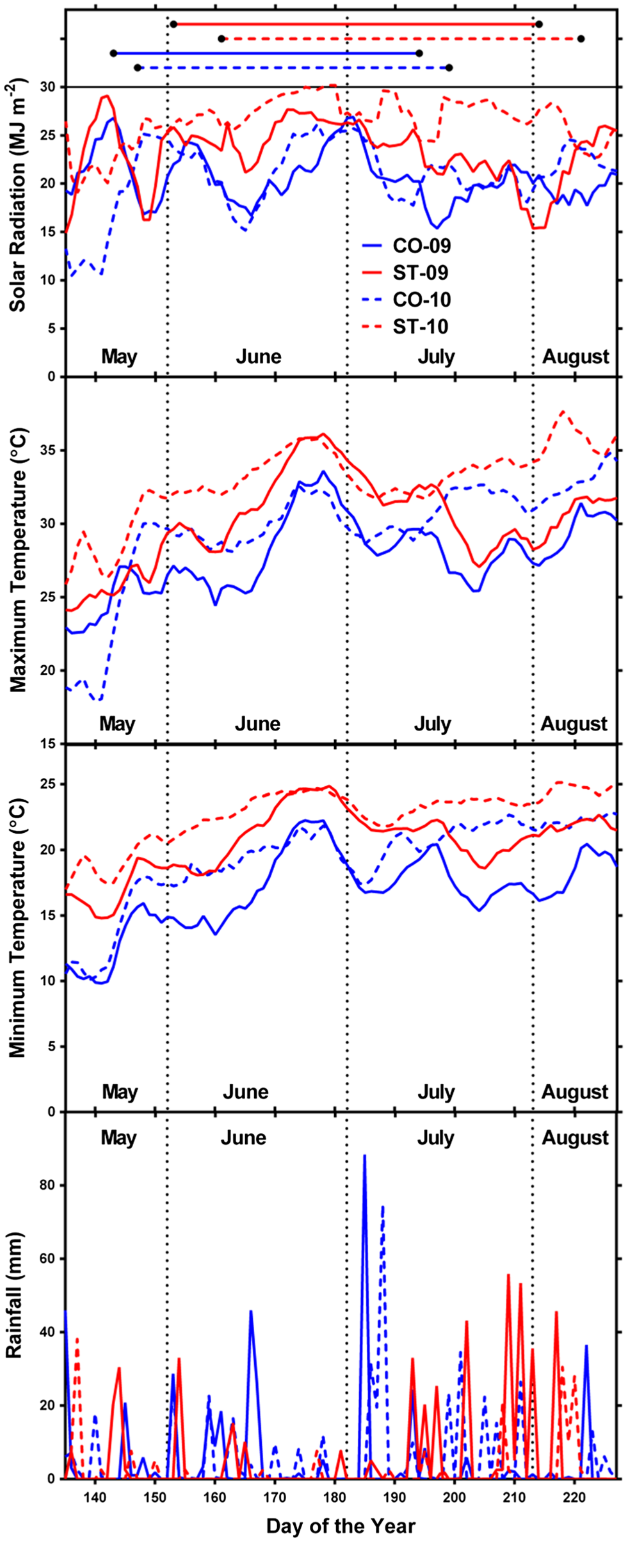


Fig. 1 Weather conditions during the growing season. *Graph* showing 7-day running averages versus day of year for solar radiation, maximum and minimum temperature and daily rainfall during the crop season for the four environments (two locations and 2 years, Columbia and Stuttgart, 2009 and 2010). *Horizontal lines* in the *top panel* indicate the growing period between planting and sampling

Table 1 Descriptive statistics of $\delta^{13}C$ (‰) for plant samples from Columbia in 2009 (CO-09) and 2010 (CO-10), and Stuttgart in 2009 (ST-09) and 2010 (ST-10)

Environments	CO-09	ST-09	CO-10	ST-10
Descriptive statistics				
N	373	373	373	373
Minimum	−29.87	−29.33	−29.82	−30.55
25 % percentile	−29.15	−28.77	−29.22	−29.76
Median	−29.00	−28.58	−29.05	−29.55
75 % Percentile	−28.80	−28.38	−28.86	−29.37
Maximum	−28.01	−27.74	−28.37	−28.85
Mean	−28.98	−28.56	−29.06	−29.56
Std. deviation	0.28	0.28	0.27	0.31
Std. error of mean	0.01	0.01	0.01	0.02

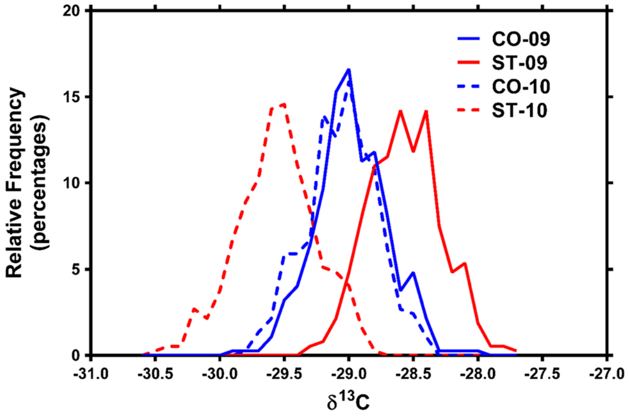


Fig. 2 Frequency distribution of the $\delta^{13}C$ values of the 373 soybean genotypes in four environments (two locations and 2 years, Columbia and Stuttgart, 2009 and 2010)

The difference between extreme mean $\delta^{13}C$ values in each environment was 1.86, 1.46, 1.59 and 1.70 ‰ for CO-09, CO-10, ST-09 and ST-10, respectively. Over all four environments the minimum $\delta^{13}C$ value was in ST-10 (−30.55 ‰) and the maximum value was in ST-09 (−27.74 ‰). Other descriptive statistics of each environment are shown in Table 1, and the frequency distribution of the $\delta^{13}C$ values of the 373 genotypes in each of the four environments is shown in Fig. 2. The ranges of values in the distributions of the 2 years at Columbia (CO-09 and CO-10) were similar, whereas there was considerable difference in the ranges of values between the 2 years at Stuttgart (ST-09 and ST-10). However, no significant ($P \leq 0.05$) skewness or kurtosis in the distributions was found in any of the four environments. Correlations of genotypic $\delta^{13}C$ values among all four locations were highly significant ($P \leq 0.001$) and ranged from $r = 0.35$ between ST-09 and ST-10 to $r = 0.61$ between CO-09 and CO-10. Analysis of variance revealed significant ($P < 0.0001$) genotype (G), environment (E), and G × E interaction effects for $\delta^{13}C$ (Table 2).

Table 2 Analysis of variance (ANOVA) results for the effect of environment (E), genotype (G) and their interaction for $\delta^{13}C$ of 373 soybean genotypes

Source of variation	$\delta^{13}C$		
	df	*F* value	*P* value
E	3	2036.90	<0.0001
Rep(E)	4	72.40	<0.0001
G	372	6.28	<0.0001
G × E	1116	1.54	<0.0001

The two locations (Columbia and Stuttgart) and 2 years (2009 and 2010) at each location were treated as four environments

AMOVA and genetic diversity

Overall, the 373 genotypes represent 11 different national sources, including 244 from South Korea, 60 from China, 41 from Japan, 11 from North Korea, 6 from Georgia, 4 from Korea (North or South Korea not recorded in GRIN), 2 each from Russia and Taiwan and 1 each from India, Mexico and Romania. Within and among components of total genetic variation were evaluated by AMOVA (Table 3). The analysis revealed that the within-population diversity explained most of the genetic diversity (80 %) when compared to among-population diversity (20 %). The distance-based methods NJ and UPGMA identified eight subclusters (C1–C8) as were identified using the model-based method subpopulation groups (G1–G8). The genotypes comprising groups G3 and G4 of the model-based method were consistent with the results of both distance-based methods. The 26 of 111 genotypes from South Korea in G5 were displayed as admixtures in the different clusters. For NJ, the 26 genotypes were clustered as follows: 1 in C1; 4 in C2; 3 in C6; 7 in C7 and 11 in C8; and for UPGMA the following pattern was observed: 1 in C1; 2 in C2; 5 in C6 and 9 in C7 and 9 in C8, in both distance-based methods. Other than these differences, model-based method and distance-based method were the same (results not shown). The 12,347 SNPs used to determine genetic diversity and for further analyses, had an average MAF value of 12.47 % (range 5.0–50.0 %). The gene diversity, heterozygosity and PIC of the 12,347 SNPs averaged 0.20, 0.003 and 0.180, with ranges of 0.05–0.50, 0–0.101 and 0.01–0.40, respectively (Fig. 3). As suggested by its consistent response across different environments and high broad-sense heritability in cowpea and wheat, CID is under tight genetic control (Condon and Richards 1992; Ehdaie

Table 3 Analysis of genetic differentiation among and within eight subpopulation groups of soybean genotypes by AMOVA

Source of variation	df	SS	MS	Est. var.	%	P value
Among subpopulations	7	5.161	0.737	0.016	20	0.0001
Within subpopulation	365	22.938	0.063	0.063	80	0.0001
Total	372	28.099		0.079	100	0.0001

df Degrees of freedom, *SS* sum of squares deviation*, MS* mean squared deviation, *Est. Var.*estimates of variance components, % percentage of total variance contributed by each component

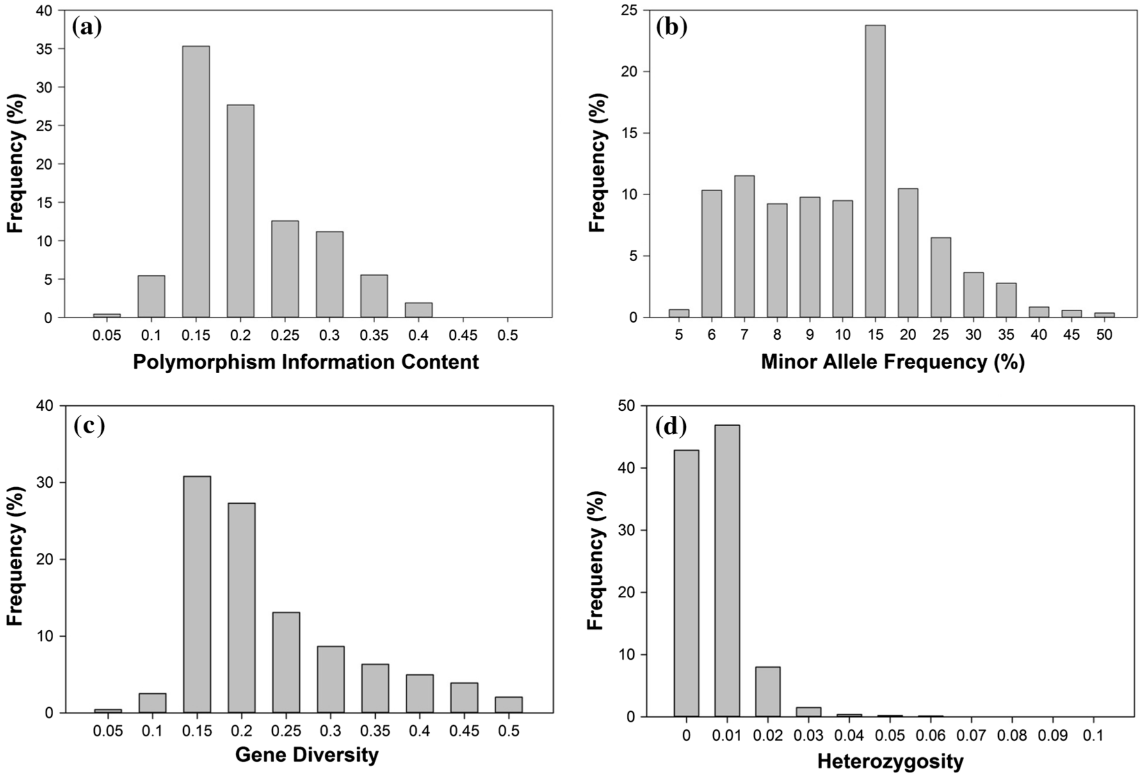


Fig. 3 Distribution of genetic diversity of 12,347 SNPs across 373 soybean genotypes. **a** Polymorphic Information Content (PIC), **b** Minor Allele Frequency (MAF), **c** gene diversity and **d** heterozygosity

et al. 1991; Hall et al. 1990). In our study, broad-sense heritability ranged from 58.68 % for Columbia 2010 to 70.57 % for Stuttgart 2010. The heritability estimate across the two Columbia environments was 76.05 % and across the two Stuttgart environments 67.01 %, and across all four environments 68.20 %.

Genetic structure and linkage disequilibrium

STRUCTURE analysis software was used to determine the population structure (i.e., genetic relatedness) subpopulations (k) of the 373 individual soybean genotypes based on the distribution of the 12,347 SNP loci evaluated in this study. The most probable number of subpopulations was determined by plotting the estimated likelihood value [LnP(D)] obtained from STRUCTURE runs against k. The, LnP(D) appeared to be an increasing function of k for all the values observed. Structure simulation demonstrated that the calculated average of LnP(D) against $k = 8$ was determined to be the optimum k, indicating that eight subpopulations (Fig. 4) could contain all individuals with the greatest probability. Hence, a k value of 8 was selected to describe the genetic structure of the 373 soybean genotypes. The estimated population structure indicated genotypes with partial membership to multiple subpopulations, with few subpopulations exhibiting distinctive identities (Fig. 5 and Supplementary file 4 Figure S2). Significant divergence among subpopulations and average distances (expected heterozygosity) among individuals in the same subpopulations was also assigned (Table 4). Among the eight subpopulations, none had individuals exclusively from one country or region within a country.

In our study, LD analysis was performed using the SNPs with MAF $\geq$ 0.05 and the 373 soybean genotypes evaluated. LD decay was much greater in the euchromatic compared to heterochromatic regions. In the euchromatic regions, the LD decayed to half of its maximum value within approximately 100 Kb, while in the heterochromatic regions, the LD did not decay to half of the maximum value within 1 Mb. Within approximately 300 Kb, LD had decayed to approximately 0.2 in the euchromatic regions, while in heterochromatic regions LD was still >0.5 at 1 Mb. These results are consistent with those reported by Hwang et al. (2014).

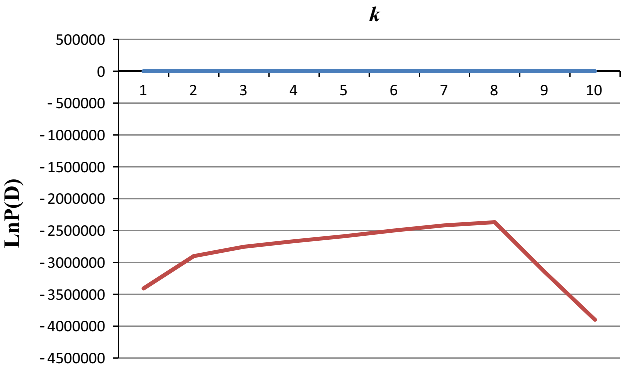


Fig. 4 Population structure results using 12,347 SNPs. Log probability data LnP(D) as function of *k* (number of groups) from the structure run. The *plateau* of the graph at $k = 8$ indicates the minimum number of subgroups possible in the panel

Genome-wide association analysis

Association analyses of 12,347 SNP markers with δ^{13}C values of the 373 genotypes were evaluated by two different models: (1) GLM model adjusted using the Q-matrix and (2) MLM model adjusted using both Q- and K-matrices. Q- and K-matrices were employed to reduce false positives derived from population structure and/or genetic relatedness. The number of potentially false-positive SNPs was also reduced by using BLUP means in both models, requiring that the putative associations be identified by both models, and that putatively associated SNPs must be identified in at least two of the four environments. Additionally, for the analysis across environments, a correction for multiple testing was applied (Benjamini and Hochberg 1995). An overview of the process employing both models by environment and across environments to reduce the 12,347 SNP to 21 putative loci is shown in Fig. 6.

Analysis using the GLM + Q model identified a total of 1,879 SNPs significantly associated ($-$Log10$P \geq 3.00$; $P \leq 0.001$) with δ^{13}C in at least one of the four environments. Of these, 1,229 were identified in at least one of

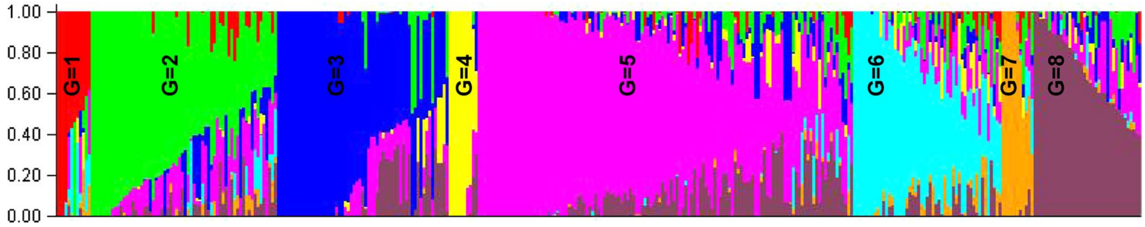


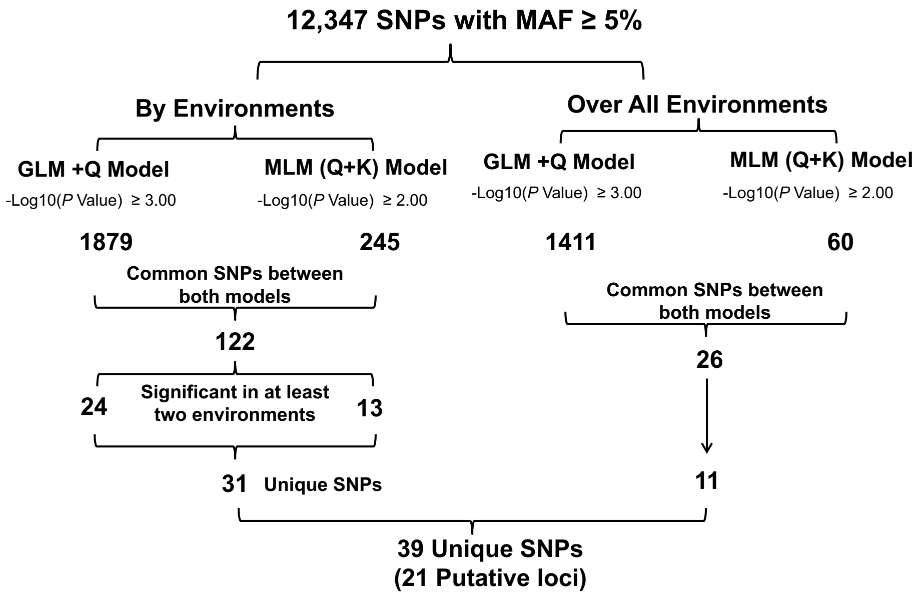
Fig. 5 Estimated population structure of 373 soybean genotypes ($k = 8$). The *y*-axis is the subgroup membership, and the *x*-axis is the genotype. G (G1–G8) stands for a subpopulation

Table 4 STRUCTURE-based analysis showing significant divergence among subpopulation and average distances (expected heterozygosity) among individuals in the same subpopulation

Subpopulation groups	F_{ST}	Heterozygosity	Number of genotypes
G1	0.2186	0.3027	12 (7 South Korea; 4 China; 1 Japan)
G2	0.6526	0.1216	64 (3 China; 2 Georgia; 36 Japan; 2 North Korea; 1 Russia; 19 South Korea; 1 Taiwan)
G3	0.9354	0.0224	59 (59 South Korea)
G4	0.9517	0.0157	10 (10 South Korea)
G5	0.7083	0.0923	127 (7 China; 1 India; 2 Japan; 3 Korea; 2 North Korea; 1 Russia; 111 South Korea)
G6	0.4565	0.2866	51 (38 China; 4 Georgia; 1 Japan; 2 North Korea; 1 Russia; 4 South Korea; 1 Taiwan)
G7	0.8598	0.0814	11 (5 China; 1 Mexico; 1 North Korea; 4 South Korea)
G8	0.5879	0.1495	39 (30 South Korea; 1 Korea; 2 China; I Japan; 3 North Korea; 1 China; 1 North Korea)

F_{ST} fixation index as a measure of genetic differentiation

Fig. 6 An overview of the process using two models to reduce the 12,347 SNPs to 21 putative loci. *Flowchart* showing final SNP selection using two models GLM + Q and MLM (Q + K) from the original 12,347 SNPs with MAF ≥ 5 % analyzed by environments and using the overall means across environments. For all analyses, the BLUP mean was used for association testing



the two Columbia environments (966 at CO-09 and 263 at CO-10) and 650 were found in at least one of the two Stuttgart environments (306 at ST-09 and 344 in ST-10). SNPs exhibiting significant associations with the GLM + Q model are shown (Supplementary File 1, Table S1) for each environment. Manhattan plots showing marker associations for the GLM + Q model for all four environments are shown in supplemental figures (Supplementary File 3, Figure S1a, b, c, d).

As indicated, the GLM model employed corrected for population structure, but not genetic relatedness. However, MLM procedures have been developed to account for both population structure and unequal relatedness (Zhang et al. 2010). The MLM (Q + K) model has been successfully applied to account for population structure in several crops (Aranzana et al. 2010; Breseghello and Sorrells 2006; Yu et al. 2006). Application of the MLM (Q + K) identified a total of 245 SNPs significantly ($-\text{Log}10P \geq 2.00$; $P \leq 0.01$) associated with δ^{13}C in at least one of the four environments. Of these, 81 SNPs were significantly associated with δ^{13}C in at least one of the two Columbia environments (37 at CO-09 and 44 at CO-10, Supplementary File 2, Table S2). The MLM (Q + K) identified 164 SNPs associated with δ^{13}C in at least one of the two Stuttgart environments (97 at ST-09 and 67 in ST-10 (Supplementary File 2, Table S2). Manhattan plots showing marker associations for the MLM (Q + K) model in each of the four environments are shown (Fig. 7a, b, c, d). Table 5 and Supplementary File 5, Table S3 show the top five individual SNPs (greatest $-\text{Log}10\ P$ value) significantly associated with δ^{13}C in each of the four environments along with corresponding probabilities ($-\text{Log}10(P)$) and r^2 values for GLM + Q and MLM (Q + K) models. Across the environments, among the top five SNP, probabilities ranged from 5.00 to 8.35 and r^2 values from the GLM + Q model ranged from 0.08 to 0.14.

Between the two models (GLM adjusted for population structure and MLM adjusted for population structure and genetic relatedness), 122 SNPs showed significant association with δ^{13}C in both models (Fig. 6, Supplementary Files 1 and 2, Tables S1 and S2). Of the 122 SNPs common to both models, 37 SNPs were identified as having a significant association in more than one environment (24 in the GLM + Q model and 13 for the MLM Q + K model). Figure 8 shows the genomic locations of the SNPs identified by each model. Six of the 37 SNPs were common between models (highlighted in yellow in Fig. 8). Thus, the number of unique SNP associations identified between the two models for at least two environments was 31. Markers with *q*FDR < 0.01 in at least two of four environments for both models were considered significant, and all 31 SNPs identified met this criterion.

Based on BLUP means across all four environments, 1,411 SNPs were significantly associated with δ^{13}C in the GLM + Q model and 60 SNPs in the MLM (Q + K) model. Of these, 26 SNPs were found in common between the two models (Supplementary File S6, Table S4). Since the use of overall means across environments precluded exploiting significance in more than one environment as a criterion, we applied a correction for multiple testing (Storey and Tibshirani 2003) to increase stringency. The 26 SNPs identified in common between the two models were subjected to multiple testing. Markers with *q*FDR < 0.01 were considered significant which reduced the number of putative candidate SNPs to 11. These 11 SNP were localized to one locus on CHR 2 and to two loci on CHR 15 (Fig. 8). The scale of Fig. 8 does not allow the separation of closely spaced SNPs,

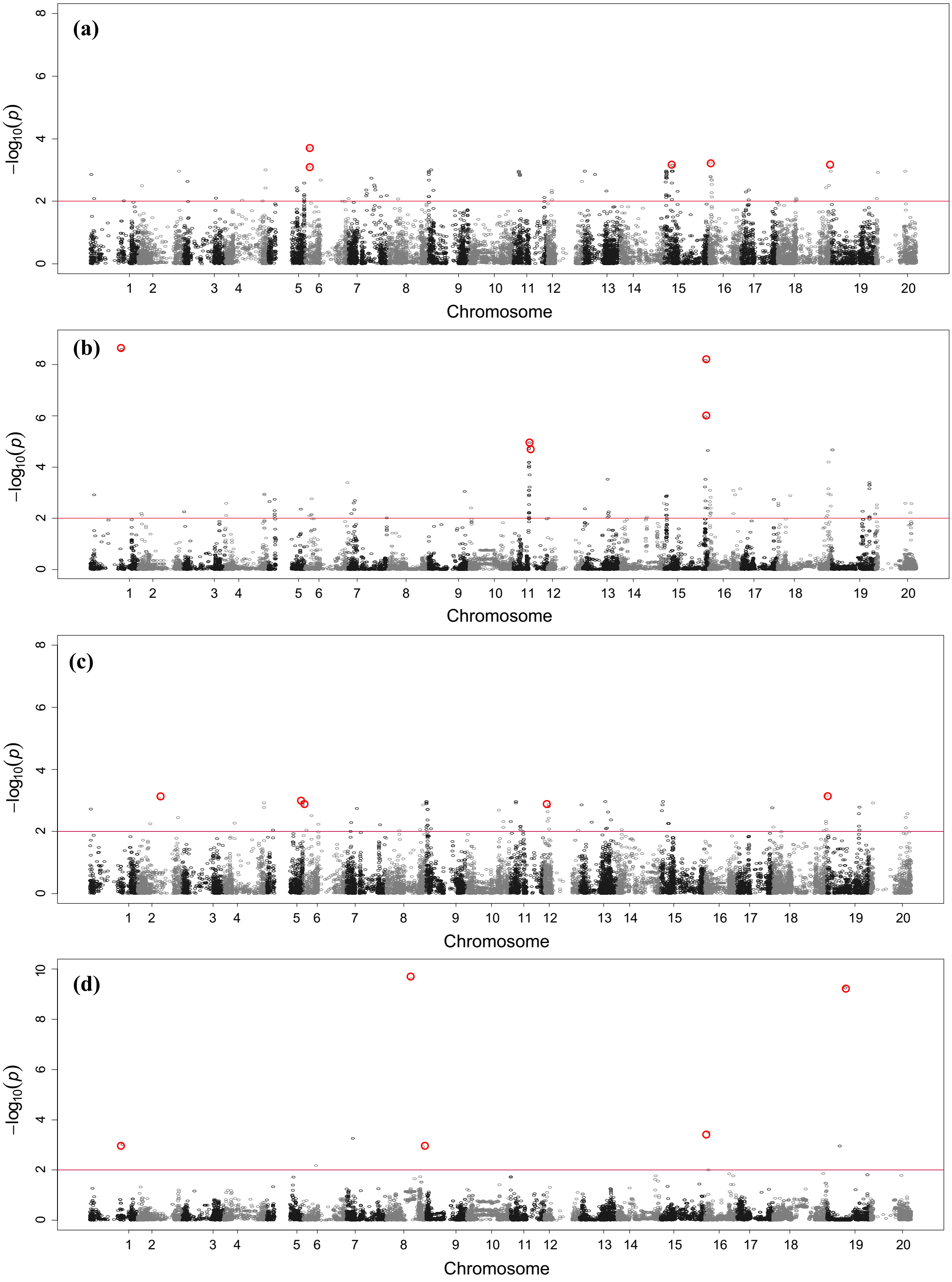


Fig. 7 Manhattan plot of −Log10 (P) vs. chromosomal position of SNP markers from MLM (Q + K) model for two locations in two consecutive years, 2009 and 2010. The plot shows −Log10 P values for each SNP against chromosomal location. (**a**) Columbia 2009; (**b**) Columbia 2010; (**c**) Stuttgart 2009; (**d**) Stuttgart 2010. *Red line* represents the association threshold (−Log 10 $P \geq 2.00$, $P \leq 0.01$)

Table 5 List of top five SNPs significantly associated with δ^{13}C in Columbia and Stuttgart in 2009 and 2010 based on GLM (−Log 10$P \geq 3.00$, $P \leq 0.001$) and corresponding P and R^2 values

SNP ID	Allele		LS Diff	MAF	GLM + Q Model	
	Major	Minor			−Log10 P value	R^2 value
Columbia 2009						
BARC_1.01_Gm04_46049853_A_G	G	A	−0.26	0.12	7.45	0.11
BARC_1.01_Gm04_46062587_A_G	G	A	−0.26	0.12	7.45	0.11
BARC_1.01_Gm14_122063_A_G	A	G	−0.20	0.09	7.11	0.10
BARC_1.01_Gm14_127315_C_T	C	T	−0.20	0.09	7.11	0.10
BARC_1.01_Gm14_67387_G_A	G	A	−0.20	0.09	7.11	0.10
Columbia 2010						
BARC_1.01_Gm13_14458750_G_T	G	T	0.10	0.25	7.87	0.12
BARC_1.01_Gm04_47099634_G_T	G	T	−0.23	0.16	6.74	0.10
BARC_1.01_Gm04_47336623_T_G	T	G	−0.20	0.20	6.74	0.12
BARC_1.01_Gm04_47016634_T_C	T	C	−0.25	0.19	6.70	0.11
BARC_1.01_Gm04_48947393_C_T	C	T	0.14	0.21	5.76	0.10
Stuttgart 2009						
BARC_1.01_Gm18_61847445_C_T	C	T	−0.17	0.30	8.35	0.14
BARC_1.01_Gm18_61850171_G_A	G	A	−0.17	0.30	8.35	0.14
BARC_1.01_Gm18_61239547_G_T	G	T	−0.19	0.28	7.79	0.14
BARC_1.01_Gm18_61867257_A_G	A	G	−0.16	0.30	7.52	0.11
BARC_1.01_Gm18_61217290_T_C	T	C	−0.19	0.28	7.45	0.12
Stuttgart 2010						
BARC_1.01_Gm18_61687642_A_G	A	G	−0.32	0.14	5.74	0.08
BARC_1.01_Gm13_40310510_T_C	T	C	−0.27	0.18	5.30	0.10
BARC_1.01_Gm20_2972792_A_G	A	G	0.15	0.08	5.27	0.09
BARC_1.01_Gm18_18921835_A_G	A	G	−0.31	0.08	5.03	0.08
BARC_1.01_Gm02_46187810_C_T	C	T	0.04	0.10	5.00	0.08

MAF minor allele frequency; *LS-Diff* major allele minus minor allele unadjusted effect (thus, the sign of the value indicates the effect of the allele)

but the locus on CHR 2 and the second locus on CHR 15 were each marked by five consecutive SNPs (only one SNP marked the first locus on CHR 15).

The 31 SNPs identified independently in at least two environments and the 11 SNPs identified using the BLUP means across environments together constituted 39 unique SNPs which were considered as putative candidate SNPs associated with δ^{13}C (Fig. 6 and Supplementary File S7, Table S5). The numbers of significant SNP associations per chromosome are summarized in Table 6 and the relative genomic locations of the SNPs are illustrated in Fig. 8. Details for each of the 39 SNPs putatively associated with δ^{13}C are shown in Table 7. Multiple, closely spaced SNPs likely identified the same putative locus associated with δ^{13}C. Visual examination of Fig. 8, indicates that, overall, 21 putative loci were identified by the process outlined in Fig. 6. The scale of Fig. 8 does not allow individual visualization of closely spaced SNPs, but as shown in Table 7, 15 of the 21 putative loci were identified by one SNP, 2 were identified by two SNPs (putative loci 7 and 10), 1 by three SNPs (putative locus 19), 1 by five SNPs (putative locus 16) and 2 by six SNPs (putative loci 2 and 4).

Potential genes associated with δ^{13}C

Based on the 60 bp sequences flanking SNPs (Supporting information Table S1; (Song et al. 2013), a blast search was conducted with default parameters in Phytozome v9.1 (http://www.phytozome.net/). The search indicated that 25 of the 39 SNPs were located in a gene (Table 7). For the 14 SNPs not located in a gene, information on the gene closest to each SNP is provided in Table 7.

Discussion

Genetic diversity and population structure of soybean genotypes

In this study, 12,347 SNPs were used to estimate the genetic diversity and population structure of 373 soybean genotypes. However, despite the large number of SNPs, gaps in the distribution of the SNPs across the genome were observed (Fig. 8). Among these 373 MG IV soybean genotypes, the gaps in SNP coverage may represent areas

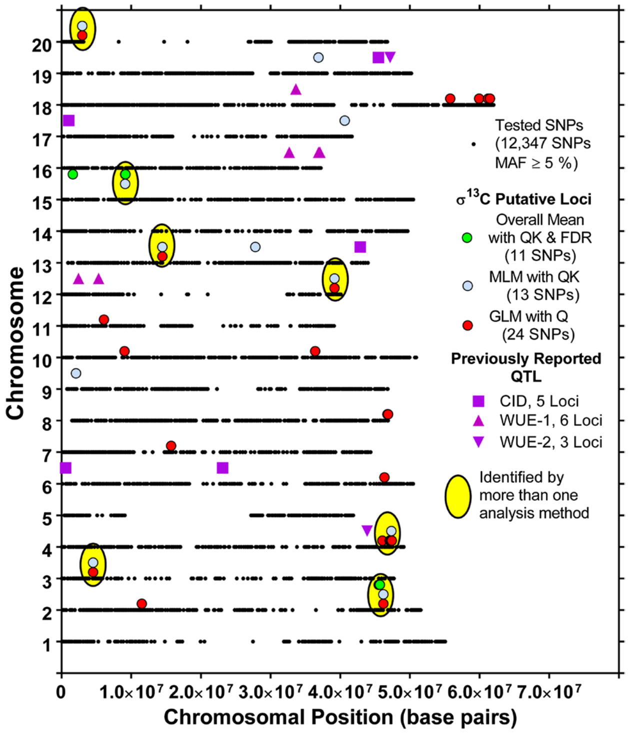


Fig. 8 Location of putative loci significantly associated with δ^{13}C in more than one environment and across environment with previously identified QTLs for CID (Specht et al. 2001) and WUE (Mian et al. 1998) as shown in Soybase (www.soybase.org, [Grant et al. 2013]). For each chromosome, the *black dots* represent the location of a SNP evaluated for association with δ^{13}C

Table 6 Number of SNPs associated with δ^{13}C and their chromosomal location

S.No	Name of chromosome	Number of SNPs
1	Chromosome 2	7
2	Chromosome 3	1
3	Chromosome 4	6
4	Chromosome 6	1
5	Chromosome 7	1
6	Chromosome 8	2
7	Chromosome 9	1
8	Chromosome 10	3
9	Chromosome 11	1
10	Chromosome 12	1
11	Chromosome 13	2
12	Chromosome 15	6
13	Chromosome 17	1
14	Chromosome 18	4
15	Chromosome 19	1
16	Chromosome 20	1

These 39 SNPs showed association in more than one environment or across environments using GLM +Q and MLM (Q + K) models

of extreme similarity. This similarity might be expected given that the SNP data used in this analysis were selected based on whole genome sequence data obtained from a diverse set of six *G. max* genotypes with a range of maturities as well as two wild soybean accessions (Song et al. 2013). Additionally, only a few lines showed a high level of heterozygosity (5 % of the SNP loci, Fig. 3d). These may be due to natural outcrossing (Ray et al. 2003) during propagation of the germplasm or other sources of contamination.

Based on the 12,347 SNPs, mean PIC was 0.18, which was less than the 0.31 reported for 191 soybean (*G. max*) landraces using 1,142 SNPs (Hao et al. 2012) (Fig. 3a). Of the 373 genotypes studied here, heterozygosity was zero in 42.82 % of the entries and average heterozygosity was 0.003, which was also less than the 0.01 reported for 191 soybean genotypes by Hao et al. (2012). The mean gene diversity coefficient was 0.20 for the 373 genotypes, which was less than the 0.39 reported for the 191 genotypes studied by Hao et al. (2012) and the 0.35 reported for 303 cultivated and wild soybean (*G. soja*) using 554 SNPs (Li et al. 2010). Since only cultivated soybean from a single maturity group and no wild soybean was examined in the present study, lower PIC, heterozygosity and genetic diversity coefficients than those found by Hao et al. (2012) and Li et al. (2010) were not surprising.

The genetic structure of soybean populations has been studied previously using both SSR and SNP markers (Hao et al. 2012; Li et al. 2010). The 373 soybean genotypes evaluated in the present study were classified into eight subpopulations with significant divergence among subpopulations. The individuals within each subpopulation were independent of their collection sites. The low genetic differentiation among genotypes could be a result of gene flow due to movement of seeds. Seed exchange among farmers is a mechanism used to enhance diversity of local germplasm, which may result in increased distribution of alleles among different populations irrespective of their geographical distance (Louette et al. 1997). The results of the present study indicated high genetic diversity within subpopulations and less genetic diversity among subpopulations. Similar results have been found in other crops using SSR markers (Aranzana et al. 2010; Belamkar et al. 2011; Cao et al. 2012; Shiferaw et al. 2012; Wang et al. 2011). However, a study of the population structure of 40 wild soybeans from China with 20 SSR markers showed contrary results to those observed in our study (Guo et al. 2012). The reason behind this may be that the wild soybean evaluated in that study was significantly differentiated from other regional soybeans, as evidenced by their low allelic richness, genetic diversity and high ratios of regionally unique and fixed alleles (Guo et al. 2012). These genetic attributes suggest that wild soybean may have undergone

Table 7 List of putative SNPs and potential genes associated with δ^{13}C based on 39 SNPs identified in more than one environment or across environments

Locus	SNP ID[a]	Model[b]	Williams 82 allele	Alternative allele	Name of gene[c]	Position in the gene	Functional annotation of the gene
1	BARC_1.01_Gm02_11500717_A_C	G–	A	C	Glyma02g13240	Intron	Uncharacterised BCR, YhbC family COG0779
2	BARC_1.01_Gm02_45509202_T_C	–O	T	C	Glyma02g40340	NA	Protein kinase/phosphorlation activity
	BARC_1.01_Gm02_45542400_A_C	–O	A	C	Glyma02g40310	NA	Translation initiation factor activity
	BARC_1.01_Gm02_45573752_A_G	–O	A	G	Glyma02g40340	NA	Protein kinase/phosphorlation activity
	BARC_1.01_Gm02_45595304_A_G	–O	A	G	Glyma02g40341	NA	Protein kinase/phosphorlation activity
	BARC_1.01_Gm02_45691551_A_G	–O	A	G	Glyma02g40460	CDS	Protein of unknown function (DUF1191)
	BARC_1.01_Gm02_46187810_C_T	GM-	C	T	Glyma02g40981	CDS	Protein kinase/phosphorlation activity
3	BARC_1.01_Gm03_4529705_A_G	GM-	A	G	Glyma03g04410	Intron	Protein kinase/phosphorlation activity
4	BARC_1.01_Gm04_46049853_A_G	G–	A	G	Glyma04g39890	NA	Dual specificity phosphatase, catalytic domain
	BARC_1.01_Gm04_46062587_A_G	G–	A	G	Glyma04g39910	CDS	PPR repeat
	BARC_1.01_Gm04_47099634_G_T	G–	G	T	Glyma04g41210	NA	PPR repeat
	BARC_1.01_Gm04_47252608_A_C	G–	A	C	Glyma04g41420	CDS	PPR repeat
	BARC_1.01_Gm04_47336623_T_G	GM-	T	G	Glyma04g41490	NA	Protein binding
	BARC_1.01_Gm04_47412879_A_G	G–	A	G	Glyma04g41540	NA	Glutamate synthase activity
5	BARC_1.01_Gm06_46350377_T_C	G–	T	C	Glyma06g43040	NA	Protein of unknown function (DUF640)
6	BARC_1.01_Gm07_15736416_G_A	G–	G	A	Glyma07g16030	NA	No annotation available
7	BARC_1.01_Gm08_46785130_T_C	G–	T	C	Glyma08g48050	NA	Structural constituent of ribosome (Intracellular)
	BARC_1.01_Gm08_46871422_G_A	G–	G	A	Glyma08g48160	NA	No annotation available
8	BARC_1.01_Gm09_2069867_G_A	-M-	G	A	Glyma09g02960	CDS	Oxidation–reduction process (Histidinol dehydrogenase)
9	BARC_1.01_Gm10_36414621_G_A	G–	G	A	Glyma10g09720	NA	Syntaxin, plant
10	BARC_1.01_Gm10_9026417_A_G	G–	A	G	Glyma10g09720	NA	Syntaxin, plant
	BARC_1.01_Gm10_9026915_T_C	G–	T	C	Glyma10g27660	NA	Metabolic process (Acotinase)
11	BARC_1.01_Gm11_6078943_A_C	G–	A	C	Glyma11g08590	CDS	No Annotation available
12	BARC_1.01_Gm12_39185574_C_A	GM-	C	A	Glyma12g36060	NA	Protein of unknown function (DUF822)
13	BARC_1.01_Gm13_14458750_G_T	GM-	G	T	Glyma13g11700	CDS	AMP-binding enzyme
14	BARC_1.01_Gm13_27801871_G_A	-M-	G	A	Glyma13g24380	NA	No annotation available
15	BARC_1.01_Gm15_1609497_C_T	–O	C	T	Glyma15g02390	Intron	Mitochondrial processing peptidase, alpha subunit
16	BARC_1.01_Gm15_9117659_T_C	–O	T	C	Glyma15g12330	NA	NADH-ubiquinone oxidoreductase complex I, 21 kDa subunit
	BARC_1.01_Gm15_9127493_G_A	-MO	G	A	Glyma15g20631	NA	No annotation available
	BARC_1.01_Gm15_9145025_A_G	-MO	A	G	Glyma15g12370	Intron	Domain of unknown function (DUF588)
	BARC_1.01_Gm15_9151938_A_C	-MO	A	C	Glyma15g12380	Intron	Stress associated endoplasmic reticulum protein (Serp1/Ramp4)
	BARC_1.01_Gm15_9162160_A_G	–O	A	G	Glyma15g12400	Intron	Protein/DNA binding
17	BARC_1.01_Gm17_40653085_T_C	-M-	T	C	Glyma17g36700	CDS	Protein of unknown function, DUF607

Table 7 continued

Locus	SNP ID[a]	Model[b]	Williams 82 allele	Alternative allele	Name of gene[c]	Position in the gene	Functional annotation of the gene
18	BARC_1.01_Gm18_55796256_T_G	G–	T	G	Glyma18g46050	NA	Leucine-rich repeat-containing protein
19	BARC_1.01_Gm18_59953429_A_G	G–	A	G	Glyma18g41760	NA	No annotation available
	BARC_1.01_Gm18_61239547_G_T	G–	G	T	Glyma18g52720	Intron	Regulation of sequence-specific DNA binding transcription factor activity
	BARC_1.01_Gm18_61500395_T_C	G–	T	C	Glyma18g53170	NA	Glutamine metabolic process
20	BARC_1.01_Gm19_36877571_T_C	-M-	T	C	Glyma19g29316	NA	Peptide deformylase.
21	BARC_1.01_Gm20_2972792_A_G	GM-	A	G	Glyma20g03275	NA	MSS1/TRME-related GTP-binding protein

CDS coding, *DNA* sequence; *NA* not applicable

[a] SNP ID includes the abbreviation for Glycine max (Gm) followed by chromosome number, genomic location and alleles

[b] Model for which SNP was found to be significantly associated, G = GLM +Q; M = MLM + QK; O = overall mean across environments

[c] Name of the gene is based on information in Soybase

severe adaptive selection for their ecogeographical conditions and had less genetic exchange with inland populations (Guo et al. 2012).

Comparison of model-based diversity and distance-based diversity and their significance are essential for population studies in plants (Guo et al. 2012). In this study, with a few exceptions, results of the model-based method (STRUCTURE) were largely in accordance with the results obtained using the distance-based method (NJ and UPGMA), (results not shown). Similar results were found for soybean, peanut (*Arachis hypogaea* L.) and peach (*Prunus persica*) by others (Belamkar et al. 2011; Cao et al. 2012; Guo et al. 2012). Comparison of the two clustering methods (neighbor joining and UPGMA) found less than 7 % of individuals falling into different clusters based on the method used (results not shown).

Although the MG IV soybean genotypes evaluated in this study were from distinct geographical regions, they did not show any regional or provincial clustering within the eight subpopulations determined from clustering with the 12,347 SNPs. Although the genotypes were originally selected to fall into two groups based on yield estimates from the germplasm database, yield values were randomly distributed among the eight subpopulations (Supplementary File 4, Figure S2).

Carbon isotope ratio

The ability of plants to respond to different levels of available water is variable and complex. Understanding the relationship between genotype and phenotype is essential for the improvement of complex traits in economically important crop species such as soybean. Across the four environments δ^{13}C values in this study ranged from −30.55 to −27.74 ‰, which is within the range of δ^{13}C values previously reported for soybean (Yoneyama et al. 2000). As indicated by the frequency distributions and descriptive statistics (Fig. 2; Table 1), the two Columbia environments (CO-09 and CO-10) were very similar, but the two Stuttgart environments (ST-09 and ST-10) were different. ST-09 had the lowest average δ^{13}C values of the four environments and ST-10 had the highest. Differences between the Columbia and the Stuttgart locations include latitude, soil type and irrigation (furrow irrigation in Stuttgart and no irrigation in Columbia), all of which may affect the growth response of soybean. These and other environmental influences (see Fig. 1) likely affected the δ^{13}C responses observed. For example, in water-limited environments, δ^{13}C would be expected to increase (become less negative) due to partial stomatal closure and an increase in WUE (Specht et al. 2001). Among the eight subpopulations, none had individuals with highest or lowest δ^{13}C exclusively clustered in it (Fig. 5).

GWAS analysis

GWAS provides a promising tool for the detection and mapping of quantitative trait loci (QTLs) underlying complex traits. Application of the GLM model with Q as corrections for population structure indicated that about a third of the markers tested had significant associations in at least one of the environments. A greater number of significant associations were identified in one of the Columbia environments than in the two Stuttgart environments (966 and 305). The greatest number of significant markers identified by the GLM + Q model was in the CO-09 environment and the fewest significant associations were identified in the CO-10 environment. Similarly, CO-10 showed the fewest significant associations using the MLM (Q + K) model, for which a total of 44 SNPs were identified as having significant associations with δ^{13}C. The greatest number of significant markers was found in the ST-09 environment (97 SNPs), followed in order by the ST-10 (67 SNPs), CO-10 (44 SNPs) and CO-09 (37 SNPs) environments. Climatic and other differences that affected tissue δ^{13}C may account for the variation in the number of markers that showed significant associations among environments.

Of all the SNPs identified by either model in the analysis by environment, 122 were identified by both models as being significantly associated with δ^{13}C. Of these 122 SNPs, 31 were identified as significant in more than one environment. In addition to the putative associations identified when analyzed by environment, 11 SNPs significantly associated with δ^{13}C were identified by analysis based on the means across environments. Together, these analyses identified 39 unique SNPs associated with δ^{13}C (three SNPs were common between the analyses).

In total, 39 unique SNPs were identified as putatively associated with δ^{13}C in more than one environment or across environments (Table 7). The genomic distribution of these SNPs revealed that several are located close together and likely mark the same locus (Fig. 8; Table 7). Thus, we putatively identified 21 genomic regions on 16 chromosomes that are highly likely to contain genes affecting δ^{13}C. While the SNPs identified as significantly associated with δ^{13}C in single environments may be important, particularly given the independence of the field experiments, those that were identified in at least two environments or from across all environments are likely the most stable. Additionally, six of the 21 putative loci were independently identified by two of the three analysis methods [GLM + Q, MLM (Q + K)], or across all environments) and one locus on CHR 2 was identified by all three methods (Fig. 8). These loci likely have a greater potential of identifying major QTLs.

Several QTLs have been identified for δ^{13}C in other species (Chen et al. 2012; Gu et al. 2012; Hervé et al. 2001; Juenger et al. 2005; Mano et al. 2005; Specht et al. 2001). For soybean, five QTLs for CID located on chromosomes 6 (2), 13, 17, and 19 (Specht et al. 2001) and nine QTLs for WUE located on chromosomes 4 (2), 12 (2), 16, 18, and 19 (Mian et al. 1996) are identified in Soybase [www.soybase.org (Grant et al. 2013)]. The putative locations of the reported CID and WUE QTLs are included in Fig. 8. None of the 21 putative loci identified in this study as being associated with δ^{13}C were located close to QTLs for CID identified by Specht et al. (Specht et al. 2001). However, one putative δ^{13}C loci was located close to a QTL for WUE (chromosome 4, Fig. 8) identified by Mian et al. (1996). Interestingly, only one of the nine WUE QTLs reported by Mian et al. (1996) was located near one of the CID QTLs reported by Specht et al. (2001) (chromosome 19, Fig. 8). The mapping conducted in these two studies (Mian et al. 1996; Specht et al. 2001) for both the CID and WUE QTLs was undertaken with a comparatively limited marker set and the locations of the QTLs were inferred based on nearest markers to the base pair sequence location presented (Fig. 8). The actual QTL location in the genome may be at considerable distance from the location shown in Fig. 8.

The lack of overlap between putative loci identified by Specht et al. (2001) and those identified in this study was not surprising. Specht et al. conducted δ^{13}C analyses on juvenile trifoliates as opposed to whole-plant samples that were used in this study and examined a population based on parental lines that were not included in this study and that did not differ in δ^{13}C. As indicated by Specht et al. (2001), except for one, the major QTLs they identified coincided with maturity and/or determinacy QTLs. Thus, the identified QTLs may have been confounded by maturity and determinacy. Interestingly, even though the parental lines of the population studied by Mian et al. (1996) were not included in this study and growing conditions (greenhouse) and a phenotyping method (gravimetric determination of WUE) differed, one putative δ^{13}C locus identified in this study is located near WUE QTLs identified by Mian et al. This may indicate the stability and importance of these putative loci and highlight the genomic regions for further investigation.

Conclusions

Even with the large number of SNPs with MAF $\geq$ 5 % (12,347), there were still areas of the genome in which differences among the 373 genotypes were not detected. Nonetheless, population analysis was able to separate the 373 soybean genotypes into eight subgroups. No relationship between the subgroups and geographic origin, yield (as reported in the germplasm collection database) or δ^{13}C was apparent. GWAS analyses using GLM and MLM models with adjustments for genetic relatedness and/or population structure were conducted on δ^{13}C data obtained from independent field experiments in four environments. Analysis by environment and on the mean across all four environments identified SNPs putatively associated with $\delta^{13}C$. In total, 39 unique SNPs were detected in at least two environments or based on the means across environments. Although the SNPs detected in single environments may be associated with genes affecting $\delta^{13}C$, we have greater confidence in the SNPs identified independently in at least two environments. Additionally, many of the 39 identified SNPs were in close proximity to each other and likely tag the same locus. Overall, results indicated 21 putative loci associated with $\delta^{13}C$ with a high level of confidence. Although these results are conservative, they identified a tractable number of putative loci for further evaluation and confirmation in biparental mapping populations as well as potential use in breeding programs.

Author contributions JRS, JDR, LCP and FBF designed the study. SKS and VHV performed field experiments in Columbia and collected leaf tissues for $\delta^{13}C$ analysis. LCP and ACK managed field experiments in Stuttgart and collected leaf tissues for $\delta^{13}C$ analysis. PBC and QS designed SoySNP50K iSelect SNP Beadchip for genotyping. APD performed SNP–trait association including other statistical analysis and co-wrote the manuscript with JDR. FBF coordinated and supervised the project. JRS, JDR, LCP and FBF critically revised the manuscript. All authors read and approved the final manuscript.

Acknowledgments This work was supported by United States Department of Agriculture–Agriculture Research Service (USDA-ARS) project number 6402-21220-010-00D and United Soybean Board project numbers 9274 and 1274. We appreciate the assistance of Dr. Randall Nelson, curator of the USDA-ARS Germplasm Collection in selecting the genotypes evaluated in this study.

Conflict of interest The authors have declared that no competing or conflicts of interest exist.

References

Abdurakhmonov I, Abdukarimov A (2008) Application of Association Mapping to Understanding the Genetic Diversity of Plant Germplasm Resources. Int J Plant Genomics. doi:10.1155/2008/574927

Ahmed IM, Dai H, Zheng W, Cao F, Zhang G, Sun D, Wu F (2013) Genotypic differences in physiological characteristics in the tolerance to drought and salinity combined stress between Tibetan wild and cultivated barley. Plant Physiol Biochem 63:49–60

Anderson MJ, Ter Braak CJF (2003) Permutations tests for multifactorial analysis of variance. J Stat Comput Simul 73:85–113

Angus J, van Herwarden AF (2001) Increasing water use and water use efficiency in dryland wheat. Agron J 93:290–298

Aranzana MJ, Abbassi EK, Howad W, Arus P (2010) Genetic variation, population structure and linkage disequilibrium in peach-commercial varieties. Genetics 11:69

Araus JL, Slafer GA, Reynolds MP, Royo C (2002) Plant breeding and drought in C3 cereals: what should we breed for? Ann Bot 89:925–940

Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21(2):263–265

Baum M, Von Korff M, Guo P, Lakew B, Hamwieh A, Lababidi S, Udupa SM, Sayed H, Choumane W, Grando S, Ceccarelli S (2007) Molecular approaches and breeding strategies for drought tolerance in barley. In: Varshney R, Tuberosa R Genomics-Assisted Crop Improvement, Volume 2: Genomics Applications in Crops. Springer, Dordrecht, pp 51–79

Belamkar V, Selvaraj MG, Ayers JL, Payton PR, Puppala N, Burow MD (2011) A first insight into population structure and linkage disequilibrium in the US peanut minicore collection. Genetica 139:411–429

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc Ser B (Methodol) 57(1):289–300

Blum A (2005) Drought resistance, water-use efficiency, and yield potential—are they compatible, dissonant, or mutually exclusive? Aust J Agric Res 56:1159–1168

Bondari K (2003) Statistical analysis of genotype x environment interaction in agricultural research. Paper SD15, SESUG: The Proceedings of the SouthEast SAS Users Group, St Pete Beach

Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32:314–331

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

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

Brüggemann N, Gessler A, Kayler Z, Keel SG, Badeck F, Barthel M, Boeckx P, Buchmann N, Brugnoli E, Esperschütz J, Gavrichkova O, Ghashghaie J, Gomez-Casanovas N, Keitel C, Knohl A, Kuptz D, Palacio S, Salmon Y, Uchida Y, Bahn M (2011) Carbon allocation and carbon isotope fluxes in the plant-soil-atmosphere continuum a review. Biogeosciences 8:3457–3489

Buckler E, Casstevens T, Bradbury P, Zhang Z (2009) Analysis byaS-Sociation, evolution and linkage (TASSEL) version 2.1. user manual. Cornell University, Ithaca

Bush WS, Moore JH (2012) Chapter 11: Genome-wide association studies. PLoS Computional Biol 8: e1002822

Cao K, Wang L, Zhu G, Fang W, Chen C, Luo J (2012) Genetic diversity, linkage disequilibrium, and association mappinganalyses of peach (Prunus persica) landraces in China. Tree Genetics Genomes 8:975–990

Cattivelli L, Rizza F, Badeck FW, Mazzucotelli E, Mastrangelo AM, Francia E, Marè C, Tondelli A, Stanca AM (2008) Drought tolerance improvement in crop plants: an integrated view from breeding to genomics. Field Crops Res 105:1–14

Chen H, He H, Zou Y, Chen W, Yu R, Liu X, Yang Y, Gao YM, Xu JL, Fan LM, Li Y, Li ZK, Deng XW (2011) Development and application of a set of breeder-friendly SNP markers for genetic analyses and molecular breeding of rice (Oryza sativa L.). Theori App Genet 123:869–879

Chen J, Chang SX, Anyia AO (2012) Quantitative trait loci for water-use efficiency in barley (Hordeum vulgare L.) measured by carbon isotope discrimination under rain-fed conditions on the Canadian Prairies. Theor Appl Genet 125:71–90

Chevenet F, Brun C, Banuls AL, Jacq B, Christen R (2006) TreeDyn: towards dynamic graphics and annotations for analyses of trees. BMC Bioinformatics 7:439

Condon AG, Richards RA (1992) Broad sense heritability and genotype environment interaction for carbon isotope discrimination in field-grown wheat. Aust J Agric Res 43:921–934

Condon AG, Farquhar GD, Richards RA (1990) Genotypic variation in carbon isotope discrimination and transpiration efficiency in wheat leaf gas exchange and whole plant studies. Aust J Plant Physiol 17:9–22

Condon AG, Richards RA, Farquhar GD (2002) Improving intrinsic water-use efficiency and crop yield. Crop Sci 42:122–131

Condon AG, Farquhar GD, Rebetzke GJ, Richards RA (2004) Breeding for high water-use efficiency. J Exp Bot 55:2447–2460

Dhanapal AP, Crisosto CH (2013) Association genetics of chilling injury susceptibility in peach (Prunus persica (L.) Batsch) across multiple years. 3 Biotech 3:481–490

Edae EA, Byrne PF, Haley SD, Lopes MS, Reynolds MP (2014) Genome-wide association mapping of yield and yield components of spring wheat under contrasting moisture regimes. Theor Appl Genetics 127(4):791–807

Ehdaie B, Hall AE, Farquhar GD, Nguyen HT, Waines JG (1991) Water-use efficiency and carbon isotope discrimination in wheat. Crop Sci 31:1282–1288

Ehleringer JR, Klassen S, Clayton C, Sherrill D, Fuller-Holbrook M, Fu Q, Cooper T (1991) Carbon isotope discrimination and transpiration efficiency in common bean. Crop Sci 31:1611–1615

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

Farquhar GD, Richards RA (1984) Isotopic composition of plant carbon correlates with water-use efficiency of wheat genotypes. Aust J Plant Physiol 11:539–552

Farquhar GD, O'Leary MH, Berry JA (1982) On the relationship between carbon isotope discrimination and the inter-cellular carbon-dioxide concentration in leaves. Aus J Plant Physiol 9: 121–137

Farquhar GD, Ehleringer JR, Hubick KT (1989) Carbon isotope discrimination and photosynthesis. Ann Rev Plant Physiol 40:503–538

Fehr WR, Caviness CE, Burmood DT, Pennington JS (1971) Stage of development descriptions for soybeans, Glycine max (L.) Merr. Crop Sci 11:929–931

Geber MA, Dawson TE (1997) Genetic variation in stomatal and biochemical limitations to photosynthesis in the annual plant Polygonum arenastrum. Oecologia 109:535–546

Gilbert ME, Zwieniecki MA, Holbrook NM (2011) Independent variation in photosynthetic capacity and stomatal conductance leads to differences in intrinsic water use efficiency in 11 soybean genotypes before and during mild drought. J Exp Bot 62:2875–2887

Grant D, Nelson RT, Cannon SC (2013) SoyBase, the USDA-ARS genetics and genomics database. [WWW document] http://soybase.org

Gu J, Yin X, Struik PC, Stomph TJ, Wang H (2012) Using chromosome introgression lines to map quantitative trait loci for photosynthesis parameters in rice (Oryza sativa L.) leaves under drought and well-watered field conditions. J Exp Bot 63:455–469

Guo J, Liu Y, Wang Y, Chen J, Li Y, Huang H, Qiu L, Wang Y (2012) Population structure of the wild soybean (Glycine soja) in China: implications from microsatellite analyses. Ann Bot 110:777–785

Hall AE, Mutters RG, Hubick KT, Farquhar GD (1990) Genotypic differences in carbon isotope discrimination by cowpea under wet and dry field conditions. Crop Sci 30:300–305

Hao D, Cheng H, Yin Z, Cui S, Zhang D, Wang H, Yu D (2012) Identification of single nucleotide polymorphisms and haplotypes associated with yield and yield components in soybean (Glycine max) landraces across multiple environments. Theor Appl Genet 124:447–458

Hervé D, Fabre F, Berrios EF, Leroux N, Al Chaarani G, Planchon C, Sarrafi A, Gentzbittel L (2001) QTL analysis of photosynthesis and water status traits in sunflower (Helianthus annuus L.) under greenhouse conditions. J Exp Bot 52:1857–1864

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

Hwang EY, Song Q, Jia G, Specht JE, Hyten DL, Costa J, Cregan PB (2014) A genome-wide association study of seed protein and oil content in soybean. BMC Genomics 15(1)

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

Ismail A, Hall A (1992) Correlation between water-use efficiency and carbon isotope discrimination in diverse cowpea genotypes and isogenic lines. Crop Sci 32:7–12

Johnson RC (1993) Carbon isotope discrimination, water relations, and photosynthesis in Tall Fescues. Crop Sci 33:169–174

Juenger TE, McKay JK, Hausmann N, Keurentjes J, Sen S, Stowe KA, Dawson TE, Simms EL, Richards JH (2005) Identification and characterization of QTL underlying whole-plant physiology in Arabidopsis thaliana: delta C-13, stomatal conductance and transpiration efficiency. Plant Cell Environ 28:697–708

Kump KL, Bradbury PJ, Wisser RJ, Buckler ES, Belcher AR, Oropeza-Rosas MA, Zwonitzer JC, Kresovich S, McMullen MD, Ware D (2011) Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. Nat Genet 43(2):163–168

Laza MR, Kondo M, Ideta O, Barlaan E, Imbe T (2006) Identification of quantitative trait loci from d13C and productivity in irrigated lowland rice. Crop Sci 46:763–773

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

Littell RC, Milliken GA, Stroup WW, Wolfinger RD (1996) SAS system for mixed models. SAS Institute Inc, Cary

Liu K, Muse SV (2005) Power Marker: integrated analysis environment for genetic marker data. Bioinformatics 21:2128–2129

Louette D, Charrier A, Berthaud J (1997) In situ conservation of maize in Mexico: genetic diversity and maize seed management in a traditional community. Econ Bot 51:20–38

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

Mano Y, Muraki M, Fujimori M, Takamizo T, Kindiger B (2005) Identification of QTL controlling adventitious root formation during flooding conditions in teosinte (Zea mays ssp. huehuetenangensis) seedlings. Euphytica 142:33–42

Mian MAR, Bailey MA, Ashley DA, Wells R, Carter TE, Parrott WA, Boerma HR (1996) A Molecular markers associated with water use efficiency and leaf ash in soybean. Crop Sci 36:1252–1257

Mian MAR, Ashley DA, Boerma HR (1998) An additional QTL for water use efficiency in soybean. Crop Sci 38:390–393

O'Leary MH (1981) Carbon isotope fractionation in plants. Phytochemistry 20:553–567

Pasam RK, Sharma R, Malosetti M, van Eeuwijk FA, Haseneyer G, Kilian B, Graner A (2012) Genome-wide association studies for agronomical traits in a world wide spring barley collection. BMC Plant Biol 12:16

Passioura JB (1977) Grain-yield, harvest index, and water-use of wheat. J Aust Inst Agric Sci 43:117–120

Passioura JB (2004) Water-use efficiency in farmers' fields. In: Bacon M (ed) Water-use efficiency in plant biology. Blackwell, Oxford, pp 302–321

Peakall R, Smouse PE, GENALEX (2006) Genetic analyses in Excel. population genetic software for teaching and research. Mol Ecol Notes 6:288–295

Piepho HP, Möhring J (2007) Computing heritability and selection response From unbalanced plant breeding trials. Genetics 177:1881–1888

Piepho HP, Möhring J, Melchinger AE, Büchse A (2008) BLUP for phenotypic selection in plant breeding and variety testing. Euphytica 161:209–228

Pinto RS, Reynolds MP, Mathews KL, McIntyre CL, Olivares- Villegas JJ, Chapman SC (2010) Heat and drought adaptive QTL in a wheat population designed to minimize confounding agronomic effects. Theor Appl Genet 121:1001–1021

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

R Development Core Team (2013) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN:3-900051-07-0, http://www.R-project.org

Ray JD, Kilen TC, Abel C, Paris RL (2003) Soybean natural cross-pollination rates under field conditions. Environ Biosaf Res 2:133–138

Rebetzke GJ, Condon AG, Richards RA, Farquhar GD (2002) Selection for reduced carbon isotope discrimination increases aerial biomass and grain yield on rainfed bread wheat. Crop Sci 42:739–745

Rebetzke GJ, Condon AG, Farquhar GD, Appels R, Richards RA (2008) Quantitative trait loci for carbon isotope discrimination are repeatable across environments and wheat mapping populations. Theor Appl Genet 118:123–137

Reynolds M, Tuberosa R (2008) Translational research impacting on crop productivity in drought-prone environments. Curr Opin Plant Biol 11:171–179

Riedelsheimer C, Lisec J, Czedik-Eysenberg A, Sulpice R, Flis A, Grieder C, Altmann T, Stitt M, Willmitzer L, Melchinger AE (2012) Genome-wide association mapping of leaf metabolic profiles for dissecting complex traits in maize. Proc Natl Acad Sci USA 109:8872–8877

Salekdeh GH, Reynolds M, Bennett J, Boyer J (2009) Conceptual framework for drought phenotyping during molecular breeding. Trends Plant Sci 14:488–496

Saranga Y, Flash I, Paterson AH, Yakir D (1999) Carbon isotope ratio in cotton varies with growth stage and plant organ. Plant Sci 142:47–56

Saranga Y, Jiang CX, Wright RJ, Yakir D, Paterson AH (2004) Genetic dissection of cotton physiological responses to arid conditions and their inter-relationships with productivity. Plant Cell Environ 27:263–277

SAS-Institute-Inc (2004) SAS/STAT User's guide version 9.2 SAS-Institute-Inc., Cary, NC

Shiferaw E, Pè ME, Porceddu E, Ponnaiah M (2012) Exploring the genetic diversity of Ethiopian grass pea (Lathyrus sativus L.) using EST-SSR markers. Mol Breeding 30:789–797

Sinclair TR (2012) Is transpiration efficiency a viable plant trait in breeding for crop improvement ? Funct Plant Biol 39:359–365

Song Q, Hyten DL, Jia G, Quigley CV, Fickus EW, Nelson RL, Cregan PB (2013) Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. PLoS One 8:e54985

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

Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci USA 100(16):9440–9445

Taji T, Seki M, Satou M, Sakurai T, Kobayashi M, Ishiyama K, Narusaka Y, Narusaka M, Zhu J, Shinozaki K et al (2004) Comparative genomics in salt tolerance between Arabidopsis and Arabidopsis-related halophyte salt cress using Arabidopsis microarray. Plant Physiol 135:1697–1709

Takai T, Fukuta Y, Sugimoto A, Shiraiwa T, Horie T (2006) Mapping of QTLs controlling carbon isotope discrimination in the photosynthetic system using recombinant inbred lines derived from a cross between two different rice (Oryza sativa L.) cultivars. Plant Prod Sci 9:271–280

Teulat B, Merah O, Sirault X, Borries C, Waugh R, This D (2002) QTLs for grain carbon isotope discrimination in field-grown barley. Theor Appl Genet 106:118–126

Tuberosa R (2013) Phenotyping for drought tolerance of crops in the genomics era. Front Physiol 3:347

Tuberosa R, Salvi S (2006) Genomics-based approaches to improve drought tolerance of crops. Trends Plant Sci 8:405–412

Tuberosa R, Gill BS, Quarrie SA (2002) Cereal genomics: ushering in a brave new world. Plant Mol Biol 48:445–449

Wang ML, Sukumaran S, Barkley NA, Chen Z, Chen CY, Guo B, Pittman RN, Stalker HT, Holbrook CC, Pederson GA, Yu J (2011) Population structure and marker–trait association analysis of the US peanut (Arachis hypogaea L.) mini-core collection. Theor Appl Genet 123:1307–1317

Wingate L, Og´ee J, Burlett R, Bosc A, Devaux M, Grace J, Loustau D, Gessler A (2010) Photosynthetic carbon isotope discrimination and its relationship to the carbon isotope signals of stem, soil and ecosystem respiration. New Phytol 188:576–589

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

Yoneyama T, Fujiwara H, Engelaar WM (2000) Weather and nodule mediated variations in delta 13C and delta 15N values in field-grown soybean (Glycine max L.) with special interest in the analyses of xylem fluids. J Exp Bot 344:559–566

Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler E (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38:203–208

Zhang Z, Ersoz E, Lai C, Todhunter R, Tiwari H, Gore M, Bradbury P, Yu J, Arnett D, Ordovas J, Buckler E (2010) Mixed linear model approach adapted for genome-wide association studies. Nat Genet 42:355–360

Zhao K, Tung CW, Eizenga GC, Wright MH, Ali ML, Price AH, Norton GJ, Islam MR, Reynolds A, Mezey J, McClung AM, Bustamante CD, McCouch SR (2011) Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa. Nature Commun 2:467