Published November, 2011

ORIGINAL RESEARCH

Genome-Wide Association Analysis Identifies Candidate Genes Associated with Iron Deficiency Chlorosis in Soybean

Sujan Mamidi, Shireen Chikara, R. Jay Goos, David L. Hyten, Deepti Annam, Samira Mafi Moghaddam, Rian K. Lee, Perry B. Cregan, and Phillip E. McClean*

Abstract

Iron deficiency chlorosis (IDC) is a significant yield-limiting problem in several major soybean [*Glycine max* (L.) Merr.] production regions in the United States. Soybean plants display a variety of symptoms that range from a slight yellowing of the leaf to interveinal chlorosis, to stunted growth that reduces yield. The objective of this analysis was to employ single nucleotide polymorphism (SNP)-based genome-wide association mapping to uncover genomic regions associated with IDC tolerance. Two populations [2005 (n = 143) and 2006 (n = 141)] were evaluated in replicated, multilocation IDC trials. After controlling for population structure and individual relatedness, and selecting statistical models that minimized false positives, 42 and 88 loci, with minor allele frequency >10%, were significant in 2005 and 2006, respectively. The loci accounted for 74.5% of the phenotypic variation in IDC in2005 and 93.8% of the variation in 2006. Nine loci from seven genomic locations were significant in both years. These loci accounted for 43.7% of the variation in 2005 and 47.6% in 2006. A number of the loci discovered here mapped at or near previously discovered IDC quantitative trait loci (QTL). A total of 15 genes known to be involved in iron metabolism mapped in the vicinity (<500 kb) of significant markers in one or both populations.

Published in The Plant Genome 4:154–164. Published 19 Aug. 2011.
doi: 10.3835/plantgenome2011.04.0011

5585 Guilford Rd., Madison, WI 53711 USA
An open-access publication

THE DEMAND FOR SOYBEAN [*Glycine max* (L.) Merr.] has grown consistently in the United States over the last decades. This has resulted in the expansion of the growing regions north and west of the traditional region. This expansion has been into soils that differ from those on which the crop was historically bred. The characteristics of the soil have led to the appearance of iron deficiency chlorosis (IDC), an important yield-limiting factor for soybeans grown on calcareous soil. Calcareous soil, with a relatively high percentage of calcium carbonate and soluble salts, is commonly present in the north-central regions of the United States and extends from eastern North Dakota and South Dakota and into central Minnesota, central Iowa, and central Nebraska and Kansas (Franzen and Richardson, 2000). For many producers in these regions, IDC is considered a major yield limiting factor. In Iowa and Minnesota alone, IDC can render losses exceeding US$10 million due to decreased soybean production (Hansen et al., 2004).

Iron deficiency chlorosis results from the inability of some genotypes to efficiently mobilize iron into the plant when it is growing in high pH calcareous soils. In these soils, ferrous iron is not readily oxidized to ferric

S. Mamidi, S. Chikara, S.M. Moghaddam, R.K. Lee, and P.E. McClean, Genomics and Bioinformatics Program, North Dakota State Univ., Fargo, ND 58102; S. Mamidi, S. Chikara, S.M. Moghaddam, R.K. Lee, and P.E. McClean, Dep. of Plant Sciences, North Dakota State Univ., Fargo, ND 58102; R.J. Goos, School of Natural Resources, North Dakota State Univ., Fargo, ND 58102; D.L. Hyten and P.B. Cregan, USDA-ARS, Soybean Genomics and Improvement Lab., Beltsville, MD 20705; D. Annam, Dep. of Statistics, North Dakota State Univ., Fargo, ND 58102; Sujan Mamidi and Shireen Chikara contributed equally to this manuscript. Received 22 Apr. 2011. *Corresponding author (phillip.mcclean@ndsu.edu).

Abbreviations: AM, association mapping; IDC, iron deficiency chlorosis; LD, linkage disequilibrium; MAF, minor allele frequency; MSD, mean square deviation; PCA, principal component analysis; pFDR, positive false discovery rate; PIC, polymorphic information content; QTL, quantitative trait loci; SNP, single nucleotide polymorphism; SSR, simple sequence repeat.

iron, and, subsequently, iron availability is limited. Under such conditions, the concentration of iron in the soil is not higher than 100 pmol (Stephan, 2002). Based on its response to Fe availability, soybean is considered a Strategy I plant (Marschner et al., 1986). Strategy I plants first release H^+ ions from the root surface into the soil by the proton pumping activity of a H^+ adenosine triphosphatase (ATPase). This lowers the soil pH, which in turn initiates the dissociation of $Fe(OH)_3$ complexes into ferrous ions. Next, Fe^{3+} is reduced by Fe^{3+} chelate reductase to the more soluble Fe^{2+}. And third, iron transporters move the Fe^{2+} into the root. Strategy I plants also increase root hair formation, thereby increasing the surface area available for iron uptake (Schmidt, 1999). Once iron has entered the root, it is then moved via membrane transporters into the xylem where it most likely chelates with citrate. The chelated form of iron then moves through the xylem stream to growing leaves. Finally, iron is mobilized from the leaves, forms a complex with nicotianamine, and is transported via the phloem to younger leaves and seed.

Excess water is another factor that accentuates IDC in calcareous soils during the early stages of soybean development. This leads to an elevated concentration of bicarbonates in the root apoplast that impedes the Fe^{3+}–chelate reductase activity necessary for the conversion of Fe^{3+} to Fe^{2+}. Bicarbonates also immobilize the movement of iron to young leaves once it is absorbed at the root level (Barker and Pilbeam, 2007).

From a genetic perspective, IDC is clearly a quantitative trait where multiple genetic factors are involved in the expression of the proteins necessary for the uptake of iron from the soil and its distribution through the plant. Therefore it was not surprising that studies designed to understand the genetic nature of the IDC response in soybean identified multiple quantitative trait loci (QTL) (Diers et al., 1992; Lin et al., 1997, 2000; Charlson et al., 2003, 2005). These original studies used biparental populations, and the QTL discovered using one population were often population specific (Diers et al., 1992; Charlson et al., 2005). From an applied plant breeding perspective this minimizes the effectiveness of the biparental marker approach. At the same time, these studies reinforce the observation that IDC is complex trait.

Association mapping (AM) is an alternative to discovering genetic factors using biparental crosses. Association mapping uses the linkage disequilibrium (LD) pattern in a large population of unrelated individuals (Risch, 2000). As such, it can identify common genetic variants that control a common phenotype. Given its complex nature, utilizing AM to discover major factors controlling the IDC response in soybean seems to be an appropriate research direction. Indeed, an early AM study in soybean using a limited number of simple sequence repeat (SSR) markers discovered two markers that were reproducibly associated with IDC in two independent populations (Wang et al., 2008). Here a genome-wide discovery effort using the universal 1536 soybean Golden Gate single nucleotide polymorphism (SNP) set (Hyten et al., 2010) with two independent populations was undertaken. Multiple genomic regions distributed throughout the soybean genome were discovered to be associated with the IDC phenotype. A number of genes known to be involved in iron metabolism were found to be closely linked to these loci. The application of regional testing trials for AM studies is also considered.

Materials and Methods

Populations, Phenotyping, and Genotyping

Two independent populations, each consisting of a unique set of advanced soybean breeding lines, developed by public and private breeding programs for 0 and early I maturity groups for the north central states of the United States, were evaluated in 2005 (n = 143) and 2006 (n = 141). The 2005 population was grown at five sites near Arthur, Ayr, Chaffee, Colfax, and Galesburg, ND. The soil at these sites had a pH varying from 7.8 to 8.1, salinity (electrical conductivity) from 0.4 to 0.2 S m^{-1}, and $CaCO_3$ contents ranged from 2 to 11%. Thirty-five seeds were planted in 1.53 m rows on 76.2 cm centers. The experimental design was a randomized complete block design with four replications at each site. Two visual observations were made at each location at the two to three and five to six trifoliolate stages. The second independent population was grown in 2006 at Arthur, Colfax, Galesburg, and Prosper, ND. The soil at these sites had pH varying from 8.1 to 8.3, salinity (electrical conductivity) from 0.02 to 0.08 S m^{-1} and $CaCO_3$ contents ranged from 2 to 8%. The experimental design and the IDC rating scales were the same as for the year 2005. The visual observations were made at two to three trifoliolate and five to six trifoliolate stages and also 2 wk later. Only two observations could be made at Prosper due to recovery of the plants from chlorosis. Ten standard varieties, listed in descending order of IDC tolerance, were included. These are 'ISU A11', 'Seeds 2000 2070', 'Traill', 'Council', 'Asgrow 0801', 'Peterson PFS 0202', 'Glacier', 'Mycogen 5072', 'Stine 0480', and 'NuTech 0505'. Iron deficiency chlorosis was ratedon a 1 to 5 scale where 1 indicates no chlorosis and a normal green plant, 2 is used when there is a slight yellowing of upper leaves and the leaf veins and interveinal area do not show a differentiation in the color, 3 shows an interveinal chlorosis in the upper leaves while no obvious stunning of growth or death of tissue (necrosis) could be observed, 4 is used when interveinal chlorosis of the upper leaves is observed along with some apparent stunning of growth or necrosis of tissue, and 5 points severe chlorosis plus stunned growth and necrosis in the youngest leaves and growing points (Wang et al., 2008). All lines were grown in the greenhouse, young leaves were harvested and stored at –80°C, and DNA was extracted using the procedure of Brady et al. (1998). Each sample was genotyped using the Illumina GoldenGate SNP assay with the Universal Soy Linkage Panel (USLP) 1.0 (Hyten et al., 2010).

Statistical Analysis

Imputation

fastPHASE 1.3 (Scheet and Stephens, 2006) was used to impute missing data for these two sets of loci using "likelihood" based imputation. The default settings were used. Of the 1265 polymorphic markers, 858 and 868 loci, for the 2005 and 2006 populations, respectively, with minor allele frequencies (MAFs) >10%, were selected for analysis. Of these, 816 markers were common to the two populations. The polymorphic information content (PIC) was estimated separately for each population using the PowerMarker software (Liu and Muse, 2005).

Pairwise Linkage Disequilibrium and Linkage Disequilibrium Decay

The extent of LD was estimated as the squared allele frequency correlation (R^2) for each populations using TASSEL v.2.1 (Bradbury et al., 2007). Linkage disequilibrium decay graphs were plotted with genetic (cM) or physical distance (Mbp) vs. R^2 for each marker pair locus located on the same chromosome using nonlinear regression as described by Remington et al. (2001). The expected decay of LD was estimated according to the following equation:

$$E(R^2) = [(10 + pd)/(2 + pd)(11 + pd)](1 + \{(3 + pd)[12 + 12pd + p^2d^2]\}/[n(2 + pd)(11 + pd)])$$

The above equation was described by Pyhajarvi et al. (2007) where n denotes the number of sequences, $p = 4N_e c$ between adjacent sites, d is the distance between the two sites of a pairwise comparison, and c is the recombination rate (Hill and Weir 1988). We fit this equation into a nonlinear regression model using NLIN procedures in SAS v. 9.2 (SAS Institute, 2002). The analyses were also performed for individual chromosomes in both the populations.

Population Structure and Kinship

Estimation of population structure (Q) and kinship relationships were derived using only loci that had pairwise R^2 values < 0.5 for all possible combinations. In the 2005 population, 306 marker loci met this criterion, while in 2006 population, the number was 303. Population structure was first characterized using STRUCTURE.2.3 to estimate subpopulation membership of each line in these two populations individually (Pritchard et al., 2000). The admixture model with correlated allele frequencies was used with a burn-in of 100,000 and 500,000 iterations for subpopulations numbers ranging from 1 to 15. Five runs for each K value were performed, and the posterior probability was determined for each run. The optimum number of subpopulations was determined by the Wilcoxon two sample t test as described by Rosenberg et al. (2001) by comparing the posterior probability for successive adjacent subpopulations numbers (K2 vs. K3, K3 vs. K4, and so on) using the NPAR1WAY procedure in SAS (SAS Institute, 2002). The smaller K value in a pairwise comparison for the first nonsignificant Wilcoxson test was chosen as the best number of subpopulations. Principal component analysis (PCA) was also used to control for population structure in the two populations individually. The PCA was performed using the PRINCOMP procedure in SAS. The number of principal components (eigenvectors per combination of SNP markers) that collectively explained 25% of the variation was selected for the analysis.

A pairwise kinship coefficient matrix (**K**-matrix) that estimates the probability of recent coancestory between genotypes (Loiselle et al., 1995) was determined using SPAGeDi 1.2 (Hardy and Vekemans, 2002). The appropriate formula is:

$$F_{ij} = (Q_{ij} - Q_m)/(1 - Q_m) \approx r_{ij},$$

where r_{ij} is the pairwise kinship coefficient, F_{ij} is an estimator of the coefficient, Q_{ij} is the probability of the identity by state between random loci for genotypes i and j, and Q_m is the average probability of identity by state for loci from random genotypes in the population used to draw i and j. The F_{ij} was calculated for all pairwise combinations in each of the two populations. Negative values for the kinship matrix was set to zero as described by Yu et al. (2006). PowerMarker software (Liu and Muse, 2005) used to estimate a second kinship coefficient matrix $\mathbf{K}^*$ (Zhao et al., 2007) that represents the proportion of shared alleles for all pairwise comparisons in each population.

Marker-Trait Association Model Testing

The IDC phenotypic data were analyzed as an adjusted entry mean using the statistical model:

$$y_{ijk} = u + g_i + l_j + r_{jk} + (gl)_{if} + \varepsilon_{ijk},$$

where y_{ijk} was the mean of the two IDC ratings for the ith genotype in the kth replication at jth location, u is an intercept term, g_i was the genetic effect of the ith genotype, l_j is the effect of jth location, r_{jk} was the effect of kth replicate at jth location, $(gl)_{if}$ was the effect of genotype × environment interaction, and ε_{ijk} was the residual. For the adjusted entry means calculation, g_i is considered to be a fixed effect. Over all locations and replicates, an adjusted entry means (M_i) was calculated for each genotype as $M_i = u' + g'_i$, where u' and g'_i denote the generalized least square estimates of u and g_i, respectively. This model is essentially the same as that used by Stitch et al. (2008).

Nine different linear regression models were tested for marker-trait association using the MIXED procedure in SAS (SAS Institute, 2002) (Table 1). Six mixed-linear models (MLMs) considered both fixed and random effects while the remaining three general linear models (GLMs) considered only the fixed effects. In these models, $\mathbf{y}$ is a vector for phenotypic observations, α is the fixed effects related to the SNP marker, β is a vector of the fixed effects related to the population structure, ν is a vector of the random effects related to the relatedness among the individuals, and ε is a vector of the residual effects. X is genotypes of the SNP markers, $\mathbf{P}$ is the matrix of the principle components, $\mathbf{K}$ is the Loiselle

Table 1. Summary of the statistical models used to test for marker-trait associations.

Model	Statistical model	Information captured in the model
Naïve	$y = X\alpha + \varepsilon$	y is related to X, without correction for structure (Q or PCA†) or relatedness (K or K^*)
K	$y = X\alpha + K\nu + \varepsilon$	y is related to X, with correction for K
K^*	$y = X\alpha + K^*\nu + \varepsilon$	y is related to X, with correction K^*
Q	$y = X\alpha + Q\beta + \varepsilon$	y is related to X, with correction for Q‡
PCA	$y = X\alpha + P\beta + \varepsilon$	y is related to X, with correction for PCA§
Q + K	$y = X\alpha + Q\beta + K\nu + \varepsilon$	y is related to X, with correction for Q‡ and K
Q + K^*	$y = X\alpha + Q\beta + K^*\nu + \varepsilon$	y is related to X, with correction for Q‡ and K^*
PCA + K	$y = X\alpha + P\beta + K\nu + \varepsilon$	y is related to X, along with correction for PCA§ and K
PCA + K^*	$y = X\alpha + P\beta + K^*\nu + \varepsilon$	y is related to X, along with correction for PCA§ and K^*

†PCA, principal component analysis.

‡Q is seven and three subpopulations for 2005 and 2006, respectively.

§Principal components (PCs) that explain ~25% variance are four in both the years.

kinship matrix, and $\mathbf{K}^*$ is the shared allele kinship matrix developed in PowerMarker (Liu and Muse, 2005). The variances of the random effects were estimated as $\text{Var}(u) = 2\mathbf{K}V_g$ and $\text{Var}(e) = \mathbf{I}V_R$, where $\mathbf{K}$ is a kinship matrix, $\mathbf{I}$ is an identity matrix with the off-diagonal elements recorded as 0 and diagonal elements is the reciprocal of the number of the observations for which the phenotypic data were obtained, V_g is the genetic variance, and V_R is the residual variance. For each marker, the positive false discovery rate (pFDR) was estimated using the PROC MULTTEST in SAS to correct for multiple marker trait association. For each model, all marker p-values were ranked from smallest to largest, and the mean square deviation (MSD) was calculated as:

$$\text{MSD} = \{\textstyle\sum_{i=1}^{n}[p_i - (i/n)^2]\}/n,$$

where i is the rank number, p_i is the probability of the ith ranked p-value, and n is the number of markers. Significant markers were selected only from the model determined to have the lowest MSD value for each year. A marker was considered to be repeatable if it was significant based on the pFDR test in each year. The multiple R^2 value for all loci were calculated using stepwise regression with forward selection using the PROC REG function in SAS. To detect the epistatic interactions between markers we used a general linear model and the p-value of interaction was used to test the significance.

Blast Analysis

All *Arabidopsis thaliana* (L.) Heynh. proteins genes demonstrated to be involved in iron metabolism (as summarized in Morrissey and Guerinot, 2009) were used as a query in a blastp analysis against the proteins defined in the 1.01 annotation of the soybean genome (Joint Genome Institute, 2010). Our search was limited the top 20 hits with an E-value cut off of 10^{-20}.

Results

Phenotypic Analysis for Iron Deficiency Chlorosis Scores for the Two Soybean Populations

Since IDC is an important yield limiting factor in soybean, the populations were evaluated in production fields where IDC was consistently noted in the past. The visual IDC scores for the 2005 population ranged from 1.5 to 3.8 with an average of 2.9, while the scores for the 2006 population ranged from 1.6 to 3.8 with an average of 2.7. Analysis of variance for the IDC scores from the 2005 and the 2006 populations showed significant genotype and location effects as well as a significant line × location interaction effect (Table 2). This further substantiates that both genetic and environmental factors influence the IDC response in soybean. Broad-sense heritability on an entry mean basis was also deduced from the ANOVA. The broad-sense heritability values were 0.99 for 2005 population and 0.97 for 2006 population. These values demonstrate the consistency of the IDC rating. The distribution of IDC scores for the two populations (Fig. 1A and 1B) was determined to be normal using the Kolmogorov-Smirnov test with p-values of 0.15 and 0.18, respectively, for the 2005 and the 2006 populations.

Single Nucleotide Polymorphism Marker Analysis

Single nucleotide polymorphism marker information was collected in each population at 1265 informative loci with the Universal Soy Linkage SNP Panel 1.0 (Hyten et al., 2010) using the IlluminaGoldenGate Assay technology. Of the 1265 SNP marker loci, 858 markers in the 2005 population and 868 markers in the 2006 population had a MAF > 10%. The Wilcoxon two-sample test was not significant ($p = 0.3439$) for the comparison of the major allele frequency in the two populations. The expected heterozygosity is generally low for SNP markers because of their biallelic nature and the selfing nature of *G. max*. Gene diversity for the 2005 genotypes ranged from 0.19 to 0.50 with an average of 0.39 and for the 2006 genotypes it ranged from 0.19 to 0.50 with an average of 0.39. The markers in both populations were polymorphic with PIC values ranging from 0.17 to 0.38 for 2005 population and 0.17 to 0.38 for 2006 population.

Linkage Disequilibrium Decay, Population Structure, and Kinship Analysis

A nonlinear regression model that estimates the decay of LD with distance was developed. Using a pairwise analysis for all 858 and 868 SNP loci, respectively, in the 2005 and 2006 populations, R^2 values were regressed on the physical distances (Fig. 2). The average decay of LD in terms of physical distance declined to $R^2 < 0.1$ at 7.0 Mbp (19.3 cM) and 5.9 Mbp (19.7 cM) in 2005 and 2006, respectively.

From the R^2 data, 306 markers from the 2005 population (367,653 SNP marker-pair comparisons) and 303 markers (376,278 SNP marker-pair comparisons) from the 2006 population had $R^2 < 0.5$ among all pairwise

comparisons. These two marker subsets were then used to decipher population structure and kinship. Population structure was estimated with the software program, STRUCTURE, using the admixture model for the multilocus genotype data (Pritchard et al., 2000) and the subpopulation number selection criteria of Rosenberg et al. (2001). This analysis determined the 2005 population consisted of seven subpopulations, while the 2006 population comprised three subpopulations. Principal component analysis was also implemented to evaluate population structure. In 2005, 28% of the variance was explained by four principal components, where 11, 6.5, 5.2, and 4.6% are the variance explained by the first to fourth components respectively. In 2006, 29% of the variance was explained by four principal components where 10, 7.6, 6.6, and 4.8% are the variance explained by the first to fourth components, respectively.

Single Nucleotide Polymorphism and Iron Deficiency Chlorosis Marker–Trait Associations

Because of the significant location, line, and location × line interaction effects, we used adjusted entry mean IDC scores in our statistical model analyses. Independent marker-trait associations were conducted for 858 and 868 markers, respectively, for the 2005 and the 2006 populations. These numbers of markers represent the ~70% of the markers with a MAF > 0.10 in each populations. The genotypic and IDC data were evaluated using nine different models described in Table 1. These models, described in Zhao et al. (2007), are often used in plant AM experiments. Model testing was performed to determine which had the least inflation of significant values at the $p < 0.05$ level (Table 3). For both years, a model that included population structure and kinship factors had the least percentage of p-values less than 0.05. The ideal model would exhibit a uniform distribution when cumulative

Table 2. Analysis of variance for iron deficiency chlorosis ratings for the two soybean populations grown at different locations.

	Population			
	2005		2006	
Source of variation	df	MS†	df	MS
Location	4	104.39***	3	133.49***
Line	142	3.76***	140	3.15***
Location × line	569	0.25***	420	0.33***
Replication × location	15	2.11	12	2.08
Error	2130	0.19	1680	0.17

***Significant difference $p \leq 0.001$.

†MS, mean square,

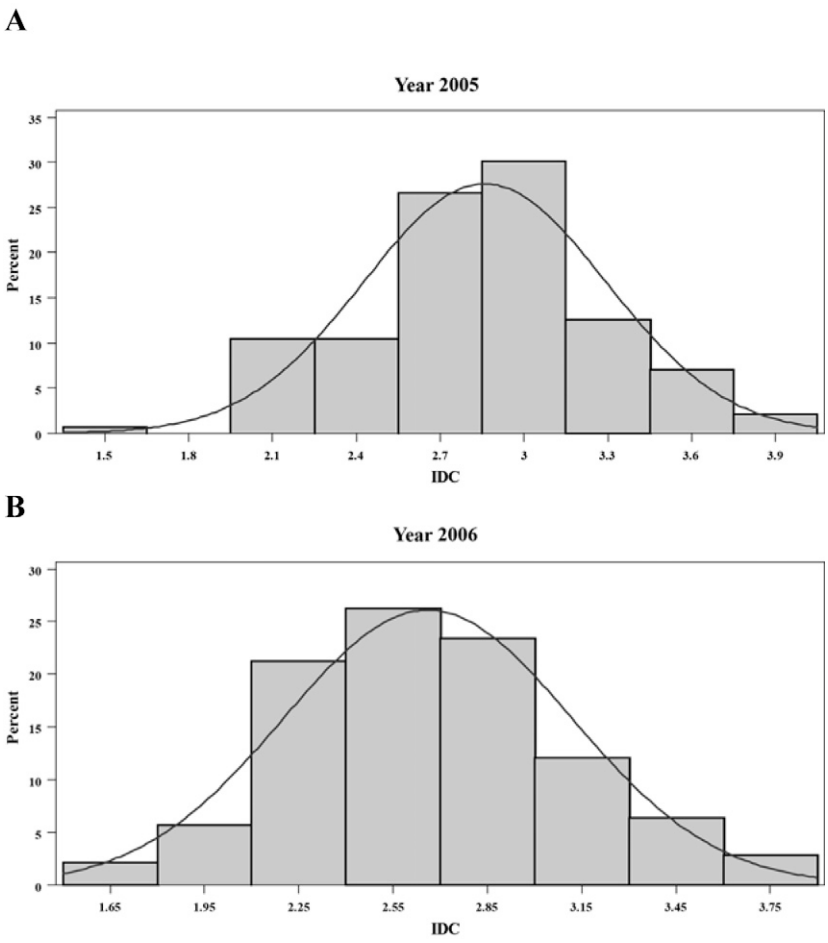

Figure 1. Phenotypic distribution of iron deficiency chlorosis (IDC) scores. (A) 2005 and (B) 2006.

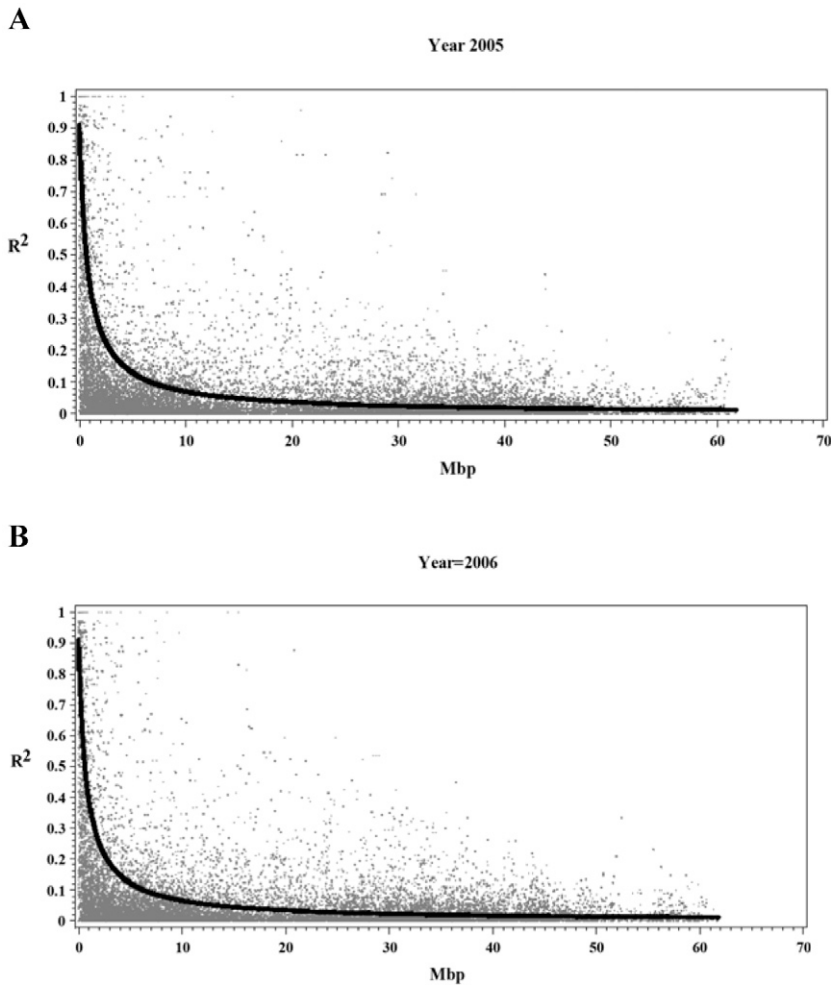

Figure 2. Genome-wide linkage disequilibrium (LD) decay plot for the two populations. Linkage disequilibrium, measured as R^2, between pairs of polymorphic marker loci is plotted against the physical distance (Mbp). (A) 2005 population. (B) 2006 population.

p-values are regressed on observed *p*-values. To observe the degree to which the statistical results for each model deviated from the expected distribution, we calculated the MSD for each model. Again, models that contained a structure component (Q in 2005 and PCA in 2006) and the shared allele kinship component had the lowest MSD values. These models were chosen to select significant SNP marker–trait associations.

Two independent soybean populations containing advanced breeding lines from public and private breeding programs were evaluated. Lines from 31 programs in 2005 and 30 programs in 2006 were included in the analysis. None of the lines contained the same SNP marker haplotype. We considered each population to be an independent reciprocal confirmation population for any significant markers discovered in the other population. A total of 42 loci met the pFDR < 0.1 criteria for the 2005 population. Of these, 28 fit into a stepwise regression with forward selection and explained 74.5% of the phenotypic variation. For the 2006 population, 88 met the criteria, and 70 of the loci explained 93.8% of the phenotypic variation. None of the markers was significant using the Bonferroni correction factor in 2005, while 13 were significant using this conservative correction factor in 2006. Those 13 accounted for 47.6% of the phenotypic variation.

Next, we used each population as a confirmation population for the other population and only selected those markers that were significant in both years and met the criteria of having a pFDR value < 0.10 in each year. Nine SNP markers, distributed over seven genomic regions on six chromosomes, met the significance criteria (Table 4). None of these nine exhibited epistatic effects in 2005, while several two and three epistatic effects were detected in 2006. Three consecutive markers located within a 408 kb (2.4 cM) window on chromosome Gm3 were significant. These three marker loci were in a high degree of LD ($R^2 > 0.70$ in each year). The MAF for most of the nine loci was >0.3 and for over half of the loci the value was >0.4 (Table 5). The trend was that the IDC means were larger in 2005 than 2006 for both the minor and major allele. This is also reflected in the phenotypic mean differences between the 2 yr. In 2005, the lower mean IDC score at any one marker was ~2.7. For this population, 40% of the entries had a mean less than this value. For the 2006 population, the trend was that the lower IDC was ~2.5, and the phenotypic rating of 44% of

Table 3. Test statistics for the nine models used to discover single nucleotide polymorphism (SNP) and iron deficiency chlorosis (IDC) tolerance marker–trait associations.

2005			2006		
Model	Percent *p*-values < 0.05	MSD†	Model	Percent *p*-values < 0.05	MSD
Q + K*	15.51	0.017	PCA‡ + K*	18.72	0.015
PCA	17.48	0.026	Q	23.39	0035
PCA + K	18.07	0.028	PCA	23.85	0.036
Q	20.86	0.032	Q + K	23.86	0.037
Q + K	22.03	0.034	PCA + K	24.45	0.038
Naïve	42.66	0.075	Naïve	38.59	0.067
K*	43.71	0.076	K	38.75	0.068
PCA + K*	50.59	0.203	K*	76.03	0.276
K	85.42	0.302	Q + K*	73.97	0.281

†MSD, mean square deviation.

‡PCA, principal component analysis.

Table 4. Single nucleotide polymorphism (SNP) marker loci significantly associated with iron deficiency chlorosis (IDC) tolerance in both 2005 and 2006 populations.

BARC SNP marker	Chromosome	SNP position (bp)	Genetic position (cM)	Minor allele	Major allele
BARC-029969-06762	2	2,454,206	18.801	A	C
BARC-044603-08734	3	45,007,967	85.822	T	A
BARC-060109-16388	3	45,391,018	86.907	A	G
BARC-016535-02085	3	45,416,307	88.225	G	A
BARC-010457-00640	6	45,281,692	108.504	T	A
BARC-039383-07310	7	7,151,246	39.939	C	A
BARC-025897-05144	13	27,145,239	49.424	A	G
BARC-055499-13329	13	31,472,325	61.354	G	A
BARC-059723-16418	19	40,357,687	56.404	A	G

the entries was equal to or less than this value. For nearly all of the loci, the R^2 value within each year was >10%. The only exception was BARC-010457-00640 located on Gm6. The multiple R^2 value was calculated for these nine makers. Only one Gm3 marker locus (BARC-044603-08734) was selected in each year to be included in the analysis. For the 2005 population, the multiple R^2 value was 43.7%, and for 2006 the value was 47.6%.

Discussion

Local or regional variety testing trials were established by public institutions to provide consistent data for a specific trait(s) of interest in support of both public and private plant breeding programs. Since these trials are often large in scale (>100 entries), they are ideal for adopting AM techniques because the collection of lines in those trials represents many rounds of recombination. Another major advantage of these trials is that they are performed by individuals highly trained for specific phenotypic evaluations, and therefore the phenotypic data generated on a year-to-year basis is often consistent. Furthermore, since the data are collected over multiple years, any two or more distinct populations can serve as a reciprocal confirmation population(s) for significant marker–trait associations in any one population. Given that the best designed trials are performed at multiple locations that often represent the environmental diversity for a specific agro-ecosystem, it is not surprising that environment × genotype interactions are often observed. Therefore, it is necessary, as we did here, to address this interaction effect by incorporating an adjusted means analysis step before searching for marker–trait associations (Stitch et al., 2008).

A major challenge for AM is to ensure any marker–trait associations are genetically significant and not the result of spurious associations due to population structure and/or relatedness. Regional trials often include multiple genotypes from multiple breeding programs, and as shown for barley (*Hordeum vulgare* L.), populations from multiple breeding programs can be structured based on the programs (Hamblin et al., 2010). To adjust for these potential confounding effects, it is now standard to evaluate genotypic and phenotypic data using several statistical models that account for population structure, genetic relatedness, coancestry based on pedigree, or some combination of these factors (Zhao et al., 2007; Stitch et al., 2008). Here we compared the results from nine different statistical models, because the effects of controlling complex structure (population structure, principal

Table 5. Test statistics for single nucleotide polymorphism (SNP) loci significantly associated with iron deficiency chlorosis (IDC) tolerance in the 2005 and 2006 populations.

		2005						2006					
BARC SNP marker	Chromosome	–log10(*p*)	pFDR†	R^2 (%)	Minor allele frequency	Minor allele mean	Major allele mean	–log10(*p*)	pFDR	R^2 (%)	Minor allele frequency	Minor allele mean	Major allele mean
BARC-029969-06762	2	2.707	0.046	11.8	49.0	3.0	2.7	7.531	0.000	24.3	40.4	2.9	2.5
BARC-044603-08734	3	3.661	0.022	14.9	40.6	2.7	3.0	4.898	0.001	15.7	45.4	2.5	2.8
BARC-060109-16388	3	3.321	0.024	16.3	46.2	2.7	3.0	3.781	0.003	13.9	47.5	2.5	2.8
BARC-016535-02085	3	2.886	0.046	15.3	46.9	2.7	3.0	3.781	0.003	13.9	47.5	2.5	2.8
BARC-010457-00640	6	2.185	0.076	0.1	39.9	2.8	2.9	4.018	0.002	2.6	44.7	2.6	2.7
BARC-039383-07310	7	3.114	0.034	17.8	21.7	2.5	3.0	4.548	0.002	13.5	21.3	2.3	2.7
BARC-025897-05144	13	2.309	0.067	15.2	38.5	3.1	2.7	4.008	0.002	15.9	37.6	2.9	2.5
BARC-055499-13329	13	2.420	0.058	9.4	31.5	3.1	2.8	2.602	0.021	13.4	27.0	2.9	2.6
BARC-059723-16418	19	3.415	0.022	15.4	44.8	2.7	3.0	4.124	0.002	10.7	31.2	2.4	2.8

†pFDR, positive false discovery rate.

components, and relative kinship matrices) varies with populations, traits, or both (reviewed in Sun et al., 2010) Given that most of the lines evaluated in the trial were from private programs, we did not have access to pedigree information that would have allowed us to include a coancestry factor. To select the appropriate model, the mean square deviation is an appropriate measure. The principle here is that the distribution of cumulative vs. observed *p*-values should approximate a uniform distribution. This implies that 1% of the marker–trait *p*-values should be less than 0.01, that 5% of the marker–trait should have a *p*-values should be less than 0.05, and so on. If a particular model fits this distribution, then the MSD should be small. In our case, in each year the MSD of a model with a structure component (PCA or Q) and the same shared allele kinship component (**K***) was found to be small and about 50% smaller than the second best model. Somewhat surprisingly, the naïve model that does not consider either structure or relatedness had a lower MSD value than several models that did consider these factors. Collectively, these results imply that model testing is necessary for whatever data set is under consideration, and a single model does not perform best in multiple years even when considering the same phenotype.

Iron metabolism in plants involves multiple genes that are associated with the acidification of the soil (using *A. thaliana* nomenclature, *AHA2*), iron reduction (*AtFRO2*), transport into the root (*AtIRT1*), sequestration of iron in the vacuole (*AtIREG2*), transport of iron carriers into the xylem (*FRD3*), distribution of chelated iron to the phloem (*YSL* transporter family), synthesis of the nicotianamine chelator (*NAS3*), carriers of iron into the seed (*ITP* carrier protein family), transport into the seed (*OPT* transporter family), transport into (*VIT1*) and out of (*NRAMP3* and *NRAMP4*) the vacuole, iron binding in the vacuole (*FER2*), and transport into the chloroplast (*FRO7* and *TIC21*) (Morrissey and Guerinot, 2009). In addition, transcription factors such as *FIT* regulate the expression of genes such as *IRT1* (Colangelo and Guerinot, 2004). Given the complex nature of iron metabolism in plants, we were expecting to observe multiple associations with our data. When we applied our cutoff criteria of pFDR < 0.10, we detected 42 significant marker–trait associations for the 2005 population and 88 marker–trait associations for the 2006 population. Given that we were using each population as a reciprocal confirmation population of any association, we next checked for those markers that met the significance criteria in both years. In this case, nine markers distributed over six chromosomes were found to be significant in each year. These loci accounted for ~45% of the variation in IDC ratings in each year. The availability of data from multiple years of a standard performance trial is an advantage for AM because it provides the necessary data and materials needed to confirm marker–trait associations. By using reciprocal confirmation populations, we are confident that some gene(s) in the vicinity of these markers are involved in the iron response of soybean grown on deficiency-inducing soils.

One of the goals of AM is to use the associations as points of departure to discover the actual genes involved in controlling the phenotype. The most extensive experiment to date was performed in *A. thaliana* where 107 traits were mapped using genotype data from a high density array (Atwell et al., 2010). For several of these associations, the peak SNP mapped at or very near the gene known to control a specific phenotype. Given that the density of markers in that experiment is much greater than in this experiment, we wished to determine if we could also discover any potential candidate genes.

The average distance between markers in this analysis was 976,723 bp. That value though includes markers found in the sparsely marked repetitive region of the genome. If we only consider markers that map in the euchromatic region of the genome, that distance is reduced to 525,640 bp. Since LD decay at $R^2 < 0.5$, a value where loci are still considered to be in LD, was ~600 kb in both years, it seemed reasonable to consider a search for candidate genes linked to significant markers within this interval distance. We performed a blastp analysis using *A. thaliana* proteins genes

Table 6. Significant (positive false discovery rate [pFDR] < 0.1) Universal Soy Linkage Panel (USLP) 1.0 markers that are the immediate neighbor to a gene known to be involved in iron metabolism. Markers without statistical information did not converge.

			2005			2006								
BARC marker	Chromosome	SNP† position (bp)	−log10 (*p*)	pFDR	R^2 (%)	−log10 (*p*)	pFDR	R^2 (%)	At‡ gene	Gm§ gene model	Start of model (bp)	Distance from SNP (bp)	E-value	Percent identity
BARC-060109-16388	3	45,391,018	3.32	0.02	16.26	3.78	0.00	13.94	*NAS3*	Glyma03g39050	45,279,921	111,097	$5.00\ E \times 10^{-109}$	62.9
BARC-053261-11776	5	937,302	2.21	0.08	0.72				*AHA2*	Glyma05g01460	960,820	23,518	0	81.1
BARC-021775-04203	5	41,114,078	0.08	0.53	0.07	2.47	0.03	9.30	*BTI2-ITP*	Glyma05g37300	40,906,083	207,995	$1.00\ E \times 10^{-95}$	65.7
BARC-054331-12480	7	8,652,831				1.72	0.09	4.78	*BTI2-ITP*	Glyma07g10870	9,082,076	429,245	$2.00\ E \times 10^{-45}$	36.8
BARC-049147-10810	9	35,895,343	0.44	0.38	0.73	2.07	0.05	1.50	*YSL7*	Glyma09g29410	36,298,317	402,974	0	54.6
BARC-062275-17736	11	38,020,165	0.02	0.56	1.64	2.63	0.02	0.18	*FER4*	Glyma11g35610	37,245,783	774,382	$7.00\ E \times 10^{-103}$	73.2
BARC-017917-02456	13	30,457,599	2.17	0.08	7.83				*FRD3*	Glyma13g27300	30,477,485	19,886	$2.00\ E \times 10^{-137}$	54.7
BARC-043041-08509	15	48,694,193	2.71	0.05	12.66	0.94	0.26	9.24	*IRT1*	Glyma15g41620	48,764,845	70,652	$8.00\ E \times 10^{-80}$	43.5
BARC-030595-06910	16	3,039,691	0.09	0.53	4.55	1.93	0.06	7.72	*FRO2*	Glyma16g03770	3,142,668	102,977	0	57.7
BARC-011625-00310	16	36,544,010	0.26	0.45	0.47	2.28	0.04	2.79	*YSL7*	Glyma16g33840	36,609,082	65,072	0	73.4
BARC-043087-08524	17	4,899,023	3.44	0.02	0.25	1.37	0.16	1.35	*AHA2*	Glyma17g06930	4,977,823	78,800	0	87.7
BARC-012289-01799	18	1,957,710				1.78	0.08	2.07	*FER4*	Glyma18g02800	1,821,344	136,366	$6.00\ E \times 10^{-102}$	75.1
BARC-016867-02359	18	56,429,447	1.26	0.19	2.77	2.11	0.05	6.63	*FRO2*	Glyma18g47060	56,712,622	283,175	0	52.2
BARC-059723-16418	19	40,357,687	3.42	0.02	15.41	4.12	0.00	10.68	*OPT1*	Glyma19g32400	40,140,993	216,694	$3.00\ E \times 10^{-146}$	47.4
BARC-042281-08231	20	343,106	0.22	0.47	2.33	1.96	0.06	8.58	*YSL7*	Glyma20g00690	418,225	75,119	0	62.8

†SNP, single nucleotide polymorphism.

‡At, *Arabidopsis thaliana* (L.) Heynh.

§Gm, *Glycine max* (L.) Merr.

demonstrated to be involved in iron metabolism (Morrissey and Guerinot, 2009) and queried the 1.01 annotation of the soybean genome (Joint Genome Institute, 2010). The results were limited to the top 20 hits with an E-value cut off of 10^{-20}. A total of 161 soybean gene models met these criteria with a median E-value of 0. We next considered only those genes that mapped within 500 kb of a significant marker. If a nonsignificant marker was located between the gene and the significant marker, that gene was excluded from further consideration as a candidate gene. A total of 15 genes met this criterion from the two populations.

We first evaluated only those iron metabolism genes linked to a marker that was significant in both years. Two gene models, Glyma03 g39050 (annotated as *NAS3*) and Glyma19 g32400 (*OPT1*) (Table 6) were identified. In *A. thaliana*, *NAS3* encodes nicotianamine synthase, an enzyme that synthesizes nicotianamine, a molecule that complexes Fe and carries it via the phloem to younger leaves and flowers. OPT1 is a protein that transports Fe into seeds. *NAS3* maps within the peak of three consecutive SNPs on Gm3. BARC-060109-16388, the SNP closest to the gene, showed the peak R^2 value while the two neighboring markers had lower R^2 values. These three markers exhibited high LD values each year ($R^2 > 0.7$) and presumably are signals for the same factor that affects IDC phenotypic expression. Since the marker linked closest to *NAS3* had the highest R^2 value in each year, it would appear *NAS3* is an important factor in IDC tolerance. *NAS3* encodes the enzyme that synthesizes the carrier that transports iron out of older leaves and via the phloem to younger leaves and flowers. Importantly, the IDC rating is made on these younger leaves. Therefore, from a physiological perspective *NAS3* would appear to be a candidate gene. This result is also consistent with previous genetic research with biparental populations that identified a major QTL on Gm3 (Lin et al., 1997, 2000). Finally, the gene maps in the QTL interval on Gm3 transferred into the soybean line Clark to develop the IDC susceptible line Iso Clark (Severin et al., 2010). Although this is compelling evidence for the role of *NAS3* in IDC tolerance, it does not preclude other genes within this region from affecting the IDC response.

We next evaluated those genes that mapped next to a significant marker in either year. This would allow us to potentially identify candidates that are unique to one of the two populations. Four unique candidates linked to significant markers ($p < 0.05$ and pFDR < 0.1) were discovered in the 2005 population, while eight were observed for the 2006 population. Most of these markers had a small effect ($R^2 < 5\%$). In 2005, the largest effect (12.7%) was noted for the marker near gene model Glyma15 g41620. This model is highly similar to *IRT1*, the gene that encodes the protein that transports iron into the root. A number of other iron metabolism genes were also observed. When all markers linked to genes involved in iron metabolism were included in a stepwise regression analysis, the R^2 value in 2005 was 37% while those markers linked in 2006 had a value of 40%.

Previous research identified IDC QTL on Gm3, Gm5, Gm12, Gm14, Gm18, Gm19, and Gm20 (Lin et al., 1997, 2000). We located these QTL on the 1.01 build of soybean (Schmutz et al., 2010) using the location of the reference SSR and restriction fragment length polymorphism (RFLP) sequences that were associated with the QTL. Those loci were compared to the results presented here. The best match was the Gm19 QTL designated by SSR Satt481. This SSR was evaluated for its utility for marker assisted selection and was found to be effective across locations to select superior IDC lines within a single breeding population. This SSR maps immediately adjacent to SNP BARC-059721-16418, a locus that was significant in both 2005 (R^2 = 15.4%) and 2006 (R^2 = 10.7%). It also maps near the *OPT1* candidate gene that is involved in movement of iron into seed. Two QTL were mapped on Gm3, but neither of these mapped at the association that centers around SNP BARC-060109-16388 at position 45,391,018 bp. Near-isogenic lines (Clark and Iso Clark) were developed that expressed different IDC ratings primarily because of different allelic states of this Gm3 QTL. This QTL was recently mapped by Severin et al. (2010) to a 9.8 Mbp interval (36.3–45.8 Mbp). This interval contains both the Satt481 SSR and BARC-060109-16388 SNP. It is quite possible that this QTL contains multiple genes that individually affect the IDC phenotype. Although none of the other previously identified IDC QTL mapped in both populations, several are located near a marker–trait association we discovered in 2006. Simple sequence repeat Satt211 on Gm5 (Charlson et al., 2003) maps near two significant SNPs (Supplemental Table S1). All of the markers are within 1.2 Mbp of an *ITP* (iron transporting protein) gene. Simple sequence repeat Satt181 maps near a Gm12 SNP (BARC-030421-06864).

Recently O'Rourke et al. (2009) evaluated the expression pattern of Clark and Iso Clark lines under iron sufficient and iron limiting conditions. Although these two lines primarily differ by the allelic state of a QTL on Gm3 (Severin et al., 2010), many genes were found to be differentially expressed in each line between the two iron nutrition conditions, but only a small number exceed the critical 2x difference in expression. Only a small portion (7–10%) of the differentially expressed genes mapped to previously defined IDC QTL. This result further supports our discovery that IDC is a complex response and poses a major challenge for molecular geneticists as they attempt to discern appropriate genetic signals that can be used to develop marker tools that will efficiently select superior IDC tolerant genotypes.

Conclusions

Here we describe the application of phenotypic and genotypic information collected from a regional nursery to study the genetic architecture of IDC in soybean. This trial provided two populations that served as reciprocal confirmation populations for the discovery of loci associated with this important agronomic trait using genome-wide AM techniques. After correcting for population structure and relatedness, multiple loci were discovered that collectively accounted for much of the variation in the phenotype. Using the extensive knowledge base of the biology and genetics of iron metabolism from the model organism *A. thaliana*, we were able to discover that a number of genes involved in this physiological network were closely linked to markers shown to be significantly associated with IDC. With these results, this ongoing regional IDC trial will continue to provide additional genetic material to further confirm and refine the results we obtained here.

Supplemental Information Available

Supplemental material is available free of charge at http://www.crops.org/publications/tpg.

Supplemental Table S1: Marker-by-marker association mapping results.

Acknowledgments

We thank Mohamed Mergoum for critically reviewing this manuscript. The project was supported by funds from the North Central Soybean Research Program.

References

Atwell, S., Y.S. Huang, B.J. Vilhjalmsson, G. Willems, M. Horton, Y. Li, D. Meng, A.A. Platt, M.A. Tarone, T.T. Hu, R. Jiang, N.W. Muliyati, X. Zhang, M.A. Amer, I. Baxter, B. Brachi, J. Chory, C. Dean, M. Debieu, J. de Meaux, J.R. Ecker, N. Faure, J.M. Kniskern, J.D. Jones, T. Michael, A. Nemri, F. Roux, D.E. Salt, C. Tang, M. Todesco, M.B. Traw, D. Weigel, P. Marjoram, J.O. Borevitz, J. Bergelson, and M. Nordborg. 2010. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature 465:627–631. doi:10.1038/nature08800

Barker, A.V., and D.J. Pilbeam (ed.). 2007. Handbook of plant nutrition. Vol. 117 ed. 1:335–337. Taylor & Francis, New York, Philadelphia, Oxford, Melbourne, Stockholm, Beijing, New Delhi, Johannesburg, Singapore and Tokyo.

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

Brady, L., M.J. Bassett, and P.E. McClean. 1998. Molecular markers associated with T and Z, two genes controlling partly colored seed coat patterns in common bean. Crop Sci. 38:1073–1075. doi:10.2135/cropsci1998.0011183X003800040031x

Charlson, D.V., T.B. Bailey, S.R. Cianzio, and R.C. Shoemaker. 2005. Molecular marker Satt481 is associated with iron deficiency chlorosis resistance in a soybean breeding population. Crop Sci. 45:2394–2399. doi:10.2135/cropsci2004.0510

Charlson, D.V., S.R. Cianzio, and R.C. Shoemaker. 2003. Associating SSR markers with soybean resistance to iron deficiency chlorosis. J. Plant Nutr. 26:2267–2276. doi:10.1081/PLN-120024280

Colangelo, E.P., and M.L. Guerinot. 2004. The essential basic helix–loop–helix protein FIT1 is required for the iron deficiency response. Plant Cell 16:3400–3412. doi:10.1105/tpc.104.024315

Diers, B.W., S.R. Cianzio, and R.C. Shoemaker. 1992. Possible identification of quantitative trait loci affecting iron efficiency in soybean. J. Plant Nutr. 15:2127–2136. doi:10.1080/01904169209364462

Franzen, D.W., and J.L. Richardson. 2000. Soil factors affecting iron chlorosis of soybean in the Red River Valley of North Dakota and Minnesota. J. Plant Nutr. 23:67–78. doi:10.1080/01904160009381998

Hamblin, M.T., T.C. Close, P.R. Bhat, S. Chao, J.G. Kling, K.J. Abraham, T. Blake, W.S. Brooks, B. Cooper, C.A. Griffey, P.M. Hayes, D.J. Hole, R.D. Horsley, D.E. Obert, K.P. Smith, S.E. Ullrich, G.J. Mucehlbauer, and J.-L. Jannink. 2010. Population structure and linkage disequilibrium in U.S. barley germplasm: Implications

for association mapping. Crop Sci. 50:556–566. doi:10.2135/cropsci2009.04.0198
Hansen, N.C., V.D. Jolley, S.L. Naeve, and R.J. Goos. 2004. Iron deficiency of soybean in the North Central U.S. and associated soil properties. Soil Sci. Plant Nutr. 50:983–987.
Hardy, Q.J., and X. Vekemans. 2002. SPAGeDi: A versatile computer program to analyse spatial genetic structure at the individual or population levels. Mol. Ecol. Res. 2:618–620.
Hill, W.S., and B.S. Weir. 1988. Variances and covariances of squared linkage disequilibria in finite populations. Theor. Popul. Biol. 33:54–78. doi:10.1016/0040-5809(88)90004-4
Hyten, D.L., I.-Y. Choi, Q. Song, J.E. Specht, T.E. Carter, R.C. Shoemaker, E.-Y. Hwang, L.K. Matukumall, and P.B. Cregan. 2010. A high density integrated genetic linkage map of soybean and the development of a 1536 universal soy linkage panel for quantitative trait locus mapping. Crop Sci. 50(3):960–968. doi:10.2135/cropsci2009.06.0360
Joint Genome Institute. 2010. Phytozome. Available at http://www.phytozome.net (verified 11 July 2011). University of California Regents, Oakland, CA.
Lin, S.F., S. Cianzio, and R. Shoemaker. 1997. Mapping genetic loci for iron deficiency chlorosis in soybean. Mol. Breed. 3:219–229. doi:10.1023/A:1009637320805
Lin, S.F., R. Shoemaker, S. Cianzio, and D. Grant. 2000. Molecular characterization of iron deficiency chlorosis in soybean. J. Plant Nutr. 23:1929–1939. doi:10.1080/01904160009382154
Liu, K., and S.V. Muse. 2005. PowerMarker: An integrated analysis environment for genetic marker analysis. Bioinformatics 21:2128–2129.
Loiselle, B.A., V.L. Sork, J. Nason, and C. Graham. 1995. Spatial genetic structure of a tropical understory shrub, Psychotriaofficinalis (Rubiaceae). Am. J. Bot. 82:1420–1425. doi:10.2307/2445869
Marschner, H., V. Romheld, and M. Kissel. 1986. Different strategies in higher plants in mobilization and uptake of iron. J. Plant Nutr. 9:3–7.
Morrissey, J., and M.L. Guerinot. 2009. Iron uptake and transport in plants: The good, the bad, and the ionome. Chem. Rev. 109:4553–4567. doi:10.1021/cr900112r
O'Rourke, J.A., R.T. Nelson, D. Grant, J. Schmutz, J. Grimwood, S. Cannon, C.P. Vance, M.A. Graham, and R.C. Shoemaker. 2009. Integrating microarray analysis and the soybean genome to understand the soybeans iron deficiency response. BMC Genomics 10:376. doi:10.1186/1471-2164-10-376
Pritchard, J.K., M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945–959.
Pyhajarvi, T., M. Rosario Garcia-Gil, T. Knurr, M. Mikkonen, W. Wachowiak, and O. Savolainen. 2007. Demographic history has influenced nucleotide diversity in European *Pinus sylvestris* populations. Genetics 177:1713–1724.
Remington, D.L., J.M. Thornsberry, Y. Matsuoka, L.M. Wilson, S.R. Whitt, J. Doebley, S. Kresovich, M.M. Goodman, and E.S. Buckler, IV. 2001. Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc. Natl. Acad. Sci. USA 98:11479–11484. doi:10.1073/pnas.201394398
Risch, N.J. 2000. Searching for genetic determination in the new millennium. Nature 405:847–855. doi:10.1038/35015718
Rosenberg, N.A., T. Burke, K. Elo, M.W. Feldman, P.J. Freidlin, M.A.M. Groenen, J. Hillel, A. Maki-Tanila, M. Tixer-Boichard, L. Vignal, K. Wimmers, and S. Weigend. 2001. Empirical evaluation of genetic clustering methods using multilocus genotypes from 20 chicken breeds. Genetics 159:699–713.
SAS Institute. 2002. The SAS system for Windows. Release 9.00. SAS Inst., Cary, NC.
Scheet, P., and M. Stephens. 2006. A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. Am. J. Hum. Genet. 78:629–644. doi:10.1086/502802
Schmidt, W. 1999. Mechanism and regulation of reduction based iron uptake in plants. New Phytol. 141:1–26. doi:10.1046/j.1469-8137.1999.00331.x
Schmutz, J., S.B. Cannon, J. Schlueter, J. Ma, T. Mitros, W. Nelson, D.L. Hyten, Q. Song, J.J. Thelen, J. Cheng, D. Xu, U. Hellsten, G.D. May, Y. Yu, T. Sakurai, T. Umezawa, M.K. Bhattacharyya, D. Sandhy, B. Valliyodan, E. Lindquist, M. Peto, D. Grant, S. Shu, D. Goodstein, K. Berry, M. Furell-Griggs, B. Abernathy, J. Du, Z. Tian, L. Zhu, N. Gill, T. Joshi, M. Libault, A. Sethuraman, X.-C. Zhang, D. Shnozaki, H.T. Nguyen, R.A. Wing, P. Cregan, J. Specht, J. Grimwood, D. Rokshar, G. Stacey, R.C. Shoemaker, and S.A. Jackson. 2010. Genome sequence of the paleopolyploid soybean. Nature 463:178–183. doi:10.1038/nature08670
Severin, A.J., G.A. Peiffer, W.W. Xu, D.L. Hyten, B. Bucciarelli, J.A. O'Rourke, Y.-T. Bolon, D. Grant, A.D. Farmer, G.D. May, C.P. Vance, R.C. Shoemaker, and R.M. Stupar. 2010. An integrative approach to genomic introgression mapping. Plant Physiol. 154:3–12. doi:10.1104/pp.110.158949
Stephan, U.W. 2002. Intra- and intercellular iron trafficking and subcellular compartmentation within roots. Plant Soil 241:19–25. doi:10.1023/A:1016086608846
Stitch, B., J. Mohring, H.P. Piepho, M. Heckenberger, E.S. Buckler, and A.E. Melchinger. 2008. Comparison of mixed-model approaches for association mapping. Genetics 178:1745–1754. doi:10.1534/genetics.107.079707
Sun, G., M.H. Kramer, S.S. Yang, W. Song, H.P. Piepho, and J. Yu. 2010. Variation explained in mixed-model association mapping. Heredity 105:333–340. doi:10.1038/hdy.2010.11
Wang, J., P.E. McClean, R. Lee, J. Goos, and T. Helms. 2008. Association mapping of iron deficiency chlorosis loci in soybean (*Glycine max* L. Merr.) advanced breeding lines. Theor. Appl. Genet. 116:777–787. doi:10.1007/s00122-008-0710-x
Yu, J., G. Pressoir, W.H. Briggs, I.V. Bi, M. Yamasaki, J.F. Doebley, M.D. McMullen, B.G. Gaut, D.M. Nielsen, J.B. Holland, S. Kresovich, and E.S. Buckler. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. 38:203–208. doi:10.1038/ng1702
Zhao, K., M.J. Aranzana, S. Kim, C. Lister, C. Shindo, C. Tang, C. Toomajian, H. Zheng, C. Dean, P. Marjoram, and N. Magnus. 2007. An *Arabidopsis* example of association mapping in structured samples. PLoS Genet 3(1):e4. doi:10.1371/journal.pgen.0030004