Genetic Mapping of Soybean Seed Protein QTLs

Piyaporn Phansak[1], Watcharin Soonsuwan[1], James E. Specht[1], George L. Graef[1], Perry B. Cregan[2], and David L. Hyten[2]

[1] Department of Agronomy and Horticulture, University of Nebraska, Lincoln, NE 68583-0915

[2] Soybean Genomics and Improvement Laboratory, USDA-ARS, BARC-West, Beltsville, MD 20705-2350

Abstract

A quantitative trait locus (QTL) is the chromosomal location of a gene governing that trait. QTL identification is the first step toward using marker-assisted selection (MAS) to introgress desirable QTL alleles into elite high-yield cultivars. Hundreds of high protein plant introductions (PIs) exist in the USDA germplasm collection and are a source of high protein alleles, but only a few PIs have been characterized for all possible QTLs or the allelic status at known QTLs. Although 86 protein QTLs are listed in SoyBase, many are repeat discoveries of the same QTL(s), given the typical 10cM confidence intervals associated with QTL positions. The additive effect for these 86 QTLs was never greater than the respective 1.2% and 0.85% estimates obtained for the Chromosome 20 [Linkage Group (LG)-I] and Chromosome 15 (LG-E) QTLs. All remaining QTLs had a smaller *apparent* additive effect (or even smaller *true* effect, given the Beavis Effect). Discovery of protein QTLs different from the Chr 20 and Chr 15 QTLs, which express a large additive effect, would be of great interest. We mated 51 high protein (48% or more) germplasm accessions to high-yield public cultivars with normal protein (42% or less) to generate 51 populations. The 240 F_2 plants in each population produced $F_{2:3}$ seed progenies in 2008 that were phenotyped for seed protein content. We used a technique called selective genotyping, or phenotypic tail analysis, to identify the lowest 10% and the highest 10%. We have since used a 1536-SNP locus assay panel to conduct a selective genotyping search for what might be completely new and heretofore unknown protein QTLs. New protein QTLs could be of great value, especially if the high-protein alleles at these QTLs have less impact on seed oil and/or yield than do the known QTLs on Chr 20 and Chr 15.

Background & Objectives

Ritchie (2003) identified 52 *Glycine max* accessions of exceptionally high seed protein in MG 000, 00, 0, I, II, III & IV germplasm, mated these to seven agronomic parents of a similar MG, and eventually obtained 41 F_2 populations (averaging about 120 F_2 plants each). $F_{2:3}$ seed progenies in the highest and lowest deciles of the protein distribution in each population were genotyped with just those SSR markers linked to four seed protein QTLs known to have a large additive effect. A selective genotyping protocol reduces genotyping costs in marker-QTL linkage studies, since only the highly informative F_2 individuals in the high and low phenotypic tails are genotyped (Lander and Botstein, 1989). Ritchie's SSR-based phenotypic tail analyses revealed that 85%, 30%, 9%, and 20% of the 41 populations segregated for the respective QTLs located on LG I, LG-E, and (top & bottom of) LG-H. The high-protein allele of the LG-I QTL was detected in 35 of 41 accessions. The phenotypic tail analysis conducted by Ritchie (2003) had limitations – just four regions of the genome were evaluated with markers (i.e., SSRs linked to the four known QTLs), the average population size of just 120 F_2s restricted the power of QTL detection (= 1 minus the imputed Type II error probability), and only 41 of the 52 accessions were examined. The objective of the research described in this poster is to mitigate those limitations. In the summer 2007, the 52 matings were re-made, 240 F_2 plants were threshed per population in summer 2008 (the greater number of progeny will increase the statistical power for QTL detection). The phenotypic analyses of the 240 $F_{2:3}$ seed progenies occurred in fall 2008, the high and low decile fractions (24 progenies each) were genotyped this spring and summer with 1536 SNP markers (Choi et al., 2007) using a new 96-well 1536-SNP marker linkage panel (Vers. 1.0). About 400 to 500 of those 1536 SNP markers were expected to be polymorphic in any given population, thereby offering suitable genome-wide coverage to scan for both known and (currently) unknown seed protein QTLs of large additive effect. Any observed gaps in genome SNP marker coverage in any population will be handled, if needed, by genotyping the decile groups with gap-located SSR and (other) SNP markers.

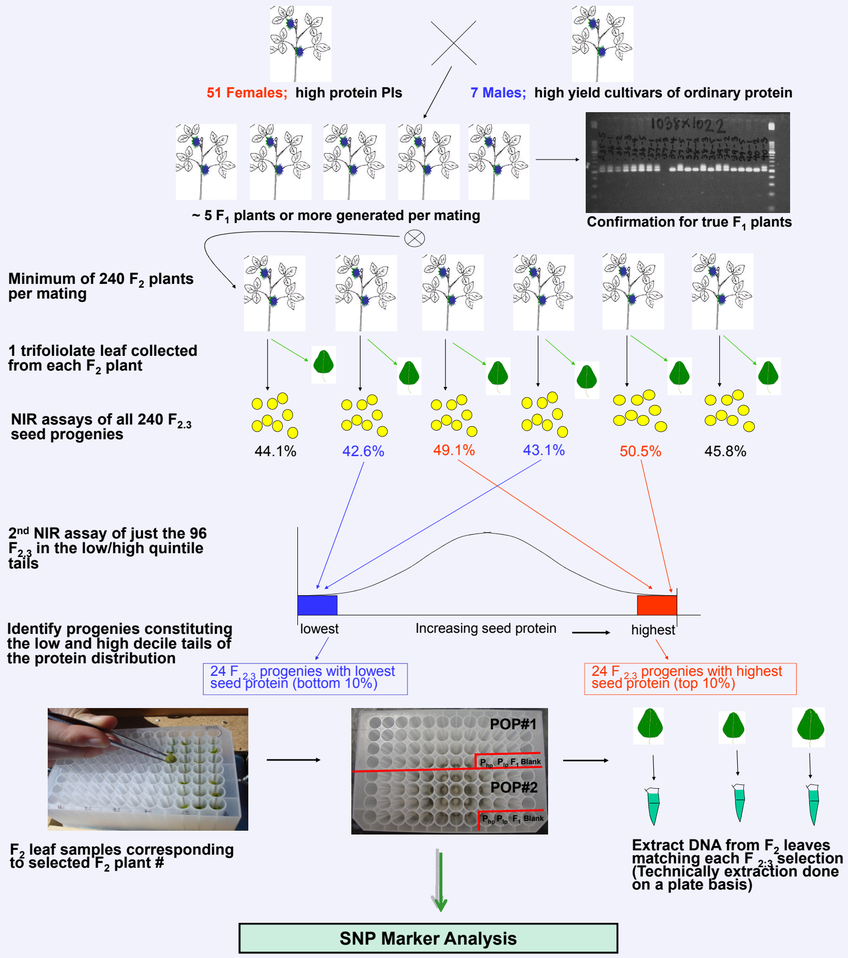
Approach & Protocols

The 52 female high protein PIs of MG 000-IV were mated to seven high-yielding cultivars of a matching MG, and the F_2 populations of 300 seeds each were planted in spring 2008 (see diagram). A small leaf collected from each (tagged) F_2 plant was freezer-stored. F_2 plant threshing in the fall resulted in ~240 $F_{2:3}$ progenies per mating that were NIR-phenotyped for seed protein. The 96 $F_{2:3}$ progenies in the lowest and highest quintiles of the seed protein distributions in each 240-F_2 population were subjected to a second NIR assay to mitigate phenotyping errors before selecting the 48 $F_{2:3}$ seed progenies (based on 2-rep means) in the lowest and highest deciles. DNA was extracted from freezer-preserved leaves of the 48 F_2 plants corresponding to the 24 low and 24 high $F_{2:3}$ seed progenies and used in the 96-well 1536-SNP plate assays. If a given segregating molecular marker is NOT associated with seed protein (which implies a protein QTL is either NOT linked to that marker, or the known QTL linked to the marker is NOT segregating in the population), then the marker A allele frequency (p) will be 0.5 in each decile. This null hypothesis of no difference between marker allele A frequencies in high *vs.* low deciles was evaluated with a two-sample t-test (Bernardo 2002) :

$$t = (pA_{hi} - pA_{lo}) / \text{sqrt}\{[p(1\text{-}p)/2N_{hi}]+[p(1\text{-}p)/2N_{lo}]\}$$

where pA = marker allele A frequency; N = progeny in the high (hi) or low (lo) deciles.
See Lebowitz et al. (1989) for details about selective genotyping and its statistical power.

Materials & Methods



The high multiplex capability of the Illumina GoldenGate assay was the technique we employed for SNP detection of $F_{2:3}$ progenies in the low and high decile groups, the two parents, and F_1 progeny using the Illumina BeadStation 500 Genotyping Platform. With this assay, 96 genotypes can be assayed with 1536 SNPs in a 3-day period (Hyten et al., 2008).

Linkage Mapping and QTL Analysis

The program MAPMAKER/EXP VER 3.0 (Lander et al., 1987; Lincoln et al., 1993) was used for determining genetic linkage (i.e., marker order and Kosambi map distance). QTL analysis was performed using R/QTL software (Broman et al., 2003).

Results & Discussion

Table 1 Protein QTLs identified by standard interval mapping using EM algorithm in 20 populations of $F_{2:3}$ progeny. QTL peaks were judged significant if their LOD scores exceeded a genome-wide 1000-permutation LOD score. The additive (a) and dominant (d) effects were those calculated for the substitution of a high protein female allele for a cultivar male allele at the indicated marker locus.

Pop. No.	Hi-Pro Female Parent	Norm-Pro Male Parent	QTL Marker/or Nearest Marker[y]	LG Name	Chr. No.	USLP 1.0 Pos. (cM)	LOD Score >3.0 MR	LOD Score >3.0 IMP	LOD Score >3.0 EM	QTL Effect a[x] (%)	QTL Effect d (%)	R^2
1121	PI 398672	Pana	ss107928195[z]	D1a	1	37.14	-	-	4.10	-0.39	-0.11	8.3
1146	PI 407823	Rend	ss107912843	D1b	2	30.67	3.17	-	3.40	-0.73	-0.04	8.7
1025	PI 153301	Jim	ss4969738[z]	D1b	2	36.12	-	-	3.01	0.31	-0.17	5.2
1121	PI 398672	Pana	ss107920774[z]	D1b	2	105.03	-	-	5.14	-0.32	-0.08	10.3
1152	PI 407773B	Rend	ss107915821[z]	N	3	38.18	-	-	4.19	-0.55	-0.35	12.5
1140	PI 424286	Rend	ss107913261	N	3	62.62	3.64	3.18	3.80	-0.59	0.22	8.8
1152	PI 407773B	Rend	ss107927543[z]	C2	6	34.40	-	-	5.53	0.69	0.16	16.1
1121	PI 398672	Pana	ss107917851[z]	C2	6	97.81	-	-	6.02	-0.52	-0.34	12.0
1108	PI 398516	Dwight	ss107917254[z]	C2	6	103.33	-	-	3.95*	-0.65	-0.23	7.7
1076	PI 437112A	Dwight	ss107930557	C2	6	104.50	4.89	-	4.86*	-0.86	-0.10	11.2
1152	PI 407773B	Rend	ss107929806[z]	C2	6	129.53	-	-	3.43	0.10	0.36	10.3
1025	PI 153301	Jim	ss107919937[z]	C2	6	130.27	-	-	3.72	0.42	0.14	6.4
1138	PI 253666A	Rend	ss107920301[z]	C2	6	132.85	-	-	3.02	-0.43	-0.47	9.4
1108	PI 398516	Dwight	ss107928306	A2	8	113.98	3.28	-	3.30	-0.39	0.31	6.5
1152	PI 407773B	Rend	ss107913292[z]	A2	8	128.80	-	-	4.64	-0.21	-0.49	13.7
2212	AC Proteus	Jim	ss107913002	K	9	47.38	3.62	-	3.20	0.34	-0.06	5.2
1110	PI 340011	Pana	ss4969791[z]	O	10	92.39	-	-	3.82	0.69	-0.05	8.1
1076	PI 437112A	Dwight	ss107919004	O	10	96.44	8.05	6.14	6.94*	0.95	-0.38	15.6
1142	PI 407877B	Rend	ss107915265	O	10	99.69	3.56	-	3.62	0.65	-0.03	7.4
1113	PI 408138C	Pana	ss107915265[z]	O	10	99.69	-	-	4.94*	0.83	-0.10	9.9
1121	PI 398672	Pana	ss107915265	O	10	99.69	5.83	3.69	5.96	0.83	-0.21	11.9
1122	PI 360843	Pana	ss107915265	O	10	99.69	5.99	4.63	5.72*	0.72	-0.15	13.0
1143	PI 398704	Rend	ss107930838[z]	O	10	117.86	-	-	3.46	0.49	-0.26	7.8
1139	PI 407788A	Rend	ss107917851	B1	11	110.63	3.37	3.19	3.46	0.45	-0.47	9.1
1076	PI 437112A	Dwight	ss107919004	O	10	96.44	8.05	6.14	6.94*	0.95	-0.38	15.6
1142	PI 407877B	Rend	ss107915265	O	10	99.69	3.56	-	3.62	0.65	-0.03	7.4
1113	PI 408138C	Pana	ss107915265[z]	O	10	99.69	-	-	4.94*	0.83	-0.10	9.9
1121	PI 398672	Pana	ss107915265	O	10	99.69	5.83	3.69	5.96	0.83	-0.21	11.9
1122	PI 360843	Pana	ss107915265	O	10	99.69	5.99	4.63	5.72*	0.72	-0.15	13.0
1143	PI 398704	Rend	ss107930838[z]	O	10	117.86	-	-	3.46	0.49	-0.26	7.8
1139	PI 407788A	Rend	ss107917851	B1	11	110.63	3.37	3.19	3.46	0.45	-0.47	9.1
1146	PI 407823	Rend	ss107912786	H	12	56.14	-	-	3.18	-0.56	-0.08	8.2
1146	PI 458256	Rend	ss107928782[z]	H	12	62.14	-	-	4.25*	0.12	0.31	8.1
1107	PI 445845	Pana	ss107919562[z]	H	12	78.39	-	-	3.79	0.40	0.35	7.6
1183	PI 458256	Rend	ss107913481	F	13	74.54	3.67	-	3.03	-0.43	0.03	5.9
1146	PI 407823	Rend	ss107930533	B2	14	13.10	3.87	2.96	3.94	-0.68	0.04	10.1
1143	PI 398704	Rend	ss107929437[z]	E	15	17.01	-	-	5.05*	0.77	-0.22	11.2
1140	PI 424286	Rend	ss107913472[z]	E	15	40.00	4.77	-	4.12*	0.60	-0.27	9.6
1025	PI 153301	Jim	ss107924165	E	15	75.94	-	-	3.00	-0.15	0.27	5.2
1022	PI 153302	Jim	ss107913754	J	16	57.70	3.30	-	3.54	0.51	0.07	6.2
1025	PI 153301	Jim	ss107913910	J	16	67.48	4.32	-	3.91	0.56	0.02	6.7
1121	PI 398672	Pana	ss107923378	G	18	0.00	-	-	3.78	-0.74	0.17	7.7
1108	PI 398516	Dwight	ss107929989[z]	G	18	0.11	-	3.26	4.56	-0.40	-0.39	8.9
1023	PI 153301	Jim	ss107915182[z]	G	18	64.26	-	-	3.39	0.36	-0.09	5.9
1023	PI 159764	Jim	ss107918449	L	19	84.53	3.80	-	3.58	0.09	0.49	6.6
1152	PI 407773B	Rend	ss107913892	I	20	19.67	-	-	3.21	0.76	0.23	9.7
1107	PI 445845	Pana	ss107913608[z]	I	20	29.56	-	-	3.54	0.62	0.03	7.1
1110	PI 340011	Pana	ss107913608[z]	I	20	29.56	-	-	7.09*	0.83	-0.05	14.5
1138	PI 253666A	Rend	ss107913608	I	20	29.56	-	-	3.89	0.93	0.20	11.9
1143	PI 398704	Rend	ss107917070	I	20	30.00	3.90	-	3.16	0.61	0.16	7.2
1022	PI 153302	Jim	ss107913844	I	20	33.21	4.00	-	3.22*	0.50	0.11	5.7
1023	PI 153301	Jim	ss107913844	I	20	33.21	4.20	2.66	4.12*	0.72	0.10	7.5
1024	PI 438415	Jim	ss107913844	I	20	33.21	4.28	-	4.47*	0.52	0.15	7.6
1025	PI 153301	Jim	ss107913844	I	20	33.21	5.85	3.98	6.02*	0.57	-0.06	10.2
1139	PI 407788A	Rend	ss107913844	I	20	33.21	-	5.94	7.69*	1.12	-0.07	19.1
1113	PI 408138C	Pana	ss107913138	I	20	44.34	5.71	3.65	5.28*	0.62	-0.43	10.5

* = Statistically significant (determined by using the 95 percentile of genome-wide maximum LOD scores of 1000 permutations).

[x] = If negative from the normal protein parent.

Results & Discussion

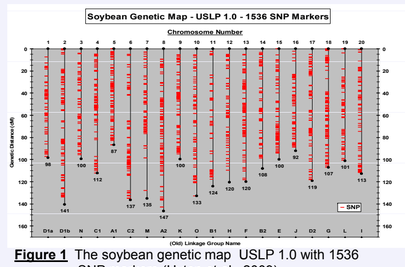


Figure 1 The soybean genetic map USLP 1.0 with 1536 SNP markers (Hyten et al., 2009)

In this study, 1536 SNPs that map across the entire soybean genome (Figure 1) were used to screen the 22 high protein and 22 low protein $F_{2:3}$ progenies in all 48 populations (3 were lost). Obviously, all 1536 SNPs were not parentally polymorphic, but we expected approximately 30-50% (460-770) of the SNPs to be polymorphic. To date, 20 of 48 populations have been analyzed. Approximately 500 SNPs were identified as polymorphic in each population. The remaining 28 populations will be analyzed later this year.

Ritchie (2003) scanned only four small genome segments in her 41 high x low protein populations for protein QTLs. In our experiment, the entire genome was scanned for possible segregating protein QTLs in those same 41 (+7 more) populations. Of particular interest were the six populations in which she observed large phenotypic protein distributions but did not detect segregation of the four known protein QTLs. Also of interest were the QTLs in her 11 failed matings. In the 20 mapping populations, 40 protein QTLs with LOD scores greater than 3 were detected and mapped on 16 linkage groups (Table 1, Figure 2). The QTL detection power of our decile-based selective genotyping protocol with a 240 F_2 population size, using a phenotypic protein standard deviation estimate of 3.0% (Ritchie, 2003), indicated that the power was 100% for QTLs with an additive effect of ≥ 0.9 % protein, and about 80% for an additive effect of ≥ 0.6% protein.

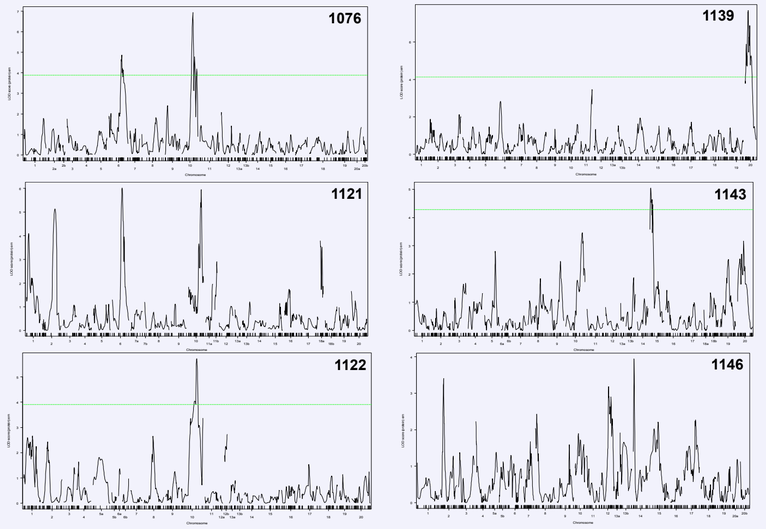


Figure 2: Protein QTLs identified by standard interval mapping using the EM algorithm in the soybean genome. The population number shown in the graphs are listed in table 1. The green lines show the threshold LOD score at the 95 percentile of the genome-wide maximum LOD scores of 1000 stratified permutation replicates calculated by the EM algorithm.

Conclusions

To date, 20 of 48 populations have been analyzed for QTLs. Relative to the six high x normal protein populations within which Ritchie (2003) found no evidence of QTLs on LGs I, E, H_{top}, H_{bottom}, we detected significant QTLs on chromosomes 2 (D1b), 10 (O), 11 (B1), 12 (H_{top}), 14 (B2), and 18 (G). Other than chromosome 12, the protein QTLs discovered on these chromosomes are not currently listed in SoyBase. For improving the seed protein content in high yielding soybean cultivars, these new protein QTLs may be useful to soybean breeders.

Acknowledgements

This research is supported by funding provided by the United Soybean Board (Project 8212) and the Nebraska Soybean Board.

References

Bernardo, R. 2002. Breeding for Quantitative Traits in Plants. Stemma Press, Woodbury, MN. (369 pages).

Broman KW, Wu H, Sen Ś, Churchill GA (2003) R/qtl: QTL mapping in experimental crosses. Bioinformatics 19:889-890.

Choi, I., D.L. Hyten, L.K. Matukumalli, Q. Song, J. Chaky, et al. 2007. A soybean transcript map: gene distribution, haplotype, and single-nucleotide polymorphism (SNP) analysis. Genetics 176: 685-696.

Hyten, D., Q. Song, I.-Y. Choi, M.-S. Yoon, J. Specht, L. Matukumalli, R. Nelson, R. Shoemaker, N. Young, and P. Cregan. 2008. High-throughput genotyping with the GoldenGate assay in the complex genome of soybean. Theor. Appl. Genet. 116: 945-952.

Lander, E., P. Green, J. Abrahamson, A. Barlow, M. Daly, S. Lincoln, and L. Newburg. 1987. MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics 1:174-181.

Lander, E.S. and D. Botstein. 1989. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 134: 943-945.

Lebowitz, R.J., M. Soller, and J.S. Beckmann. 1987. Trait-based analyses for the detection of linkage between marker loci and quantitative trait loci in crosses between inbred lines. Theor Appl. Genet. 73:556-562.

Lincoln, S., M. Daly, and E. Lander. 1993. Constructing genetic maps with MAPMAKER/EXP version 3.0: a tutorial and reference manual: pp 97.

Ritchie, R. A. 2003. High-protein plant introductions: selective genotyping to detect soybean protein QTL. M.S. thesis. Univ. of Nebraska, Lincoln.

Xu, S. 2003. Theoretical basis of the Beavis Effect. Genetics 165: 2259-22.