CellPress
PARTNER JOURNAL

Molecular Plant
Research Article

Genome-wide Scan for Seed Composition Provides Insights into Soybean Quality Improvement and the Impacts of Domestication and Breeding

Jiaoping Zhang[1,7], Xianzhi Wang[1,6,7], Yaming Lu[1], Siddhi J. Bhusal[1], Qijian Song[2], Perry B. Cregan[2], Yang Yen[3], Michael Brown[4] and Guo-Liang Jiang[5,*]

[1]Plant Science Department, South Dakota State University, Brookings, SD 57006, USA

[2]Soybean Genomics and Improvement Laboratory, US Department of Agriculture, Agricultural Research Services (USDA-ARS), 10300 Baltimore Avenue, Beltsville, MD 20705, USA

[3]Department of Biology and Microbiology, South Dakota State University, Brookings, SD 57006, USA

[4]Department of Natural Resource Management, South Dakota State University, Brookings, SD 57006, USA

[5]Agricultural Research Station, Virginia State University, PO Box 9061, Petersburg, VA 23806, USA

[6]Present address: School of Agriculture, Yunnan University, Kunming, Yunnan Province, China

[7]These authors contributed equally to this article.

***Correspondence: Guo-Liang Jiang (gjiang@vsu.edu)**

https://doi.org/10.1016/j.molp.2017.12.016

ABSTRACT

The complex genetic architecture of quality traits has hindered efforts to modify seed nutrients in soybean. Genome-wide association studies were conducted for seed composition, including protein, oil, fatty acids, and amino acids, using 313 diverse soybean germplasm accessions genotyped with a high-density SNP array. A total of 87 chromosomal regions were identified to be associated with seed composition, explaining 8%–89% of genetic variances. The candidate genes *GmSAT1*, *AK-HSDH*, *SACPD-C*, and *FAD3A* of known function, and putative *MtN21 nodulin*, *FATB*, and steroid-5-α-reductase involved in N_2 fixation, amino acid biosynthesis, and fatty acid metabolism were found at the major-effect loci. Further analysis of additional germplasm accessions indicated that these major-effect loci had been subjected to domestication or modern breeding selection, and the allelic variants and distributions were relevant to geographic regions. We also revealed that amino acid concentrations related to seed weight and to total protein had a different genetic basis. This helps uncover the in-depth genetic mechanism of the intricate relationships among the seed compounds. Thus, our study not only provides valuable genes and markers for soybean nutrient improvement, both quantitatively and qualitatively, but also offers insights into the alteration of soybean quality during domestication and breeding.

Key words: genome-wide association study, GWAS, soybean, seed composition, candidate genes, quality improvement, domestication and breeding

Zhang J., Wang X., Lu Y., Bhusal S.J., Song Q., Cregan P.B., Yen Y., Brown M., and Jiang G.-L. (2018). Genome-wide Scan for Seed Composition Provides Insights into Soybean Quality Improvement and the Impacts of Domestication and Breeding. Mol. Plant. **11**, 460–472.

INTRODUCTION

Soybean (*Glycine max*) is the number one oilseed crop, accounting for 60% of the world oilseed production (US Department of Agriculture Foreign Agricultural Service, 2017). Soybean seed is rich in both protein and oil. Generally, it contains 40% protein and 20% oil. Soybean seed protein is composed of 18 amino acids, including all 10 essential amino acids. Among them, Trp, Met, and Cys are the most deficient. Soybean oil is high in unsaturated fatty acids, and it typically consists of 12% palmitic

Published by the Molecular Plant Shanghai Editorial Office in association with Cell Press, an imprint of Elsevier Inc., on behalf of CSPB and IPPE, SIBS, CAS.

acid (16:0), 4% stearic acid (18:0), 23% oleic acid (18:1), 53% linoleic acid (18:2), and 8% linolenic acid (18:3). The content of seed composition has a striking effect on the quality and uses of soybean products such as protein meal and baking oil.

Using various population types and different mapping techniques, previous studies have identified 304 and 221 quantitative trait loci (QTL) governing seed protein and oil concentration in soybean, respectively (SoyBase, https://soybase.org/). Some of them showed a pleiotropic effect, such as the hotspot on chromosome 20 (Gm20) that is associated with both protein and oil content but in opposite manner (Diers et al., 1992; Sebolt et al., 2000; Chung et al., 2003). Recent studies leveraging the power of genome-wide association mapping confirmed and further narrowed down this QTL from 8.4 Mb (Bolon et al., 2010) to less than 3 Mb containing few genes (Hwang et al., 2014; Vaughn et al., 2014; Bandillo et al., 2015). Through a genome-wide association study (GWAS) and comparing the genetic differential among soybean wild relatives, landraces, and cultivars, Zhou et al. (2015) and Han et al. (2016) identified novel loci for protein and oil and a broad overlap between oil QTL and selective sweeps, indicating the extensive effect of domestication and breeding activities on oil content in soybean.

Soybean quality not only depends on the concentration of total oil and protein, but it also relies on the profiles of fatty acids and amino acids. A large number of QTL associated with fatty acids have been reported and can be found on SoyBase (https://soybase.org/). In addition, many novel loci have been identified in recent studies with diverse wild soybeans (Leamy et al., 2017) or a population derived from cultivated soybean (Li et al., 2017). Previous studies also identified some key regulators of fatty acid biosynthesis in soybean, such as *3-Keto-Acyl-ACP synthase II* (Aghoram et al., 2006), *stearoyl-ACP desaturases* (*SACPD-A*, *B*, and *C*) (Byfield et al., 2006; Zhang et al., 2008; Gillman et al., 2014), fatty acid desaturases (*FAD*-2, 3, and 7) (Heppard et al., 1996; Bilyeu et al., 2005; Andreu et al., 2010) and *fatty acyl-ACP thioesterases* (*FAT-A* and *B*) (Hitz and Yadav, 1992; Cardinal et al., 2007), which are involved in carbon-chain elongation, dehydrogenation, and termination of fatty acids. Progress in investigating the molecular basis of fatty acid biosynthesis has led to improved soybean genotypes with dramatically altered oil composition. For instance, high oleic (>80%) and ultra-low linolenic (~1%) soybeans have been developed through pyramiding *FAD2* and *FAD3* (also known as *fan loci*) genes (Ross et al., 2000; Pham et al., 2012). However, an analysis of soybean genome identified over 250 gene homologs governing seed oil as well as fatty acid storage and metabolism (Schmutz et al., 2010), indicating that our understanding of the complex genetic architecture of related traits in soybean is far from complete.

Compared with our understanding of fatty acid biosynthesis, knowledge of molecular mechanisms of amino acid biosynthesis in soybean was mainly limited to genetic mapping (Panthee et al., 2006a; Carlson, 2011; Fallen et al., 2013; Wang et al., 2015). Recent studies have enhanced our understating of the genetic architecture of amino acids in soybean. Vaughn et al. (2014) conducted GWAS for Met, Thr, Cys, and Lys using data available on the Germplasm Resources Information Network (GRIN; https://www.ars-grin.gov/) and identified multiple loci on Gm01 and Gm08, which were not reported previously, associated with multiple amino acids across different populations. Qiu et al. (2014) predicted 12 candidate genes based on the synteny between 113 genes of sulfur-containing amino acid synthases and 33 related QTL in soybean, and bioinformatic analyses. Although soybean lines with improved amino acid profiles can be achieved through bioengineering (Falco et al., 1995; Krishnan, 2005), breeding efforts for the improvement of soybean seed amino acids have rarely been reported.

Soybean has been widely used in the food, feed, and fuel industries. The diverse and increasing demands for soybean (Grundy, 1986; Wardlaw and Snook, 1990; Ramos et al., 2009) have created considerable impetus for exploring the genetic mechanisms controlling soybean seed composition. To obtain a comprehensive understanding of genetic architecture of soybean quality, we carried out an extensive and in-depth investigation into seed composition through molecular and genomic approaches. We conducted a GWAS for soybean seed compounds, including protein, oil, five fatty acids, and 18 amino acids, using a diverse panel genotyped with the SoySNP50K BeadChip. Using the genotypic data of 13 433 germplasm accessions, we also explored the allelic variation and the impact of domestication and modern breeding on soybean quality. A total of 138 QTL associated with seed composition and nine candidate genes at the major-effect loci were identified. The results revealed the different genetic basis between amino acid concentration related to seed weight and to total protein content, and demonstrated the profound impact of domestication and modern breeding on soybean seed composition.

RESULTS

Phenotypic Variations and Correlations among Traits

All of the seed composition traits involved in this study exhibited a continuous distribution (Supplemental Figure 1) and were quantitatively inherited. Seed protein and oil content varied from 36.3% to 46.3% and from 15.7% to 22.5%, respectively (Supplemental Table 1). In comparison, linolenic acid and His exhibited the largest relative variations, while most of the amino acids had small variations (Supplemental Figure 2). Interestingly, the sulfate-containing amino acid Cys had a greater relative variation than that of protein or oil. Estimates of heritability indicated that, for all the traits, the genetic component dominated the phenotypic performance (Supplemental Table 1). The well-documented inverse correlation between protein and oil concentrations was also observed in this study with a correlation coefficient $r = -0.77$ (Supplemental Figure 3). However, a two-segment linear regression model ($r^2 = 0.67$), which partitions the dataset into two parts at 19.6% of oil concentration, showed a better fit to the data than a regular regression model ($r^2 = 0.59$) (Supplemental Figure 3). The oil content was positively correlated with saturated fatty acids but negatively correlated with the unsaturated ones (Supplemental Figure 4).

Positive phenotypic and genetic correlations were detected between total protein and all amino acids, and among amino acids in most cases without correcting for total protein content (Supplemental Figure 5A). By calculating amino acids relative to

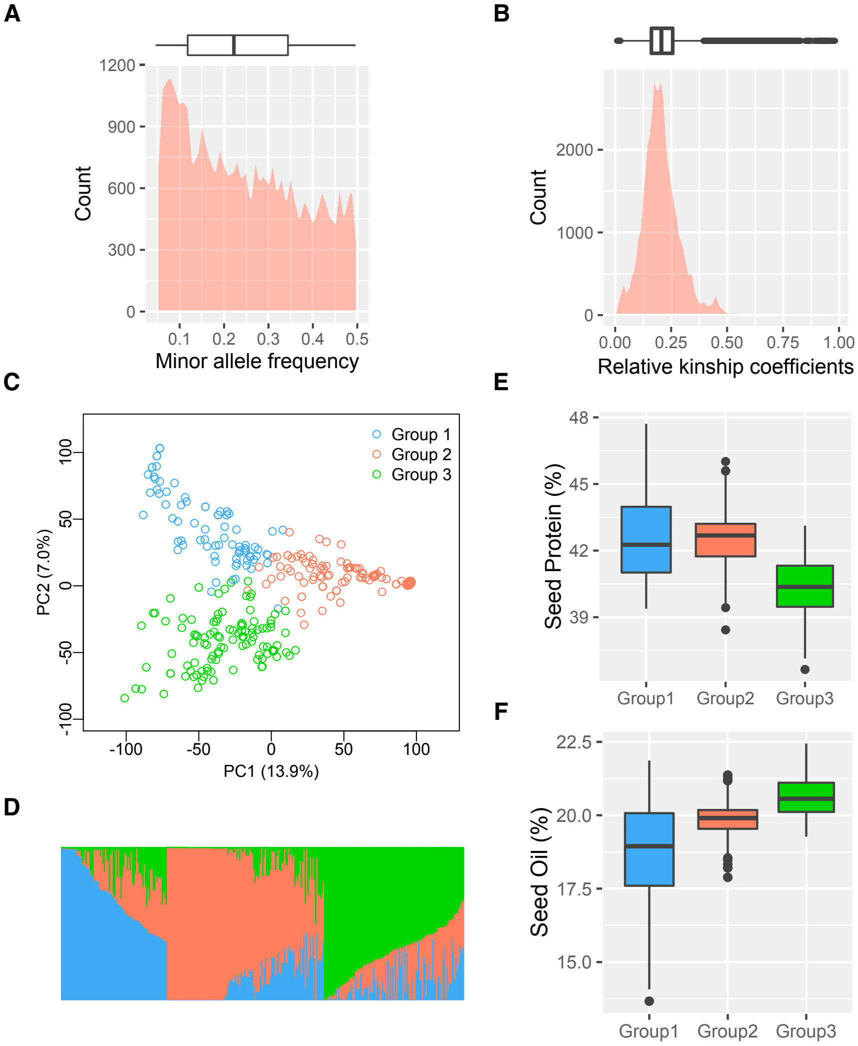

Figure 1. Genetic Diversity and Population Structure Analysis of 313 Soybean Germplasm Accessions.

(A) Distribution of the minor allele frequency of the SNPs used in this study.

(B) Distribution of the pairwise relative kinship coefficients of the association panel.

(C) Plot of the first two components of the principal component analysis of the population genetic variation. The three subgroups are defined by STRUCTURE analysis.

(D) STRUCTURE analysis for the population stratification with k = 3.

(E and F) Association between the population structure and the seed protein **(E)** and oil **(F)** concentration.

total protein content, however, different relationships among them were found from those without correcting for protein (Supplemental Figure 5B). Discounting insignificant ones, roughly half of the correlation coefficients among the adjusted amino acids were positive and half were negative. More interestingly, most of the protein-corrected amino acids exhibited a negative correlation with total protein content (Supplemental Figure 5B). For some pairs of protein-corrected amino acids, the genetic correlation was also different from the phenotypic correlation, such as Leu_P and Gly_P, Leu_P and Glu_P, and Met_P and Leu_P. The results suggested that an increase of total protein may improve the amount of amino acids in terms of the absolute content, but not necessarily change the amino acid profile in a given mode, which is crucial to the quality of soybean meal, since protein content was not always positively correlated with the adjusted content of amino acids, which exhibited a more complicated relationship.

Genetic Diversity and Population Structure

The SNPs with minor allele frequency (MAF) $\geq$0.05 but <0.10 accounted for 18% (5751) of the total SNPs used for association analysis (Figure 1A). Almost half of the pairwise kinship estimates were close to 0.20, representing a moderate level of familial relationships within the population (Figure 1B), which is expected in soybean, a self-pollinated crop with a relative slow linkage disequilibrium (LD) decay rate (Zhang et al., 2015b).

The first two principal components (PCs) explained 20.9% of the total variation (Figure 1C) and were applied to control the population stratification as suggested by the model fitness analysis with the Bayesian information criterion (BIC). The individuals were assigned into three subgroups using the STRUCTURE program (Figure 1D) (Pritchard et al., 2000). The subgroups were well captured by the first two PCs (Figure 1C). The population differentiation (F_{st}) was estimated at 0.19, slightly higher than that of a diverse collection of *Oryza sativa indica* landraces (F_{st} = 0.17) reported in a previous study (Huang et al., 2010). Significant differences in protein and oil contents were discovered across subgroups (Figure 1E and 1F), suggesting an association between the population structure and the variation of seed composition traits and the necessity of involving population structure in the association analysis.

GWAS for Seed Composition and Candidate Genes

A total of 87 chromosomal regions on 19 of the 20 soybean chromosomes were identified to be associated with seed composition (Supplemental Table 2). Of them, 18 exhibited pleiotropic effects. As a result, 138 QTL associated with the traits were detected, three for protein, 25 for oil, 18 for fatty acids, and 39 and 53 for amino acids on a dry-weight (DW) and a protein (P) basis, respectively (Figure 2, Supplemental Figure 6, Supplemental Table 2). Among them, 95 and 85 QTL were identified by regular mixed linear model (MLM) and multi-locus mixed model (MLMM), respectively, and 42 QTL were detected by both approaches (Supplemental Table 2). However, only four of the 92 amino acid QTL were associated with both DW- and P-based amino acids, indicating that different genetic architecture existed between the amino acids

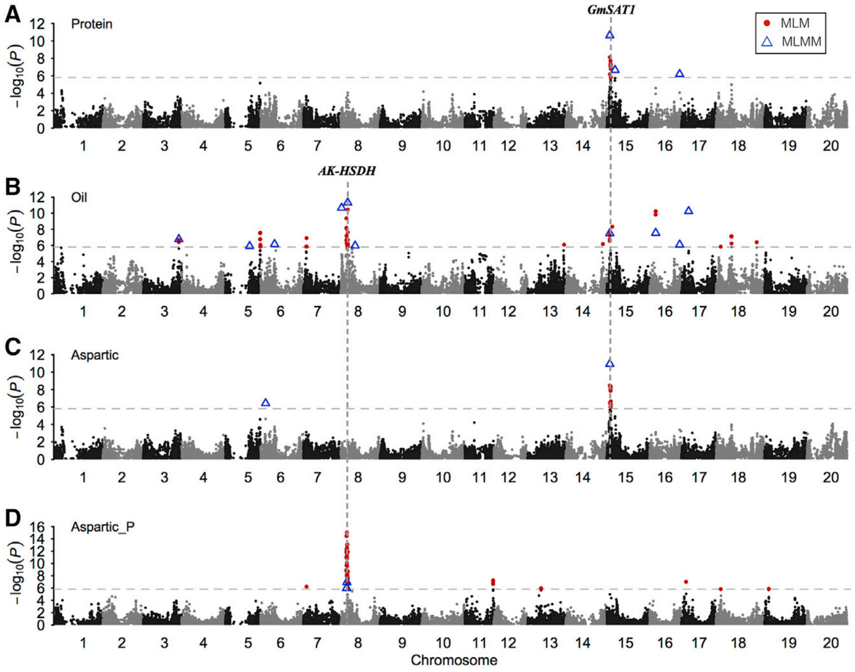

Figure 2. Manhattan Plots of Genome-wide Association Analyses with Mixed Linear Model (MLM) and Multi-locus MLM (MLMM).

(A) Seed protein.

(B) Seed oil.

(C) Aspartic acid relative to dry weight.

(D) Aspartic acid relative to protein concentration.

Negative $\log_{10}$-transformed P values from a genome-wide scan by using MLM and MLMM are plotted against positions on each of the 20 chromosomes. The significant traits-associated SNPs ($P = 1.57 \times 10^{-6}$) are distinguished by the horizontal dashed threshold line and highlighted in red dots and blue triangles. The candidate genes are also given.

determined in the two distinct measurements. No QTL was detected for Pro, Cys_P, His_P, Ile_P, Leu_P, Met_P, and Ser_P. The genetic variances explained by the identified loci varied from 8% for linoleic and Tyr_P to 89% for Trp_P (Supplemental Figure 7). Eighteen candidate genes, including known or putative functions involved in nitrogen fixation, amino acid biosynthesis, and fatty acid metabolism were identified in close proximity to the peak SNPs of the loci (Supplemental Table 2).

A major-effect QTL with broad impact on seed composition was identified at 3.8–4.8 Mb on Gm15. It was led by *ss715622170* or *ss715621777*, depending on the trait. They were in high LD (r^2 = 0.61) and were 839 kb apart from each other. This locus was concurrently associated with the content of protein, oil, linolenic acid, and 11 amino acids on a DW basis and six amino acids on a P basis (Supplemental Figure 8). It explained 22% and 14% of the genetic variation for protein and oil content, respectively, and represented the strongest trait-marker association for most seed storage compounds. The lines with *AA* genotype at *ss715622170* exhibited 1.9% higher protein and 1.1% lower oil than the lines carrying *GG* genotype, resulting in a protein/oil ratio of 1.6 (Figure 3A–3B). *Glycine max symbiotic ammonium transporter 1* (*GmSAT1*) was found 45.9 kb downstream of *ss715622170* on Gm15 (Figure 3C). *GmSAT1* has been specifically detected in N_2-fixing nodules and is located on the plant-derived peribacteroid membrane that encloses N_2-fixing bacteroids in nodules (Kaiser et al., 1998). In addition, candidate genes *Glyma.15g049700*, *Glyma.15g049800*, and *Glyma.15g049900* encoding putative nodulin MtN21 family proteins were also identified at 69.7 kb away from *ss715621777* (Supplemental Figure 9). In *Arabidopsis*, the nodulin MtN21 gene *SIAR1* is known as a bidirectional amino acid transporter and plays an important role in organic nitrogen allocation (Ladwig et al., 2012). *SIAR1* may function in loading amino acids into the xylem in the root, which is substantial for the efficiency of N_2 fixation and seed development (Carter and Tegeder, 2016).

Another locus at the 8.3–9.3 Mb on Gm08, tagged by *ss715602750*, was associated with seven amino acids on a DW basis and six on a P basis, and with the content of total oil and oleic acid. This locus accounted for 15%–67% of the genetic variances. *Glycine Max Aspartokinase-homoserine dehydrogenase* (*GmAK-HSDH*) was 19.8 kb away from *ss715602750* and next to *Rhg4*, one of the major cyst nematode resistance genes in soybean (Figure 4A). AK-HSDH is a bifunctional enzyme catalyzing the key steps of Asp phosphatization and aspartate-semialdehyde to homoserine by which aspartate family amino acids (Lys, Thr, Met, and Ile) are synthesized in plants (Zhu-Shimoni and Galili, 1998) (Figure 4B). Accordingly, this locus was also detected to be associated with Asp_P, Thr, and Lys_P (Figure 4C–4E). Asp is one of the major agents playing an important role in nitrogen transport and storage and is involved in the biosynthesis of other amino acids (Lam et al., 1994). Concordantly, this locus was also associated with variation of Tyr, Arg_P, and Phe_P (Figure 4F–4H). However, no association with the sulfate-containing amino acids was detected at this locus. Further investigating the dataset of a recently published resequencing study in soybean identified a non-synonymous SNP *ss.99218558* at the third exon of *AK-HSDH* that was in high LD with *ss715602750* (also known as *ss.99218329*), r^2 = 0.59, in 302 diverse lines (Supplemental Figure 10) (Zhou et al., 2015). In addition, a group of SNPs at the promoter region of *AK-HSDH* was also identified in high LD with *ss715602750* (r^2 > 0.70). The above results strongly suggested that *AK-HSDH* was very likely the causal gene of this QTL.

The locus on Gm05 specifically associated with oil concentration accounted for 40% of the genetic variation. A 2.9% difference was observed between genotypes segregating at this locus (Supplemental Figure 11A). The candidate gene *Glyma.05g245000*, encoding a putative steroid-5-α-reductase, was pinpointed by *ss715591633* (P = 8.7 × 10^{-5}) at the 3′ untranslated region, which is 28.8 kb apart and in high LD (r^2 = 0.90) with the leading SNP *ss715591638* (Supplemental Figure 11B). It has been reported that the steroid-5-α-reductase family protein is involved in elongation of very long chain fatty

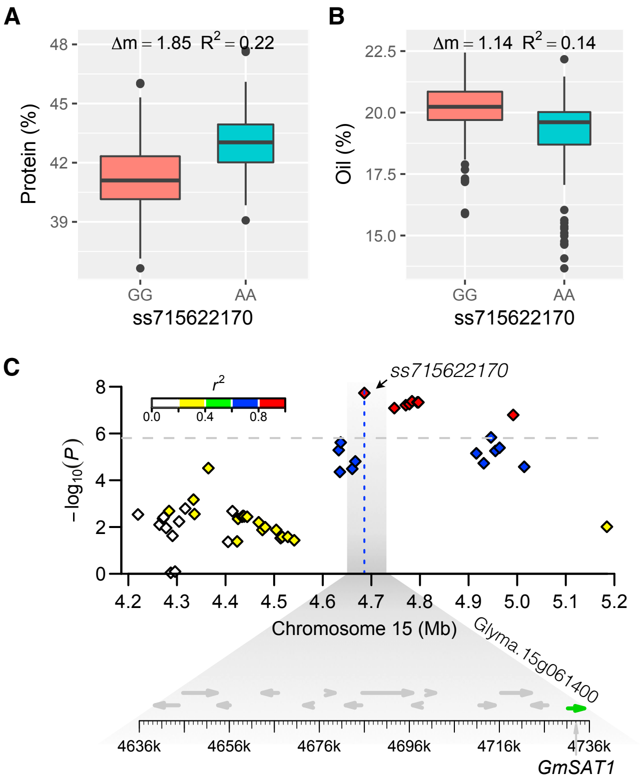

Figure 3. Difference in Seed Protein and Oil Concentration between Lines Segregating at the Lead SNP *ss715622170* Associated with Seed Composition in Soybean and the Candidate Gene.

(A and B) The boxplot shows the differences in the genotypic values of seed protein **(A)** and oil **(B)** over three environments between lines segregating at *ss715622170*. The box shows the first, second (median), and third quartile. The width of the box is proportional to the square root of the number of individuals with each allele. The vertical lines extend to 1.5 times the interquartile or the data extreme, whichever is smaller. The differences in mean (Δm) and R^2 are also given.

(C) The top of the panel shows regional associations for the protein concentration using MLM for the indicated region. The color of each SNP indicates its r^2 value with the peak SNP as shown in the color intensity index on the top left. The bottom panel shows all putative genes within the 50 kb adjacent region on each side of the peak SNP as indicated by the shadow. The candidate gene is indicated by a green arrow.

acids and metabolism of storage lipids (Zheng et al., 2005). Also on Gm05, the locus at the 1.1 Mb position led by *ss715592503* or *ss715592495* was associated with both palmitic and stearic acids. The two SNPs were in high LD (r^2 = 0.91) and 130 kb apart. The candidate gene *Glyma.05g012300* was located at the middle of the two SNPs (Supplemental Figure 12A). It encodes an elongated fatty acyl-ACP thioesterase B (FATB) with an acyl-ATP thioesterase domain at the N terminus, which differs from the reported FATB1a and FATB1b. In soybean, four isoforms of FATB (FATB1a and 1b, FATB2a and 2b) were identified, and FATB1a has been proven to account for the saturated fatty acid content in seed oil (Cardinal et al., 2007). A difference of 0.37% and 0.15% in palmitic and stearic acid concentrations was detected between lines segregating for the leading SNPs, which accounted for 11% and 21% of the genetic variation, respectively (Supplemental Figure 12B and 12C). *SACPD-C* on Gm14 was tagged by *ss715618427* (Supplemental Figure 12D). It was associated with palmitic acid and stearic acid (Supplemental Table 2) and explained 18% and 35% of the genetic variation (Supplemental Figure 12E and 12F). A strong epistatic effect on saturates was detected between the loci of *FATB* and *SACPD-C* (Supplemental Figure 12G). In addition, also on Gm14, the *FAD3A* locus was identified at 32 kb upstream of *ss715619179* that was associated with linolenic acid (Bilyeu et al., 2003, 2005) and explained 7% of the genetic variation (Supplemental Figure 13). Additional candidate genes encoding transcription factors, amino acid transporter, and lipid metabolism-related proteins were identified and detailed information is shown in Supplemental Table 2. The information on the major QTL is presented in Table 1.

Repeatability of the results for different GWAS was relatively low, especially for complicated traits. To confirm the reliability of QTL, we compared our results with similar studies. One of the three lead SNPs associated with protein and three of the 25 lead SNPs for oil identified in this study were identical or in close proximity (<50 kb) to the SNPs reported in previous GWAS (Supplemental Figure 14, Supplemental Table 2). However, more extensive overlaps were found between the loci identified in the present study and those previously detected by linkage mapping studies for protein and oil. Two of the three QTL associated with protein and 20 of the 25 QTL for oil were situated in the chromosomal regions of reported QTL (Supplemental Table 2). In addition, nearly 45% (8 of 18) of the loci associated with fatty acids were located in previously identified QTL (Supplemental Table 2). However, no overlap was found between the amino acid loci identified in the present study and previously reported QTL.

Allelic Analysis Suggests the Impact of Soybean Domestication and Modern Breeding on Seed Quality

The large effects and broad impact of the QTL on seed compounds identified in this study encouraged us to further investigate their allelic changes during soybean domestication and modern breeding. Of the six major-effect QTL, *ss715622170* underwent domestication and *ss715618427* resided in *SACPD-C*. Their association with the traits identified in the present study using cultivated soybean could refer to wild progenitors. A marked impact of selection was observed at the *ss715622170* locus during domestication and modern breeding (Figure 5). The *G* allele at *ss715622170* associated with low protein, high oil, and amino acid profile was rare in wild soybean (MAF <0.05), but was successively and highly selected during domestication and modern breeding, resulting in almost elimination of the alternative allele in the elite cultivars released in North America during the 1990s. The high selection at *ss715622170* during domestication was also detected in other studies (Zhou et al., 2015; Valliyodan et al., 2016), where extensive correlation between reported oil QTL and selection sweeps was observed using diverse germplasm accessions from multiple countries. On the contrary, *ss715618427* in *SACPD*-C associated with palmitic acid did not exhibit a noticeable change between wild soybean and landrace populations but showed an obvious alternation during modern breeding (Figure 5). Loci *ss715591638* and *ss715619179* were also highly selected and the *G* alleles associated with high

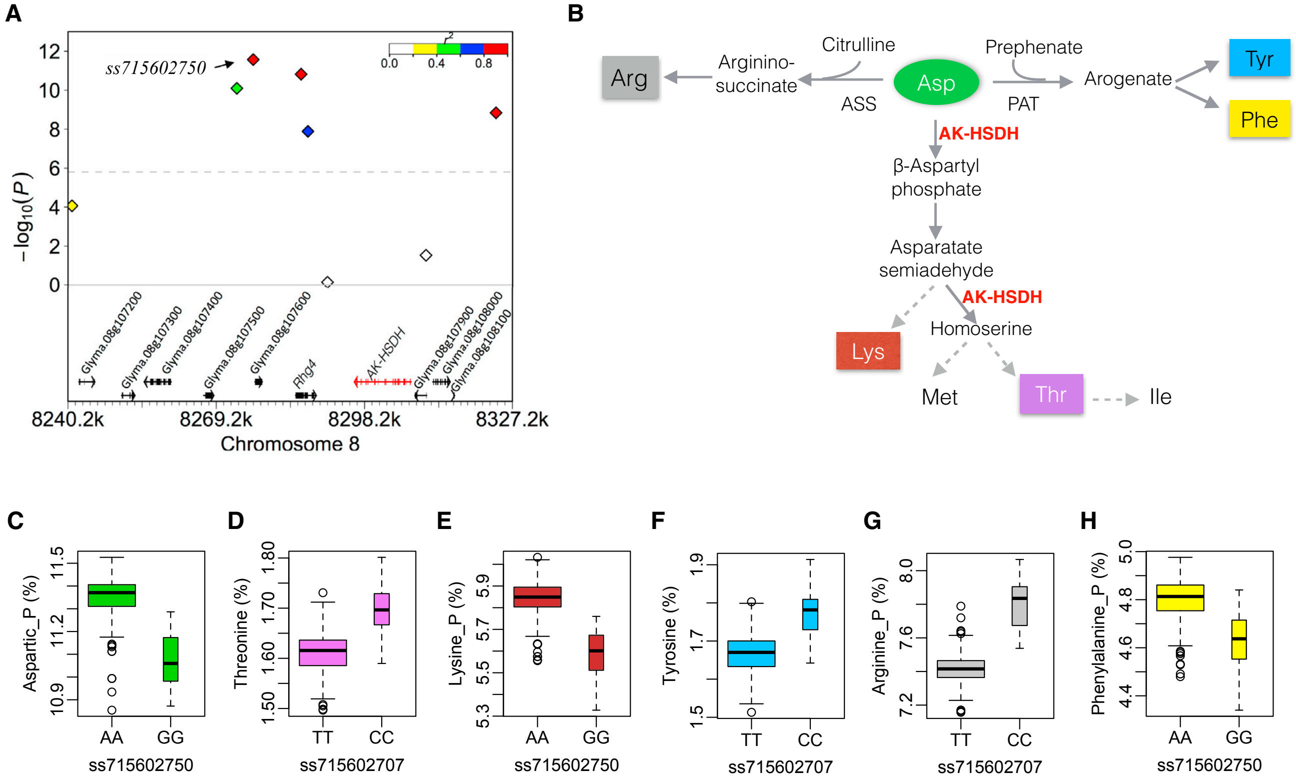

Figure 4. Candidate Gene, Aspartate-Related Amino Acid Biosynthesis Pathways, and the Difference in the Genotypic Values of Multiple Amino Acids Associated with the Major-Effect Locus on Chromosome 8.

(A) Regional Manhattan plot with candidate gene *AK-HSDH* of the QTL led by *ss715602750* associated with aspartate content based on protein.

(B) Aspartate-related amino acid biosynthesis pathways in plant. The dashed arrows indicate multi-step reactions and the solid arrows indicate single-step reactions. ASS, argininosuccinate synthetase; PAT, prephenate aminotransferase.

(C–H) The difference in genotypic values of the aspartate-related amino acid concentration based on seed dry weight or on protein (_P) associated with the segregation of *AK-HSDH* locus that was tagged by SNP *ss715602750* and *ss715602707* in LD (r^2 = 0.65). **(C)** aspartate; **(D)** threonine; **(E)** lysine; **(F)** tyrosine; **(G)** arginine; **(H)** phenylalanine. The *P* values for *ss715602707* associated with threonine **(D)** (6.54×10^{-6}) and tyrosine **(F)** (1.37×10^{-5}) were slightly lower than the significance level of Bonferroni correction ($P < 1.57 \times 10^{-6}$) but were significant with the false discovery rate ($q < 0.05$).

oil or low linolenic acid were fixed or prevalent in the elite cultivars released in the United States (US) (Supplemental Figure 15). However, no selection was observed at *ss715602750* during modern breeding, which was associated with the content of total oil, oleic acid, and several amino acids. Notably, the ratio of CC genotype at *ss7155992503* associated with low saturates was substantially increased in the US elite cultivars compared with that in Asian landraces (Supplemental Figure 15). These results suggest that soybean domestication and modern breeding had an extensive influence on the seed composition, especially the seed oil and protein concentration, although selection for yield-related traits was the primary consideration. It also implied great potential for improvement of seed nutrients such as saturated fatty acids in modern elite cultivars.

During the past 80 years of soybean breeding in North America, a steady increase in yield of 29 kg ha^{-1} yr^{-1} was achieved (Rincker et al., 2014). Subsequently, this also resulted in modern cultivars with lower seed protein and higher oil concentration (Rincker et al., 2014). To investigate the possible effect of the major QTL associated with quality traits identified in this study on soybean yield, their physical positions were projected to the Soybean-GmComposite2003 linkage map (SoyBase http://soybase.org/) to identify overlap with known yield QTL. The results showed that *ss715622170* and *ss715591638*, associated with protein and oil and highly selected for the high oil allele during modern breeding (Figure 5 and Supplemental Figure 15), were located at similar regions of the yield QTL reported previously (Kabelka et al., 2004; Li et al., 2008) (Supplemental Table 3). No overlap with yield QTL was observed for other major-effect loci.

Geographic Distribution of the Major-Effect Alleles Associated with Seed Composition

Geographic analysis was performed to obtain detailed information about the distribution and transition of the allele associated with seed nutrients during soybean domestication and breeding selection. A scan of the USDA Soybean Germplasm Collection for *ss715622170* identified 30 (26 homozygotes and four heterozygotes) of the 996 *G.soja* (3.0%) carrying *G* at this locus. They originated from South Korea (6/30) and Japan (24/30) (Figure 6A). An investigation using the previous soybean resequencing dataset (Zhou et al., 2015) identified two additional wild accessions carrying heterozygous *ss715622170*

Lead SNP	Chromosome	Position	MAF	Traits associated	Candidate gene	Annotation
ss715592503	5	1049939	0.42	Palmitic and Stearic acid	*Glyma.05g012300*	Elongated FATB with an acyl-ATP thioesterase domain
ss715591638	5	41883826	0.11	Oil	*Glyma.05g245000*	3-Oxo-5-α-steroid 4-dehydrogenase
ss715602750	8	8276381	0.09	Oil, oleic, Gly, Ser, Trp, Arg_P, Asp_P, Gly_P, Lys_P, Phe_P, and Trp_P	*AK-HSDH*	Aspartate-kinase–homoserine-dehydrogenase (Zhu-Shimoni and Galili, 1998)
ss715618427	14	17499955	0.33	Palmitic and stearic acid	*SACPD-C*	Stearoyl-ACP desaturase isoform C (Gillman et al., 2014)
ss715619179	14	45971865	0.34	Linolenic acid	*FAD3A*	Omega-3 fatty acid desaturase A (Bilyeu et al., 2003; Bilyeu et al., 2005)
ss715622170	15	4685574	0.33	Protein, oil, linolenic, Arg, Asp, Cys, Glu, His, Ile, Leu, Lys, Met, Phe, Trp, Ala_P, Arg_P, Glu_P, Pro_P, Thr_P, and Val_P	*SAT1*, *Glyma.15g049700-Glyma.15g049900*	Nodule growth and NH_4^+ transport (Kaiser et al., 1998; Chiasson et al., 2014), Nodulin MtN21 (Ladwig et al., 2012)

Table 1. Major-Effect QTL Associated with Seed Composition and Their Candidate Genes.

in Zhejiang Province, South China (Figure 6A). The *G* allele associated with low protein and high oil was then selected and spread to the entire soybean growing region in East Asian (Figure 6B). In the US, however, the *AA* genotype at this locus existed mainly in the modern cultivars from southeastern regions, but was rarely observed in other regions except Minnesota (Figure 6C). This indicated that the selection of *GG* in the southeastern regions was not as strict as in other regions during modern breeding in North America. Different from *ss715622170*, *ss715618427* tagging *SACPD-C* did not undergo domestication selection (Figure 6D and 6E); the genotype *AA* associated with high saturates was common in Northern China, but was rare in South Korea and Japan (Figure 6E). In the US, both genotypes *AA* and *CC* were common in the elite cultivars across the entire soybean growing region (Figure 6F).

For the other major-effect loci associated with the seed compositional traits, geographic changes in the allele frequency were also evidently found during modern breeding. For instance, the genotype *TT* of *ss715591638* related to low oil was nearly eliminated in the North American elite cultivars and the landraces in Northeast China, Korea, and Japan, but prevalent in South China (Supplemental Figure 16A and 16B). In contract, both homozygous genotypes at the locus *ss715619179* harboring *FAD3A* were spread across the regions of East Asia (Supplemental Figure 16C), but the *AA* genotype associated with high polyunsaturated fatty acids predominantly existed in five Midwest states in the US (Supplemental Figure 16D). For the loci *ss715602750* and *ss715592503*, a distinctive geographic distribution pattern was observed in the US cultivars. Both loci exhibited differential selection between the Northern and Southern breeding programs in the US (Supplemental Figure 16E–16H).

DISCUSSION

Development of cultivars high in both seed protein and oil content is one of the most challenging tasks in soybean improvement because of their negative correlation. Based on the variation and the protein/oil transfer ratio observed in the present dataset, our results suggest that cultivars with 20%–22% oil and 41%–43% protein are achievable. In other words, within a certain range, simultaneously increasing both protein and oil should be possible. Further increase of either oil or protein beyond these ranges would be at the cost of the other dramatically. Improvement of total protein leads to an increased amount of overall amino acids, which has a complicated effect on the amino acid profile as indicated by the broad positive genetic correlations but intricate relationships when corrected for protein. However, genetically modifying the proportion of certain amino acids in soybean protein may produce a broad effect on the seed nutrient profile. Some positive genetic correlations on a seed dry basis became negative after adjusting the amino acid content with total protein content, such as Cys and Glu, which were negatively correlated with most of the remaining protein-adjusted amino acids (Supplemental Figure 5). Some significant correlations between amino acids before correcting for protein were insignificant after correction, and vice versa. The results implied different genetic architectures for amino acids with and without protein-based adjustment in soybean.

Instead of fitting a single locus at a time in MLM, MLMM fits multiple loci through stepwise regression, which makes MLMM more powerful to detect QTL with a small effect for complex traits (Segura et al., 2012). In this study, both regular MLM and MLMM were applied for association analysis. For most of the traits investigated, the major QTL were consistently detected by both methods, and MLMM revealed additional loci after conditioning loci of larger effect. Compared with other traits, oil presented more discrepancy in the results between MLM and MLMM. Notably, loci with a small effect accounted for the primary divergence, and most of the oil QTL detected were located at oil QTL reported previously, indicating the complementarity of MLM and MLMM.

The identification of the locus led by *ss715622170* and *ss715621777* on Gm15 may shed light on the genetic

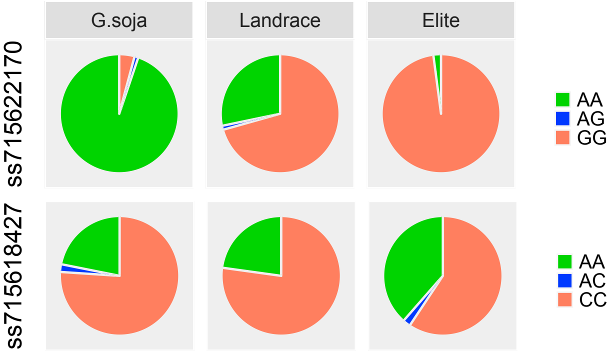

Figure 5. Allele Frequency of the Lead SNPs *ss715622170* and *ss715618427* of the Major-Effect QTL Associated with Seed Composition in Wild Soybean (*Glycine soja*), Asian Landrace, and Elite Cultivars.

A total of 96, 92, and 96 accessions of *G. soja*, Asian landrace, and elite cultivars released in the United States, respectively, were involved in the analysis. The panels were used in a previous study (Song et al., 2013) and each had broad genetic diversity.

mechanism underlying the negative correlation between soybean seed protein and oil. This locus associated with protein and oil, as well as other 18 seed compounds. The candidate genes *GmSAT1* and the putative nodulin *MtN21* are essential for nodule development, effective symbiotic N_2 fixation, and NH_4^+ transport (Kaiser et al., 1998; Ladwig et al., 2012; Chiasson et al., 2014). Symbiotic N_2 fixation is the major contributor to seed protein (Imsande, 1992; Leffel et al., 1992; Fabre and Planchon, 2000; Salvagiotti et al., 2008), while the content of soybean seed protein is related to the N:C ratio (Allen and Young, 2013). The efficiency of biological N_2 fixation conditioned by *GmSAT1* may alter the N:C ratio and thus have an impact on the concentration of protein and amino acids in soybean seed as well. The large effect of *ss715621777* or similar region on protein and oil was also reported in a recent GWAS using the entire or a subgroup of the USDA Soybean Germplasm Collection with dataset on GRIN (Vaughn et al., 2014; Bandillo et al., 2015). Using backcross-derived lines, Kim et al. (2016) fine mapped this QTL to a 535 kb region from BARCSOYSSR_15_0161 at 3 587 104 to BARCSOYSSR_15_0194 at 4 122 592 on Gm15, which partially overlaps the range determined by *ss715622170* and *ss715621777*. The large effect and the stability of this locus across environments and genetic backgrounds indicated its importance and potential use in soybean quality improvement.

However, alteration of the oil concentration in soybean seed attributed to the locus led by *ss715622170* may relate to carbon availability. Studies in the model organism *Chlamydomonas* have shown that starch is the preferred form of carbon and energy storage rather than triacylglycerol (TAG), the major storage compound of oil (Fan et al., 2012). Substantial TAG accumulation only occurs when starch biosynthesis or N metabolism is blocked or is saturated by the available carbon source (Gaude et al., 2007; Fan et al., 2012). Nodule development and the maintenance of morphological structure and function are carbon and energy costly (Walsh et al., 1987). The carbon source and energy required for these processes are provided by current photosynthesis in the form of sucrose during daylight or by the starch pool in the leaf in the dark (Kouchi et al., 1986). Thus, efficient symbiotic N_2 fixation may restrict carbon availability in at least two ways: (1) by increasing the use of sucrose/starch as demanded by nodule growth and function maintenance; and (2) by increasing the demand for carbon in N metabolism. Both limit carbon flux toward oil biosynthesis.

Consistent with previous reports (Zhou et al., 2015; Valliyodan et al., 2016), the present study demonstrated that *ss715622170* underwent domestication selection. This locus was associated with protein, oil, fatty acids, and amino acids, and thus it may serve as an excellent case for further exploring the impact of domestication on soybean quality. In addition, investigation into the geographic distribution of the rare allele of *ss715622170* provided insight into the origin center of cultivated soybean. In the USDA Germplasm Collection, none of the wild accessions carrying the rare allele of *ss715622170*, which commonly existed in cultivated soybeans, was located in China. This appears to be in contradiction to the general idea that China is the origin of cultivated soybean (Fukuda, 1933; Hymowitz and Newell, 1981). Notably, the USDA Germplasm Collection lacks coverage of wild soybean from South China, which was documented as one of the origin centers of cultivated soybean (Gai et al., 1999; Guo et al., 2010). Identification of the two wild accessions carrying the domesticated allele of *ss715622170* originated from Zhejiang Province (South China) by investigating a previously re-sequenced panel (Zhou et al., 2015) supports the hypothesis of South China as an origin center of soybean.

We noticed that the amino acid QTL, *ss715602750*, on Gm08 harboring *AK-HSDH* was not associated with protein concentration. A similar result was also observed in a recent study, in which this region was tagged by *Gm08_8462762* (also known as *ss715602763*) and was identified to be associated with Cys, Lys, and Thr but not with total protein (Vaughn et al., 2014). It is highly consistent with the underlying regulation mechanisms of *AK-HSDH.* In *Arabidopsis*, the expression of *AK-HSDH* is induced by the photosynthesis-related metabolite sucrose; while nitrogen, which is important for seed protein accumulation, has no effect on the expression of *AK-HSDH* (Zhu-Shimoni and Galili, 1998). During the day, photosynthesis-produced sucrose stimulates *AK-HSDH* expression and facilities the utilization of this carbon source for the biosynthesis of Asp-related amino acids. However, as discussed above, the synthesis of fatty acids is sensitive to carbon source availability. Interestingly, strong associations of this locus with oil and oleic acid were also detected in this study (Table 1). The independence of this locus from total protein makes it an ideal target for the modification of the amino acid profile in soybean seed, and thus useful for soybean meal improvement as well. Further analysis indicated that the miscoding mutation at the 3rd exon or mutation(s) at the promoter of *AK-HSDH* might be the causal genetic variant. However, introgression of the favorable *AK-HSDH* into desired genotypes might be problematic because of the tight linkage with the soybean cyst nematode resistance gene *Rhg4*, if they are in repulsion linkage phase. Genome editing that directly modifies the target genes could be a potential tool in solving the problem.

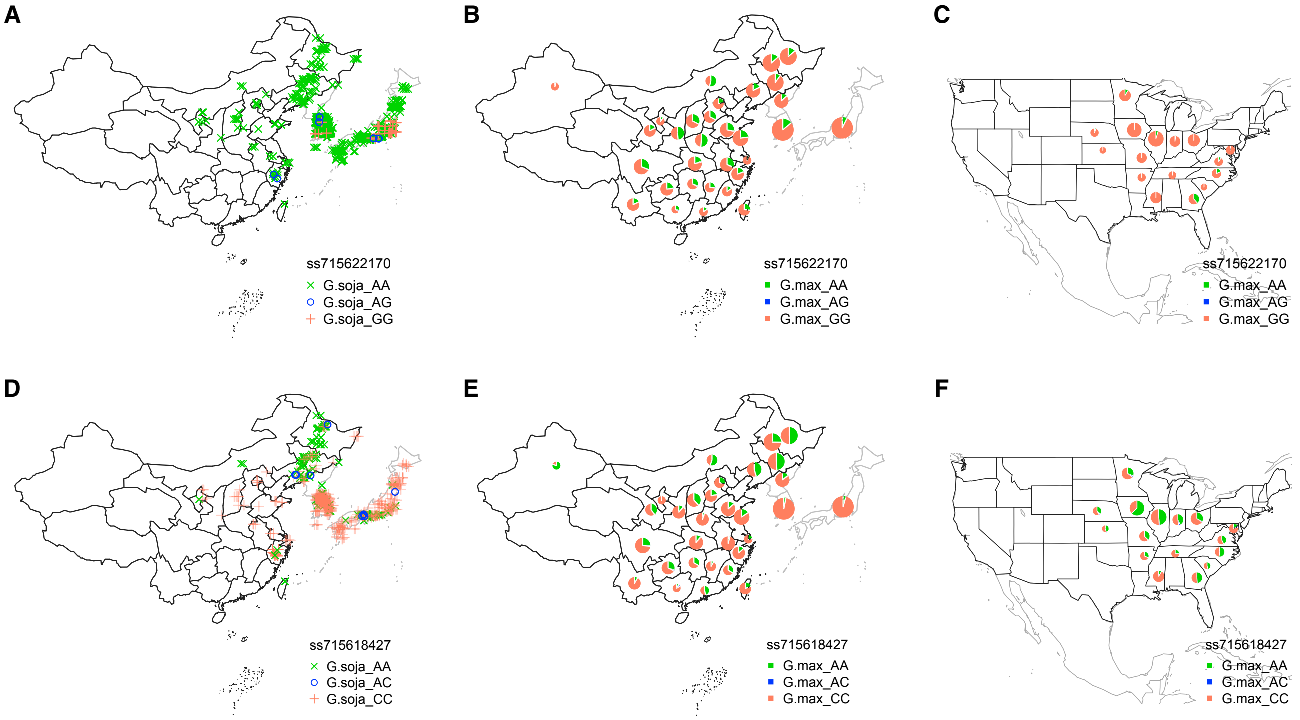

Figure 6. Origin and Geographic Distribution of the Alleles at *ss715622170* Related to Domestication and *ss715618427* in *SACPD-C* Associated with Soybean Seed Composition.

Shown are the allele composition and distribution of related loci: **(A and D)** 712 of 758 *G. soja* accessions in the USDA Germplasm Collection that have geo-location information and originated from China, South Korea, and Japan, and 12 wild progenitors from a previous study (Zhou et al., 2015) that originated from Zhejiang Prince, China, which have no detailed geo-location information and are plotted around the center of Zhejiang Province. **(B and E)** 11 577 soybean landraces (*G. max*) across China, North and South Korea, and Japan; and **(C and F)** 860 modern cultivars across the US based on their origin/released location. Only the provinces/states that had more than 20 landraces or 10 modern cultivars are plotted. The radius of the pie diagram indicates one half of the $\log_{10}$-transformed number of lines originating from the region.

The concentrations of seed protein and amino acids are naturally highly correlated because protein consists of amino acids. In previous studies on the genetic architecture of seed amino acids, commonly used measurements were based on seed weight and thus the impact of protein could not be eliminated in discovering genetic variants specific for the amino acid profile, the quality determinant of soybean meal (Panthee et al., 2006b; Carlson, 2011; Vaughn et al., 2014). From an applied perspective, the relative amino acid composition in soybean protein meal is more of interest to producers and animal feed industries. In the present study, the genetic basis of amino acids was also investigated using the estimates related to protein concentration, in addition to dry-weight-based measurements. We attempted to see if there were differences between the normalized with protein content and non-normalized amino acids. The results indicated that of the 92 QTL associated with amino acids, only four were associated with both DW- and P-based amino acids. The major-effect QTL harboring *AK-HSDH* was identified to be associated with both Gly and Gly_P as well as Try and Try_P. However, it was not associated with Arg, Asp, Lys, and Phe, but associated with Arg_P, Asp_P, Lys_P, and Phe_P (Supplemental Table 2). Therefore, it could be reasoned that there is an independent genetic control of amino acids from total protein. It also suggests that the amino acid profile could be improved without changing the total protein content, which may incongruously affect soybean yield.

Yield is always a paramount trait for breeding in soybean as well as other crops. However, cloning yield genes is challenging because it is determined by various related traits, and it is hard to identify the candidate genes. Given the relationship between seed yield and the seed protein and oil concentration in soybean, identification of causal genes for seed protein and oil might bring a breakthrough in cloning yield genes. Similar localization of the previously identified yield QTL and the major-effect loci, *ss715622170* and *ss715591638*, associated with protein and oil detected in the present study implies promising candidate genomic regions for genetic manipulation of soybean yield. On the other hand, the lack of overlap between the major-effect loci associated with fatty acids and amino acids identified in this study and the yield QTL reported previously suggests great potential in soybean seed nutrient improvement without sacrificing yield. In addition, given that the germplasm accessions of the GWAS panel in this study were all sampled from the early maturity groups, it would be of interest to see if the seed protein and oil contents correlate with days to maturity in the panel. A computation indicated that the correlation between protein or oil and days to maturity was very weak and not significant, with a correlation coefficient of 0.001 (P = 0.98) or −0.013 (P = 0.82). This suggested that seed protein and oil contents might not be genetically affected by days to maturity.

Artificial selection had an extensive impact on the performance of traits during crop domestication and modern breeding. In this study, we explored the molecular evidence underlying trait evolution in soybean through analysis of allelic alternation and geographic distribution of the major-effect loci associated with seed composition. The results indicated that soybean domestication and modern breeding produced a broad influence on seed composition, particularly on the seed oil and protein concentration. Moreover, evolutionary changes of alleles at these loci were also relevant to geographic regions. This phenomenon might be attributed to different ecological conditions and cropping management as well as region-focused breeding programs. The geographic distribution of the favorable alleles sheds light on the potential of soybean improvement for regional foci.

In summary, this study is of significance for deciphering the mechanism underlying the complex relationship among seed compounds, especially the inverse correlation between protein and oil. It reveals the genetic basis of quality alterations that occurred during soybean domestication and the long-term yield-driven breeding in North America. This study also provides insights into the manipulation of seed compounds to meet the requirements of soybean quality for diverse end uses through marker-assisted selection or genome-assisted breeding. The loci or candidate genes identified in this study and confirmed in other studies would have potential uses in genetic improvement of soybean. Further validation of the causal genetic variants at the major-effect loci associated with soybean quality and their relationship with yield is of great interest.

METHODS

Plant Materials and Field Trials

A population of 321 *G. max* plant introductions (PIs) (one with green cotyledon) was used in this study. The PIs were randomly selected from the USDA Soybean Germplasm Collection but limited to early maturity group 0 and 00 given that soybean is highly photoperiod-sensitive and a single experiment could not accommodate a wide range of maturities. The population was planted in a randomized complete block design with three replications at three locations: Aurora (2011), Brookings (2012) and Watertown (2012), South Dakota. According to the GRIN (http://www.ars-grin.gov/), 92% of the PIs are MG 0 and 8% MG 00, and 91% originated from China. The field experimental design and plot information were described in previous reports (Zhang et al., 2015a, 2015b).

Phenotypic Evaluation and Statistical Analysis

All the plots were bulk harvested individually after full maturity (R8 stage), and then the seeds were dried in an air drier. Soybean seeds were milled with a Perten Laboratory Mill 3600. The concentrations of seed protein, oil, five fatty acids, and 18 amino acids were estimated by near-infrared reflectance (NIR) spectroscopy DA-7200 (Perten Instruments, Sweden) using a ground sample on a DW basis (0% moisture), except fatty acids, which were determined as a percentage relative to the total oil content. In addition, the percentages of amino acids were also calculated based on protein. The NIR calibrations for ground soybean samples were developed by Perten Instruments and updated yearly. The 2011 calibrations for ground soybean samples, which were based on the analysis of 915–3641 samples depending on the seed component, were used. The correlation coefficients between the NIR estimates and standard analysis data ranged from 0.77 for stearic acid to 0.97 for protein, with an average of 0.90 over all the 25 seed compositions determined in this study (Supplemental Table 4). The model for the phenotypic trait and the calculation of entry-mean-based heritability were described in previous studies (Zhang et al., 2015a, 2015b). The comparison of phenotypic variation among traits was conducted using the range of standardized phenotypic values that were calculated by dividing the trait phenotypic values with the trait mean.

Genotyping, Quality Control and Genome-wide LD

The SNP data of the association panel using the *Glyma.Wm82.a2* soybean reference genome was retrieved from SoyBase (http://soybase.org/), prepared by Song et al. (2013, 2015) using the Illumina Infinium SoySNP50K BeadChip, which contains 52 041 SNPs that were chosen from 209 903 SNPs identified throughout both euchromatic and heterochromatic regions of the soybean genome by applying multiple filters and selections. However, eight of the 321 PIs phenotyped had not been genotyped, and thus only 313 PIs were used in GWAS. Quality control and LD estimation were performed as described previously (Zhang et al., 2015b). Finally, a total of 31 850 SNPs with MAF $\geq$5% were used for association analyses.

Genome-wide Association Analysis

The best linear unbiased predictors (BLUPs) of the genotypic values for each trait were calculated as described previously (Zhang et al., 2015b) and were used for association analyses with both the single-locus MLM (Yu et al., 2006) and the MLMM (Segura et al., 2012). The association analysis with single-locus MLM was implemented in GAPIT (Zhang et al., 2010; Lipka et al., 2012), and the association analysis in MLMM was conducted in R (https://cynin.gmi.oeaw.ac.at/home/resources/mlmm). Both Kinship and principal component analyses were based on the entire set of SNPs. The first two PCs were used to capture the population structure as suggested by the BIC model fitness test and exhibited three groups. The STRUCTURE was further employed to determine the three subgroups (K = 3) by using a randomly selected set of 1000 SNPs from the genome-wide SNPs with 10 runs. Each run consisted of a burn-in period of 15 000 steps followed by 20 000 MCMC repeats. Population differentiation (F_{st}) was calculated using the R package snpStats weighted by the number of chromosomes in each group (Clayton and Clayton, 2012). Bonferroni correction was used to identify significant association ($P < 1.57 \times 10^{-6}$). The lead SNP was chosen to represent the QTL, and the nearby significant SNPs identified by using MLM with LD $r^2 \geq 0.60$. The BLUPs of trait genotypic values were also used to fit the general linear model containing all the QTL identified for a trait and to estimate the proportion of genetic variance attributed to all the identified loci.

QTL Alignment and Prediction of Candidate Genes

Most of the reported QTL had been mapped by SSR markers using biparental segregating populations. Therefore, we transferred each locus into flanked SSR markers of the peak SNPs on the Consensus 4.0 sequence order (SoyBase; http://soybase.org/). The flanked markers were used for loci alignment analysis. Genes annotated in Glyma1.0, Glyma1.1, and NCBI RefSeq gene models in SoyBase (www.soybase.org) were used as the source of identifying candidate genes.

Allelic Analysis of Major-Effect QTL Associated with Seed Composition

The SNP dataset of the 96 wild soybeans (*G. soja*), 92 landrace lines, and 96 elite cultivars, which represent a wide genetic diversity of soybean wild progenitors, landraces, and modern cultivars (Song et al., 2013), was retrieved from SoyBase (http://soybase.org/). Allele frequencies of the six major-effect QTL associated with seed composition were computed for each of the panels.

Origin and Geographic Distribution of Major-Effect Alleles Associated with Seed Composition

A total of 13 195 accessions, a broad representative sample of the wild and cultivated soybeans, were used in the analysis. They included all

758 *G. soja* accessions originating from China (158), South Korea (307), and Japan (293); 11 577 *G. max* accessions originating from China (5096), North Korea (224), South Korea (3380) and Japan (2877); and 860 US modern cultivars (after removing the isogeneic lines), stored in the USDA Soybean Germplasm Collection and genotyped with the SoySNP50K BeadChip (Song et al., 2015). The genotypic data for all the accessions was retrieved from SoyBase (http://soybase.org/) as described above. The origin information was obtained from GRIN (http://www.ars-grin.gov/). Maps of related regions were created using the maps R package (Becker et al., 2013). For a rare allele locus identified, phylogenetic analysis was performed based on 21 SNPs (*ss715622170* and 10 adjacent SNPs on each side) representing the chromosomal region of 4.53–4.93 Mb on Gm15 by using the R package ape (Paradis et al., 2004).

SUPPLEMENTAL INFORMATION

Supplemental Information is available at *Molecular Plant Online*.

FUNDING

This work was supported in part by the USDA-NIFA/CSREES Hatch project and Even-Allen project, and grants from the United Soybean Board (USB) and the South Dakota Soybean Research and Promotion Council (SDSRPC) awarded to G.-L. Jiang.

AUTHOR CONTRIBUTIONS

J.Z., X.W., and G.-L.J. designed the research experiments. J.Z., X.W. and S.J.B. implemented the field experiments. J.Z., X.W., and Y.L. performed phenotyping. Q.J. and P.B.C. performed genotyping. Y.Y. and M.B. contributed to the SDSRPC project proposal. J.Z. analyzed data. G.-L.J. designed the overall project. J.Z., X.W., and G.-L.J. wrote the manuscript. All authors reviewed the manuscript.

ACKNOWLEDGMENTS

The authors are grateful to Nicholas Hall, Marci Green, Sivananda Tirumalaraju, and Mukhtar Agoub for assistance in implementing the field experiments. The authors also thank Dr. Randall Nelson, the curator of the USDA-ARS Germplasm Collection, for providing germplasm accessions, and Dr. Richard Joost, Director of Supply Programs, the United Soybean Board (USB), for reviewing the manuscript. No conflict of interest declared.

Received: July 24, 2017
Revised: December 20, 2017
Accepted: December 24, 2017
Published: January 1, 2018

REFERENCES

Aghoram, K., Wilson, R.E., Burton, J.W., and Dewey, R.E. (2006). A mutation in a 3-keto-acyl-ACP synthase II gene is associated with elevated palmitic acid levels in soybean seeds. Crop Sci. **46**:2453–2459.

Allen, D.K., and Young, J.D. (2013). Carbon and nitrogen provisions alter the metabolic flux in developing soybean embryos. Plant Physiol. **161**:1458–1475.

Andreu, V., Lagunas, B., Collados, R., Picorel, R., and Alfonso, M. (2010). The GmFAD7 gene family from soybean: identification of novel genes and tissue-specific conformations of the FAD7 enzyme involved in desaturase activity. J. Exp. Bot. **61**:3371–3384.

Bandillo, N., Jarquin, D., Song, Q.J., Nelson, R., Cregan, P., Specht, J., and Lorenz, A. (2015). A population structure and genome-wide association analysis on the USDA soybean germplasm collection. Plant Genome **8** https://doi.org/10.3835/plantgenome2015.04.0024.

Becker, R.A., Wilks, A.R., Brownrigg, R., and Minka, T.P. (2013). Maps: draw geographical maps. R package version 2. https://cran.r-project.org/web/packages/maps/maps.pdf.

Bilyeu, K., Palavalli, L., Sleper, D., and Beuselinck, P. (2005). Mutations in soybean microsomal omega-3 fatty acid desaturase genes reduce linolenic acid concentration in soybean seeds. Crop Sci. **45**:1830–1836.

Bilyeu, K.D., Palavalli, L., Sleper, D.A., and Beuselinck, P.R. (2003). Three microsomal omega-3 fatty-acid desaturase genes contribute to soybean linolenic acid levels. Crop Sci. **43**:1833–1838.

Bolon, Y.T., Joseph, B., Cannon, S.B., Graham, M.A., Diers, B.W., Farmer, A.D., May, G.D., Muehlbauer, G.J., Specht, J.E., Tu, Z.J., et al. (2010). Complementary genetic and genomic approaches help characterize the linkage group I seed protein QTL in soybean. BMC Plant Biol. **10**:41.

Byfield, G.E., Xue, H., and Upchurch, R.G. (2006). Two genes from soybean encoding soluble Delta 9 stearoyl-ACP desaturases. Crop Sci. **46**:840–846.

Cardinal, A.J., Burton, J.W., Camacho-Roger, A.M., Yang, J.H., Wilson, R.F., and Dewey, R.E. (2007). Molecular analysis of soybean lines with low palmitic acid content in the seed oil. Crop Sci. **47**:304–310.

Carlson, C.M. (2011). Genetic control of protein and amino acid content in soybean determined in two genetically connected populations. PhD thesis, North Carolina State University, Raleigh, NC.

Carter, A.M., and Tegeder, M. (2016). Increasing nitrogen fixation and seed development in soybean requires complex adjustments of nodule nitrogen metabolism and partitioning processes. Curr. Biol. **26**:2044–2051.

Chiasson, D.M., Loughlin, P.C., Mazurkiewicz, D., Mohammadidehcheshmeh, M., Fedorova, E.E., Okamoto, M., McLean, E., Glass, A.D., Smith, S.E., and Bisseling, T. (2014). Soybean SAT1 (Symbiotic Ammonium Transporter 1) encodes a bHLH transcription factor involved in nodule growth and NH4+ transport. Proc. Natl. Acad. Sci. USA **111**:4814–4819.

Chung, J., Babka, H.L., Graef, G.L., Staswick, P.E., Lee, D.J., Cregan, P.B., Shoemaker, R.C., and Specht, J.E. (2003). The seed protein, oil, and yield QTL on soybean linkage group I. Crop Sci. **43**:1053–1067.

Clayton, D., and Clayton, M.D. (2012). Package 'snpStats'. http://bioconductor.org/packages/snpStats/.

Diers, B.W., Keim, P., Fehr, W.R., and Shoemaker, R.C. (1992). RFLP analysis of soybean seed protein and oil content. Theor. Appl. Genet. **83**:608–612.

Fabre, F., and Planchon, C. (2000). Nitrogen nutrition, yield and protein content in soybean. Plant Sci. **152**:51–58.

Falco, S.C., Guida, T., Locke, M., Mauvais, J., Sanders, C., Ward, R.T., and Webber, P. (1995). Transgenic canola and soybean seeds with increased lysine. Biotechnology (N. Y.) **13**:577–582.

Fallen, B.D., Hatcher, C.N., Allen, F.L., Kopsell, D.A., Saxton, A.M., Chen, P., Kantartzi, S.K., Cregan, P.B., Hyten, D.L., and Pantalone, V.R. (2013). Soybean seed amino acid content QTL detected using the universal Soy linkage panel 1.0 with 1,536 SNPs. J. Plant Genome Sci. **1**:68–79.

Fan, J.L., Yan, C.S., Andre, C., Shanklin, J., Schwender, J., and Xu, C.C. (2012). Oil accumulation is controlled by carbon precursor supply for fatty acid synthesis in *Chlamydomonas reinhardtii*. Plant Cell Physiol. **53**:1380–1390.

Fukuda, Y. (1933). Cytogenetical studies on the wild and cultivated Manchurian soybeans (*Glycine* L.). Jpn. J. Bot. **6**:489–506.

Gai, J., Xu, D., Gao, Z., Shimamoto, Y., Abe, J., Fukushi, H., and Kitajima, S. (1999). Studies on the evolutionary relationship among eco-types of *G. max* and *G. soja* in China. Zuo wu xue bao **26**:513–520.

Gaude, N., Brehelin, C., Tischendorf, G., Kessler, F., and Dormann, P. (2007). Nitrogen deficiency in *Arabidopsis* affects galactolipid composition and gene expression and results in accumulation of fatty acid phytyl esters. Plant J. **49**:729–739.

Gillman, J.D., Stacey, M.G., Cui, Y., Berg, H.R., and Stacey, G. (2014). Deletions of the SACPD-C locus elevate seed stearic acid levels but also result in fatty acid and morphological alterations in nitrogen fixing nodules. BMC Plant Biol. **14**:143.

Grundy, S.M. (1986). Comparison of monounsaturated fatty acids and carbohydrates for lowering plasma cholesterol. N. Engl. J. Med. **314**:745–748.

Guo, J., Wang, Y., Song, C., Zhou, J., Qiu, L., Huang, H., and Wang, Y. (2010). A single origin and moderate bottleneck during domestication of soybean (*Glycine max*): implications from microsatellites and nucleotide sequences. Ann. Bot. **106**:505–514.

Han, Y., Zhao, X., Liu, D., Li, Y., Lightfoot, D.A., Yang, Z., Zhao, L., Zhou, G., Wang, Z., Huang, L., et al. (2016). Domestication footprints anchor genomic regions of agronomic importance in soybeans. New Phytol. **209**:871–884.

Heppard, E.P., Kinney, A.J., Stecca, K.L., and Miao, G.H. (1996). Developmental and growth temperature regulation of two different microsomal omega-6 desaturase genes in soybeans. Plant Physiol. **110**:311–319.

Hitz, W., and Yadav, N. (1992). Nucleotide sequences of soybean acyl-ACP thioesterase genes. International patent no. WO1992011373.

Huang, X.H., Wei, X.H., Sang, T., Zhao, Q.A., Feng, Q., Zhao, Y., Li, C.Y., Zhu, C.R., Lu, T.T., Zhang, Z.W., et al. (2010). Genome-wide association studies of 14 agronomic traits in rice landraces. Nat. Genet. **42**:961–U976.

Hwang, E.Y., Song, Q.J., Jia, G.F., Specht, J.E., Hyten, D.L., Costa, J., and Cregan, P.B. (2014). A genome-wide association study of seed protein and oil content in soybean. BMC Genomics **15**:1.

Hymowitz, T., and Newell, C. (1981). Taxonomy of the genus *Glycine*, domestication and uses of soybeans. Econ. Bot. **35**:272–288.

Imsande, J. (1992). Agronomic characteristics that identify high-yield, high protein soybean genotypes. Agron. J. **84**:409–414.

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

Kaiser, B.N., Finnegan, P.M., Tyerman, S.D., Whitehead, L.F., Bergersen, F.J., Day, D.A., and Udvardi, M.K. (1998). Characterization of an ammonium transport protein from the peribacteroid membrane of soybean nodules. Science **281**:1202–1206.

Kim, M., Schultz, S., Nelson, R.L., and Diers, B.W. (2016). Identification and fine mapping of a soybean seed protein QTL from PI 407788A on chromosome 15. Crop Sci. **56**:219–225.

Kouchi, H., Akao, S., and Yoneyama, T. (1986). Respiratory utilization of ^{13}C-labelled photosynthate in nodulated root systems of soybean plants. J. Exp. Bot. **37**:985–993.

Krishnan, H.B. (2005). Engineering soybean for enhanced sulfur amino acid content. Crop Sci. **45**:454–461.

Ladwig, F., Stahl, M., Ludewig, U., Hirner, A.A., Hammes, U.Z., Stadler, R., Harter, K., and Koch, W. (2012). Siliques are Red1 from *Arabidopsis* acts as a bidirectional amino acid transporter that is crucial for the amino acid homeostasis of siliques. Plant Physiol. **158**:1643–1655.

Lam, H.M., Peng, S.S., and Coruzzi, G.M. (1994). Metabolic regulation of the gene encoding glutamine-dependent asparagine synthetase in *Arabidopsis thaliana*. Plant Physiol. **106**:1347–1357.

Leamy, L.J., Zhang, H.Y., Li, C.B., Chen, C.Y., and Song, B.H. (2017). A genome-wide association study of seed composition traits in wild soybean (*Glycine soja*). BMC Genomics **18**:18.

Leffel, R.C., Cregan, P.B., Bolgiano, A.P., and Thibeau, D.J. (1992). Nitrogen-metabolism of normal and high-seed-protein soybean. Crop Sci. **32**:747–750.

Li, B., Fan, S., Yu, F., Chen, Y., Zhang, S., Han, F., Yan, S., Wang, L., and Sun, J. (2017). High-resolution mapping of QTL for fatty acid composition in soybean using specific-locus amplified fragment sequencing. Theor. Appl. Genet. **130**:1467–1479.

Li, D.D., Pfeiffer, T.W., and Cornelius, P.L. (2008). Soybean QTL for yield and yield components associated with *Glycine soja* alleles. Crop Sci. **48**:571–581.

Lipka, A.E., Tian, F., Wang, Q.S., Peiffer, J., Li, M., and Bradbury, P.J. (2012). GAPIT: genome association and prediction integrated tool. Bioinformatics **28**:2397–2399.

Panthee, D.R., Pantalone, V.R., Sams, C.E., Saxton, A.M., West, D.R., Orf, J.H., and Killam, A.S. (2006a). Quantitative trait loci controlling sulfur containing amino acids, methionine and cysteine, in soybean seeds. Theor. Appl. Genet. **112**:546–553.

Panthee, D.R., Pantalone, V.R., Saxton, A.M., West, D.R., and Sams, C.E. (2006b). Genomic regions associated with amino acid composition in soybean. Mol. Breed. **17**:79–89.

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. Bioinformatics **20**:289–290.

Pham, A.T., Shannon, J.G., and Bilyeu, K.D. (2012). Combinations of mutant FAD2 and FAD3 genes to produce high oleic acid and low linolenic acid soybean oil. Theor. Appl. Genet. **125**:503–515.

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

Qiu, H., Hao, W., Gao, S., Ma, X., Zheng, Y., Meng, F., Fan, X., Wang, Y., Wang, Y., and Wang, S. (2014). Gene mining of sulfur-containing amino acid metabolic enzymes in soybean. Yi Chuan **36**:934–942.

Ramos, M.J., Fernandez, C.M., Casas, A., Rodriguez, L., and Perez, A. (2009). Influence of fatty acid composition of raw materials on biodiesel properties. Bioresour. Technol. **100**:261–268.

Rincker, K., Nelson, R., Specht, J., Sleper, D., Cary, T., Cianzio, S.R., Casteel, S., Conley, S., Chen, P., and Davis, V. (2014). Genetic improvement of US soybean in maturity groups II, III, and IV. Crop Sci. **54**:1419–1432.

Ross, A.J., Fehr, W.R., Welke, G.A., and Cianzio, S.R. (2000). Agronomic and seed traits of 1%-linolenate soybean genotypes. Crop Sci. **40**:383–386.

Salvagiotti, F., Cassman, K.G., Specht, J.E., Walters, D.T., Weiss, A., and Dobermann, A. (2008). Nitrogen uptake, fixation and response to fertilizer N in soybeans: a review. Field Crop Res. **108**:1–13.

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

Sebolt, A.M., Shoemaker, R.C., and Diers, B.W. (2000). Analysis of a quantitative trait locus allele from wild soybean that increases seed protein concentration in soybean. Crop Sci. **40**:1438–1444.

Segura, V., Vilhjalmsson, B.J., Platt, A., Korte, A., Seren, U., Long, Q., and Nordborg, M. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat. Genet. **44**:825–830.

Song, Q.J., Hyten, D.L., Jia, G.F., Quigley, C.V., Fickus, E.W., Nelson, R.L., and Cregan, P.B. (2013). Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. PLoS One **8**:e54985.

Song, Q.J., Hyten, D.L., Jia, G.F., Quigley, C.V., Fickus, E.W., Nelson, R.L., and Cregan, P.B. (2015). Fingerprinting soybean germplasm and its utility in genomic research. G3 (Bethesda) **5**:1999–2006.

US Department of Agriculture Foreign Agricultural Service. (2017). Oilseeds: World Production Markets and Trade Reports (Washington, DC: FAS).

Valliyodan, B., Dan, Q., Patil, G., Zeng, P., Huang, J., Dai, L., Chen, C., Li, Y., Joshi, T., Song, L., et al. (2016). Landscape of genomic diversity and trait discovery in soybean. Sci. Rep. **6**:23598.

Vaughn, J.N., Nelson, R.L., Song, Q., Cregan, P.B., and Li, Z. (2014). The genetic architecture of seed composition in soybean is refined by genome-wide association scans across multiple populations. G3 (Bethesda) **4**:2283–2294.

Walsh, K.B., Vessey, J.K., and Layzell, D.B. (1987). Carbohydrate supply and N2 fixation in soybean the effect of varied daylength and stem girdling. Plant Physiol. **85**:137–144.

Wang, X.Z., Jiang, G.L., Song, Q.J., Cregan, P.B., Scott, R.A., Zhang, J.P., Yen, Y., and Brown, M. (2015). Quantitative trait locus analysis of seed sulfur-containing amino acids in two recombinant inbred line populations of soybean. Euphytica **201**:293–305.

Wardlaw, G., and Snook, J. (1990). Effect of diets high in butter, corn oil, or high-oleic acid sunflower oil on serum lipids and apolipoproteins in men. Am. J. Clin. Nutr. **51**:815–821.

Yu, J.M., Pressoir, G., Briggs, W.H., Bi, I.V., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Holland, J.B., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. **38**:203–208.

Zhang, J., Song, Q., Cregan, P.B., and Jiang, G.L. (2015a). Genome-wide association study, genomic prediction and marker-assisted selection for seed weight in soybean (*Glycine max*). Theor. Appl. Genet. **129**:117–130.

Zhang, J., Song, Q., Cregan, P.B., Nelson, R.L., Wang, X., Wu, J., and Jiang, G.-L. (2015b). Genome-wide association study for flowering time, maturity dates and plant height in early maturing soybean (*Glycine max*) germplasm. BMC Genomics **16**:1–11.

Zhang, P., Burton, J.W., Upchurch, R.G., Whittle, E., Shanklin, J., and Dewey, R.E. (2008). Mutations in a Δ-stearoyl-ACP-desaturase gene are associated with enhanced stearic acid levels in soybean seeds. Crop Sci. **48**:2305–2313.

Zhang, Z.W., Ersoz, E., Lai, C.Q., Todhunter, R.J., Tiwari, H.K., and Gore, M.A. (2010). Mixed linear model approach adapted for genome-wide association studies. Nat. Genet. **42**:355–360.

Zheng, H., Rowland, O., and Kunst, L. (2005). Disruptions of the *Arabidopsis* enoyl-CoA reductase gene reveal an essential role for very-long-chain fatty acid synthesis in cell expansion during plant morphogenesis. Plant Cell **17**:1467–1481.

Zhou, Z., Jiang, Y., Wang, Z., Gou, Z., Lyu, J., Li, W., Yu, Y., Shu, L., Zhao, Y., Ma, Y., et al. (2015). Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nat. Biotechnol. **33**:408–414.

Zhu-Shimoni, J.X., and Galili, G. (1998). Expression of an *Arabidopsis* aspartate kinase/homoserine dehydrogenase gene is metabolically regulated by photosynthesis-related signals but not by nitrogenous compounds. Plant Physiol. **116**:1023–1028.