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 the25 seed compositions determined in this study (Supplemental Table4). 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 Bead Chip, 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 R5% 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 000MCMCrepeats.Population differentiation (Fst)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 a significant association (P<1.57310 6). The lead SNP was chosen to represent the QTL, and the nearby significant SNPs were identified by using MLM with LD r2 R 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).