2016

Genome-wide Association Mapping of Qualitatively Inherited Traits in a Germplasm Collection

Nonoy B. Bandillo
University of Nebraska-Lincoln

Aaron J. Lorenz
University of Nebraska-Lincoln, lore0149@umn.edu

George L. Graef
University of Nebraska-Lincoln, ggraef1@unl.edu

Diego Jarquin
University of Nebraska-Lincoln, jhernandezjarquin2@unl.edu

David L. Hyten
University of Nebraska-Lincoln, david.hyten@unl.edu

See next page for additional authors

Bandillo, Nonoy B.; Lorenz, Aaron J.; Graef, George L.; Jarquin, Diego; Hyten, David L.; Nelson, Randall L.; and Specht, James E., "Genome-wide Association Mapping of Qualitatively Inherited Traits in a Germplasm Collection" (2016). *Agronomy & Horticulture -- Faculty Publications*. 1183.
https://digitalcommons.unl.edu/agronomyfacpub/1183

Published July 13, 2017

Genome-wide Association Mapping of Qualitatively Inherited Traits in a Germplasm Collection

Nonoy B. Bandillo, Aaron J. Lorenz, George L. Graef, Diego Jarquin, David L. Hyten, Randall L Nelson, and James E. Specht*

Abstract

Genome-wide association (GWA) has been used as a tool for dissecting the genetic architecture of quantitatively inherited traits. We demonstrate here that GWA can also be highly useful for detecting many major genes governing categorically defined phenotype variants that exist for qualitatively inherited traits in a germplasm collection. Genome-wide association mapping was applied to categorical phenotypic data available for 10 descriptive traits in a collection of ~13,000 soybean [*Glycine max* (L.) Merr.] accessions that had been genotyped with a 50,000 single nucleotide polymorphism (SNP) chip. A GWA on a panel of accessions of this magnitude can offer substantial statistical power and mapping resolution, and we found that GWA mapping resulted in the identification of strong SNP signals for 24 classical genes as well as several heretofore unknown genes controlling the phenotypic variants in those traits. Because some of these genes had been cloned, we were able to show that the narrow GWA mapping SNP signal regions that we detected for the phenotypic variants had chromosomal bp spans that, with just one exception, overlapped the bp region of the cloned genes, despite local variation in SNP number and nonuniform SNP distribution in the chip set.

Published in Plant Genome
Volume 10. doi: 10.3835/plantgenome2016.06.0054

Core Ideas

- Genome-wide association (GWA) is usually aimed at quantitative (but not so much at qualitative) traits.
- Germplasm collections have extensive data on qualitatively inherited descriptor traits.
- Positional location of classical genes is lacking in most crop genome sequence maps.
- Genome-wide association easily generates high-resolution genome sequence map positions for classical loci.
- Genome-wide association-based gene positions are attainable even for traits governed by digenic epistasis.

IN the United States, there are 30 USDA-ARS National Plant Germplasm System sites (http://www.ars-grin.gov/npgs/sitelist.html, accessed 1 May 2017), which were established for the collection, preservation, and distribution of plant species accessions of national interest. A substantial amount of phenotypic data has been collected in many of these germplasm collections. The soybean repository is located at Urbana, IL (https://npgsweb.ars-grin.

N. Bandillo, A. Lorenz, G. Graef, D. Jarquin, D. Hyten, and J. Specht, Dep. of Agronomy & Horticulture, Univ. of Nebraska-Lincoln, Keim Hall Lincoln, NE 68583-0915; R. Nelson, USDA-ARS, Soybean/Maize Germplasm, Pathology, and Genetics Research Unit and Dep. of Crop Sciences, Univ. of Illinois, 1101 West Peabody Drive, Urbana, IL 61801-0000. Received 8 June 2016. Accepted 15 Feb. 2017. *Corresponding author (jspecht1@unl.edu). Assigned to Associate Editor Jesse Poland.

Abbreviations: Chr; chromosome; GRIN, Germplasm Resource Information Network; GBS, genotyping by sequencing; GWA, genome-wide association; MG, maturity group; MLM, mixed linear model; NIL, near-isogenic line; QTL, quantitative trait locus; SNP, single nucleotide polymorphism.

gov/gringlobal/site.aspx?id=24, accessed 1 May 2017) and it contains accessions of two annual species– the wild *Glycine soja* Siebold & Zucc. and the cultivated *G. max*, plus the accessions of 19 perennial *Glycine* species.

Nearly all of the annual *Glycine* accessions have been characterized by the collection's curation staff for many descriptive traits. Of particular interest to soybean breeders and geneticists are the descriptor traits: maturity group; stem termination; flower color; pubescence color, form, and density; pod color; seed coat luster and color; and hilum color. At least two and often several phenotype variants are listed as categories for each trait. The phenotypic category names and codes for each descriptor trait can be found at https://npgsweb.ars-grin.gov/gringlobal/descriptors.aspx (accessed 1 May 2017); select soybean, then click on any given descriptor name.

Phenotypic variants in most of these soybean descriptor traits are known to be qualitatively inherited in a monogenic or a digenic (or sometimes even in a trigenic or tetragenic) manner. Because intergenic (qualitative) epistasis plays a role in some cases, the number of phenotypes can be fewer than the number of genotypes. Past soybean inheritance studies involving qualitatively inherited traits have led to the assignment of gene symbols to the alleles at each of the loci that were inferred to govern the trait. Palmer et al. (2004) listed 251 soybean genes and also noted that 72 of these were members of 21 classical (i.e., nonmolecular) linkage groups. On the basis of molecular marker genotyping of the biparental mapping populations in which some of those 72 genes were segregating (e.g., Shoemaker and Specht, 1995), 19 of those 21 classical linkage groups (68 of the 72 genes) were assigned to molecular linkage groups that were labeled A1 to O (Cregan et al., 1999). The number of genes assigned to the molecular linkage groups has now increased from 68 to 77 [SoyBase (www.soybase.org, accessed 1 May 2017); Grant et al., 2010]. Obviously, the majority of known soybean genes have yet to be mapped. Moreover, even the genetically mapped genes have low-resolution cM map positions, except for a few cloned genes that now have a specified chromosomal bp position on the 'Williams 82' reference genome.

Establishing a chromosomal bp map position for all soybean genes using molecular-marker-genotyped biparental mapping populations would be a laborious and expensive effort. However, two recent publications suggested to us that gene mapping of qualitative traits could be accomplished via an alternative approach. Sonah et al. (2015) used genotyping-by-sequencing (GBS) to generate 47,702 SNPs, which they used to genotype 304 soybean lines spanning maturity groups (MGs) 000 to II. After performing a population structure analysis, they conducted a GWA analysis on just 139 MG 0 lines that they had characterized for five agronomic and seed traits in six field environments. Their primary goal was to discover SNPs associated with these five quantitative traits but they stated that "to validate our GWA approach", they also applied GWA to the flower, pubescence, and hilum color phenotypes that they had also recorded for those 139 lines. These authors also stated that they detected "a towering distribution of many (significant) SNPs" in the chromosomal regions corresponding to four classical genes known to control those three traits. Subsequently, Wen et al. (2015), using 342 landraces and 1062 cultivars released during 2007 to 2012, used the soybean 50K SNP chip to apply GWA to 1402 lines differing in flower color (two phenotypes), pubescence color (two phenotypes), and seed coat color (six phenotypes). In this set of MG I, II, and III genotypes, they detected strong SNP associations for these three traits in the same chromosomal regions as those reported by Sonah et al. (2015).

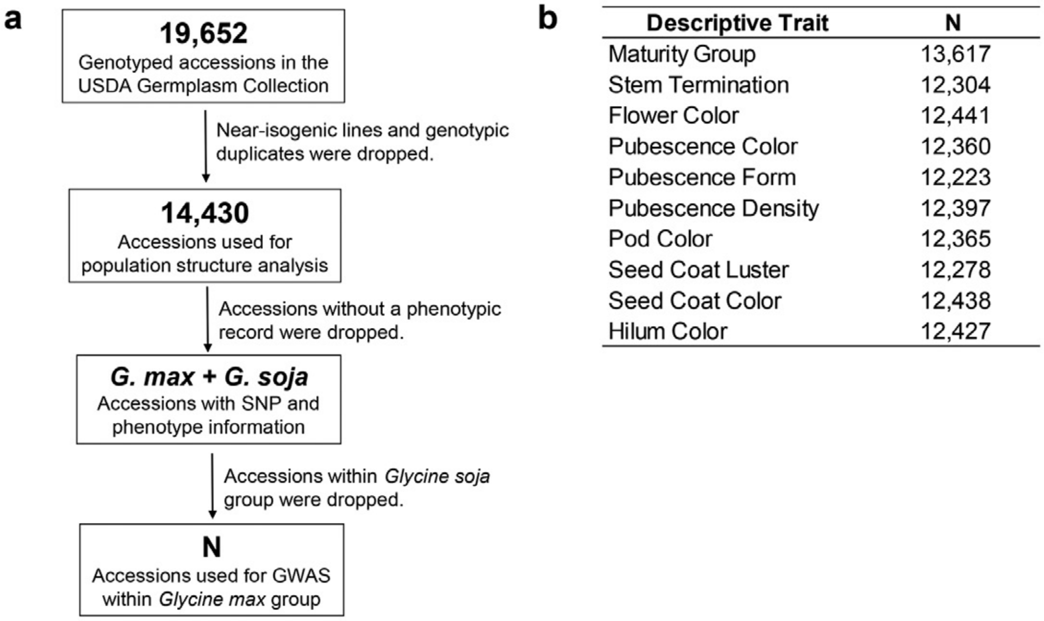
These two reports indicated that GWA could be used for quickly "mapping" many of the simply inherited classical genes that are known to govern traits qualitatively, and for which extensive phenotypic data exist in many germplasm collections. More importantly, the application of GWA to classical traits can result in immediate high-resolution, chromosomal bp map positions for the controlling genes, which would be useful for researchers interested in cloning any given classical gene of scientific or commercial interest.

To test this thesis more thoroughly, we conducted a GWA analysis using phenotypic category data for 10 soybean descriptive traits listed in Germplasm Resource Information Network (GRIN) for ~13,000 *G. max* accessions genotyped with a 50,000 SNP chip (Song et al., 2013). A GWA on a panel of accessions of this magnitude can offer substantially greater statistical power and mapping resolution than the smaller panels used by Sonah et al. (2015) and Wen et al. (2015). Our primary objective was to assess the use of GWA as a tool for chromosomal bp positional mapping of (known and unknown) genes controlling the major phenotypic variants associated with each of the 10 soybean descriptive traits. Of interest were three issues: (i) the degree of SNP signal resolution obtainable when a 50,000 SNP chip is used in a GWA to identify a chromosomal bp position of a gene locus controlling a given pair of categorical phenotypic variants *vis-à-vis* a cloned gene bp sequence; (ii) using GWA for digenic qualitative gene mapping when the population contains only three instead of four phenotypes as a result of classical digenic epistasis (i.e., F_2 phenotypic ratios of 12:3:4 or 9:3:4); and (iii) creating two-phenotype-only subsets of multiphenotype populations to clarify which GWA signal corresponds to one of the two known gene loci. The results generated in this study relative to those three issues are likely to be of interest to researchers interested in high-resolution GWA mapping of genes governing qualitatively inherited traits in their specific crop species of interest.

Materials and Methods

Plant Materials

The accessions used in this study are maintained in the USDA Soybean Germplasm Collection and have been described previously (Bandillo et al., 2015; Song et al.,


a
19,652
Genotyped accessions in the USDA Germplasm Collection
Near-isogenic lines and genotypic duplicates were dropped.
14,430
Accessions used for population structure analysis
Accessions without a phenotypic record were dropped.
G. max + G. soja
Accessions with SNP and phenotype information
Accessions within Glycine soja group were dropped.
N
Accessions used for GWAS within Glycine max group
b

Descriptive Trait	N
Maturity Group	13,617
Stem Termination	12,304
Flower Color	12,441
Pubescence Color	12,360
Pubescence Form	12,223
Pubescence Density	12,397
Pod Color	12,365
Seed Coat Luster	12,278
Seed Coat Color	12,438
Hilum Color	12,427

Fig. 1. The stepwise filtering of *G. max* accessions held in the USDA Germplasm Collection for genome-wide association mapping of 10 descriptor traits.

2015). As of 20 Nov. 2016, this collection contained 22,143 accessions of the 21 species in the genus *Glycine* (https://npgsweb.ars-grin.gov/gringlobal/site.aspx?id=24, accessed 1 May 2017), which included 1181 wild annual *G. soja* accessions, 19,956 domesticated annual *G. max* accessions, and 1006 accessions of the 19 perennial species.

Extraction of Genotype and Phenotype Data

Song et al. (2015) used an Infinium SoySNP50K iSelect Beadchip (Illumina, San Diego, CA) to Genotype 19,652 accessions of the two annual species. On the basis of a pairwise genetic similarity analysis of 18,840 *G. max* accessions genotyped with 42,509 SNPs, they discovered that 1682 accessions were 100% (and another 4206 were at least 99.9%) identical to at least one other accession. Relative to the 1168 *G. soja* accessions, 95 were 100% (and another 362 were 99.9%) identical. In the *G. max* collection, there also are 600 near-isogenic line (NIL) accessions (not including the recurrent parents). Bandillo et al. (2015) removed the SNP-identical duplicates and the NILs to conduct a population structure analysis of the two annual *Glycine* species, and then removed the *G. soja* accessions for a subsequent GWA analysis that targeted just two quantitatively inherited traits: soybean seed protein and oil. The stepwise filtering process conducted by Bandillo et al. (2015) resulted in 13,624 *G. max* accessions, which is the same accession set used in the present study for the GWA mapping of 10 descriptive traits. Any SNP with a minor allele frequency of <0.01 was removed from the genotype dataset for the GWA mapping. The SNP genotype dataset is publicly available at http://www.soybase.org/dlpages/index.php (accessed 1 May 2017).

The phenotypic data used in this study were obtained from the USDA Soybean Germplasm Collection general evaluation trials in which data were collected for morphological, agronomic, and seed quality traits. The trials were grown where the accessions were adapted; most cases, there was one replication in each of two successive years. For a comprehensive listing of all of the phenotypic categories and their codes relative to the 10 descriptor traits, see the GRIN website (https://npgsweb.ars-grin.gov/gringlobal/descriptors.aspx, 1 May 2017); enter SOYBEAN, then click on these (abbreviated) descriptor names: MatGroup, StemTerm, FlwrColor, PubColor, PubForm, PubDensity, PodColor, SCoatLuster, SCoatColor, and HilumColor. The genotyped accessions and their 10-trait phenotypes were filtered (see Fig. 1) to create a final data file of accessions and their phenotype categories by trait (see Supplemental Table S1). The phenotypic categories in each descriptive trait are quite distinct, and accession phenotypic calls between replications or trials are rarely different (i.e., phenotypic call errors). The few phenotypic call errors detected in the phenotypic data were set as missing. Because of missing phenotype scores for some traits in some accessions, the total number filtered accessions varied by trait (Fig. 1).

Genome-Wide Association Analysis

An intensive comparison of various GWA methods conducted by Wang et al. (2012) demonstrated that the mixed linear model (MLM) is the most promising for analyzing either binary, categorical or continuous traits in crops exhibiting a population structure. The MLM has been used in GWA mapping of continuous and binary or categorical traits in model plant species (Atwell et al., 2010), and in crop species such as rice (*Oryza sativa* L.) (Huang et al., 2010), corn (*Zea mays* L.) (Romay et al., 2013), barley (*Hordeum vulgare* L.) (Wang et al., 2012), and soybean (Sonah et al., 2015; Wen et al., 2015; Rincker et al., 2016). In our study, as in these prior studies, the

MLM was used for GWA mapping of either binary or categorical traits to handle the confounding effects caused by the strong population structure present in the soybean germplasm collection. For each trait, marker–trait associations were tested using the $Q + K$ model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{C}\gamma + \mathbf{Z}\mathbf{u} + e$, where $\mathbf{y}$ is a vector of phenotypic responses for $i = 1, \ldots, N$ accessions for the analyzed trait; $\boldsymbol{\beta}$ is a vector of fixed marker effects; $\boldsymbol{\gamma}$ is a vector of subpopulation effects; and $\mathbf{u}$ is a vector of polygenic effects caused by relatedness assuming a genomic representation of the random effect of the i^{th} accession as a linear combination between p markers and their corresponding marker effects as:

$$\mathbf{u}_i = \sum_{j=1}^{p} x_{ij} b_j , \qquad [1]$$

with $b_j \sim N\left(0,\sigma_{\mathrm{b}}^2\right)$, where σ_{b}^2 is the marker effect. By the properties of the multivariate normal distribution, $\mathbf{u} \sim MVN(0,\mathbf{K}\sigma_u^2)$ for $\mathbf{K} = \dfrac{\mathbf{XX}`}{p}$ and $\sigma_u^2 = p\sigma_{\mathrm{b}}^2$. The vector of residuals is $\mathbf{e} = \left\{e_i\right\}$ where $e_i \overset{IID}{\sim} N\left(0,\sigma_e^2\right)$, in which *IID* signifies independent and identically distribution and σ_e^2 is the residual variance. $\mathbf{X}$ is a marker matrix, $\mathbf{C}$ is an incidence matrix containing membership proportions to each of the five genetic clusters identified by the ADMIXTURE analysis (Bandillo et al., 2015; Alexander et al., 2009), $\mathbf{Z}$ is the corresponding design matrix for u in the case of replicated accessions, and $\mathbf{K}$ is the realized genomic relationships matrix describing genetic similarities between pairs of individuals, which is estimated internally in the Factored Spectrally Transformed Linear Mixed Models using the SNP data (Lippert et al., 2011). This model was implemented using the Factored Spectrally Transformed Linear Mixed Models algorithm, which is a program designed to accommodate large datasets with reduced computational time. Association analyses were conducted across groups of accessions classified either by MG class or by world region. Genome-wide association mapping across all groups was conducted using only SNPs with a minor allele frequency of >0.01, with population structure accounted for by using the respective fixed and random effects of γ and u. The qqman R package (Turner, 2014) was used to visualize quantile–quantile plots and the genomic inflation parameter Λ, a metric of the degree of inflation of p-values (Devlin and Roeder, 1999), was calculated.

We used an error value of $-\log_{10}P = 5.17$ (i.e., 6.75×10^{-6}) for the detection of significant SNP associations, which was determined by Bandillo et al. (2015) in 13,624 *G. max* accessions to correspond to an experiment-wise Type I error value of $\alpha = 0.05$. Briefly, the correlation matrix and eigenvalue decomposition among 42,509 SNPs were calculated to determine the effective number of independent tests (M_{eff}) (Li and Ji, 2005). The significance test criteria were then adjusted using the M_{eff}, with the correction (Sidak, 1967):

$$\alpha_p = 1 - (1 - \alpha_a)^{\frac{1}{M_{eff}}} , \qquad [2]$$

where α_{p} is the computed comparison-wise error rate but α_{e} is the inputted desired experiment-wise error rate (i.e., 0.05). The stringent $-\log_{10}P = 5.17$ significance value was used for all GWA scans in this study because our primary focus was on mapping major SNP signals corresponding to major qualitative trait genes (rather than modifier genes with a modest effect). The impact of a lower N in some GWA population subsets was minimal (see the tabulated p-values for various values of N in the Supplementary File S1). Multiple-linear regression was used to estimate the proportion of phenotypic variance accounted for by significant SNPs after accounting for population structure effects.

Determining the Global Distribution of Allelic Variations

Accessions were grouped into subpopulations defined by world region, which is a major determinant of population structure within the soybean germplasm collection, as reported by Bandillo et al. (2015). World region subpopulations consisted of eight major manageable countries or regions of origin: China (36%), North and South Korea (19%), Japan (17%), North and South America (9%), South and Southeast Asia (8%), Europe (5%), Russia (5%), and Others (Bandillo et al., 2015). On the basis of the results of GWA mapping, the closest SNP that tagged a classical gene locus was used to estimate the frequency of the two alleles at that locus. Allele frequencies were estimated within each subpopulation using the CrossTable function in the gmodels package, implemented in R software version 3.2.1 (R Development Core Team, 2014). At each SNP locus, Fisher's exact test was used to test the null hypothesis that the frequency of the allele conferring a trait of interest was the same across world regions. The allele frequency output from CrossTable was then used to make plots using the pie function in R.

Candidate Gene Annotations

Gene annotations were extracted using the *G. max* cv. Williams 82 gene models (Glyma.Wm82.a1.v1.1) downloaded from Phytozome (http://phytozome.jgi.doe.gov/pz/portal.html, accessed 1 May 2017) and were displayed using a chromosome visualization tool (Cannon and Cannon, 2011). A 250-kb sliding-window approach (125 kb upstream and 125 kb downstream from the most significant SNP position) was used to search for functional genes; this was implemented in BEDTools (Quinlan and Hall, 2010). Candidate genes included (i) soybean genes of known function related to the trait, (ii) genes with orthologs with known function in *Arabidopsis thaliana* (L.) Heynh., or both. Annotation data are presented only for noncloned classical genes and new loci for which a GWA signal was detected in this study (Supplemental Table S2). SoyBase provides an easy tool to look up the name correspondence between the Glyma.Wm82.a1.v1.1 annotation

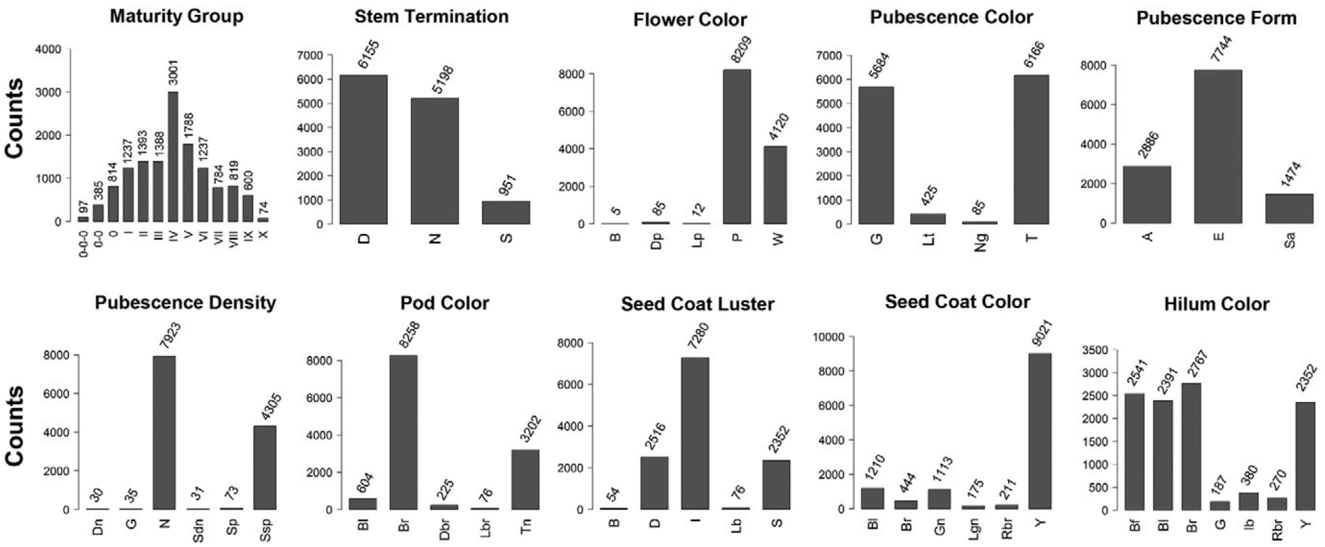


Fig. 2. Frequency distributions of multiple phenotypic variants that were available in each of the 10 soybean descriptive traits relative to the genome-wide analysis conducted on each trait (i.e., the *N* value in Fig. 1 for the *G. max* accessions). Abbreviations used in each histogram are: Maturity Group: Roman numerals from 000 (early) to X (late); Stem Termination: determinate (D), indeterminate (N), and semi-determinate (S); Flower Color: blue (B), dark purple (Dp), light purple (Lp), purple (P), and white (W); Pubescence Color: grey (G), light tawny (Lt), near grey (Ng), and Tawny (T); Pubescence Form: appressed (A), erect (E), and semi-appressed (Sa); Pubescence Density: dense (D), glabrous (G), normal (N), semi-dense (Sdn), sparse (S), and semi-sparse (Ssp); Pod Color: black (Bl), brown (Br), dark brown (Dbr), light brown (Lbr), and tan (Tn); Seed Coat Luster: bloom (B), dull (D), intermediate (I), light bloom (Lb), and shiny (S); Seed Coat Color: black (Bl), brown (Br), green (Gn), light green (Lgn), red-brown (Rbr), and yellow (Y); Hilum Color: buff (Bf), black (Bl), brown (Br), grey (G), imperfect black (Ib), red-brown (Rbr), and yellow (Y).

used here (and in the soybean papers cited) and the Glyma.Wm82.a2.v1 annotation (http://soybase.org/about-genomenomenclature.php, accessed 1 May 2017).

Results and Discussion

From a population of ~21,000 soybean accessions originating from 84 different countries, we extracted a large association panel for mapping genes governing the phenotypes of the 10 qualitative descriptive traits. The filtered population sizes (*N* in Fig. 1) ranged from 13,617 to 12,223. For the initial GWA conducted on each trait, the available multiple phenotypic categories (Fig. 2) ranged from 13,617 to 10,888 accessions (Supplemental Fig. S1–Supplemental Fig. S10). The phenotypic category names are abbreviated in the text and in all figures (see Supplemental File S1for a tabulated listing of the full names). Using the MLM that corrects for the effects of population structure and genetic relatedness, our GWA mapping identified a total of 723 significant SNPs ($-\log_{10}P >$ 5.17) in 61 genomic regions among all 10 traits (Table 1). Overall, the GWA Manhattan plots documented significant SNP signals corresponding to 24 known classical genes, 11 of which have been cloned (Supplemental Fig. S1 to Supplemental Fig. S10; Fig. 3). Several strong SNP signals that may correspond to heretofore unknown (i.e., nonsymbolized) qualitative genes were also detected. The large population size (~13,000 accessions), coupled with the substantial genetic diversity in the soybean germplasm collection, resulted in our GWA analyses providing high mapping resolution relative to pinpointing the chromosomal bp position of the genes controlling the phenotypic variation associated with these qualitatively inherited traits. In addition, the magnitudes of the $-\log_{10}P$ scores for the SNPs identifying qualitatively inherited genes obtained in this study were substantially higher than any previous GWA or quantitative trait locus (QTL) mapping study conducted to date in soybean (www.soybase.org, accessed 1 May 2017). To leverage the fine bp mapping resolution obtainable from GWA, we assembled a list of of annotated candidate genes (Wm82.a1.v1.1 version; Supplemental Table S2) located within 250-kb regions centered on each GWA-detected SNP peak signal (but not for SNP signals associated with already cloned genes). The data in Supplemental Table S2 allow us to assess the plausibility of potential candidate genes for the SNP signals that corresponded to known named Mendelian (but not yet cloned) genes. Here, we document the chromosomal bp positions of the significant SNP signal regions that overlapped the coding sequence bp positions of the 10 cloned loci of: *E1–e1*, *E2–e2*, *E3–e3*, *Dt1–dt1*, *Dt2–dt2*, *W1–w1*, *T–t*, *Hps*, *I–i*, and *R–r* (see their bp positions shown in bold in Table 1). Overlapping SNP signals were not detected for the cloned *E4–e4* or the fine-mapped *L1–l1* loci, and a few SNP signals did not correspond to any known classical gene locus. Our GWA map findings for each of the 10 descriptor traits are successively presented in the next 10 subsections. For tabulated information about the classical (symbolized) gene loci known to control the trait phenotypes, see Supplemental File S1.

Table 1. Summary of single nucleotide polymorphisms (SNP) association signals that exceeded an experiment-wise significance criterion of $-\log_{10}P > 5.17$ in a genome-wide association analysis (GWA) using the *Q* + *K* model performed on 13,624 *G. max* accessions for 10 soybean descriptor traits. The significant SNP associations detected in this study are ordered by trait, then by chromosome (Chr), and then within each chromosome according to the bp regions of the significant SNPs.

				Significant SNP–trait associations‡						Cloned gene information					
Descriptive trait	Supp. Fig. No.†	Chr	LG	SNPs	First	Last	Max.	$-\log_{10}P$	Gene locus	Glyma name	Other name	Start	End	Size	References
				n	bp							bp			
Maturity group	S1a, S1b	6	C2	16	18,724,519	21,426,047	19,709,089	23.97	***E1–e1***¶	*Glyma06g23026*	*RPR*	**2,000,6973**	**20,007,810**	838	Xia et al. (2012)
	S1a, S1b	8	A2	1	3,642,671	3,642,671	3,642,671	5.19	–	–	–	–	–	–	–
	S1a, S1b, S1c	10	O	20	44,476,584	45,294,441	44,743,315	49.73	***E2–e2***¶	*Glyma10g36600*	*GmGIa*	**44,716,720**	**44,738,268**	21,549	Watanabe et al. (2011)
	S1a, S1b, S1c	11	B1	4	10,721,006	11,572,077	11,269,310	18.83	–	–	–	–	–	–	SoyBase Pod Maturity QTL 17–2 (Lu et al., 2015)
	S1a, S1b, S1c, S1e	12	H	6	5,491,240	5,786,241	5,491,240	14.56	–	–	–	–	–	–	SoyBase Pod Maturity QTL 26–2 (Lu et al., 2015)
	S1a, S1b	13	F	1	36,616,135	36,616,135	36,616,135	5.69	–	–	–	–	–	–	–
	S1b	18	G	1	59,902,680	59,902,680	59,902,680	7.02	–	–	–	–	–	–	SoyBase Pod Maturity QTL29–8 (Wen et al. 2015)
	S1a, S1b, S1c	19	L	7	47,270,486	48,037,479	47,510,130	21.54	***E3–e3***¶	*Glyma19g41210*	*GmPhyA3*	**47,511,246**	**47,519,957**	8,712	Watanabe et al. (2009)
	S1d	20	I	2	3,145,294	3,150,963	3,150,963	5.36	–	–	–	–	–	–	–
	S1d	20	I	1	28,550,287	28,550,287	28,550,287	5.85	–	–	–	–	–	–	–
	S1d	20	I	5	34,434,402	34,462,359	34,437,459	6.39	***E4–e4***¶	*Glyma20g22160*	*GmPhyA2*	**32,087,580**	**32,093,266**	5,687	Liu et al. (2008)
Stem termination	S2b, S2c	6	C2	1	38,948,190	38,948,190	38,948,190	6.80	–	–	–	–	–	–	–
	S2a, S2b	18	G	4	**59,902,680**	**60,380,782**	59,902,680	17.27	***Dt2–dt2***¶	*Glyma18g50910*	*Loc100788956*	**59,918,841**	**59,927,027**	8,187	Ping et al. (2014)
	S2a	19	L	1	43,080,829	43,080,829	43,080,829	5.26	–	–	–	–	–	–	–
	S2a, S2b, S2c	19	L	55	**44,329,464**	**45,525,374**	45,000,827	238.79	***Dt1–dt1***¶	*Glyma19g37890*	*PEPB; TFL1*	**44,979,743**	**44,981,385**	1,643	Liu et al. (2010); Tian et al. (2010)
	S2a	19	L	2	47,069,443	47,076,497	47,069,443	6.22	***E3–e3***¶	*Glyma19g41210*	*GmPhyA3*	**47,511,246**	**47,519,957**	8,712	Watanabe et al. (2009)
Flower color	S3a, S3b	13	F	2	1,629,730	1,756,025	1,756,025	8.05	–	–	–	–	–	–	–
	S3a, S3b	13	F	20	2,493,212	3,044,754	2,833,623	24.25	–	–	–	–	–	–	–
	S3a, S3b	13	F	23	3,175,654	3,825,644	3,657,853	152.79	–	–	–	–	–	–	–
	S3a, S3b	13	F	29	**4,173,955**	**5,480,962**	4,559,799	169.79	***W1–w1***¶	*Glyma13g04210*	*W1*	**4,552,711**	**4,557,371**	4,661	Zabala and Vodkin (2007); Sonah et al. (2015); Wen et al. (2015)
	S3b	19	L	1	36,603,029	36,603,029	36,603,029	14.81	***L1–l1***	*Glyma19g27460*	–	34,750,891	34,752,190	1,300	Bernard (1967); He et al. (2015)¶
Pubescence color	S4a, S4b	3	N	65	46,332,845	47,702,654	47,307,916	95.83	***Td–td***	–	–	–	–	–	Bernard (1975a); Behm et al. (2011); Wen et al. (2015)
	S4a, S4b	6	C2	26	**17,258,654**	**19,815,389**	18,252,495	298.24	***T–t***¶	*Glyma06g21920*	*T*	**18,534,619**	**18,541,464**	6,846	Zabala & Vodkin (2003); Sonah et al. (2015); Wen et al. (2015)
	S4a	6	C2	1	25,762,003	25,762,003	25,762,003	7.45	–	–	–	–	–	–	–

(cont'd.)

Table 1. Continued.

Descriptive trait	Supp. Fig. No.[†]	Chr	LG	Significant SNP–trait associations[‡]					Gene locus	Cloned gene information					References
				SNPs	First	Last	Max.	$-\log_{10}P$		Glyma name	Other name	Start	End	Size	
				n	bp							bp			
Pubescence color, cont'd.	S4a	6	C2	2	30,106,667	30,163,816	30,163,816	7.08	–	–	–	–	–	–	–
	S4b	6	C2	1	38,948,190	38,948,190	38,948,190	5.23	–	–	–	–	–	–	–
	S4a	12	H	1	18,537,212	18,537,212	18,537,212	8.24	–	–	–	–	–	–	–
	S4a	12	H	1	21,317,830	21,317,830	21,317,830	9.06	–	–	–	–	–	–	–
	S4a, S4b	14	B2	1	46,968,410	46,968,410	46,968,410	5.40	–	–	–	–	–	–	–
	S4a	20	I	1	13,081,929	13,081,929	13,081,929	8.45	–	–	–	–	–	–	–
Pubescence form	S5a, S5b	12	H	16	37,036,017	37,786,243	37,356,120	251.32	***Pa1–pa1***	–	–	–	–	–	Bernard (1975b); Lee et al. (1999)
	S5a, S5b	13	F	11	30,181,642	30,708,708	30,708,708	26.87	***Pa2–pa2***	–	–	–	–	–	Bernard (1975b); Lee et al. (1999)
Pubescence density	S6a, S6b, S1c	1	D1a	4	55,493,281	55,523,014	55,523,014	8.14	***Pd1–pd1***	–	–	–	–	–	Bernard and Singh (1969); Pubescence Density QTL 1–1 (Komatsu et al., 2007)
	S6a	7	M	1	17,153,201	17,153,201	17,153,201	5.31	–	–	–	–	–	–	–
	S6a, S6b, S1c	12	H	53	34,477,297	35,525,603	34,877,806	224.28	***Ps*–*Ps*^{s}–*ps***	–	–	–	–	–	Bernard (1975c); Pubescence Density QTL 1–2 (Komatsu et al., 2007)
	S6a, S6b	14	B2	1	4,934,894	4,934,894	4,934,894	69.92	–	–	–	–	–	–	–
	S6c	9	K	11	43,686,430	44,855,340	44,348,623	21.41	***P1–p1***	*Glyma09g38410*	–	–	–	–	Bernard and Singh (1969); Zabala and Vodkin (2014); Hunt et al. (2011)
	S6c	9	K	8	45,344,827	46,694,731	46,139,114	18.93	–	–	–	–	–	–	–
Pod color	S7a, S7b	1	D1a	3	52,249,479	52,263,952	52,253,980	6.64	–	–	–	–	–	–	–
	S7a, S7b	3	N	27	246,658	1,338,018	537,774	147.86	***L2–l2***	–	–	–	–	–	Bernard (1967)
	S7a, S7b, S7c	19	L	48	36,397,778	38,521,183	37,649,854	127.49	***L1–l1***	*Glyma19g27460*	–	34,750,891	34,752,190	1,300	Bernard (1967); He et al. (2015)[¶]
Seed coat luster	S8b	8	A2	1	6,897,932	6,897,932	6,897,932	6.25	–	–	–	–	–	–	–
	S8b	8	A2	7	**8,272,057**	**8,656,325**	8,462,762	12.78	***I*–*i*^{k}–*i*^{i}–*i*[¶]**	Inverted repeats[§]	*CHS1–3-4*[§]	**8,462,596**	**8,469,679**	7,084	Todd and Vodkin (1996); Tuteja et al. (2009)**[§]**
	S8a, S8b, S8c	9	K	1	1,456,482	1,456,482	1,456,482	34.25	***B?–b?***	–	–	–	–	–	Woodworth (1932); this study (proposed the locus symbols)
	S8b	11	B1	1	15,920,433	15,920,433	15,920,433	6.80	–	–	–	–	–	–	–
	S8a	13	F	1	34,089,340	34,095,060	34,095,060	7.68	***B1–b1***	–	–	–	–	–	Woodworth (1932); Tang and Tai (1962); Chen & Shoemaker (1998)
	S8a, S8c	15	F	1	7,861,342	7,861,342	7,861,342	6.24	–	–	–	–	–	–	–
	S8a, S8c	15	E	17	**8,512,905**	**8,941,824**	8,893,988	9.26	***Hps*[††]**	*Glyma15g11970*	*Hps*[#]	**8,868,741**	**8,875,714**	6,974	Gijzen et al. (1999, 2003a, 2003b, 2006**[#]**)

(cont'd.)

Table 1. Continued.

Descriptive trait	Supp. Fig. No.†	Chr	LG	Significant SNP–trait associations‡					Cloned gene information						References
				SNPs	First	Last	Max.	–log$_{10}P$	Gene locus	Glyma name	Other name	Start	End	Size	
				n	bp							bp			
Seed coat luster, cont'd	S8a, S8b, S8c	15	E	93	9,012,474	10,475,708	10,416,352	110.19	***B–b***	–	–	–	–	–	Lorenzen et al. (1995); Gijzen et al. (2003a, 2003b)
Seed coat color	S9a, S9b	1	D1a	30	5,225,3980	52,493,627	52,253,980	274.08	***G–g***	–	–	–	–	–	Woodworth (1921)
	S9a	6	C2	5	**18,033,759**	**18,810,733**	18,766,611	11.12	***T–t***¶	*Glyma06g21920*	*T*	**18,534,619**	**18,541,464**	6,846	Zabala and Vodkin (2003)
	S9a	8	A2	7	4,802,080	5,113,384	5,003,648	8.64	***O–o***	–	–	–	–	–	–
	S9a	8	A2	16	**8,013,021**	**9,120,830**	8,462,762	37.47	**I–i^k–i^i–i**¶	Inverted repeats§	*CHS1–3-4*§	**8,462,596**	**8,469,679**	7,084	Todd and Vodkin (1996); Tuteja et al. (2009)§
	S9a	9	K	1	42,974,503	42,974,503	42,974,503	5.80	**R–r^m–r**¶	*Glyma09g36983*	*R2R3 MYB*	**42,562,649**	**42,564,660**	2,012	Gillman et al. (2011); Zabala and Vodkin (2014)
Hilum color	S10a	6	C2	16	**17,567,713**	**18,810,733**	18,766,611	100.30	***T–t***¶	*Glyma06g21920*	*T*	**18,534,619**	**18,541,464**	6,846	Zabala and Vodkin (2003)
	S10c	8	A2	11	4,800,584	5,113,384	4,802,080	35.00	***O–o***	–	–	–	–	–	–
	S10a, S10b, S10c	8	A2	20	**7,418,586**	**8,836,971**	8,572,686	96.55	**I–i^k–i^i–i**¶	Inverted repeats§	*CHS1–3-4*§	**8,462,596**	**8,469,679**	7,084	Todd and Vodkin (1996); Tuteja et al. (2009)§
	S10a, S10c	8	A2	3	9,111,316	9,130,626	9,111,316	8.46	–	–	–	–	–	–	–
	S10b	9	K	55	**41,660,046**	**43,669,720**	42,974,503	80.58	**R–r^m–r**¶	*Glyma09g36983*	*R2R3 MYB*	**42,562,649**	**42,564,660**	2,012	Gillman et al. (2011); Zabala and Vodkin (2014)
	S10b	12	H	1	487,052	487,052	487,052	7.57	–		–	–	–	–	–
	S10a	13	F	2	4,559,799	4,562,505	4,559,799	8.62	***W1–w1***¶	*Glyma13g04210*	*W1*	**4,552,711**	**4,557,371**	4,661	Zabala and Vodkin (2007); Sonah et al. (2015); Wen et al. (2015)

† Data in each row originate from the underlined supplemental figure (i.e., the GWA analysis Manhattan plot). This column lists the supplemental figure numbers and panel labels (i.e., a, b, c, etc.) displaying the GWA Manhattan plots generated for each trait.

‡ These columns list the detected number of significant SNPs within a contiguous region, the bp positions of the first and last SNP spanning that region, and the bp position of the SNP in that region that exhibited the maximum –log$_{10}P$-value. If an entry is underlined, the corresponding row's data were used for the magnified single-chromosome GWA Manhattan plot presented in the Fig. 3 panels.

§ The dominant alleles at the I–i^k–i^i–i locus consist of two inverted repeats of the three contiguously arranged chalcone synthetase genes of *CHS1*, *CHS2*, and *CH3*, so a single Glyma number is not available for this locus. For specific details, see Tuteja et al. (2009).

¶ This column lists the detected classical loci (cloned or not). The *L1-l1* locus has not yet been cloned, but was recently fine-mapped by He et al. (2015) to a Chr 19 region of 13 potential candidate genes. Of those 13, the Glyma gene listed in this table for *L1–l1* was inferred by He et al. (2015) as the causal gene but that inference has not yet been experimentally verified. Listed in the adjacent columns is information for genes that have been cloned to date, and includes the cloned gene Glyma name, plus any other name, the start and stop bp positions of the coding sequence, and the gene bp length. If the cloned gene start–stop bp positions were bracketed by the first to last bp positions of GWA SNPs, the bp regions of the former and latter are shown in bold typeface.

# The *Hps* gene consists of a tandem array of reiterated 8.6-kb coding units. Each unit is a single open reading frame for the hydrophobic protein from soybean. A null *Hps* allele has not been identified but the copy number variants constitute multiple alleles (i.e., many copies in the dull seed coat genotypes but fewer copies in the shiny seed coat genotypes). For specific details, see Gijzen et al. (2006).

†† LG, linkage group; QTL, quantitative trait locus.

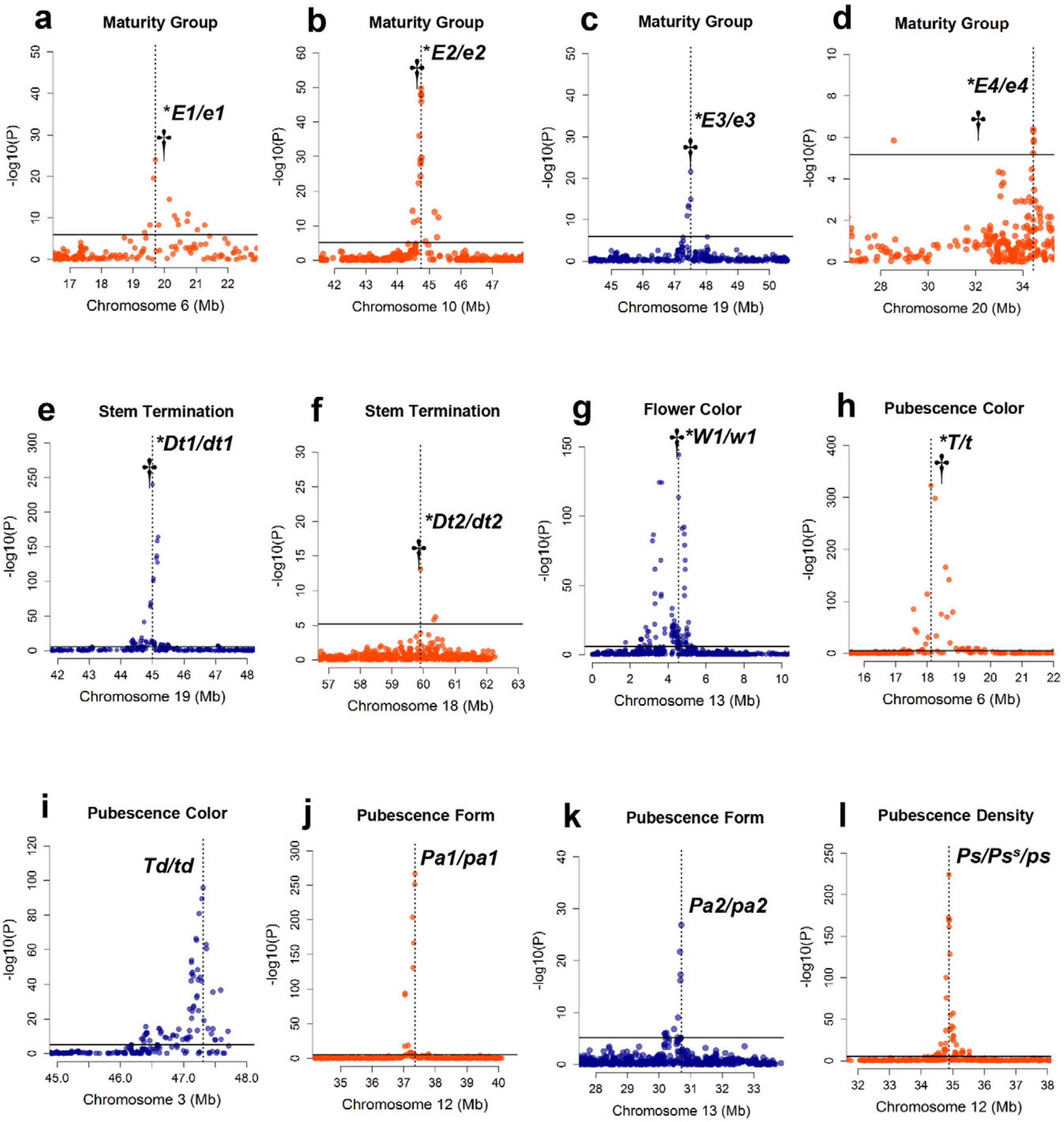


Fig. 3. (continued on next page) Genome-wide association mapping and functional annotations of associated genomic regions for 10 descriptive soybean traits. The magnified regions of the GWA Manhattan plots for the 21 SNP association signals corresponding to each classical gene locus are displayed in panels: a–d, maturity group; e–f, stem termination type; g, flower color; h–l, pubescence color; j–k, pubescence form; l–n, pubescence density; o–p, pod color; q, seed coat luster; r, seed coat color; s–u, hilum color. Qualitatively inherited genes known to control a given trait are shown in the dominant/recessive allele format (Table 1) and are asterisked if cloned. The dagger symbols denote the bp position of the coding sequence of cloned genes or the fine-mapped location of genes not yet cloned. The $-\log_{10}P$-values are plotted against the physical position (bp) on each of the 20 chromosomes. The solid horizontal line indicates the calculated threshold value ($-\log_{10}P > 5.17$) for determining a significant association; the dashed vertical lines indicate the most significant associations detected.

Maturity Group

A GWA analysis of all 13,617 accessions spanning 13 MG groups (Fig. 2) generated a Manhattan plot that exhibited five highly significant signals, a moderately significant signal, and two other signals of borderline significance (Supplemental Fig. S1a). Notably, three of the five MG signals had SNP bp ranges spanning the chromosome (Chr) 6, 10, and 19 bp positions of the cloned *E1–e1*, *E2–e2*, and *E3–e3* loci (Table 1; see the Supplemental File S1 for the known *E–e* genes). For a magnified single Chr view

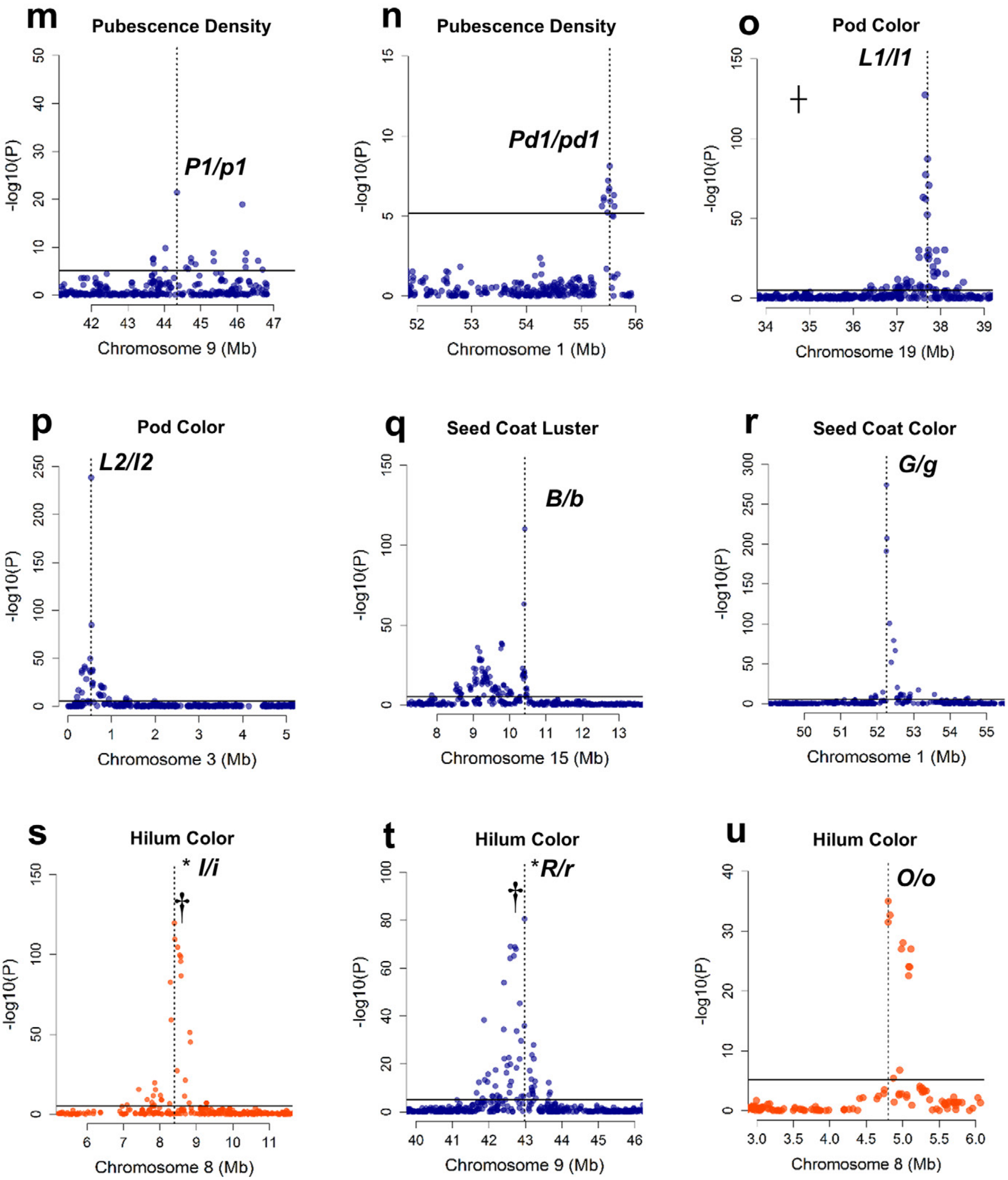


Fig. 3. Continued.

of each of these signals, see Fig. 3a–c, in which the dagger symbols point to the bp position of the cloned gene. The *E7–e7* locus is closely linked to the *E1–e1* locus, so their two GWA signals may be commingled (Fig. 3a). The other two major signals did not correspond to any cloned or mapped *E–e* genes but had map positions near two highly significant maturity QTLs. The pod maturity QTL 26–2 on Chr 12 detected by Li et al. (2008) had a large 6-d additive effect on maturity. The pod maturity QTL 17–2 on Chr 11 detected by Gai et al. (2007) had a large 7-d additive effect. Lu et al. (2015) recently noted that the two QTLs had positions near simple sequence repeat locus *Satt442* (6,361,515–6,361,774 on Chr 12) and near simple sequence repeat locus *Satt519* (13,984,414–13,984,651 on

Chr 11), which were close to our Chr 12 and 11 GWA SNP positions (Table 1). Lu et al. (2015) considered these two QTLs to be of major importance for breeder-manipulated adaptations in China. The moderately significant single SNP at Gm18:59902680 had a nearly identical position to the 59,603,446 SNP signal on Chr 18 detected by Wen et al. (2015), which those authors associated with the SoyBase Pod Maturity QTL 29–8.

Soybean adaptation to latitudes distal from the equator generally requires dominant *E* alleles that condition late maturity to be replaced with recessive *e* alleles conditioning early maturity (Tsubokura et al., 2012, 2013; Zhai et al., 2014; Zhao et al., 2016). This led us to conduct a GWA targeting just the 8315 high-latitude adapted accessions of MGs 000 to IV (Fig. 2) and a contrasting GWA targeting the 5302 low-latitude adapted accessions of MGs V to IX (Fig. 2). The SNP signals for the *E1–e1*, *E2–e2*, and *E3–e3* loci and those on Chr 11 and 12 were again detected in the high-latitude MG set (Supplemental Fig. S1b), though at diminished $-\log_{10}P$-values, except for the *E2–e2* signal, whose $-\log_{10}P$-value signal was strengthened twofold. The original three borderline significant SNPs disappeared. In contrast, in the low-latitude MG set (Supplemental Fig. S1c), maturity class variation attributable to the *E1–e1* and *E3–e3* loci was not detected and the *E2–e2* signal and Chr 11 signal did not change appreciably. This result led us to infer that very few accessions, if any, in the low-latitude MG V to X classes were homozygous recessive at the *E1–e1* and *E3–e3* loci. Apparently, the attainment of a finer degree of latitudinal photoperiod adaptation *within* the five southern US MGs arises solely from the *E2–e2* locus and from the two (yet to be cloned and named) *E–e* loci that underlie the Chr 11 and 12 QTLs. Bernard (1971) reported that *E1* and *E2* delayed maturity and flowering by 18 and 23 d for *E1* and 14 and 7 d for *E2* respectively. Using additional NILs and more replications, McBlain et al. (1987) reported that *E1*, *E2*, and *E3* delayed maturity and flowering by 11 and 16 d for *E1*, 11 and 7 d for *E2*, and 6 and 6 d for *E3*, respectively. Because the *E2–e2* locus has a smaller allelic effect on flowering date than on maturity date, it offers a distinct advantage over the other two loci when breeders seeking latitudinal photoperiod adaptation want to delay or advance the date of Stage R7 (physiological maturity) without an equal (i.e., *E3–e3*) or larger (i.e., *E1–e1*) delay or advance in the date of soybean Stage R1 (first flower). This may explain why maturity variation at the *E2–e2* locus has a stronger signal than the other two loci, not only within the MG 000 to IV classes (Supplemental Fig. S1b) but also within the MG V to X classes (Supplemental Fig. S1c).

The cloned maturity gene locus *E4–e4* on Chr 20 was not detected in the 13,617-accession GWA of all MGs (Supplemental Fig. S1a), nor was it detected in the 8537-accession GWA of MGs 000 to IV (Supplemental Fig. S1b). This may not be surprising, because this locus may not come into play except in soybean crop production areas that have rapidly developed in ever-higher latitudes, where breeders have been replacing the dominant *E4* allele (late flowering or maturity) with the recessive *e4* allele (early flowering or maturity) to create new cultivars with a suitable photoperiod adaptation (Zhai et al., 2014; Zhao et al., 2016). A GWA for just 1199 accessions of MGs 00 and 0 (Fig. 2) displayed two significant Chr 20 SNP signals (Supplemental Fig. S1d) but neither of these overlapped the *E4–e4* coding sequence. A more significant SNP was located ~2.3 Mbp upstream (Table 1; Fig. 3d). This low-*N* GWA may not have had sufficient statistical power for optimally resolving the *E4* gene position (Supplemental Fig. S1d). However, adding the 1237 MG I and 000 set to the 1199 MG 00 to 0 set resulted in the disappearance of the Chr 20 GWA signal (data not shown). The reason may be that *E3* is epistatic to *e4* (Saindon et al., 1989a; 1989b).

A long juvenile period, which is produced in genotypes that are homozygous for recessive genes of *e6* and *j* (Cober et al., 2010) and in genotypes that are homozygous for the recessive gene *e9* located on Chr 16 (Zhao et al., 2016), provides a means for delaying the onset of flowering of genotypes adapted to nonequatorial environments to attain greater yield potential. To determine if we could detect any of these three loci, we conducted a GWA for just the 2277 accessions of the late MGs VII to X (Fig. 2); however, we detected only a Chr 12 signal (Supplemental Fig. S1e): again, Pod Maturity QTL 17–2 (Li et al., 2008). No information is available as to whether the early maturity allele for this QTL is dominant, as it would have to be if it corresponds to either the *E6–e6* or a *J–j* locus.

Stem Termination Type

Genome-wide association mapping, using all 12,034 accessions that had been classified as having a stem termination phenotypes of determinate, semi-determinate, or indeterminate (i.e., 6155, 951, and 5198 accession frequency, respectively; see Fig. 2), resulted in the detection of two major SNP signals on Chr 19 and 18, whose positions corresponded to the respective cloned *Dt1–dt1* (Liu et al., 2010; Tian et al., 2010) and *Dt2–dt2* (Ping et al., 2014) genes (Supplemental Fig. S2a; Table 1; Fig. 3e, 3f; Supplemental File S1). The low-level significant GWA signal detected on Chr 19 was close to the *E3–e3* maturity locus (Fig. 3e); the latter is located ~2.5 Mbp downstream from the *Dt1–dt1* locus (~25 cM in Figure 1 of Watanabe et al., 2009). Bernard (1972) reported that (i) the *Dt1–dt1* gene locus was responsible for the stem termination phenotypic extremes of determinate (*dt1dt1* genotypes) and indeterminate stem termination types (*Dt1Dt1* genotypes), (ii) the dominant allele at the *Dt2–dt2* locus converted the indeterminate phenotype (*Dt1Dt1dt2dt2*) into a semi-determinate phenotype (*Dt1Dt1Dt2Dt2*), and (iii) the recessive *dt1* gene suppressed the expression of the semi-determinate phenotype in a *dt1dt1Dt2Dt2* genotype, leading to a 9:3:4 semi-determinate/intermediate/determinate F_2 segregation ratio (i.e., recessive epistasis). To mitigate the impact of the epistatic effect of *dt1dt1* on *Dt2* expression and to determine how closely the semi-determinate and indeterminate phenotypic classifications (which are based on the presence of a terminal raceme and the degree of stem tapering) correspond to the actual genotype, we restricted our next GWA to the 6149 accessions scored by the germplasm collection staff as having either an semi-determinate or an indeterminate phenotype. This GWA was partly successful in strengthening the signal for the *Dt2–dt2* locus (Supplemental Fig. S2b); however, compared with the initial GWA (Supplemental Fig. S2a), the Chr 19 signal (near the *E3–e3* locus) disappeared and a new signal appeared on Chr 6. Moreover, the *Dt1–dt1* signal did not disappear, indicating that a phenotype-based definition of interdeterminate and semi-determinate stem type (as defined above) does not always correspond to the genotypic-based definitions of *Dt1Dt1dt2dt2* (indeterminate) and *Dt1Dt1Dt2Dt2* (semi-determinate). Next, we used the most significant SNP nearest the *Dt1* gene as a “tag” to perform discriminant analysis with manual checking, which revealed that 27% (261 out of 951) of the semi-determinate phenotypes and 7% (356 out of 5198) of indeterminate phenotypes might actually be *dt1dt1* genotypes. In a new GWA conducted with these 261 semi-determinate and 365 indeterminate accessions omitted (Supplemental Fig. S2c), the Chr 18 *Dt2* signal was strengthened (Table 1; Fig. 3f), as was the Chr 6 signal, suggesting that the latter may be a “genetic background factor” that influences the phenotypic distinction between indeterminate and semi-determinate. Though the Chr 19 *Dt1* signal was further weakened, it was not eliminated, confirming that the GRIN-listed phenotypes for stem termination cannot automatically be assumed to have the corresponding two-locus *Dt* genotypes reported by Bernard (1972).

Flower Color

Regarding flower color in the 12,431 accessions that we used for an initial GWA, five of the nine known phenotypic categories were present (for their frequencies, see Fig. 2; for their codes, see Supplemental File S1): blue (*W1W1w2w2*), dark purple (*W1W1W3W3W4W4*), light purple (*W1W1W3W3w4w4*), purple (*W1W1w3w3W4W4*), near-white (*W1W1w3w3w4w4*), and white (*w1w1*) [see Yan et al. (2014) for phenotype–genotype details]. However, just four significant regions were detected in this GWA (Supplemental Fig. S3a) and all four were on Chr 13 not far from each other (Table 1; for a magnified view of the major SNP signal and dagger symbol indicating the cloned *W1–w1* locus position, see Fig. 3g). Limiting the GWA mapping to the 12,329 accessions that were just purple (8209) or white (4120) led to the detection of the same four Chr 13 signals (Supplemental Fig. S3b), though a new significant SNP signal appeared on Chr 19 at 36,603,029 bp (Table 1). This location is not consistent with the known chromosomal locations of all other flower color loci, though it is very close to the Chr 19 location of the *L1–l1* gene locus, whose dominant allele gives rise to a black versus brown or tan pod color phenotype (pod color is discussed later). If the Chr 19 flower color signal is not a false positive, then this Chr 19 signal could only have been detected if the underlying gene had an epistatic impact on the purple versus white phenotypic calls. In any event, the most significant SNP regions identified in the GWA analyses overlapped the cloned gene (Table 1). The other nearby significant SNP regions in Chr 13 resided about 1 to 2 Mb upstream of *W1* (Fig. 3g) and might simply arise from the extensive linkage disequilibrium in this region. Using fewer accessions, Sonah et al. (2015) detected 14 significant SNPs in a single region spanning 2.5 and 4.8 Mb (though their SNP maximum was 8.1 kb downstream of *W1–w1*), whereas Wen et al. (2015) reported five separate significant SNP signals ranging from 2,833,623 to 4,559,799 bp; the latter was their SNP maximum and it was the same SNP maximum detected in our study (Table 1).

Pubescence Color

Our initial GWA mapping of 12,360 accessions that included all tawny, light tawny, near-gray, and gray phenotypes (i.e., 6166, 425, 85, and 5684 accession frequencies, respectively; see Fig. 2 and see Supplementary File S1 for known genes) resulted in the identification of a strong SNP signal on Chr 6 but much weaker ones on Chr 3, 12, 14, and 20 (Supplemental Fig. S4a; Table 1). Of the 26 significant SNP signals located on Chr 6 from 17,258,654 to 19,815,389 bp (with a maximum SNP at 17,567,713 bp); one was at 18,252,495 bp, which colocalized with the position of the cloned *T–t* gene locus (Fig. 3h). For the same reasons noted in the stem termination and flower color sections, we attempted to mitigate epistasis by conducting a GWA without the gray pubescence color accessions (Supplemental Fig. S4b). We expected the *T–t* locus signal to disappear and the *Td–td* locus signal to be amplified because the GWA would then be focused solely on the pubescence color phenotypic variants inferred to arise from just a *TT* <u>*TdTd*</u> versus *TT* <u>*tdtd*</u> genotypic comparison. In that regard, we were nearly successful: the Chr 3 signal (which we infer to be the *Td–td* locus) was amplified 200-fold (Table 1; Fig. 3i) and the *T–t* locus signal was nearly extinguished. Our inference that the Chr 3 signal corresponds to the *Td–td* locus is supported by the findings of Wen et al. (2015), who in their GWA of 1402 lines for pubescence color, detected not only the *T–t* locus signal at 18,118,558 bp but also a significant Chr 3 signal at 47,244,893 bp (see their Figure 6A); however, they offered no commentary about that signal. Sonah et al. (2015) did not detect a Chr 3 signal in their GWA with 139 accessions but did detect a large region comprising 68 significant SNPs (i.e., 17,313,874–21,182,692 bp) associated with the cloned *T* locus, though their two closest SNPs consisted of one SNP 18.7 kb away and another 100 kb more distant. The borderline significant genomic regions on Chr 12, 14, and 20 detected in the initial GWA (Supplemental Fig. S4a) were deemed to be false positives, given that those signals disappeared in the second GWA analysis (Supplemental Fig. S4b).

Pubescence Form

In our initial GWA mapping of 12,104 accessions that were erect (7744), semierect (1474), or appressed (2886) (Fig. 2), two hig- resolution SNP signals were detected that we inferred to correspond to the *Pa1–pa1* and *Pa2–pa2* loci on Chr 12 and Chr 13, respectively (Supplemental Fig. S5a; see Supplemental File S1for known genes). To mitigate the impact of the epistasis and to better amplify the *Pa2–pa2* signal, only the 4360 accessions possessing the phenotypes of semierect (inferred to be *pa1pa1Pa2Pa2* genotypes) and appressed (inferred to be *pa1pa1pa2pa2*) were included in the next GWA (Supplemental Fig. S5b). Though the *Pa1–pa1* signal did not diminish much, the *Pa2–pa2* signal was amplified fourfold. The mapping resolution for the *Pa2–pa2* locus was remarkable, in that GWA SNP signal pinpointed a region of less than 50 kb (i.e., between 30,665,757 and 30,708,708 on Chr 13), whereas the *Pa1* locus mapped to a location between 37,036,017 and 37,786,243 bp on Chr 12 that spanned 750 kb (Table 1; for a magnified view of the two signals, see Fig. 3j,k). These two loci have not been cloned, so the SNP signals detected here could be useful starting points for researchers interested in doing so. The discrepancy between the Chr 11 map position that Lee et al. (1999) reported for the *Pa1–pa1* locus and the Chr 12 map position for that locus that we report here may be because these two chromosomes are highly homeologous (Lee et al., 2001). It is possible that the molecular markers used by Lee et al. (1999) may not have been homeology-specific, leading them to position *Pa1–pa1* on Chr 11 instead of Chr 12, where we mapped it.

Pubescence Density

Our GWA mapping of 12,397 accessions that exhibited six phenotypic categories of dense, glabrous, normal, slightly dense, sparse, or semisparse (for their frequencies, see Fig. 2) resulted in the identification of several SNP signals (Supplemental Fig. S6a; Table 1; Fig. 3l,m), two of which corresponded with the (SoyBase) linkage map positions of *Ps*–Ps^s–*ps* and *Pd1–pd1* loci (see Supplemental File S1 for known genes). These two signals translate into 34,877,806 bp on Chr 12 and 55,523,014 bp on Chr 1, respectively. This finding was consistent with the detection of two significant Pubescence Density QTLs: 1–2 (SoyBase: 35,314,290–37,138,680 on Chr 12) and 1–1 (SoyBase: 52,767,178–55,838,478 on Chr 1) reported by Komatsu et al. (2007) in a Japanese mapping population derived from a mating of a densely pubescent, insect-resistant cultivar with a sparsely pubescent, insect-susceptible cultivar. A strongly significant GWA signal was also detected at the top of Chr 14, along with a borderline significant signal on Chr 7 (Table 1), but neither one can be the *Pd2–pd2* locus, which is known to map to Chr 11 (Devine, 2003; Seversike et al., 2008). Limiting the GWA analysis to just the 12,301 accessions with the three phenotypes of sparse (73), semisparse (4305), and normal (7923) did not appreciably change the GWA results (Fig. 2): the *Pd1–pd1* and *Ps*–Ps^s–*ps* signals and the Chr 14 signal remained (Supplemental Fig. S6b), though the borderline significant Chr 7 signal disappeared. To narrow the GWA's focus on the *P1–p1* locus detected in the first GWA, we conducted a GWA using only the glabrous and normal accessions (35 and 7660, respectively; Fig. 2) and identified two separate but closely located significant regions of 11 and 8 SNPs corresponding to the *P1–p1* locus, one with a SNP maximum of 4,424,863 on Chr 9 and the other with a SNP maximum in the other region (46,139,114 on Chr 9) (Fig. 6c; Table 1); for a magnified view of these two adjacent positions, see Fig. 3n). Because of a very low frequency (just 35) of the glabrous accessions (Fig. 2), this GWA exhibited substantial noise but the Chr 9 *P1–p1* locus signals still stood out from that noise in terms of the strong $-\log_{10}P$-values (Supplemental Fig. S6c). Hunt et al. (2011) conducted transcriptional profiling of the two NILs, Clark-*p1p1* and Clark-*P1P1*, but mainly focused on *Glyma04g35130*, which was overexpressed in Clark-*p1p1*. They offered no commentary about *Glyma09g38410* (calreticulin-3 precursor) that was listed in their Table 1 as being overexpressed in Clark-*P1P1*. It has a Chr 9 bp position (i.e., 43,780,130–43,785,822) that falls within our Chr 9 SNP signal region (43,686,430–4,485,534), thus making *Glyma09g38410* a plausible candidate gene for *P1–p1* (Supplemental Table S2). Our GWA mapping results will be useful to those wishing to clone the *Ps*–Ps^s–*ps*, *Pd1–pd1*, and *P1–p1* loci, as well as the unknown gene locus corresponding to the strong Chr 14 SNP signal, particularly given that SoyBase lists no pubescence density QTLs on Chr 14.

Pod Color

A GWA using the 12,365 accessions that exhibited five pod color phenotypic variants of black (604), dark brown (225), brown (8258), light brown (76), and tan (3202) (Fig. 2) produced two highly significant SNP signals on Chr 19 and Chr 3, plus a significant SNP signal on Chr 1 (Supplemental Fig. S7a; Table 1; Fig. 3o,p; see Supplemental File S1 for known genes). The Chr 3 SNP signal corresponding to the *L2–l2* locus (brown and tan) spanned a 1091-kb region (i.e., 246,658–1,338,018 bp on Chr 3). The Chr 19 SNP signal corresponding to the *L1–l1* locus (black or brown + tan) spanned a 618-kb region (i.e., 37,503,524–38,121,212 bp on Chr 19). Interestingly, the Chr 1 SNP signal had a bp position nearly identical to the SNP signal detected for a gene locus governing green versus yellow seed coat color (discussed later). To determine if we could improve the resolution of each of the two main signals, we conducted a GWA with the 12,064 accessions exhibiting just the black, brown, and tan pod colors (Fig. 2); the results (Supplemental Fig. S7b) did not change much, suggesting that the 225 dark brown and 76 light brown phenotypes, which were included in the prior GWA but were omitted in this GWA, were simply slightly darker or lighter variants of the nominal brown phenotype. Our final GWA targeted only the accessions that had black (604) or brown (8258) pod colors (i.e., the respective genotypes of *L1L1*––/*l1l1 L2L2*) (Supplemental Fig. S7c); as expected, it resulted in the detection of only the *L1–l1* signal at a high significance level (Table 1; Fig. 3o). He et al. (2015) inferred that of the 13 gene candidates located in their fine-mapped Chr 19 *L2–l2* region, *Glyma19g27460* was the most likely candidate; however, that candidate gene has a bp position located ~2.75 Mbp downstream from our region of 48 significant SNPs (i.e., 36,397,778–38,521,183 on Chr 19). The reason for this substantive localization difference between our study and their study is not clear.

Seed Coat Luster

In GRIN, six phenotypic categories for this trait are listed: dense bloom, bloom, light bloom, dull, intermediate, and shiny. However, most accessions belong to the dull, intermediate, or shiny categories (Fig. 2; see Supplemental File S1 for known genes). We conducted an initial GWA using 12,278 accessions exhibiting five of those six categories (Supplemental Fig. S8a), and detected two separate significant signals on Chr 15 (Fig. 3q) corresponding to the tightly linked loci of *B–b* and the cloned *Hps–hps* (Gijzen et al., 2006). The strong signal at the top of Chr 9 was symbolized as a unnumbered *B?–b?* locus to distinguish it from the mapped *B1–b1* locus that corresponds to the weaker signal on Chr 13 (see Supplemental File S1 for known genes). The intermediate seed coat luster phenotype accounts for more than half (i.e., 7280) of the total accession set (Fig. 2) but the intermediate accessions tend not to have a consistent luster phenotype when grown in different environments. For that reason, we conducted another GWA omitting all of the intermediate accessions (Fig. 2), thereby using just the accessions that were classified as bloom (54), light bloom (76), dull (2516), and shiny (2352) phenotypes (Supplemental Fig. S8b) and it resulted in a reduction in the Chr 15 and Chr 9 signals (probably because of the loss of statistical power when going from about 12,000 to 5000 accessions). The Chr 13 (*B1–b1*) signal disappeared, whereas a new signal appeared on Chr 8 corresponding to gene locus *I–i*. The reason for the disappearance of the *B1–b1* signal is not clear, given that Chen and Shoemaker (1998) mapped *B1–b1* on the basis of dull versus shiny segregation. The appearance of the *I–i* signal may be related to the fact that the seed luster phenotype is more easily observed or called when the seed coats are fully pigmented (as in case of *ii* genotypes) as opposed to yellow seed coats (as in *II* genotypes). A borderline significant signal also appeared on Chr 11. We conducted a final GWA using just the accessions that exhibited just the dull (2516) or shiny (2352) phenotypes (Supplemental Fig. S8c). The Chr 15 (*B–b*) and Chr 9 (*B?–b?*) signals were restrengthened by this targeting (Table 1) and the significant SNP bp region (8,512,905–8,941,824 on Chr 15) just left of the *B–b* region overlapped the 8,868,741–8,875,714 bp region on Chr 15 of the cloned *Hps–hps* gene (Fig. 3q).

Seed Coat Color and Hilum Color

Only six seed coat color variants and only six hilum color variants had a phenotypic frequency of >0.01 (Fig. 2; see Supplemental File S1 for known genes). One GWA was focused on just the seed coat color variants (Supplemental Fig. S9a; *N* = 12,174) and another GWA was focused on just the hilum color variants (Supplemental Fig. S10a; *N* = 10,888), with both producing the expected signals corresponding to the respective *T–t* and $I–i^{k}–i^{i}–i$ loci on Chr 6 (18,766,611 bp) and Chr 8 (8,396,392 bp). Green seed coat color arises when chlorophyll does not degrade at seed maturity as it does in yellow seed coats (Woodworth, 1921). The green and yellow phenotypes are controlled by the single gene *G–g*, which has been mapped to Chr 1 (Cregan et al., 1999). We conducted a GWA restricted to just the 10,134 accessions exhibiting yellow (9021) or green (1113) seed coats (Supplemental Fig. S9b). As expected, the *T–t* and *I–i* signals disappeared and there was a threefold amplification of the $-\log_{10}P$-value of the Chr 1 (*G–g*) signal, which spanned a very small 25-kb region near the SNP maximum located at 52,253,980 bp on Chr 1 (Table 1; Fig. 3r). Relative to the cloned *I–i* locus (Todd and Vodkin, 1996; Tuteja et el., 2009), the seed coat color GWA (Supplemental Fig. S9a; Table 1) produced a signal that was weaker than the high-resolution stronger signal generated in the hilum color GWA (Supplemental Fig. S10; Table 1; Fig. 3s), primarily because the number of yellow seed coat accessions was far greater than the number of nonyellow accessions (Fig. 2). For hilum color, Sonah et al. (2015) reported 10 SNPs associated with the *I–i* locus on Chr 8, with a SNP maximum at 84,803,396 bp, which was not far from our SNP maximum at 8,572,686 bp. A signal for the *R–r* locus was detected in the initial GWA of hilum color, so we conducted a GWA targeting only the accessions with black (2391) or brown (2767) hilum color phenotypes (Fig. 2) whose respective inferred genotypes are *RR–rr.* That GWA amplified the Chr 9 *R–r* signal (Supplemental Fig. S10b; Table 1; Fig. 3t). The SNP region (i.e., 41,660,046–43,669,720 on Chr 9) brackets the cloned *R–r* locus bp position (Table 1). A low-level significant signal detected on Chr 12 may represent an unknown gene locus that somehow impacts the black versus brown classification. Finally, to locate the gene locus *O–o*, a GWA was performed on just the accessions with brown (2767) or red-brown (270) phenotypes (Supplemental Fig. S10c). The *I–i* and *O–o* loci are known to be linked (about 18 cM) on Chr 8 (Palmer et al., 2004); indeed, the GWA signal for *O–o* was identified at a comparable bp positon (i.e., 312 kb; 4,800,584–5,113,384 on Chr 8) just upstream from *I–i* (Table 1; Fig. 3u).

Phenotypic Variance and Distribution of the Mapped Genes

For each trait, the SNP signal with the largest effect explained the largest fraction of the total phenotypic variance, ranging from 11 to 59% (Fig. 4a), with the highest percentages observed for flower color (59%), pubescence color (52%), and hilum color (33%). For hilum and seed coat color, the cumulative effects of five genes (e.g., *G*, *I*, *O*, *R*, and *T)* explained up to 79 and 77% of total phenotypic variance, respectively. Overall, the cumulative contributions of all significant SNP signals to phenotypic variance explained about 48% on average, though it varied from 11 to 83% depending on the trait, which is comparable to the SNP associations identified in *A. thaliana* (Atwell et al., 2010), rice (Huang et al., 2010), and corn (Romay et al., 2013). In our study, the identified SNPs conferring new loci explained additional variation that ranged from 4 to 15%. For MG, the *E* genes (e.g., *E1*, *E2*, *E3*, and *E4*) explained 16% of phenotypic variance, whereas the non-*E* genes controlled an additional 7% of phenotypic variance. Population structure also explained some portion of total phenotypic variance, which ranged from 4 to 50%, with the highest proportion observed for MG (50%), probably because population structure is closely related to the latitudinal photoperiod sensitivity of soybean (Bandillo et al., 2015). It is possible, of course, that some portion of phenotypic variance may have arisen from imperfect linkage disequilibrium, imperfect

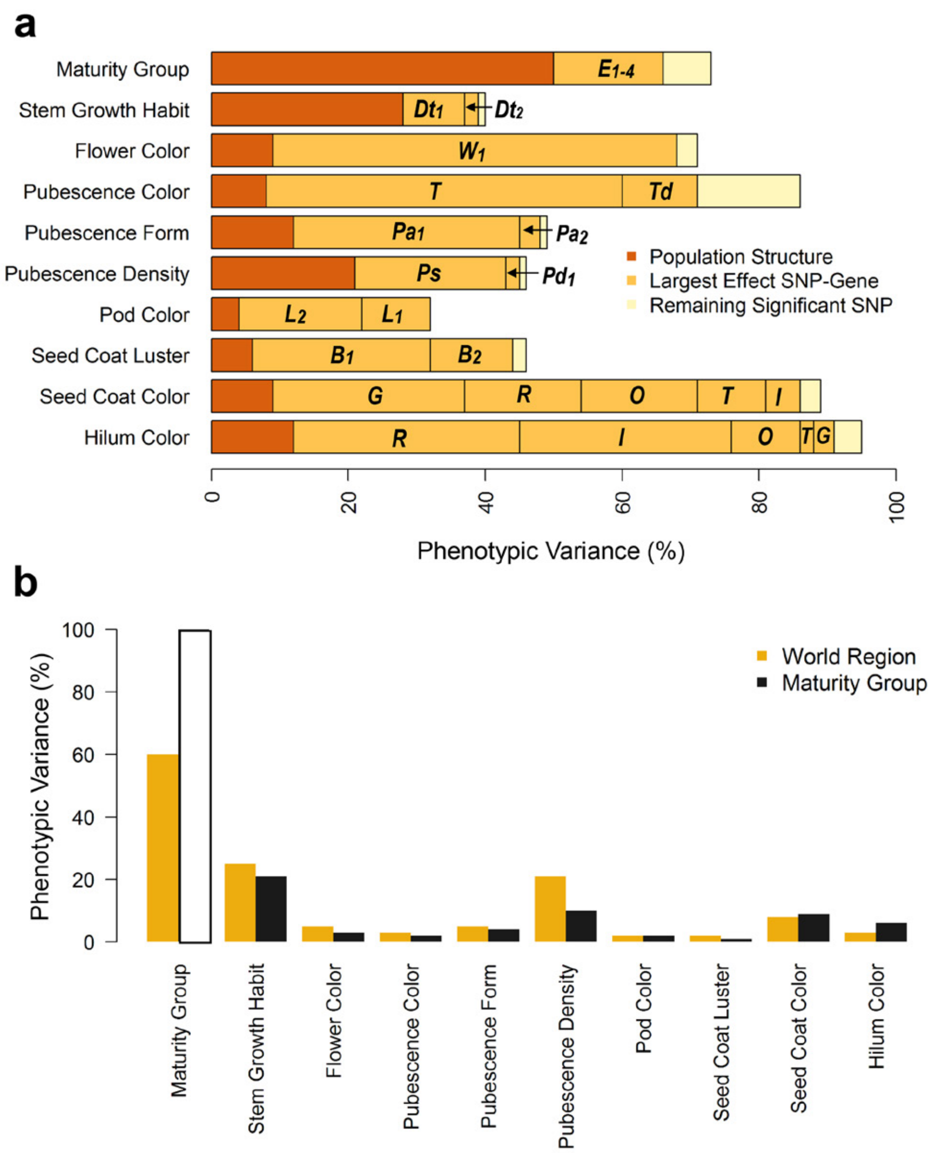


Fig. 4. (a) Contributions of significant single nucleotide polymorphisms (SNPs) and population structure (defined by ADMIXTURE *K* = 5) to phenotypic variance of each of 10 descriptive soybean traits. The proportion of phenotypic variance accounted for by significant SNPs was partitioned into large-effect SNPs (tagging known or candidate genes) and small-effect SNPs and calculated after accounting for population structure effects. (b) Contribution of world region and maturity group, which are the major determinants of population of structure within the collection, to phenotypic variance.

phenotyping, digenic epistasis, (undetected) small-effect modifier loci, or a combination of these.

The global distributions of narrowed loci revealed an essential pattern of allelic variation in gene loci that probably reflects a geographical-based difference in the history of soybean breeding (Supplemental Fig. S11). Several traits were found to be more highly correlated with world region other than MGs, indicating that *G. max* subpopulations are structured more by geography than by MG class (see Fig. 4b). The overall distribution patterns of allelic frequency at the various gene loci illustrate how accessions originating from China and North America (i.e., the United States and Canada) diverged from accessions originating from Japan and Korea. Between Japan and Korea, however, the allelic frequency spectrum was almost the same except for loci associated with maturity *(E3–e3*), flower color (*W1–w1*), pubescence (*Pa1–pa1*, *Pa2–pa2*, and *Ps–ps*), and hilum color (*O–o*, *R–r*). Similarly, China and North America differed in their allele frequency spectrum of loci associated with breeding and genetic improvement such as maturity (*E1–e1*, *E2–e2*, and *E3–e3*), stem growth habit (*Dt1* and *Dt2*), pubescence (*Pa1–pa1*, *Pa2–pa2, Ps–ps*, and *Pa1–pa1*), and seed coat or hilum color (*I–i* and *O–o*).

Breeding objectives are also factors contributing to the substantial allelic variation observed. For example, the degree of trichome density and the trichomes' orientation on the epidermal surfaces of soybean plants have been used in breeding aimed at deterring insect feeding or impairing the viability of insect larvae (Hulburt et al., 2004; Kanno, 1996; Lambert et al., 1992). Judging by the global distributions of *Ps–ps*, accessions in Japan and Korea are predominantly *Ps*, whereas America and China had predominantly *ps* (frequency >0.85). This is not surprising, given the fact that Japan has used sparse pubescence in their breeding programs as a key insect control strategy (Komatsu et al., 2007; Oki et al., 2012), whereas in the United States, erect and normal pubescence are needed to deter feeding by the potato leaf hopper (Broersma et al., 1972), which migrates in early summer northward from the Gulf Coast states where it overwinters each year (Illinois College of Agricultural, Consumer, and Environmental Sciences, 2017). Substantial allelic variation at some loci also might reflect cultural preferences and farming practices.

Summary

Genome-wide association analysis is nominally treated as a tool to be used mainly for dissecting the genetic architecture of quantitatively inherited traits. However, as documented here, GWA can also serve as a highly useful tool for detecting major qualitative genes governing categorically defined phenotype variants that exist for given traits in a germplasm collection. Indeed, we used GWA to identify the chromosomal bp positions of 24 classical genes governing the phenotypic variants listed for 10 key soybean descriptive traits. Because some classical genes have been cloned, we were able to show that the high-resolution SNP signal regions we detected for the trait phenotypic variants had chromosomal bp positions closely bracketing 22 of the 24 cloned genes; the two exceptions were the cloned *E4–e4* in Fig. 3d and the fine-mapped but not yet cloned *L1–l2* in Fig. 3o). Relative to the classical genes for which only imprecise cM map positions are available, GWA mapping resulted in strong, narrowly bounded SNP signals that essentially fine-mapped these genes (i.e., *Td–td*, *Pa1–pa1*, *Pa2–pa2*, *Ps–Ps*^{s}*–ps*, *P1–p1*, and *Pd1–pd1* in Fig. 3i –n; and *L2–l2*, *B–b*, *G–g*, and *O–o* on Fig. 3p–r, u). Researchers interested in cloning these genes will welcome this higher resolution mapping data as a good starting point. For traits governed by digenic epistasis, GWA was used to fine-map each of the two loci separately by creating phenotypic subsets of the accessions (e.g., Supplemental Fig. S4a,b). The power of using GWA to map qualitative trait genes is clearly substantial when categorical phenotype miss-calls are rare. Finally, it is interest to note that strong SNP signals were detected even for strong QTLs that have not yet been fully characterized (i.e., Chr 11 and Chr 12 maturity QTLs).

This demonstration that GWA mapping aimed at qualitatively inherited traits can be used to quickly generate high-resolution positions for the controlling genes on a genome sequence map is likely to be of interest to researchers in other crop species that have germplasm collections for which extensive data for qualitatively inherited traits also exist. We are now applying the GWA qualitative gene mapping protocol to all other qualitatively inherited soybean descriptor traits, and the results, when complete, will be documented in a forthcoming publication.

Supplemental Material

Table S1 is a spreadsheet listing the 13,624 *G. max* accessions and the corresponding phenotypic codes for each of 10 descriptive traits. Table S2 is a spreadsheet listing of (Glyma) candidate genes in a 250-kb window centered on each significant GWA SNP signal detected (excluding those for cloned genes). Supplemental File S1 tabulates the phenotypic variant names and category scores and presents a table of *N*-specific significance thresholds and information on the known genes governing the 10 descriptor traits. Supplemental Figure PDF contains Figures S1 to S11, which show the 10 Manhattan plots for each descriptor trait (maturity group, stem growth habit, flower color, pubescence color, pubescence form, pubescence density, pod color, seed coat luster, seed coat color, and hilum color) and the allele frequencies for the genes of for all 10 traits.

Conflict of Interest Disclosure

The authors declare that they have no conflicts of interest.

References

Alexander, D.H., J. Novembre, and K. Lange. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19:1655–1664. doi:10.1101/gr.094052.109

Atwell, S., Y.S. Huang, B.J. Vilhjalmsson, G. Willems, M. Horton, Y. Li, et al. 2010. Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. Nature 465:627–631. doi:10.1038/nature08800

Bandillo, N., D. Jarquin, Q. Song, R. Nelson, P. Cregan, J. Specht, et al. 2015. A population structure and genome-wide association analysis on the USDA soybean germplasm collection. Plant Genome 8:1–13. doi:10.3835/plantgenome2015.04.0024

Behm, J., L. Cerny, T. Floyd, J. Hall, and D. Wooten. 2011. Methods and compositions for increased yield. US Patent Application Publication. 2011/0010793:A1. Date issued: 13 Jan. 2011.

Bernard, R.L. 1967. The inheritance of pod color in soybeans. J. Hered. 58:165–168. doi:10.1093/oxfordjournals.jhered.a107575

Bernard, R.L. 1971. Two major genes for time of flowering and maturity in soybeans. Crop Sci. 11:242–244. doi:10.2135/cropsci1971.0011183X001100020022x

Bernard, R.L. 1972. Two genes affecting stem termination in soybeans. Crop Sci. 12:235–239. doi:10.2135/cropsci1972.0011183X001200020028x

Bernard, R.L. 1975a. The inheritance of near-gray pubescence color. Soybean Genet. Newsl. 2:31–33.

Bernard, R.L. 1975b. The inheritance of semi-sparse pubescence. Soybean Genet. Newsl. 2:33–34.

Bernard, R.L. 1975c. The inheritance of appressed pubescence. Soybean Genet. Newsl. 2:34–36.

Bernard, R.L., and B.B. Singh. 1969. Inheritance of pubescence type in soybeans: Glabrous, curly, dense, sparse, and puberulent. Crop Sci. 9:192–197. doi:10.2135/cropsci1969.0011183X000900020025x

Broersma, D.B., R.L. Bernard, and W.H. Luckmann. 1972. Some effects of soybean pubescence on populations of the potato leafhopper. J. Econ. Entomol. 65:78–82. doi:10.1093/jee/65.1.78

Cannon, E.K., and S.B. Cannon. 2011. CViT: "Chromosome Visualization Tool"—A whole-genome viewer. Int. J. Plant Genomics. 373875. doi:10.1155/2011/373875

Chen, Z., and R.C. Shoemaker. 1998. Four genes affecting seed traits in soybeans map to linkage group F. J. Hered. 89:211–215. doi:10.1093/jhered/89.3.211

Cober, E.R., S.J. Molnar, M. Harette, and H.D. Voldeng. 2010. A new locus for early maturity in soybean. Crop Sci. 50:524–527. doi:10.2135/cropsci2009.04.0174

Cregan, P.B., T. Jarvik, A.L. Bush, R.C. Shoemaker, K.G. Lark, A.L. Kahler, et al. 1999. An integrated genetic linkage map of the soybean. Crop Sci. 39:1464–1490. doi:10.2135/cropsci1999.3951464x

Devine, T.E. 2003. The *Pd2* and *Lf2* loci define soybean linkage group 16. Crop Sci. 43:2028–2030. doi:10.2135/cropsci2003.2028

Devlin, B., and K. Roeder. 1999. Genomic control for association studies. Biometrics 55:997–1004. doi:10.1111/j.0006-341X.1999.00997.x

Gai, J., Y. Wang, X. Wu, and S. Chen. 2007. A comparative study on segregation analysis and QTL mapping of quantitative traits in plants—with a case in soybean. Front. Agric. China 1:1–7. doi:10.1007/s11703-007-0001-3

Gijzen, M., S.S. Miller, K. Kuflu, R.I. Buzzell, and B.L.A. Miki. 1999. Hydrophobic protein synthesized in the pod endocarp adheres to the seed surface. Plant Physiol. 120:951–959. doi:10.1104/pp.120.4.951

Gijzen, M., R. Gonzales, D. Barber, and F. Polo. 2003a. Level of airborne Gly m 1 in regions of soybean cultivation. J. Allergy Clin. Immunol. 12:803–804. doi:10.1016/S0091-6749(03)01884-0

Gijzen, M., C. Weng, K. Kuflu, L. Woodrow, K. Yu, and V. Poysa. 2003b. Soybean seed lustre phenotype and surface protein cosegregate and map to linkage group E. Genome 46:659–664. doi:10.1139/g03-047

Gijzen, M., K. Kuflu, and P. Moy. 2006. Gene amplification of the *Hps* locus in *Glycine ma*x. BMC Plant Biol. 6:6. doi:10.1186/1471-2229-6-6

Gillman, J.D., A. Tetlow, J.-D. Lee, J.G. Shannon, and K. Bilyeu. 2011. Loss-of-function mutations affecting a specific *Glycine max* R2R3 MYB transcription factor result in brown hilum and brown seed coats. BMC Plant Biol. 11:155. doi:10.1186/1471-2229-11-155

Grant, D., R.T. Nelson, S.B. Cannon, and R.C. Shoemaker. 2010. SoyBase, the USDA-ARS soybean genetics and genomics database. Nucleic Acids Res. 38:D843–D846. doi:10.1093/nar/gkp798

He, Q., H. Yang, S. Xiang, D. Tian, W. Wang, T. Ahao, et al. 2015. Fine mapping of the genetic locus *L1* conferring black pods using a chromosome segment substitution line population of soybean. Plant Breed. 134:437–445. doi:10.1111/pbr.12272

Huang, X., X. Wei, T. Sang, Q. Zhao, Q. Feng, Y. Zhao, et al. 2010. Genome-wide association studies of 14 agronomic traits in rice landraces. Nat. Genet. 42:961–967. doi:10.1038/ng.695

Hulburt, D.J., H.R. Boerma, and J.N. All. 2004. Effect of pubescence tip on soybean resistance to lepidopteran insects. J. Econ. Entomol. 97:621–627. doi:10.1093/jee/97.2.621

Hunt, M., N. Kaur, M. Stomvik, and L. Vodkin. 2011. Transcript profiling reveals expression differences in wild-type and glabrous soybean lines. BMC Plant Biol. 11:145. doi:10.1186/1471-2229-11-145

Kanno, H. 1996. Role of leaf pubescence in soybean resistance to the false melon beetle, *Atrachya menetriesi* Faldermann (Coleoptera: Chrysomelidae). Appl. Entomol. Zool. (Jpn.) 31:597–603.

Illinois College of Agricultural, Consumer, and Environmental Sciences Extension and Outreach. 2017. Potato leafhopper. Illinois College of Agricultural, Consumer, and Environmental Sciences Extension and Outreach. http://extension.cropsciences.illinois.edu/fieldcrops/alfalfa/potato_leafhopper/ (accessed 1 May 2017).

Komatsu, K., S. Okuda, M. Takahashi, R. Matsunaga, and Y. Nakazawa. 2007. Quantitative trait loci mapping of pubescence density and flowering time of insect-resistant soybean (*Glycine max* L. Merr.). Genet. Mol. Biol. 30:635–639. doi:10.1590/S1415-47572007000400022

Lambert, L., R.M. Beach, T.C. Kilen, and J.W. Todd. 1992. Soybean pubescence and its influence on larval development and oviposition preference of lepidopterous insects. Crop Sci. 32:463–466. doi:10.2135/cropsc i1992.0011183X003200020035x

Lee, J.M., A.L. Bush, J.C. Specht, and R.C. Shoemaker. 1999. Mapping of duplicate genes in soybeans. Genome 42:829–836. doi:10.1139/g99-008

Lee, J.M., D. Grant, C.E. Valejos, and R.C. Shoemaker. 2001. Genome organization in dicots. II. *Arabidopsis* as a 'bridging species' to resolve genome evolution events among legumes. Theor. Appl. Genet. 103:765–773.

Li, J., and L. Ji. 2005. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. Heredity 95:221–227. doi:10.1038/sj.hdy.6800717

Li, W., D. Zheng, K. Van, and S. Lee. 2008. QTL mapping for major agronomic traits across two years in soybean (*Glycine max* L. Merr.). J. Crop Sci. Biotech. 11:171–190.

Lippert, C., J. Listgarten, Y. Liu, C.M. Kadie, R.I. Davidson, and D. Heckerman. 2011. FaST linear mixed models for genome-wide association studies. Nat. Methods 8:833–835. doi:10.1038/nmeth.1681

Liu, B., A. Kanazawa, H. Hatsumura, R. Takashashi, K. Harada, and J. Abe. 2008. Genetic redundancy in soybean photoresponses associated with duplication of the Phytochrome A gene. Genetics 180:995–1007. doi:10.1534/genetics.108.092742

Liu, B., S. Watanabe, T. Uchiyama, F. Kong, A. Kanazawa, Z. Xia, et al. 2010. The soybean stem growth habit gene *Dt1* is an ortholog of *Arabidopsis TERMINAL FLOWER1*. Plant Physiol. 153:198–210. doi:10.1104/pp.109.150607

Lorenzen, L.L., S. Boutin, N. Young, J.E. Specht, and R.C. Shoemaker. 1995. Soybean pedigree analysis using map-based molecular markers: I. Tracking RFLP markers in cultivars. Crop Sci. 35:1326–1336. doi:10.2135/cropsci1995.0011183X003500050012x

Lu, S., Y. Li, J. Wang, H. Nan, D. Cao, X. Li et al. 2015. Identification of additional QTLs for flowering time by removing the effect of maturity gene E1 in soybean. J. Integr. Agric. 15:60345.

McBlain, B.A., J.D. Hesketh, and R.L. Bernard. 1987. Genetic effects on reproductive phenology in soybean isolines differing in maturity genes. Can. J. Plant Sci. 67:105–116. doi:10.4141/cjps87-012

Oki, N., K. Komatsu, T. Sayama, M. Ishimoto, M. Takahashi, and M.M. Takahashi. 2012. Genetic analysis of antixenosis resistance to the common cutworm (*Spodoptera litura* Fabricius) and it relationship with pubescence characteristics in soybean (*Glycine max* (L.) Merr.). Breed. Sci. 61:608–617. doi:10.1270/jsbbs.61.608

Palmer, R.G., T.W. Pfeiffer, G.R. Buss, and T.C. Kilen. 2004. Qualitative genetics. In: H.R. Boerma and J.E. Specht, editors. Soybeans: Improvement, production, and uses. 3rd ed. Agron. Monogr. 16. ASA, CSSA, and SSSA, Madison, WI. p. 137–233.

Ping, J., Y. Liu, L. Sun, M. Zhao, Y. Li, M. She, et al. 2014. *Dt2* is a gain-of-function MADS-domain factor gene that specifies semideterminacy in soybean. Plant Cell 26:2831–2842. doi:10.1105/tpc.114.126938

Quinlan, A.R., and I.M. Hall. 2010. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics 26(6):841–842. doi:10.1093/bioinformatics/btq033

R Development Core Team. 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing.

Rincker, K., A.E. Lipka, and B.W. Diers. 2016. Genome-wide association study of brown Stem rot resistance in soybean across multiple populations. Plant Genome 9(2). doi:10.3835/plantgenome2015.08.0064

Romay, M.C., M.J. Millard, J.C. Glaubitz, J.A. Peiffer, K.L. Swarts, T.M. Casstevens et al. 2013. Comprehensive genotyping of the USA national maize inbred seed bank. Genome Biol. 14:R55. doi:10.1186/gb-2013-14-6-r55

Saindon, G., W.E. Beversdorf, and H.D. Voldeng. 1989a. Adjustment of the soybean phenology using the E4 locus. Crop Sci. 29:1361–1365. doi:10.2135/cropsci1989.0011183X002900060006x

Saindon, G., H.D. Voldeng, W.E. Beversdorf, and R.I. Buzzell. 1989b. Genetic control of long daylength response in soybean. Crop Sci. 29:1436–1439. doi:10.2135/cropsci1989.0011183X002900060021x

Seversike, T.M., J.D. Ray, J.L. Shultz, and L.C. Purcell. 2008. Soybean molecular linkage group B1 corresponds to classical linkage group 16 based on map location of the *lf2* gene. Theor. Appl. Genet. 117:143–147. doi:10.1007/s00122-008-0759-6

Shoemaker, R.C., and J.E. Specht. 1995. Integration of the soybean molecular and classical genetic linkage groups. Crop Sci. 35:436–446. doi:10.2135/cropsci1995.0011183X003500020027x

Sidak, Z. 1967. Rectangular confidence regions for the means of multivariate normal distributions. J. Am. Stat. Assoc. 62:626–633.

Sonah, H., L. O'Donoughue, E. Cober, I. Rajcan, and F. Belzile. 2015. Identification of loci governing eight agronomic traits using a GBS-GWAS approach and validation by QTL mapping in soya bean. Plant Biotechnol. J. 13:211–221. doi:10.1111/pbi.12249

Song, Q., D.L. Hyten, G. Jia, C.V. Quigley, E.W. Fickus, R.L. Nelson, and P.B. Cregan. 2013. Development and evaluation of SoySNP50K, a high density genotyping array for soybean. PLoS ONE 8:e54985. doi:10.1371/journal.pone.0054985

Song, Q., D.L. Hyten, G. Jia, C.V. Quigley, E.W. Ficus, R.L. Nelson, et al. 2015. Fingerprinting soybean germplasm and its utility in genomic research. G3 (Bethesda) 5:1999–2006. doi:10.1534/g3.115.019000

Tang, W.T., and G. Tai. 1962. Studies on the qualitative and quantitative inheritance of an interspecific cross of soybean, *Glycine max* × *G. formosana*. Bot. Bull. Acad. Sin. 3:39–54.

Tian, Z., X. Wang, R. Lee, Y. Li, J.E. Specht, R.L. Nelson, et al. 2010. Artificial selection for determinate growth habit in soybean. Proc. Natl. Acad. Sci. USA 107:8563–8568. doi:10.1073/pnas.1000088107

Todd, J.J., and L.O. Vodkin. 1996. Duplications that suppress and deletions that restore expression from a chalcone synthase multigene family. Plant Cell 8:687–699. doi:10.1105/tpc.8.4.687

Tsubokura, Y., H. Matsumura, M. Xu, B. Liu, H. Nakashima, T. Anai, et al. 2012. Genetic variation in the maturity locus *E4* is involved in adaptation to high latitudes in soybean. Agronomy 3:117–134. doi:10.3390/agronomy3010117

Tsubokura, Y., S. Watanabe, Z. Xia, H. Kanamori, H. Yamagata, A. Kaga, et al. 2013. Natural variation in the genes responsible for maturity loci *E1, E2, E3,* and *E4* in soybean. Ann. Bot. (Lond.) 113:429–441. doi:10.1093/aob/mct269

Turner, S.D. 2014. qqman: An R package for visualizing GWAS results using Q-Q and Manhattan plots. bioRxiv. http://biorxiv.org/content/early/2014/05/14/005165 (accessed 2 May 2017).

Tuteja, J.H., G. Zabata, K. Varala, M. Hudson, and L.O. Vodkin. 2009. Endogenous, tissue-specific short interfering RNAs silence the chalcone synthase gene family in *Glycine max* seed coats. Plant Cell 21:3063–3077. doi:10.1105/tpc.109.069856

Wang, M., N. Jiang, T. Jia, L. Leach, J. Cockram, R. Waugh, et al. 2012. Genome-wide association mapping of agronomic and morphological traits in highly structured populations of barley cultivars. Theor. Appl. Genet. 124:233–246. doi:10.1007/s00122-011-1697-2

Watanabe, S., R. Hideshima, Z. Xia, Y. Tsubokura, S. Sato, Y. Nakamoto, et al. 2009. Map-based cloning of the gene associated with the maturity gene locus *E3*. Genetics 182:1251–1262. doi:10.1534/genetics.108.098772

Watanabe, S., K.Z. Xia, R. Hideshima, Z. Xia, Y. Tsubokura, S. Sato, et al. 2011. A map-based cloning strategy employing a residual heterozygous line reveals that the *GIGANTEA* gene is involved in soybean maturity and flowering. Genetics 188:395–407. doi:10.1534/genetics.110.125062

Wen, Z., J.F. Boyse, Q. Song, P.B. Cregan, and D. Wang. 2015. Genomic consequences of selection and genome-wide association mapping in soybean. BMC Genomics 16:671. doi:10.1186/s12864-015-1872-y

Woodworth, C.M. 1921. Inheritance of cotyledon, seed-coat, hilum, and pubescence colors in soybeans. Genetics 6:487–553.

Woodworth, C.M. 1932. Genetics and breeding in the improvement of the soybean. Bull. Agric. Exp. Stn. (Illinois) 384:297–404.

Xia, Z., S. Watanabe, T. Yamada, Y. Tsubokura, H. Nakashima, H. Zhai, et al. 2012. Positional cloning and characterization reveal the molecular basis for soy bean maturity locus *E1* that regulates photoperiodic flowering. Proc. Natl. Acad. Sci. USA 109:E2155–E2164. doi:10.1073/pnas.1117982109

Yan, F., S. Di, F.R. Rodas, T.R. Torrico, Y. Murai, T. Iwashina, et al. 2014. Allelic variation of soybean flower color gene *W4* encoding dihdroflavonol 4-reductase 2. BMC Plant Biol. 14:58. doi:10.1186/1471-2229-14-58

Zabala, G., and L.O. Vodkin. 2003. Cloning of the pleiotropic *T* locus in soybean and two recessive alleles that differentially affect structure and expression of the encoded flavonoid 3′ hydroxylase. Genetics 163:295–309.

Zabala, G., and L.O. Vodkin. 2007. A rearrangement resulting in small tandem repeats in the F3‘5’H genes of white flower genotypes is associated with the soybean *W1* locus. Plant Genome 2:113–124.

Zabala, G., and L.O. Vodkin. 2014. Methylation affects transposition and splicing of a large CACTA transposon from a MYB transcription factor regulating anthocyanin synthase genes in soybean seed coats. PLoS ONE 9:e111959. doi:10.1371/journal.pone.0111959

Zhai, H., S. Lu, Y. Wang, X. Chen, H. Ren, J. Yang, et al. 2014. Allelic variation at four major maturity *E* genes and transcriptional abundance of the *E1* gene are associated with flowering time and maturity of soybean cultivars. PLoS One 9:e97636. doi:10.1371/journal.pone.0097636

Zhao, C., R. Takeshima, J. Zhu, M. Xu, M. Sato, S. Watanabe, et al. 2016. A recessive allele for delayed flowering at the soybean maturity locus *E9* is a leaky allele of *FT2a*, a *FLOWERING LOCUS T* ortholog. BMC Plant Biol. 16:20. doi:10.1186/s12870-016-0704-9