


# Meta-GWAS for quantitative trait loci identification in soybean

Johnathon M. Shook,<sup>1</sup> Jiaoping Zhang,<sup>1</sup> Sarah E. Jones,<sup>1</sup> Arti Singh,<sup>1</sup> Brian W. Diers,<sup>2</sup> and Asheesh K. Singh <sup>1,\*</sup>

<sup>1</sup>Department of Agronomy, Iowa State University, Ames, IA 50011, USA

<sup>2</sup>Department of Crop Sciences, University of Illinois, Urbana, IL 61801, USA

\*Corresponding author: Email: singhak@iastate.edu

## Abstract

We report a meta-Genome Wide Association Study involving 73 published studies in soybean [*Glycine max* L. (Merr.)] covering 17,556 unique accessions, with improved statistical power for robust detection of loci associated with a broad range of traits. *De novo* GWAS and meta-analysis were conducted for composition traits including fatty acid and amino acid composition traits, disease resistance traits, and agronomic traits including seed yield, plant height, stem lodging, seed weight, seed mottling, seed quality, flowering timing, and pod shattering. To examine differences in detectability and test statistical power between single- and multi-environment GWAS, comparison of meta-GWAS results to those from the constituent experiments were performed. Using meta-GWAS analysis and the analysis of individual studies, we report 483 peaks at 393 unique loci. Using stringent criteria to detect significant marker-trait associations, 59 candidate genes were identified, including 17 agronomic traits loci, 19 for seed-related traits, and 33 for disease reaction traits. This study identified potentially valuable candidate genes that affect multiple traits. The success in narrowing down the genomic region for some loci through overlapping mapping results of multiple studies is a promising avenue for community-based studies and plant breeding applications.

**Keywords:** meta-analysis; GWAS; agronomic traits; seed composition traits; disease resistance

## Introduction

Genome-wide association studies (GWAS) analyze the association between a trait of interest and thousands of genetic variants throughout the genome. The general approach has benefited from the development of greatly increased numbers of markers due to the advent of next-generation sequencing approaches, and increased sample size with the formation of biobanks, such as the 100,000 Genomes Project (2019). Plant scientists now routinely conduct GWAS in crop species, including soybean [*Glycine max* (L.) Merr.]. Increased marker data availability and development of new statistic methods provided great opportunities to gain new knowledge from existing data and address the previous lacuna of GWAS experiments (Bandillo et al. 2015, 2017; Zhang et al. 2015, 2017; Zhou et al. 2015; Chang et al. 2016; Chang and Hartman 2017; de Azevedo Peixoto et al. 2017; Zeng et al. 2017).

Researchers have recognized that while single environment GWAS such as those conducted in the greenhouse are powerful for genetic studies and candidate gene identification, their extrapolation in field environment applications require further validation (Zhang et al. 2015; Coser et al. 2017; de Azevedo Peixoto et al. 2017; Moellers et al. 2017). When comparing separate studies of the same trait, significant differences in results are often found. These differences may be caused by allele frequency variation between populations, inadequate control of population structure, or environmental dependencies (Bubeck et al. 1993). With the availability of standardized marker data across the

USDA soybean germplasm collection (Song et al. 2015), several studies have mapped important major effect quantitative trait loci (QTL) using historical records and GWAS analysis: for example, insect resistance (Chang and Hartman 2017), disease resistance (Chang et al. 2016), descriptive traits such as flower and pubescence color (Bandillo et al. 2017), and seed oil and protein content (Bandillo et al. 2015). However, for many quantitative traits such as seed composition or plant height, using raw measurements from differing environments introduces bias, which may erode the power of detection for significant QTL (Chen et al. 2010). While results from within the same environment(s) share a common environmental component, attempting to combine multiple panels grown in different environments leads to an improper assignment of environmental effects to the differences between genetics of the panels involved (Zhao et al. 2019). Meta-analysis provides an attractive alternative to address the above-mentioned challenges of individual GWAS, and this analysis can be performed on results from independent studies using statistical approaches such as those provided by the analysis program METAL (Willer et al. 2010).

Quantitative traits, in contrast with qualitative traits, are controlled by many genes and environmental factors. To provide greater understanding of the genetic and metabolic networks that regulate these traits, interactions between previously discovered genes and new candidate genes can be added to the existing models. Directly measured traits often comprise only a portion of

Received: October 12, 2020. Accepted: April 02, 2021

© The Author(s) 2021. Published by Oxford University Press on behalf of Genetics Society of America.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs licence (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial reproduction and distribution of the work, in any medium, provided the original work is not altered or transformed in any way, and that the work is properly cited. For commercial re-use, please contact [journals.permissions@oup.com](mailto:journals.permissions@oup.com)

the information about a biological pathway, necessitating the identification of pleiotropic effects (on correlated traits) for an increased biological understanding of the phenotype. Genes may exhibit pleiotropy either through control of a common pathway such as the influence of *Dt1* on both plant height and lodging (Diers *et al.* 2018), or through multiple effects of a chemical as seen in the effect of *T* locus that has a dual role in pigmentation and chilling tolerance through isoflavones (Takahashi and Asanuma 1996). Identifying genes that control multiple phenotypes of importance can either suggest candidates for fixation, in cases where both effects are positive, or may identify possible penalties associated with incorporating particular alleles and improve multi-trait selection results (Bolormaa *et al.* 2014).

Meta-analyses include separately analyzing each individual experiment in order to determine experiment-specific *P*-value and allele effect estimates, rather than performing a combined analysis to leverage extensive data (Bandillo *et al.* 2015). Further genetic insights can be gleaned through an ease in the identification of pleiotropic effects due to the analysis of a wide range of traits. Moreover, the ability to compare the results from a meta-analysis with those from separate analyses of individual studies allows for the identification of both environment-dependent associations and for the enrichment and detection of rare alleles from more unique but diverse populations. Previous results have shown the effectiveness of combined panels to identify minor genes that were missed in a single study (Chang *et al.* 2017). Due to the need for adequate representation of minor alleles in GWAS, rare alleles that are predominant in a small zone of adaptation may be absent or undetectable within individual studies. The agronomic screenings for the USDA soybean germplasm collection are arranged based on the influx of new germplasm into the United States, and therefore serve as a semi-randomized subset of global soybean variation and spatiotemporal patterns in the origins of new accessions enabling potential detection of rare variants, which may be enriched in one of these geographical regions (Trotta *et al.* 2016).

While combined analyses for disease and insect resistance (Chang *et al.* 2016; Chang and Hartman 2017) and seed composition (Bandillo *et al.* 2015) have previously been reported, we perform a large-scale meta-analysis utilizing individual studies in soybean. Our study takes a two-pronged approach: first, each individual germplasm characterization study is subjected to traditional GWAS analysis, with the inclusion of quantitative traits of agronomic importance, stress tolerance, and seed composition. Following the initial GWAS analysis, studies of the same trait undergo meta-analysis. The multitude of traits examined with our study facilitates the detection of co-localized peaks indicative of potential pleiotropic effects of genes across a diverse range of phenotypes. Subsequent study of pleiotropic genes and reporting on gene-rich clusters can be useful when attempting to introgress favorable alleles into breeding lines (Cameron *et al.* 2017), as it improves the understanding of potential complications of introgression. Loci associated with multiple traits identified within this study require additional functional validation, as GWAS are not designed to definitively differentiate between pleiotropy and lack of recombination. We included results from reports published from 1964 to 2009 for a total of 73 individual studies. The design of this study was intended to identify co-localization of peaks for multiple traits, as well as to identify previously overlooked genes through meta-analysis approaches. The meta-GWAS approach differs from the original GWASs in that it can be continually added to as new studies are performed, with each new study increasing the power to detect marker-trait

associations by increasing the sample size. Two main approaches have been used for performing meta-analysis on GWAS results: a *P*-value and effect-sign based approach, and an effect size approach (Zeggini and Ioannidis 2009). Using meta-GWAS analysis and analysis of individual studies, we report 393 unique peaks including 66 candidate genes across important traits and provide confirmation of many previously reported genes. This study provides targets for functional characterization and introgression of previously untapped diversity for many important traits.

## Materials and methods

### Genotypic data and quality control

Marker data from the testing of 20,087 *G. max* and *G. soja* accessions from the USDA Soybean Germplasm Collection with the SoySNP50K iSelect BeadChip (Song *et al.* 2013) were downloaded from SoyBase ([www.soybase.org/dlpages/#snp50k](http://www.soybase.org/dlpages/#snp50k); last accessed April 13, 2021). A data imputation pipeline based on Java implementation of Beagle 5.0 (Browning and Browning 2016) was utilized to impute missing data for the 42,080 SNP markers that were aligned to the Williams 82 reference genome v2 assembly. Markers aligned to scaffolds but not assigned to a chromosome were removed prior to processing. Ten burn-in iterations and five phasing iterations were used to impute missing markers, which accounted for 0.64% of all markers. For each test, markers remaining after applying cutoffs of minor allele frequency  $\geq 0.05$  for studies involving  $300 \leq n \leq 1000$  accessions, or 0.01 for studies involving  $n \geq 1001$  accessions, were selected for further analysis.

### Phenotypic data and genetic accessions

Quantitative phenotypic data from USDA reports were compiled from the U.S. National Plant Germplasm System website (<http://npgsweb.ars-grin.gov/gringlobal/descriptors.aspx>, Descriptors for Soybean 2019). Subsets of accessions that were part of historical USDA germplasm characterization trials with a size  $n \geq 300$  were selected for further analysis. Information on the design of the original trials is available from the technical bulletins in which they were originally published. These technical bulletins are available online in part at <https://pubs.nal.usda.gov/sites/pubs.nal.usda.gov/files/tb.htm> (Miller 2003). Alternatively, PDFs of the technical bulletins are available on our GitHub (<https://github.com/SoylabSingh/META-GWAS>). Additional traits, such as disease resistance and amino acid composition, were downloaded from the NPGS website.

### Genome-wide association analysis

Each experiment was analyzed separately with a mixed linear model implemented using GAPIT in R (Lipka *et al.* 2012) to prevent confounding of environmental effects with marker effects, which would be expected for several traits (*i.e.*, flowering time, oil, protein, and so on). Population structure was controlled using the first three PCAs based on the marker data. This resulted in 585 combinations of experiment/trait analyses. Analysis was subsequently performed for combined panels for each trait. The Bonferroni threshold (Neyman and Pearson 1928) was employed to minimize the likelihood of false positives in declaring significance. The significant SNPs were compiled for further analysis (Supplementary Table S1).

### Meta-analysis

For each trait with two or more studies, a meta-analysis was performed using METAL (Willer *et al.* 2010). The *P*-value, direction of effect, and sample size were utilized to carry out a sample size

weighted analysis, with additional genomic control correction performed based on the difference between the median test statistic and that expected by chance.

## Candidate gene identification

Initial peak calling was performed trait-by-trait based on marker position. Subsequently, peaks for related traits (such as flowering date and maturity date) with substantial overlap were merged, resulting in fewer unique peaks than originally called. Local LD decay analysis was used to further clarify between separate or overlapping peaks.

Markers that were significant for multiple traits and experiments, or were identified during meta-analysis of the results, were examined for nearby candidate genes. Candidate genes were identified by examining annotated genes within linkage disequilibrium (LD) of the leading SNP with  $r^2 > 0.7$  for each experiment and peak (de Azevedo Peixoto et al. 2017). Candidate gene identification was performed based on previously characterized genes, gene family function, and distance from lead SNP <100 kbp. For candidate causal genetic variant analysis, we utilized the SNP dataset from the genome resequencing study of 302 soybean lines (Zhou et al. 2015) and searched the possible causal mutants at the identified candidate genes. We first identified the lead SNP from peaks of interest in the resequencing dataset, then calculated the pairwise LD  $r^2$  values between the lead SNP and the SNPs covering the locus of candidate gene. All other analyses here within were aligned to the Wm82.a2 reference genome (<https://soybase.org/gb2/gbrowse/gmax2.0/>; last accessed April 13, 2021). The R package “circlize” was employed to generate the circular visualizations of significant SNPs for multiple traits throughout the genome (Gu et al. 2014). Study names have been shortened for convenience within the text; a reference file is provided to find the initial source of phenotypic data used in this work (Supplementary Table S2). Trait definitions, as well as the number of peaks and candidate genes identified for each trait, are provided in Supplementary Table S3. A process workflow diagram can be found as Figure 1.

## Data availability

The authors affirm that all data necessary for confirming the conclusions of the article are present within the article, figures, and tables. Raw data, supplementary, and code files are available at <https://github.com/SoylabSingh/Soy-Meta-GWAS>.

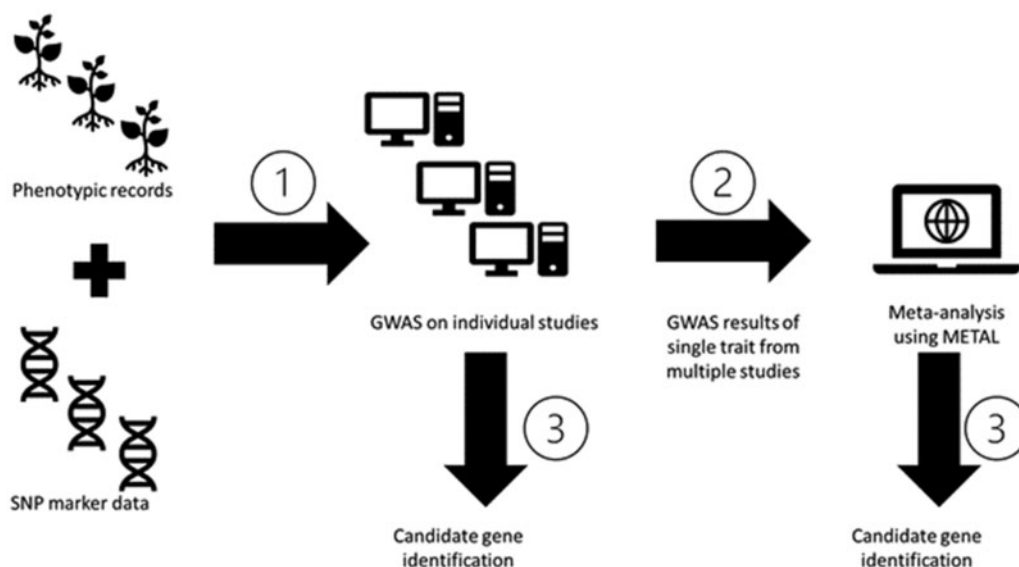
## Results and discussion

From the individual study GWAS and meta-GWAS 4919 significant SNPs were detected, of which 787 were reported from the meta-GWAS analysis. Complete listing of the significant SNP identified using individual study GWAS and meta-GWAS are provided in Supplementary Table S1. Among these 787 SNPs identified using meta-GWAS, 110 were associated with agronomic traits, 106 with seed composition traits, and 571 with disease resistance traits. Overall, candidate genes were assigned for 65 unique loci; and these included genes with moderate to large effects. We focus our results on loci that were associated with multiple traits.

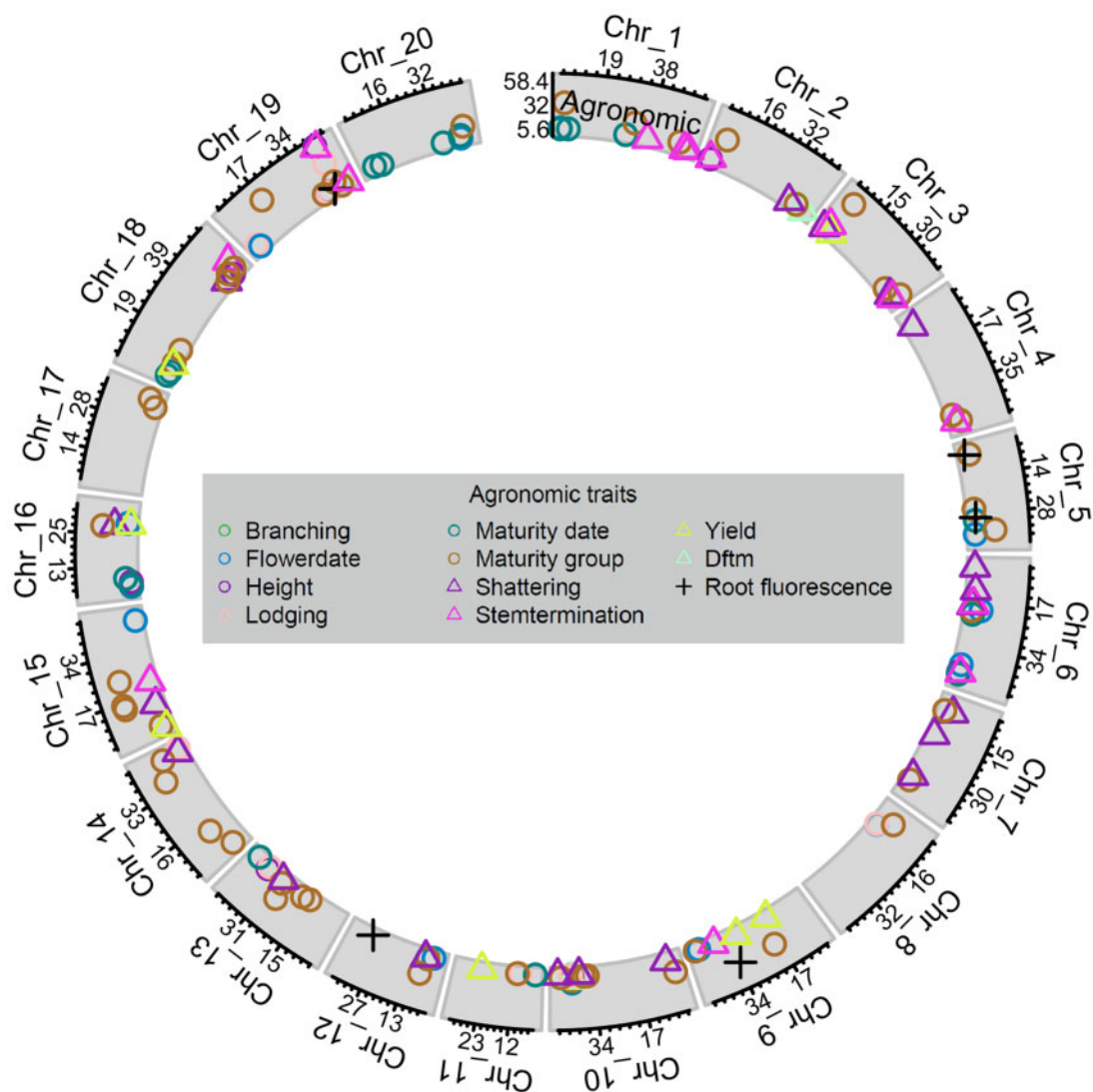
### Agronomic traits

Among agronomic traits, we identified 1422 marker-trait associations with traditional GWAS studies, as well as 110 SNPs associated with agronomic traits when analyzed across studies by meta-GWAS. In all, 115 peaks across 20 chromosomes were identified, with 17 candidate genes (Figure 2, Table 1, Supplementary Tables S1 and S3).

In our approach, we used results from individual studies to detect overlapping genomic regions for the purpose of locating candidate genes for traits, including for genes previously cloned. The locus harboring Dt1 (*Glyma.19g194300*) (Liu et al. 2010), the major gene conditioning stem termination in soybean, was significantly associated with oleic acid and linoleic acid content, as well as plant height, stem termination, and stem lodging (Supplementary Table S1). By comparing the mapping results of four studies, we were able to limit the candidate genomic region to a 125 kb fragment harboring previously cloned Dt1 (from



**Figure 1** Process workflow. In step 1, GWAS is performed on each trait/study combination. In step 2, the P-value, allele effect, and standard error from each GWAS for a given trait are subjected to meta-analysis using METAL. In step 3, candidate gene identification is performed in parallel for peaks detected in individual studies and the meta-analysis.



**Figure 2** Significant SNPs from GWAS from individual studies and meta-GWAS for agronomic and developmental traits. Symbol position along the x-axis shows the position (in Mb) along the chromosome, while y-axis symbol position shows the LOD score of the lead SNP for each QTL. The x-axis labels indicate position (in Mb) of tertile points, while y-axis labels show minimum, maximum, and middle of LOD score range for the given trait class. Shape and color correspond to unique traits.

ss715635422 to ss715635460) (Supplementary Figure S1). These results highlight the advantages of meta-GWAS for finer mapping the candidate gene region. A nonsynonymous SNP (SNP\_19\_44980087), in high LD ( $r^2 = 0.5$ ) with the leading SNP ss715635424 (also known as SNP\_19\_45000827), was found at the fourth exon of *Dt1* that changes amino acid R (Arg) to W (Trp) (Supplementary Figure S2). This SNP is identical to the R166W mutation previously identified (Liu et al. 2010).

On chromosome 19, we identified a peak for stem lodging which was on the opposite end of the chromosome as *Dt1*. Stem lodging is associated with plant height and this has been reported in multiple crops (Flint-Garcia et al. 2003; Diers et al. 2018; Singh et al. 2019). As lodging causes significant yield and quality losses, the development of the shorter statured wheat and rice were promoted which could better handle high input agriculture. However, this solution is not universally applicable. In soybean, pods are arranged at nodes on the stem, so a reduction in height through decreased node number may reduce yield potential. Leveraging four studies, we report a peak for tolerance to stem lodging with the candidate gene *Glyma.19g016400*, an ABC

transporter on chromosome 19. This locus was found to affect lodging tolerance but was not found to be associated with plant height, thereby making it a useful target to develop lodging resistant soybean cultivars without decreasing stem length and yield potential. While this is the first genome-wide association study identifying this gene, additional evidence towards its validity comes from several recent patents (US Patents #8697941, 8748695, and 9675071) that relate to molecular markers in the region of interest and include *Glyma.19g016400* as one of the candidate genes for PPO inhibitor tolerance in soybean. Significant effects of this region for seed yield, lodging, and plant height were reported from the SoyNAM project (Diers et al. 2018). The results from Hulting et al. (2001) on PPO inhibitor tolerance and our findings on stem lodging susceptibility suggest a tradeoff between PPO inhibitor tolerance and lodging susceptibility. The soybean accessions highly tolerant to sulfentrazone contain alleles associated with increased lodging in our study, necessitating further studies to validate these observations.

On chromosome 6, a significant SNP peak was identified that co-located with the *T* gene, a flavonoid 3' hydroxylase (Toda et al.

**Table 1** List of candidate genes identified for agronomic traits using GWAS from individual studies and meta-GWAS

| Chromosome | Likely gene                 | Meta-GWAS | Individual studies GWAS | Trait(s)   | Studies source  |
|------------|-----------------------------|-----------|-------------------------|--|---|
| 5          | <i>Glyma.05G200100</i>      |           | *                       | Flower date, Maturity date, Maturity group                           | 4il87, ms1999.01, ms923   |
| 6          | E1                          | *         | *                       | Flower date, Maturity date, Maturity group, Stem termination         | 1il64, 1il66, 2il81.1, 2il81.2, 4il87, 5il90, il0102, il989, meta, mn945  |
|            | <i>Glyma.06G068900</i>      | *         | *                       | Seed mottling  | 3mn83.2, meta   |
|            | <i>Glyma.06g134400</i>      | *         | *                       | Pod shattering (early), Pod shattering (late)                        | 4il87   |
|            | T                           | *         | *                       | Seed mottling  | 3il84, meta, ms1999.01, ms923, ms967  |
| 7          | <i>Glyma.07g049800</i>      | *         | *                       | Pod shattering (early), Pod shattering (late)                        | 3il84, meta, ms1999.01, ms923   |
| 8          | I                           | *         | *                       | Seed mottling  | 1il66, 2ky81, 4il87, il0102, ms923  |
| 9          | <i>fr1</i>                  |           | *                       | Root fluorescence  | fluorjt97   |
|            | <i>Glyma.09g090600</i>      | *         | *                       | Seed mottling  | 1il66, 4il87, meta  |
|            | <i>Glyma.09g266200</i>      | *         | *                       | Flower date, Maturity group  | ms923, ms1999.01  |
| 10         | E2                          | *         | *                       | Branching, Flower date, Height, Maturity date, Maturity group, Yield | 1il64, 1il66, 2il81.1, 3il83.1, 3il84, il0102, il989, meta, ms1999.01, ms967  |
| 11         | K1/AGO                      | *         | *                       | Seed mottling  | 3mn83.2, il0102, meta, ms923, ms967   |
| 13         | Rsv1                        | *         | *                       | Seed mottling  | 1il66, 2il81.1, 2il81.2, 5il90, meta, ms1999.01, ms2000.02, ms923   |
| 14         | <i>fan1</i>                 |           | *                       | Seed quality   | 2ky81   |
| 15         | <i>Glyma.15g139800</i>      | *         | *                       | Pod shattering (early), Pod shattering (late)                        | 1il66, 2il81.2, 2ky81, meta   |
| 16         | E9                          | *         | *                       | Flower date, Maturity group  | 2il81.1, 3il83.1, meta, ms1999.01   |
|            | <i>Pdh1</i>                 | *         | *                       | Pod shattering (early), Pod shattering (late)                        | 1il64, 2il81.1, 4il87, il0102, meta, ms1999.01, ms2000.02, ms923, ms967   |
| 18         | Dt2                         | *         | *                       | Stem termination   | meta, mn945, ms923  |
| 19         | ABC, <i>Glyma.19g016400</i> | *         | *                       | Lodging  | 1il66, 2ky81, ms923, 3il84, meta  |
|            | Dt1, <i>Glyma.19g194300</i> | *         | *                       | Height, Lodging, Stem termination                                    | 1il64, 1il66, 2il81.1, 2il81.2, 2ky81, 3il83.1, 3il84, 3mn83.2, 4il87, 5il90, il0102, meta, mn945, ms1999.01, ms2000.02, ms923, ms967 |
|            | E3                          |           | *                       | Maturity group   | 2il81.2   |

\* Bonferroni corrected P-value threshold [p-value 0.05 / number of markers].

2002). This region was significant for arginine, cysteine, isoleucine, and leucine levels, as well as for seed mottling (Figure 3). The cloned E2 locus (Watanabe et al. 2011) was significantly associated with flowering and maturity date, maturity group, days from flowering to maturity, plant height, and seed yield (Figure 2). The associations between E2 and these traits have been previously reported (Fang et al. 2017).

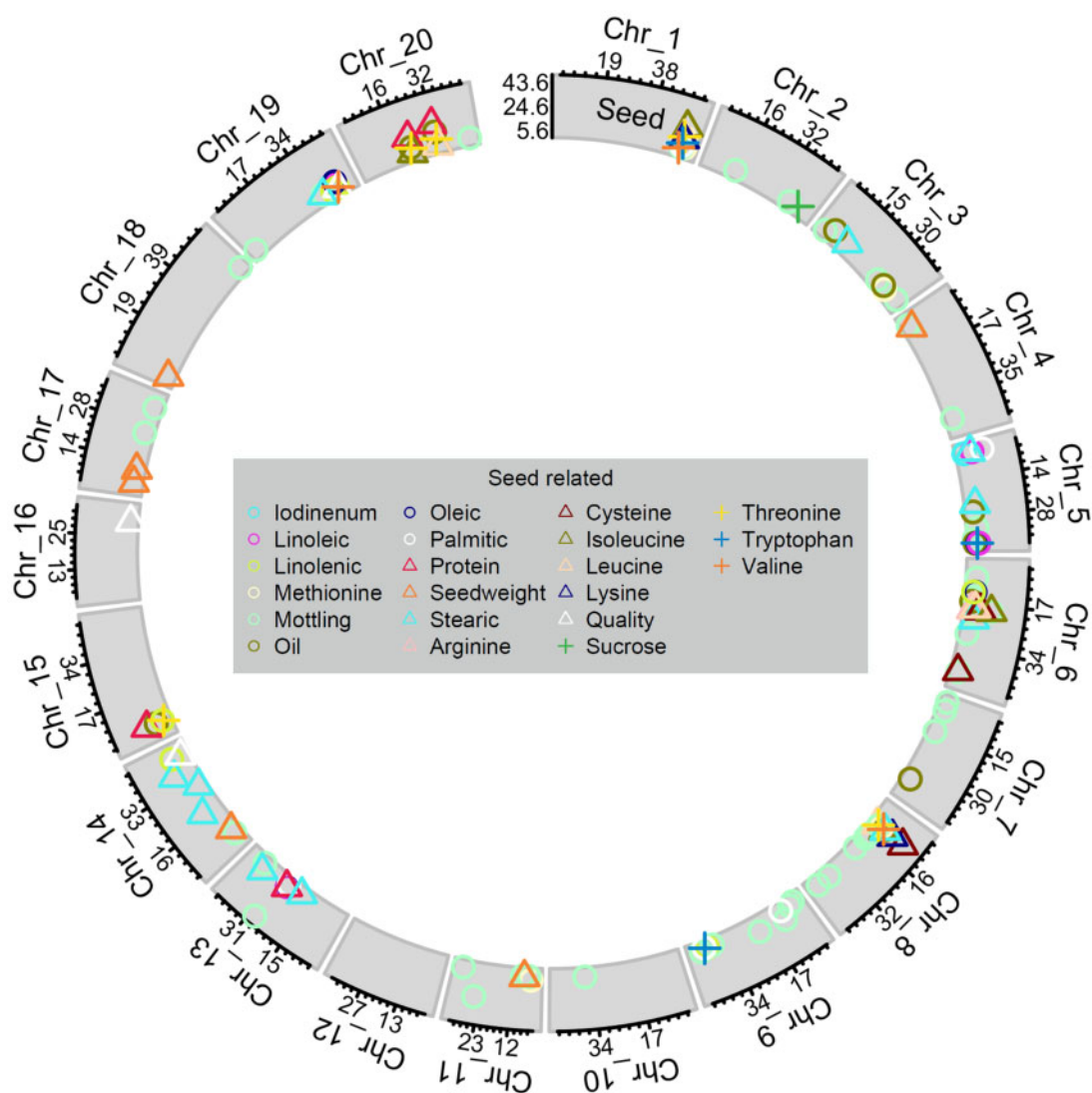
### Seed composition traits

Among seed composition traits, we identified 1364 marker-trait associations with traditional GWAS studies, as well as 106 SNPs associated with compositional traits when analyzed across studies by meta-GWAS. SNPs associated with composition were found on chromosomes 1–9, 11, 13–15, 17, and 19–20, resulting in 88 peaks with 19 candidate genes (Figure 3, Table 2, Supplementary Tables S1 and S3).

A cluster of candidate genes for seed composition, including isoleucine, methionine, leucine, tryptophan, threonine, lysine, and palmitic acid, were located in a region of 30 kb on chromosome 1 between 53.13 and 53.16 Mb, 4 a cysteine desulfurase

(*Glyma.01g197100*) and a malate and lactate dehydrogenase gene (*Glyma.01g197700*) (Supplementary Figure S1). Further targeted analysis will be necessary to determine which gene is influencing each trait, as a single enzyme is unlikely responsible for multiple steps in the metabolic pathway. We found significant SNPs in high LD ( $r^2 > 0.5$ ) with the detected leading SNP at the promoter of *Glyma.01g197700*, but not in the coding region of the gene (Supplementary Figure S2).

A region including the I locus on chromosome 8 (Clough et al. 2004) was associated with seed mottling, as well as oil, cysteine, isoleucine, leucine, linoleic acid, lysine, methionine, palmitic acid, stearic acid, threonine, and valine levels in the seed (Figure 3). The most likely candidate gene for the observed differences in amino acids levels, AK-HDSH (aspartokinase homoserine dehydrogenase, *Glyma.08g107800*) is a bifunctional enzyme catalyzing the key steps of asparagine phosphatization and the aspartate-semialdehyde to homoserine conversion by which aspartate family amino acids (lysine, threonine, methionine, and isoleucine) are synthesized (Zhu-Shimoni and Galili 1998). However, amino acid data were generated using Near Infrared Reflectance,



**Figure 3** Significant SNPs from GWAS from individual studies and meta-GWAS for seed-related traits. Symbol position along the x-axis shows the position (in Mb) along the chromosome, while y-axis symbol position shows the LOD score of the lead SNP for each QTL. The x-axis labels indicate position (in Mb) of tertile points, while y-axis labels show minimum, maximum, and middle of LOD score range for the given trait class. Shape and color correspond to unique traits.

which may have low precision in estimating amino acid composition when there is variability in seed coat color (Baianu et al. 2011). Therefore, further validation is needed to establish the association between the *AK-HDSH* or *I* loci and the amino acid profile.

*SACPD-C* (*Glyma.14g121400*) was the primary candidate to explain differences in stearic acid content within seed oil and has been previously functionally validated (Gillman et al. 2014). Using the Wm82.a2 reference genome build, this appeared as three separate peaks; however, a single peak was observed when using the Wm82.a1 version. We postulate a possible assembly error in the region surrounding the *SACPD-C* locus in the soybean reference genome Wm82.a2, due to conflicting results (Supplementary Table S4). We attempted to identify false peaks generated due to genome mis-assembly by fitting the lead SNP as a covariate in the GWAS model, and then observed lower *P*-values for the remaining SNPs and detected a weaker signal from surrounding SNPs indicative of a single gene. Presence of stronger signals in surrounding SNPs would have indicated that two separate genes

are in play. In addition, the  $r^2$  between SNPs in all three regions was greater than 0.7, suggesting physical linkage. The Wm82.a1 results (SNP effects, physical location, LD) provide the most plausible explanation for the presence of a single gene in this genomic region and suggest that Wm82.a2 has unresolved errors in scaffold positioning.

A peak on chromosome 5 associated with palmitic acid content was detected in 3 different studies. Using data from the “2mn81” study, the locus mapped to a region of over 600 kb. However, two other studies (2ky81 and ms2000.02) mapped this locus within a smaller region of 130 kb (ss715592495-ss715592503) and 182 kb (ss715592491-ss715592500), respectively, with an overlap of about 88 kb (ss715592495-ss715592500) (Supplementary Figure S2). The candidate gene *FATB1a* (*Glyma.05g012300*) (Wilson et al. 2001) was identified in the overlap. However, no SNP in LD ( $r^2 \geq 0.5$ ) with the leading SNP of the locus was identified at the coding region or promoter of *FATB1a* based on analysis of resequencing data (Zhou et al. 2015) except the synonymous SNP\_5\_7995427 (Supplementary Figure S1). Causal variants have

**Table 2** List of candidate genes identified for seed composition traits using GWAS from individual studies and meta-GWAS

| Chromosome | Likely gene                           | Meta-GWAS | Individual studies GWAS | Trait(s)   | Studies source  |
|------------|---------------------------------------|-----------|-------------------------|--|---|
| 1          | BCAT/MDH                              | *         | *                       | Isoleucine, Leucine, Lysine, Methionine, Palmitic acid, Threonine, Tryptophan  | aa op sugar fa 2009, il0102, meta, ms967  |
| 3          | <i>Glyma.03g173400</i>                |           | *                       | Methionine   | aa op sugar fa 2009   |
| 5          | <i>fap3</i>                           | *         | *                       | Iodine number, Palmitic acid, Stearic acid   | aa op sugar fa 2009, 1il64, 2il81.1, 2il81.2, 2ky81, 2mn81, 3il83.1, 3il84, 3il87, il0102, meta, ms1999.01, ms2000.02, ms923, ms967 |
|            | MTFL                                  |           | *                       | Linoleic acid, Seed oil, Oleic acid, Tryptophan  | aa op sugar fa 2009, 2il81.1, il0102, ms1999.01, ms967  |
| 6          | <i>Glyma.06G214800</i>                | *         | *                       | Stearic acid   | meta, ms1999.01, ms2000.02  |
|            | <i>Glyma.06g275100</i><br>T           |           | *                       | Cysteine   | aa op sugar fa 2009   |
|            |                                       |           | *                       | Arginine, Cysteine, Isoleucine, Leucine  | aa op sugar fa 2009   |
| 8          | I/AK-HDSH                             | *         | *                       | Cysteine, Isoleucine, Leucine, Linoleic acid, Lysine, Methionine, Seed oil, Palmitic acid, Stearic acid, Threonine, Valine | aa op sugar fa 2009, meta, ms967  |
| 9          | <i>Glyma.09g090600</i><br>R           | *         | *                       | Palmitic acid  | il0102, meta  |
| 13         | <i>Glyma.13g149700</i>                | *         | *                       | Tryptophan   | aa op sugar fa 2009   |
|            |                                       |           | *                       | Oleic acid, Palmitic acid, Seed protein  | meta, ms2000.02   |
| 14         | <i>fan1</i>                           | *         | *                       | Linolenic acid   | 2mn81, 3il83.1, il0102, meta, mn945, ms967  |
| 15         | <i>Glyma.15g049200</i><br>"GmSWEET15" | *         | *                       | Linolenic acid, Seed oil, Seed protein, Threonine  | aa op sugar fa 2009, 2ky81, 3il83.1, 3il84, il989, meta, ms1999.01, ms923   |
| 19         | Dt1, <i>Glyma.19g194300</i>           |           | *                       | Linoleic acid, Oleic acid, Valine  | aa op sugar fa 2009, ms1999.01, ms2000.02   |
| 20         | CHR20OP                               | *         | *                       | Seed oil, Seed protein   | aa op sugar fa 2009, 2il81.1, meta, ms1999.01, ms967  |
| 14 (3)     | SACPD-C                               | *         | *                       | Stearic acid   | 1il66, 2il81.1, 2mn81, 3il83.1, 4il87, 5il90, il0102, meta, mn945, ms923  |

\* Bonferroni corrected P-value threshold [p-value 0.05 / number of markers].

been identified in mutagenized breeding material (Bachleda et al. 2016; Goettel et al. 2016; Thapa et al. 2016), but naturally occurring variations are not well characterized.

### Disease resistance traits

Among disease traits, we identified 1346 marker-trait associations with traditional GWAS studies, as well as 571 SNPs associated with disease traits when analyzed across studies by meta-GWAS. 213 peaks mapped to all 20 chromosomes, with 33 candidate genes or QTL identified (Figure 4, Table 3, Supplementary Tables S1 and S3). Meta-analysis in several instances narrowed the genomic region for peaks. For example, the association between the *Rps3* region and resistance to race 1 of *Phytophthora* root rot was mapped to a 144 kb region in the meta-analysis, compared to a 1 Mb region in individual studies (Supplementary Table S1). This reduces the search space for causal genes and allows for greater accuracy when identifying candidate genes.

We found a peak that was associated with resistance to races 1, 2, 3, 4, 5, 7, 10, and 17 of *Phytophthora sojae* that mapped to the position of the *Rps1* locus (Gao and Bhattacharyya 2008). A previously unreported peak for soybean cyst nematode resistance identified on chromosome 11 was mapped to *Glyma.11g234500*, an alpha-soluble N-ethylmaleimide-sensitive

factor (NSF) attachment protein ( $\alpha$ -SNAP). Notably, the candidate genes *GmSNAP11* (*Glyma.11g234500*) and *GmSNAP14* (*Glyma.14g054900*) (Lakhssassi et al. 2017), identified at 7 and 84 kb apart from lead SNPs ss715610420 and ss715618859, respectively, are paralogs and encode a Soluble NSF Attachment Protein (SNAP). Another soybean SNAP gene on chromosome 18, *GmSNAP18*, has been reported to play a role in resistance to SCN (Cook et al. 2012). On chromosome 1, the locus for seed composition co-localized with a bacterial pustule resistance peak. This peak does not correspond to the previously identified *Rxp* locus, instead, a candidate gene *Glyma.01g197800* is identified as the potential underlying gene. A peak on chromosome 3 at 34.24–35.18 Mb was found to be significantly associated with iron deficiency chlorosis tolerance and *Pythium irregularare* resistance. This region has previously been investigated as the source of IDC tolerance in "Isoclark" (Stec et al. 2013; Assefa et al. 2020). The GWAS analysis identified previously unreported genomic regions that were associated with resistance to bean pod mottle virus, brown stem rot, frogeye leaf spot, *Phytophthora* root rot, and soybean cyst nematode (Figure 4). A full list of identified SNPs and candidate genes for these traits, as well as for all other traits examined in this study using both combined analyses and analysis of individual experiments are provided in Supplementary Table S1.

**Table 3** List of candidate genes identified for disease resistance/stress tolerance traits using GWAS from individual studies and meta-GWAS

| Chromosome | Likely gene                         | Meta-GWAS | Individual studies GWAS | Trait(s)                        | Studies source            |
|------------|-------------------------------------|-----------|-------------------------|---------------------------------|---------------------------|
| 1          | RLK3                                |           | *                       | Bacterial pustule               | bp488001                  |
| 3          | Glyma.03g127100                     |           | *                       | Pythium root rot                | PYU.11002                 |
|            | Glyma.03g130600                     | *         | *                       | Iron deficiency chlorosis       | lssleepyeye04, meta       |
|            | Glyma.03g262500                     | *         | *                       | SCN races: 14                   | Meta                      |
|            | Rps1                                | *         | *                       | Phytophthora root rot           | meta, PRR1, PRR1.10001,   |
|            |                                     |           |                         | races: 1, 2, 3, 4, 5, 7, 10, 17 | PRR1.10002, PRR1.10004,   |
|            |                                     |           |                         |                                 | PRR1.11002, PRR1.11003,   |
|            |                                     |           |                         |                                 | PRR1.461592,              |
|            |                                     |           |                         |                                 | PRR1.488001,              |
|            |                                     |           |                         |                                 | PRR1.492577,              |
|            |                                     |           |                         |                                 | PRR1.492990, PRR10,       |
|            |                                     |           |                         |                                 | PRR17, PRR17.491404,      |
|            |                                     |           |                         |                                 | PRR17.492448,             |
|            |                                     |           |                         |                                 | PRR17.492990, PRR2,       |
|            |                                     |           |                         |                                 | PRR3, PRR3.492577,        |
|            |                                     |           |                         |                                 | PRR3.492990, PRR4,        |
|            |                                     |           |                         |                                 | PRR4.492990, PRR5,        |
|            |                                     |           |                         |                                 | PRR5.492990, PRR7,        |
|            |                                     |           |                         |                                 | PRR7.491404,              |
|            |                                     |           |                         |                                 | PRR7.492448,              |
|            |                                     |           |                         |                                 | PRR7.492990, prrdl96_1,   |
|            |                                     |           |                         |                                 | prrdl96_3, prrfs04_17,    |
|            |                                     |           |                         |                                 | prrfs04_7, prrrs01_1      |
|            | Rps7                                | *         | *                       | Phytophthora root rot           | meta, PRR1, PRR1.10002,   |
|            |                                     |           |                         | races: 1, 2, 3, 4, 5, 7, 10, 17 | PRR1.10003, PRR1.10004,   |
|            |                                     |           |                         |                                 | PRR1.11003, PRR1.488001,  |
|            |                                     |           |                         |                                 | PRR1.492577,              |
|            |                                     |           |                         |                                 | PRR1.492990, PRR10,       |
|            |                                     |           |                         |                                 | PRR17, PRR17.491404,      |
|            |                                     |           |                         |                                 | PRR17.492448,             |
|            |                                     |           |                         |                                 | PRR17.492990, PRR2,       |
|            |                                     |           |                         |                                 | PRR3, PRR3.492990, PRR5,  |
|            |                                     |           |                         |                                 | PRR5.492990, PRR7,        |
|            |                                     |           |                         |                                 | PRR7.491404,              |
|            |                                     |           |                         |                                 | PRR7.492448,              |
|            |                                     |           |                         |                                 | PRR7.492990, prrfs04_17,  |
|            |                                     |           |                         |                                 | prrfs04_7                 |
| 4          | Glyma.04g190400                     | *         | *                       | SCN races: 3, 4, 14             | meta, SCN14, soysc-       |
|            |                                     |           |                         |                                 | nyoung94_3                |
|            | Glyma.04g227900                     |           | *                       | Brown stem rot                  | bsrcodeall                |
| 5          | Glyma.05g137500/<br>Glyma.05g137800 |           | *                       | Brown stem rot                  | bsr97, bsrcode492477      |
| 6          | Glyma.06g199600/<br>Glyma.06g197800 | *         | *                       | Frogeye leaf spot, race 2       | 2ky91, Fe2, meta          |
| 7          | Glyma.07g192200                     | *         | *                       | SCN races: 1, 3, 5, 14          | meta, SCN14,              |
|            |                                     |           |                         |                                 | SCN14.491576,             |
|            |                                     |           |                         |                                 | SCN14code.491576,         |
|            |                                     |           |                         |                                 | soyscnaand_3, soyscna-    |
|            |                                     |           |                         |                                 | nand_5, soysc-            |
|            |                                     |           |                         |                                 | nyoung94_3, soysc-        |
|            |                                     |           |                         |                                 | nyoung94_5                |
| 8          | Glyma.08g231100                     | *         | *                       | SCN races: 3, 5, 14             | meta, SCN14, soysc-       |
|            |                                     |           |                         |                                 | nyoung94_5, soysc-        |
|            |                                     |           |                         |                                 | nyoung94_14               |
|            | Rhg4                                | *         | *                       | SCN races: 1, 3, 5, 14          | meta, SCN1, SCN14, soysc- |
|            |                                     |           |                         |                                 | nyoung94_3                |
| 10         | Glyma.10g273300/<br>276600          | *         | *                       | SCN races: 14                   | meta, SCN14,              |
|            |                                     |           |                         |                                 | SCN14.491576,             |
|            |                                     |           |                         |                                 | SCN14code.491576,         |
|            |                                     |           |                         |                                 | soyscnyoung94_14          |
| 11         | Glyma.11g233500                     |           | *                       | Phytophthora root rot           | PRR17.492990              |
|            | Glyma.11g234500<br>(SNAP11)         | *         | *                       | SCN races 1, 3, 4, 14           | meta, SCN14, sojascnar-   |
|            |                                     |           |                         |                                 | elli00, soyscnaand_5,     |
|            |                                     |           |                         |                                 | soyscnyoung88_5, soysc-   |
|            |                                     |           |                         |                                 | nyoung94_5, soysc-        |
|            |                                     |           |                         |                                 | nyoung94_14               |
| 12         | Glyma12g22660                       |           | *                       | SCN races: 1                    | SCN1                      |
| 13         | Glyma.13g222300                     | *         | *                       | SCN races: 1, 3, 14             | meta, SCN14, sojascnar-   |
|            |                                     |           |                         |                                 | elli00, soyscnyoung94_14  |

(continued)



**Table 3.** (continued)

| Chromosome | Likely gene              | Meta-GWAS | Individual studies GWAS | Trait(s)   | Studies source  |
|------------|--------------------------|-----------|-------------------------|--|---|
|            | Rag2/Rag5<br>Rps3        | *         | *                       | Soybean aphid<br>Phytophthora root rot<br>races: 1, 4, 12, 20, 25  | aphidcm02<br>PRR1, PRR1.10004,<br>PRR1.11003, PRR1.492990,<br>PRR12, PRR20, PRR25,<br>PRR25.491404,<br>PRR25.492990, PRR4,<br>PRR4.492990, meta |
| 14         | Rsv1<br>Glyma.14g098900  |           | *                       | Peanut mottle virus<br>Brown stem rot                              | pmv<br>bsr97, bsrcode492477   |
| 15         | NSC14<br>Glyma.15g052000 |           | *                       | Northern stem canker<br>Phytophthora root rot<br>races: 2          | NSC, NSC.491493<br>PRR2   |
| 16         | Glyma.16g096900          |           | *                       | Phytophthora root rot<br>races: 2                                  | PRR2  |
|            | Rag3<br>Rbs1/ Rbs2/ Rbs3 |           | *                       | Soybean aphid<br>Brown stem rot                                    | aphidcm02<br>bsr97, bsr491584, bsrall,<br>bsrcodeall  |
|            | Rcs3<br>Rps2             | *         | *                       | Frogeye leaf spot, race 2<br>Phytophthora root rot<br>races: 2, 25 | 2il81.1, Fe2, meta<br>PRR2, meta  |
| 17         | Glyma.17g090200          |           | *                       | Bean pod mottle virus  | bpmvall   |
| 18         | Glyma.18g138700          |           | *                       | Phytophthora root rot<br>races: 5                                  | PRR5, PRR5.492990   |
|            | Rhg1                     | *         | *                       | SCN races: 3, 4, 5, 14   | meta, SCN14, soyscna-<br>nand_3, soysc-<br>nyoung88_5, soysc-<br>nyoung94_3, soysc-<br>nyoung94_14  |
|            | Rps4                     | *         | *                       | Phytophthora root rot<br>races: 1, 3, 4, 25                        | meta, PRR1, PRR1.10001,<br>PRR1.10002, PRR1.10004,<br>PRR1.488001, PRR25,<br>PRR25.491404, PRR4   |

\* Bonferroni corrected P-value threshold [p-value 0.05 / number of markers].

The majority of studies included in this study for disease resistance were germplasm screenings, where many entries were tested to find new sources of resistance. Such germplasm screening studies were not originally intended for GWAS; for example, multiple rating systems, ordinal rating scales, and noninteger ratings used in the studies complicates result comparisons and are not easily amenable to linear statistical models. Standardization of screening protocols across research groups and inclusion of key data for comparison of studies such as those suggested by the MIAPPE checklist (Cwiek-Kupczyńska et al. 2016) will be key for future research into plant disease resistance. In addition, an increased utilization of image-based phenotyping will play a key role, allowing for digital disease severity ratings on a continuous scale (Naik et al. 2017; Zhang et al. 2017), minimal inter- and intra-rater variability in measurements through hyperspectral camera and ML-based analysis (Nagasubramanian et al. 2018, 2019). It will also enable the comparison of results across studies by facilitating reanalysis of previous experiments with new rating systems or approaches, as long as needed input variables are available.

### Implications of pleiotropy vs linked genes

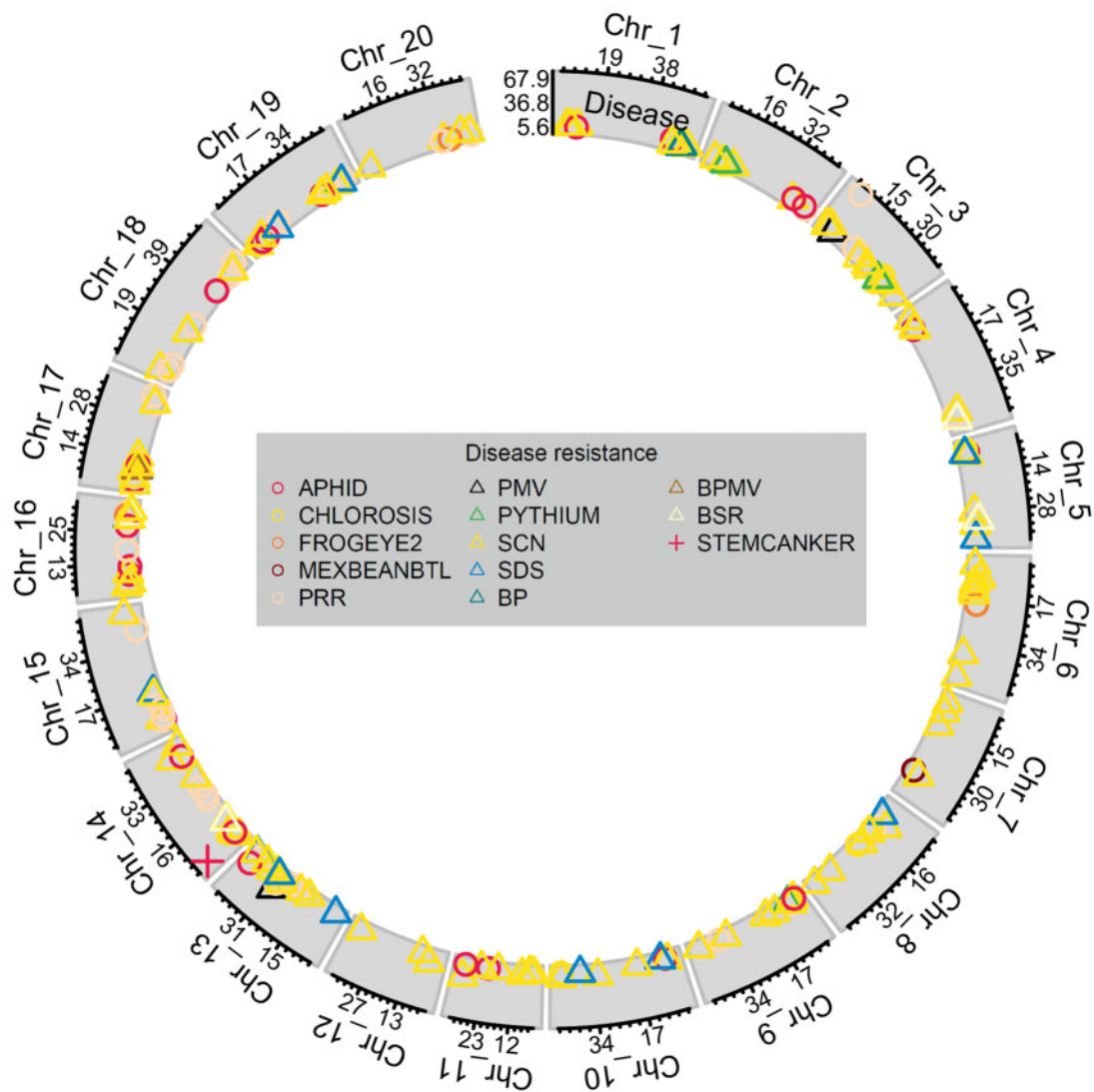
While repeated crossing or careful selection of the donor parent can break linkage drag, negative pleiotropic effects associated with a gene of interest are more problematic. Candidate gene analysis was aided by tissue-specific gene expression data available at SoyBase. The use of a blend of individual and meta-analyses provided improved resolution through examining overlapping peaks and utilizing the increased power in larger panels in the

meta-analysis. However, when investigating the peak on chromosome 1 for fatty acid and amino acid composition, a convincing distinction between pleiotropy and linkage could not be made. This was due to the presence of multiple strong candidate genes. While meta-GWAS approaches are very beneficial for improving map resolution, they are still limited in their inference in regions with strong LD. Meta-GWAS results outputs still require follow-up molecular and functional validation to confirm the candidate genes as well as to confirm pleiotropy vs linkage.

Pleiotropic effects of major genes significantly alter multiple traits simultaneously, creating a situation of either rapid improvement across traits, or of tradeoffs, such as is found in most soybean protein/oil content QTL. Genetic improvement utilizing pleiotropic effects may be limited in applicability to specific geographic regions if they affect key adaptation genes such as the maturity loci or stem termination. Therefore, it will be necessary for breeders to independently determine whether a gene with pleiotropic effects is a good fit for their variety development goals. In cases where pleiotropy is associated with a tradeoff between multiple traits, such as between seed protein and oil content, breeders will need to weigh the importance of each trait or identify combinations of genes affecting the trait that can provide an adequate phenotype for each trait considered.

### Motivation for the use of meta-analysis

For many important row crop species, such as soybean, corn, wheat, and sorghum, it is impractical or impossible to evaluate the full breadth of the available germplasm at a single location. This is due to space limitations, availability of labor or funding



**Figure 4** Significant SNPs from GWAS from individual studies and meta-GWAS for disease tolerance/resistance traits. Symbol position along the x-axis shows the position (in Mb) along the chromosome, while y-axis symbol position shows the LOD score of the lead SNP for each QTL. The x-axis labels indicate position (in Mb) of tertile points, while y-axis labels show minimum, maximum, and middle of LOD score range for the given trait class. Shape and color correspond to unique traits.

for phenotyping, or irreconcilable differences between genotypes preventing them from growing in the same place, such as differences in photoperiod sensitivity or vernalization requirements. To capture the breadth of the genetic and phenotypic diversity, it is necessary to test each variety with a similarly adapted cohort. The separate analysis of each environment can increase the odds of detecting alleles which are near fixation in the population or are environmentally dependent (Singh et al. 2014; Sherman et al. 2019).

For simple, qualitative traits such as pubescence color in soybean, there is often little benefit in meta-GWAS due to the consistency with which the gene can be mapped and the lack of environmental dependence on trait expression. When studying environmentally dependent traits, such as agronomic, disease resistance, and seed composition traits including seed oil or protein content, meta-GWAS provides advantages particularly in increasing the likelihood of finding small-effect genes. Studies sharing an environmental factor, such as high nighttime temperature, can be preferentially grouped for meta-analysis to identify these small effect genes within the context of the shared

environmental cues. When comparing individual experiments results (Supplementary Figure S3A) with the combined meta-analysis (Supplementary Figure S3B), additional significant peaks were observed in meta-analysis. For example, the SNP marker *ss715614263* was previously associated with seed protein using mega-analysis (Bandillo et al. 2015). The same locus was found to be associated with protein, palmitic, and oleic acid content in an individual panel in the current study (*ms2000.02*), but was associated with protein and linoleic acid content in the meta-analysis (Supplementary Table S1). Identification of an association with multiple related traits is a strong signal that the association may merit additional study to identify a strong candidate gene and further explore the possible pleiotropic effects this locus is exhibiting, especially when stringent cut-offs are used to declare significance.

While meta-analysis identified fewer traits in the specific instance of *ss715614263*, the association with an additional trait (compared to individual analysis) still encourages its use, as each newly associated trait may provide guidance in identifying putative causal genes. A full listing of candidate genes detected in

each study is provided as Supplementary Table S5, which also provides a reference to candidate genes detected either only in individual studies or only via meta-analysis. To maximize the effectiveness of soybean breeding programs, we sought to identify as many genes as possible for numerous traits, ensuring that multiple paths are available for further cultivar improvement. To this extent, we perform the first published GWAS analysis for many, but not all, of these studies. By maximizing the identified links between markers and phenotypes of interest, meta-GWAS aids efforts to bridge the gap between genotype and phenotype, allowing for improvements not only in trait prediction and selections, but also in modeling the interactions between multiple genes in overall trait performance.

### Future mapping, validation, and integration with phenomics studies

Traditional fine mapping through creating lines sharing homogeneous genetic background, such as near isogenic lines, is a powerful tool to uncover the causal genetic variants. However, it is time consuming to develop new near-isogenic lines in multiple backgrounds to reduce the potential influence of background-specific effects. In this study, large variation of LD architecture was observed across populations. This enables substantially shortening of the candidate chromosomal regions of specific peaks by comparing mapping results from separate studies using different populations. Considering almost all accessions in the USDA Soybean Germplasm Collection were genotyped by SoySNP50K BeadChip and are publicly accessible, parents with divergent haplotypes at specific genomic regions of interest can be selected for fine mapping. The consistent identification of major genes, including those affecting multiple traits of interest, suggests that further improvements in mapping ability would likely require a model with the major genes treated as covariates. While it is currently possible to account for the effects of major genes by using SNPs linked to the gene of interest as covariates, this approach is only an approximation due to incomplete linkage between common SNPs and the underlying gene. Instead, allele-specific markers should be developed and deployed across both wild-type germplasm and breeding material.

In the future, similar studies will benefit by incorporating weather, soil, or management parameters to explain differences in marker effects between individual studies and in Meta-GWAS (Cook et al. 2017). In this scenario, access to standardized, quality-controlled records will be needed to tease apart the GxE component and identify the architecture of environmentally mediated expression and decipher associations between genetics and environmental signals for the traits of interest. The establishment of standardized tests enabled with advanced sensors and high-throughput phenotyping should improve the opportunity to identify additional genes influencing traits of interest through the analysis of previously ignored component traits, such as leaf expansion rate or chlorophyll density in the case of yield, (Dhondt et al. 2013) which may lead to an increased understanding of the genetic architecture of these traits and responses to environmental and management conditions (Parmley et al. 2019).

### Conclusions

Combined analysis of all investigated traits found 65 loci that corresponded to previously reported QTL, characterized genes, and new reported loci backed up with strong candidate genes conditioning the observed phenotypes. Several of the previously

identified loci (for example, *Dt1*, *E2*) were associated with multiple traits, identifying putative pleiotropic effects of the underlying genes. Differences between results in individual trials and the combined analyses confirm the importance of multi-environment testing for the identification of key traits, but also provide a strong motivation to create a community database that can be queried for scientific advancement. Continued publication of raw phenotypic values from screenings will increase the power for identification of important genes for both mean and plastic responses to reduce the financial and time burden on any individual program while benefiting future breeders and researchers. For example, the sharing of phenotypic information across research programs both nationally and globally, as currently ongoing with multi-states and -institutions uniform soybean tests and other cooperatively run tests in other crops.

### Acknowledgments

Authors sincerely thank all researchers past and present who generated data for individual studies and set up a community resource for advancing soybean research and development. They thank David Blystone and Dr. David Grant (USDA-ARS, retired), which greatly helped the manuscript.

A.S. and A.K.S. conceptualized the study; J.S., A.S., and A.K.S. designed the study; J.S. conducted statistical analysis with contributions from A.K.S. and J.Z.; Figures were prepared by J.Z. with inputs from J.S.; J.S. interpreted the results with contributions from J.Z., S.J., A.S., B.D., and A.K.S.; J.S. wrote the first draft with A.K.S.; all authors contributed in writing, reviewing, and approving the manuscript.

### Funding

The authors thank the Iowa Soybean Association (to A.K.S.), R. F. Baker Center for Plant Breeding (to A.K.S.), Bayer Chair in Soybean Breeding (to A.K.S.), USDA IOW04714, and National Science Foundation NRT-DESE: P3 (to J.M.S.) for the financial support.

### Conflicts of interest

Authors declare no conflict of interest.

### Literature cited

- Assefa T, Zhang J, Chowda-Reddy RV, Moran Lauter AN, Singh A, et al. 2020. Deconstructing the genetic architecture of iron deficiency chlorosis in soybean using genome-wide approaches. *BMC Plant Biol.* 20:42.
- Bachleda N, Pham A, Li Z. 2016. Identifying *FATB1a* deletion that causes reduced palmitic acid content in soybean N87-2122-4 to develop a functional marker for marker-assisted selection. *Mol Breed.* 36:45.
- Baianu I, Guo J, Nelson R, You T, Costescu D. 2011. NIR calibrations for soybean seeds and soy food composition analysis: total carbohydrates, oil, proteins and water contents. *Nat Preced.* v.2. <https://doi.org/10.1038/npre.2011.6611.2>
- Bandillo N, Jarquin D, Song Q, Nelson R, Cregan P, et al. 2015. A population structure and genome-wide association analysis on the USDA soybean germplasm collection. *Plant Genome.* 8: 1–13.

- Bandillo NB, Lorenz AJ, Graef GL, Jarquin D, Hyten DL, et al. 2017. Genome-wide association mapping of qualitatively inherited traits in a germplasm collection. *Plant Genome*. 10: 1–18.
- Bolormaa S, Pryce JE, Reverter A, Zhang Y, Barendse W, et al. 2014. A multi-trait, meta-analysis for detecting pleiotropic polymorphisms for stature, fatness and reproduction in beef cattle. *PLoS Genet*. 10:e1004198.
- Browning BL, Browning SR. 2016. Genotype imputation with millions of reference samples. *Am J Hum Genet*. 98:116–126.
- Bubeck DM, Goodman MM, Beavis WD, Grant D. 1993. Quantitative trait loci controlling resistance to gray leaf spot in maize. *Crop Sci*. 33:838–847.
- Cameron JN, Han Y, Wang L, Beavis WD. 2017. Systematic design for trait introgression projects. *Theor Appl Genet*. 130:1993–2004.
- Chang D, Nalls MA, Hallgrímsson IB, Hunkapiller J, van der Brug M, et al. 2017. A meta-analysis of genome-wide association studies identifies 17 new Parkinson's disease risk loci. *Nat Genet*. 49: 1511–1516.
- Chang H-X, Hartman GL. 2017. Characterization of insect resistance loci in the USDA soybean germplasm collection using genome-wide association studies. *Front Plant Sci*. 8:670.
- Chang H-X, Lipka AE, Domier LL, Hartman GL. 2016. Characterization of disease resistance loci in the USDA soybean germplasm collection using genome-wide association studies. *Phytopathology*. 106:1139–1151.
- Chen X, Zhao F, Xu S. 2010. Mapping environment-specific quantitative trait loci. *Genetics*. 186:1053–1066.
- Clough SJ, Tuteja JH, Li M, Marek LF, Shoemaker RC, et al. 2004. Features of a 103-kb gene-rich region in soybean include an inverted perfect repeat cluster of CHS genes comprising the I locus. *Genome*. 47:819–831.
- Cook DE, Lee TG, Guo X, Melito S, Wang K, et al. 2012. Copy number variation of multiple genes at Rhg1 mediates nematode resistance in soybean. *Science*. 338:1206–1209.
- Cook J, Mahajan A, Morris A. 2017. Guidance for the utility of linear models in meta-analysis of genetic association studies of binary phenotypes. *Eur J Hum Genet*. 25:240–245.
- Coser SM, Chowda Reddy RV, Zhang J, Mueller DS, Mengistu A, et al. 2017. Genetic architecture of charcoal rot (*Macrophomina phaseolina*) resistance in soybean revealed using a diverse panel. *Front Plant Sci*. 8:1626.
- Cwiek-Kupczyńska H, Altmann T, Arend D, Arnaud E, Chen D, et al. 2016. Measures for interoperability of phenotypic data: minimum information requirements and formatting. *Plant Methods*. 12:44.
- de Azevedo Peixoto L, Moellers TC, Zhang J, Lorenz AJ, Bhering LL, et al. 2017. Leveraging genomic prediction to scan germplasm collection for crop improvement. *PLoS One*. 12:e0179191.
- Descriptors for Soybean. 2019. U.S. National Plant Germplasm System. <https://npgsweb.ars-grin.gov/gringlobal/cropdetail?type=descriptor&id=51>; last accessed April 13, 2021.
- Dhondt S, Wuyts N, Inzé D. 2013. Cell to whole-plant phenotyping: the best is yet to come. *Trends Plant Sci*. 18:428–439.
- Diers BW, Specht J, Rainey KM, Cregan P, Song Q, et al. 2018. Genetic architecture of soybean yield and agronomic traits. *G3 (Bethesda)*. 8:3367–3375.
- Fang C, Ma Y, Wu S, Liu Z, Wang Z, et al. 2017. Genome-wide association studies dissect the genetic networks underlying agronomical traits in soybean. *Genome Biol*. 18:161.
- Flint-Garcia SA, Jampatong C, Darrach LL, McMullen MD. 2003. Quantitative trait locus analysis of stalk strength in four maize populations. *Crop Sci*. 43:13–22.
- Gao H, Bhattacharyya MK. 2008. The soybean-*Phytophthora* resistance locus Rps1-k encompasses coiled coil-nucleotide binding-leucine rich repeat-like genes and repetitive sequences. *BMC Plant Biol*. 8:29.
- Gillman JD, Stacey MG, Cui Y, Berg HR, Stacey G. 2014. Deletions of the SACPD-C locus elevate seed stearic acid levels but also result in fatty acid and morphological alterations in nitrogen fixing nodules. *BMC Plant Biol*. 14:143.
- Goettel W, Ramirez M, Upchurch RG, An YQ. 2016. Identification and characterization of large DNA deletions affecting oil quality traits in soybean seeds through transcriptome sequencing analysis. *Theor Appl Genet*. 129:1577–1593.
- Gu Z, Gu L, Eils R, Schlesner M, Brors B. 2014. Circlize implements and enhances circular visualization in R. *Bioinformatics*. 30: 2811–2812.
- Hulting AG, Wax LM, Nelson RL, Simmons FW. 2001. Soybean (*Glycine max* (L.) Merr.) cultivar tolerance to sulfentrazone. *Crop Protection*. 20:679–683.
- Lakhssassi N, Liu S, Bekal S, Zhou Z, Colantonio V, et al. 2017. Characterization of the soluble NSF Attachment Protein gene family identifies two members involved in additive resistance to a plant pathogen. *Sci Rep*. 7:45226.
- Lipka AE, Tian F, Wang Q, Peiffer J, Li M, et al. 2012. GAPIT: genome association and prediction integrated tool. *Bioinformatics*. 28: 2397–2399.
- Liu B, Watanabe S, Uchiyama T, Kong F, Kanazawa A, et al. 2010. The soybean stem growth habit gene Dt1 is an ortholog of Arabidopsis TERMINAL FLOWER1. *Plant Physiol*. 153:198–210.
- Miller EK. 2003. Index to USDA Technical Bulletins, edited by USDA/ARS. National Agricultural Library. <https://pubs.nal.usda.gov/sites/pubs.nal.usda.gov/files/tb.htm>; last accessed April 13, 2021.
- Moellers TC, Singh A, Zhang J, Brungardt J, Kabbage M, et al. 2017. Main and epistatic loci studies in soybean for *Sclerotinia sclerotiorum* resistance reveal multiple modes of resistance in multi-environments. *Sci Rep*. 7:3554.
- Nagasubramanian K, Jones S, Sarkar S, Singh AK, Singh A, et al. 2018. Hyperspectral band selection using genetic algorithm and support vector machines for early identification of charcoal rot disease in soybean stems. *Plant Methods*. 14:86.
- Nagasubramanian K, Jones S, Singh AK, Sarkar S, Singh A, et al. 2019. Plant disease identification using explainable 3D deep learning on hyperspectral images. *Plant Methods*. 15:98.
- Naik HS, Zhang J, Lofquist A, Assefa T, Sarkar S, et al. 2017. A real-time phenotyping framework using machine learning for plant stress severity rating in soybean. *Plant Methods*. 13:23.
- Neyman J, Pearson ES. 1928. On the use and interpretation of certain test criteria for purposes of statistical inference. Part I. *Biometrika*. 20A:175–240.
- Parmley K, Nagasubramanian K, Sarkar S, Ganapathysubramanian B, Singh AK. 2019. Development of optimized phenomic predictors for efficient plant breeding decisions using phenomic-assisted selection in soybean. *Plant Phenomics*. 2019:5809404.
- Sherman RM, Forman J, Antonescu V, Puiu D, Daya M, et al. 2019. Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat Genet*. 51:30–35.
- Singh A, Knox RE, DePauw RM, Singh AK, Cuthbert RD, et al. 2014. Stripe rust and leaf rust resistance QTL mapping, epistatic interactions, and co-localization with stem rust resistance loci in spring wheat evaluated over three continents. *Theor Appl Genet*. 127:2465–2477.
- Singh D, Wang X, Kumar U, Gao L, Noor M, et al. 2019. High-throughput phenotyping enabled genetic dissection of crop lodging in wheat. *Front Plant Sci*. 10:394.

- Song Q, Hyten DL, Jia G, Quigley CV, Fickus EW, et al. 2013. Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. *PLoS One*. 8:e54985.
- Song Q, Hyten DL, Jia G, Quigley CV, Fickus EW, et al. 2015. Fingerprinting soybean germplasm and its utility in genomic research. *G3 (Bethesda)*. 5:1999–2006.
- Stec AO, Bhaskar PB, Bolon Y-T, Nolan R, Shoemaker RC, et al. 2013. Genomic heterogeneity and structural variation in soybean near isogenic lines. *Front Plant Sci*. 4:104–104.
- Takahashi R, Asanuma S. 1996. Association of T gene with chilling tolerance in soybean. *Crop Sci*. 36:559–562.
- Thapa R, Carrero-Colón M, Hudson KA. 2016. New Alleles of *FATB1A* to reduce palmitic acid levels in Soybean. *Crop Sci*. 56:1076–1080.
- The 100,000 Genomes Project. 2019. *GenomicsEngland*. <https://www.genomicsengland.co.uk/about-genomics-england/the-100000-genomes-project/>; last accessed April 13, 2021.
- Toda K, Yang D, Yamanaka N, Watanabe S, Harada K, et al. 2002. A single-base deletion in soybean flavonoid 3'-hydroxylase gene is associated with gray pubescence color. *Plant Mol Biol*. 50:187–196.
- Trotta L, Hautala T, Hämäläinen S, Syrjänen J, Viskari H, et al. 2016. Enrichment of rare variants in population isolates: single AICDA mutation responsible for hyper-IgM syndrome type 2 in Finland. *Eur J Hum Genet*. 24:1473–1478.
- Watanabe S, Xia Z, Hideshima R, Tsubokura Y, Sato S, 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.
- Willer CJ, Li Y, Abecasis GR. 2010. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*. 26:2190–2191.
- Wilson R F, Marquardt T C, Novitzky W P, Burton J W, Wilcox J R, et al. 2001. Metabolic mechanisms associated with alleles governing the 16:0 concentration of soybean oil. *J Amer Oil Chem Soc*. 78:335–340. 10.1007/s11746-001-0265-4
- Zeggini E, Ioannidis JP. 2009. Meta-analysis in genome-wide association studies. *Pharmacogenomics*. 10:191–201.
- Zeng A, Chen P, Korth K, Hancock F, Pereira A, et al. 2017. Genome-wide association study (GWAS) of salt tolerance in worldwide soybean germplasm lines. *Mol Breed*. 37:30.
- Zhang J, Naik HS, Assefa T, Sarkar S, Reddy RVC, et al. 2017. Computer vision and machine learning for robust phenotyping in genome-wide studies. *Sci Rep*. 7:44048.
- Zhang J, Singh A, Mueller DS, Singh AK. 2015. Genome-wide association and epistasis studies unravel the genetic architecture of sudden death syndrome resistance in soybean. *Plant J*. 84:1124–1136.
- Zhao J, Sauvage C, Zhao J, Bitton F, Bauchet G, et al. 2019. Meta-analysis of genome-wide association studies provides insights into genetic control of tomato flavor. *Nat Comm*. 10:1534.
- Zhou Z, Jiang Y, Wang Z, Gou Z, Lyu J, et al. 2015. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat Biotechnol*. 33:408–414.
- Zhu-Shimoni JX, Galili G. 1998. Expression of an Arabidopsis aspartate Kinase/Homoserine Dehydrogenase gene is metabolically regulated by Photosynthesis-related signals but not by Nitrogenous compounds. *Plant Physiol*. 116:1023–1028.

Communicating editor: P. Brown