2018

Genetic Architecture of Soybean Yield and Agronomic Traits

Brian W. Diers
*University of Illinois*, bdiers@illinois.edu

Jim Specht
*University of Nebraska-Lincoln*

Katy Martin Rainey
*Purdue University*

Perry Cregan
*USDA, Agricultural Research Service*

Qijian Song
*USDA, Agricultural Research Service*

See next page for additional authors
Follow this and additional works at: https://digitalcommons.unl.edu/agronomyfacpub

Part of the Agricultural Science Commons, Agriculture Commons, Agronomy and Crop Sciences
Commons, Botany Commons, Horticulture Commons, Other Plant Sciences Commons, and the Plant
Biology Commons

Diers, Brian W.; Specht, Jim; Rainey, Katy Martin; Cregan, Perry; Song, Qijian; Ramasubramanian, Vishnu;
Graef, George; Nelson, Randall L.; Schapaugh, William; Wang, Dechun; Shannon, Grover; McHale, Leah;
Kantartzi, Stella K.; Xavier, Alencar; Mian, Rouf; Stupar, Robert M.; Michno, Jean-Michel; Charles An, Yong-
Qiang; Goettel, Wolfgang; Ward, Russell; Fox, Carolyn; Lipka, Alexander E.; Hyten, David; Cary, Troy; and
Beavis, William D., "Genetic Architecture of Soybean Yield and Agronomic Traits" (2018). *Agronomy &
Horticulture -- Faculty Publications*. 1221.
https://digitalcommons.unl.edu/agronomyfacpub/1221

This Article is brought to you for free and open access by the Agronomy and Horticulture Department at
DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Agronomy & Horticulture --
Faculty Publications by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.
Authors
Brian W. Diers, Jim Specht, Katy Martin Rainey, Perry Cregan, Qijian Song, Vishnu Ramasubramanian, George Graef, Randall L. Nelson, William Schapaugh, Dechun Wang, Grover Shannon, Leah McHale, Stella K. Kantartzi, Alencar Xavier, Rouf Mian, Robert M. Stupar, Jean-Michel Michno, Yong-Qiang Charles An, Wolfgang Goettel, Russell Ward, Carolyn Fox, Alexander E. Lipka, David Hyten, Troy Cary, and William D. Beavis
ABSTRACT

Soybean is the world’s leading source of vegetable protein and demand for its seed continues to grow. Breeders have successfully increased soybean yield, but the genetic architecture of yield and key agronomic traits is poorly understood. We developed a 40-mating soybean nested association mapping (NAM) population of 5,600 inbred lines that were characterized by single nucleotide polymorphism (SNP) markers and six agronomic traits in field trials in 22 environments. Analysis of the yield, agronomic, and SNP data revealed 23 significant marker-trait associations for yield, 19 for maturity, 15 for plant height, 17 for plant lodging, and 29 for seed mass. A higher frequency of estimated positive yield alleles was evident from elite founder parents than from exotic founders, although unique desirable alleles from the exotic group were identified, demonstrating the value of expanding the genetic base of US soybean breeding.

KEYWORDS

soybean

Genetic Improvement

Yield

Genetic Mapping

Multiparental populations
key physiological mechanisms, whose phenotypic/genotypic variance underpins much of the variance in yield. Phenology is critical because greater yield can be positively correlated with later maturity in soybean. Of the mapped quantitative trait loci (QTL) listed in SoyBase (Grant et al. 2010), there are 79 yield and 82 maturity non-redundant QTL, but 40 of each are closely linked (<5 to 0 cM) or possibly pleiotropic. Though QTL mapping is relatively powerful for identifying associations between segregating alleles and phenotypes, it cannot provide high-resolution map positions when recombination events are rare.

Only a subset of the total number of QTL affecting a trait can be detected when using segregating lines derived from a single cross between two parents. Alternatively, greater map resolution of marker-trait associations (MTAs) can be achieved using genome-wide association studies (GWAS), which exploit historical recombination events. However, GWAS is not as powerful as QTL linkage mapping (Hirschhorn and Daly 2005; Kingsmore et al. 2008), nor can it detect rare alleles (even those of large effect), as documented in both theory (Korte and Farlow 2013) and practice (Bandillo et al. 2015). In contrast, the opportunity to detect rare alleles increases when QTL mapping is applied to scores of bi-parental matings, and when the lines used for initial crosses have been purposefully selected to be phenotypically diverse (Phansak et al. 2016).

Nested association mapping (NAM) takes advantage of both linkage and association mapping to increase map resolution and statistical power (Yu et al. 2008). In the NAM design, a common inbred parent is mated to a diverse group of homozygous founders to create thousands of recombinant inbred lines (RILs) in half-sib families. In the maize NAM project, the inbred B73 was crossed with 25 distinct lines adapted to a wide range of latitudes from the tropics to Canada (Buckler et al. 2009; McMullen et al. 2009). The maize NAM families have been used to map QTL for several traits including disease resistance (Kump et al. 2011; Poland et al. 2011; Olukolu et al. 2014) flowering time (Buckler et al. 2009), kernel composition (Cook et al. 2012) and plant type and architecture (Tian et al. 2011; Peiffer et al. 2013; Brown et al. 2011). The NAM population structure also was used to map QTL for stem rust in wheat (Bajgain et al. 2016), and flowering time and grain weight in barley (Maurer et al. 2015; Maurer et al. 2016). The objectives of our project were to map maker-trait associations for yield and other important agronomic traits in a soybean NAM population with a goal of improving our understanding of the genetic basis of these traits and to identify exotic sources of genes that can improve yield. To our knowledge, no prior NAM populations have been used to study the genetic architecture of yield, the most important trait for most major field crops.

MATERIALS AND METHODS

Germplasm development

The photoperiod sensitivity of soybean forces breeders to develop cultivars and lines adapted to specific latitudes. To minimize the confounding effect of non-adaptation in the soybean NAM population developed for this study, 40 soybean founder lines were selected for adaption to the major North American soybean production region occupying a latitudinal zone of ca. 37° to 43° N, and were mated to the high-yielding common parent IA3023, which is adapted to ca. 40° N. By limiting the latitude adaptation of the parents, we were able to develop recombinant inbred lines that could be evaluated for agronomic traits in common environments to reduce the confounding effect of photoperiod responses on these traits. Many potential founders were initially nominated by soybean breeders and a limited SNP-based diversity analysis of these nominees was used to identify a final set of 40 founders (Song et al. 2017). These consisted of 17 high-yielding elite cultivars and breeding lines (EL), 15 breeding lines selected for yield and exotic diversity (BX), and eight plant introductions (PI) that yielded well under severe drought in field trials (Song et al. 2017) (Table 1). From each of the 40 matings, 140 RILs were derived using single-seed descent from the F2 to F3 generation, resulting in 5,600 RILs.

Field evaluation

The common parent and founder lines, their RILs and check cultivars were grown in two-row field plots (0.76 m spacing; ca. 4 m long) replicated across 22 environments in eight Midwestern USA states from 2011 to 2013. RILs from each family were sub-divided into four sets of 35, with each set augmented with the family’s two parents and three check cultivars chosen for adaptation to the field environments. If sufficient seed was not available for a RIL entry, the plot was planted with a check variety. The entries within each set were randomized and the 40-entry sets were randomized within field sites. All 5,600 RILs (160 sets) were evaluated at eight sites, while at the remaining sites only 25 to 100 sets, with an average of 50 sets, were evaluated due to capacity limitations of the cooperators (Table 2).

The plot-combined seed weight and seed moisture content were measured electronically, and these data were used to calculate seed yield (kg ha⁻¹ on a 130 g kg⁻¹ moisture basis). Plant maturity was rated as days from planting to stage R8 (95% of pods fully mature), plant height as the cm distance from the soil surface to the top node on the main stem, plant lodging score rated visually on a scale of 1 = all plants erect to 5 = all plants prostrate and the yield component of seed mass as g (100 seed)⁻¹. Any plot data discarded by the cooperating scientist who judged that data to be of poor quality were treated as missing data. The number of plots evaluated for each trait ranged from 52,000 to over 66,000 (Table 2).

Marker evaluation

NAM RILs were genotyped with SNP markers (Song et al. 2016). Briefly, the common parent and 40 founder lines were sequenced with an Illumina HiSeq 2000 to identify SNP loci segregating in at least 28 (70%) of the 40 families based on marker allele differences between the common and founder lines. A total of 6,000 SNPs were selected from those that met this criterion and were submitted to Illumina, which identified 5,303 that were capable of associating with bead types on a BeadChip designed for the project (SoyNAM6K). The chip successfully detected 4,312 SNPs in the NAM parents and RILs. Markers that were non-polyorphic, or exhibited severe segregation distortion (i.e., a minor allele frequency of less than 10%), in any family were eliminated, resulting in 2,470 to 3,791 SNP loci segregating within individual families.

The NAM common parent and founder lines (but not RILs) were evaluated with the SoySNP50K BeadChip (Song et al. 2013) that detected the segregation of 42,509 SNP markers among these parents. The framework of the mapped SoyNAM6K markers was then used to project the segregating SoySNP50K markers onto the NAM RILs using the
An initial analysis of the SNP-genotyped RILs was conducted to identify RILs that deviated from the expected marker segregation (Song et al. 2017), and this led to 424 RILs being discarded because they had a SNP genotype identical with the female founder (i.e., were likely inadvertent female-parent self-pollinations), or they segregated for alleles that did not match the parent alleles. Most of the RILs from family N46 (PI307618B) fell into the latter category, indicating that a line other than the intended founder PI had been used in the mating with IA3023. Therefore, all lines from the N46 family were removed from the dataset.

### Data analyses

**Agromonic phenotypes:** Field plots were planted when conditions were suitable. Due to protracted and excessive rain some environments experienced late planting. Planting dates spanned four weeks in both 2012 and 2013. Therefore, Julian planting dates were used as a covariate in the analyses of agronomic trait values. In addition, environmental conditions due to soils, pests and diseases were inconsistent between years, locations and even within location-year combinations. Incomplete blocks within environments were augmented with IA3023 and at least three additional check varieties to provide estimates of block effects for purposes of adjusting genotypic values for agronomic traits. Two check varieties, IA3023 and U06-100052 were included in all blocks, two were included in 77.5% of the blocks and the remaining six checks were unevenly distributed among the blocks. Due to the imbalance of check varieties assigned to blocks as well as the variable number of blocks evaluated among environments, we used shrunken, i.e., best linear unbiased predicted values, rather than average values for field block effects that were based on the mixed linear model:

\[
Y = X\beta + Ck + B\pi + \epsilon,
\]

\[
\pi \sim N(0, I\sigma^2_{\pi}),
\]

\[
\epsilon \sim N(0, I\sigma^2_{\epsilon}),
\]

\[
\text{Cov}(\pi, \epsilon) = 0
\]

For yield, kg/ha, \( \gamma \) consisted of a vector of 7537 values for ten check varieties evaluated in 1726 blocks, \( X \) is a vector of planting dates for the checks in each block and is treated as a fixed effect, \( \beta \) is the slope and intercept for the covariate, planting date, \( C \) represents an incidence matrix for the check varieties, \( k \) is the vector of fixed effects represented by the check varieties, \( B \) is an incidence matrix indicating whether the \( \gamma \) value was obtained from a block, \( \pi \) is the vector of random effects represented by each block, \( I \) is the identity matrix, \( \sigma^2_{\beta} \) is the variance among blocks, \( \epsilon \) represents a residual effect, not accounted for in the model, and \( \sigma^2_{\epsilon} \) is the variance among residual values. Best Linear Unbiased Estimates (BLUE’s) of slope and intercept for planting date and the check variety values as well as BLUP values for block effects were obtained using the lmer package in R (Bates et al. 2015).

RIL agronomic phenotypes were evaluated in an unbalanced design across environments. Most RILs were evaluated for yield in seven environments, while subsets of RILs were evaluated for agronomic traits in the remaining environments (Table 2). Yield data were not collected from some of the 22 planted environments due to data quality issues caused by environmental problems at field locations. For example, yield data were not collected at one of the eight environments that all RILs were planted because of damage from a hail storm. Due to the imbalance of RIL’s evaluated among environments and variable numbers of blocks within environments a mixed linear model was used to analyze agronomic traits among the non-check entries:

| NAM Family | Parent | Origin | Group* |
|------------|--------|--------|--------|
| Hub        | IA3023 | Iowa State Univ. | Common parent |
| N02        | TN05-3027 | Univ. of Tenn. | EL |
| N03        | 4J105-3-4 | Purdue Univ. | EL |
| N04        | SM20-2-5-2 | Purdue Univ. | EL |
| N05        | CLO1055-4-6 | Purdue Univ. | EL |
| N06        | CLO1173-6-8 | Purdue Univ. | EL |
| N08        | HS6-3976 | Ohio State Univ. | EL |
| N9         | Prohio | USDA-ARS, Wooster, OH | EL |
| N10        | LD00-3309 | Univ. of Illinois | EL |
| N11        | LD01-5907 | Univ. of Illinois | EL |
| N12        | LD02-4485 | Univ. of Illinois | EL |
| N13        | LD02-9050 | Univ. of Illinois | EL |
| N14        | Magellan | Univ. of Missouri | EL |
| N15        | Maverick | Univ. of Missouri | EL |
| N17        | S06-13640 | Univ. of Missouri | EL |
| N18        | NE3001 | Univ. of Nebraska | EL |
| N22        | Skylla | Mich. State Univ. | EL |
| N23        | U03-100612 | USDA-ARS, Urbana, IL | EL |
| N24        | LG03-2979 | USDA-ARS, Urbana, IL | BX |
| N25        | LG02-3191 | USDA-ARS, Urbana, IL | BX |
| N26        | LG04-4717 | USDA-ARS, Urbana, IL | BX |
| N27        | LG05-4292 | USDA-ARS, Urbana, IL | BX |
| N28        | LG05-4317 | USDA-ARS, Urbana, IL | BX |
| N29        | LG05-4464 | USDA-ARS, Urbana, IL | BX |
| N30        | LG05-4832 | USDA-ARS, Urbana, IL | BX |
| N31        | LG90-2550 | USDA-ARS, Urbana, IL | BX |
| N32        | LG92-1255 | USDA-ARS, Urbana, IL | BX |
| N33        | LG94-1128 | USDA-ARS, Urbana, IL | BX |
| N34        | LG94-1906 | USDA-ARS, Urbana, IL | BX |
| N36        | LG97-7012 | USDA-ARS, Urbana, IL | BX |
| N37        | LG98-1605 | USDA-ARS, Urbana, IL | BX |
| N38        | LG00-3372 | USDA-ARS, Urbana, IL | BX |
| N39        | LG04-6000 | USDA-ARS, Urbana, IL | BX |
| N40        | PI 398.881 | South Korea | PI |
| N41        | PI 4427136 | South Korea | PI |
| N42        | PI 437169B | Russia | PI |
| N46        | PI S07681B | Uzbekistan | PI |
| N48        | PI 518751 | Serbia | PI |
| N50        | PI 561370 | China | PI |
| N54        | PI 404188A | China | PI |
| N64        | PI 574486 | China | PI |

*Founder group designations are EL = Elite, BX = breeding lines with exotic ancestry, and PI = plant introduction.

Wiltiams 82 reference genome (Wm82.a2_v1) bp positions for both the 6K and 50K chip SNP markers (Tian et al. 2011). The combined dataset of SoyNAM6K and SoySNP50K markers were then used to identify MTAs throughout the genome.

Linkage mapping was not conducted for the SoySNP50K SNP markers with the SoyNAM RILs. Instead, estimated cM linkage map positions of the SNPs were established using a linkage map derived from a Williams 82 x G. soja PI 479752 (WxP) population of 1083 RILs genotyped with 21,000 SoySNP50K SNPs (Song et al. 2016). The genetic location (cM) of any NAM SNP not present in the WxP map was inferred by linear interpolation between the NAM SNP physical position relative to the physical positions of the flanking WxP mapped SoySNP50K SNPs.
Table 2 The number of field plots, environments and blocks used to obtain BLUP values for the agronomic traits, estimates of broad-sense heritability (H) on an entry mean basis with 95% confidence intervals (CI), and the proportion of genotypic variance attributable to variance among families ($\sigma^2_{\text{res}}$ among families)

| Trait                              | Number of plots | Number of environments (and field Blocks) | H (95% CI)          | $\sigma^2_{\text{res}}$ among families |
|------------------------------------|-----------------|------------------------------------------|---------------------|----------------------------------------|
| Yield (kg ha$^{-1}$)               | 66,684          | 17 (1726)                                | 0.822 (0.815, 0.829)| 0.57                                   |
| Maturity (days)                    | 58,714          | 15 (1541)                                | 0.935 (0.932, 0.937)| 0.25                                   |
| Lodging (1-5 scale)               | 57,420          | 15 (1351)                                | 0.824 (0.817, 0.831)| 0.18                                   |
| Plant height (cm)                  | 57,822          | 15 (1490)                                | 0.938 (0.936, 0.941)| 0.28                                   |
| Seed mass [g (100 seed)$^{-1}$]    | 52,703          | 10 (1222)                                | 0.939 (0.936, 0.941)| 0.26                                   |

\[
y = \mathbf{X}\beta + \mathbf{Z}\mathbf{v} + \mathbf{e},
\]

\[
\beta = [\beta_{\text{plate}}, \beta_{\text{blk}}],
\]

\[
\mathbf{v} \sim N(0, \mathbf{I}\sigma^2_{\text{res}}),
\]

\[
\mathbf{e} \sim N(0, \mathbf{I}\sigma^2_{\text{res}}),
\]

\[
\text{Cov}(\mathbf{v}, \mathbf{e}) = 0
\]

(2)

where $y$ is a vector of measured phenotypic trait values for entries consisting of RILs and their founder parents. The length of $y$ as well as the dimensions of the matrices, $\mathbf{X}$ and $\mathbf{Z}$, depend on the agronomic traits (Table 2). For yield, $y$ consists of a vector of 66,684 values for RILs and their founder parents evaluated at 17 environments in 1726 blocks. $\mathbf{X}$ consists of two vectors representing covariates for planting date and the estimated shrunken block values obtained from (1). Both covariates are treated as fixed effects in (2) and $\beta$ are the slopes and intercepts for planting date and block effects. $\mathbf{Z}$ is an incidence matrix for entries (RILs and founder lines) indicating whether the measured trait value, $y$, for the entry was evaluated in a block, and $\mathbf{v}$ is the vector of random effects for entries, $I$ is the identity matrix, $\sigma^2_{\text{res}}$ is the genotypic variance among entries and $\mathbf{e}$ represents the residual value, not accounted for in the model and $\sigma^2_{\text{res}}$ is the variance among residual values.

Estimates of variance components for model (2) were obtained using the lmer package (Bates et al. 2015). Because the variance component estimates were from unbalanced numbers of environments, the estimated phenotypic variance was used to calculate broad sense heritability on both a family mean and an RIL entry mean basis. In these calculations, estimates of variance components were divided by the harmonic means for the number of environments in which the entries were evaluated (Holland et al. 2003). Harmonic means were likewise used to approximate confidence intervals for the estimated heritabilities (Knapp et al., 1985).

GWAS analyses of agronomic phenotypes: We used the BLUP genotypic values from model (2) for the analyses of genome wide associations in which we identified marker trait associations (MTAs) as random effects dependent on family background:

\[
\mathbf{v} = \mathbf{\mu} + \mathbf{W} \alpha + \phi + \mathbf{e},
\]

\[
\alpha \sim N(0, \mathbf{I}\sigma^2_{\alpha}),
\]

\[
\phi \sim N(0, \mathbf{K}\sigma^2_{\phi}),
\]

\[
\mathbf{e} \sim N(0, \mathbf{I}\sigma^2_{\text{res}}),
\]

\[
\text{Cov}(\phi, \mathbf{e}) = 0
\]

(3)

Estimates of the model parameters were obtained using an empirical Bayes algorithm (Xavier et al. 2015) implemented in the R NAM package (https://CRAN.R-project.org/package=NAM). Genotypic values, $\mathbf{v}$, were BLUP values from (2). The polygenic term, $\phi$, accounted for genetic structure among entries through the genomic relationship matrix, $\mathbf{K}$. The model allowed each family, indicated by $\mathbf{W}$, to have a unique estimated effect of allele substitution. The predicted values for allelic substitution, $\alpha$, were interpreted as estimates of marker substitution effects from each of the founder lines. Because there were 39+1 founders, there were potentially 40 distinctive allelic substitution effects, conditioned by polygenic background ($\phi$), at each of the marker loci. The polygene was estimated with the genomic relationship matrix $\mathbf{K}$ that captured the additive relationship among individuals, thereby accounting for population structure.

Implementation of (3) in the NAM package provided an option to exclude markers linked to the marker of interest from the polygenic background using a ‘linkage window’ (Xu and Atchley 1995). If this option is not used, then the estimated effects at each marker will be adjusted for polygenic background effects that include tightly linked markers, thus reducing power to detect significant associations. At the other extreme, if the linkage window is too large, then polygenic background effects will lead to false positive associations. For exploratory analyses of experimental data, it is unlikely that there is a single best size window. Based on the evaluation of several possible window sizes we chose to report results from a window size of 5 cM. MTAs for 5 cM regions were considered significant if the $-\log_{10}(p$-value) $\geq 3$, which corresponds to a Bonferroni-corrected experiment-wide false positive probability of no greater than 0.1 for each trait. In effect, the window size of 5 cM enabled us to define unique genomic regions for multiple MTAs that were closely linked among segregating progeny in multiple families. GWAS with a window size of 20 cM resulted in a similar number of MTAs identified for yield (data not shown).

Candidate gene identification

To identify candidate genes for the co-expression network analysis, plants of the NAM common and founder lines were grown in the Danforth Center Greenhouse (St Louis, MO) and seeds at the mid-maturation stage were harvested and used for purification of total RNA as previously described (Goettel et al. 2014). RNA-seq libraries were prepared with the TruSeq RNA Sample Preparation Kit v2 following the manufacturer’s instruction (Illumina, Inc., San Diego, CA), and sequenced on the Illumina HiSeq2000. Sequence reads for each sample were independently aligned to the soybean reference genome (Ws62.a2.v1) (Schmutz et al. 2010) guided by the soybean gene annotation in Phytozome v10 using TopHat 2 (v2.0.10) (Kim et al. 2013). Cufflinks (v2.2.1) (Trapnell et al. 2012) was then run on each sample assembly to determine and normalize gene expression as the total fragments per kilobase of transcript per million mapped reads (FPKM).

To generate weighted co-expression networks, the NAM common and founder line expression matrix served as input into Camoco (Schefer et al. 2018) with the following parameters; max_gene_missing_data = 0.5,
For more than one interval, download from SoyBase (https://soybase.org/SoyNAM/index.php). The quality-assured phenotypic and SNP genotypic data are available for data and germplasm availability connection after transforming the weights were further investigated. Edges with a Z-score less than two. Genes that retained at least one edge for genes were transformed into binary variables, eliminating genes. Given the stringency of these filters, a more permissive FDR threshold of 0.35 was applied, similar to what Schaefer et al. (2018) previously used to successfully discover and validate a candidate gene flanking gene combination. Weighted max accession_missing_data = 0.4, min_single_sample_expr = 1, min_expr = 0.001, and max_val = 300. The Camoco overlap command was used to identify genes enriched for density to other genes underlying MTA peaks using 1000 bootstraps. Fifteen different networks were generated using factorial combinations of interval sizes (within 10, 20, 50 100, and 500 kb of the most significant marker, respectively) and the number of genes surrounding the most significant marker (1, 2, and 5 flanking genes, respectively). Genes with a false discovery rate (FDR) less than 0.35 were deemed significant. Our decision criteria also required that a discovered gene be significant across at least two different parameters and have a direct connection to other candidate genes. Given the stringency of these filters, a more permissive FDR threshold of 0.35 was applied, similar to what Schaefer et al. (2018) previously used to successfully discover and validate a candidate gene in maize. Gene lists were filtered for genes that appeared as significant for more than one interval/flanking gene combination. Weighted edges for genes were transformed into binary variables, eliminating edges with a Z-score less than two. Genes that retained at least one connection after transforming the weights were further investigated.

Data and germplasm availability
The quality-assured phenotypic and SNP genotypic data are available for download from SoyBase (https://soybase.org/SoyNAM/index.php). The site also contains forms for requesting seed samples of the NAM RILs and founder parents, and also images of the field-grown parents. Supplemental material available at Figshare: https://doi.org/10.25387/g3.6970496.

RESULTS AND DISCUSSION
In a combined analysis of yield across 17 environments, the yield of the common parent IA3023 was superior to the yields of all founders except for N03 (4f105-3-4), N10 (LD00-3309) and N27 (LG05-4292) (Figure 1, Table 1). As expected, median family yields were uniformly greater for families descending from EL founders, generally lower for families from BX founders and even lower for families from PI founders.

Broad-sense heritability estimates for yield and agronomic traits on an entry mean (RIL) basis across families ranged from 0.82 to 0.93 (Table 2). Slightly more than half (0.57) of the yield genotypic variance among RILs was attributed to variance among family means while variance among RILs within families accounted for the remainder. For other agronomic traits, this proportion ranged from 0.18 to 0.28 (Table 2). Genotypic correlation estimates were 0.40 for yield and maturity, -0.20 for yield and lodging, but near-zero for yield and either plant height or seed mass.

A 5-cM linkage window size was used to distinguish multiple linked MTAs from broader background polygenic effects (Xu and Atchley 1995), and to establish distinct genomic regions containing linked MTAs for multiple traits (Table S1). Ultimately, 23 unique genomic regions on 16 chr for plant height, 17 on 10 chr for plant lodging, and 29 on 18 chr for seed mass (Figures S1–4).

Additive allelic effects for MTAs were estimated by family relative to IA3023. No allelic effects were estimable for any MTA-family combination for any trait in which the SNP marker was not segregating (Table S1; Figure 2; Figures S1–4). The number of families with additive effect alleles per MTA ranged from as few as one to as many as 38, and the allelic effects varied in sign and magnitude (Figure S5). For yield, negative allelic effects were detected in all founder families for five MTA regions (i.e., two on chr 3, and one each on chr 13, 19, and 20), indicating that no founder allele was superior to the IA3023 allele at these loci (Figure 2). The allelic effect also was negative for all but one founder in four other regions (i.e., chr 10, 15, 18, and 19) and for all but two in one other region (i.e., chr 17). Conversely, a positive allelic effect was detected in all families in four significant MTA regions (i.e., chr 3, 11, 16, and 19), implying that the yield potential of IA3023 might be
improved by introgression of the corresponding estimated effect for the allele from founder parent alleles. The estimated allelic effects were split between the common allele and the founder alleles for the remaining nine MTAs, with negative allelic effects generally outnumbering positive allelic effects from the non-IA3023 founder, except for MTAs on chr 9 and 16. This implies that two different founder alleles might exist, with one being more, and the other less, favorable than the IA3023 allele. A sizeable contrast in the +/− allele effect magnitude was observed in some cases, such as the chr 9 MTA, for which N06 (CL0J173-6-8) contributed an allele associated with a +239 kg ha−1 effect, whereas N50 (PI 561370) contributed an allele associated with a −326 kg ha−1 effect (Figure 2; Table S1). We suggest several possible explanations for inconsistent sign and magnitude of estimated allelic effects among the families. Loci with contrasting effects for yield may be segregating near the same marker in different families due to historical recombination between these loci and markers in founder parents. Also, marker alleles may tag multi-allelic functional haplotypes, most likely in cis-regulatory regions, with distinctly different quantitative impacts on the trait (Swinnen et al. 2016; Rodríguez-Leal et al. 2017). Third, genomic backgrounds could influence expression of cis-regulatory regions of MTAs (Rodríguez-Leal et al. 2017).

The estimated average allelic effect of MTAs for yield across all families was negative (i.e., −16.7 kg ha−1), which was expected given that IA3023 is higher yielding than all but three founders. When averaged by founder class, the mean allelic effect was +3.4 kg ha−1 for EL founder families compared to −19.2 kg ha−1 and −52.1 kg ha−1 for BX and PI founder families, respectively (Figure 2; Table S1).

A key project goal was to identify positive effect alleles for yield in significant MTAs that were present in PI or BX founders, but non-existent, rare, or small in effect in EL founders. No such alleles were
detected in any PI founder family (Figure 2, Table S1), but were detected in the BX founder family N38 (LG00-3372), which notably contributed 14 positive yield effect alleles (vs. IA3023 allele) – three of which were EL-absent. BX founder N39 (LG04-6000) also was unique for contributing a rare but very large (+157 kg ha\(^{-1}\)) effect allele on chr 8. These two founder lines were derived from recently introduced germplasm from China. It would be of interest to determine if an estimated yield effect of that size would be maintained were it to be introgressed into IA3023. Several EL founders also could be used for the improvement of IA3023 yield, notably N06 (CL0J173-6-8), N03 (4J105-3-4), and N04 (5M20-2-5-2), which contributed many more positive effect alleles.

To identify MTAs for two or more traits that could be either a single multi-trait pleiotropic QTL, or closely linked QTL exhibiting linkage-phased allelic effects, significant MTAs for yield, maturity, height, lodging, or seed weight that clustered within 5 cM of each other were binned (Table 3; Tables S1-S2). Cases in which yield QTL and maturity MTAs that were tagged with a common SNP that mapped within 200 kb of this gene (Table 4; Table S1-S2). GLYMA.09G001300

Table 3 Significant marker trait associations (MTA) for seed yield (Yd), seed mass (Ms), date of maturity (Mt), plant height (Ht), and lodging (Lg) that were placed in separate bins when the MTA peaks were greater than 5 cM apart. The 21 bins with MTAs for more than one trait are highlighted with gray. See Table S2 for detailed information including interval locations and MTA p-values

| Chr | Bin | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|-----|-----|---|---|---|---|---|---|---|
| 1   | Ms  | Ms |   |   |   |   |   |   |
| 2   | Yd  | Ms |   |   |   |   |   |   |
|     |     | Mt |   |   |   |   |   |   |
|     |     | Ht |   |   |   |   |   |   |
| 3   | Mt  | Yd | Ldg| Yd | Ht | Ldg| Yd |   |
|     |     | Ht | Ms |   |   |   | Ms |   |
| 4   | Ms  | Ldg| Ms |   | Mt |   |   |   |
|     |     | Ht |   |   |   |   |   |   |
|     |     | Lg |   |   |   |   |   |   |
| 5   | Ms  | Lg |   |   | Ms |   |   |   |
| 6   | Ms  | Mt | Yd |   | Mt |   | Ms |   |
|     |     | Ht |   |   |   |   |   |   |
| 7   | Mt  | Ht | Ms | Ms | Ms | Ldg| Yd |   |
|     |     | Lg |   |   |   |   |   |   |
| 8   | Mt  | Yd |   |   |   |   |   |   |
|     |     | Ht |   |   |   |   |   |   |
| 9   | Yd  | Ms | Ms | Ht | Mt |   |   | Lg |
|     |     | Ht |   |   |   |   |   | Lg |
| 10  | Ms  | Yld| Ht |   |   |   |   | Lg |

In chr 10 bin 2, the large effect MTAs for yield and each of the other traits had a common SNP that mapped within 200 kb of GLYMA.10G221500, which is the cloned maturity gene E2 (Watanabe et al. 2011). Seven NAM families segregated for the late maturity E2 allele vs. the early maturity e2 allele (Figure S1), which was expected based on a founder parent E2 allele analysis (Langewisch et al. 2014). The E2 allele for later maturity was associated with greater yield (with one exception), taller plants, more lodging and smaller seed mass compared to the e2 allele for early maturity (Table S1), an indication of potential multi-trait pleiotropism. Coupling phased positive yield - positive maturity also was evident in chr 9 bin 1 for all founders except one BX and three PI. In the other three bins, positive/negative allele effects for maturity and yield were inconsistently phase-paired, implying separate QTL for each trait. This inconsistency was also the case for five bins with yield and seed mass, four bins with yield and lodging, and three bins with yield and plant height (Table 3, Tables S1-S2). The common SNP-tagged maturity-height MTAs in chr 12 bin 2 and the common SNP-tagged height-lodging MTAs on chr 19 bin 1 exhibited coupling phased allelic effects that were \(-/-\) in the former (two exceptions) and \(+/+\) in the latter. Surprisingly, chr 18 bin 3 contained a common SNP that tagged the lodging and seed mass MTAs, and a nearby SNP tagging a yield MTA, and all three exhibited estimates of allelic effects that were consistently positive.

To identify candidate genes underlying the yield MTAs, a weighted co-expression network analysis was performed (Schaefer et al. 2018). This analysis identifies subsets of genes residing near the significant markers, factoring for both genomic distance and number of genes near the most significant marker (see Methods section). Genes within these subnetworks were tested for significant co-expression by comparing their connectivity “density” to all genes in the MTA subnetwork vs. the connectivity of genes in random subnetworks equal in size (via resampled distribution). Fifteen different subnetworks were developed. Genes significant in at least two different subnetworks and displaying strong connectivity with one another were deemed candidate genes for the yield MTAs. Ten such genes were identified, connected in two distinct modules (Table 4; Fig. S6).

The candidate gene GLYMA.09G001300 corresponds to chr 9 bin 1 yield and maturity MTAs that were tagged with a common SNP that maps within 8 kb of this gene (Table 4; Table S1-S2). GLYMA.09G001300
Table 4 Candidate genes underlying marker trait association (MTA) peaks resulting from co-expression analysis

| Module | Candidate gene     | MTA chr | MTA position | Pvalue (-log10) | Annotation                                           | Arabidopsis ortholog |
|--------|-------------------|---------|--------------|----------------|-----------------------------------------------------|----------------------|
| 1      | GLYMA.09G001300   | 9       | 106196       | 10.77          | FAR1 DNA-binding domain                             | AT4G15090            |
| 1      | GLYMA.06G325300   | 6       | 5131639      | 9.29           | Methylenetratrihydrofolate reductase                | AT2G44160            |
| 1      | GLYMA.12G242300   | 12      | 40077424     | 4.21           | Dynamin family                                      | AT1G59610            |
| 1      | GLYMA.12G242500   | 12      | 40077424     | 4.21           | Methyltransferase domain                            | AT1G22800            |
| 1      | GLYMA.11G106500   | 11      | 8106372      | 3.92           | Plasma-membrane choline transporter family protein  | AT4G38640            |
| 2      | GLYMA.03G022900   | 3       | 2428982      | 3.44           | SEC7-like guanine nucleotide exchange family protein | AT3G60860            |
| 2      | GLYMA.10G219500   | 10      | 45110885     | 40.77          | RNA-binding protein                                 | AT1G71800            |
| 2      | GLYMA.16G000500   | 16      | 34320        | 15.04          | Ribonucleoprotein-related                           | AT4G24270            |
| 2      | GLYMA.13G278500   | 13      | 37965132     | 5.66           | Nucleolar GTP-binding family protein                | AT1G52980            |
| 2      | GLYMA.11G245000   | 11      | 33885696     | 3.43           | Regulator of chromosome condensation (RCC1)         | AT3G26100            |

ACKNOWLEDGMENTS

This research was supported by soybean checkoff funds from the United Soybean Board.

LITERATURE CITED

Baigain, P., M. N. Rouse, T. J. Tsio, G. K. Macharia, S. Bhanvani et al., 2016 Nested association mapping of stem rust resistance in wheat using genotyping by sequencing. PLoS One 11: e0155760. https://doi.org/10.1371/journal.pone.0155760

Bandillo, N., D. Jarquin, Q. Song, R. Nelson, P. Cregan et al., 2015 A population structure and genome-wide association analysis on the USDA soybean germplasm collection. Plant Genome 8: 1–13. https://doi.org/10.3833/plantgenome.2015.04.0024

Bates, D., M. Maechler, B. Bolker, and S. Walker, 2015 Fitting linear mixed-effect models using lme4. J. Stat. Softw. 67: 1–48. https://doi.org/10.18637/jss.v067.i01

Bernardo, R., 2002 Breeding for quantitative traits in plants, Stemma Press, Woodbury, MN.

Brown, P. J., N. Upadayya, G. S. Mahone, F. Tian, P. J. Bradbury et al., 2011 Distinct genetic architectures for male and female inflorescence traits of maize. PLoS Genet. 7: e1002383. https://doi.org/10.1371/journal.pgen.1002383

Buckler, E. S., J. B. Holland, C. B. Acharya, P. J. Brown et al., 2009 The genetic architecture of maize flowering time. Science 325: 714–718. https://doi.org/10.1126/science.1174276

Cook, J. P., M. D. McMullen, J. B. Holland, F. Tian, P. Bradbury et al., 2012 Genetic architecture of maize kernel composition in the nested association mapping and inbred association panels. Plant Physiol. 158: 824–834. https://doi.org/10.1104/pp.111.185033

Favery, B., L. A. Chelysheva, M. Lebris, F. Jammes, A. Marmagne et al., 2004 Arabidopsis formin AtFH6 is a plasma membrane-associated protein upregulated in giant cells induced by parasitic nematodes. Plant Cell 16: 2529–2540. https://doi.org/10.1105/tpc.104.024372

Goettel, W., E. Xia, R. Uphuchur, M. L. Wang, P. Chen et al., 2014 Identification and characterization of transcript polymorphisms in soybean lines varying in oil composition and content. BMC Genomics 15: 299. https://doi.org/10.1186/1471-2164-15-299

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. https://doi.org/10.1093/nar/gkp798

Hartman, G. L., E. D. West, and T. K. Herman, 2011 Crops that feed the world 2. Soybean-worldwide production, use, and constraints caused by pathogens and pests. Food Secur. 3: 5–17. https://doi.org/10.1007/s12571-010-0108-x

Hirschhorn, J. N., and M. J. Daly, 2005 Genome-wide association studies for common diseases and complex traits. Nat. Rev. Genet. 6: 95–108. https://doi.org/10.1038/nrg1521

Holland, J. B., W. E. Nyquist, and C. T. Cervantes-Martinez, 2003 Estimating and interpreting heritability for plant breeding: an update. Plant Breed. Rev. 22: 9–112.

Kim, K. S., B. W. Diers, D. L. Hyten, M. A. Rouf Mian, J. G. Shannon et al., 2012 Identification of positive yield QTL alleles from exotic soybean germplasm in two backcross populations. Theor. Appl. Genet. 125: 1353–1369. https://doi.org/10.1007/s00122-012-1944-1
Kim, D., G. Perteau, C. Trapnell, H. Pimentel, R. Kelley et al., 2013 TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 14: R36. https://doi.org/10.1186/gb-2013-14-4-r36

Kingsmore, S., I. E. Lindquist, J. Mudge, D. D. Gessler, and W. D. Beavis, 2008 Genome-wide association studies: progress and potential for drug discovery and development. Nat. Rev. Drug Discov. 7: 221–230. https://doi.org/10.1038/nrd2519

Knapp, S. J., W. W. Stroup, and W. M. Ross, 1985 Exact con-

Korte, A., and A. Farlow, 2013 The advantages and limitations of trait analysis with GWAS: a review. Plant Methods 9: 29. https://doi.org/10.1186/1746-4811-9-29

Kump, K. L., P. J. Bradbury, R. J. Wisser, E. S. Buckler, A. R. Belcher et al., 2011 Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. Nat. Genet. 43: 163–168. https://doi.org/10.1038/ng.747

Langewisch, T., H. Zhang, R. Vincent, T. Joshi, D. Xu et al., 2017 Construction of high resolution genetic linkage maps to improve the soybean genome sequence assembly Glyma1.01. BMC Genomics 17: 33. https://doi.org/10.1186/s12864-015-2344-0

Langewisch, T., H. Zhang, R. Vincent, T. Joshi, D. Xu et al., 2017 Construction of high resolution genetic linkage maps to improve the soybean genome sequence assembly Glyma1.01. BMC Genomics 17: 33. https://doi.org/10.1186/s12864-015-2344-0

Li, J., S. Gao, C. Martínez, G. He, Z. Zhou et al., 2010 Arabidopsis transcrip-
tion factor ELONGATED HYPOCOTYL5 plays a role in the feed-

Liu, S., C. Fu, J. Gou, L. Sun, D. Huhman et al., 2017 Simultaneous downregulation of MTHFR and COMT in switchgrass affects plant per-
formance and induces lesion-mimic cell death. Front. Plant Sci. 8: 982. https://doi.org/10.3389/fpls.2017.00982

Maurer, A., V. Draba, Y. Jiang, F. Schnaittahm, R. Sharma et al., 2015 Modelling the genetic architecture of flowering time control in barley through nested association mapping. BMC Genomics 16: 290. https://doi.org/10.1186/s12864-015-1459-7

Maurer, A. V., Arba, Y. Jiang, F. Schnaittahm, R. Sharma et al., 2015 Modelling the genetic architecture of flowering time control in barley through nested association mapping. BMC Genomics 16: 290. https://doi.org/10.1186/s12864-015-1459-7

Maurer, A., V. Draba, and K. Pillen, 2016 Genomic dissection of plant development and its impact on thousand grain weight in barley through nested association mapping. J. Exp. Bot. 67: 2507–2518. https://doi.org/10.1093/jxb/erv070

McMullen, M. D., S. Kresovich, H. S. Villa, P. Bradbury, H. Li et al., 2009 Genetic properties of the maize nested association mapping population. Science 325: 737–740. https://doi.org/10.1126/science.1174320

Olkoku, B. A., G. F. Wang, V. Vontimitta, B. P. Venkata, S. Marla et al., 2014 A genome-wide association study of the maize hypersensitive defense response identifies genes that cluster in related pathways. PLoS Genet. 10: e1004562. https://doi.org/10.1371/journal.pgen.1004562

Peiffer, I., A. M., C. Romay, M. A. Gore, S. A. Flint-Garcia, Z. Zhang et al., 2013 The genetic architecture of maize stalk strength. PLoS One 8: e67066. https://doi.org/10.1371/journal.pone.0067066

Phansak, P., W. Soonsuwon, D. L. Hyten, Q. Song, P. B. Cregan et al., 2016 Multi-Population Selective genotyping to identify soybean [Glycine max (L.) Merr.] seed protein and oil QTLs. G3 (Bethesda) 6: 1635–1648. https://doi.org/10.1534/g3.116.027656

Poland, J. A., P. J. Bradbury, E. S. Buckler, and R. J. Nelson, 2011 Genome-

Ray, D. K., N. D. Mueller, P. C. West, and J. A. Foley, 2013 Yield trends are insufficient to double global crop production by 2050. PLoS One 8: e66428. https://doi.org/10.1371/journal.pone.0066428

Rodríguez-Leal, D., Z. H. Lemmon, J. Man, M. E. Bartlett, and A. B. Lippman, 2017 Engineering quantitative trait variation for crop improvement by genome editing. Cell 171: 470–480.e8. https://doi.org/10.1016/j.cell.2017.08.030

Schafer, R. J., J. M. Michno, J. Jeffers, O. Hoekenga, B. Dilkies et al., 2018 Integrating co-expression networks with GWAS detects genes driving elemental accumulation in maize seeds. bioRxiv 221655.

Schmutz, J., S. B. Cannon, J. Schluter, J. Ma, T. Mitros et al., 2010 Genome seq-

Song, Q., D. L. Hyten, G. Jia, C. V. Quigley, E. W. Fickus et al., 2013 Development and evaluation of SoySNP50K, a high-density gen-

Song, Q., J. Jenkins, G. Jia, D. L. Hyten, V. Pantalone et al., 2016 Construction of high resolution genetic linkage maps to improve the soybean genome sequence assembly Glyma1.01. BMC Genomics 17: 33. https://doi.org/10.1186/s12864-015-2344-0

Taylor, N., 2011 A role for Arabidopsis dynamin related proteins DRP2A/B in endocytosis: DRP2 function is essential for plant growth. Trends Plant Sci. 21: 506–515. https://doi.org/10.1016/j.tplants.2016.01.014

Tian, F., P. J. Bradbury, P. J. Brown, H. Hung, Q. Sun et al., 2011 Genome-

Watanabe, S., Z. Xia, R. Hideshima, 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. https://doi.org/10.1038/nbt.2450

Xavier, A., B. Hall, A. A. Hearst, K. A. Cherkauer, and K. M. Rainey, 2017 Genetic architecture of phenomic-enabled canopy coverage in Glycine max: Genetics 206: 1081–1089. https://doi.org/10.1534/genetics.111.199713

Xavier, A., S. Xu, W. M. Muir, and K. M. Rainey, 2015 NAM: association studies in multiple populations. Bioinformatics 31: 46–53. https://doi.org/10.1093/nbtl.2450

USDA, 2016 World agricultural production. USDA Foreign Agricultural Service, Circular Series WAP. 6–16.

Communicating editor: D. J. de Koning