G3 Genes | Genomes | Genetics

INVESTIGATION

Quantitative Genomic Dissection of Soybean Yield Components

Alencar Xavier*,† and Katy M Rainey*,1

*Department of Agronomy, Purdue University, West Lafayette IN 47907 and †Department of Biostatistics, Corteva Agrisciences, Johnston IA 50131

ORCID ID: 0000-0001-5034-9954 (A.X.)

ABSTRACT Soybean is a crop of major economic importance with low rates of genetic gains for grain yield compared to other field crops. A deeper understanding of the genetic architecture of yield components may enable better ways to tackle the breeding challenges. Key yield components include the total number of pods, nodes and the ratio pods per node. We evaluated the SoyNAM population, containing approximately 5600 lines from 40 biparental families that share a common parent, in 6 environments distributed across 3 years. The study indicates that the yield components under evaluation have low heritability, a reasonable amount of epistatic control, and partially oligogenic architecture: 18 quantitative trait loci were identified across the three yield components using multi-approach signal detection. Genetic correlation between yield and yield components was highly variable from family-to-family, ranging from -0.2 to 0.5. The genotype-by-environment correlation of yield components ranged from -0.1 to 0.4 within families. The number of pods can be utilized for indirect selection of yield. The selection of soybean for enhanced yield components can be successfully performed via genomic prediction, but the challenging data collections necessary to recalibrate models over time makes the introgression of QTL a potentially more feasible breeding strategy. The genomic prediction of yield components was relatively accurate across families, but less accurate predictions were obtained from within family predictions and predicting families not observed included in the calibration set.

KEYWORDS
soybean
genomic prediction
GWAS
GxE
yield
yield components
heritability
SoyNAM

Soybean is a field crop of major importance due to its seed composition, containing approximately 40% protein and 20% oil. Its unique composition and scalable production make soy a key crop to world-wide food security (Qiu *et al.* 2013). However, soybean germplasm has narrow genetic basis (Carter *et al.* 2004, Mikel *et al.* 2010) that has limited the rate of genetic gains of yield grain to 29 kg/ha/year in North America (Rincker *et al.* 2014). Better breeding strategies are needed to explore soybeans' full genetic potential (Specht *et al.* 1999, 2014), and a possible approach to increase grain yield is through trait dissection, breaking down yield into yield components. In fact, whereas modern cultivars have around 30 pods per plant (Kahlon *et al.* 2011), some accessions have as many as 200 pods per plant (Zhang *et al.* 2015).

doi: https://doi.org/10.1534/g3.119.400896
Manuscript received September 23, 2019; accepted for publication December 6, 2019; published Early Online February 1, 2020.

[1]Corresponding author: E-mail: krainey@purdue.edu

Kahlon and Board (2012) contrasted cultivars released over the past few decades and observed that grain yield increases may have been triggered by changes in yield components over time, particularly in pods and nodes. Suhre *et al.* (2014) found that the number of nodes and pods per node have steadily increased in cultivars released from 1920 to 2010. The number of pods and nodes are key yield-driver (Robinson *et al.* 2009) that reflects the efficiency of the complex physiological process (Board and Tan 1995). These yield components can be increased at farming levels with good agronomic practices and high-end genetics (Board and Kahlon 2011, Kahlon *et al.* 2011). However, the labor-intensive phenotyping of counting soybean pods and nodes can restrict the number of entries and most studies have been conducted with a small number of genotypes (Egli and Bruening 2006, Robinson *et al.* 2009, Kahlon *et al.* 2011, Nico *et al.* 2019).

The first large-scale genetic assessment of complex traits was performed in the SoyNAM population, where 5600 genotypes from 40 biparental families sharing a common parent were phenotyped for various agronomic traits (Xavier *et al.* 2016, Xavier *et al.* 2017a, Diers *et al.* 2018). Whereas soybeans have constrained genetic diversity (Carter *et al.* 2004), the SoyNAM is a relatively rich panel of locally adapted

genotypes that represents an invaluable resource for the breeding community.

From a preliminary analysis in the SoyNAM population, Xavier *et al.* (2017a) found that grain yield presents strong genetic correlation to yield components, canopy development, and the length of the reproductive period. The latter is a function of days to flowering and days to maturity, both traits controlled by a few major genes (Watanabe *et al.* 2009, 2011, Xia *et al.* 2012, Langewisch *et al.* 2014). The genetic architecture of canopy development has been recently described by Xavier *et al.* (2017b) and Kaler *et al.* (2018). However, the in-depth genetic architecture of yield components had not been characterized with sufficient power and resolutions.

This study aims to conduct a set of quantitative genetic analyses performed with genome-wide markers to unravel the underlying architecture of yield components and assess potential breeding applications. Our evaluation approach includes comparing different strategies for genomic prediction within and across family; perform genomic covariance analysis to uncover the pleiotropy between yield and yield components, as well as the amount of genetic variation attributed to epistasis and genotype-by-environment interactions; and multi-approach association studies to identify regions containing QTL with the potential to be deployed for marker-assisted selection.

METHODS

Population

The panel under evaluation is a nested association panel, namely the SoyNAM populations, where the standard parent IA3023 (Dairyland DSR365 x Pioneer P9381) was crossed to 40 founder parents that attempt to capture the diversity of public germplasm, each family comprising approximately 140 individuals. Among the 40 founder parents, 17 lines are U.S. elite public germplasm, 15 have diverse ancestry, and eight are planted introductions. The descriptions of parents are available https://www.soybase.org/SoyNAM/. The population's maturity ranged from late maturity group II to early maturity group IV. More details about the population composition are available in Diers *et al.* (2018) and Xavier *et al.* (2018). After quality control based on segregation patterns, 5363 individuals were used for this study.

Experimental design

The experiment was conducted under a modified augmented design, with a 7:1 lines-to-check ratio, in two Purdue University research centers: Throckmorton-Purdue Agricultural Center (TPAC) located in Throckmorton, Indiana, and at the Agronomy Center for Research and Education (ACRE) in West Lafayette, Indiana. The experiments were planted during the third week of May in two-row plots (2.9m × 0.76m), at a density of approximately 36 plants m^{-2}. The phenotypes were collected in 10 field blocks, these being distributed as 4 adjacent blocks in 2013, 4 adjacent blocks in 2014 and 2 field blocks in different locations in 2015. In 2013 and 2014 the experiments were conducted at the ACRE farm, where each field block contained all 40 families with 35 recombinant inbred lines (RIL) per family, that is, one-quarter of the total number of RILs. In 2013 and 2014 RILs were not replicated, but the same checks were used across fields. In 2015, the experiments were conducted on 6 of the 40 SoyNAM families in two locations, ACRE and TPAC, with two replicates per location.

Phenotyping

The number of pods and nodes was counted in the main stem, between phenological stages R5 and R7, averaging the counts of 3, 6 and 4 representative plants per plot in 2013, 2014 and 2015 respectively. The variable number of subsamples varied according to the resources available each year. The number of pods per node was obtained by the ratio. Grain yield was collected at harvest, converting the grain weight from individual plots to bushels per acre adjusted to 13% grain moisture. The number of days to maturity (Fehr *et al.* 1971) was collected by scoring the plots every 3 days from the time where the first mature plot was observed, using back-and-forth scoring to assign the plots that matured between scoring dates.

Genotyping

The genetic information was collected from Illumina SoyNAM BeadChip SNP array specially designed for SoyNAM, comprising 5305 SNP markers selected from the sequencing of all 41 parental lines (Song *et al.* 2017). Missing loci were imputed using a hidden Markov model and removed markers with minor allele frequency below 0.05 using the R package NAM (Xavier *et al.* 2015). A total of 4240 SNPs were used for genomic analysis.

Genetic merit

The genetic values were estimated as the best linear unbiased predictors (BLUP), as a random term of a mixed model. The mixed linear model was fitted with variance components based on restricted maximum likelihood (REML), computed using the R package lme4 (Bates *et al.* 2014). The linear model used to model genetic values:

$$y = \mu + f(s) + Zu + Wg + e$$

Where the response variable y was modeled as a function of an intercept μ, spatial covariate $f(s)$ based on a moving-average of neighbor plots as described by Lado *et al.* (2013) implemented in the functions NNscr/NNcov of the R package NAM (Xavier *et al.* 2015), a random effect Zu to capture the genetic effects of individual lines, namely the genetic effects, assumed to be normally distributed as $u \sim N(0, \sigma^2{}_u)$, a nuisance random effect Wg to capture the local environment effects, as normally distributed as $g \sim N(0, \sigma^2{}_g)$, and a vector $e$of residuals, normally distributed $e \sim N(0, \sigma^2{}_e)$. The inverse phenotypic variance was computed for each environment and used as observation weights to account for the heteroscedasticity among trials. Although the checks were not explicitly included in the genetic merit model, these were invaluable for the spatial correction of the field plot variation. Broad-sense heritability (H) was estimated from the REML variance components as:

$$H = \frac{\sigma^2{}_u}{\sigma^2{}_u + r^{-1}\sigma^2{}_e}.$$

Where r is the average number of replicates per entry. The reliability of the j^{th} genotype (H_j) was used to deregress (Garrick *et al.* 2009) its corresponding BLUP (u_j) in order to obtain the genetic values in natural scale ($y_j = u_j/H_j$). This procedure of unshrink BLUPs precludes the downstream analyses to be performed upon a vector of phenotypes with heterogeneous degree of shrinkage, which may lead to biased results.

The narrow-sense heritability estimates (h^2) were based on the following SNP-BLUP model:

$$y = \mu + Ma + \varepsilon$$

Where y correspond to the genetic values, modeled as a function of an intercept μ, matrix with SNP information and marker effects (Ma), and the vector of residuals (ε). Both marker effects and residuals were assumed to be normally distributed with variances σ^2_a and σ^2_ε,

respectively. The narrow-sense heritability was computed under two scenarios: 1) deploying all markers and 2) only with the markers found to be associated with yield components. The narrow-sense heritability was estimated as follows:

$$h^2 = \frac{\sigma^2{}_a \ \times 2 \sum_{j=1}^{J} p_j(1-p_j)}{\sigma^2{}_a \ \times 2 \sum_{j=1}^{J} p_j(1-p_j) + \sigma^2{}_\varepsilon}.$$

Polygenic epistasis

We performed a within-family variance component analysis to determine the amount of variability jointly explained by additive and additive-by-additive epistasis. For that, we fit a kernel-based model referred to as the G2A model (Zeng *et al.* 2005). Variance components were estimated using REML estimates (Misztal 2008). The analysis followed the linear model:

$$y = \mu + \psi + \omega + \varepsilon$$

$$\psi \sim N(0, K\sigma^2{}_\psi)$$

$$\omega \sim N(0, Q\sigma^2{}_\omega)$$

$$\varepsilon \sim N(0, I\sigma^2{}_\varepsilon)$$

Where y correspond to the genetic values, modeled as a function of an intercept (μ), additive genetic values (ψ), additive epistatic value (ω), and the vector of residuals, (ε). The relationship matrices were built in accordance to Zeng *et al.* (2005) and Xu (2013). The additive genetic relationship matrix was obtained by the cross-product of the centralized marker matrix (M) with centralized trace, thus K=MM'α with α=n×Tr{(MM')^{{−1}}}, and the additive epistatic relationship matrix was computed by the additive Hadamard product with centralized trace, thus Q=(MM'\##MM')α with a normalizing factor α=n×Tr{(MM'\##MM')^{{−1}}}.

Multivariate analysis of pleiotropy and stability

Multivariate analysis, namely genetic and additive genetic correlations, allows exploring the interaction between traits across years (pleiotropy) or within trait between years (stability or genotype-by-environment correlation). The genetic correlations within-family were obtained as the Pearson's correlation between the BLUPs of yield and yield components for pleiotropy analysis, as well as the correlations of yield components from year to year for stability analysis. We estimated the additive genetic correlation between yield and yield components for pleiotropy analysis, and yield components across years for stability analysis, for each of the 40 families using a multivariate GBLUP model. The GBLUP model was fit with REML variance components. For the multivariate polygenic analysis, we fitted the following multi-trait model:

$$y = \mu + \psi + \varepsilon$$

$$\psi \sim N(0, K \otimes \Sigma_\psi)$$

$$\varepsilon \sim N(0, I \otimes \Sigma_\varepsilon)$$

Where, under multivariate settings, $y = \{y_1, y_2, ..., y_k\}$ correspond to the genetic merits, modeled as a function of their corresponding intercepts, $\mu = \{\mu_1, \mu_2, \ldots, \mu_k\}$, the additive genetic values, $\psi = \{\psi_1, \psi_2, \ldots, \psi_k\}$, and the residuals, $\varepsilon = \{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_k\}$. With respect to the model variances, $K$is the relationship matrix defined in the previous model, the additive covariances Σ_ψ is a dense $k \times k$ matrix where the *ij* cell corresponds to the additive genetic covariance $\sigma_\psi(i,j)$ between i^{th} and j^{th} traits, and the residual covariance was assumed to be diagonal $\Sigma_\varepsilon = diag(\sigma^2{}_{\varepsilon 1}, \sigma^2{}_{\varepsilon 2}, \ldots, \sigma^2{}_{\varepsilon k})$. Additive genetic correlations were estimated from the covariance components as $\rho_\psi(i,j) = \sigma_\psi(i,j)/[\sigma_\psi(i)\sigma_\psi(j)]$. From the genetic correlations and heritabilities, the efficiency of indirect selection (Falconer and Mackay 1996) using i^{th} trait to select the j^{th} trait was estimated as $E = h_j^{-2} \ h_i^{\ 2} \ \rho_\psi(i,j)$.

Association studies

Since various signal detection strategies may capture different QTL (Yang *et al.* 2018), three complementary methodologies of genome-wide association studies were deployed in this study: Single marker analysis, implemented in the R package NAM (Xavier *et al.* 2015), whole-genome regression BayesCpi (Habier *et al.* 2011) implemented in the R package bWGR (Xavier *et al.* 2019), and random forest implemented in the R package ranger (Wright and Ziegler 2015). A brief description of the methods is provided below.

Mixed Linear Model (MLM): This method of an association study is based on the likelihood ratio between a model containing the marker of interest (full model) and a model without the marker (reduced model). Both models include a polygenic term that accounts for the population structure. The statistical model that describes this association study is tailored to NAM populations (Xavier *et al.* 2015) and follows the linear model:

$$y = \mu + X\beta + \psi + e$$

Where the genetic values (y) are modeled as a function of an intercept (μ), the matrix containing the interaction between the SNP information and family for the target marker under evaluation (X), the vector of marker effect within family $\beta \sim N(0, \sigma^2{}_\beta)$, the vector of independent residuals, $\varepsilon \sim N(0, \sigma^2{}_\varepsilon)$, and the polygenic term defined previously, $\psi \sim N(0, K\sigma^2{}_\psi)$, which parametrizes the genetic covariance among individuals through the full-ranking genomic relationship matrix K. Bonferroni thresholds were utilized to account for multiple testing and mitigate false-positives, yielding a two-sided threshold of -log10(0.025/4240)=5.23. The association model was fit with REML variance components.

Whole-genome regression (WGR): Designed primarily for prediction, WGR methods fit all markers at once. The prior distribution of marker effects follows a mixture of distributions to perform feature selection. The association statistics are based on the posterior probability of each marker to be included in the model, or "model frequency". The model of choice, BayesCpi, assumes each marker has a probability π of being included in the model, where the parameter π is estimated in each MCMC iteration. Markers reached statistical significance if 1-π was smaller than a two-sided threshold of $\alpha = 0.05$, which translates into a threshold for the Manhattan plot of -log10(0.025)=1.6. The linear model that describes BayesCpi is the following:

$$y = \mu + Ma + e$$

Where y correspond to the genetic values, modeled as a function of an intercept (μ), the matrix containing the all SNP information (M) and

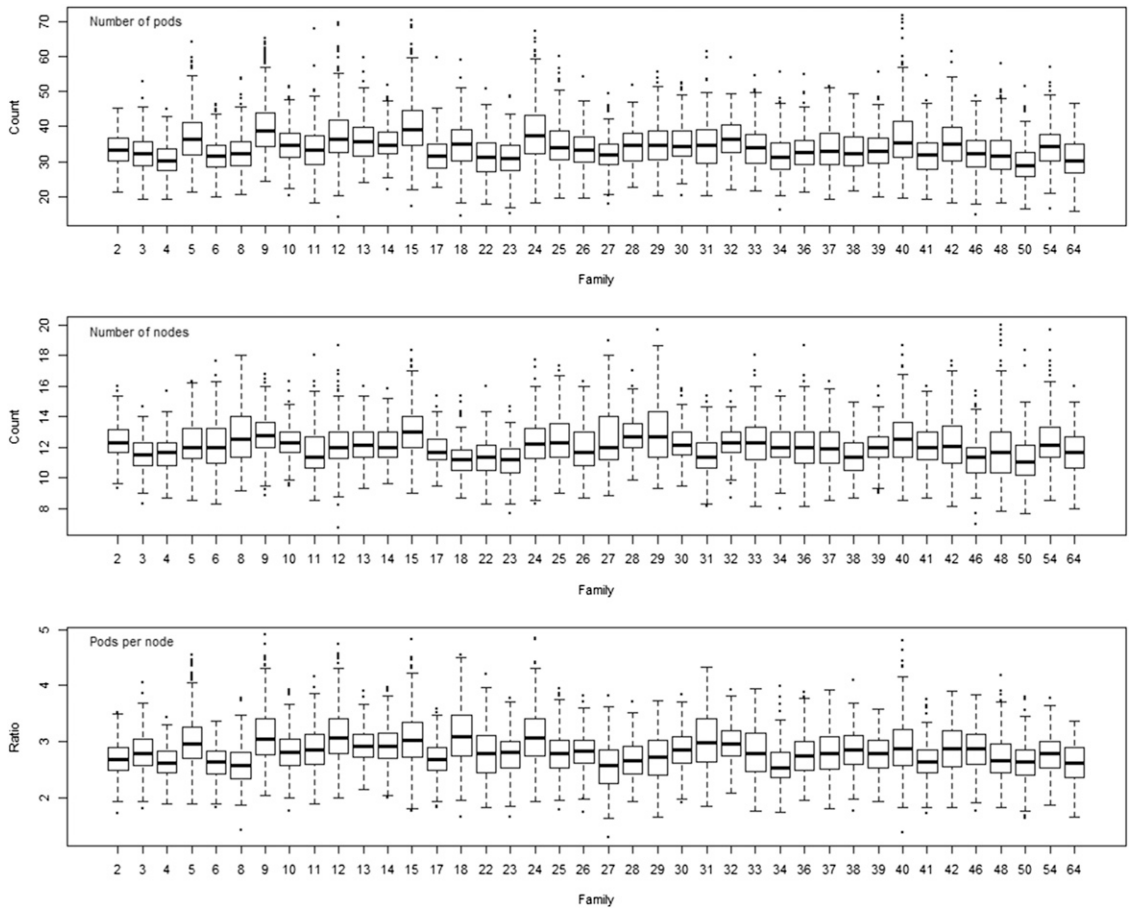

Figure 1 Phenotypic distribution of the pod number (top), node number (center) and pods per node (bottom). Families had elite (2-23), diverse (2-39) and exotic (40-64) genetic background.

the vector of all marker effects jointly estimated (a), which followed a mixture of distributions, having probability π of having null effect and probability $1-\pi$ or being normally distributed as $N(0,\sigma^2_\beta)$, and the vector of independent residuals, $\varepsilon \sim N(0,\sigma^2_\varepsilon)$. The marker and residual variances were assumed to follow an inverse scaled chi-squared distribution, $\sigma^2_\beta \sim \chi^2(S_\beta,\nu_0)$and $\sigma^2_\varepsilon \sim \chi^2(S_\varepsilon,\nu_0)$, assuming $\nu_0 = 5$ prior degrees of freedom and shape parameters computed assuming prior heritability of 0.5 (Pérez and de Los Campos 2014), thus $S_\beta = 0.5\ \sigma^2_y\ MSx^{-1}(1-\pi)^{-1}$and $S_\varepsilon = 0.5\ \sigma^2_y$. The model was fit with 20000 MCMC iterations, discarding the initial 2000 iterations, and no thinning, such that the posterior means were computed by averaging 18000 MCMC iterations.

Random forest regression (RFR): Random forest is a non-parametric regression derived from the bootstrapping aggregation of decision trees built from subsets of data and parameters. The association statistics of RFR is based on feature importance (Botta *et al.*, 2014). The forest was grown with 10000 decision trees. The trees were built having as starting point $\sqrt{m} = 65$ SNPs sampled at random with replacement. The metric of variable importance was the 'impurity' index, which is a measure of the out-of-bad explained variance. Because there is no objective way of defining an association threshold for significant SNPs, we estimated the global empirical threshold (Doerge and Churchill 1996) based on 1000 permutations ($\alpha = 0.05$), thus making no assumptions about the distribution of the associations.

Cross-validation studies

Cross-validations were performed for each yield component. Due to the known population structure of the SoyNAM, three types of cross-validations were performed: (1) within-family, (2) across-family, and (3) leave-family-out. Within- and across-family validations were performed as fivefold cross-validation, randomly selecting 80% of the data as a calibration set, and using the remaining 20% as a prediction target. The sampling and prediction procedure is repeated 25 times. Leave-family-out validation use 39 families to predict the family left out, and the procedure is performed to all 40 families. The prediction statistic is the predictive ability (PA), as the correlation between predicted and observed values.

The cross-validation was performed using the functions *emCV* of the R package bWGR (Xavier *et al.* 2019). In accordance with the genomic prediction benchmark proposed by Daetwyler *et al.* (2013), two statistical models evaluated in this study were GBLUP (VanRaden 2008) and BayesB (Meuwissen *et al.* 2001). The GBLUP model was fitted as a ridge regression with REML variance components, and the BayesB assumes that markers effects follow a mixture of distribution, where the j^{th} marker had probability $\pi = 0.95$ of having null effect and probability $1-\pi$ of being normally distributed as $N(0,\sigma^2_{\beta j})$, variances were assumed to follow an inverse scaled chi-squared distribution, $\sigma^2_{\beta j} \sim \chi^2(S_\beta,\nu_0)$and $\sigma^2_\varepsilon \sim \chi^2(S_\varepsilon,\nu_0)$, assuming $\nu_0 = 10$ prior degrees of freedom and shape parameters computed as $S_\beta = 0.5\ \sigma^2_y\ MSx^{-1}$and $S_\varepsilon = 0.5\ \sigma^2_y$.

Table 1 Trait distribution (mean and standard deviation) and genetic metrics: broad-sense heritability (H), narrow-sense heritability (h^2) estimated using all SNPs and the subset of significant SNPs

Trait	Mean	Std. Dev.	H	h^2(all SNPs)	h^2 (QTL SNPs)
Nodes	12.085	1.090	0.352	0.159	0.069
Pods	34.046	4.955	0.361	0.110	0.095
P/N	2.819	2.819	0.301	0.064	0.142
Yield	66.557	14.345	0.334	0.280	0.093

Data availability

All phenotypic and genotypic data are available in the R package SoyNAM available on CRAN. To access the data, install the SoyNAM package (CRAN.R-project.org/package = SoyNAM), then load the Indiana dataset with the following command in R: data(soyin, package='SoyNAM')

RESULTS

The SoyNAM provided reasonable variation for the three yield components. The phenotypic distributions of the yield components for each of the SoyNAM families is presented in Figure 1. The mean and standard deviation across families is provided in Table 1, alongside the broad- and narrow-sense heritability estimated across families.

The broad-sense heritability of the number of pods and nodes was slightly higher than the broad-sense heritability of yield, however, the narrow-sense heritability of yield was almost twice as large and the number of nodes, and almost three times higher than the narrow-sense heritability of the number of pods. The narrow-sense heritability estimated from the 18 markers found associated with yield components recovered almost entirely the narrow-sense heritability of the number of pods, but just a third of the heritability of the number of nodes and grain yield. And, surprisingly, the narrow-sense heritability of the ratio of pods per node was higher when only the significant markers were used.

Association analysis

The genome-wide screening for segments associated to yield components is presented in Figure 2. Regions associated with the number of pods were located in chromosomes 3, 5, 14 and 19; significant associations for node number were observed in chromosomes 2, 3, 5, 6, 14, 18 and 19; and regions associated with pods per node were detected in chromosomes 3, 7, 12 and 19. The summary of the associated regions is presented in Table 2, alongside the impact of each significant marker on the yield components, grain yield and days to maturity. With the exception of the association between the marker Gm02_6396340 and the number of nodes, our study did not find any other consensus QTL detected by all three association methods for any of the yield components. All three yield components had significant associations in chromosomes 3 and 19, and the marker Gm19_1587494 was associated with all three traits. From the associated markers, Gm13_14346156 had the highest impact on grain yield, potentially increasing yield as much as 0.6 bushels per acre.

Polygenic architecture

The proportion of variance explained by additivity and epistasis for individual families is presented in Figure 3. The additive fraction of the genetic variance computed using G2A kernels is comparable to the narrow-sense heritability estimated across families (Table 1). All three yield components presented similar average polygenic architecture, having the additive and epistatic components ranging from 0 to approximately 50%, but the estimates where highly variable from family

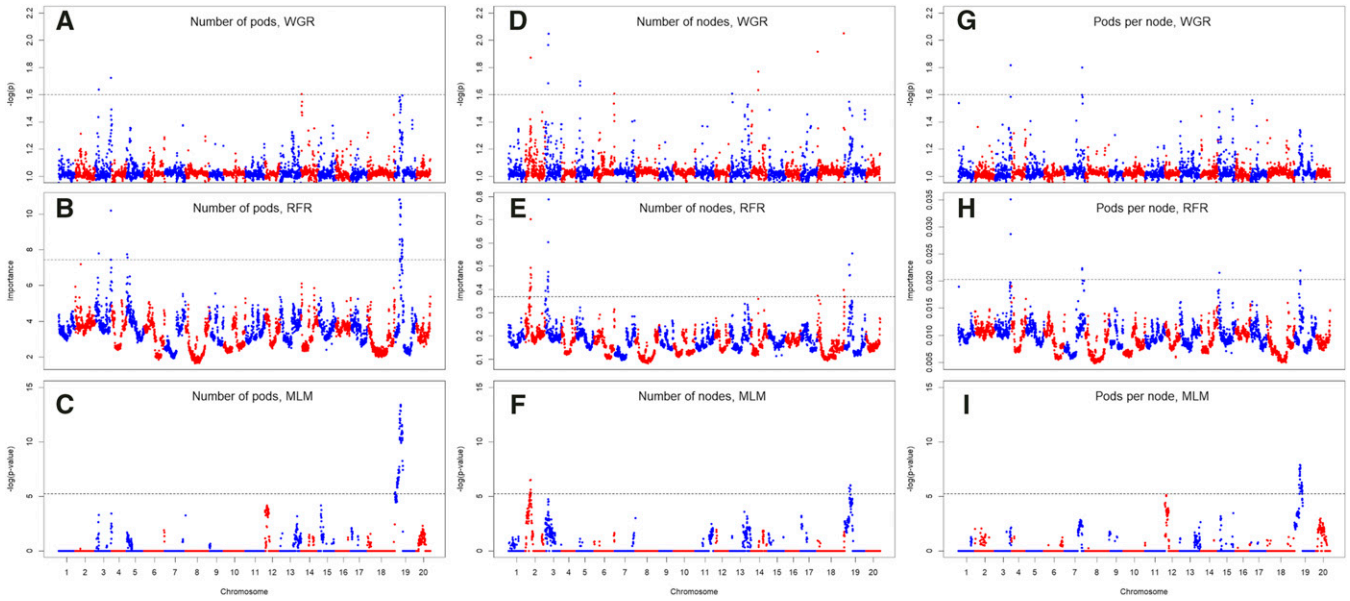

Figure 2 Genome-wide association studies of pod number (A,B,C), node number (D,E,F) and pods per node (G,H,I), performed through three methodologies: WGR whole-genome regression (A,D,G), RFR random forest regression (B,E,H), and MLM mixed linear model (C,F,I). RFR significance is defined by permutation threshold; MLM significance is adjusted for multiple testing with Bonferroni threshold; WGR does not require adjustment for multiple testing.

Table 2 Summary of association studies: SNP at the peak of each QTL; corresponding trait and method from which the QTL was identified, and the least squared effect of the SNP for each yield components, yield and days to maturity. Negative values indicate the desirable allele is inherited from founder parents

SNP	GWAS (Figure 2)	Number of pods	Number of node	Ratio pods ped node	Yield (bu/ac)	Days to Maturity
Gm02_6396340	B,D,E,F	−0.26	−0.41	−0.03	−0.47	−0.16
Gm03_2182974	A,D,E	−0.25	−0.38	−0.03	−0.18	−0.33
Gm03_46533591	A,B,G,H	−0.32	−0.12	−0.34	0.07	0.02
Gm05_914933	B	−0.20	−0.17	−0.10	0.15	0.02
Gm05_3661638	B,C	0.13	0.25	−0.05	−0.23	0.30
Gm06_47199506	D	0.10	0.21	−0.05	0.22	0.06
Gm07_7868756	G,H	−0.02	0.18	−0.18	0.05	−0.03
Gm12_2838455	I	0.07	0.15	−0.04	−0.09	0.08
Gm13_14346156	D	0.19	0.26	0.05	0.62	0.07
Gm14_743883	A	0.23	0.16	0.18	0.18	0.13
Gm14_917668	B	0.23	0.19	0.15	0.11	0.22
Gm14_2322106	D	0.17	0.27	−0.01	0.40	0.32
Gm15_5446785	H	−0.09	0.14	−0.23	−0.24	−0.02
Gm18_2357823	D,E	−0.11	−0.24	0.05	−0.27	−0.02
Gm18_57370051	D,E	0.23	0.28	0.08	0.14	0.22
Gm19_1496625	B,E,I	−0.43	−0.38	−0.28	−0.34	0.02
Gm19_1587494	B,C,F,H,I	−0.43	−0.33	−0.31	−0.15	0.09
Gm19_1991181	E	−0.36	−0.34	−0.21	−0.15	0.10

to family. The additive component averaged 7.46%, 9.03%, and 6.18%; the epistatic component averaged 7.92%, 7.02%, and 7.77%, and the total genomic heritability (additive + epistatic components) averaged 15.38%, 16.05%, and 13.95% for the number of pods, nodes, and pods per node, respectively. Many families provided near-zero genetic control for yield components, in agreement with the low within-family predictive ability (Figure 4).

Prediction analysis

The outcome of the prediction analysis is presented in Figure 4. Predictions within-family provided lower correlations than leave-family-out, and across-family predictions yielded the most predictive scenario. All three yield components had similar heritabilities (Table 1) and, consequently, similar prediction accuracies. For the different cross-validation scenarios, correlations around 0.05, 0.08 and 0.21 were observed for predictions within-family, leave-family-out, and across-families, respectively. BayesB provided a slightly higher predictive ability than GBLUP across cross-validation scenarios, providing an increase in predictability of as much as 0.02. However, the differences in predictive ability were negligible, in agreement with previous results (Xavier *et al.* 2016). The slightly advantageous performance of BayesB suggests that some QTL contribute to the prediction of yield components, but a polygenic model captures most of the genomic signal.

Genetic correlations and indirect selection

The within-family genetic and additive genetic correlations between yield components and yield, as well as yield components stability, are presented in Figure 5. Whereas the average correlations between yield components and yield are relatively small (Figure 5A), there is a large variation from family to family, which indicates that some families could benefit from the selection of yield components. From the three yield components, the number of pods was the only trait with the efficiency of indirect selection that departed from zero (data not presented), so we broke down the efficiency of indirect selection based on pod counts by the genetic background of the SoyNAM founder (Figure 5C). Families with non-elite genetic backgrounds are more likely to benefit, and the indirect selection based on pods was more effective than on yield itself in 10 families ($E > 1$).

DISCUSSION

The dissection of yield components using multiple quantitative genetic approaches using genomic information provides an insight on how such traits can be utilized for breeding purposes. For that, we performed a wide range on analysis, including checking the heritability in broad- and narrow-sense, whether there were major genes involved, whether these genes are captured by different approaches of association analysis, whether the genetic control is influenced by epistatic factors, the trait stability across years, and genomic predictive ability in different settings, different models, within and across families. The collective interpretation of these analysis contributes to the construct the big picture of the genetic architecture of these traits.

Brief overview of the architecture

The number of pods and nodes, as the ratio pods per node, are key yield components in soybean (Herbert and Litchfield 1982) reported to be yield drivers (Kahlon and Board 2012, Suhre *et al.* 2014). Understanding how such traits work may provide insight into better strategies to increase yield and yield stability (Xavier *et al.* 2017a). In soybeans, we found that these yield components have low heritability, both in the broad and narrow sense, and have partially oligogenic architecture, where the genomic control is jointly explained by a set of QTL and polygenic terms (Figures 2 and 3). In addition, within-family analysis indicates that some populations display more epistatic than additive control under the polygenic model (Figure 3), whereas other families presented no genetic control whatsoever.

QTL

Successful mapping of markers associated with complex traits relies on the size and variability of the mapping population. Our study was conducted on the SoyNAM, a large population designed to optimize power and resolution. Yet, only a small number of QTL were detected. Previous mapping studies on yield components have relied on non-experimental panels with a highly diverse genetic background. The studies of Hao *et al.* (2012), Hu *et al.* (2014), Zhang *et al.* (2015) and Fang *et al.* (2017) assessed 191, 113, 219 and 809 genotypes, respectively, including landraces and wild accessions. Among the studies on diverse backgrounds, Fang *et al.* (2017) found a QTL for pod and node

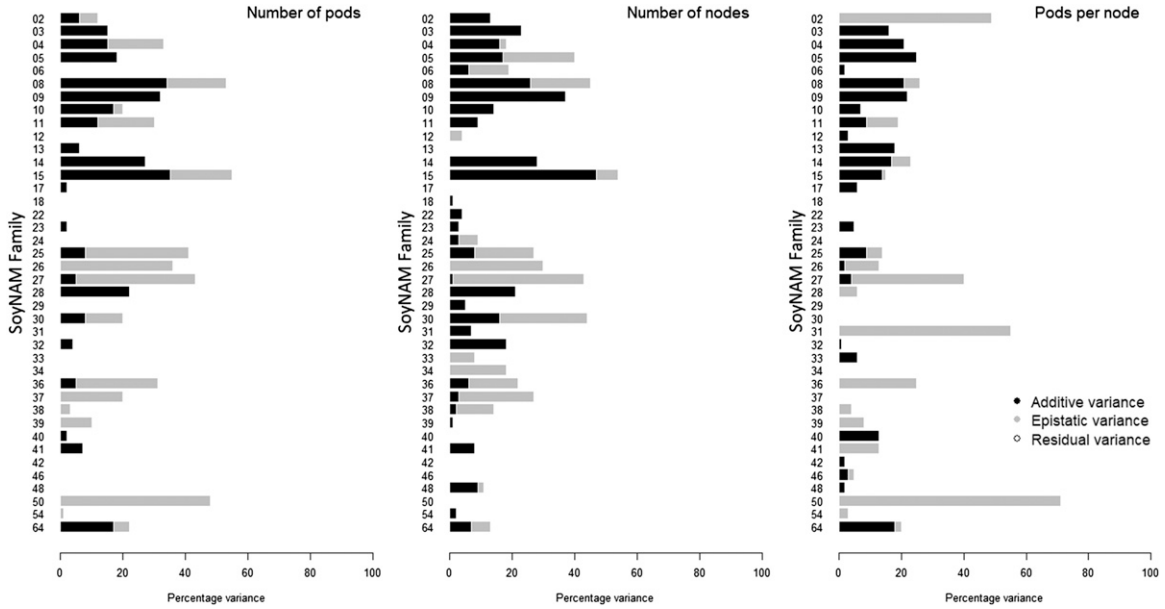

Figure 3 Barplot of the proportion of variance explained by different genetic components of pod number (left), node number (center) and pods per node (right) by family. Additive (black), epistatic (gray) and residual (white) variances. Families had elite (2-23), diverse (2-39) and exotic (40-64) genetic background.

numbers in close proximity to our QTL peak on chromosome 06, marker Gm06_47199506. The pod number QTL detected by Hu *et al.* (2014) were located in chromosomes 3, 5 and 6, in overlapping regions to signals Gm03_2382974, Gm03_46533591, Gm05_914933, Gm05_3661638, Gm06_47199506, and Gm07_7868756.

The significant markers found from this study do not overlap with the signals found for grain yield (Diers *et al.* 2018) and yield stability (Xavier *et al.* 2018) in the SoyNAM population. However, markers Gm02_639640, Gm07_7868756 and Gm12_2838455 are in close proximity to seed size QTL reported by Diers *et al.* (2018). Two markers, Gm19_1587494 and Gm18_57370051, were found to be associated with important traits from previous studies. The marker Gm19_1587494 was also found to be the key association to canopy coverage (Xavier *et al.* 2017b), which means that canopy coverage could be associated with the three yield components. The marker Gm18_57370051 is linked to the stem termination gene Dt2 (Bernard *et al.* 1972), which has been previously detecting in NAM families by Ping *et al.* (2014). In previous studies, Hao *et al.* (2011) and Fang *et al.* (2017) found that Dt2 is an influential gene on the number of pods and nodes. The Dt2 gene is also believed to have played a role in the soybean domestication (Sedivy *et al.* 2017).

The markers that were found to be associated to yield components in this study had little to no impact in maturity, which can be a major limiting factor to their use in breeding as most QTL that improve yield often increase the number of days to maturity (Table 2). However, the QTL peaks also had a limited impact on grain yield across family, with effects ranging from -0.46 to 0.62 bu/ac.

It is important to point out that Table 2 presents an average effect of allele substitution for simplicity. However, two association methods deployed in this study do not directly estimate the allele effects: The MLM utilized in this study computes the significance from within-family effects, hence capturing signal in different linkage phases between marker and QTL. The RFR also does not necessarily provide an allele effect, instead it computes recursive decision trees that would capture QTL with additive, dominant or epistatic effect. Therefore, the intend of this study was mostly focused on tracking which markers are likely associated to the yield components rather than inferring from which parent the desirable alleles are inherited from.

Genomic selection

Markers are informative in two levels for genomic predictions: they can inform the relationship and detect markers linked to, or under linkage disequilibrium with, the quantitative trait loci (Habier *et al.* 2007). Within-family predictive ability solely relies on the linkage disequilibrium (LD) between markers and QTL, as the relationship among individuals is constant. The predictions of families not included in the training set (leave-family-out) can yield mixed results since the training set often holds families with shared ancestry. Of course, the controlled ancestry is a key property of NAM populations since all families share a common parent and, therefore, the outcome predictive ability is higher than the non-experimental population where neither parent has offspring in the calibration set. Predictions performed across family are presumably the most likely to be accurate, as they capture relationships among families and disequilibrium between markers and QTL.

Figure 4 depicts well the expected predictive ability, as within-family predictions hold a high degree of uncertainty, with correlations averaging from 0.064 across yield components, followed by leave-family-out predictions, with an average correlation of 0.092, and the most predictive was across-family predictions, with average correlations above 0.224. Predictive abilities computed from leave-family-out and within-family can be penalized from the fact that some families presented had near-zero heritability and hence no variation for yield components. Within-family predictions may be further penalized due to the small population size to calibrate the genomic models. However, across family predictions are relatively more accurate as those capture both relationship and LD information, and the lower dispersion of the predictions can be attributed to the fact that the prediction model is large, containing a large number of full- and half-siblings. Results from the enhancement in predictive ability due to the joint availability of LD and relationship information have been previously presented from a

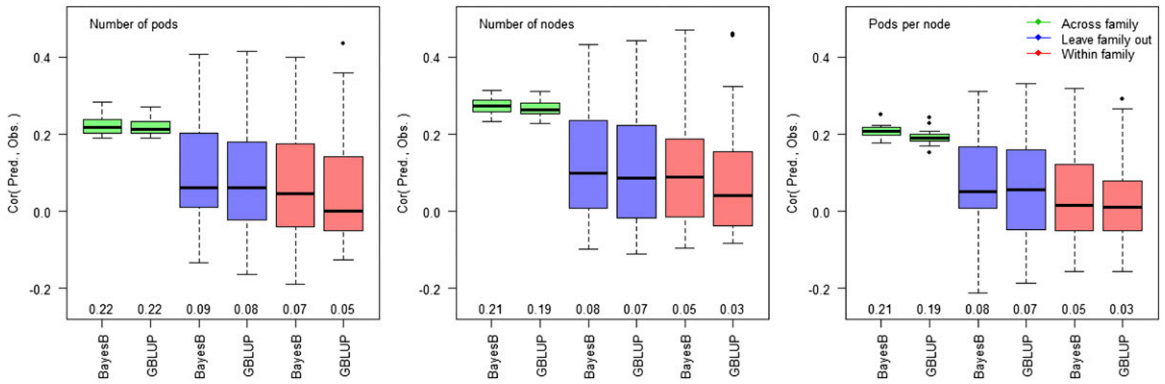

Figure 4 Boxplot of predictive ability of pod number (left), node number (center) and pods per node (right), where two prediction models (BayesB and GBLUP) tested three cross-validations strategies: across-family (green), leave-family-out (blue) and within-family (red). The three cross-validations schemes provide an insight on across-family selection (across family), prediction and selection of individuals from unobserved family (leave family out), and within family selection that capture only QTL segregating in the family under evaluation (within family).

theoretical standpoint by Habier *et al.* (2013) and Schopp *et al.* (2017), and similar results in real data were reported by Ogut *et al.* (2015) in the maize NAM population. In hybrid maize study, Lehermeier *et al.* (2014) claimed 375 half-siblings to provide the same predictive ability of 50 full-siblings, but emphasized the degree of relatedness among families would also play a key role in the predictive ability.

Prediction accuracies estimated across families can also have a misleading interpretation, as these are subject to the Simpson paradox (Chipman and Braun 2017), where the model is able to detect large differences across families, but the predicted families may display negative correlation within-family. Such limitations could be addressed if the cross-validations across-family were performed sampling 20% of individuals from each family and training with the remaining set comprising all families, then estimating the average within-family predictions. However, across-family validations have two advantages: (1) these indicate the predictive potential of selections performed across populations and (2) provide results that can be more easily compared to other literature reports, as most studies performed cross-validations disregarding within-population studies.

The difference in predictive ability between GBLUP and BayesB, which translated into an average improvement of 0.02 going from GBLUP to BayesB, is due the larger flexibility the BayesB model, which is more likely to capture large effects and perform variable selection (Meuwissen *et al.* 2001, Habier *et al.* 2011, Pérez and de los Campos 2014, Xavier *et al.* 2016). Having a comparison between GBLUP and BayesB can provide an insight into the genetic architecture of the trait under evaluation (Daetwyler *et al.* 2013). In this study, we expected BayesB to outperform GBLUP since we uncover a partially oligogenic architecture from the association analysis, but a key piece of information that the genomic prediction analysis provides is discrepancy between GBLUP and BayesB, which inform the degree to which the genetic architecture of the traits under evaluation depart from a polygenic architecture.

Note that the advantage provided by changing the model from GBLUP to BayesB is nowhere comparable to the difference in predictability between cross-validation methods (*i.e.*, within-family, leave-family-out, and across-family). The reason why this phenomenon occurs is that different methods may improve how well the model detects the genetic architecture, but different types of cross-validation provide different information. Thus, gains associated to the choice of a prior are often considered negligible in comparison to increases in population size, better experimental practices, or more representative calibration sets (de los Campos *et al.* 2013, Xavier *et al.* 2016). A possible way of capturing more information for genomic prediction is the explicit modeling of other sources of genetic information, such as dominance and epistasis (Xu 2013). As presented in this study, yield components in some populations have a greater influence of epistasis than the additive background and, on average, the within-family variance decomposition indicates that additive genetics explains as much of the yield components phenotypes as epistasis (Figure 3).

Stability and plasticity

When assessing genotype-by-environment, the total genetic correlation was larger the additive genetic correlation for all yield components (Figure 5B), approximately ranging from 0 to 0.4, whereas the additive genetic correlations ranged from 0 to 0.25. The discrepancy between genetic and additive genetic correlations is attributed to the genetic control due to QTL and non-additive polygenic genetic background.

For the families with near-zero genotype-by-environment correlation, performing selections with a single year of data may not reflect into observable genetic gains in the coming years, and that collecting data from more environments may not necessarily increase the predictive ability of the yield components. Particularly for yield components, low genotype-by-environment correlations is not necessarily bad since the soybean yield plasticity relies on reallocating resources among yield components, which serves as a physiological response to mitigate yield losses under stress (Board and Tan 1995, Board *et al.* 1997, Pedersen and Lauer 2004, Zhang *et al.* 2004). Whereas yield components are mainly responsible for the yield formations, these are not necessarily the best linear yield predictors (Board and Modali 2005). For example, Board and Harville (1993) showed that the number of pods serves as the mechanism by which seed production increases in response to greater light interception.

Our previous study (Xavier *et al.* 2017a) assessed the association among soybean agronomic traits and yield components in the Soy-NAM population-based on undirected graphical models. The graphical models depicted genetic and environmental interdependence among yield components. That means that interactions among yield components occur due to genetic forces as well as a response to environmental stimuli and agronomic practices. Such a phenomenon is also described in a summary of agronomic studies on soybean yield components authored by Board and Kahlon (2011). The interactions among yield components play a key role in the redistribution of resources and yield stability (Ball *et al.* 2000). It is possible that breeding any given yield

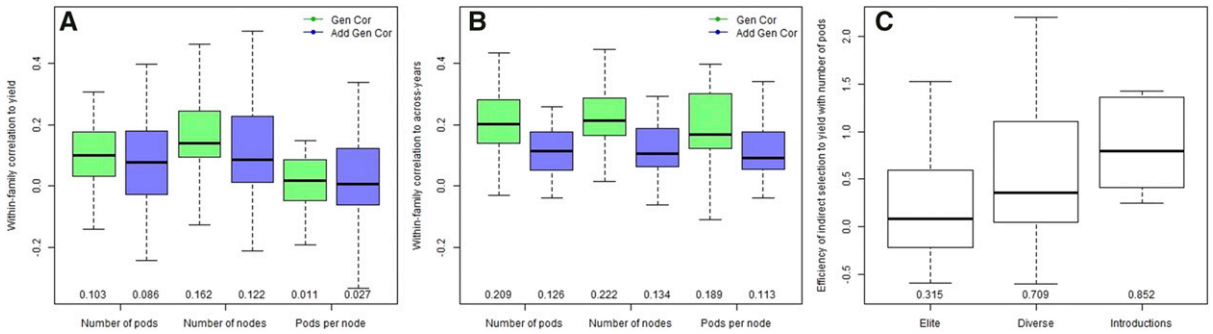

Figure 5 Pleiotropy yield and yield component (A), genotype-by-environment correlation (B) and efficiency of indirect selection (C). Boxplot displaying the dispersion of within-family genetic and additive-genetic correlation between yield components and grain yield (A); the within-family genetic and additive genetic genotype-by-environment correlation (B), where more means more stable across years; the efficiency of indirect selection to yield using pods, breakdown by germplasm background (C).

component toward extreme values may result in a compromised ability of soybeans to compensate yield under stress (Malausa *et al.* 2005).

Yield increases

From the standpoint of trait decomposition, the breaking down of grain yield into pods and nodes does not seem to be an effective approach since there is no strong evidence that these yield components are more heritable than yield (Table 1) or strong genetic correlation to yield (Figure 5A) that would justify the selection based on yield components. With the exception of a few families, yield components are not good proxies for grain yield (Figure 5C). It is possible that the genetic architecture of the yield components under evaluation is just as complex as grain yield itself, not justifying predicting yield components instead of yield *per se*.

In our previous genomic prediction study (Xavier *et al.* 2016), we assessed how a variety of different genomic prediction models would predict the agronomic traits and yield components under the following scenario: within year and across-population. Even though that study did not provide in-depth insight into the genetic architecture of yield components, it was found that genomic prediction models that can jointly account for large effect QTL and epistasis were advantageous over simpler prediction approaches. That study also found that predicting yield is easier than predicting yield components. Those results were further confirmed by the current study, where we assessed the architecture of yield components with more data and under different approaches.

Phenotyping

A major challenge of working with yield components is the data collection as the counting is highly subjective to human error, lowering the trait heritability and affecting the signal detection in downstream analysis. As deep learning methods for computer vision become increasingly population for phenotyping morphological traits (Singh *et al.* 2018), the current limitations with data collection could be addressed by an automated high-throughput phenotyping instead of human counts, that would likely increase both accuracy and scalability of the process. A recent study by Zhang *et al.* (2019) provides a procedure using computer vision for counting soybean pod under experimental settings that would address the phenotypic limitation of this study. Similarly, Uzal *et al.* (2018) and Li *et al.* (2019) recently proposed an imagery system for counting seeds directly from images of soybean pods, yet another yield component limited by the challenging phenotyping. Technologies that enable better, faster and cheaper data collection remain a key limiting factor for the research in yield components.

ACKNOWLEDGMENTS

We thank the SoyNAM collaborators for their contributions to the experiment. William Beavis for experimental design, Qijan Song and Perry Cregan for genotyping, and Jim Specht and Brian Diers for creating the germplasm resource. We thank Chris Hoagland and Curtis Brackett for managing the experiments and contributed to collect the phenotypes. We thank the United Soybean Board for funding the field experiment in 2013 and Dow AgroSciences (now Corteva Agrisciences) for funding the data collection in 2013 and 2014, and the field experiment in 2015.

LITERATURE CITED

Bates, D., M. Mächler, B. Bolker, and S. Walker, (2014). Fitting linear mixed-effects models using lme4. arXiv preprint arXiv:1406.5823.

Bernard, R. L., 1972 Two Genes Affecting Stem Termination in Soybeans 1. Crop Sci. 12: 235–239. https://doi.org/10.2135/cropsci1972.0011183X001200020028x

Board, J. E., and C. S. Kahlon, 2011 *Soybean Yield Formation: What controls it and how it can be improved. Soybean Physiology and Biochemistry*, INTECH Open Access Publisher, Rijeka, Croatia.

Board, J. E., and H. Modali, 2005 Dry matter accumulation predictors for optimal yield in soybean. Crop Sci. 45: 1790–1799. https://doi.org/10.2135/cropsci2004.0602

Board, J. E., M. S. Kang, and B. G. Harville, 1997 Path analyses identify indirect selection criteria for yield of late-planted soybean. Crop Sci. 37: 879–884. https://doi.org/10.2135/cropsci1997.0011183X003700030030x

Board, J. E., and Q. Tan, 1995 Assimilatory capacity effects on soybean yield components and pod number. Crop 35: 846–851. https://doi.org/10.2135/cropsci1995.0011183X003500030035x

Board, J. E., and B. G. Harville, 1993 Soybean yield component responses to a light interception gradient during the reproductive period. Crop Sci. 33: 772–777. https://doi.org/10.2135/cropsci1993.0011183X003300040028x

Botta, V., G. Louppe, P. Geurts, and L. Wehenkel, 2014 Exploiting SNP correlations within random forest for genome-wide association studies. PLoS One 9: e93379. https://doi.org/10.1371/journal.pone.0093379

Carter, T. E., Nelson, R. L., Sneller, C. H., Cui, Z., Boerma, H. R., & Specht, J. E. (2004). Genetic diversity in soybean. Soybeans: Improvement, production, and uses.

Chipman, J., and D. Braun, 2017 Simpson's paradox in the integrated discrimination improvement. Stat. Med. 36: 4468–4481. https://doi.org/10.1002/sim.6862

Daetwyler, H. D., M. P. Calus, R. Pong-Wong, G. de los Campos, and J. M. Hickey, 2013 Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. Genetics 193: 347–365. https://doi.org/10.1534/genetics.112.147983

de los Campos, G., J. M. Hickey, R. Pong-Wong, H. D. Daetwyler, and M. P. Calus, 2013 Whole-genome regression and prediction methods applied

to plant and animal breeding. Genetics 193: 327–345. https://doi.org/10.1534/genetics.112.143313

Diers, B. W., J. Specht, K. M. Rainey, P. Cregan, Q. Song *et al.*, 2018 Genetic Architecture of Soybean Yield and Agronomic Traits. G3: Genes, Genomes. Genetics 8: 3367–3375.

Doerge, R. W., and G. A. Churchill, 1996 Permutation tests for multiple loci affecting a quantitative character. Genetics 142: 285–294.

Egli, D. B., and W. P. Bruening, 2006 Temporal profiles of pod production and pod set in soybean. Eur. J. Agron. 24: 11–18. https://doi.org/10.1016/j.eja.2005.04.006

Falconer, D. S., and T. F. C. Mackay, 1996 *Introduction to quantitative genetics*, Longman Group, Essex, UK.

Fang, C., Y. Ma, S. Wu, Z. Liu, Z. Wang *et al.*, 2017 Genome-wide association studies dissect the genetic networks underlying agronomical traits in soybean. Genome Biol. 18: 161. https://doi.org/10.1186/s13059-017-1289-9

Fehr, W. R., C. E. Caviness, D. T. Burmood, and J. S. Pennington, 1971 Stage of development descriptions for soybeans, Glycine Max (L.) Merrill 1. Crop Sci. 11: 929–931. https://doi.org/10.2135/cropsci1971.0011183X001100060051x

Garrick, D. J., J. F. Taylor, and R. L. Fernando, 2009 Deregressing estimated breeding values and weighting information for genomic regression analyses. Genet. Sel. Evol. 41: 55. https://doi.org/10.1186/1297-9686-41-55

Habier, D., R. L. Fernando, and D. J. Garrick, 2013 Genomic BLUP decoded: a look into the black box of genomic prediction. Genetics 194: 597–607. https://doi.org/10.1534/genetics.113.152207

Habier, D., R. L. Fernando, K. Kizilkaya, and D. J. Garrick, 2011 Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics 12: 186. https://doi.org/10.1186/1471-2105-12-186

Habier, D., R. L. Fernando, and J. C. Dekkers, 2007 The impact of genetic relationship information on genome-assisted breeding values. Genetics 177: 2389–2397. https://doi.org/10.1534/genetics.107.081190

Hao, D., H. Cheng, Z. Yin, S. Cui, D. Zhang *et al.*, 2012 Identification of single nucleotide polymorphisms and haplotypes associated with yield and yield components in soybean (Glycine max) landraces across multiple environments. Theor. Appl. Genet. 124: 447–458. https://doi.org/10.1007/s00122-011-1719-0

Herbert, S. J., and G. V. Litchfield, 1982 Partitioning Soybean Seed Yield Components 1. Crop Sci. 22: 1074–1079. https://doi.org/10.2135/cropsci1982.0011183X002200050044x

Hu, Z., D. Zhang, G. Zhang, G. Kan, D. Hong *et al.*, 2014 Association mapping of yield-related traits and SSR markers in wild soybean (Glycine soja Sieb. and Zucc.). Breed. Sci. 63: 441–449. https://doi.org/10.1270/jsbbs.63.441

Kahlon, C. S., J. E. Board, and M. S. Kang, 2011 An analysis of yield component changes for new *vs.* old soybean cultivars. Agron. J. 103: 13–22. https://doi.org/10.2134/agronj2010.0300

Kahlon, C. S., and J. E. Board, 2012 Growth dynamic factors explaining yield improvement in new *vs.* old soybean cultivars. J. Crop Improv. 26: 282–299. https://doi.org/10.1080/15427528.2011.637155

Kaler, A. S., J. D. Ray, W. T. Schapaugh, M. K. Davies, C. A. King *et al.*, 2018 Association mapping identifies loci for canopy coverage in diverse soybean genotypes. Mol. Breed. 38: 50. https://doi.org/10.1007/s11032-018-0810-5

Lado, B., I. Matus, A. Rodríguez, L. Inostroza, J. Poland *et al.*, 2013 Increased genomic prediction accuracy in wheat breeding through spatial adjustment of field trial data. G3: Genes, Genomes. Genetics 3: 2105–2114.

Langewisch, T., H. Zhang, R. Vincent, T. Joshi, D. Xu *et al.*, 2014 Major soybean maturity gene haplotypes revealed by SNPViz analysis of 72 sequenced soybean genomes. PLoS One 9: e94150. https://doi.org/10.1371/journal.pone.0094150

Lehermeier, C., N. Krämer, E. Bauer, C. Bauland, C. Camisan *et al.*, 2014 Usefulness of multiparental populations of maize (Zea mays L.) for genome-based prediction. Genetics 198: 3–16. https://doi.org/10.1534/genetics.114.161943

Li, Y., J. Jia, L. Zhang, A. M. Khattak, S. Sun *et al.*, 2019 Soybean Seed Counting Based on Pod Image Using Two-Column Convolution Neural Network. IEEE Access 7: 64177–64185. https://doi.org/10.1109/ACCESS.2019.2916931

Malausa, T., T. Guillemaud, and L. Lapchin, 2005 Combining genetic variation and phenotypic plasticity in tradeoff modelling. Oikos 110: 330–338. https://doi.org/10.1111/j.0030-1299.2005.13563.x

Mikel, M. A., B. W. Diers, R. L. Nelson, and H. H. Smith, 2010 Genetic diversity and agronomic improvement of North American soybean germplasm. Crop Sci. 50: 1219–1229. https://doi.org/10.2135/cropsci2009.08.0456

Misztal, I., 2008 Reliable computing in estimation of variance components. J. Anim. Breed. Genet. 125: 363–370. https://doi.org/10.1111/j.1439-0388.2008.00774.x

Meuwissen, T., B. Hayes, and M. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819–1829.

Nico, M., D. J. Miralles, and A. G. Kantolic, 2019 Natural post-flowering photoperiod and photoperiod sensitivity: Roles in yield-determining processes in soybean. Field Crops Res. 231: 141–152. https://doi.org/10.1016/j.fcr.2018.10.019

Ogut, F., Y. Bian, P. J. Bradbury, and J. B. Holland, 2015 Joint-multiple family linkage analysis predicts within-family variation better than single-family analysis of the maize nested association mapping population. Heredity 114: 552–563. https://doi.org/10.1038/hdy.2014.123

Pedersen, P., and J. G. Lauer, 2004 Response of soybean yield components to management system and planting date. Agron. J. 96: 1372–1381. https://doi.org/10.2134/agronj2004.1372

Pérez, P., and G. de Los Campos, 2014 Genome-wide regression and prediction with the BGLR statistical package. Genetics 198: 483–495. https://doi.org/10.1534/genetics.114.164442

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

Rincker, K., R. Nelson, J. Specht, D. Sleper, T. Cary *et al.*, 2014 Genetic improvement of US soybean in maturity groups II, III, and IV. Crop Sci. 54: 1419–1432.

Robinson, A. P., S. P. Conley, J. J. Volenec, and J. B. Santini, 2009 Analysis of high yielding, early-planted soybean in Indiana. Agron. J. 101: 131–139. https://doi.org/10.2134/agronj2008.0014x

Schopp, P., D. Müller, Y. C. Wientjes, and A. E. Melchinger, 2017 Genomic prediction within and across biparental families: means and variances of prediction accuracy and usefulness of deterministic equations. G3: Genes, Genomes. Genetics 7: 3571–3586.

Sedivy, E. J., F. Wu, and Y. Hanzawa, 2017 Soybean domestication: the origin, genetic architecture and molecular bases. New Phytol. 214: 539–553. https://doi.org/10.1111/nph.14418

Singh, A. K., B. Ganapathysubramanian, S. Sarkar, and A. Singh, 2018 Deep learning for plant stress phenotyping: trends and future perspectives. Trends Plant Sci. 23: 883–898. https://doi.org/10.1016/j.tplants.2018.07.004

Specht, J. E., Diers, B. W., Nelson, R. L., Francisco, J., de Toledo, F., Torrion, J. A., & Grassini, P. (2014). Soybean. Yield gains in major US field crops, 311–356.

Specht, J. E., D. J. Hume, and S. V. Kumudini, 1999 Soybean yield potential: a genetic and physiological perspective. Crop Sci. 39: 1560–1570. https://doi.org/10.2135/cropsci1999.3961560x

Song, Q., L. Yan, C. Quigley, B. D. Jordan, E. Fickus *et al.*, 2017 Genetic characterization of the soybean nested association mapping population. Plant Genome 10 https://doi.org/10.3835/plantgenome2016.10.0109

Suhre, J. J., N. H. Weidenbenner, S. C. Rowntree, E. W. Wilson, S. L. Naeve *et al.*, 2014 Soybean yield partitioning changes revealed by genetic gain and seeding rate interactions. Agron. J. 106: 1631–1642. https://doi.org/10.2134/agronj14.0003

Qiu, L. J., L. L. Xing, Y. Guo, J. Wang, S. A. Jackson *et al.*, 2013 A platform for soybean molecular breeding: the utilization of core collections for food security. Plant Mol. Biol. 83: 41–50. https://doi.org/10.1007/s11103-013-0076-6

Uzal, L. C., G. L. Grinblat, R. Namías, M. G. Larese, J. S. Bianchi *et al.*, 2018 Seed-per-pod estimation for plant breeding using deep learning. Comput. Electron. Agric. 150: 196–204. https://doi.org/10.1016/j.compag.2018.04.024

VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. J. Dairy Sci. 91: 4414–4423. https://doi.org/10.3168/jds.2007-0980

Xavier, A., W. M. Muir, and K. M. Rainey, 2019 bWGR: Bayesian Whole-Genome Regression. Bioinformatics https://doi.org/10.1093/bioinformatics/btz794 13

Xavier, A., D. Jarquin, R. Howard, V. Ramasubramanian, J. E. Specht *et al.*, 2018 Genome-Wide analysis of grain yield stability and environmental interactions in a multiparental soybean population. G3: Genes, Genomes. Genetics 8: 519–529.

Xavier, A., B. Hall, S. Casteel, W. Muir, and K. M. Rainey, 2017a Using unsupervised learning techniques to assess interactions among complex traits in soybeans. Euphytica 213: 200. https://doi.org/10.1007/s10681-017-1975-4

Xavier, A., B. Hall, A. A. Hearst, K. A. Cherkauer, and K. M. Rainey, 2017b Genetic architecture of phenomic-enabled canopy coverage in Glycine max. Genetics 206: 1081–1089. https://doi.org/10.1534/genetics.116.198713

Xavier, A., W. M. Muir, and K. M. Rainey, 2016 Assessing predictive properties of genome-wide selection in soybeans. G3: Genes, Genomes. Genetics 6: 2611–2616.

Xavier, A., S. Xu, W. M. Muir, and K. M. Rainey, 2015 NAM: association studies in multiple populations. Bioinformatics 31: 3862–3864.

Xu, S., 2013 Mapping quantitative trait loci by controlling polygenic background effects. Genetics 195: 1209–1222. https://doi.org/10.1534/genetics.113.157032

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/10.1534/genetics.110.125062

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

Wright, M. N., and A. Ziegler, (2015). Ranger: a fast implementation of random forests for high dimensional data in C++ and R. arXiv preprint arXiv:1508.04409.

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

Yang, J., R. K. Ramamurthy, X. Qi, R. L. Fernando, J. C. Dekkers *et al.*, 2018 Empirical Comparisons of Different Statistical Models To Identify and Validate Kernel Row Number-Associated Variants from Structured Multi-parent Mapping Populations of Maize. G3: Genes, Genomes. Genetics 8: 3567–3575.

Zeng, Z. B., T. Wang, and W. Zou, 2005 Modeling quantitative trait loci and interpretation of models. Genetics 169: 1711–1725. https://doi.org/10.1534/genetics.104.035857

Zhang, H., D. Hao, H. M. Sitoe, Z. Yin, Z. Hu *et al.*, 2015 Genetic dissection of the relationship between plant architecture and yield component traits in soybean (Glycine max) by association analysis across multiple environments. Plant Breed. 134: 564–572. https://doi.org/10.1111/pbr.12305

Zhang, W. K., Y. J. Wang, G. Z. Luo, J. S. Zhang, C. Y. He *et al.*, 2004 QTL mapping of ten agronomic traits on the soybean (Glycine max L. Merr.) genetic map and their association with EST markers. Theor. Appl. Genet. 108: 1131–1139. https://doi.org/10.1007/s00122-003-1527-2

Zhang, Z., and S. Ghosal *et al.*, (2019) A deep vision based approach to real-time detection and counting of soybean pods. Presented at the Machine Learning and Cyber Agriculture Symposium, Iowa State University. 14

Communicating editor: A. Lipka

GGG February (2020 Xavier, Rainey)
Author query sheet Xavier (GGG_400896)

QA1 If you or your coauthors would like to include an ORCID ID in this article, please provide your respective ORCID IDs along with your corrections.

Note: If you do not yet have an ORCID ID and would like one, you may register for this unique digital identifier at https://orcid.org/register.

1 Please supply a mailing address for the corresponding author.

2 Any alternations between capitalization and/or italics in genetic and taxonomic nomenclature have been retained per the original manuscript. *G3* style is for genes and alleles to be italicized; please confirm that all nomenclature has been formatted properly throughout. Uppercase Greek letters should remain roman per journal style even when appearing in a term where the overall style is italic (e.g., a gene name such as *kap108*Δ). Note that headings are set all roman or all italics based on journal style and should not be changed.

3 Please verify styling of Greek and math symbols in text and equations throughout article. Check carefully for correct use of boldface, italics, operators, qualifiers, spacing, superscripts, and subscripts. Journal style includes Greek letters set roman (not italic), math variables set in italic, and variable modifiers set in roman.

4 Please confirm or update any and all URLs in your article.

5 Please check all figure legends carefully to confirm that any and all labels, designators, directionals, colors, etc. are represented accurately in comparison with the figure images.

6 If you or your coauthors provided an ORCID ID, please verify the accuracy of the ID number(s) on the proof as presented. If you or your coauthors would like to include an ORCID ID in this article, please provide your respective ORCID IDs along with your corrections. Note: If you do not yet have an ORCID ID and would like one, you may register for this unique digital identifier at https://orcid.org/register.

7 Please confirm that your Data Availability statement is accurate, or if not, update to include the details of where your data can be found. Details are available in the Materials and Methods instructions at http://www.genetics.org/content/prep-manuscript#text.

8 In-text citation "Rinker *et al.* (2014)" has been changed to "Rincker *et al.* (2014)" to match reference with "Literature Cited" section. Kindly check.

9 In-text citation "Hao *et al.* (2011)" has been changed to "Hao *et al.* (2012)" to match reference with "Literature Cited" section. Kindly check.

10 In-text citation "Perez and de los Campos (2014)" has been changed to "Pérez and de los Campos (2014)" to match reference with "Literature Cited" section. Kindly check.

11 Please add "Ball *et al.* (2000)" to Literature Cited or delete citation from text.

12 Please provide page range for the reference "Song *et al.* (2017)."

13 Please provide page range and volume for the reference "Xavier *et al.* (2019)."

14 Please provide complete details for the reference "Zhang *et al.* (2019)."