




## ORIGINAL ARTICLE

## Crop Breeding &amp; Genetics

# An assessment of the interaction between sucrose content and seed quality traits in soybeans

Diana M. Escamilla<sup>1</sup>  | Alencar Xavier<sup>1,2</sup>  | Tri D. Vuong<sup>3</sup> | Henry T. Nguyen<sup>3</sup> | Katy Martin Rainey<sup>1</sup> 

<sup>1</sup>Department of Agronomy, Purdue University, West Lafayette, Indiana, USA

<sup>2</sup>Department of Biostatistics, Corteva Agriscience, Johnston, Iowa, USA

<sup>3</sup>Division of Plant Sciences and Technology and National Center for Soybean Biotechnology, University of Missouri, Columbia, Missouri, USA

## Correspondence

Katy M. Rainey, Department of Agronomy, Purdue University, 915 West State Street, West Lafayette, IN 47907, USA.  
Email: [krainey@purdue.edu](mailto:krainey@purdue.edu)

Assigned to Associate Editor William Schapaugh.

## Funding information

United Soybean Board, Grant/Award Number: 1820-152-0101

## Abstract

Soybean (*Glycine max* [L.] Merr.) cultivars with increased protein and sucrose content are desirable to enhance soybean meal quality; however, the complex interactions between soybean traits, mainly trade-offs, make the plant breeders' job challenging because traits cannot simultaneously be enhanced. Information about the interactions of seed carbohydrate composition with other seed traits remains inconclusive and contradictory; thus, this study explored the interactions among soluble carbohydrate content with other seed traits, flowering time, and maturity in a diverse panel of 1096 soybean accessions from the USDA Soybean Germplasm Collection. Trait interactions were explored through phenotypic, additive genetic, and residual correlations. We did not identify traits that can be useful in the indirect selection of sucrose. Additive genetic correlations of sucrose with flowering and maturity time suggest that obtaining cultivars with high sucrose content might be easier in earlier flowering and later maturing cultivars. Sucrose correlations with other seed traits were primarily driven by their additive genetic correlations, which make them more stable, whereas correlations of raffinose with oil and maturity were population and environment specific. The two main obstacles to enhancing soymeal quality are the trade-offs of protein with sucrose and oil.

## 1 | INTRODUCTION

The United States is the second-largest soybean (*Glycine max* [L.] Merr.) producer in the world, with soybean seeds mainly used for soybean meal and oil (Soy Stats, 2020). Of the soybeans grown in the United States, 70% goes to the animal feed industry. Soymeal production in the United States for 2019–2020 was 46 million metric tons (MT), with poultry and swine

as primary consumers (United Soybean Board, 2021); thus, soybean has a significant role in feeding people directly or indirectly, and its production and consumption will increase as the world population increases (Hartman et al., 2011).

Soybean is an excellent source of energy, amino acids, and protein for animal feed formulations (Kerley & Allee, 2003), and is principally composed of protein (40%), oil (20%), and carbohydrates (34%). About 1.6% of carbohydrates are soluble carbohydrates (Burton, 1997; Hymowitz & Collins, 1974; Liu, 1997). Sucrose is the primary soluble sugar, followed by raffinose family oligosaccharides (RFOs) that include raffinose, stachyose, and verbascose (Obendorf & Gorecki, 2012).

**Abbreviations:** BLUP, best linear unbiased prediction; ME, metabolizable energy; MG, maturity group; PCA, principal component analysis; QTL, quantitative trait loci; RFO, raffinose family oligosaccharide; SNP, single-nucleotide polymorphism.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2023 The Authors. *Crop Science* published by Wiley Periodicals LLC on behalf of Crop Science Society of America.

Monogastric animals do not completely digest RFOs when consumed as a component of soybean meal and they provide little metabolizable energy (ME) (Dierking & Bilyeu, 2009b; Kumar et al., 2010); ME is the gross energy consumed minus the gross energy contained in the excreta, representing the energy available for growth and reproduction (Sibbald, 1980). Sucrose, on the other hand, increases soybean meal's ME, sweetness, and flavor (Gupta & Manjaya, 2022); therefore, increasing sucrose levels in soybean seeds could improve the energy density of soybean meal (Graham et al., 2002; Kerley & Allee, 2003).

Developing soybeans with modified seed composition that improves soymeal digestibility and animal growth efficiency in animal production systems is crucial to maintaining and growing the soybean presence in the animal feed market (Kerley & Allee, 2003). Reduction of RFOs was observed with induced mutations in the *RS2* locus, a gene encoding an enzyme involved in the biosynthesis of raffinose and stachyose, and with CRISPR/Cas9 knockouts in two galactinol synthase genes (Dierking & Bilyeu, 2008, 2009a; Le et al., 2020). Increasing sucrose, on the other hand, is a more complex task. Strategies to increase sucrose content include the exploration of diverse soybeans to identify new variants, incremental increase through breeding cycles via selection for increased sucrose as a quantitative trait, selection for increased sucrose using known quantitative trait loci (QTL) and molecular markers associated with sucrose content, and the generation of new variation through mutagenesis. Previous studies have reported over 44 QTL for seed sucrose content and 15 QTL for oligosaccharide content in soybean (<https://www.soybase.org>); however, to our knowledge, the causal genes for these QTL have not been identified or validated by experiments.

Improving the soybean carbohydrate profile without compromising other seed quality traits is important for soybean breeders; however, physiological pathways interconnect most plant traits, where some traits have regulatory effects on other traits (Xavier et al., 2017). From this complex network of interactions, trade-offs are the most challenging for plant breeders because they limit the breeder's ability to effect simultaneous improvement. Trade-offs, defined as the situation when "one trait cannot increase without a decrease in another trait (or vice versa)" (Garland, 2014), help to maintain relative fitness under unpredictable conditions and maximize reproductive success (Dwivedi et al., 2021). Contrarily, positive correlations can be advantageous in soybean breeding, allowing the improvement of primary traits through selecting secondary traits (Bernardo, 2014; Xavier et al., 2017), and facilitating simultaneous trait improvement. Some of the current breeding strategies available for simultaneous trait improvement include the pyramiding of several QTL through marker-assisted selection (Patil et al., 2017), tandem selection, indirect selection, selection indexes (Bernardo, 2014), and multitrait genomic prediction models, which are being

### Core Ideas

- Trade-offs of protein with sucrose and oil hamper efforts to improve soybean seed quality.
- There are no strong correlations between sucrose and other seed traits for indirect selection.
- Sucrose correlations with protein and seed size are primarily driven by their additive genetic correlations.
- The high-sucrose phenotype might be easier to obtain in cultivars with earlier flowering and later maturity.

recently incorporated into animal and plant breeding (Jean et al., 2021; Manzanilla-Pech et al., 2020; Sun et al., 2017).

In soybean seed composition, the most recognized trade-off is the negative association of protein and oil content (Chung et al., 2003; Patil et al., 2018). According to previous studies, sucrose and RFOs have positive correlations with seed yield, oil content, and seed size, while protein content and soybean maturity have negative correlations with sucrose and RFOs (Cicek et al., 2006; Jauregui et al., 2011; Jiang et al., 2018; Wilcox & Shibles, 2001). At the same time, other studies reported nonsignificant and negative correlations between sucrose and oil, a positive association between sucrose and protein, and a nonsignificant association between carbohydrates and yield (Cicek et al., 2006; Kim et al., 2005; Li et al., 2012). In addition, correlations among sucrose, raffinose, and stachyose seem to be positive or nonsignificant (Cicek et al., 2006; Jauregui et al., 2011; Kumar et al., 2010). According to the above complex network of interactions, there is a lack of consensus regarding the relationship between carbohydrate traits and other seed traits, and changes in the carbohydrate profile of soybean seed would likely affect other traits of economic importance for soybeans, such as protein, oil, and seed size.

Most researchers measure trait relationships by phenotypic correlations that, at the same time, are determined by genetic and nongenetic factors; thus, some plant scientists use phenotypic, additive genetic, and nonadditive genetic correlations among traits to explore their interactions (Reich et al., 2003; Xavier et al., 2017). There are estimates of phenotypic and genetic correlations for many traits and populations; however, many of them are rough estimates since there is a need for larger datasets to obtain reliable estimates of the genetic variances and covariances required to describe the relationships among multiple traits (Hill, 2013). Sample sizes of populations used previously to study correlations between carbohydrate contents and other seed traits did not exceed 500 individuals. Obtaining improved estimates of the phenotypic and genetic correlations of carbohydrate contents and other

seed traits is necessary to obtain more strategic crossing and progeny selections and to use genetic resources.

Publicly available datasets, such as the high-density marker data (Song et al., 2013) and passport data of the USDA Soybean Germplasm Collection, as well as the large carbohydrate content dataset previously generated by Qiu et al. (2015), can be employed to support scientific research and improve our understanding of soybean seed trait interactions. Here, we explored the interactions among seed traits, flowering, and maturity using data from a large panel of *G. max* accessions of the USDA Soybean Germplasm Collection. We included flowering and maturity time in this study because they are crucial in soybean productivity and seed quality. Our primary aim was to determine the phenotypic, additive genetic, and residual associations of sucrose, raffinose, and stachyose with flowering, maturity, and other seed traits.

## 2 | MATERIALS AND METHODS

### 2.1 | Plant material

We studied a diverse panel of 1096 accessions from the USDA Soybean Germplasm Collection from maturity groups (MGs) I–VIII. This panel was previously evaluated for sucrose, raffinose, and stachyose content (Qiu et al., 2015). The 1096 accessions are all *G. max* predominantly from MGs IV (47%), III (21%), V (13%), and II (10%) (Figure S1). Most of the accessions originated in Korea (23%), Japan (20%), China (20%), the United States (9%), and Russia (6%). The remaining originated from other countries, such as Brazil, Bulgaria, Germany, India, Indonesia, Israel, Romania, Serbia, Sweden, Taiwan, and Vietnam (Figure S2).

### 2.2 | Genotypic data

Song et al. (2013) genotyped the entire USDA Soybean Germplasm Collection with the Illumina Infinium SoySNP50K iSelect Bead chip that contains 42,509 single-nucleotide polymorphisms (SNPs). Xavier et al. (2018) retrieved genotypic data from the SoyBase website (<https://soybase.org/snps/index.php>) and coded allelic genotype {AA, Aa, aa} as {0,1,2}. Missing genotypic data imputation and removal of redundant SNPs and markers with minor allele frequency lower than 0.15 were carried out using the R package NAM (Xavier et al., 2015).

### 2.3 | Phenotypic data from the USDA soybean germplasm collection

Existing passport data per accession for the percentage of seed oil content (Oil), percentage of seed protein content (Prot),

the weight of 100 seeds, flowering date, and maturity date were initially provided by Randall Nelson and made available by Xavier et al. (2018). Phenotypic data collected by USDA Agricultural Research Service (ARS) germplasm curation staff and their collaborators resulted from field evaluations conducted in various locations, where accessions from one or more MG classes are adapted, and such field trials often span several years (Bandillo et al., 2015). Protein and oil are measured as milligram per gram. Flowering or growth stage R1 is when 50% of the plants have at least one flower (month–day), and maturity or R8 is when 95% of the pods have reached final color (month–day) (Fehr et al., 1971). Seed weight or seed size is the centigrams per seed equivalent to g/100 seeds based on a 100-seed sample (Hill et al., 2005). Further details relative to the methods used to evaluate the USDA soybean germplasm collection are described by Hill et al. (2005). The passport data made available by Xavier et al. (2018) correspond to the mean values across locations and years.

### 2.4 | Carbohydrate content data

Sucrose (Suc), raffinose (Raf), and stachyose (Sta) phenotypic data of the studied panel were accessed from the Nguyen Laboratory, the University of Missouri. Briefly, carbohydrate data were measured from seeds harvested in replicated field experiments at four environments (location–year) (Table S1). There were carbohydrate data for 1245 accessions in the dataset, including 475 and 394 accessions grown at Columbia, MO in 2012 and 2015, respectively, and 43 accessions evaluated in both years. A total of 277 accessions were grown at Fayetteville and Stuttgart, AR 2015 with other 39 accessions grew only at Fayetteville and two at Stuttgart. There was a total of 15 accessions grown across three of the four environments. A randomized complete block design was used to grow accessions in Columbia MO, 2015, and Fayetteville and Stuttgart, AR, 2015. Accessions in Columbia MO, 2012 were grown without replications. Sucrose, raffinose, and stachyose quantification was conducted using the Agilent HPLC detection system and a stablished protocol as previously described (Qiu et al., 2015).

### 2.5 | Data analysis

To estimate genotypic values, the replicated carbohydrate phenotypes were fitted using the restricted maximum likelihood (REML) algorithm, implemented in the R package “lme4” (Bates et al., 2015). We combined location and year into one environment term resulting in four environments. The model implemented was

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{k(j)} + (\alpha \times \beta)_{ij} + e_{ijk}, \quad (1)$$

where  $y_{ijk}$  is the phenotype,  $\mu$  is the mean,  $\alpha_i$  is the random genotype effect with  $\alpha_i \sim N(0, \sigma_\alpha^2)$  where  $\sigma_\alpha^2$  is the genetic variance,  $\beta_j$  ( $j = 1, \dots$ , number of environments) is the random effect of environments with  $\beta_j \sim N(0, \sigma_\beta^2)$  where  $\sigma_\beta^2$  is the environmental variance,  $\gamma_{k(j)}$  is the random effect of  $k$ th block nested within the  $j$ th environment with  $\gamma_{k(j)} \sim N(0, \sigma_\gamma^2)$  where  $\sigma_\gamma^2$  is the variance due to the blocks,  $(\alpha \times \beta)_{ij}$  is the random genotype  $\times$  environment interaction effect with  $(\alpha \times \beta)_{ij} \sim N(0, \sigma_{\alpha \times \beta}^2)$  where  $\sigma_{\alpha \times \beta}^2$  is the genotype  $\times$  environment variance, and  $e_{ijk}$  is the residual term with  $e_{ijk} \sim N(0, \sigma_e^2)$  where  $\sigma_e^2$  is the residual variance. The solutions for  $\alpha_i$  for each trait here are defined as best linear unbiased predictions (BLUPs). BLUPs were adjusted as  $\mu + \alpha_i$  for sucrose, raffinose, and stachyose content, to express them in the units of the phenotypes. Residuals of sucrose, raffinose, and stachyose models from Equation (1) were normally distributed. The adjusted BLUPs for sucrose, raffinose, and stachyose were merged with passport data, which corresponds to the mean value per accession across environments and years. Of the 1245 accessions with carbohydrate data, only 1096 had passport data, so only those 1096 were included in the final dataset (Table S2) used for estimating trait's correlations.

## 2.6 | Phenotypic correlations

From the final dataset containing flowering and maturity time, seed weight, sucrose, raffinose, stachyose, protein, and oil content, we estimated pairwise Pearson and Spearman correlations. Pearson's correlation coefficient measures the strength of the linear relationship between two variables. In contrast, Spearman's rank correlation evaluates the monotonic function between variables (Kossowski & Hauke, 2011; Xavier et al., 2017). We transformed the traits into standard normal random variables to estimate Pearson and Spearman phenotypic correlations and computed them using built-in functions in R (R Core Team, 2018).

## 2.7 | Relatedness and population structure

We built the genomic relationship ( $\mathbf{G}$ ) matrix as described by VanRaden (2008) using the R package NAM (Xavier et al., 2015).

$$\mathbf{G} = \frac{(\mathbf{M} - \mathbf{P})(\mathbf{M} - \mathbf{P})'}{2 \sum_j p_j (1 - p_j)}, \quad (2)$$

where  $\mathbf{M}$  is a matrix with the marker allele's information, whose dimensions are the number of individuals ( $n$ ) by the number of loci ( $m$ ).  $\mathbf{P}$  is a matrix containing the mean allele frequencies. The denominator is a normalizing fac-

tor computed as the sum of the loci variances based on allele frequencies (VanRaden, 2008; Xavier et al., 2017). In addition, we performed a cluster analysis to identify the underlying population structure of the 1096 accessions. We estimated dissimilarities among genotypes using the Euclidean distance metric as implemented in the R package NAM (Xavier et al., 2015). Euclidean distance between individuals  $d_{(a,b)}$  is defined as

$$d_{(a,b)} = \sqrt{\sum_{j=1}^m (a_j - b_j)^2}, \quad (3)$$

where  $a_j$  and  $b_j$  are allele scores at the  $j$ th locus in two accessions under consideration and  $m$  refers to the number of loci (Reif et al., 2005). We used Ward's  $D^2$  as agglomeration method, available in the built-in R function *hclust* (Murtagh & Legendre, 2014). We clustered the 1096 accessions into 13 clusters (Figure S3) based on the Ball–Hall index (Ball & Hall, 1965).

## 2.8 | Quantitative analysis

We obtained narrow-sense heritabilities and genetic and nongenetic correlations from the covariance components estimated using a multivariate mixed model implemented in GIBBS3F90 (Misztal et al., 2015). The final dataset (Table S2), containing multitrait single values per accession, was used to fit the multivariate model. This model fits multiple traits simultaneously and the model for each trait can be described as

$$\mathbf{y}_k = \mu + \mathbf{Z}_k \mathbf{u}_k + \mathbf{e}_k, \quad (4)$$

where  $\mathbf{y}_k$  is the vector of observations for the  $k$ th trait,  $\mu$  is the overall mean of the  $k$ th trait,  $\mathbf{Z}_k \mathbf{u}_k$  is the genetic term treated as a random effect, with  $\mathbf{Z}_k$  as the incidence matrix of the random effect (i.e., genotypes) and  $\mathbf{u}_k$  as the vector of regression coefficients or breeding values for the  $k$ th trait, and  $\mathbf{e}_k$  is the vector of residuals. The variance of the response variables ( $\mathbf{y} = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_k\}$ ) is described as

$$\text{Var}(\mathbf{y}) = \mathbf{Z} \left( \mathbf{G} \otimes \sum_a \right) \mathbf{Z}' + \mathbf{I} \otimes \sum_e, \quad (5)$$

where  $\mathbf{G}$  is the relationship matrix calculated from molecular markers,  $\sum_a$  is the additive genetic covariance matrix with the additive genetic variances ( $\sigma_{a_i}^2$ ) of the traits in the diagonal and the pairwise covariances ( $\sigma_{a_{ij}}$ ) among traits in the off-diagonal,  $\mathbf{I}$  is an identity matrix,  $\sum_e$  is the residual covariance matrix with residual variances ( $\sigma_{e_i}^2$ ) in the main diagonal and residual covariances in the off-diagonal ( $\sigma_{e_{ij}}$ ), and  $\otimes$  denotes the Kronecker product. Using the elements of  $\sum_a$  and  $\sum_e$ ,



we estimated the additive genetic (Equation 6) and residual correlations (Equation 7) as

$$\rho_{a_{ij}} = \frac{\sigma_{a_{ij}}}{\sqrt{\sigma_{a_i}^2 \sigma_{a_j}^2}}, \quad (6)$$

$$\rho_{e_{ij}} = \frac{\sigma_{e_{ij}}}{\sqrt{\sigma_{e_i}^2 \sigma_{e_j}^2}}. \quad (7)$$

We estimated trait narrow-sense heritabilities using the formula

$$h_i^2 = \frac{\sigma_{a_i}^2}{\sigma_{a_i}^2 + \sigma_{e_i}^2}, \quad (8)$$

where  $\sigma_{a_i}^2$  and  $\sigma_{e_i}^2$  were described previously.

## 2.9 | Graphical representation of traits relationships

We used principal component analysis (PCA) to display the trait relationships detected in the phenotypic, genetic, and residual correlation matrices. We computed the principal components by eigendecomposition of the phenotypic, genetic, and residual correlation matrices using the R function `princomp` (R Core Team, 2018). This multivariate technique extracts the most crucial information from the data to display the patterns of similarity of the observations and variables as points in maps (Bartholomew, 2010). Traits projected in the same direction have similar properties, while traits projected in opposite directions are antagonistic (Xavier et al., 2017).

Undirected Gaussian graphical models represent the covariance between two random variables in multivariate Gaussian distribution in terms of sums of components related to individual paths between the variables in an underlying graphical model (Jones & West, 2005). We computed undirected graphical models from the additive genetic, residual, and phenotypic correlation matrix to infer the connection among traits. We used a Gaussian undirected graphical model based on neighborhood selection with the graphical least absolute shrinkage and selection operator (GLASSO) algorithm as proposed by Meinshausen and Bühlmann (2006) and implemented by Zhao et al. (2012) using the “huge” package in R. Undirected graphical models have been shown to be useful for uncovering patterns of interaction among soybean traits (Lopez et al., 2021; Xavier et al., 2017). These models are particularly useful when studying variables that are

highly correlated, as is the case with plant traits, to infer causal relationships between them. and their use can also enable the identification of key traits that may have a large impact on the overall trait interaction network in multitrait interaction studies (Hastie et al., 2005; Xavier et al., 2017).

## 2.10 | Indirect selection and indexes of selection

We calculated the efficiency of indirect selection over the direct phenotypic selection of carbohydrate traits according to Xavier et al. (2017) using the formula

$$EF = \frac{|r_g| h_x^2}{h_y^2}, \quad (9)$$

where  $|r_g|$  is the absolute value of genetic correlation,  $h_x^2$  is the heritability of the secondary trait, and  $h_y^2$  is the heritability of the trait of interest. In addition, we constructed a selection index for seed compositional traits in soybeans. Multitrait selection index is a method that allows the simultaneous improvement of traits by selecting the individuals with the highest overall merit based on observable traits (Bouchet et al., 2017; Céron-Rojas & Crossa, 2022). A linear phenotypic selection index ( $I$ ) can be written as

$$I = \mathbf{b}'\mathbf{x} = b_1x_1 + b_2x_2 + \dots + b_nx_n, \quad (10)$$

where  $\mathbf{b}$  corresponds to the vector of coefficients and  $\mathbf{x}$  corresponds to vector of traits phenotypic values, where the coefficients are computed as a function of the variance-covariance matrix of the trait's phenotypes ( $\mathbf{P}$ ), the genetic covariance matrix ( $\mathbf{C}$ ), and the economic weights ( $\mathbf{w}$ ) as  $\mathbf{b} = \mathbf{P}^{-1}\mathbf{C}\mathbf{w}$  (Hazel, 1943; Smith, 1936). The two fundamental parameters associated with a selection index are the selection response ( $R_I$ ) and the expected ( $E$ ) gain per trait (equations 9 and 10 in Céron-Rojas & Crossa, 2022), which are defined as

$$R_I = k_I \frac{\mathbf{b}'\mathbf{C}\mathbf{w}}{\sqrt{\mathbf{b}'\mathbf{P}\mathbf{b}}}, \quad (11)$$

$$E = k_I \frac{\mathbf{C}\mathbf{b}}{\sqrt{\mathbf{b}'\mathbf{P}\mathbf{b}}}, \quad (12)$$

where  $k_I$  is a standardized selection differential that corresponds to 2.063 if the selection intensity is 5% (Falconer & Mackay, 1996), and  $\mathbf{b}$ ,  $\mathbf{P}$ , and  $\mathbf{C}$  were described previously. The selection index was constructed using the R package “selection.index” (Goyani, 2021).

**TABLE 1** Phenotypic correlations: Spearman's correlation coefficients (upper right diagonal elements) and Pearson's correlation coefficients (lower left diagonal) coefficients.

Trait	Suc	Raf	Sta	SW	Flo	Mat	Oil	Prot
Suc		0.32****	0.23****	0.26****	−0.10**	0.06*	0.18****	−0.17****
Raf	0.32****		−0.02	0.07*	−0.14****	−0.16****	0.15****	−0.08**
Sta	0.21****	−0.05		−0.02	0.05	0.08**	−0.10**	0.11***
SW	0.19****	0.07*	−0.03		−0.12****	0.16****	0.32****	−0.11***
Flo	−0.10***	−0.12****	0.03	−0.08**		0.74****	−0.37****	0.08**
Mat	0.06	−0.15****	0.11***	0.19****	0.56****		−0.26****	0.10**
Oil	0.16****	0.15****	−0.11***	0.26****	−0.32****	−0.22****		−0.54****
Prot	−0.18****	−0.06*	0.11***	−0.08**	0.05	0.09**	−0.54****	

Abbreviations: Flo, flowering data; Mat, maturity date; Oil, oil content; Prot, protein content; Raf, raffinose; Sta, stachyose; Suc, sucrose; SW, seed weight.

\*\*\*\*Significant at the 0.0001 probability level; \*\*\*significant at the 0.001 probability level; \*\*significant at the 0.01 probability level; \*significant at the 0.05 probability level.

### 3 | RESULTS

#### 3.1 | Phenotypic variability and genetic diversity

The grouping of the studied diverse panel into 13 cluster groups is presented in Figure S3. The frequency count of the off-diagonal elements of the genomic relationship matrix (Figure S4B) and the corresponding heatmap (Figure S4A) showed that there are no large blocks of high genomic relationship, revealing genetic diversity among the studied accessions. The smaller ranges observed in seed compositional traits indicate a narrow variation for these seed traits. A complete summary statistic of the 1096 *G. max* accessions is presented in Table S3.

#### 3.2 | Trait correlations

Pearson's and Spearman's correlation coefficients (Table 1) had similar values across pairwise correlations revealing linear relationships among traits. Nonlinear (monotonic) associations occur when Spearman correlations are greater than Pearson correlations (Xavier et al., 2017). Correlations among flowering and maturity (0.56), and protein and oil (−0.54), a well-known soybean trade-off, were the highest significant correlations. Sucrose, stachyose, and raffinose exhibited correlations lower than 0.2 with other traits. Stachyose content was significantly correlated with maturity (0.11), protein (0.11), and oil (−0.11); sucrose was significantly correlated with seed weight (0.19), oil (0.16), protein (−0.18), and flowering (−0.10); and raffinose was correlated to seed weight (0.07), oil (0.15), maturity (−0.15), protein (−0.06), and flowering (−0.12). Among carbohydrate traits, there was a significant correlation of sucrose with raffinose (0.32) and stachyose (0.21), and no significant correlation between raffinose and stachyose (−0.05).

Table 2 shows additive genetic and residual correlations among traits. Significant additive genetic correlations ranged from 0.07 to 0.68. Protein and oil (−0.68) and flowering and maturity (0.66) had the highest additive genetic correlations. Genetic correlations suggested that selection for low raffinose content could negatively affect oil content (0.20). Stachyose had correlations lower than 0.2 with other traits. Sucrose had positive additive genetic correlations with maturity, while raffinose had negative additive genetic correlations with maturity. Sucrose and raffinose had negative additive genetic correlations with flowering time. The additive genetic correlation of sucrose and protein content (−0.33) was negative and positive for sucrose with seed weight (0.18) and oil content (0.23). Among the carbohydrates, the highest correlation was between sucrose and raffinose (0.25). Raffinose and stachyose, however, had a positive but low correlation (0.07).

Significant residual correlations ranged from 0.06 to 0.54 (Table 2). In this study, residual correlations represented the proportion of variance that two traits shared due to non-additive genetic factors, such as epistasis and dominance and environmental factors not accounted for in the statistical model. Residual correlations of sucrose with raffinose (0.38) and flowering with maturity (0.54) were the highest residual correlations. Most pairwise phenotypic and additive genetic correlations shared the same direction of the relationships between traits. In contrast, pairwise residual correlations showed a different direction of the relationships between traits compared to phenotypic and additive genetic correlations.

#### 3.3 | Heritability

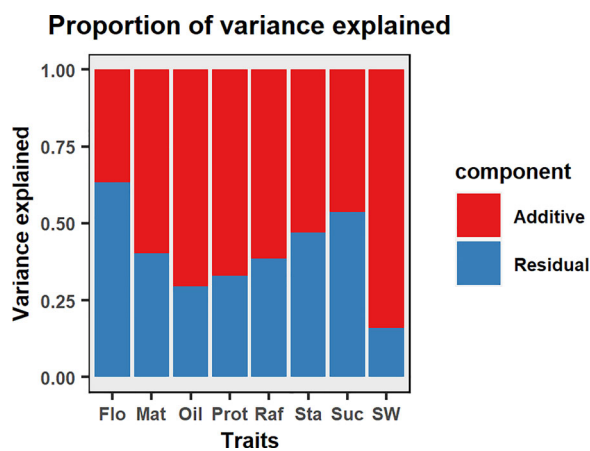
Narrow sense heritability (Table 2) across traits ranges from 0.36 to 0.83. Flowering (0.36), sucrose (0.46), stachyose (0.53), maturity (0.59), and raffinose (0.61) showed the lowest heritabilities compared to protein (0.67), oil (0.70), and seed weight (0.83). Knowledge of narrow sense heritabilities

**TABLE 2** Residual correlation coefficients (upper right diagonal elements), additive genetic correlation coefficients (lower left diagonal), and narrow-sense heritabilities (main diagonal, bold letters).

Trait	Suc	Raf	Sta	SW	Flo	Mat	Oil	Prot
Suc	<b>0.46</b>	0.38****	0.24****	−0.06*	0.01	−0.07**	−0.08**	−0.03
Raf	0.25****	<b>0.61</b>	−0.01	0.03	0.01	0.13****	−0.19****	0.11***
Sta	0.18****	0.07*	<b>0.53</b>	−0.04	0.01	0.07**	−0.13****	−0.01
SW	0.18****	−0.12****	0.10*	<b>0.83</b>	0.29****	0.29****	−0.22****	0.11***
Flo	−0.10***	−0.18****	−0.01	−0.07*	<b>0.36</b>	0.54****	−0.22****	−0.21****
Mat	0.23****	−0.36****	0.05*	0.14****	0.66****	<b>0.59</b>	−0.21****	−0.15****
Oil	0.23****	0.20****	−0.03	0.21****	−0.24****	−0.13****	<b>0.70</b>	−0.16****
Prot	−0.33****	−0.08**	0.12****	−0.05	0.18****	0.13****	−0.68****	<b>0.67</b>

Abbreviations: Flo, flowering data; Mat, maturity date; Oil, oil content; Prot, protein content; Raf, raffinose; Sta, stachyose; Suc, sucrose; SW, seed weight.

\*\*\*\*Significant at the 0.0001 probability level; \*\*\*significant at the 0.001 probability level; \*\*significant at the 0.01 probability level; \*significant at the 0.05 probability level.



**FIGURE 1** The plot of the proportion of variance explained by additive genetic and residual components. Suc, sucrose; Raf, raffinose; Sta, stachyose; SW, seed weight; Flo, flowering data; Mat, maturity date; Oil, oil content; Prot, protein content.

allows for estimating the genetic gain and defining the best strategies for trait selection in plant-breeding programs (Fellahi et al., 2018).

### 3.4 | Polygenic architecture

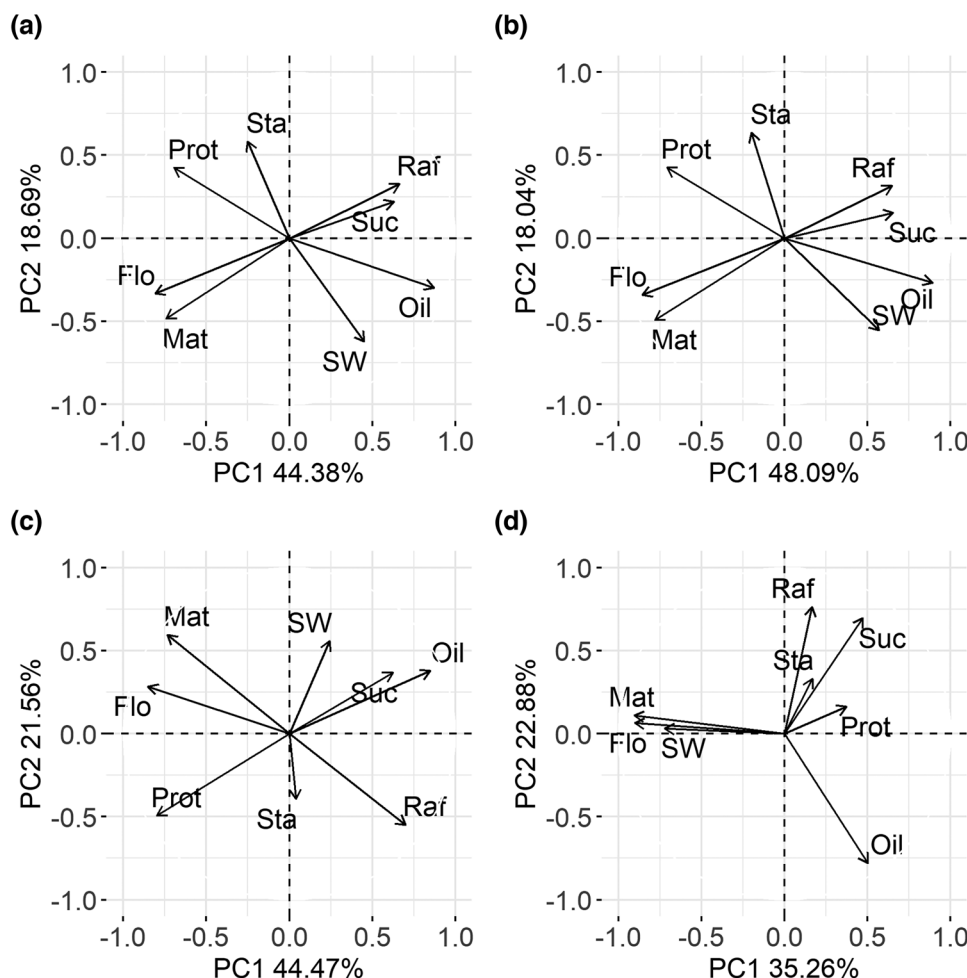
Figure 1 shows the proportion of genetic variance explained by the additive genetic and residual components across all traits. Traits' residual variance in this study represents the variance explained by nonadditive genetic components such as epistasis and dominance and environmental factors not accounted for in the statistical model used to estimate the BLUPs for each trait per accession. The proportion of genetic variance explained by additive genetic and residual factors ranged from 0.36 to 0.83 and from 0.16 to 0.63, respectively. Phenotypic variances of protein, oil, and seed weight are mainly due to additive genetic factors. In contrast, traits

such as sucrose and flowering have a higher residual component indicating a significant effect of environmental and nonadditive genetic factors in trait variation.

### 3.5 | PCA and graphical models

The Pearson and Spearman correlations' biplot representations (Figure 2a,b) of PCA were similar. They displayed positive phenotypic associations between raffinose, sucrose, seed weight, and oil and between protein, stachyose, flowering, and maturity. Traits displaying negative associations include raffinose and sucrose with flowering and maturity, and protein and stachyose with seed weight and oil. Sucrose and oil, and flowering and maturity displayed the strongest positive additive genetic correlations (Figure 2c). In contrast, we observed negative additive genetic correlations for protein with sucrose, seed weight, and oil, and for raffinose with flowering and maturity. Residual correlations (Figure 2d) revealed positive associations between raffinose, stachyose, and sucrose, and between flowering, seed weight, and maturity. Biplot representations of phenotypic, residual, and additive genetic correlations showed similar patterns of associations for protein with oil and flowering with maturity. Traits with stronger phenotypic correlations are more likely to show similar trends of associations across additive and residual correlations.

Figure 3 shows the structure and dependence among variables as a network for phenotypic (Figure 3a,b), additive genetic (Figure 3c), and residual correlations (Figure 3d). Traits with the strongest phenotypic correlations, such as protein and oil, and flowering and maturity, showed interdependence in phenotypic and additive genetic networks. Flowering and maturity also showed strong interdependence in the residual network. Carbohydrate traits did not show interdependence with other traits in the phenotypic and residual networks but did in the additive network. The additive



**FIGURE 2** Principal component analysis of (a) phenotypic Pearson, (b) phenotypic Spearman, (c) additive genetic, and (d) residual correlations of soybean traits. The variation explained by each principal component is presented on the axes. Suc, sucrose; Raf, raffinose; Sta, stachyose; SW, seed weight; Flo, flowering data; Mat, maturity date; Oil, oil content; Prot, protein content.

genetic network revealed interdependence between raffinose, maturity, and flowering time. At the same time, protein shared additive genetic connections with sucrose and oil. In the residual network, sucrose interacted with raffinose, seed weight interacted with maturity and flowering, and stachyose, protein, and oil did not show interdependence with other traits.

### 3.6 | Indirect selection and selection index

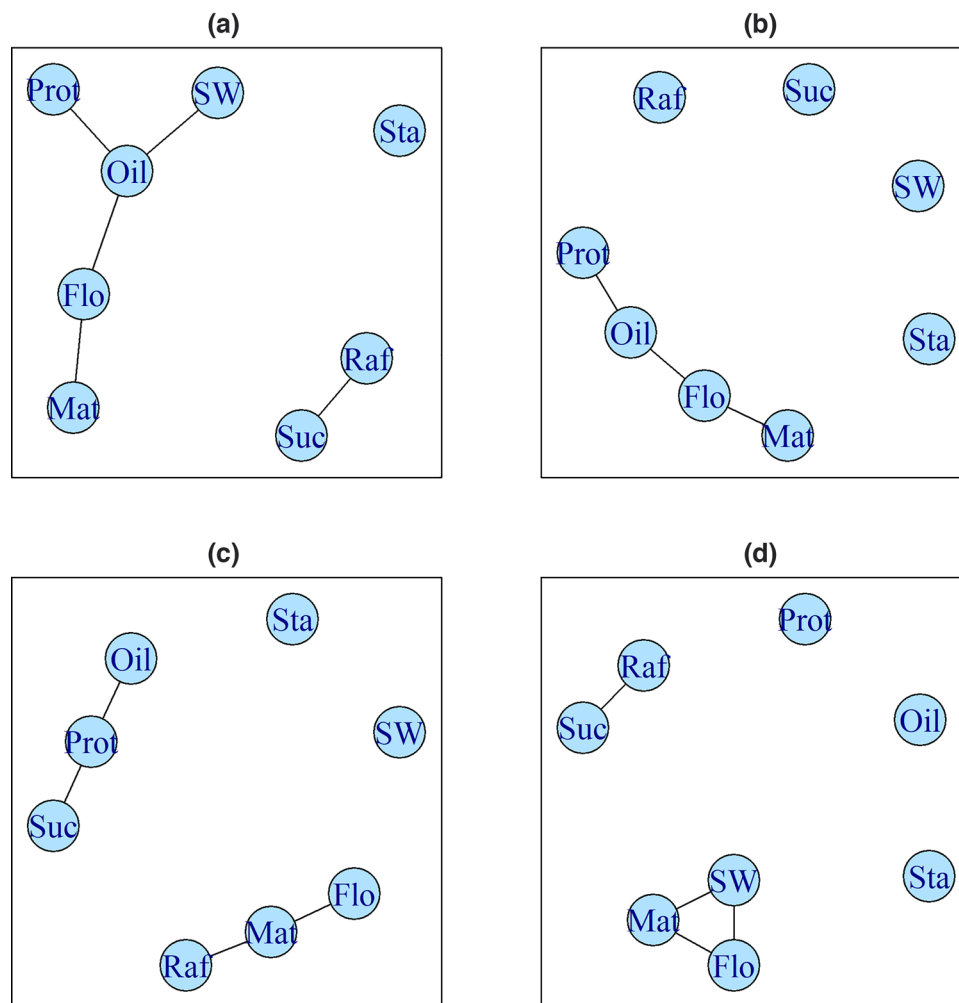
In this study, we did not identify any trait with high indirect selection efficiency for carbohydrate content, with efficiencies ranging from 0.08 to 0.48 (Table S4). When considering an economic weight of zero for raffinose and stachyose and equal economic weights (1) for sucrose, protein, and oil contents, the constructed selection index was  $I = -0.004 \cdot \text{Suc} + 0.206 \cdot \text{Raf} + 0.0132 \cdot \text{Sta} + 0.11 \cdot \text{Oil} + 0.054 \cdot \text{Prot}$ , which had a selection response (0.586) four times higher than the selection response for sucrose content (0.132).

The expected change in individual traits ranged from  $-0.01$  to  $0.08$  and is presented in Figure S5.

## 4 | DISCUSSION

The basis of crop improvement and evolution relies on the large genetic diversity of plant populations (Bhandari et al., 2017); however, in this study, there were narrow ranges of variation in seed compositional traits among the studied *G. max* accessions, which indicated lower variation in these traits (Table S3). Similar ranges in the content of seed compositional traits have been observed in previous studies (Hou et al., 2009; Zhang et al., 2010). The narrow variation of seed compositional traits in *G. max* accessions is likely the result of genetic selection and domestication, which could be a significant barrier to future soybean meal quality improvement. Available strategies to increase genetic diversity in breeding populations include exploring natural genetic variation in





**FIGURE 3** Undirected graphical models of (a) phenotypic Pearson, (b) phenotypic Spearman, (c) additive genetic, and (d) nonadditive genetic correlations using the least absolute shrinkage and selection operator (LASSO) algorithm proposed by Meinshausen and Bühlmann (2006). Suc, sucrose; Raf, raffinose; Sta, stachyose; SW, seed weight; Flo, flowering data; Mat, maturity date; Oil, oil content; Prot, protein content.

extensive wild and domesticated soybean collections, induced mutation, and transgenic approaches (Hyten et al., 2006; Patil et al., 2017; Valliyodan et al., 2016).

Sucrose, raffinose, and stachyose content had lower narrow-sense heritabilities (Table 2), with a more significant contribution of environmental and nonadditive genetic factors in the trait's variation compared to protein, oil, and seed weight (Figure 1); however, high narrow- and broad-sense heritabilities for carbohydrate traits have also been previously reported in soybean recombinant inbred line populations and smaller subsets of the USDA soybean germplasm collection (Cicek et al., 2006; Jiang et al., 2018; Kim et al., 2006). The difference in heritabilities among studies is due to the genetic diversity, size of the populations, and the effect of nonadditive genetic factors and the environments. A plant breeder's primary concern is additive genetic variance, as it contributes significantly to selection response (Fellahi et al., 2018; Hem et al., 2021). Contrarily, breeders usually ignore the nonadditive genetic variance because of its complex calculation,

the difficulty in using it in practice for parent selection, and the difficulty of separating it from the environmental variance (Jiang & Reif, 2015; Varona et al., 2018). A substantial contribution from nonadditive genetic effects on sucrose variation could slow down genetic progress when selecting for sucrose content unless sources of greater additive genetic variation exist.

In soybean populations, the trade-off between oil and protein is strong and stable, making developing cultivars with increased oil and protein contents challenging (Bueno et al., 2018; Chung et al., 2003; Jiang et al., 2018; Patil et al., 2018; Sun et al., 2017). In the panel of this study, protein and oil had high phenotypic and additive genetic correlations (Tables 1 and 2). High additive genetic correlations between two traits occur mainly due to the pleiotropic effects of individual genes, shared biochemical pathways between traits, or linkage disequilibrium between genes that affect only each one of the traits (Bennett et al., 2014; Hill, 2013). Despite their massive efforts, plant scientists have been unable to break the trade-off

between protein and oil; however, a recent study suggests that it may be possible to overcome the inverse correlation between the two by examining the C and N partition during seed development (Kambhampati et al., 2020). Due to the importance of protein and oil for soybean seed quality, researchers will continue to try to understand and modify the trade-off between the two.

Soybean flowering and maturity are crucial traits in soybean productivity and seed quality. Overlap between loci controlling flowering and maturity in soybeans indicates that they might share some genetic basis (Fang et al., 2017; Lee et al., 1996; Zhang et al., 2004, 2015), which leads to high correlations between the two traits, and similar interactions with other seed traits as shown in this study (Figure 2; Table 2). A study on the compositional change of soybeans during seed development and maturity revealed that oil content accumulates rapidly in the early stages, protein decreases during the first weeks but later increases gradually, sucrose decreases over time, and oligosaccharides remained low at first and then increase toward maturity (Saldivar et al., 2011). Phenotypic correlation between maturity and sucrose was low (0.06) and insignificant (Table 1). Genetic correlation among traits (Table 2) indicated that accessions with later flowering accumulate lower sucrose and raffinose content on the seeds, while later maturity accessions accumulate higher sucrose and lower raffinose contents. Considering Saldivar et al. (2011) and our results, sucrose will likely increase under a more extended reproductive period, while the already accumulated raffinose may begin to decrease. Flowering time had a stronger influence on protein and oil contents than maturity, whereas later flowering accessions had high protein and low oil contents. Previous studies found similar correlation trends between flowering and maturity with carbohydrate traits (Bachlava et al., 2009; Bellaloui et al., 2009; Cicek et al., 2006; Recker et al., 2014); therefore, the observed correlation trends suggest that obtaining the high-sucrose-content phenotype might be easier in soybeans with later maturity and earlier flowering, which could be a target of soybean breeding programs.

Modifying carbohydrate seed composition could alter protein and oil contents as sucrose had negative correlations with protein, and sucrose and raffinose had positive correlations with oil. Stachyose content exhibited a positive association with protein, as described by Jiang et al. (2018), and a low insignificant association with oil content. Despite the contradictory results for carbohydrate correlations with protein and oil from previous research (Jiang et al., 2018; Kim et al., 2005; Li et al., 2012), there is a greater agreement that oil content correlates positively with carbohydrate traits, while protein content correlates negatively with carbohydrate traits (Bueno et al., 2018; Cicek et al., 2006; Hymowitz et al., 1972; Jaureguy et al., 2011; Wilcox & Shibbles, 2001). Scientists believe that a limited supply of carbon during seed maturation is the principal cause of the observed trade-offs among

carbohydrates, protein, and oil contents, where the metabolic processes in the developing embryo and the supply of amino acids and carbohydrates from maternal sources establish the proportion of carbohydrates, protein, and oil accumulated in soybean seeds (Allen & Young, 2013; Kambhampati et al., 2020; Truong et al., 2013). Kambhampati et al. (2020) suggested that in late seed development, carbon derived from the turnover of lipids and proteins contributes to the synthesis of RFOs; thus, the repartitioning of carbon among storage reserves likely contributes to the observed correlations among carbohydrates, protein, and oil. Our results indicate that increasing sucrose content could compromise protein content and breeders need to fine-tune the proportion of protein and sucrose through conventional breeding and transgenic approaches.

The positive association of sucrose and RFOs suggested concurrent accumulations of sucrose, raffinose, and stachyose in soybean seeds, which agree with the RFOs metabolic pathway that depends on sucrose as substrate (Elango et al., 2022; Tian et al., 2019). Stachyose showed a smaller correlation with sucrose than raffinose, which could be associated with their order of synthesis since raffinose is the first RFO formed in the biosynthetic pathway, followed by stachyose and verbascose (Elango et al., 2022). Previous reports also found positive correlations between sucrose, raffinose, and stachyose (Cicek et al., 2006; Hymowitz et al., 1972; Jaureguy et al., 2011); however, in this study, correlations between sucrose and RFOs were low, which indicates that increasing sucrose without increasing concentrations of RFOs could be possible. It is also noteworthy that previous studies have already been able to significantly reduce the RFOs content in soybean seeds (Dierking & Bilyeu, 2008, 2009a; Kambhampati et al., 2020; Le et al., 2020). In addition, the positive association between seed size and sucrose content could be helpful when selecting cultivars with larger seeds since previous reports found that seed size is also positively associated with yield (Cicek et al., 2006; Maestri et al., 1998).

The correlation of carbohydrate traits with oil and maturity content had contrary trends at the additive genetic and residual levels. Correspondence of genetic and residual correlations was mainly observed in traits highly correlated at the phenotypic level, as suggested by a previous study where the concurrence of genetic and residual correlations often was associated with high phenotypic correlations (Xavier et al., 2017). Opposite additive genetic and residual correlations could result from strong genotype  $\times$  environment interactions affecting trait phenotypes (Falconer & MacKay, 1996; Moreira et al., 2019; Xavier et al., 2017); thus, it appears that correlations between raffinose with oil content and maturity are more specific to populations and environments and may not be generalized to the whole species like other well-documented correlations such as protein and oil. In contrast, sucrose residual correlations with protein, seed

weight, flowering, maturity, and oil were low, indicating that the association between these traits is less influenced by nonadditive genetic factors, making them more stable.

Accurate measurements of genetic and environmental covariances between pairs of complex traits are crucial to characterize their genetic and environmental architectures (Gao et al., 2021); however, reasonable estimates of genetic and environmental covariances require large datasets (Hill, 2013) and accurate measurements of the phenotypes. To our knowledge, we explored carbohydrate interactions with other traits in a more extensive dataset than the ones used in previous studies; therefore, this study provided valuable insight into the correlation trends among traits. It is important to note that carbohydrate trait associations with other traits were low ( $<0.33$ ), that variance components are population dependent, and that results from this study may not hold in other populations. Although correlation estimates are not all consistent among different studies, some patterns emerge; for instance, there is a significant agreement in the positive association of sucrose and raffinose with oil, the negative association between sucrose and protein, and the positive association between carbohydrates.

Pyramiding (or introgression) of favorable alleles and desirable traits into a single cultivar is usually the goal of plant breeders. When traits have low heritabilities and complex interactions with other traits, as is the case of carbohydrate contents, a breeder could consider indirect selection and the simultaneous improvement of correlated traits through selection indexes (Bernardo, 2014). In soybean breeding, estimates of the additive genetic correlations and trait heritabilities are helpful for determining the indirect response of traits to selection, which is particularly important if there are genetic interactions among multiple traits (Kwon & Torrie, 1964; Recker et al., 2014; Xavier et al., 2017). In this study, we did not identify any trait with high indirect selection efficiency to be helpful in effectively selecting for high sucrose content (Table S4). If negative correlations exist, indirect selection of secondary traits is not feasible unless it is desirable to lower one of the traits. On the other hand, the multitrait selection index is a method that allows the simultaneous improvement of traits by selecting the individuals with the highest overall merit based on observable traits (Bouchet et al., 2017; Céron-Rojas & Crossa, 2018). Our results suggest that using a selection index for simultaneous improvements of seed compositional traits could lead to simultaneous increases in carbohydrate and oil contents at the expense of seed protein content (Figure S4); however, these results might vary between populations depending on the observed correlations among traits in soybeans. In addition, plant breeders could fine-tune the desired gains per trait by assigning different economic weights to the traits in the index. Multitrait genomic prediction is the gold standard for simultaneous trait's improvement, which has proved to improve prediction

accuracy of low heritable traits in animal and plant breeding (Jean et al., 2021; Manzanilla-Pech et al., 2020; Sun et al., 2017); however, it has not been applied yet for improving sucrose content in soybeans. Further research is also necessary to explore the environmental, physiological, molecular, and genetic factors that influence seed trait interactions during seed development and maturation to identify candidate genes for high-quality soybean germplasm production.

## 5 | CONCLUSIONS

This study reveals the interactions of seed sucrose with RFOs, other seed traits, flowering, and maturity times at the phenotypic, additive genetic, and residual levels. The additive genetic correlation of flowering and maturity time with sucrose suggests that obtaining cultivars with high sucrose content may be enhanced in earlier flowering and later maturity cultivars. Our results suggest a concurrent accumulation of sucrose, raffinose, and stachyose in soybean seeds; however, they have low correlations, indicating that increasing sucrose without increasing content of RFOs is possible. Sucrose correlations with protein, oil, seed weight, flowering, and maturity are primarily driven by their additive genetic correlations, which makes them more stable associations. In contrast, correlations of raffinose with oil and maturity are more population and environment specific. Our results suggest that the major constraints for improving soybean seed quality are the trade-offs of protein with oil and sucrose, and increasing sucrose content may lead to sacrifices in protein content. It is important to note that carbohydrate trait associations with other traits were low ( $<0.33$ ), that variance components are population dependent, and that results from this study may not hold in other populations. Although correlation estimates are not all consistent among different studies, some patterns emerge; for instance, there is a significant agreement in the positive association of sucrose and raffinose with oil, the negative association between sucrose and protein, and the positive association between carbohydrates. We were unable to estimate correlation coefficients within environments given the limitations of the dataset studied. It could thus be useful to conduct further research to identify genetic correlations specific to the environments in order to gain a better understanding of how environmental factors influence seed compositional traits.

## AUTHOR CONTRIBUTIONS

**Diana M. Escamilla:** Conceptualization; methodology; formal analysis, writing—original draft; writing—review and editing. **Henry T. Nguyen:** Data curation. **Tri D. Vuong:** Data curation. **Alencar Xavier:** Methodology. **Katy Martin Rainey:** Conceptualization; project administration; resources; writing—review and editing.

## ACKNOWLEDGMENTS

The authors wish to acknowledge the funding support by the United Soybean Board (USB project 1820-152-0101). We thank Haiying Shi for her technical assistance in the HPLC analysis.

## CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.


## DATA AVAILABILITY STATEMENT

The authors confirm that the carbohydrate and passport data supporting the findings of this study are available within Tables S1 and S2.

## ORCID

Diana M. Escamilla  <https://orcid.org/0000-0001-9293-9581>

Alencar Xavier  <https://orcid.org/0000-0001-5034-9954>

Katy Martin Rainey  <https://orcid.org/0000-0001-8541-5851>

## REFERENCES

- Allen, D. K., & Young, J. D. (2013). Carbon and nitrogen provisions alter the metabolic flux in developing soybean embryos. *Plant Physiology*, 161(3), 1458–1475. <https://doi.org/10.1104/pp.112.203299>
- Bachlava, E., Dewey, R. E., Burton, J. W., & Cardinal, A. J. (2009). Mapping and comparison of quantitative trait loci for oleic acid seed content in two segregating soybean populations. *Crop Science*, 49(2), 433–442. <https://doi.org/10.2135/cropsci2008.06.0324>
- Ball, G. H., & Hall, D. J. (1965). *ISODATA, a novel method of data analysis and pattern classification*. Stanford Research Institute.
- Bandillo, N., Jarquin, D., Song, Q., Nelson, R., Cregan, P., Specht, J., & Lorenz, A. (2015). A population structure and genome-wide association analysis on the USDA Soybean Germplasm Collection. *The Plant Genome*, 8(3). <https://doi.org/10.3835/plantgenome2015.04.0024>
- Bartholomew, D. J. (2010). Principal components analysis. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International Encyclopedia of Education* (Vol. 2, pp. 374–377). John Wiley & Sons, Inc. <https://doi.org/10.1016/B978-0-08-044894-7.01358-0>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Bellaloui, N., Smith, J. R., Ray, J. D., & Gillen, A. M. (2009). Effect of maturity on seed composition in the early soybean production system as measured on near-isogenic soybean lines. *Crop Science*, 49(2), 608–620. <https://doi.org/10.2135/cropsci2008.04.0192>
- Bennett, G. L., Pollak, E. J., Kuehn, L. A., & Snelling, W. M. (2014). Breeding: Animals. In N. K. Van Alfen (Ed.), *Encyclopedia of agriculture and food systems* (pp. 173–186). Academic Press. <https://doi.org/10.1016/B978-0-444-52512-3.000228-X>
- Bernardo, R. (2014). *Essentials of plant breeding*. Stemma Press. <http://lib.ugent.be/catalog/rug01:002156374>
- Bhanu, A. N. (2017). Assessment of genetic diversity in crop plants - An overview. *Advances in Plants & Agriculture Research*, 7(3), 279–286. <https://doi.org/10.15406/apar.2017.07.00255>
- Bouchet, S., Olatoye, M. O., Marla, S. R., Perumal, R., Tesso, T., Yu, J., Tuinstra, M., & Morris, G. P. (2017). Increased power to dissect adaptive traits in global sorghum diversity using a nested association mapping population. *Genetics*, 206(2), 573–585. <https://doi.org/10.1534/genetics.116.198499>
- Bueno, R. D., Borges, L. L., God, P. I. V. G., Piovesan, N. D., Teixeira, A. I., Cruz, C. D., & Barros, E. G. D. (2018). Quantification of anti-nutritional factors and their correlations with protein and oil in soybeans. *Anais Da Academia Brasileira de Ciencias*, 90(1), 205–217. <https://doi.org/10.1590/0001-3765201820140465>
- Burton, J. W. (1997). Soyabean (*Glycine max* (L.) Merr.). *Field Crops Research*, 53(1–3), 171–186. [https://doi.org/10.1016/S0378-4290\(97\)00030-0](https://doi.org/10.1016/S0378-4290(97)00030-0)
- Céron-Rojas, J. J., & Crossa, J. (2018). The linear phenotypic selection index theory. In J. J. Céron-Rojas & J. Crossa (Eds.), *Linear selection indices in modern plant breeding* (pp. 15–42). Springer.
- Céron-Rojas, J. J., & Crossa, J. (2022). The statistical theory of linear selection indices from phenotypic to genomic selection. *Crop Science*, 62, 537–563. <https://doi.org/10.1002/csc2.20676>
- Chung, J., Babka, H. L., Graef, G. L., Staswick, P. E., Lee, D. J., Cregan, P. B., Shoemaker, R. C., & Specht, J. E. (2003). The seed protein, oil, and yield QTL on soybean linkage group I. *Crop Science*, 43(3), 1053–1067. <https://doi.org/10.2135/cropsci2003.1053>
- Cicek, M. S., Chen, P., Saghai Maroof, M. A., & Buss, G. R. (2006). Interrelationships among agronomic and seed quality traits in an inter-specific soybean recombinant inbred population. *Crop Science*, 46(3), 1253–1259. <https://doi.org/10.2135/cropsci2005.06-0162>
- Dierking, E. C., & Bilyeu, K. D. (2008). Association of a soybean raffinose synthase gene with low raffinose and stachyose seed phenotype. *The Plant Genome Journal*, 1(2), 135–145. <https://doi.org/10.3835/plantgenome2008.06.0321>
- Dierking, E. C., & Bilyeu, K. D. (2009a). New sources of soybean seed meal and oil composition traits identified through TILLING. *BMC Plant Biology*, 9, 1–11. <https://doi.org/10.1186/1471-2229-9-89>
- Dierking, E. C., & Bilyeu, K. D. (2009b). Raffinose and stachyose metabolism are not required for efficient soybean seed germination. *Journal of Plant Physiology*, 166(12), 1329–1335. <https://doi.org/10.1016/j.jplph.2009.01.008>
- Dwivedi, S. L., Reynolds, M. P., & Ortiz, R. (2021). Mitigating tradeoffs in plant breeding. *iScience*, 24(9), 102965. <https://doi.org/10.1016/j.isci.2021.102965>
- Elango, D., Rajendran, K., Van Der Laan, L., Sebastiar, S., Raigne, J., Thaiparambil, N. A., El Haddad, N., Raja, B., Wang, W., Ferela, A., Chiteri, K. O., Thudi, M., Varshney, R. K., Chopra, S., Singh, A., & Singh, A. K. (2022). Raffinose family oligosaccharides: Friend or foe for human and plant health? *Frontiers in Plant Science*, 13, 829118. <https://doi.org/10.3389/fpls.2022.829118>
- Falconer, D. S., & Mackay, T. F. C. (1996). *Introduction to quantitative genetics* (4th ed.). Longmans Green. <https://www.pearson.com/us/higher-education/program/Falconer-Introduction-to-Quantitative-Genetics-4th-Edition/PGM194806.html>
- Fang, C., Ma, Y., Wu, S., Liu, Z., Wang, Z., Yang, R., Hu, G., Zhou, Z., Yu, H., Zhang, M., Pan, Y., Zhou, G., Ren, H., Du, W., Yan, H., Wang, Y., Han, D., Shen, Y., Liu, S., ... Tian, Z. (2017). Genome-wide association studies dissect the genetic networks underlying agronomical traits in soybean. *Genome Biology*, 18(1), 161. <https://doi.org/10.1186/s13059-017-1289-9>
- Fehr, W. R., Caviness, C. E., Burmood, D. T., & Pennington, J. S. (1971). Stage of development descriptions for soybeans, *Glycine max*



- (L.) Merrill. *Crop Science*, 11(6), 929–931. <https://doi.org/10.2135/cropsci1971.0011183X001100060051x>
- Fellahi, Z. E. A., Hannachi, A., & Bouzerzour, H. (2018). Analysis of direct and indirect selection and indices in bread wheat (*Triticum aestivum* L.) segregating progeny. *International Journal of Agronomy*, 2018, 8312857. <https://doi.org/10.1155/2018/8312857>
- Gao, B., Yang, C., Liu, J., & Zhou, X. (2021). Accurate genetic and environmental covariance estimation with composite likelihood in genome-wide association studies. *PLoS Genetics*, 17(1), e1009293. <https://doi.org/10.1371/journal.pgen.1009293>
- Garland, T. (2014). Trade-offs. *Current Biology*, 24(2), R60–R61. <https://doi.org/10.1016/j.cub.2013.11.036>
- Goyani, Z. (2021). Package “selection.index”. Analysis of selection index in plant breeding. R package.
- Graham, K., Kerley, M., Firman, J., & Allee, G. (2002). The effect of enzyme treatment of soybean meal on oligosaccharide disappearance and chick growth performance. *Poultry Science*, 81(7), 1014–1019. <https://doi.org/10.1093/ps/81.7.1014>
- Gupta, S. K., & Manjaya, J. G. (2022). Advances in improvement of soybean seed composition traits using genetic, genomic and biotechnological approaches. *Euphytica*, 218(7), 99. <https://doi.org/10.1007/s10681-022-03046-4>
- Hartman, G. L., West, E. D., & Herman, T. K. (2011). Crops that feed the World 2. Soybean-worldwide production, use, and constraints caused by pathogens and pests. *Food Security*, 3(1), 5–17. <https://doi.org/10.1007/s12571-010-0108-x>
- Hastie, T., Tibshirani, R., Friedman, J., & Franklin, J. (2005). The elements of statistical learning: Data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2), 83–85.
- Hauke, J., & Kossowski, T. (2011). Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data. *Quaestiones Geographicae*, 30(2), 87–93. <https://doi.org/10.2478/v10117-011-0021-1>
- Hazel, L. N. (1943). The genetic basis for constructing selection indexes. *Genetics*, 28(6), 476–490. <https://doi.org/10.1093/genetics/28.6.476>
- Hem, I. G., Selle, M. L., Gorjanc, G., Fuglstad, G.-A., & Riebler, A. (2021). Robust modeling of additive and nonadditive variation with intuitive inclusion of expert knowledge. *Genetics*, 217(3), iyab002. <https://doi.org/10.1093/genetics/iyab002>
- Hill, J. L., Peregrine, E. K., & Sprau, G. L. (2005). Evaluation of the USDA soybean germplasm collection: Maturity groups 000-IV (PI 507670-PI 574486). U.S. Department of Agriculture Technical Bulletin No. 1914. <https://www.ars.usda.gov/oc/np/soybeangermplasm/soybeangermplasmintro/>
- Hill, W. G. (2013). Genetic correlation. In S. Maloy & K. Hughes (Eds.), *Brenner's encyclopedia of genetics* (2nd ed., pp. 237–239). Academic Press. <https://doi.org/10.1016/B978-0-12-374984-0.00611-2>
- Hou, A., Chen, P., Shi, A., Zhang, B., & Wang, Y.-J. (2009). Sugar variation in soybean seed assessed with a rapid extraction and quantification method. *International Journal of Agronomy*, 2009, 484571. <https://doi.org/10.1155/2009/484571>
- Hymowitz, T., & Collins, F. I. (1974). Variability of sugar content in seed of. *Agronomy Journal*, 66, 239–240. <https://doi.org/10.2134/agronj1974.0002196200600020017x>
- Hymowitz, T., Collins, F. I., Panczner, J., & Walker, W. M. (1972). Relationship between the content of oil, protein, and sugar in soybean seed 1. *Agronomy Journal*, 64(5), 613–616. <https://doi.org/10.2134/agronj1972.00021962006400050019x>
- Hyten, D. L., Song, Q., Zhu, Y., Choi, I.-Y., Nelson, R. L., Costa, J. M., Specht, J. E., Shoemaker, R. C., & Cregan, P. B. (2006). Impact of genetic bottlenecks on soybean genome diversity. *Proceedings of the National Academy of Sciences of the United States of America*, 103(45), 16666–16671. <https://doi.org/10.1073/pnas.0604379103>
- Jauregui, L. M., Chen, P., & Scaboo, A. M. (2011). Heritability and correlations among food-grade traits in soybean. *Plant Breeding*, 130(6), 647–652. <https://doi.org/10.1111/j.1439-0523.2011.01887.x>
- Jean, M., Cober, E., O'donoghue, L., Rajcan, I., & Belzile, F. (2021). Improvement of key agronomical traits in soybean through genomic prediction of superior crosses. *Crop Science*, 61(6), 3908–3918. <https://doi.org/10.1002/csc.2.20583>
- Jiang, G.-L., Chen, P., Zhang, J., Florez-Palacios, L., Zeng, A., Wang, X., Bowen, R. A., Miller, A., & Berry, H. (2018). Genetic analysis of sugar composition and its relationship with protein, oil, and fiber in soybean. *Crop Science*, 58(6), 2413–2421. <https://doi.org/10.2135/cropsci2018.03.0173>
- Jiang, Y., & Reif, J. C. (2015). Modeling epistasis in genomic selection. *Genetics*, 201(2), 759–768. <https://doi.org/10.1534/genetics.115.177907>
- Jones, B., & West, M. (2005). Covariance decomposition in undirected Gaussian graphical models. *Biometrika*, 92(4), 779–786. <https://doi.org/10.1093/biomet/92.4.779>
- Kambhampati, S., Aznar-Moreno, J. A., Hostetler, C., Caso, T., Bailey, S. R., Hubbard, A. H., Durrett, T. P., & Allen, D. K. (2020). On the inverse correlation of protein and oil: Examining the effects of altered central carbon metabolism on seed composition using soybean fast neutron mutants. *Metabolites*, 10(1), 18. <https://doi.org/10.3390/metabo10010018>
- Kerley, M. S., & Allee, G. L. (2003). Modifications in soybean seed composition to enhance animal feed use and value: Moving from a dietary ingredient to a functional dietary component. *AgBioForum*, 6(1-2), 14–17.
- Kim, H.-K., Kang, S.-T., Cho, J.-H., Choung, M.-G., & Suh, D.-Y. (2005). Quantitative trait loci associated with oligosaccharide and sucrose contents in soybean (*Glycine max* L.). *Journal of Plant Biology*, 48(1), 106–112. <https://doi.org/10.1007/BF03030569>
- Kim, H. K., Kang, S. T., & Oh, K. W. (2006). Mapping of putative quantitative trait loci controlling the total oligosaccharide and sucrose content of *Glycine max* seeds. *Journal of Plant Research*, 119(5), 533–538. <https://doi.org/10.1007/s10265-006-0004-9>
- Kumar, V., Rani, A., Goyal, L., Dixit, A. K., Manjaya, J. G., Dev, J., & Swamy, M. (2010). Sucrose and raffinose family oligosaccharides (RFOs) in soybean seeds as influenced by genotype and growing location. *Journal of Agricultural and Food Chemistry*, 58(8), 5081–5085. <https://doi.org/10.1021/jf903141s>
- Kwon, S., & Torrie, J. (1964). Heritability of and Interrelationship among traits of two soybean populations. *Crop Science*, 4, 196–198. <http://dx.doi.org/10.2135/cropsci1964.0011183X000400020023x>
- Le, H., Nguyen, N. H., Ta, D. T., Le, T. N. T., Bui, T. P., Le, N. T., Nguyen, C. X., Rolletschek, H., Stacey, G., Stacey, M. G., Pham, N. B., Do, P. T., & Chu, H. H. (2020). CRISPR/Cas9-mediated knockout of galactinol synthase-encoding genes reduces raffinose family oligosaccharide levels in soybean seeds. *Frontiers in Plant Science*, 11, 612942. <https://doi.org/10.3389/fpls.2020.612942>
- Lee, S. H., Bailey, M. A., Mian, M. A. R., Carter, T. E., Ashley, D. A., Hussey, R. S., Parrott, W. A., & Boerma, H. R. (1996). Molecular markers associated with soybean plant height, lodging, and

- maturity across locations. *Crop Science*, 36(3), 728–735. <https://doi.org/10.2135/cropsci1996.0011183X003600030035x>
- Li, Y. S., Du, M., Qi, H., Zhang, Q. Y., Wang, G. H., Liu, X. B., & Hashemi, M. (2012). Greater differences exist in seed protein, oil, total soluble sugar and sucrose content of vegetable soybean genotypes [*Glycine max* (L.) Merrill] in Northeast China. *Australian Journal of Crop Science*, 6(12), 1681–1686. [https://www.researchgate.net/profile/Yansheng\\_Li2/publication/265728874\\_Greater\\_differences\\_exist\\_in\\_seed\\_protein\\_oil\\_total\\_soluble\\_sugar\\_and\\_sucrose\\_content\\_of\\_vegetable\\_soybean\\_genotypes\\_Glycine\\_max\\_L\\_Merrill\\_in\\_Northeast\\_China/links/541a306c0cf2218008b](https://www.researchgate.net/profile/Yansheng_Li2/publication/265728874_Greater_differences_exist_in_seed_protein_oil_total_soluble_sugar_and_sucrose_content_of_vegetable_soybean_genotypes_Glycine_max_L_Merrill_in_Northeast_China/links/541a306c0cf2218008b)
- Liu, K. (1997). *Soybeans: Chemistry, technology, and utilization*. Chapman and Hall.
- Lopez, M. A., Freitas Moreira, F., & Rainey, K. M. (2021). Genetic relationships among physiological processes, phenology, and grain yield offer an insight into the development of new cultivars in soybean (*Glycine max* L. Merr). *Frontiers in Plant Science*, 12, 651241. <https://doi.org/10.3389/fpls.2021.651241>
- Maestri, D. M., Guzmán, G. A., & Giorda, L. M. (1998). Correlation between seed size, protein and oil contents, and fatty acid composition in soybean genotypes. *Grasas y Aceites*, 49, 450–453. <https://dialnet.unirioja.es/servlet/articulo?codigo=1977456>
- Manzanilla-Pech, C. I. V., Gordo, D., Difford, G. F., Løvendahl, P., & Lassen, J. (2020). Multitrait genomic prediction of methane emissions in Danish Holstein cattle. *Journal of Dairy Science*, 103(10), 9195–9206. <https://doi.org/10.3168/jds.2019-17857>
- Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34(3), 1436–1462. <https://doi.org/10.1214/009053606000000281>
- Misztal, I., Tsuruta, S., Lourenco, D., Aguilar, I., Legarra, A., & Vitezica, Z. (2015). *BLUPF90 family of programs*. University of Georgia. [http://nce.ads.uga.edu/wiki/lib/exe/fetch.php?media=blupf90\\_all2.pdf](http://nce.ads.uga.edu/wiki/lib/exe/fetch.php?media=blupf90_all2.pdf)
- Moreira, F. F., Hearst, A. A., Cherkauer, K. A., & Rainey, K. M. (2019). Improving the efficiency of soybean breeding with high-throughput canopy phenotyping. *Plant Methods*, 15, 1–9. <https://doi.org/10.1186/s13007-019-0519-4>
- Murtagh, F., & Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: Which algorithms implement ward's criterion? *Journal of Classification*, 31(3), 274–295. <https://doi.org/10.1007/s00357-014-9161-z>
- Obendorf, R. L., & Górecki, R. J. (2012). Soluble carbohydrates in legume seeds. *Seed Science Research*, 22(4), 219–242. <https://doi.org/10.1017/S0960258512000104>
- Patil, G., Mian, R., Vuong, T., Pantalone, V., Song, Q., Chen, P., Shannon, G. J., Carter, T. C., & Nguyen, H. T. (2017). Molecular mapping and genomics of soybean seed protein: A review and perspective for the future. *Theoretical and Applied Genetics*, 130(10), 1975–1991. <https://doi.org/10.1007/s00122-017-2955-8>
- Patil, G., Vuong, T. D., Kale, S., Valliyodan, B., Deshmukh, R., Zhu, C., Wu, X., Bai, Y., Yungbluth, D., Lu, F., Kumpatla, S., Shannon, J. G., Varshney, R. K., & Nguyen, H. T. (2018). Dissecting genomic hotspots underlying seed protein, oil, and sucrose content in an inter-specific mapping population of soybean using high-density linkage mapping. *Plant Biotechnology Journal*, 16(11), 1939–1953. <https://doi.org/10.1111/pbi.12929>
- Qiu, D., Vuong, T., Valliyodan, B., Shi, H., Guo, B., Shannon, J. G., & Nguyen, H. T. (2015). Identification and characterization of a stachyose synthase gene controlling reduced stachyose content in soybean. *Theoretical and Applied Genetics*, 128(11), 2167–2176. <https://doi.org/10.1007/s00122-015-2575-0>
- R Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <http://www.r-project.org/>
- Recker, J. R., Burton, J. W., Cardinal, A., & Miranda, L. (2014). Genetic and phenotypic correlations of quantitative traits in two long-term, randomly mated soybean populations. *Crop Science*, 54(3), 939–943. <https://doi.org/10.2135/cropsci2013.07.0447>
- Reich, P. B., Wright, I. J., Cavender-Bares, J., Craine, J. M., Oleksyn, J., Westoby, M., & Walters, M. B. (2003). The evolution of plant functional variation: Traits, spectra, and strategies. *International Journal of Plant Sciences*, 164(S3), S143–S164. <https://doi.org/10.1086/374368>
- Reif, J. C., Melchinger, A. E., & Frisch, M. (2005). Genetical and mathematical properties of similarity and dissimilarity coefficients applied in plant breeding and seed bank management. *Crop Science*, 45(1), 1–7. <https://doi.org/10.2135/cropsci2005.0001>
- Saldívar, X., Wang, Y.-J., Chen, P., & Hou, A. (2011). Changes in chemical composition during soybean seed development. *Food Chemistry*, 124(4), 1369–1375. <https://doi.org/10.1016/j.foodchem.2010.07.091>
- Sibbald, I. R. (1980). Metabolizable energy in poultry nutrition. *Bioscience*, 30(11), 736–741. <https://doi.org/10.2307/1308333>
- Smith, H. F. (1936). A discriminant function for plant selection. *Annals of Eugenics*, 7(3), 240–250. <https://doi.org/10.1111/j.1469-1809.1936.tb02143.x>
- Song, Q., Hyten, D. L., Jia, G., Quigley, C. V., Fickus, E. W., Nelson, R. L., & Cregan, P. B. (2013). Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. *PLoS ONE*, 8(1), 54985. <https://doi.org/10.1371/journal.pone.0054985>
- Soy Stats. (2020). *US Soy Crop Statistics*. <http://soystats.com/2011/Default-frames.htm>
- Sun, J., Rutkoski, J. E., Poland, J. A., Crossa, J., Jannink, J.-L., & Sorrells, M. E. (2017). Multitrait, random regression, or simple repeatability model in high-throughput phenotyping data improve genomic prediction for wheat grain yield. *The Plant Genome*, 10(2). <https://doi.org/10.3835/plantgenome2016.11.0111>
- Tian, C., Yang, J., Zeng, Y., Zhang, T., Zhou, Y., Men, Y., You, C., Zhu, Y., & Sun, Y. (2019). Biosynthesis of raffinose and stachyose from sucrose via an in vitro multienzyme system. *Applied and Environmental Microbiology*, 85(2), AEM.02306–18. <https://doi.org/10.1128/AEM.02306-18>
- Truong, Q., Koch, K., Yoon, J. M., Everard, J. D., & Shanks, J. V. (2013). Influence of carbon to nitrogen ratios on soybean somatic embryo (cv. Jack) growth and composition. *Journal of Experimental Botany*, 64(10), 2985–2995. <https://doi.org/10.1093/jxb/ert138>
- United Soybean Board. (2021). *Soybean meal*. <https://www.unitedsoybean.org/topics/soybean-meal/>
- Valliyodan, B., Dan Qiu, Patil, G., Zeng, P., Huang, J., Dai, L., Chen, C., Li, Y., Joshi, T., Song, L., Vuong, T. D., Musket, T. A., Xu, D., Shannon, J. G., Shifeng, C., Liu, X., & Nguyen, H. T. (2016). Landscape of genomic diversity and trait discovery in soybean OPEN. *Scientific Reports*, 6, 23598. <https://doi.org/10.1038/srep23598>
- Vanraden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, 91(11), 4414–4423. <https://doi.org/10.3168/jds.2007-0980>

- Varona, L., Legarra, A., Toro, M. A., & Vitezica, Z. G. (2018). Non-additive effects in genomic selection. *Frontiers in Genetics*, 9, 78. <https://doi.org/10.3389/fgene.2018.00078>
- Wilcox, J. R., & Shibbles, R. M. (2001). Interrelationships among seed quality attributes in soybean. *Crop Science*, 41(1), 11–14. <https://doi.org/10.2135/cropsci2001.41111x>
- Xavier, A., Hall, B., Casteel, S., Muir, W., & Rainey, K. M. (2017). Using unsupervised learning techniques to assess interactions among complex traits in soybeans. *Euphytica*, 213(8), 1–18. <https://doi.org/10.1007/s10681-017-1975-4>
- Xavier, A., Thapa, R., Muir, W. M., & Rainey, K. M. (2018). Population and quantitative genomic properties of the USDA soybean germplasm collection. *Plant Genetic Resources: Characterisation and Utilisation*, 16(6), 513–523. <https://doi.org/10.1017/S1479262118000102>
- Xavier, A., Xu, S., Muir, W. M., & Rainey, K. M. (2015). NAM: Association studies in multiple populations. *Bioinformatics*, 31(23), 3862–3864. <https://doi.org/10.1093/bioinformatics/btv448>
- Zhang, B., Chen, P., Florez-Palacios, S. L., Shi, A., Hou, A., & Ishibashi, T. (2010). Seed quality attributes of food-grade soybeans from the U.S. and Asia. *Euphytica*, 173(3), 387–396. <https://doi.org/10.1007/s10681-010-0126-y>
- Zhang, J., Song, Q., Cregan, P. B., Nelson, R. L., Wang, X., Wu, J., & Jiang, G.-L. (2015). Genome-wide association study for flowering time, maturity dates and plant height in early maturing soybean (*Glycine max*) germplasm. *BMC Genomics*, 16(1), 1–11. <https://doi.org/10.1186/s12864-015-1441-4>
- Zhang, W.-K., Wang, Y.-J., Luo, G.-Z., Zhang, J.-S., He, C.-Y., Wu, X.-L., Gai, J.-Y., & Chen, S.-Y. (2004). QTL mapping of ten agronomic traits on the soybean (*Glycine max* L. Merr.) genetic map and their association with EST markers. *Theoretical and Applied Genetics*, 108(6), 1131–1139. <https://doi.org/10.1007/s00122-003-1527-2>
- Zhao, T., Liu, H., Roeder, K., Lafferty, J., & Wasserman, L. (2012). The huge package for high-dimensional undirected graph estimation in R. *Journal of Machine Learning Research*, 13, 1059–1062.

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Escamilla, D. M., Xavier, A., Vuong, T. D., Nguyen, H. T., & Rainey, K. M. (2023). An assessment of the interaction between sucrose content and seed quality traits in soybeans. *Crop Science*, 1–15. <https://doi.org/10.1002/csc2.21027>