Euphytica (2017)213:200
DOI 10.1007/s10681-017-1975-4

Using unsupervised learning techniques to assess interactions among complex traits in soybeans

Alencar Xavier · Benjamin Hall · Shaun Casteel · William Muir · Katy Martin Rainey

Received: 28 February 2016 / Accepted: 20 July 2017

Abstract Soybean yield components and agronomic traits are connected through physiological pathways that impose tradeoffs through genetic and environmental constraints. Our primary aim is to assess the interdependence of soybean traits by using unsupervised machine learning techniques to divide phenotypic associations into environmental and genetic associations. This study was performed on large scale, jointly analyzing 14 quantitative traits in a large multi-parental population designed for genetic studies. We collected phenotypes from 2012 to 2015 from a soybean nested association panel with 40 families of approximately 140 individuals each. Pearson and Spearman correlations measured phenotypic associations. A multivariate mixed linear model provided genotypic and environmental correlations. To evaluate relationships among traits, the study used principal component and undirected graphical models from phenotypic, genotypic, and environmental correlation matrices. Results indicate that high phenotypic correlation occurs when traits display both genetic and environmental correlations. In genetic terms, length of reproductive period, node number, and canopy coverage play important roles in determining yield potential. Optimal grain yield production occurs when the growing environment favors faster canopy closure and extended reproductive length. Environmental associations found among yield components give insight into the nature of yield component compensation. The use of unsupervised learning methods provides a good framework for investigating interactions among various quantitative traits and defining target traits for breeding.

Keywords Mixed models · SoyNAM · Unsupervised learning · Variance decomposition

Abbreviations

DAP	Days after planting
LASSO	Least absolute shrinkage and selection operator
MG	Maturity group
NAM	Nested association mapping
PCA	Principal component analysis
QTL	Quantitative trait loci
RIL	Recombinant inbred line
SNP	Single nucleotide polymorphism

Electronic supplementary material The online version of this article (doi:10.1007/s10681-017-1975-4) contains supplementary material, which is available to authorized users.

A. Xavier · B. Hall · S. Casteel · K. M. Rainey (✉)
Department of Agronomy, Purdue University, West Lafayette, IN 47907, USA
e-mail: krainey@purdue.edu

W. Muir
Department of Animal Science, Purdue University, West Lafayette, IN 47907, USA

Published online: 03 August 2017

Introduction

In the genetic and agronomic context, a tradeoff represents an interaction of multiple traits in which each trait provides a regulatory effect on others, striking a balance between the desirable and mutually incompatible states of traits. Identifying and managing tradeoffs among traits, such as grain yield and seasonal maturity, is challenging in the breeding and production of soybean (*Glycine max* L. Merr.) (Mansur et al. 1993; Chung et al. 2003). Whereas most studies focus on interactions among genotype, environment, and agronomic management (Concibido et al. 2003; Pedersen and Lauer 2004; Zhang et al. 2010; Board and Kahlon 2011; Hu 2011), few studies have investigated trait interactions in terms of the synergetic or antagonistic influence they have on each other. All traits interact to greater or lesser degrees through physiological pathways that involve tradeoffs imposed by genetic and environmental constraints (Recker et al. 2014). But a better understanding of trait interactions provides important insight for developing breeding and management strategies to overcome the grain yield limitations imposed by genetic and agronomic factors (Lynch and Walsh 1998; Panthee et al. 2005; Wortman et al. 2013).

Soybeans have an inferred grain yield potential of 8 Mg ha^{-1} (Specht et al. 1999). However the current production is only slightly over 4 Mg ha^{-1} (Rowntree et al. 2013; Rincker et al. 2014). To achieve full grain yield potential requires optimization of every yield-affecting biotic or abiotic factor (Carpenter and Board 1997), including a favorable environment, improved genetics, and optimized management practices. Increases in soybean grain yields are generally attributed either to seed quantity or seed size (Board and Kahlon 2011). While the contribution of seed size provides inconsistent results (Ball et al. 2000; Soares et al. 2013), seed quantity is considered the most reliable trait for yield improvement in soybeans (Sudaric et al. 2002). The measurement of seed quantity is seed.m^{-2}, which can be further divided into four subcomponents: plants.m^{-2}, nodes.$plant^{-1}$, pods.$node^{-1}$, and seeds.pod^{-1} (Lesoing and Francis 1999). The first subcomponent refers to the population density and is mostly determined by management practices and environmental conditions (Fehr et al. 1973) while genetic factors make a smaller contribution to germination and emergence (Spear and Fehr 2007). The three others; nodes.$plant^{-1}$, pods.$node^{-1}$, and seeds.pod^{-1}; along with seed weight, are known as yield components (Hu 2011). Yield components are inter-correlated and highly dependent on genetics, management, and environment.

Grain yield is, therefore, a composite trait, sensitive to interactions among its components (Board and Tan 1995; Board and Kahlon 2011; Recker et al. 2013, 2014) and to higher-order interactions among environment, management, and genetics (Carpenter and Board 1997; Yan and Rajcan 2003; Pedersen and Lauer 2004; Piepho et al. 2008). It is possible to allocate photosynthates to different yield components to provide yield compensation and stable production even when seasonal stresses occur during the reproductive period (Ball et al. 2000; Board 2000; Pedersen and Lauer 2004). But for such complex statistical problems, machine learning methods are necessary to extract patterns of association from data; in other words, to determine the directional relationship among the variables. Going a step further, we used a multi-trait mixed model to obtain genetic and environmental covariance components, such that we could apply machine learning techniques to infer phenotypic, genetic, and environmental associations.

Machine learning has two major classes: supervised and unsupervised (James et al. 2013). Supervised methods are used for modeling, prediction, and classification, problems for which a response variable exists. Unsupervised methods are used to cluster and to infer causality from an observed correlation when there are multiple variables, for example, interactions among agronomic traits and yield components. Unsupervised methods, such as principal component analysis (PCA), are commonly used to study tradeoffs or characterize interactive phenomena in agronomic studies (Ali et al. 2015) and to evaluate population structure in genetic studies (Jombart et al. 2010). PCA identifies patterns through the orthogonal transformations of relationship matrices. PCA projects all variable into a reduced dimensional space, where variables with similar properties are projected in the same direction and antagonistic variable are projected in opposite directions. Our study uses PCA to provide insight into tradeoffs with regard to the correlated response of traits to genetic improvement (i.e. breeding) and environmental stimuli (i.e. management).

However, breeding studies rarely exploit more sophisticated unsupervised methods to study complex interactions (Steinsland and Jensen 2010), such as undirected graphical models (Hastie et al. 2005). Undirected graphical models are used to infer causal associations, revealing the structure and dependence among variables as a network. Undirected graphical models are based on Markovian principles, in that they assume that variables that appear to be correlated are, in reality, independent when conditional to other variables.

The primary aim of this study is to assess the interdependence of soybean agronomic traits and yield components. It seeks to expose the mutually dependent state of complex traits evaluated from phenotypic, genotypic, and environmental correlations (Searle 1961) using a dataset collected over multiple years from a unprecedentedly large and diverse nested association mapping population. The study employs correlations to evaluate interactions and associations among agronomic traits and yield components using unsupervised learning approaches for multivariate analysis (Hastie et al. 2005), specifically, correlations, principal component analysis, and undirected graphical models.

Materials and methods

Genetic resources

The SoyNAM population (soynam.org) is a nested association mapping panel that comprises nearly 5600 recombinant inbred lines (RILs); including determinate, indeterminate, and semi-determinate genotypes from maturity groups (MG) ranging from late MG II to early MG IV. The population was derived from 40 biparental crosses sharing the cultivar IA3023 as the standard parent, containing approximately 140 RILs each. Of the other 40 founder parents, 17 lines are elite public germplasm from different regions, 15 have diverse ancestry, and eight are plant introductions (Diers 2014). The design of the SoyNAM population employs a diverse panel to dissect the genetic architecture of complex traits and to map yield-associated quantitative trait loci (QTL).

Lines were genotyped in the F_5 generation with the Illumina SoyNAM BeadChip SNP array (Song et al. 2017), designed for this population, which called 5305 single nucleotide polymorphism (SNP) markers from the genomic sequencing of the 41 parental lines. We imputed missing loci using random forest (Xavier et al. 2016) and removed SNPs with a minor allele frequency lower than 0.15 and redundant markers using the R package NAM (Xavier 2015). The project genotyped a total of 5555 lines and identified 196 lines among these that had high genomic similarity ($\geq$95% identical by state).

Experimental design

Phenotypic data was collected from the SoyNAM population in West Lafayette, Indiana in 2012, 2013, 2014, and 2015, as well as from a second location in 2015. The experiment used a unreplicated design from 2012 to 2014, where each family of approximately 140 individuals were split into four family blocks of 35 individuals. The entries within the family blocks were randomized, and the family blocks were randomly allocated into the experimental field. In 2015, the experiment was conducted as complete block design with two replications in two location.

Lines were planted on May 17, 2012; May 20, 2013; May 24, 2014; and May 23, 2015 at the Purdue University Agronomy Center for Research and Education (ACRE). The second growing site in 2015 was located at Throckmorton Purdue Agricultural Center where the experiment was planted on May 22. Experimental units were based on two-row plots (0.76 m $\times$ 2.90 m) at a density of approximately 35 plants.m^{-2}. All 6400 SoyNAM entries were grown from 2012 to 2014, and just families NAM5, NAM12, NAM15, NAM24 and NAM40 were grown in 2015. The experimental fields experienced partial drought in 2012, and the experiment located at ACRE in 2015 suffered mild flood damage.

Phenotypes

The traits under evaluation were grain yield, lodging score, seed weight, days to flowering, days to maturity, length of reproductive period, leaflet shape, pod number, node number, pods per node, internode length, plant height, average canopy closure and rate of canopy closure. The normality of the each trait was confirmed using the Shapiro–Wilk test (p *value* < 0.01) implemented in the built in R function *shapiro.test* (R Core Team 2016).

Phenotypic measurements were collected as follows. Grain yield was collected from 2012 to 2015 and measured in grams per plot adjusted to 0.13 g kg^{-1} seed moisture. Lodging was scored on a scale from 1 to 5 before harvest, where 1 indicates erect plants and 5 means all plants down. Seed weight, collected in 2012 and 2013, was measured as the mass of 100 seeds, determined by sampling and weighing 350 seeds.

Flowering and maturity were collected twice a week as days after planting (DAP), back and forward scoring plots that flowered and matured between the intervals. The criterion for a plot to achieve flowering (R1) and maturity (R8) was 50% of the plants with open flowers on the main stem while the criterion for maturity (R8) was 95% of pods mature (Fehr et al. 1971). The project collected flowering data in 2013 and 2014 and maturity data in all environments. We determined the length of the reproductive period by subtracting DAP to flowering from DAP to maturity.

Yield components were collected in two SoyNAM families in 2012, in all families in 2013 and 2014, and in six families from both locations in 2015. Counts included the number of reproductive nodes (i.e. nodes with at least one pod) and pods from the main stem during R7–R8 (first to full physiological maturity), measuring from three representative plants per plot in 2012 and 2013, from six representative plants in 2014, and from four representative plants in 2015. These counts also allowed assessment of pods per node.

We measured leaflet shape and plant height using a barcode ruler, measuring three plants per plot. In 2015, we collected plant height from four plants per plot with a regular ruler. We evaluated leaflet shape in 2013 and 2014, calculated as the ratio between length and width of the central leaflet from the fifth node from the top, thus higher values represent narrower leaflets. We surveyed plant height in all environments, measured as the distance from the base of the stem to the apical meristem and calculated internode length as the ratio between plant height and node number.

Canopy closure was collected in 2013 and 2014, measured weekly through ground-based images from the second week after emergence until flowering in accordance with the methods suggested by Hall (2015) and Purcell (2000). Digital image analysis provided two phenotypes, the average value of canopy coverage (%) across sampling dates, and the rate of canopy closure (% day^{-1}) as represented by the slope obtained from regressing canopy closure by days after planting.

Multivariate analysis

We derived associations among soybean agronomic traits and yield components from phenotypic, genetic, and environmental correlations. The statistical significance of correlation coefficients was inferred using single-tailed asymptotic t-statistics with $n-2$ degrees of freedom. Table 1 shows the number of pairwise observations in this study used to calculate correlations. After computing phenotypic, genetic and environmental correlations, we used two methods of unsupervised machine learning: principal component analysis and undirected graphical models.

Correlations

We calculated phenotypic correlations using pairwise Pearson correlation and Spearman correlation from phenotypic values normalized by environment. While Pearson correlation traditionally quantifies linear associations, Spearman correlation is a non-parametric measure that evaluates a monotonic function between variables based on their rank order, which is not necessarily linear. Simultaneous analysis of both types of correlations enables investigation of the nature of trait associations. We used the built-in functions in R to compute Pearson and Spearman correlations (R Core Team 2016).

We inferred genetic and environmental correlations from covariance components calculated through a multivariate mixed linear model computed in Bayesian framework (Sorensen and Gianola 2002). Markov chain Monte Carlo solved the model with the Gibbs sampler implemented in GIBBS3F90 (Misztal 2002). The model fit k traits simultaneously, with each trait in the linear model described by

$$\mathbf{y}_k = \mathbf{X}_k\mathbf{b}_k + \mathbf{Z}_k\mathbf{u}_k + \boldsymbol{\varepsilon}_k \quad (1)$$

where $\mathbf{y}$ is the vector of observations of the kth trait, $\mathbf{X}_k$ and $\mathbf{Z}_k$ are the incidence matrices of fixed effects and random effect (i.e. genotypes), $\mathbf{b}_k$ is the vector of regression coefficient of fixed effects that accommodate the terms imposed by the experimental design, $\mathbf{u}_k$ is the polygenic effect associated with each line, and $\boldsymbol{\varepsilon}_k$ is the residual term.

Table 1 Number of times that each pairwise combination of traits was observed together

Trait	Yld	Flo	Mat	Rep	Hgt	Ldg	Acc	Rcc	LSh	Node	Pod	P/N	SW	Int
Yld	**15,643**	9992	15,638	9990	15,640	11,082	11,061	11,059	11,096	11,331	11,331	11,331	10,058	11,331
Flo	–	**10,005**	10,000	10,003	10,002	9993	9970	9968	10,005	10,005	10,005	10,005	4426	10,005
Mat	–	–	**19,012**	10,001	19,009	14,451	11,070	11,068	11,105	14,700	14,700	14,700	10,063	14,700
Rep	–	–	–	**10,004**	10,001	9994	9969	9967	10,004	10,004	10,004	10,004	4424	10,004
Hgt	–	–	–	–	**19,014**	14,449	11,072	11,070	11,107	14,702	14,702	14,702	10,065	14,702
Ldg	–	–	–	–	–	**14,452**	11,060	11,058	11,095	14,452	14,452	14,452	5518	14,452
Acc	–	–	–	–	–	–	**11,075**	11,073	11,075	11,075	11,075	11,075	5529	11,075
Rcc	–	–	–	–	–	–	–	**11,073**	11,073	11,073	11,073	11,073	5528	11,073
LSh	–	–	–	–	–	–	–	–	**11,110**	11,110	11,110	11,110	5529	11,110
Node	–	–	–	–	–	–	–	–	–	**14,705**	14,705	14,705	5762	14,705
Pod	–	–	–	–	–	–	–	–	–	–	**14,705**	14,705	5762	14,705
P/N	–	–	–	–	–	–	–	–	–	–	–	**14,705**	5762	14,705
SW	–	–	–	–	–	–	–	–	–	–	–	–	**10,065**	5762
Int	–	–	–	–	–	–	–	–	–	–	–	–	–	**14,705**

Main diagonal represents the total number of observations for each trait (*bold*)

Yld grain yield, *Flo* flowering, *Mat* maturity, *Rep* length of reproductive period, *Hgt* plant height, *Ldg* lodging score, *Acc* average canopy coverage, *Rcc* rate of canopy closure, *LSh* leaflet shape, *Node* number of reproductive nodes, *Pod* pods in the main stem, *P/N* pods per node, *SW* 100-seed weight, *Int* internode length

The multivariate covariances of the model are described as

$$\begin{aligned}\mathbf{U} &\sim \mathrm{MVN}(0, \mathbf{G} \otimes \mathbf{\Sigma_a}) \\ \mathbf{E} &\sim \mathrm{MVN}(0, \mathbf{I} \otimes \mathbf{\Sigma_e})\end{aligned} \quad (2)$$

where $\mathbf{U} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u_k}\}$ represents the set of regression coefficients for all traits, $\mathbf{\Sigma_a}$ is the genomic covariance matrix that contains the variance of the traits in the diagonal and the pairwise genetic covariance among traits in the off diagonal. In similar manner, $\mathbf{E} = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e_k}\}$ represents the set of residuals and $\mathbf{\Sigma_e}$ the residuals covariance matrix, which informs about the pairwise environmental covariance among traits. $\mathbf{G}$ defines the additive genomic relationship matrix built as described by VanRaden (2008)

$$\mathbf{G} = \frac{(\mathbf{M} - \mathbf{P})(\mathbf{M} - \mathbf{P})'}{2\sum_j p_j(1 - p_j)} \quad (3)$$

where $\mathbf{M}$ is the genotypic matrix where rows represent the individuals and columns represent the markers. The matrix $\mathbf{P}$ adjust the markers by the allele frequencies. The denominator is the normalizing factor computed as the sum of the loci variances, based on binomial distribution of alleles frequencies (p).

Genetic correlations were estimated from the covariance components of the additive genetic term, and environmental correlations from the residual covariances. Trait heritabilities (narrow-sense) were computed as

$$h^2 = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2} \quad (4)$$

Genetic, environmental and phenotypic correlation tables are illustrated as heatmaps in the supplementary material.

Principal components

The study computed principal components as the Eigenvectors of each correlation matrix corresponding to phenotypic (Pearson and Spearman), genetic, and environmental correlations. We used the R built-in function *eigen* for the Eigendecomposition (R Core Team 2016). In this, an axis represents each soybean trait and interpretation of PCA is based on the length and direction of the axes. Variables with similar properties are projected in the same direction while antagonistic variables are projected in opposite directions. Therefore, PCA provides directionality and an indication of the tradeoffs observed in the phenotype and imposed by genetic and environmental causes. PCA of genetic correlations provides an insight into the genetic tradeoffs faced by soybean breeding programs aiming to improve multiple traits simultaneously (Ecochard and Ravelomanantsoa 1982). Likewise, PCA of environmental correlations gives insight into how agronomic practices could optimize productivity by using management to change the environment in which plants grow.

Graphical models

Graphical models were computed from the genetic, environmental and phenotypic correlation matrices to infer the connection among traits as networks. We used a Gaussian undirected graphical model based on neighborhood selection with the least absolute shrinkage and selection operator (LASSO) algorithm as proposed by Meinshausen and Bühlmann (2006) and implemented by Zhao et al. (2012).

The use of the Meinshausen-Bühlmann algorithm in this study aims to generate sparsity among variables by minimizing the LASSO loss function, which provides a robust but not necessarily unique network. Graphical models, also known as Gaussian Markov random fields (Rue and Held 2005), are commonly used to generate networks for the identification of patterns of relationships (Pellet and Elisseeff 2008). This approach is especially useful when all variables (in this case the soybean traits) are highly correlated but frequently show conditional independence (Hastie et al. 2005).

Results

Correlation analyses

Table 2 presents phenotypic correlations in terms of Pearson and Spearman coefficients. The phenotypic correlations express the product of multiple interactions between genetics and environment through the observed phenotype. When Pearson and Spearman correlations have similar values, it indicates that trait relationships work mostly in linear fashion, while non-linear associations occur in cases where the Spearman correlation is greater than Pearson. For example, the model infers the correlation between lodging and yield

Table 2 Phenotypic correlation: Pearson's correlation (upper-right diagonal) and Spearman's correlation (lower-left diagonal)

Trait	Yld	Flo	Mat	Rep	Hgt	Ldg	Acc	Rcc	LSh	Node	Pod	P/N	SW	Int
Yld	–	−0.06***	0.31***	0.31***	0.13***	0.01	0.31***	0.13***	0.12***	0.20***	0.18***	0.06***	0.07***	−0.06***
Flo	−0.05***	–	0.21***	−0.53***	0.19***	0.07***	−0.01	0.02*	−0.06***	−0.001	−0.04***	−0.06***	0.05**	0.13***
Mat	0.30***	0.30***	–	0.59***	0.42***	0.17***	0.18***	0.05***	−0.003	0.21***	0.10***	−0.07***	0.03**	0.15***
Rep	0.41***	−0.24***	0.75***	–	0.21***	0.10***	0.13***	0.03***	0.02*	0.22***	0.13***	−0.01	−0.02	0.01
Hgt	0.12***	0.23***	0.40***	0.30***	–	0.35***	0.44***	0.25***	−0.05***	0.34***	0.28***	−0.01	−0.02**	0.42***
Ldg	0.03**	0.05***	0.18***	0.13***	0.38***	–	0.30***	0.21***	−0.13***	0.19***	0.19***	0.07***	0.002	0.11***
Acc	0.30***	−0.01	0.18***	0.17***	0.43***	0.31***	–	0.53***	−0.13***	0.30***	0.24***	0.06***	0.09***	0.09***
Rcc	0.12***	0.05***	0.06***	0.03**	0.24***	0.23***	0.50***	–	−0.05***	0.21***	0.14***	0.02*	0.03*	0.05***
LSh	0.15***	−0.03**	−0.001	−0.003	−0.05***	−0.14***	−0.11***	−0.04***	–	−0.03***	−0.03**	−0.01	−0.03*	−0.03**
Node	0.20***	−0.01	0.23***	0.30***	0.39***	0.24***	0.29***	0.20***	−0.05***	–	0.51***	−0.03***	−0.01	−0.27***
Pod	0.18***	−0.05***	0.10***	0.18***	0.28***	0.21***	0.23***	0.15***	−0.01	0.60***	–	0.78***	-0.06***	−0.20***
P/N	0.07***	−0.06***	−0.06***	−0.02*	0.02*	0.08***	0.06***	0.03**	0.02**	0.03***	0.77***	–	−0.06***	−0.04***
SW	0.08***	0.08***	0.05***	0.01	−0.02	0.01	0.10***	0.02*	−0.05***	−0.02	−0.06***	−0.06***	–	0.06***
Int	−0.05***	0.16***	0.15***	0.01	0.43***	0.12***	0.10***	0.05***	−0.03**	−0.32***	−0.21***	−0.04***	0.06***	–

Yld grain yield, *Flo* flowering, *Mat* maturity, *Rep* length of reproductive period, *Hgt* plant height, *Ldg* lodging score, *Acc* average canopy closure, *Rcc* rate of canopy closure, *LSh* leaflet shape, *Node* number of reproductive nodes, *Pod* pods in the main stem, *P/N* pods per node, *SW* 100-seed weight, *Int* internode length

* Significant at the 0.05 probability level

** Significant at the 0.01 probability level

*** Significant at the 0.001 probability level

to be non-linear because it is only significant in the Spearman correlation.

In phenotypic terms, yield appears mostly correlated to maturity, length of reproductive period, average canopy closure, and reproductive nodes (Table 2), which supports the relevance of these traits in both genetic improvement and agronomic management practices when aiming to increase yield. However, whether any yield improvement attained in this case would be the result of genetics, management, or both depends on the strength of the genetic and environmental correlations.

Table 3 presents trait heritabilities, and genetic and environmental correlations. Heritabilities ranged from 0.40 to 0.88, where seed weight was the least heritable trait and plant height was the most heritable trait. The supplementary material contains heritability data for the individual families. Genetic correlations ranged from −0.54 to 0.90, and environmental correlations ranged from −0.64 to 0.78. All traits were significantly correlated to yield in both genetic and environmental terms. Environmental correlations might be deflated in this study due to similar growing conditions across environments, whereby most discrepancies are likely due to microenvironmental variation (Vieira and Paz-Gonzalez 2003).

Multidimensional and graphical associations

Figure 1 presents the result of the principal components biplot. In the multidimensional plane, the overlap of the axes of phenotypic principal components shows a strong phenotypic association between yield and reproductive period (Fig. 1a, b), a trend also observed in both genetic and environmental analysis (Fig. 1c, d). This indicates that there are strong phenotypic associations when traits display both genetic and environmental associations.

Traits including lodging, days to maturity, plant height, rate, and average canopy closure appear strongly associated in genetic terms (Fig. 1c). Yield overlaps with length of reproductive period both in terms of direction and magnitude. In this PCA biplot, yield is located between two clusters of traits, one with yield components and a second with canopy traits, lodging, maturity, and height. This trend indicates that genetic enhancement of these traits favors grain yield. This information could be exploited through approaches such as index selection or indirect selection.

Flowering, seed size, and internode length appear as a cluster of traits in the phenotypic and genetic biplots (Fig. 1a, c), and leaflet shape seems unconnected to any cluster but shows negative effects on plant height and maturity. In environmental terms (Fig. 1d), however, leaflet shape appears correlated to flowering and seed size in all instances, and internode length is negatively associated with the following yield components: pods, nodes, and pods per node. The remaining yield component, seed size, is positively associated with internode length. However, it displays the shortest axis in all cases, indicating poor influence of this trait over others, which we attribute to an overall lack of variation in seed size among the population. Figure 1d shows that yield appears in a cluster of agronomic traits that have strong overlap, including reproductive length, canopy traits, lodging, height, and maturity.

Figure 2 shows undirected graphical models. This analysis identifies nodes or 'bubbles' of interdependent traits (Pellet and Elisseeff 2008). Since phenotypic interactions are rooted in genetic and environmental causes, when nodes of interactions occur in the phenotypic networks they are also likely to appear in either the genetic or environmental networks, or both, according to the original nature of the interaction.

Discussion

Heritability, genetics, and environment

Heritabilities (Table 3) indicate the effectiveness of breeding for individual traits. Nonetheless, the heritability of individual traits may act upon the magnitude of the genetic and environmental relationships between traits (Searle 1961; Crabbe et al. 1990), consequently affecting breeding value estimates and the efficacy of selection for multiple traits (Hazel 1943; Falconer 1952). Associations between traits can create distinct patterns in genetic and environmental terms (Searle 1961). Yet the concurrence of high genetic and environmental associations often relates to high phenotypic associations. Genetic association among traits can be interpreted as a measure of pleiotropy (Sorensen and Gianola 2002; Ramachandra

Table 3 Genetic correlation (upper-right diagonal), environmental correlation (lower-left diagonal) and heritabilities (main diagonal, *bold letters*)

Trait	Yld	Flo	Mat	Rep	Hgt	Ldg	Acc	Rcc	LSh	Node	Pod	P/N	SW	Int
Yld	**0.63**	−0.29***	0.69***	0.80***	0.55***	0.50***	0.73***	0.53***	0.08***	0.58***	0.44***	0.15***	0.09***	0.08***
Flo	−0.05***	**0.70**	0.21***	−0.54***	0.39***	0.32***	0.04***	−0.07***	−0.33***	−0.07***	−0.12***	−0.21***	0.13***	0.42***
Mat	0.34***	0.13***	**0.82**	0.71***	0.86***	0.71***	0.61***	0.29***	−0.14***	0.47***	0.19***	−0.17***	0.21***	0.49***
Rep	0.25***	−0.64***	0.54***	**0.72**	0.45***	0.38***	0.50***	0.30***	0.10***	0.48***	0.26***	−0.01	0.11***	0.08***
Hgt	0.27***	0.13***	0.47***	0.16***	**0.88**	0.89***	0.77***	0.52***	−0.29***	0.39***	0.22***	−0.07***	0.21***	0.67***
Ldg	0.09***	0.03**	0.23***	0.11***	0.35***	**0.66**	0.83***	0.65***	−0.42***	0.57***	0.45***	0.15***	0.07***	0.41***
Acc	0.36***	−0.06***	0.18***	0.13***	0.46***	0.29***	**0.73**	0.90***	−0.36***	0.54***	0.43***	0.17***	0.21***	0.31***
Rcc	0.20***	−0.01	0.06***	0.05***	0.21***	0.14***	0.50***	**0.60**	−0.32***	0.38***	0.30***	0.12***	0.16***	0.20***
LSh	0.10***	−0.05***	−0.02*	0.01	−0.06***	−0.15***	−0.15***	−0.03***	**0.59**	−0.03**	−0.04***	−0.04***	−0.08***	−0.27***
Node	0.22***	−0.003	0.22***	0.15***	0.36***	0.22***	0.31***	0.19***	−0.05***	**0.82**	0.83***	0.38***	−0.07***	−0.42***
Pod	0.20***	−0.04***	0.10***	0.10***	0.20***	0.19***	0.23***	0.09***	−0.03**	0.63***	**0.84**	0.83***	−0.23***	−0.48***
P/N	0.08***	−0.06***	−0.06***	0.002	−0.05***	0.06***	0.05***	−0.02**	0.003	0.00	0.78***	**0.75**	−0.32***	−0.41***
SW	0.02*	−0.10***	−0.07***	−0.07***	−0.01	−0.03*	0.04**	−0.01	0.001	−0.05***	−0.04**	−0.02	**0.40**	0.27***
Int	0.05***	0.12***	0.23***	0.03***	0.57***	0.12***	0.14***	0.04***	−0.01	−0.54***	−0.38***	−0.06***	0.03*	**0.85**

Yld grain yield, *Flo* flowering, *Mat* maturity, *Rep* length of reproductive period, *Hgt* plant height, *Ldg* lodging score, *Acc* average canopy closure, *Rcc* rate of canopy closure, *LSh* leaflet shape, *Node* number of reproductive nodes, *Pod* pods in the main stem, *P/N* pods per node, *SW* 100-seed weight, *Int* internode length

* Significant at the 0.05 probability level

** Significant at the 0.01 probability level

*** Significant at the 0.001 probability level

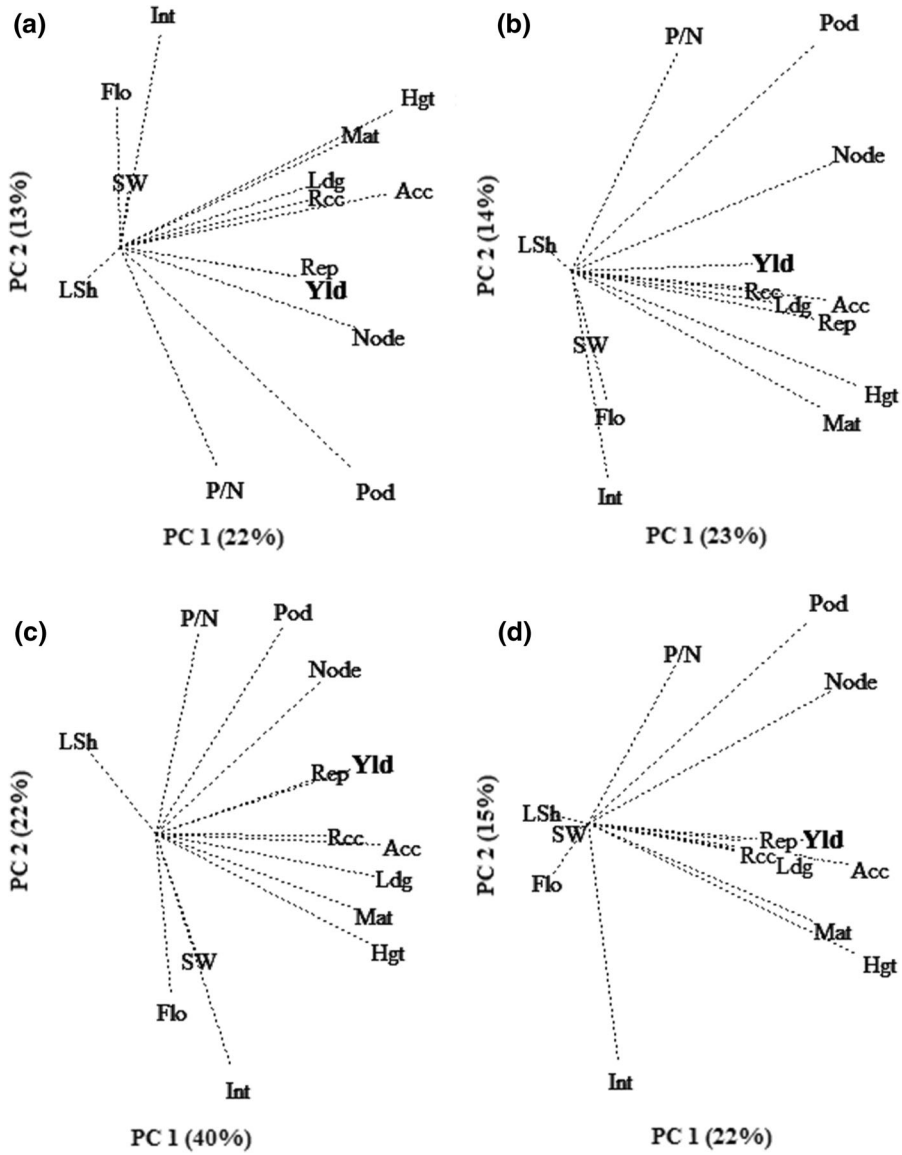

Fig. 1 Principal component analysis of **a** phenotypic Pearson, **b** phenotypic Spearman, **c** genetic, and **d** environmental correlations of soybean traits. The variation explained by each principal component is presented in parenthesis. Traits: grain yield (*Yld*, *bold*), flowering (*Flo*), maturity (*Mat*), length of reproductive period (*Rep*), plant height (*Hgt*), lodging (*Ldg*), average canopy closure (*Acc*), rate of canopy closure (*Rcc*), leaflet shape (*LSh*), node number (*Node*), pod number (*Pod*), pods per node (*P/N*), seed weight (*SW*), and internode length (*Int*)

et al. 2015), whereas environmental associations are indicators of the joint response of traits to environmental factors.

Along with trait heritability, additive genetic correlation estimates are relevant for soybean breeding to determine the indirect response of traits to selection (Kwon and Torrie 1964; Recker et al. 2014). This is important in cases of genetic interdependency among multiple traits for which extra attention to tradeoffs is necessary to improve target traits (Johnson et al. 1955; Herbert and Litchfield 1982; Board et al. 1997). There can be a discrepancy in the correlations among traits between the genetic and environmental terms (Searle 1961) as observed in node number and pods per node (Table 3). As such, some studies focus exclusively on the environmental relationship among traits (Ordas et al. 2008). Such studies work best when exposing genetic panels to various management practices or distinct growing environments (Falconer 1952), thereby imposing strong environmental stimuli to allow for investigation of the joint response of traits to higher-order interactions.

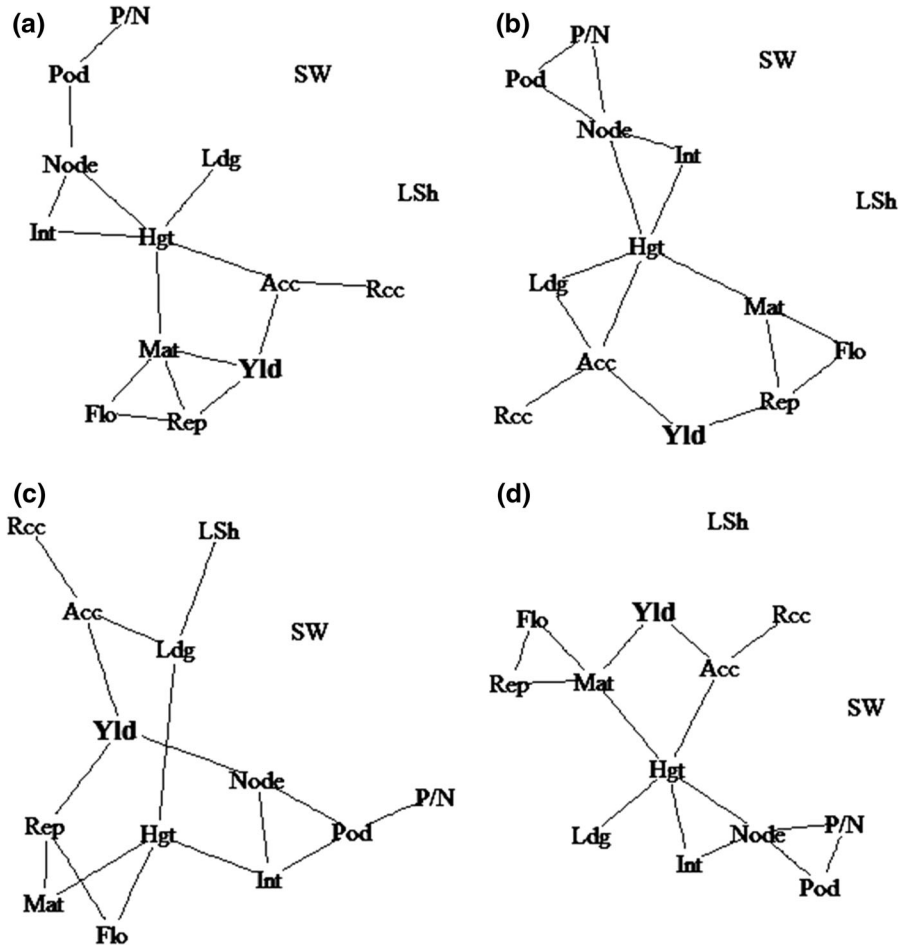

Fig. 2 Graphical modeling of **a** phenotypic Pearson, **b** phenotypic Spearman, **c** genetic, and **d** environmental correlations using the graphical LASSO of Meinshausen and Bühlmann (2006). Soybean traits include grain yield (*Yld*, *bold*), flowering (*Flo*), maturity (*Mat*), length of reproductive period (*Rep*), plant height (*Hgt*), lodging (*Ldg*), average canopy closure (*Acc*), rate of canopy closure (*Rcc*), leaflet shape (*LSh*), node number (*Node*), pod number (*Pod*), pods per node (*P/N*), seed weight (*SW*), and internode length (*Int*)

Canopy coverage and rate

In all graphical models (Fig. 2), we found a relevant relationship in the connection between yield and the canopy traits of average coverage and rate of closure. This indicates that canopy development, along with reproductive period, are the traits likely to have the most impact on grain yield. These traits have the potential for exploitation through agronomic practices and for genetic improvement through plant breeding (Xavier at al. 2017).

Yield and canopy traits were linked in all graphical analyses (Fig. 2), which is commonly attributed to an increase in light interception in the canopy that causes a positive balance in the source-sink ratio (Wells 1991; Board and Harville 1993). Thus, more energy captured across the growing season results in stronger sources of the photosynthates allocated to grain yield (Board and Tan 1995; Board et al. 1997; Purcell 2000). From the agronomic standpoint, increased light interception during the vegetative stages (i.e. prior to flowering) results in an increased number of nodes (Board et al. 1992) and pods (Board and Tan 1995), whereas stresses associated with light interception during the reproductive period (R1–R7); such as shadowing, defoliation, and cloud cover; mostly reduce yield by decreasing the number of pods per reproductive node (Board et al. 1997).

Genetic gains in soybean grain yield potential have been attributed to increases in the radiation intercepted by the plant canopy (Board and Kahlon 2012; Koester et al. 2014), increases in the growth rate, and in the net assimilation rate (Dornhoff and Shibles 1970; Gay et al. 1980; Larson et al. 1981; Frederick et al. 1989; Board and Kahlon 2011). The improvement of canopy traits seems to be one of the most feasible strategies to increase the source capacity in soybean (Richards 2000; Borrás et al. 2004; Ramachandra 2015).

Associations with yield

The traits with the highest genetic correlations to grain yield were reproductive period, maturity, average canopy coverage, and reproductive nodes on the main stem. Except for maturity, these traits were also the traits genetically connected to yield in the genetic graphical model (Fig. 2c). Due to their high heritability, they are promising traits for breeders to exploit for improvement of grain yield potential. It will soon be feasible to phenotype canopy coverage, flowering, and maturity on a large scale using new phenomic technologies, such as images from unmanned aerial systems (Ghanem 2014; Giglioti et al. 2015); however, there is still no high-throughput phenotyping method to determine node number.

Breeders may perform indirect selection for a complex trait via its components; this is known as trait dissection and is a common strategy to improve grain yield potential (Paterson 1995; Cui et al. 2008; Board and Kahlon 2011). In our study, most agronomic traits and yield components display positive genetic correlations to grain yield (Table 3). Our estimation of heritabilities and genetic correlations also provides breeders with valuable insights for indirect selection to improve soybean yield potential.

For a same selection intensity ($i_x = i_y$), the indirect selection efficiency (E), based on the correlated response to selection (CR) in respect to direct response to selection (R), is a function of the narrow-sense heritability of primary (h_y^2) and secondary (h_x^2) traits, and the additive genetic correlation (ρ_{xy}), described by the equation

$$E = \frac{CR}{R} = \frac{h_x^2 \rho_{xy}}{h_y^2}. \quad (5)$$

In this study, we observed that yield is moderately heritable, while length of reproductive period is more heritable (0.716) and highly correlated to yield (0.798). As such, indirect selection of yield through the length of reproductive period ($h_x^2 \rho_{xy} = 0.716 \times 0.798 = 0.571$) is almost as effective as selecting for yield itself ($h_y^2 = 0.632$), corresponding to an indirect selection efficiency of $E = 0.571/0632 = 0.903$. Although that would suggest breeding for earlier flowering and later maturity, delays in maturity are undesirable. Instead, indirect selection for yield through the average canopy coverage ($CR = h_x^2 \rho_{xy} = 0.726 \times 0.729 = 0.529$) does not present that tradeoff (Hall 2015) while providing a relatively efficient indirect selection ($E = 0.529/0.632 = 0.837$).

We found the traits most environmentally correlated to grain yield to be maturity and average canopy coverage, followed by plant height, reproductive period, and node number (Table 3). In environmental terms, the strong associations between canopy closure and yield (shown in Figs. 1, 2) indicate that management practices for faster canopy closure play a role in increasing these traits together (Board and Kahlon 2012; Kahlon and Board 2012). Wells (1991) described that the combination of population density and row spacing have direct influence on how fast the canopy closes. Early closure increases growth rate during the vegetative and early reproductive periods, increasing the reproductive nodes per area (Board et al. 1992). Likewise, changes to soybean phenological stages are controlled by photoperiod and temperature (Board and Hall 1984; Cober et al. 2001). Thus, adjusting the planting date allows management of the number of days to flowering and maturity by enhancing the reproductive window, allowing more time for node production prior to flowering (Rowntree et al. 2013). In addition, faster canopy closure combined with an extended reproductive period may be particularly beneficial by providing greater light interception during grain fill periods.

Environmental associations to soybean grain yield potential are relevant for agronomic practices because farmers maximize production by providing the most favorable environment for development and growth. Management practices that have been reported to influence agronomic traits and yield components include planting date (Board et al. 1997; Pedersen and Lauer 2004; Rowntree et al. 2013), density and row spacing (Wells 1991; Board et al. 1992; DeBruin and Pedersen 2008; Epler and Staggenborg 2008), application of chemical inputs (Swoboda and Pedersen 2009), crop rotation (Lesoing and Francis 1999), irrigation (El-Mohsen et al. 2013), tillage (Elmore 1990; Frederick et al. 2001; Pedersen and Lauer 2004), and fertilizer application (Wilson et al. 2014). However, physiological traits, plant architecture, source capacity, and sink strength are, in general, not manageable at the agronomic level (Ramachandra 2015).

Associations among yield components

Despite significant positive correlations in both Spearman and Pearson correlations, yield components do not seem directly connected to yield in the phenotypic graphical model (Fig. 2a, b). However, this association is observed in the genetic network (Fig. 2c) and in the phenotypic and genetic principal components (Fig. 1a–c). Among the yield components, reproductive nodes is the trait most correlated to grain yield (Table 2), one which has been described as a yield indicator from the physiological standpoint because it shares a genetic basis with yield (Simpson and Wilcox 1983; Zhang et al. 2004) and has a similar response to a variety of stresses (Board and Harville 1993; Board and Tan 1995; Board et al. 1997).

While there have been many consensus QTL for agronomic traits reported in the past two decades (Hu 2011), it is notable that few genetic studies have examined yield components or their interaction (Board and Kahlon 2011). Nevertheless, the heritability and genetic control of complex traits such as yield arises from combinations of simpler and more heritable traits (Mansur et al. 1993). The idea of decomposing soybean yield into more heritable traits is not new, but it has not been exploited (Johnson et al. 1955; Kwon and Torrie 1964; Ecochard and Ravelomanantsoa 1982). Past reports have indicated that the number of pods per node is a yield estimator based on its genetic associations, since it is less sensitive to environmental stimuli (Board and Tan 1995; Board et al. 1997). In agreement with the literature, Table 3 shows that the association we found between pods per node and yield was almost twice as large in genetic terms as it was in environmental terms.

In agreement with Board et al. (1997), the phenotypic graphical model in Fig. 2a, b indicates that pods per node and pod number are directly connected. In the Pearson correlation of phenotypes (Fig. 2a), pod number appears to be the link between pods per node and reproductive nodes, showing that, in linear terms, these two traits are conditionally independent in terms of their observable phenotypes.

The fact that the phenotypic correlation of pod number with grain yield is weaker than that of reproductive nodes with yield could arise from the indirect effect of pods on branches as an alternative allocation of resources (Herbert and Litchfield 1982; Frederick et al. 2001; Zera and Harshman 2001). Kahlon and Board (2012) also reported similar results.

There is an interdependency among pods, nodes, and pods per node (Fig. 2b, d). The three-way interaction among yield components we observed in the Spearman correlations and environmental networks indicates that yield component compensation is not linear and occurs at environmental levels. Similarly, Malausa et al. (2005) found that yield component compensation occurs in response to environmental factors. This interaction among yield components represents a mechanism of yield compensation at the pod level (Ball et al. 2000) that confers physiological flexibility to seed production (Ball et al. 2000; Board 2000; Pedersen and Lauer 2004). This was also captured by the path analysis presented by Board et al. (1997).

Phenotypic plasticity of seed production in soybean is an evolutionary mechanism occurring when suboptimal growing conditions affect a specific yield component, thereby triggering another yield component to compensate the grain yield (Peirson 2015). Genotypes with extreme values for any given yield component may have a compromised compensation ability by losing the plasticity afforded by alternative allocation of resources (DeJong and VanNoordwijk 1992). The plasticity to attain full grain yield potential is intrinsic to the plant's physiological response to environmental stimuli (Zera and Harshman 2001), which makes it easier to exploit through agronomic management.

Certain yield components, such as seeds per pod and pods per node, are less sensitive to environmental stresses and management (Board et al. 1997), while number of nodes.m^2 is the cause of reductions in grain yield during biotic and abiotic stresses, reducing the number of pods and consequently the number of seeds per m^2 (Herbert and Litchfield 1982; Pedersen and Lauer 2004; Board and Kahlon 2011). Board and Tan (1995) described the improvement of pods per node as a breeding strategy less sensitive to environmental factors and thus promoting yield stability. Both correlations (Table 3) and PCA (Fig. 1d) indicated a weak environmental association between yield and number of pods on the main stem, suggesting that environmental stimuli affect the number of pods on branches.

Associations in agronomic traits

Principal component analysis indicated a strong association between maturity, height, and lodging (Fig. 1a–d); connections that all the networks also captured (Fig. 2a–d). In contrast, graphical models indicated that maturity and lodging are conditionally independent. Associations among these three agronomic traits have been reported to have both morphological and physiological origins (Wilcox and Sediyama 1981; Lee et al. 1996a, b; Mansur et al. 1996). High values of phenotypic correlation like those shown in Table 2, occur in traits with a physiological role (DeJong and VanNoordwijk 1992) that share genetic and environmental origins.

Maturity showed a high genetic correlation to plant height, flowering, and length of reproductive period, similar to results reported by Wu et al. (2015). Reports suggest these agronomic traits share a similar genetic basis, possibly related to growth habit (Lee et al. 1996a, b; Mansur et al. 1996), and that they are relevant to yield, protein, and oil seed content (Simpson and Wilcox 1983). Height, maturity, and lodging are moderately to highly correlated with reproductive nodes and average canopy coverage in phenotypic, genetic, and environmental terms (Tables 2, 3), which supports the idea that agronomic traits also indirectly affect grain yield through their effect on canopy development.

Over the years, soybean breeding has improved grain yield while keeping maturity constant (Ustun et al. 2001; Jin et al. 2010). Because of the strong relationship between the length of the reproductive period and yield, there is also a tradeoff in soybean breeding regarding grain yield and maturity. A possible solution to overcome this relationship is to focus on traits that do not give tradeoffs, such as the number of pods on the main stem and pods per node, as suggested by Board and Kahlon (2011). These two traits are also genetically correlated to yield (Table 3) without sharing a genetic basis with maturity, height, and lodging, with which they form a 90° angle in the PCA (Fig. 1) and lack connection in the graphical models (Fig. 2).

Maturity has a moderate genetic association with grain yield within the SoyNAM maturity range (II–IV). Similar results have been reported in random mating populations (Recker et al. 2014). Patterns in the Pearson phenotypic graphical model (Fig. 2a) and environmental model (Fig. 2d) indicate a direct phenotypic association between maturity and grain yield, which could be attributed to environmental causes, or the indirect effect of maturity on the length of the reproductive period, or both. Our results indicate that it is possible to genetically improve grain yield and maturity independently, supporting other studies achieving similar yields across different maturity groups (Egli 1993; Edwards and Purcell 2005).

Leaflet shape

Leaflet shape did not display even moderate values ($\geq$30%) of correlation with most traits (Table 2), it was not connected to any trait through any graphical model (Fig. 2) and it did not have a large magnitude in the principal component analysis (Fig. 1). This agrees with the results reported by Mandl and Buss (1981) and Mansur et al. (1996). Many traits were significantly correlated to leaflet shape, but results from PCA and graphical models indicated a lack of causation.

We found the strongest phenotypic correlations (Spearman) with leaflet shape to be with yield (0.151) and lodging (−0.141). This association with yield might be due to the contribution of light interception (Board and Kahlon 2012). The negative genetic associations with lodging and height though may be attributed to the existence of genetic material in the SoyNAM population with diverse pedigrees that is prone to be taller, to lodge, and to have round leaves (Rincker et al. 2014; Wu et al. 2015). In addition, we found the lanceolate gene *Ln* to be segregating in two families, NAM28 and NAM34 (see supplementary material). Dinkins et al. (2002) reported *Ln* to increase the number of seeds per pod with some tradeoffs with regard to other yield components.

Conclusions

Soybean grain yield improvement associates with a variety of agronomic traits and yield components including pod number, pods per node, flowering, and maturity (Hu 2011; Palomeque et al. 2009a, b; Kahlon and Board 2011, 2012; Wu et al. 2015). In this study we identified patterns of association among soybean traits that could provide insight into the tradeoffs

imposed by genetic and environmental factors, emphasizing associations that could lead to yield improvement. At the phenotypic level, we found the strength of associations to be a function of both genetic and environmental causes.

Days to maturity, length of the reproductive period, average canopy coverage and the number of reproductive nodes were the traits most correlated to yield at the phenotypic, environmental and genetic levels. Genetic correlations indicated that length of reproductive period, average canopy coverage, and reproductive nodes play an important role in yield improvement and could be targeted by soybean breeders, while maturity *per se* appears associated to yield through environmental factors so that it can be kept static as yield increases.

Environmental associations support that environmental forces are the driving factor of soybean yield plasticity (Zera and Harshman 2001; Pedersen and Lauer 2004). The strong environmental association of average canopy coverage and reproductive period with yield indicates that management practices that improve canopy coverage (i.e. row spacing and planting density) and extend reproductive period (i.e. early planting data) can enhance grain yield.

Funding United Soybean Board funded the SoyNAM experiment from 2012 to 2013. Dow AgroScience funded the SoyNAM experiment from 2014 to 2015 in Indiana, and the data collection of yield component data from 2013 to 2015.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Ali F, Kanwal N, Ahsan M, Ali Q, Bibi I, Niazi NK (2015) Multivariate analysis of grain yield and its attributing traits in different maize hybrids grown under heat and drought stress. Scientifica 2015:1–6

Ball RA, Purcell LC, Vories ED (2000) Short-season soybean yield compensation in response to population and water regime. Crop Sci 40(4):1070–1078

Board JE (2000) Light interception efficiency and light quality affect yield compensation of soybean at low plant populations. Crop Sci 40(5):1285–1294

Board JE, Hall W (1984) Premature flowering in soybean yield reductions at nonoptimal planting dates as influenced by temperature and photoperiod. Agron J 76(4):700–704

Board JE, Harville BG (1993) Soybean yield component responses to a light interception gradient during the reproductive period. Crop Sci 33(4):772–777

Board JE, Kahlon CS (2011) Soybean yield formation: what controls it and how it can be improved? Soybean Physiol Biochem. doi:10.5772/17596

Board JE, Kahlon CS (2012) A proposed method for stress analysis and yield prediction in soybean using light interception and developmental timing. Crop Management 11(1):22

Board JE, Tan Q (1995) Assimilatory capacity effects on soybean yield components and pod number. Crop Sci 35(3):846–851

Board JE, Kamal M, Harville BG (1992) Temporal importance of greater light interception to increased yield in narrow-row soybean. Agron J 84(4):575–579

Board JE, Kang MS, Harville BG (1997) Path analyses of the yield formation process for late-planted soybean. Agron J 91(1):128–135

Borrás L, Slafer GA, Otegui ME (2004) Seed dry weight response to source-sink manipulations in wheat, maize and soybean: a quantitative reappraisal. Field Crops Res 86(2):131–146

Carpenter AC, Board JE (1997) Branch yield components controlling soybean yield stability across plant populations. Crop Sci 37(3):885–891

Chung J, Babka HL, Graef GL, Staswick PE, Lee DJ, Cregan PB, Specht JE (2003) The seed protein, oil, and yield QTL on soybean linkage group I. Crop Sci 43(3):1053–1067

Cober ER, Stewart DW, Voldeng HD (2001) Photoperiod and temperature responses in early-maturing, near-isogenic soybean lines. Crop Sci 41(3):721–727

Concibido V, LaVallee B, Mclaird P, Pineda N, Meyer J, Hummel L, Wang J, Wu K, Delannay X (2003) Introgression of a quantitative trait locus for yield from *Glycine soja* into commercial soybean cultivars. Theor Appl Genet 106(4):575–582

Crabbe JC, Phillips TJ, Kosobud A, Belknap JK (1990) Estimation of genetic correlation: interpretation of experiments using selectively bred and inbred animals. Alcohol Clin Exp Res 14(2):141–151

Cui S, He X, Fu S, Meng Q, Gai J, Yu D (2008) Genetic dissection of the relationship of apparent biological yield and apparent harvest index with seed yield and yield related traits in soybean. Crop Pasture Sci 59:86–93

DeBruin JL, Pedersen P (2008) Soybean seed yield response to planting date and seeding rate in the Upper Midwest. Agron J 100(3):696–703

DeJong G, VanNoordwijk AJ (1992) Acquisition and allocation of resources: genetic (co) variances, selection, and life histories. Am Nat 139(4):749–770

Diers, B.W., 2014. SoyNAM Project Update. Soybean Breeders Workshop, St. Louis MO. http://soybase.org/meeting_presentations/soybean_breeders_workshop/SBW_2014/presentations/Diers_SBW2014.pdf

Dinkins RD, Keim KR, Farno L, Edwards LH (2002) Expression of the narrow leaflet gene for yield and agronomic traits in soybean. J Hered 93(5):346–351

Dornhoff GM, Shibles RM (1970) Varietal differences in net photosynthesis of soybean leaves. Crop Sci 10(1):42–45

Ecochard R, Ravelomanantsoa Y (1982) Genetic correlations derived from Full-sib relationships in soybean (Glycine max Merr.). Theor Appl Gen 63(1):9–15

Edwards JT, Purcell LC (2005) Soybean yield and biomass responses to increasing plant population among diverse maturity groups. Crop Sci 45(5):1770–1777

Egli DB (1993) Cultivar maturity and potential yield of soybean. Field Crops Res 32(1):147–158

El-Mohsen AAA, Mahmoud GO, Safina SA (2013) Agronomical evaluation of six soybean cultivars using correlation and regression analysis under different irrigation regime conditions. J Plant Breed Crop Sci 5(5):91–102

Elmore RW (1990) Soybean cultivar response to tillage systems and planting date. Agron J 82(1):69–73

Epler M, Staggenborg S (2008) Soybean yield and yield component response to plant density in narrow row systems. Crop Manag. doi:10.1094/CM-2008-0925-01-RS

Falconer DS (1952) The problem of environment and selection. Am Nat 86(830):293–298

Fehr WR, Caviness CE, Burmood DT, Pennington JS (1971) Stage of development descriptions for soybeans, *Glycine max* (L. Merrill). Crop Sci 11(6):929–931

Fehr WR, Burris JS, Gilman NA (1973) Soybean emergence under field conditions. Agron J 65(5):740–742

Frederick JR, Alm DM, Hesketh JD (1989) Leaf photosynthetic rates, stomatal resistances, and internal CO_2 concentrations of soybean cultivars under drought stress. Photosynthetica 23(4):575–584

Frederick JR, Camp CR, Bauer PJ (2001) Drought-stress effects on branch and mainstem seed yield and yield components of determinate soybean. Crop Sci 41(3):759–763

Gay S, Egli DB, Reicosky DA (1980) Physiological aspects of yield improvement in soybeans. Agron J 72(2):387–391

Ghanem ME, Marrou H, Sinclair TR (2014) Physiological phenotyping of plants for crop improvement. Trends Plant Sci 20:139–144

Giglioti ÉA, Sumida CH, Canteri MG (2015) Disease phenomics. Phenomics. Springer, Berlin, pp 101–123

Hall B (2015) Quantitative characterization of canopy coverage in the genetically diverse soybean population. M.Sc. Thesis, Department of Agronomy, Purdue University

Hastie T, Tibshirani R, Friedman J, Franklin J (2005) The elements of statistical learning: data mining, inference and prediction. Math Intell 27(2):83–85

Hazel LN (1943) The genetic basis for constructing selection indexes. Genetics 28(6):476–490

Herbert SJ, Litchfield GV (1982) Partitioning soybean seed yield components. Crop Sci 22(5):1074–1079

Hu G, Liu C, Jiang H, Wang J, Chen Q, Qi Z (2011) Integration of major QTLs of important agronomic traits in soybean. INTECH, Rijeka

James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning. Springer, New York

Jin J, Liu X, Wang G, Mi L, Shen Z, Chen X, Herbert SJ (2010) Agronomic and physiological contributions to the yield improvement of soybean cultivars released from 1950 to 2006 in Northeast China. Field Crops Res 115(1):116–123

Johnson HW, Robinson HF, Comstock RE (1955) Estimates of genetic and environmental variability in soybeans. Agron J 47(7):314–318

Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet 11(1):94

Kahlon CS, Board JE (2012) Growth dynamic factors explaining yield improvement in new versus old soybean cultivars. J Crop Improv 26(2):282–299

Koester RP, Skoneczka JA, Cary TR, Diers BW, Ainsworth EA (2014) Historical gains in soybean (*Glycine max* Merr.) seed yield are driven by linear increases in light interception, energy conversion, and partitioning efficiencies. J Exp Bot 65(12):3311–3321

Kwon SH, Torrie JH (1964) Heritability of and interrelationships among traits of two soybean populations. Crop Sci 4(2):196

Larson EM, Hesketh JD, Woolley JT, Peters DB (1981) Seasonal variations in apparent photosynthesis among plant stands of different soybean cultivars. Photosynth Res 2(1):3–20

Lee SH, Bailey MA, Mian MAR, Carter TE, Ashley DA, Hussey RS, Parrott WA, Boerma HR (1996a) Molecular markers associated with soybean plant height, lodging, and maturity across locations. Crop Sci 36(3):728–735

Lee SH, Bailey MA, Mian MAR, Shipe ER, Ashley DA, Parrott WA, Hussey RS, Boerma HR (1996b) Identification of quantitative trait loci for plant height, lodging, and maturity in a soybean population segregating for growth habit. Theor Appl Genet 92(5):516–523

Lesoing GW, Francis CA (1999) Strip intercropping effects on yield and yield components of corn, grain sorghum, and soybean. Agron J 91(5):807–813

Lynch M, Walsh B (1998) Genetics and analysis of quantitative traits, vol 1. Sinauer, Sunderland

Malausa T, Guillemaud T, Lapchin L (2005) Combining genetic variation and phenotypic plasticity in tradeoff modelling. Oikos 110(2):330–338

Mandl FA, Buss GR (1981) Comparison of narrow and broad leaflet isolines of soybean. Crop Sci 21(1):25–27

Mansur LM, Lark KG, Kross H, Oliveira A (1993) Interval mapping of quantitative trait loci for reproductive, morphological, and seed traits of soybean (*Glycine max* L.). Theor Appl Genet 86(8):907–913

Mansur LM, Orf JH, Chase K, Jarvik T, Cregan PB, Lark KG (1996) Genetic mapping of agronomic traits using recombinant inbred lines of soybean. Crop Sci 36(5):1327–1336

Meinshausen N, Bühlmann P (2006) High-dimensional graphs and variable selection with the lasso. Ann Stat 34:1436–1462

Misztal I, Tsuruta S, Strabel T, Auvray B, Druet T, Lee DH (2002) BLUPF90 and related programs (BGF90). In: Proceedings of the 7th World Congress on Genetics Applied to Livestock Production, August, 2002. Session 28. Institut National de la Recherche Agronomique (INRA), Montpellier, France, pp 1–2

Ordas B, Malvar RA, Hill WG (2008) Genetic variation and quantitative trait loci associated with developmental stability and the environmental correlation between traits in maize. Genet Res 90(5):385

Palomeque L, Li-Jun L, Li W, Hedges B, Cober ER, Rajcan I (2009a) QTL in mega-environments: I. Universal and specific seed yield QTL detected in a population derived

from a cross of high-yielding adapted x high-yielding exotic soybean lines. Theor Appl Genet 119(3):417–427
Palomeque L, Li-Jun L, Li W, Hedges B, Cober ER, Rajcan I (2009b) QTL in mega-environments: II. Agronomic trait QTL co-localized with seed yield QTL detected in a population derived from a cross of high-yielding adapted × - high-yielding exotic soybean lines. Theor Appl Genet 119(3):429–436
Panthee DR, Pantalone VR, West DR, Saxton AM, Sams CE (2005) Quantitative trait loci for seed protein and oil concentration, and seed size in soybean. Crop Sci 45(5):2015–2022
Paterson AH (1995) Molecular dissection of quantitative traits: progress and prospects. Genome Res 5(4):321–333
Pedersen P, Lauer JG (2004) Response of soybean yield components to management system and planting date. Agron J 96(5):1372–1381
Peirson BE (2015) Plasticity, stability, and yield: the origins of Anthony David Bradshaw's model of adaptive phenotypic plasticity. Stud Hist Philos Sci C 50:51–66
Pellet JP, Elisseeff A (2008) Using Markov blankets for causal structure learning. J Mach Learn Res 9:1295–1342
Piepho HP, Möhring J, Melchinger AE, Büchse A (2008) BLUP for phenotypic selection in plant breeding and variety testing. Euphytica 161(1–2):209–228
Purcell LC (2000) Soybean canopy coverage and light interception measurements using digital imagery. Crop Sci 40(3):834–837
R Core Team (2016) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/
Ramachandra D, Madappa S, Phillips J, Loida P, Karunanandaa B (2015) Breeding and biotech approaches towards improving yield in soybean. In: Davey MR, Daniell H, Azhakanandam K, Silverstone A (eds) Recent advancements in gene expression and enabling technologies in crop plants. Springer, New York, pp 131–192
Recker JR, Burton JW, Cardinal A, Miranda L (2013) Analysis of quantitative traits in two long-term randomly mated soybean populations: I. Genetic Variances. Crop Sci 53(4):1375–1383
Recker JR, Burton JW, Cardinal A, Miranda L (2014) Genetic and phenotypic correlations of quantitative traits in two long-term, randomly mated soybean populations. Crop Sci 54(3):939–943
Richards RA (2000) Selectable traits to increase crop photosynthesis and yield of grain crops. J Exp Bot 51(suppl 1):447–458
Rincker K, Nelson R, Specht J, Sleper D, Cary T, Cianzio SR, Diers B (2014) Genetic improvement of US soybean in maturity groups II, III, and IV. Crop Sci 54(4):1419–1432
Rowntree SC, Suhre JJ, Weidenbenner NH, Wilson EW, Davis VM, Naeve SL, Casteel SN, Diers BW, Esker PD, Specht JE, Conley SP (2013) Genetic gain x management interactions in soybean: I. Planting date. Crop Sci 53(3):1128–1138
Rue H, Held L (2005) Gaussian Markov random fields: theory and applications. CRC Press, Baco Raton
Searle SR (1961) Phenotypic, genetic and environmental correlations. Biometrics 17(3):474–480
Simpson AM, Wilcox JR (1983) Genetic and phenotypic associations of agronomic characteristics in four high protein soybean populations. Crop Sci 23(6):1077–1081
Soares MM, Oliveira GL, Soriano PE, Sekita MC, Sediyama T (2013) Performance of soybean plants as function of seed size: II. Nutritional stress. J Seed Sci 35(4):419–427
Song Q, Yan L, Quigley C, Jordan BD, Fickus E, Schroeder S, Song BH, Charles An YQ, Hyten D, Nelson R, Rainey KM, Beavis WD, Specht JE, Diers BW, Cregan P (2017) Genetic characterization of the soybean nested association mapping population. Plant Genome 10(2):1–14
Sorensen D, Gianola D (2002) Likelihood, Bayesian, and MCMC methods in quantitative genetics. Springer, New York
Spear JD, Fehr WR (2007) Genetic improvement of seedling emergence of soybean lines with low phytate. Crop Sci 47(4):1354–1360
Specht JE, Hume DJ, Kumudini SV (1999) Soybean yield potential: a genetic and physiological perspective. Crop Sci 39(6):1560–1570
Steinsland I, Jensen H (2010) Utilizing Gaussian Markov random field properties of Bayesian animal models. Biometrics 66(3):763–771
Sudaric A, Vrataric M, Duvnjak T (2002) Quantitative genetic analysis of yield components and grain yield for soybean cultivars. Poljoprivreda 2(8):11–15
Swoboda C, Pedersen P (2009) Effect of fungicide on soybean growth and yield. Agron J 101(2):352–356
Ustun A, Allen FL, English BC (2001) Genetic progress in soybean of the US Midsouth. Crop Sci 41(4):993–998
VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91(11):4414–4423
Vieira SR, Paz-Gonzalez A (2003) Analysis of the spatial variability of crop yield and soil properties in small agricultural plots. Bragantia 62(1):127–138
Wells R (1991) Soybean growth response to plant density: relationships among canopy photosynthesis, leaf area, and light interception. Crop Sci 31(3):755–761
Wilcox JR, Sediyama T (1981) Interrelationships among height, lodging and yield in determinate and indeterminate soybeans. Euphytica 30(2):323–326
Wilson EW, Rowntree SC, Suhre JJ, Weidenbenner NH, Conley SP, Davis VM, Diers BW, Naeve SL, Esker PD, Specht J, Casteel SN (2014) Genetic gain × management interactions in soybean: II. Nitrogen utilization. Crop Sci 54(1):340–348
Wortman SE, Francis CA, Galusha TD, Hoagland C, VanWart J, Baenziger PS, Johnson M et al (2013) Evaluating cultivars for organic farming: maize, soybean, and wheat genotype by system interactions in Eastern Nebraska. Agroecol Sust Food Syst 37(8):915–932
Wu T, Sun S, Wang C, Lu W, Sun B, Song X, Han T (2015) Characterizing changes from a century of genetic improvement of soybean cultivars in Northeast China. Crop Sci 55(5):2056–2067
Xavier A, Xu S, Muir WM, Rainey KM (2015) NAM: association studies in multiple populations. Bioinformatics 31:3862–3864
Xavier A, Muir WM, Rainey KM (2016) Impact of imputation methods on the amount of genetic variation captured by a

single-nucleotide polymorphism panel in soybeans. BMC Bioinform 17(1):17–55

Xavier A, Hall B, Hearst A, Cherkauer KA, Rainey KM (2017) Genetic architecture of phenomic-enabled canopy coverage in glycine max. Genetics 206(2):1081–1089

Yan W, Rajcan I (2003) Prediction of cultivar performance based on single-versus multiple-year tests in soybean. Crop Sci 43(2):549–555

Zera AJ, Harshman LG (2001) The physiology of life history trade-offs in animals. Annu Rev Ecol Syst 32:95–126

Zhang WK, Wang YJ, Luo GZ, Zhang JS, He CY, Wu XL, Chen SY 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(6):1131–1139

Zhang D, Cheng H, Wang H, Zhang H, Liu C, Yu D (2010) Identification of genomic regions determining flower and pod numbers development in soybean (*Glycine max* L.). J Genet Genom 37(8):545–556

Zhao T, Liu H, Roeder K, Lafferty J, Wasserman L (2012) The huge package for high-dimensional undirected graph estimation in R. J Mach Learn Res 13(1):1059–1062