8-2016

# Learning from data: Plant breeding applications of machine learning

Alencar Xavier
*Purdue University*

**PURDUE UNIVERSITY**
**GRADUATE SCHOOL**
**Thesis/Dissertation Acceptance**

This is to certify that the thesis/dissertation prepared

By Alencar Xavier

Entitled

LEARNING FROM DATA: PLANT BREEDING APPLICATIONS OF MACHINE LEARNING

For the degree of Doctor of Philosophy

Is approved by the final examining committee:

| | |
|---|---|
| Katy Martin Rainey<br>Chair | Bruce Craig |
| William Muir<br>Co-chair | |
| Shaun Casteel | |
| Tobert Rocheford | |

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University's "Policy of Integrity in Research" and the use of copyright material.

Approved by Major Professor(s): Katy Martin Rainey

Approved by: Joseph Anderson 6/7/2016

Head of the Departmental Graduate Program Date

# LEARNING FROM DATA:

# PLANT BREEDING APPLICATIONS OF MACHINE LEARNING

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Alencar Xavier

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

August 2016

Purdue University

West Lafayette, Indiana

# ACKNOWLEDGMENTS

I would like to thank my advisors Katy Martin Rainey and William Muir for the solid academic training, for the opportunity of having a scientific degree, for the motivations and for allowing me to expand my horizons. I also would like to express my gratitude to professors Shizhong Xu and William Beavis for the support and strong contributions to my theoretical background.

I am grateful to Chris Hoagland and Curtis Brackett for the availability and willingness to help, for the organization of the field experiments and data collection, support and enthusiasm. I want to thank Benjamin Hall for the sharing of ideas and extensive discussions about our joint research on soybean genetics.

I would like to thank Dow AgroSciences for funding the field experiments, and I am grateful to Erica Bakker, Maqsood Rehman and Sam Reddy for the important industry insight in my research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

# LIST OF ABBREVIATIONS

| | |
|---|---|
| ANOVA | Analysis of variance |
| BGS | Bayesian Gibbs sampling |
| BL / BLASSO | Bayesian LASSO |
| BLUP | Best linear unbiased predictor |
| BRR | Bayesian ridge regression |
| CI | Credibility intervals |
| DAP | Days after planting |
| EBV | Estimated breeding values |
| EM | Expectation Maximization |
| EMMA | Efficient mixed model association |
| GBLUP | Genomic BLUP |
| GBS | Genotyping-by-sequencing |
| GDV | Genomic direct values |
| GEBV | Genomic enhanced breeding values |
| GP | Gaussian process |
| GRAMMAR | Genome-wide association using mixed model and regression |
| GRM | Genomic relationship matrix |
| GS | Genomic selection |
| GSRU | Gauss-Seidel residual update |

| | |
|---|---|
| GWAS | Genome-wide association studies |
| GWP | Genome-wide prediction |
| HMM | Hidden Markov models |
| LASSO | Least absolute shrinkage and selection operator |
| LOD | Log of odds |
| LRT | Likelihood ratio test |
| MAF | Minor allele frequency |
| MAGIC | Multi-parent advanced generation intercross |
| MCMC | Markov chain Monte Carlo |
| MG | Maturity group |
| MRF | Markov Random Fields |
| NAM | Nested association mapping |
| Ne | Effective population size |
| NGP | Next-generation populations |
| OLS | Ordinary least square |
| P3D | Population parameter previously determined |
| PA | Prediction accuracy |
| PCA | Principal component analysis |
| QTL | Quantitative trait loci |
| R1 | Reproductive stage 1 (flowering) |
| R2 | Reproductive stage 2 (full bloom) |
| R3 | Reproductive stage 3 (pod development) |
| R4 | Reproductive stage 4 (fully developed pods) |

| | |
|---|---|
| R5 | Reproductive stage 5 (seed development) |
| R6 | Reproductive stage 6 (fully developed seeds) |
| R7 | Reproductive stage 7 (beginning to mature) |
| R8 | Reproductive stage 8 (fully mature) |
| REML | Restricted maximum likelihood |
| RIL | Recombinant inbred line |
| RKHS | Reproducing kernel Hilbert spaces |
| RR | Ridge regression |
| SNP | Single nucleotide polymorphism |
| SoyNAM | soybean nested association mapping population |
| SVM | Support vector machine |
| UMM | Unified mixed model |
| VCA | Variance component analysis |

ABSTRACT

Xavier, Alencar. Ph.D., Purdue University, August 2016. Learning from Data: Plant Breeding Applications of Machine Learning. Major Professor: Katy Martin Rainey.

Increasingly, new sources of data are being incorporated into plant breeding pipelines. Enormous amounts of data from field phenomics and genotyping technologies places data mining and analysis into a completely different level that is challenging from practical and theoretical standpoints. Intelligent decision-making relies on our capability of extracting from data useful information that may help us to achieve our goals more efficiently. Many plant breeders, agronomists and geneticists perform analyses without knowing relevant underlying assumptions, strengths or pitfalls of the employed methods. The study endeavors to assess statistical learning properties and plant breeding applications of supervised and unsupervised machine learning techniques. A soybean nested association panel (*aka.* SoyNAM) was the base-population for experiments designed *in situ* and *in silico*. We used mixed models and Markov random fields to evaluate phenotypic-genotypic-environmental associations among traits and learning properties of genome-wide prediction methods. Alternative methods for analyses were proposed.

# CHAPTER 1: PHENOTYPIC, GENETIC AND ENVIRONMENTAL ASSOCIATIONS AMONG SOYBEAN TRAITS

ABSTRACT

Soybean yield components and agronomic traits are connected through physiological pathways and tradeoffs are imposed by genetic and environmental constrains. The main goal of this study is to assess the interdependence of soybean traits by stratifying the phenotypic associations into environmental and genetic associations using unsupervised machine learning techniques. Phenotypic data was collected from 2012 to 2015 in West Lafayette, Indiana, from a soybean nested association panel containing 40 families. Phenotypic associations were measured by Pearson and Spearman correlations. Genotypic and environmental correlations were obtained through mixed model solved by MCMC. Relationships among traits were evaluated using principal component and undirected graphical models computed from phenotypic, genotypic and environmental correlation matrices. Results indicate that (1) high phenotypic correlation occurs when traits display simultaneously genetic and environmental correlations; (2) length of reproductive period, node number and average canopy closures could be further exploited by breeders to improve yield; (3) environmental associations indicate optimal yield production under growing conditions that favor faster canopy closure and extended reproductive length; and that (4) the nature of the yield compensation in soybeans was captured by environmental correlation among yield components.

## 1.1 Introduction

All traits are somehow connected through physiological pathways that imply tradeoffs imposed by genetic and environmental constrains (Recker et al. 2014). The understanding of these interactions is important to overcome yield limitations (Lynch and Walsh 1998) from both genetic and agronomic standpoint (Panthee et al. 2005, Wortman et al. 2013). Identifying and managing tradeoffs of traits such as yield, maturity and protein, is a major concern in soybean breeding and production (Mansur et al. 1993, Chung et al. 2003). Whereas most studies focus on interaction among genotypes, environment and management (Concibido et al. 2003, Pedersen and Lauer 2004, Zhang et al. 2010, Board and Kahlon 2011, Hu et al. 2011), few studies are dedicated to the investigation of interaction among traits.

Soybeans have an attainable yield of inferred 8 Mg/ha (Specht et al. 1999). To achieve high yield standards, an optimization of every yield-affecting biotic or abiotic factor is required (Carpenter and Board 1997), including a favorable environment, good genetic and proper management practices. Increases in soybeans yield are either associated to seed quantity or seed size (Board and Kahlon 2011). While the contribution of seed size has provided controversial results (Ball et al. 2000, Soares et al. 2013), seed quantity is considered the most reliable traits for yield improvement in soybeans (Sudaric et al. 2003). Seed quantity is measured in terms of seed.$m^{-2}$ and can be further divided into four subcomponents (Lesoing and Francis 1999), such as plants.$m^{-2}$, nodes.$plant^{-1}$, pods.$node^{-1}$ and seeds.$pod^{-1}$. The first factor refers to the population density and is most determined by management practices and environmental conditions (Fehr et al. 1973) with some contribution of genetic factors to germination and emergence (Spear and Fehr 2007). The three others, nodes.plant$^{-}$

$^{1}$, pods.node$^{-1}$ and seeds.pod$^{-1}$, are known as yield components along with seed weight (Hu et al. 2011). Thus, 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, Recker et al. 2014) and interactions among environment, management and genetics (Carpenter and Board 1997, Yan and Rajcan 2003, Pedersen and Lauer 2004, Piepho et al. 2008). Yield components can exchange resources (i.e., photosynthates) which confers yield compensation and stable production, even under seasonal stresses during the reproductive period (Ball et al. 2000, Board 2000, Pedersen and Lauer 2004).

A better understanding of these interactions is essential to learn about the tradeoffs that occur at physiological level (De Jong and Van Noordwijk 1992) and necessary to uncover new breeding and managements trends for yield improvement. The main goal of this study is to assess the interdependence of soybean agronomic traits and yield components through phenotypic, genotypic and environmental correlations. Connection and association among agronomic traits and yield components were evaluated from the correlations and investigated through unsupervised methods for multivariate analysis (Friedman et al. 2001), more specifically, principal component analysis and undirected graphical models.

## 1.2 Materials and Methods

### 1.2.1 Population

The SoyNAM population (soynam.org) is a nested association mapping panel that comprises nearly 5600 recombinant inbred lines (RILs), including determined, undetermined and semi-determined genotypes with maturity ranging from late MG II to

early MG IV, derived from 40 biparental populations. Each biparental population approximately contains 140 individuals and all families share IA3023 as standard parent. From the other 40 founder parents, 17 lines are elite public germplasm from different regions, 15 have diverse ancestry and 8 are plant introductions. The SoyNAM population was designed with the purpose of dissecting the genetic architecture of complex traits and mapping yield-related genes using a diverse panel.

SoyNAM represents a particularly useful population for genetic association analyses of agronomic traits, yield, and yield components, provided that genetic resources for yield improvement in soybean is mostly associated to exotic elite cultivars (Kabelka et al. 2004, Guzman et al. 2007, Palomeque et al. 2009a), to germplasm from different regions (Orf et al. 1999a 1999b, Reyna and Sneller 2001) and with diverse background (Concibido et al. 2003, Wang et al. 2004, Kim et al. 2012).

Lines were genotyped with a 5k SNPchip especially designed for these populations, where 5305 single nucleotide polymorphism (SNP) markers were called from the genomic sequencing of the parental lines. Missing loci were imputed using random forest (Stekhoven and Buhlmann 2012) and SNPs with minor allele frequency lower than 0.15 and redundant markers were removed. A total of 5555 lines were genotyped and 196 lines were identified as having high genomic similarity (≥95% identical). The computation of the quality control of genotypic data was performed using the R package NAM (Xavier et al. 2015).

### 1.2.2 Experimental design

Phenotypic data was collected from the SoyNAM population in 2012, 2013, 2014 and two locations in 2015 in West Lafayette, Indiana. The experiment was conducted as a modified

augmented design from 2012 to 2014 and as augmented complete block design in both location of 2015, with two replications each. Lines were planted May 17, 20, 24, and 23 in 2012, 2013, 2014 and 2015, respectively, at the Purdue University Agronomy Center for Research and Education (ACRE). The second growing site of 2015 was located at Throckmorton Purdue Agricultural Center where the experiment was planted on May 22. Experimental units were based on two-row plots, 0.76m × 2.90m, at a density of approximately 35 plants.$m^{-2}$. All 6400 SoyNAM entries were grown from 2012 to 2014 and just the six families with the highest mean and variance of yield components were grown in 2015. The experimental fields of 2012 and one location of 2015 were subject to partial drought and flood damage, respectively.

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 in a scale from 1 to 5 right before harvest, where one represents erect and five means all plants down. Seed size was collected in 2012 and 2013, measured in term of mass of 100 seeds, sampling and weighting 350 seeds.

Flowering and maturity were collected twice a week in terms of 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 and 95% of mature pods, respectively (Fehr et al. 1971). Flowering was collected in 2013 and 2014 and maturity in all environments. Length of the reproductive period was obtained 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 of 2015. Number of reproductive nodes (i.e., nodes with at least one pod) and pods from the main stem were counted during R7-R8 (first to full physiological maturity), measuring from 3 representative plants per plot in 2012 and 2013, 6 representative plants in 2014 and 4 representative plants in 2015. Pods per node were obtained by the ratio.

Leaflet shape and plant height were measured during R4-R5 (full pod to first seed) and R6-R7 (full pod to first physiological maturity), respectively, three plants per plot with a barcode ruler. In 2015, plant height was collected from four plants per plot with a regular ruler. Leaflet shape was collected in 2013 and 2014, calculated as the ratio between length and width of the central leaflet, thus higher values represent narrower leaflets. Plant height was collected in all environments and measured as the distance from the base of the stem to the apical meristem. Internode length was obtained by 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 to Hall (2015) and Purcell (2000). Two phenotypes were obtained from the digital image analysis, the average value of canopy closure (%) across sampling dates, and rate of canopy closure ($\%.day^{-1}$) as the slope from regressing canopy closure by days after planting. For the statistical analysis, observations of all traits were normalized by environment.

### 1.2.3 Multivariate analysis

Evaluation of associations among soybean agronomic traits and yield components were based on phenotypic, genetic and environmental correlations. Statistical significance of

correlation coefficients was inferred by single-tailed asymptotic t-statistics with $n-2$ degrees of freedom. The number of pairwise observations in this study used to calculate correlations is shown in Table 1. After computing phenotypic, genetic and environmental correlations, we used two methods of unsupervised machine learning to assess the correlations, principal component analysis (PCA) and undirected graphical models.

Phenotypic correlations were calculated though pairwise Pearson correlation and Spearman correlation. While Pearson correlation is traditionally used to quantify linear association, Spearman correlation is a non-parametric measure that evaluates a monotonic function between variable based on the rank order, which is not necessarily linear. Simultaneous analysis of both types of correlations enable the investigation of the nature of association. Pearson and Spearman correlations were computed by build-in functions in R (R Core Team 2015).

Genetic and environmental correlations were inferred from the covariance components calculated through a multivariate mixed linear model computed in Bayesian framework (Sorensen and Gianola 2002). The model fits $k$ traits simultaneously, for each traits the linear model is described by $y_k = X_k b_k + Z_k u_k + \varepsilon_k$, where $y$ is the vector of observations of the $k^{th}$ trait, $X_k$ and $Z_k$ are the incidence matrices of fixed effects and random effect (ie. genotypes), $b_k$ is the vector of regression coefficient of fixed effects, $u_k$ is the polygenic effect associated to each line and $\varepsilon_k$ is the residual term.

Regression coefficient of the random term are normally distributed $u_k \sim \mathrm{N}(0, A\sigma^2_{a_k})$, where $A$ is the relationship matrix and $\sigma^2_{a_k}$ is the additive genetic variance associated to the $k^{th}$ trait. Genetic correlations were based on the additive genetic term while environmental

correlation were computed from the residuals. Trait heritabilities were computed as $h^2 = \frac{\sigma_a^2}{\sigma_a^2+\sigma_e^2}$. The model was solved by MCMC with the Gibbs sampler implemented in GIBBS3F90 (Misztal et al. 2002) that uses genomic information to describe the genetic relationship among genotypes.

Principal component analysis (PCA) was used to identify patterns through the orthogonal transformations of relationship matrices, reducing the dimensionality of complex interactions for visual interpretation. Principal components were computed as the Eigenvectors of each correlation matrices corresponding to phenotypic (Pearson and Spearman), genetic and environmental correlations. We used the R build-in function *eigen* for the Eigendecomposition (R Core Team 2015). Each soybean trait is represented by an axis and the interpretation of PCA is based on the length and direction of the axes. Variables with similar properties are likely to be projected in the same direction while antagonistic variables would appear in opposite sense. In this study, PCA provides directionality and an indication of tradeoffs observed in the phenotype and imposed by genetic and environmental causes.

Undirected graphical models were required to analyze causal structure learning, in other words, the structure and dependence among soybean traits at phenotypic, genotypic and environmental level (Fig1). For this study we chose to use Gaussian 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 Meinshausen-Bühlmann algorithm used in this study aims to generate sparsity among variable by minimizing the LASSO loss function, which

provides a robust but not necessarily unique network. Graphical models, also known as Markov random fields, 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 conditionally independent (Friedman et al. 2001).

## 1.3 Results

### 1.3.1 Correlation analyses

Phenotypic correlations in terms of Pearson and Spearman coefficients is presented in Table 2. The phenotypic correlations express the product of multiple interactions among genetics and environment through the observed phenotype. Similar values between Pearson and Spearman correlations indicate that relationships work mostly in linear fashion, likewise non-linear association is observed in cases where Spearman correlation is greater than Pearson. For example, the correlation between lodging and yield is inferred as non-linear because it is only significant in the Spearman correlation.

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 for both breeding and management aiming to increase yield. However, whether the improvement should be associated to breeding or management (or both) depends on the strength of genetic and environmental correlations.

Genetic and environmental correlations are presented in Table 3. Genetic association among traits can be interpreted as a measure of pleiotropy (Sorensen and Gianola 2002, Ramachandra et al. 2015). Analysis of genetic correlations is relevant from the breeding stand point to determine the indirect response of traits to selection (Recker et al. 2014).

Genetic interdependency among traits imply that extra care is necessary for breeders to deal with tradeoffs (Johnson el at 1955, Herbert and Litchfield 1982, Board et al. 1997).

Environmental correlations may be deflated in this study due to the lack of environmental contrasts, where most discrepancies are due to field plot variation and macroenvironment (ie. year and location). The field plot variation, or microenvironmental variation, is due to naturally occurring soil variability, which has been reported to be a major source of yield variation in soybean (Vieira and Gonzalez 2003). This variation of soil properties has been reported to impact soybean growth, development, yield and yield components (Harper 1974, Sinclair 1986, Coale and Grove 1990, Board and Tan 1995, Gan et al. 2003, Malik et al. 2006, Pettigrew 2008, Fernández et al. 2009).

The number of pairwise environmental associations with statistical significances in Table 3 indicates that the existing field variability trigger sufficient environmental stimuli for the evaluation of environmental relationships. Some correlations between traits are even stronger in environmental terms than genetic terms, such as reproductive period with flowering and leaflet shape with yield. However, we recognize that the exposure of this population to distinct management practices could induce more environmental stimuli for the study of environmental relationship among traits, which would allow for studies of higher order interactions, such as genotype by environment by management.

### 1.3.2 Multidimensional and graphical associations

The result of the principal components biplot is presented in Figure 2. Together, the first two principal components explain 35%, 37%, 62% and 37% of the total variation for phenotypic Pearson and Spearman, genetic and environmental correlations, respectively. These relatively low values indicate interactions with high complexity among traits and the

use of additional principal components would be necessary to better represent the interactions among soybean traits. A three-dimensional version of principal component analysis is presented in the Figure 4.

Into the multidimensional plane, the overlap of the axis in phenotypic principal components shows a strong phenotypic association between yield and reproductive period (Fig.2 a-b) and a similar trend is observed in both genetic and environmental analysis (Fig.2 c-d), indicating that strong phenotypic associations are observed when traits display both genetic and environmental associations.

PCA of genetic correlation provides a good insight of genetic tradeoffs faced by breeding soybeans aiming to improve multiple traits simultaneously. Some traits appeared strongly associated in genetic terms (Fig.2c). Yield overlaps with length of reproductive period in terms of direction and magnitude. In this PCA biplot, yield is located between two clusters of traits, one with yield components and another with canopy traits, lodging, maturity and height. This trend indicates that the genetic enhancement of these traits are favorable to yield and this information could be exploited through approaches such as selection index or indirect selection.

Flowering, seed size and internode length appear as a cluster of traits in phenotypic and genetic biplots (Fig.2 a-c) and leaflet shape seems unconnected to any cluster but with negatively affecting plant height and maturity. Whereas in environmental terms appear correlated to flowering and seed size while internode length does not. In all instances, internode length is negatively associated to the yield components pods, nodes and pods per node. The remaining yield component, seed size, is positively associated to internode

length whereas it displays the shortest axis in all cases which indicates poor influence of this trait over the others.

Principal components of environmental correlations are relevant for better understanding how agronomic practices could optimize the productivity by changing the environment where plants grow through management. It is observed in Figure 2d that yield appears in a cluster of agronomic traits with strong overlap, including reproductive length, canopy traits, lodging, height and maturity.

Undirected graphical models are presented in Figure 3. This analysis can identify nodes or 'bubbles' of interdependent traits (Pellet and Elisseeff 2008). Since all phenotypic interaction are rooted into genetic and environmental causes, when nodes of interactions are observed in the phenotypic networks they are also likely to appear in either genetic or environmental network, or both, according to the original nature of the interaction.

## 1.4 Discussion

### 1.4.1 Canopy closure

A relevant relationship shown in all graphical models (Fig.3) is the connections between yield and canopy closure. Indicating that canopy closure along with reproductive period are likely to be the most impactful to yield, with potential to be exploited in agronomic practices and for the genetic improvement through plant breeding.

Yield and canopy closure traits were linked together in all graphical analysis (Fig.3) and that is commonly attributed to the increase in light interception (Wells 1991, Board and Harville 1993) that causes a positive balance in the source-sink ratio. Thus, more energy captured across the growing season reflects into stronger sources of photosynthates that

can be allocated into the grain yield (Board and Tan 1995, Board et al. 1997, Purcell 2000). From the agronomic standpoint, higher light interception during the vegetative stages (ie. prior to flowering) results in increased number of nodes (Board et al. 1992) and pods (Board and Tan 1995), whereas stresses associated to light interception during the reproductive period (R1-R7) mostly reduce yield through the number of pods per reproductive node (Board et al. 1997).

Genetic gains in soybean yield have been historically associated to intercepting more radiation by the plant canopy (Board and Kahlon 2012, Koester et al. 2014) and photosynthetic process associated to the canopy development, more specifically growth rate and 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 is considered one of the most feasible strategies to increase the source capacity in soybean (Richards 2000, Borrás et al. 2004, Ramachandra et al. 2015).

### 1.4.2 Associations with yield

The most genetically correlated traits to yield were reproductive period, maturity, average canopy closure and reproductive nodes on the main stem. Except for maturity, these traits were also the traits genetically connected to yield in genetic graphical model (Fig3c). High heritability of these traits also make them interesting targets for breeders to exploit for yield improvement. The feasibility of phenotyping canopy closure, flowering and maturity in large scale is expected from forthcoming phenomic technologies such as drone-based images (Ghanem et al. 2014, Giglioti et al. 2015), however node number still lacks in high-throughput phenotyping methods.

Breeders often perform indirect selection to complex trait by its subcomponents, so-called trait dissection. Trait dissection is a common strategy to improve yield (Paterson 1995, Cui et al. 2008, Board and Kahlon 2011) and, in fact, most agronomic traits and yield components display positive genetic correlation to yield (Table 3). Once heritabilities and genetic correlation are estimated, breeders have a valuable insight for indirect selection.

In this study, we observed that yield is moderately heritable and length of reproductive period is more heritable (0.716) and highly correlated to yield (0.798), 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 ($h_y^2 = 0.632$). However, that would imply in breeding for earlier flowering and later maturity but changes in maturity are usually undesirable in soybean breeding. Alternatively, the indirect selection for yield through the average canopy closure does not imply in any tradeoff and it is also represents a relatively efficient indirect selection ($h_x^2\rho_{xy} = 0.726 \times 0.729 = 0.529$).

The traits most environmentally correlated to yield were observed to be maturity and average canopy closure, followed by plant height, reproductive period and node number (Table 3). In environmental terms, the strong associations among canopy closure with yield shown in Figures 2 and 3 indicate that management practices for a faster canopy closure can play an important role to increase 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 reflects into increases in growth rate during vegetative and early reproductive periods, which results in reproductive nodes per area (Board et al. 1992). Likewise, changes on soybean phenological stage are controlled by photoperiod and temperature (Board and Hall 1984,

Cober et al. 2001). Thus planting date is used to manage the number of days to flowering and maturity by enhancing the reproductive window, which allowed more time for node production prior to flowering (Rowntree et al. 2014). In addition, faster canopy closure combined with extended reproductive period may be particularly beneficial to late planted soybeans and greater light interception during grain fill periods.

Environmental associations to yield are relevant for agronomic practices because, at farming level, the maximization of production is attained by providing soybean 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. 2014), density and row spacing (Wells 1991, Board et al. 1992, De Bruin 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 not manageable at agronomic level (Ramachandra et al. 2015).

### 1.4.3 Association among yield components

Despite the significant correlation in both Spearman and Pearson correlations, yield components do not seem directly connected to yield in the phenotypic graphical model (Fig.3 a-b). However, this association is observed in the genetic network (Fig.3c) and in the phenotypic and genetic principal components (Fig.2 a-c). Among the yield components, reproductive nodes has the highest correlation to yield (Table 2), and it has been described as good yield indicator from the physiological standpoint because it shares genetic basis

with yield (Simpson and Wilcox 1983, Zhang et al. 2004) and have similar response to different stresses (Board and Harville 1993, Board and Tan 1995, Board et al. 1997).

Many consensus QTL of agronomic traits have reported in the past two decades (Hu et al. 2011), but it is remarkable that few genetic studies were performed on yields components or their interaction (Board and Kahlon 2011). Yet, the heritability and genetic control of any complex traits, such as yield, is due to the combination 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). The number of pods per node has been reported as good yield estimators based on genetic associations, once it is less sensitive to environmental stimuli (Board and Tan 1995, Board et al. 1997). In accordance to the literature, Table 3 shows that the associated between pods per node and yield is almost twice as large in genetic terms than in environmental terms.

In agreement with Board et al. (1997), the phenotypic graphical model in Figure 3 (a-b) indicates that pods per node and pod number are directly connected. In the Pearson correlation of phenotypes (Fig.3a), pod number appears as the link between pods per node and reproductive nodes, showing these two traits as conditionally independent in terms of observable phenotype in linear terms.

The fact that the phenotypic correlation pods and yield is weaker than reproductive nodes and yield could be attributed to the indirect effect of branch pods as an alternative allocation of resources (Herbert and Litchfield 1982, Frederick et al. 2001, Zera and Harshman 2001), although similar results were also reported by Kahlon and Board (2012). Remarkably, seed size does not appear connected to any other yield component or agronomic trait in the graphical models (Fig.3) nor seems to impact yield or other traits on multidimensional plant

represented by principal components (Fig.2). Nevertheless, this negatively correlation to the other yield components is significant and it may suggest another possible mechanism of yield compensation (Table 2).

There exist an interdependency among pods, node and pods per node (Fig.3 b,d). The three-way interaction among yield components observed in Spearman and environmental networks supports that the compensation among yield components is not linear and occurs at environmental levels. Malausa et al. (2005) observed similar findings that yield compensation at yield components level would act mostly by environmental forces. This interaction among yield components can represent a mechanisms of yield compensation at pod level (Ball et al. 2000) that confers physiological flexibility to seed production (Ball et al. 2000, Board 2000, Pedersen and Lauer 2004), also captured by the path analysis presented by Board et al. (1997).

Genotypes with extreme values for any given yield component may have a compromised compensation ability by losing the plasticity of reallocating resources (De Jong and Van Noordwijk 1992). Yield plasticity is intrinsic to the physiological response to environmental stimuli (Zera and Harshman 2001) and hence can be better exploited from the agronomic standpoint.

Some 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 causative of yield drag 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 that would be stable across environments.

Environmental correlations (Table 3) and environmental PCA (Fig.2d) indicate weak association between yield and pods in the main stem, suggesting that environmental stimuli may affect the amount of pods located on branches.

1.4.4 Association in agronomic traits

Principal components analysis indicate a strong association between maturity, height and lodging (Fig.2 a-d), connection also captured by all networks (Fig.3 a-d), and nonetheless graphical models indicate that maturity and lodging are conditionally independent. Associations among these three agronomic traits have been reported to have both morphological and physiological origins with influence of growth habit (Wilcox and Sediyama 1981, Lee et al. 1996a 1996b, Mansur et al. 1996). High values of phenotypic correlation (Table 2) are observed in traits related physiological role (De Jong and Van Noordwijk 1992), often sharing genetic and environmental origins.

Maturity displays a high genetic correlation to plant height, flowering and length of reproductive period, similar to results reported by Wu et al. (2015). These agronomic traits have been also reported to share similar genetic basis possibly related to growth habit (Lee et al. 1996a 1996b, Mansur et al. 1996), and to be relevant to yield, protein and oil seed content (Simpson and Wilcox 1983). Height, maturity and lodging are moderately-high correlated to reproductive nodes and average canopy closure in phenotypic, genetic and environmental terms (Table 2 and 3), which supports that agronomic traits also indirectly affect yield through these two traits.

Over the years, soybean breeding has attempted improving grain yield while keeping maturity constant (Ustun et al. 2001, Jin et al. 2010). Because of the strong relationship between the length of reproductive period and yield, there exist a major tradeoff in soybean

breeding regards yield and maturity. A possible solution to overcome this issue is to focus on traits that do not imply in major 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 genetically correlated to yield (Table 3) without sharing genetic basis with maturity, height and lodging as shown be the $90^0$ angle in the PCA (Fig.2) and lack of connection in the graphical models (Fig.3).

Maturity has a moderate genetic association to yield within the SoyNAM maturity range (II to IV) and similar results were reported in random mating populations (Recker et al. 2014). Patterns in the Pearson phenotypic graphical model (Fig.3a) and environmental model (Fig.3d) indicate direct phenotypic association between maturity and yield, which could be attributed to environmental causes or through the indirect effect of maturity in length of reproductive period. Our results supports that yield and maturity could be genetically improved independently, supporting other studies where similar yield can be achieved across different maturity groups (Egli 1993, Edwards and Purcell 2005).

### 1.4.5 Leaflet shape

Leaflet shape does not display moderate values ($\geq 30\%$) of correlation to most traits (Table 2), it is not connected to any trait through any graphical model (Fig.3) and it does nod display large magnitude in the principal component anlysis (Fig.2), in accordance to the results reported by Mandl and Buss (1981) and Mansur et al. (1996). Many traits are significantly correlated to leaflet shape but results from PCA and graphical models indicate the lack of causation.

The strongest phenotypic correlations (Spearman) with leaflet shape were found to be with yield (0.151) and lodging (-0.141). The association to yield might due to the contribution

to light intercept (Board and Kahlon 2012). Stronger correlation were observed in genetic terms, where leaflet shape is negatively correlated to height, lodging and canopy closure traits. Higher values of leaflet shape indicate elongated or lanceolate leaves, thus, our data supports that round leaves are more related to canopy closure. The negative associations with lodging and height through genetics may be attributed to the existence of genetic material in the SoyNAM population with diverse background that is prone to be taller, lodge and have round leaves (Rincker et al. 2014, We et al. 2015) and, therefore, leaflet shape could be an indicator of diversity and less adapted background.

It has been observed that the association between leaflet shape and yield varies among families (data not shown), we speculate that is may be due to the existence of a major gene called *Ln* found to be segregating in some families. Further investigation in this subpopulations would be required for more consistent associations. Ln gene is known for increasing the number of seeds per pod, although tradeoff with other yield components has been reported (Dinkins et al. 2002).

## 1.5 Conclusions

Yield improvement has been associated to different agronomic traits and yield components over the years, including pod number, pods per node, flowering and maturity (Hu et al. 2011, Palomeque et al. 2009a, 2009b, Kahlon and Board 2011, 2012, Wu et al. 2015). In this study we attempted to identify patterns of association among soybean traits that could provide an insight of the tradeoffs imposed by genetics and environmental factors, emphasizing associations that could lead to yield improvement. At phenotypic level, the strength of associations was found to be a function of both genetic and environmental causes.

Days to maturity, length of reproductive period, average canopy closure and the number of reproductive nodes were the most correlated traits to yield at phenotypic, environmental and genetic level. The high genetic correlations to yield indicate that, length of reproductive period, average canopy closure and reproductive nodes have a great potential to be exploited by breeder, while maturity is more associated to yield through environmental factors and can be kept static as yield increases.

Environmental associations support that environmental forces may be the driving factor of soybean yield plasticity (Zera and Harshman 2001, Pedersen and Lauer 2004). The strong environmental association of average canopy closure and reproductive period with yield indicate that management practices that improve canopy closure (i.e., row spacing and planting density) and extend reproductive period (i.e., early planting date) can have a good potential to increase yield.

References

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

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

Board JE, Kahlon CS (2011) Soybean Yield Formation: What Controls It and How It Can Be Improved? Soybean Physiology and Biochemistry.

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

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

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

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

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

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

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 Research 86(2): 131-146.

Carpenter AC, Board JE (1997) Branch yield components controlling soybean yield stability across plant populations. Crop Science 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 science 43(3): 1053-1067.

Coale FJ, Grove JH (1990) Root distribution and shoot development in no-till full-season and double-crop soybean. Agronomy journal 82(3): 606-612.

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

Concibido V, La Vallee 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. Theoretical and Applied Genetics 106(4): 575-582.

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 and Pasture Science 59(1) 86-93.

De Bruin JL, Pedersen P (2008) Soybean seed yield response to planting date and seeding rate in the Upper Midwest. Agronomy Journal 100(3): 696-703.

De Jong G, Van Noordwijk AJ (1992) Acquisition and allocation of resources: genetic (co) variances, selection, and life histories. American Naturalist 139(4): 749-770.

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

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

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

Egli DB (1993) Cultivar maturity and potential yield of soybean. Field Crops Research 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. Journal of plant breeding and crop science 5(5): 91-102.

Elmore RW (1990) Soybean cultivar response to tillage systems and planting date. Agronomy Journal 82(1): 69-73.

Epler M, Staggenborg S (2008) Soybean yield and yield component response to plant density in narrow row systems. Crop Management: 7(1).

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

Fehr WR, Burris JS, Gilman, DF (1973) Soybean emergence under field conditions. Agronomy Journal 65(5): 740-742.

Fernández FG, Brouder SM, Volenec JJ, Beyrouty CA, Hoyum R (2009) Root and shoot growth, seed composition, and yield components of no-till rainfed soybean under variable potassium. Plant and soil 322(1-2): 125-138.

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

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.

Friedman J, Hastie T, Tibshirani R (2001) The elements of statistical learning (Vol. 1). Springer, Berlin: Springer series in statistics.

Gan Y, Stulen I, van Keulen H, Kuiper PJ (2003) Effect of N fertilizer top-dressing at various reproductive stages on growth, N 2 fixation and yield of three soybean (Glycine max (L.) Merr.) genotypes. Field Crops Research 80(2): 147-155.

Gay S, Egli DB, Reicosky DA (1980) Physiological aspects of yield improvement in soybeans. Agronomy Journal 72(2): 387-391.

Ghanem ME, Marrou H, Sinclair TR (2014) Physiological phenotyping of plants for crop improvement. Trends in plant science.

Giglioti ÉA, Sumida CH, Canteri MG (2015) Disease Phenomics. In Phenomics (pp. 101-123). Springer International Publishing.

Guzman PS, Diers BW, Neece DJ, St Martin SK, LeRoy AR, Grau CR, Hughes TJ, Nelson RL (2007) QTL associated with yield in three backcross-derived populations of soybean. Crop science 47(1): 111-122.

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

Harper JE (1974) Soil and symbiotic nitrogen requirements for optimum soybean production. Crop Science 14(2): 255-260.

Herbert SJ, Litchfield GV (1982) Partitioning soybean seed yield components. Crop Science 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 Open Access Publisher.

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 Research 115(1): 116-123.

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

Kabelka EA, Diers BW, Fehr WR, LeRoy AR, Baianu IC, You T, Neece DJ, Nelson RL (2004) Putative alleles for increased yield from soybean plant introductions. Crop science 44(3): 784-791.

Kahlon CS, Board JE (2012) Growth Dynamic Factors Explaining Yield Improvement in New Versus Old Soybean Cultivars. Journal of Crop Improvement 26(2): 282-299.

Kim KS, Diers BW, Hyten DL, Mian MR, Shannon JG, Nelson RL (2012) Identification of positive yield QTL alleles from exotic soybean germplasm in two backcross populations. Theoretical and Applied Genetics 125(6): 1353-1369.

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. Journal of experimental botany 65(12): 3311-3321.

Larson EM, Hesketh JD, Woolley JT, Peters DB (1981) Seasonal variations in apparent photosynthesis among plant stands of different soybean cultivars. Photosynthesis research 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 science 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. Theoretical and applied genetics 92(5): 516-523.

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

Li D, Pfeiffer TW, Cornelius PL (2008) Soybean QTL for yield and yield components associated with alleles. Crop Science 48(2): 571-581.

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

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

Malik MA, Cheema MA, Khan HZ, Wahid MA (2006) Growth and yield response of soybean (Glycine max L.) to seed inoculation and varying phosphorus levels. Journal of Agricultural Research 44(1): 47-53.

Mandl FA, Buss GR (1981) Comparison of narrow and broad leaflet isolines of soybean. Crop Science 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.*). Theoretical and Applied Genetics 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 Science 36(5): 1327-1336.

Meinshausen N, Bühlmann P (2006) High-dimensional graphs and variable selection with the lasso. The Annals of Statistics; 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, Montpellier, France, August, 2002. Session 28. (pp. 1-2). Institut National de la Recherche Agronomique (INRA).

Orf JH, Chase K, Jarvik T, Mansur LM, Cregan PB, Adler FR, Lark KG (1999a). Genetics of soybean agronomic traits: I. Comparison of three related recombinant inbred populations. Crop Science 39(6): 1642-1651.

Orf JH, Chase K, Adler FR, Mansur LM, Lark KG (1999b). Genetics of soybean agronomic traits: II. Interactions between yield quantitative trait loci in soybean. Crop science 39(6): 1652-1657.

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. Theoretical and Applied Genetics 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. Theoretical and applied genetics 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 Science 45(5): 2015-2022.

Paterson AH (1995) Molecular dissection of quantitative traits: progress and prospects. Genome Research 5(4): 321-333.

Pedersen P, Lauer JG (2004) Response of soybean yield components to management system and planting date. Agronomy Journal 96(5): 1372-1381.

Pellet JP, Elisseeff A (2008) Using Markov blankets for causal structure learning. The Journal of Machine Learning Research 9:1295-1342.

Pettigrew WT (2008) Potassium influences on yield and quality production for maize, wheat, soybean and cotton. Physiologia plantarum 133(4): 670-681.

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 Science 40(3): 834-837.

R Core Team (2015) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL 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 Recent Advancements in Gene Expression and Enabling Technologies in Crop Plants (pp. 131-192). Springer New York.

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 Science 54(3): 939-943.

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 Science 53(4): 1375-1383.

Reyna N, Sneller CH (2001) Evaluation of marker-assisted introgression of yield QTL alleles into adapted soybean. Crop Science 41(4): 1317-1321.

Richards RA (2000) Selectable traits to increase crop photosynthesis and yield of grain crops. Journal of Experimental Botany, 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 Science 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 × management interactions in soybean: I. Planting date. Crop Science 53(3): 1128-1138.

Simpson AM, Wilcox JR (1983) Genetic and phenotypic associations of agronomic characteristics in four high protein soybean populations. Crop Science 23(6): 1077-1081.

Sinclair TR (1986) Water and nitrogen limitations in soybean grain production I. Model development. Field Crops Research 15(2): 125-141.

Soares MM, Oliveira GL, Soriano PE, Sekita MC, Sediyama T (2013) Performance of soybean plants as function of seed size: II. Nutritional stress. Journal of Seed Science 35(4): 419-427.

Sorensen D, Gianola D (2002) Likelihood, Bayesian, and MCMC methods in quantitative genetics. Springer.

Spear JD, Fehr WR (2007) Genetic improvement of seedling emergence of soybean lines with low phytate. Crop science 47(4): 1354-1360.

Specht JE, Hume DJ, Kumudini SV (1999) Soybean yield potential-a genetic and physiological perspective. Crop Science 39(6): 1560-1570.

Stekhoven DJ, Bühlmann P (2012) MissForest - nonparametric missing value imputation for mixed-type data. Bioinformatics 28(1): 112-118.

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. Agronomy Journal 101(2): 352-356.

Ustun A, Allen FL, English BC (2001) Genetic progress in soybean of the US Midsouth. Crop Science 41(4): 993-998.

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.

Wang D, Graef GL, Procopiuk AM, Diers BW (2004) Identification of putative QTL that underlie yield in interspecific soybean backcross populations. Theoretical and Applied Genetics 108(3): 458-467.

Wells R (1991) Soybean growth response to plant density: Relationships among canopy photosynthesis, leaf area, and light interception. Crop Science 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 Science 54(1): 340-348.

Wortman SE, Francis CA, Galusha TD, Hoagland C, Van Wart J, Baenziger PS, Johnson M, et al (2013) Evaluating cultivars for organic farming: maize, soybean, and wheat genotype by system interactions in Eastern Nebraska. Agroecology and Sustainable Food Systems 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 Science 55(5): 2056-2067.

Xavier A, Xu S, Muir WM, and Rainey KM (2015) NAM: Association Studies in Multiple Populations. Bioinformatics, btv448.

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

Zhao T, Liu H, Roeder K, Lafferty J, Wasserman L (2012) The huge package for high-dimensional undirected graph estimation in R. The Journal of Machine Learning Research 13(1): 1059-1062.

Zera AJ, Harshman LG (2001) The physiology of life history trade-offs in animals. Annual review of Ecology and Systematics, 95-126.

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.). Journal of Genetics and Genomics 37(8): 545-556.

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. Theoretical and Applied Genetics 108(6): 1131-1139.

# CHAPTER 2: WALKING THROUGH STATISTICAL BLACK BOXES IN PLANT BREEDING

ABSTRACT

Intelligent decision making relies on our capability of extracting useful information from data that may help us to achieve our goals more efficiently. Many plant breeders and geneticists perform statistical analyses without knowing the underlying assumptions of the methods and their strengths or pitfalls. In other words, they treat these statistical methods (software and programs) like black boxes. Black boxes represent complex pieces of machinery with contents that are not fully understood by the user. The user sees the inputs and outputs without knowing how the outputs are generated. By providing a general background on statistical methodologies, the objectives of this review are (1) to introduce basic concepts of machine learning and its applications to plant breeding; (2) to link classical selection theory to current statistical approaches; (3) to show how mixed models are solved and to extend their application to pedigree-based and genomic-based prediction; and (4) to clarify how the algorithms of genome-wide association studies work, including their assumptions and limitations.

## 2.1 Introduction

Inferences and models can be of empirical or experimental design. Empirical methods work best for well-characterized phenomena, for which the solution can be found analytically, whereas experimental methods are necessary to make inferences from data and use algorithms to identify patterns in the data. The science that studies these algorithms is known as machine learning. Machine learning also includes the area of artificial intelligence dedicated to building and studying algorithms that are capable of learning from data, endeavoring to find an optimal solution that minimizes a given loss. This makes these machine learning algorithms much more flexible than logical algorithms.

Genetics widely exploits two particular branches of machine learning, so-called *supervised* and *unsupervised* learning. Supervised techniques help solve problems for which we have explanatory and response variables. This commonly applies to quantitative genetics for prediction, selection, and classification. Unsupervised procedures are used when no response variable exists. Population genetics often uses unsupervised procedures for problems associated with clustering genotypes and to find admixture in populations.

Due to the quantitative nature of most traits of interest, Gaussian process (GP) is the most employed type of supervised learning algorithm in plant and animal breeding (Rasmussen 2004, Lynch and Walsh 1998). Fisher's infinitesimal model, which forms the basis of the principles of breeding, states that an infinite number of stochastic processes control the observed phenotype (Orr 2005, Farrall 2004), which converges to a Gaussian distribution according to the central limit theorem. GP represents the basis of selection theory, breeding values, and association studies (Sorensen and Gianola 2002).

Classification procedures are important for the genetic improvement of categorical traits and decision making. For instance, breeding programs develop products specifically for different markets (Acquaah 2009, Cleveland and Soleri 2002) and classification models determine the boundaries of the qualities that define these market niches (Lim 1997). In soybeans, adaptation zones define which maturity group (MG) can be cultivated in each region according to the latitude, soil, and climate; in other words, they determine the target environment for breeders (Dardanelli et al. 2006). For example, Zhang et al. (2007) suggest that soybean adaptation zones have misclassification issues because the growing zone for MG IV to MG VI is much larger than originally thought.

The main goal of this paper is to reveal the inner workings of the black boxes of statistical analysis in plant breeding by explaining the theory and applications of machine learning in statistical genetics, focusing on widely applied mixed linear models designed for prediction, selection, and inference.

## 2.2. Gaussian Process

In one way or another, quantitative traits follow a distribution pattern. For example, categorical traits with two classes follow a binomial distribution, as with the color of flowers in soybeans, which are either white or purple. If a third flower color existed, the trait would follow a multinomial distribution. Counting (ie. discrete) traits, such as the number of days until flowering, could be modeled using a Poisson distribution. Traits like grain yield and plant height are continuous and often follow a normal distribution. The heritability of the traits, discussed later in this review, can assume any value between zero and one, thus a beta distribution is often best to characterize this process. Variance components discussed in the coming section should always have positive values on a

continuous scale, and thus they can be described in terms of a gamma distribution or chi-squared distribution.

In general terms, all distributions have two very important coefficients derived from their moment generation function: these coefficients are expectation ($E[X]$) and variance ($E[X^2] - (E[X])^2$). The normal distribution has a sigmoidal shape, like a bell. The expectation of any normal random variable is its mean, notated by the Greek letter mu ($\mu$), and the deviance from the expectation is the variance, notated by the square of Greek letter sigma ($\sigma^2$). The square root of the variance is the standard deviation $\sigma$, which represents the deviance in the same scale as the observations. The proper notation of a random variable ($y$) normally distributed is $y \sim N(\mu, \sigma)$.

In plant and animal breeding, it is very important to know how to handle a normal distribution, since most quantitative genetic theory assumes normality. For example, the equation by which one can calculate the probability of finding a plant that yields *x* bu/ac from a given population is called a probability density function (PDF, $\phi$), and the probability of finding any plant with yield equal or lower than *x* is called a cumulative density function (CDF, $\Phi$). The probability density function is, therefore, the first derivative of the cumulative density function. The function that defines the normal PDF is $\phi(x) = (2\pi)^{(-0.5)}\sigma^{-1}\exp(-0.5\sigma^{-2}(x - \mu)^2)$. A description of a Gaussian distribution is shown in Figure 5.

The so-called standard normal, which is notated as $Z$, is a special case of normal distribution with a mean of zero and variance of one. The following transformation can

standardize any normal: $Z = (x - \mu)/\sigma$. The sum of $\nu$ squared standard Gaussians ($Z^2$) is called a chi-squared ($\chi^2$) distribution with $\nu$ degrees of freedom.

There are several methods to estimate parameters of a distribution. These include likelihood methods, such as maximum likelihood (ML) and restricted maximum likelihood (REML), and Bayesian procedures. What differentiates these methods is their so-called *loss function*. For example, the least square procedure aims to minimize the squared error while likelihood methods maximize the likelihood function. For now, we will focus on likelihood methods.

Since each observation contains some information about the unknown parameters, more data can provide more accurate and precise estimates of mean and variance. Likelihood methods search for the parameters that maximize either the likelihood (L) or log-likelihood ($l$). The normal PDF defines the joint probability $p(\mathbf{y}; \boldsymbol{\theta})$, where $\mathbf{y}$ represents the observed data and the Greek letter theta ($\boldsymbol{\theta}$) represents one or more unknown parameters, here $\boldsymbol{\theta} = (\mu, \sigma^2)$. Thus, assuming Gaussian data, the marginal log-likelihood for each observation is given by $l(\theta, y_i) = -0.5\ln(\pi) - 0.5\ln(\sigma^2) - (2\sigma^2)^{-1}(y_i - \mu)^2$.

The ML estimator for each element of $\boldsymbol{\theta}$ adjusts iteratively by means of a gradient that is the vector of the first-order partial derivatives of the log-likelihood for each element (ie. mean and variance), here notated as the $S(\theta|y)$ that satisfies $S(\hat{\theta}|y) = 0$. Estimation of the mean and variance of a normal random variable is the simplest example because the conversion is satisfied in the first iteration. In this case, these estimators are said to have a closed-form solution:

$$S(\mu|\mathbf{y}) = \partial L/\partial \mu = 0 \therefore \hat{\mu} = \sum y/n$$

$$S(\sigma^2|\mathbf{y}) = \partial L/\partial\sigma^2 = 0 \therefore \hat{\sigma}^2 = \sum(y-\mu)^2/n$$

Multidimensional problems are solved using linear algebra (ie. matrix framework). In this case, parameter estimation requires the second derivative of the log-likelihood, called the Hessian matrix. The negative expectation of the Hessian matrix yields the Fisher information matrix. Hessian and Fisher information matrices are further discussed in later sections.

2.3. Infinitesimal Model and Selection Theory

For a normally distributed trait in a population, *directional selection* occurs when a breeder induces the mean to move in the desired direction over generations (Fig6). To achieve that, the breeder must impose a selection threshold. The breeder selects individuals above this threshold as the progenitors of the next generation under the assumption that those individuals provide better genetic properties. In self-pollinated species, male-sterility is a common tool that makes directional selection possible (Recker et al. 2014).

The genetic properties that affect the phenotype involve alleles with positive and negative effect. Alleles are versions of genes that represent the genetic effect over a given trait. Alleles can interact within the locus, across loci, and by external stimuli; these phenomena are called *gene action*, *epistasis*, and *expression*, respectively. The number of alleles carried by a locus depends on the ploidy level of the individual. This review focuses on diploid organisms, those with two alleles at each locus.

Selection intensity (*i*) represents the number of standard deviations that defines the cutoff of the population, known as the truncation point, above which selected individuals remain in the breeding population as progenitors. The population of selected individuals

characterizes a one-sided truncated normal distribution. It is possible to estimate the expectation of this distribution ($\mu^*$) using the mean ($\mu$) and standard deviation ($\sigma$) of the original distribution and, of course, the truncation point ($t = \hat{\mu} + i\hat{\sigma}$) (Wricke and Weber 1986). The expected mean of a selected population is estimated as $E[\mu^*|t] = \mu + \sigma[\phi(\alpha)/(1 - \Phi(\alpha)]$, where $\phi$, $\Phi$ and $\alpha$ represent the normal PDF CDF, and the standardized truncation point ($\alpha = (t - \mu)/\sigma$), respectively, as shown in Figure 7.

Breeders obtain larger short-term genetic gains by increasing selection intensity; however, this practice sacrifices long-term gains unless, of course, breeders continuously introduce exogenous sources of genetic variability into the breeding population.

The next generation will not have the expected mean $\mu^*$, since the phenotype is not exclusively due to genetic factors (Nyquist and Baker 1991). Despite the fact that alleles interact in a very complex fashion, their expression is a function of environmental stimuli (aka. genotype by environment interaction). This is called realized heritability ($h_r^2$): the rate between the observed mean of the new generation ($\mu^{(t+1)}$) and its expected mean ($\mu^*$) based on the selected progenitors.

Fisher (1918) proposed that, for a given quantitative trait, there are an infinite number of genes with minor additive contributions affecting the phenotype, the so-called infinitesimal model. In selection theory, the general goal of breeders is to increase the frequency of desirable alleles in a population over time, under the assumption that allele effect works in additive fashion. Exceptions to this include the gains associated with heterosis as exploited by programs that develop hybrids (eg. maize), or by clonally propagated species (eg. potato). According to Fisher's model, the outcome of each gene is additive and is measured

by the effect of an allelic substitution. In this sense, the model matches the definition of a Gaussian process that consists of normally distributed random variables as elements of some infinite-dimensional space (aka. Hilbert spaces) or, in other words, a multivariate normal with an infinite number of kernels.

When applied to finite breeding populations, Fisher's model is confronted with population genetic issues. For example, finite populations can maintain only a limited number of alleles (Kimura and Crow 1964). Furthermore, multiple evolutionary forces will be acting simultaneously, such as various types of selection and long-term random genetic drift, which triggers continuous bottlenecks (Wright 1930). This extension of the infinitesimal model is called the Wright-Fisher Markov Chain model. The selection pressure applied over generations in a finite population implies a major trade-off between the response to selection and genetic gains over time (Fig8).

From the standpoint of statistical genetics, most field crops breeding populations follow the definition of a stochastic Fisher-Wright process (Imhof and Nowak 2006): finite populations with non-overlapping generations, diploid behavior, and ongoing frequency-dependent selection. Frequency-dependent selection occurs when breeders endeavor to improve fitness-related traits, breeding populations where the main goal is to increase grain yield or resistance to pests and disease.

Crow and Kimura (1970) pointed out that the fluctuations that Fisher defined as noise, Sewall Wright defined as (a slow) evolution. The stability of genetic gain over time relies on selection intensity, mutation rate, and total ($n$) and effective ($\mathrm{N_e}$) population size. Effective population size is a major limiting factor for efficient selection in plant breeding

programs, with serious implications for traditional and genomic-based selection techniques (MacLeod et al. 2014). According to Zeng and Hill (1986), the optimal selection intensity occurs when new haplotypes arise at the same frequency with which alleles undergo fixation (known as a convergence rate), such that the population does not exhaust its diversity.

Self-pollinated species are more likely to run out of genetic resources due to their reproductive nature. For example, the effective population size of soybeans in the United States is equivalent to 27 lines (St. Martin 1982) and, not surprisingly, soybean production is reaching a yield plateau (Egli 2008a) that is nearly half of the field potential (Specht et al. 1999) due to these limited genetic resources (Egli 2008b). However, new breeding tools in the *"omics generation"* are bringing hope to this currently limited scenario (Rincker et al. 2014).

## 2.4. Variance Decomposition and Parsimony

The phenotype of a quantitative trait is in a non-deterministic state. Therefore, it requires a stochastic model to approximate an infinite population; in other words, a model with random variables defining which variance components are of interest. The first model to express variation in the phenotype was the infinitesimal model, in which the phenotypic variance ($\sigma^2_y$) is a function of genetics ($\sigma^2_G$) and environmental variances ($\sigma^2_E$), so that $\sigma^2_P = \sigma^2_G + \sigma^2_E$.

Variance component analysis (VCA) is a very common practice in plant breeding and agronomic studies. Two of the most common methods to perform variance decomposition are the analysis of variance (ANOVA) and restricted maximum likelihood (REML). Studying the variance due to genotype and environment in soybeans, Carvalho et al. (2008)

suggest that both methods provide similar variance components under a balanced experimental design, but that under unbalanced conditions, the ANOVA method becomes biased while REML still provides consistent variance components and the best linear unbiased predictions (BLUPs) (Henderson 1975). This makes REML procedures the most deployed method for VCA in breeding studies with BLUPs used for variety selection (Piepho et al. 2008).

For Fisher, all variation not explained by genetics was due to environment. In plant breeding in which replications allow us to measure the variation due to environment, the variance of the phenotype can be further decomposed. Thereby it is possible, for example, to estimate the interaction between genotype and environment ($\sigma^2_{G\times E}$) and isolate the pure error ($\sigma^2_{\varepsilon}$). Each term can undergo further decomposition. Environmental variance can include year ($\sigma^2_{Y}$), location ($\sigma^2_{L}$), and management ($\sigma^2_{M}$), which reflects the controllable environment. In soybeans, Yan and Rajcan (2003) conducted a genotype by environment analysis, decomposing $\sigma^2_{E}$ into $\sigma^2_{Y}$ and $\sigma^2_{L}$ with all possible interaction terms (ie. $\sigma^2_{G\times Y\times L}, \sigma^2_{G\times L}, \sigma^2_{G\times Y}$). They concluded that most variance associated with environment is due to year rather than location.

If genotypic information is available by genotyping with co-dominant molecular markers, such as single nucleotide polymorphism (SNP), then breeders and geneticists are able to subdivide genetic variance terms. The first decomposition of genetic variation yields the additive genetic variance ($\sigma^2_{A}$), the dominance genetic variance ($\sigma^2_{D}$), and epistasis ($\sigma^2_{I}$). Likewise, the epistasis represents the interaction among loci that comprises the following terms: additive-by-additive ($\sigma^2_{AA}$), additive-by-dominant ($\sigma^2_{AD}$), and dominant-by-dominant ($\sigma^2_{DD}$).

At this point, it is very important to introduce two concepts: *narrow-* ($h^2$) and *broad-sense* (H) heritability (Acquaah 2009). In statistical terms, heritability is known as the intra-class correlation coefficient, a term that refers to the amount of total variation due to one of its components. Broad-sense heritability is the amount of variation due to genetics ($H = \sigma^2_G/\sigma^2_P$), also known as *repeatability* (Nyquist and Baker 1991). It illustrates 'nature-versus-nurture', distinguishing between what is due to genetics and what is due to environment. Narrow-sense heritability is the fraction of phenotypic variance due to the additive genetic variance only ($h^2 = \sigma^2_A/\sigma^2_P$) associated with the variance transmitted over generations. The latter is the most important for breeding quantitative traits because it describes how accurately breeding values, generated from the additive relationship between individuals, correspond to the phenotype. Because of this, narrow-sense heritability is used to predict the offspring performance.

Genetic variance component estimation typically starts with building Wright's numerator relationship matrix (aka. kinship or kernel) and then proceeds by solving the Henderson's equation (Henderson 1984). The Henderson's equation refers to a generalized mixed linear model for genetic prediction purposes. This model treats controllable elements, such as those imposed by experimental designs, as a fixed effect and treats the term that defines genetic components as a random effect with non-independent observations. The interdependence among observations is expressed by the so-called kernel matrix.

There are multiple types of kernel matrices used to represent the relationship among genotypes, including: the pedigree matrix (**A**) as originally proposed by Wright (1922); the genomic relationship (G) expressed as a linear kernel obtained by the cross product of the genotypic matrix containing the marker information ($\mathbf{MM}'$); and distance-based kernels,

such as the Gaussian ($\exp[-\mathbf{E}^2/h]$) and exponential kernels ($\exp[-\mathbf{E}/h]$) that use Euclidean distance $\mathbf{E}$ to describe the genetic distance among individuals based on molecular markers and a bandwidth parameter $h$. The term support vector machine (SVM) is often used to define GP that use regularized kernels for prediction or classification.

The dimensions of a genotypic matrix depend on the number of markers ($p$) for the columns and the number of individuals ($n$) for the rows. Therefore, each cell in this matrix represents a locus of an individual. Xu (2013) coded {AA, Aa, aa} using {1, 0, -1} to build a linear kernel that describes the additive-relationship matrix with molecular data and {0, 1, 0} to build the dominance-relationship matrix. Although there are many other ways one can code the molecular genotype of an allele (Strandén and Christensen 2011, VanRaden 2008). The resulting cross product of genotypic matrices is always a square symmetric matrix ($n \times n$) where each cell describes the relationship between individuals in the corresponding row and column. Although it is possible to add as much complexity to the variance decomposition model as the geneticist or breeder desires, there are two principles that one must take into account: the *hierarchical principle* and the *sparsity principle*. The first states that lower order terms are generally more important than higher order ones. In other words, epistasis may contribute little to the total genetic variance and at a high computational cost. The second principle reinforces the statistical parsimony in which a few terms explain most variation. In practical terms not all of the genome contributes to all traits, but rather a reduced number of regions contribute most. These regions are known as quantitative trait loci (QTL). Lander and Botstein (1989) defined the phenotypic variance of a quantitative trait as a Gaussian process after figuring out that the phenotypic

distribution considered to comprise a single normal distribution was actually a mixture of distributions associated with combinations of QTL (Fig9).

The identification of QTL occurs by comparing the log-likelihood of two models (Yan et al. 2014). The first is the *null model*, which contains the polygenic term corresponding to the effect of background genetics, often computed through a kernel regression (Xu 2013, de los Campos et al. 2010). The second is the *full model*. It is a mixture model including the polygenic term and the candidate genomic fraction, which is a marker or an interval between markers. The statistical test is called the likelihood ratio test (LRT). The hypothesis testing supporting the association between any point in the genome being and the trait in study can be expressed in terms of LRT itself, as p-values ($LRT \sim \chi^2_{v=1}$) or as a logarithm of odds (LOD score) by dividing LRT by 4.61 (Lynch and Walsh 1998).

The practice of QTL mapping occurs in both experimental and random populations. There are two major methods to find QTL: linkage mapping and association mapping. Linkage mapping is a method of tracking QTL as a map function of known genetic distance between markers. It is commonly performed in experimental populations designed for this purpose, with no need for kinship in either the full or reduced model. Association mapping, also known as linkage disequilibrium mapping, is a test of single markers across the whole genome for experimental or random populations with extra scrutiny for the existence of subpopulations. In both methods, undetected regions will bias the number of QTL downward and the average effect of QTL upward due to a phenomenon called the *Beavis effect*. This is because the precision and accuracy of finding real QTL relies extensively on the population size (Beavis 1998) and implicit assumptions associated with the population type (Xu 2003a, Nyquist and Baker 1991).

## 2.5. Breeding Values, Kinship and Regression

Breeders select only a fraction of the breeding population to develop into the release of a commercial line. They base their selection of top-ranked genotypes either on the values of one trait at a time (ie. tandem selection), multiple quantitative traits simultaneously (ie. independent culling), or on the combination of traits (ie. index selection). Nonetheless, there are four possible values they use to select a quantitative trait: phenotypic value, genetic value, estimated breeding value, or direct genomic value. While selection based on phenotypic values uses the phenotypes in a straightforward manner, the estimation of the latter three requires the implementation of mixed linear models with various relationship matrices.

Mixed model theory is the life's work of the geneticist Charles Henderson, who was motivated to implement and apply Wright's pedigree-based kinship matrix to breeding and selection, a technique which later expanded to generalized expressions and to the genomic level. A mixed model occurs when the response variable ($\mathbf{y}$) is a function of a fixed effect term ($\mathbf{Xb}$) and one or more random effects ($\mathbf{Zu}$) other than the residuals ($\mathbf{e}$). Random effects have a mean of zero. The correlation between their observations is expressed by the variance-covariance matrix ($\mathbf{V}$), which is a function of the residual correlation ($\mathbf{R}$), residual variance, one or more kinship matrices ($\mathbf{K}$), and the variances associated with each random effect ($\mathbf{V} = \sum \mathbf{ZKZ}\sigma_a^2 + \mathbf{R}\sigma_e^2$). Random terms can be independent as well, and if so, any $\mathbf{K}$ and/or $\mathbf{R}$ are replaced by an identity matrix $\mathbf{I}$.

In linear algebra terminology, capital letters express matrices while lowercase letters are vectors and scalars. Vector and matrices are written in bold letters and constant scalars are written in italic. The common notation of a mixed model is given by the linear model $\mathbf{y} =$

$\mathbf{Xb} + \mathbf{Zu} + \mathbf{e}$. The $\mathbf{X}$ and $\mathbf{Z}$ matrices are $n \times p$ incidence matrices of fixed and random effects, respectively, while b and u are the regression coefficients of each fixed and random parameter. Likewise $\sigma_a^2$ and $\sigma_e^2$ are the random effect and residual variances, and $\mathbf{K}$ and $\mathbf{R}$ are the kernels of random effects and residuals used to define the relationship among observations.

The simplest case in breeding is the so-called animal model. The animal model is an implementation of Fisher's variance decomposition that attributes everything that is not due to the genetic term to error, since it is possible to include controllable environmental factors in the model as fixed effects. A random effect shrinks based on its regression coefficient by the factor of a regularization term notated by lambda (λ), which is the ratio between error variance and random term variance ($\lambda = \sigma_e^2/\sigma_a^2$). Henderson further simplified the mixed model equation (MME) by assuming that residuals are uncorrelated ($\mathbf{R} = \mathbf{I}$). This is known as *Henderson's method III*, reducing it to a $\mathbf{Cg} = \mathbf{r}$ problem, thus:

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{Z}'\mathbf{X} \\ \mathbf{X}'\mathbf{Z} & \mathbf{Z}'\mathbf{Z} + \lambda\mathbf{K}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix} \therefore \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \end{bmatrix} \therefore \mathbf{Cg} = \mathbf{r}$$

The kernel relationship matrix $\mathbf{K}$ will define what type of value the model yields for selection purposes. If $\mathbf{K}$ is an identity matrix then $\mathbf{u}$ is a vector of genetic values. If $\mathbf{K}$ is Wright's numerator matrix built with pedigree information then $\mathbf{u}$ is a vector of estimated breeding values, and if $\mathbf{K}$ is based on molecular information then $\mathbf{u}$ is a vector of genomic direct values, also known as genomic enhanced breeding values. In order to avoid conflicting terminology, from this point the term “breeding value” denotes the random effect coefficients $\mathbf{u}$.

If $\sigma_e^2$ and $\sigma_a^2$ were known quantities, finding the coefficients $\mathbf{b}$ and $\mathbf{u}$ would not be a problem. However it is necessary to estimate coefficients and variance components from the data simultaneously. The parameters estimated by Henderson's method are *Empirical Bayes estimators* because the prior estimation depends directly on the data (Zhou and Stephens 2014, Gianola et al. 1986). Sorensen and Gianola (2002) showed the Bayesian nature of the model by expressing $\mathbf{X'X}$ as an additional random effect ($\mathbf{X'X} + \lambda\mathbf{K}^{-1}$) that does not undergo regularization (ie. shrinkage) due to the prior knowledge of $\sigma_b^2 \to \infty$, which results in a null shrinkage ($\lambda = \sigma_e^2/\sigma_b^2 = \sigma_e^2/\infty = 0$) with independent terms ($\lambda\mathbf{K}^{-1} = 0 \times \mathbf{K}^{-1} = 0$). Under the frequentist framework, the probabilistic description of $\mathbf{y}$ is defined as $\mathbf{y}$~N($\mathbf{Xb}, \mathbf{ZKZ}\sigma_a^2 + \mathbf{I}\sigma_e^2$), whereas under the Bayesian framework it becomes $\mathbf{y}$~N($\mathbf{Xb} + \mathbf{Zu}, \mathbf{I}\sigma_e^2$).

To simplify the notation, let $\mathbf{W}$ represent the design matrices $[\mathbf{X}, \mathbf{Z}]$, and $\mathbf{g}$ represent the regression coefficients $[\mathbf{b}, \mathbf{u}]$, and $\mathbf{\Sigma}$ represent the matrix of covariances that would accommodate $\lambda\mathbf{K}^{-1}$ in the position $\mathbf{C_{22}}$. Thereby $\mathbf{C} = \mathbf{W'W} + \mathbf{\Sigma}$ and $\mathbf{r} = \mathbf{W'y}$.

If there is a known residual correlation between observations that can be described by a $n \times n$ residual relationship matrix $\mathbf{R}$, then it is possible to build the model with a minor modification to accommodate heteroskedasticity: $\mathbf{C} = \mathbf{W'R^{-1}W} + \mathbf{\Sigma}$ and $\mathbf{r} = \mathbf{W'R^{-1}y}$.

For genotype prediction $\mathbf{u}^*$, breeders must estimate the properties of a non-existent distribution based on observed populations and, in this case, they will have to fit stochastic models for events that are yet to occur (Sorensen and Gianola 2002). In cases such as these, when the computation of breeding values requires estimation of $\lambda$, there are several approaches that can help to find an optimal value for $\lambda$.

This raises a question: how can one find the λ that provides a robust prediction? The main tool of supervised machine learning is its use of cross validation to find the tuning parameters λ that provide the best prediction. Cross validation works by dividing the dataset into $k$ subsets and testing the predictability for a wide range of values for λ. The predictability can be computed as the mean square prediction error (lower is better) or the correlation between the predicted and observed (higher is better). A three-fold cross validation would work as follows:

1. Divide the observed data into three groups $(\mathrm{A}, \mathrm{B}, \mathrm{C})$;
2. Propose a value for λ;
3. Use AB to predict C, AC to predict B, and BC to predict A;
4. Compute the mean predictability for this given value of λ;
5. Repeat the previous two steps for a wide range of λ;
6. Use the value of λ that provides the highest predictability.

The λ parameter controls the complexity of the model and, consequently, the tradeoff between bias and variance. Increases in λ mean that bias is being added to reduce the complexity of the model, which often creates a more consistent prediction.

As an alternative to cross validation, it is possible to compute λ to provide the best linear unbiased prediction. There are three popular kinship-based methods used for estimating variance components in order to obtain a robust value of λ as $\sigma_e^2/\sigma_b^2$ (Robinson 1991): restricted maximum likelihood (Patterson and Thompson 1971), Bayesian Gibbs sampling (BGS) (Wang et al. 1993), and an alternative re-parameterization by reproducing kernel Hilbert spaces (RKHS) (Gianola et al. 2006). The next section will present some whole-

genome regression methods that do not require explicit kernels to provide an equivalent BLUP solution.

### 2.5.1 REML Algorithm

REML is probably the most employed method for general-purpose estimation of variance components and regression coefficients. It is relatively unbiased when the number of observations is greater than the number of parameters ($n > p$) and much work has been done to make computationally feasible algorithms (Zhou and Stephens 2014, Kang et al. 2008, Lee and van der Werf 2006, Misztal et al. 2002).

There are a variety of algorithms to compute the REML variance components. This can be seen as a numerical optimization problem in which the main goal is to find the variance components and regression coefficients that maximize the restricted maximum likelihood of the data. Popular algorithms include the derivation-free algorithm (Meyer 1989); first-derivative methods, such as expectation-maximization (EM) (Dempster et al. 1977); and second-derivative or Newton-type methods, such as Newton-Raphson (NR), Fisher Scoring (FS), and Average Information (AI) (Gilmour et al. 1995). First- and second-derivative methods have an iterative-analytical solution but can be also solved numerically via Monte Carlo (Matilainen et al. 2013).

As previously mentioned, the restricted log-likelihood function is expressed by $l = -0.5[\log|\mathbf{C}| + \log|\mathbf{K}| + \mathrm{n_r}\log(\sigma_\mathrm{a}^2) + \mathrm{n}\log(\sigma_\mathrm{e}^2) + \mathbf{y}'\mathbf{P}\mathbf{y}]$ (Searle 1979), in which $\mathrm{n_r}$ is the length of $\mathbf{u}$, $\mathbf{C}$ is from the simplified MME representation ($\mathbf{C}\mathbf{g} = \mathbf{r}$), and $\mathbf{P}$ is the parametrization matrix that corresponds to the covariance matrix ($\mathbf{V}$) adjusted by the number of degrees of freedom of fixed effects. The parameterization matrix is computed as $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}\mathbf{V}^{-1}$.

The derivation-free approach implemented by Meyer (1989) finds the variance components and coefficients by minimizing the restricted log-likelihood through a heuristic method of minimization called the simplex method (Nelder and Mead 1965). This method is considered inefficient for complex models with large data. Despite the obsolescence of the simplex method, Kang et al. (2008) reintroduced the use of alternative numerical optimizers to efficiently solve mixed models in the so-called efficient mixed model association (EMMA) algorithm.

Henderson (1984) presented the expectation maximization (EM-REML) solution based on the EM-ML algorithm of Dempster et al. (1977), using the first derivative of the restricted log-likelihood as simplified by Searle (1979). The principle of EM is to iteratively update residuals, variances, and coefficients as follows: coefficients $\mathbf{g}$ are obtained by solving the MME as $\mathbf{g} = \mathbf{C^{-1}r}$ and residuals as $\mathbf{e} = \mathbf{y} - \mathbf{Wg}$. The residual variance is obtained by $\sigma_e^2 = n^{-1}[\mathbf{e'e} + \mathrm{tr}(\mathbf{WC^{-1}W'})\sigma_e^2]$ and the random effect variance is calculated as $\sigma_a^2 = n_r^{-1}[\mathbf{u'A^{-1}u} + \mathrm{tr}(\mathbf{A^{-1}C^{22}})\sigma_e^2]$, where $\mathbf{C^{22}}$ represents the $\mathbf{C_{22}}$ term from $\mathbf{C^{-1}}$. EM is a very consistent algorithm, but it converges slowly and it requires the inversion of $\mathbf{C}$ every round to find the regression coefficients. Some numerical strategies can help with solving the MME, such as Cholesky decomposition and Gauss-Seidel algorithm (Legarra and Misztal 2008).

Newton-type methods work by using the gradient $S(\boldsymbol{\theta}|\mathbf{y})$ of the second derivative, as described in the first section. This gradient is generated by a Taylor series converging toward the direction in which the parameters maximize the log-likelihood (Hofer 1998). All Newton-type methods have a similar framework to update parameters $\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t + H^t S^t$. The parameters being updated ($\boldsymbol{\theta}^{t+1}$) here are the variance components ($\boldsymbol{\theta} =$

$[\sigma_a^2, \sigma_e^2]$), while $H^t(\boldsymbol{\theta}|\mathbf{y})$ is the hessian matrix at the time $t$. The hessian matrix is employed for NR-REML. It represents the observed matrix information ($H = \partial^2 l / \partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'$).

In the FS-REML, the hessian is replaced by its negative expectation, the so-called Fisher Information matrix $I(\boldsymbol{\theta}|y) = E[-H(\boldsymbol{\theta}|\mathbf{y})]$. The average of the observed information and expected information $AI(\boldsymbol{\theta}) = 0.5(H(\boldsymbol{\theta}) + E[H(\boldsymbol{\theta})])$ provides the AI-REML proposed by Gilmour et al. (1995). The iterative algorithm AI-REML uses to find variance components in the animal model is:

$$\begin{bmatrix}\sigma_a^2\\ \sigma_e^2\end{bmatrix}^{t+1}$$

$$= \begin{bmatrix}\sigma_a^2\\ \sigma_e^2\end{bmatrix}^{t} + 0.5\begin{bmatrix}\mathrm{tr}(\mathbf{y}'\mathbf{PZZ}'\mathbf{PZZ}'\mathbf{Py})\sigma_e^2 & \mathrm{tr}(\mathbf{y}'\mathbf{PZZ}'\mathbf{Py})\sigma_e^4\\ \mathrm{tr}(\mathbf{y}'\mathbf{PZZ}'\mathbf{P})\sigma_e^4 & \mathrm{tr}(\mathbf{y}'\mathbf{Py})\sigma_e^6\end{bmatrix}^{-1}\begin{bmatrix}\mathrm{tr}(\mathbf{PZZ}') - \mathbf{y}'\mathbf{PZZ}'\mathbf{Py}\\ \mathrm{tr}(\mathbf{P}) - \mathbf{y}'\mathbf{PPy}\end{bmatrix}$$

The AI-REML is computationally demanding, but it converges within a few iterations to a consistent result. This algorithm has been widely implemented for breeding applications (Gilmour et al. 2009, Meyer 2007, Misztal et al. 2002). The most time-consuming operation for this method is to update the $\mathbf{P}$ matrix because it involves inversion of the covariance matrix. However, it is possible to substantially reduce this computational burden through the spectral decomposition or Eigendecomposition of $\mathbf{K}$ to speed up the inversion of $\mathbf{V}$ (Kang et al. 2008, Lippert et al. 2011). Any positive-definite square matrix can be Eigendecomposed into eigenvectors ($\mathbf{U}$) and eigenvalues ($\mathbf{D}$), thus $\mathbf{K} = \mathbf{UDU}'$. Then, one can obtain $\mathbf{V}^{-1} = \mathbf{ZU}[\mathbf{D} \times (\sigma_a^2\sigma_e^{-2}) + 1]^{-1}\mathbf{U}'\mathbf{Z}'\sigma_e^{-2}$ and the only inversion required is the vector of Eigenvalues.

### 2.5.2 BGS Algorithm

Bayesian Gibbs sampling (BGS) is a Monte Carlo Markov Chain (MCMC) algorithm proposed by Gelman and Gelman (1984) to generate posterior distributions by sampling from the conditional probability distribution of each parameter. The main idea is to generate samples based on the expectation and deviance of one parameter at a time and then use the mean, median, or mode of the distribution as the final parameter estimate.

The posterior distribution is especially useful for making inferences about the parameters. Iterations of Gibbs samplers converge to a point with "stable randomness" called *entropy* (a term named in accordance with its meaning in thermodynamics). The term *burn in* denotes the removal of iterations prior to entropy. Wang et al. (1993) proposed the first Gibbs sampler algorithm to solve mixed models in the breeding context, where coefficients follow a normal distribution ($N_{\mu,\sigma}$) and variance components follow an inverse Gamma ($\gamma_{\nu,S}^{-1}$) distribution, ensuring positive values for variance components. Nowadays, variance components are more commonly described in terms of a scaled inverse chi-squared distribution ($\chi_{\nu,S}^{-2}$), regulated by degrees of freedom ($\nu$) and scale (S). This is simply a special case of inverse gamma.

The sampling process from $\chi_{\nu,S}^{-2}$ works by dividing the sum of squares by a sample of chi-squared distributions. In this case, $\sigma_a^2 = (\mathbf{a}'\mathbf{A}^{-1}\mathbf{a} + S^*\nu^*)/(\chi^2_{n_a+\nu^*})$ and $\sigma_e^2 = (\mathbf{e}'\mathbf{e} + S^*\nu^*)/\chi^2_{n+\nu^*}$, where $S^*$ and $\nu^*$ represent the *priors* (García-Cortés and Sorensen 1996, Sorensen and Gianola 2002). Regression coefficients $\mathbf{g}$ have a closed form (ie. do not depend on priors). They are sampled from a normal distribution, one at a time, as $g_i \sim N(\mu =$

$g_i^*, \sigma^2 = \sigma_e^2 \mathbf{C_{ii}^{-1}})$, where $g_i^* = (r_i - \mathbf{C_{i,-i}} g_{-i}) \mathbf{C_{ii}^{-1}}$. As opposed to REML procedures, there is no need for inversion of $\mathbf{C}$.

Flat priors are used to express the total unawareness about the expected response based upon *Laplace's principle of uniform ignorance*. Flat priors are often used to provide results equivalent to those of frequentist analysis. For that, one can set $S^* = 0$ and $\nu^* = -2$. It is important to point out that flat priors can be improper, which means that they do not integrate out to one. However, improper priors often yield proper posteriors.

As opposed to its use in REML methods, the term *update* applies differently to BGS iterations because it is necessary to store the value of all coefficients and variance components from each round to generate the posteriori distribution of each parameter. Once the posteriori distribution is calculated, it is easy to infer credibility intervals (CI) by simply computing the percentiles that correspond to the boundaries of interest -- usually 0.025 and 0.975 based on the two-sigma rule (95% CI).

### 2.5.3 RKHS Algorithm

The reproducing kernels Hilbert spaces (RKHS) algorithm is another alternative to solve mixed effect models with known covariance structure (eg. animal model) that also yields the BLUP solution. The idea of this method is to replace the random term Zu with $u \sim N(0, \mathbf{K}\sigma_a^2)$ by a straight solver of kernels, comparable to a ridge regression of Eigenvectors.

Because they capture different levels of interaction among individuals, for the purpose of omic prediction, it is preferable to use Gaussian kernels ($\exp[-\mathbf{E}^2/h]$) over the linear kernel that commonly describes the genomic relationship matrix (Gianola et al. 2006).

Besides the Euclidean distance among genotypes $\mathbf{E}$, Gaussian kernels also require a bandwidth parameter $h$ that can be defined through cross validation or replaced by a normalizing factor, such as the mean of the distance matrix. To avoid the cross validation step, González-Camacho et al. (2012) used three Gaussian kernels computed with distinct bandwidth parameters.

The example with the animal model will help to illustrate the RKHS algorithm proposed by de los Campos et al. (2010). The first step is the spectral decomposition of the relationship matrix, $\mathbf{K} = \mathbf{UDU}'$. The incidence matrix of random effect ($\mathbf{Z}$) will be replaced by the Eigenvector matrix $\mathbf{U}$, or $\mathbf{ZU}$ in the case of replicated trials. The precision matrix previously computed as the inverse relationship matrix $\mathbf{K^{-1}}$ is replaced by the diagonal matrix of inverse Eigenvalues $\mathbf{D}^{-1}$. The model is solved with a BGS algorithm and the variance of the random effect is sampled from $\chi^{-2}_{\nu,S}$ as $(\mathbf{u}'\mathbf{D^{-1}}\mathbf{u} + S^{*}\nu^{*})/(\chi^{2}_{n_a+\nu^{*}})$.

The computational advantage of RKHS with a linear kernel in comparison to the ridge regression procedure comes from not having to regress the markers individually. This is especially important when there are more markers than observations and it also provides a nice framework to solve problems with multiple kinships. In addition, two computation strategies can help speed up the computation of the regression coefficients: (1) After $\mathbf{U}'\mathbf{U}$ yields an identity matrix, it is possible to sample a given regression coefficient $u_i$ from a normal distribution with mean $\mathbf{U}_i'(y - \mathbf{X}_{-i}b_{-i})/(1 + \lambda/\mathbf{D}_i)$ and variance $\sigma^2_e/(1 + \lambda/\mathbf{D}_i)$ or (2) one can employ strategies like Gauss-Seidel algorithm (Legarra and Misztal 2008) for solving linear equations.

The major pitfall of RKHS is the computational burden associated with the re-parameterization of the relationship kernel into Eigenvalues and Eigenvectors, especially for problems with multiple kernels and a large number of observations. This limitation can be overcome if just a partial number of Eigenpairs is considered sufficient. Then computational strategies such as the Lanczos algorithm become feasible. The Lanczos algorithm is an adaptation of power methods implemented in the Fortran package ARPACK.

### 2.5.4 WGR algorithm

As previously discussed, it is also possible to obtain BLUP estimates of breeding values and variance components without kinship matrices This is especially useful when *omic* information is available (de los Campos et al. 2013; VanRaden 2008) for a more reliable inference of breeding values (Bernardo and Nyquist 1998). These are called *whole-genome regression* (WGR) methods. Methods used for WGR are flexible so that they can accommodate high-dimensional problems; in other words, models with more parameters than observations.

In the WGR framework, the additive value of each marker is computed and breeding values are obtained by taking the sum of all marker values. The breeding value $\boldsymbol{u}$ of the $i^{th}$ genotype can be represented by $u_i = \mathbf{x}_i b$, where $\mathbf{x_i}$ represents a vector containing the marker information of the individual $i$, and $b$ is the value of each marker. If markers are coded as {-1, 0, 1} or {0, 1, 2} representing {AA, Aa, aa}, then the vector of regression coefficients **b** represents the additive value of each allele substitution (Xu 2013).

The simplest WGR model is called ridge regression (RR) or Tikhonov regularization, a Gaussian process compressing $p$ stochastic processes, where $p$ is the number of parameters

(ie. markers) in the model, that provides a result equivalent to kernel methods when using an genomic relationship.

The loss function that most WGR methods attempt to minimize is represented by $\text{argmin}(\mathbf{e}'\mathbf{e} + \lambda\mathbf{b}'\mathbf{b})$. Notice that this loss function comprises two terms: the sum of squares ($\mathbf{e}'\mathbf{e}$) and the complexity term $\lambda\mathbf{b}'\mathbf{b}$. The squared penalization of coefficients ($\lambda\mathbf{b}'\mathbf{b}$) is called $L_2$ penalization, while $L_1$ penalization denotes the use of the absolute sum ($\lambda||\mathbf{b}||$). The latter is also known as least absolute shrinkage and selector operator (LASSO) loss (Tibshirani 1996).

Let us begin by recalling the simplest univariate solution: the ordinary least squared (OLS). For a given model $\mathbf{y} = \mathbf{x}b + \mathbf{e}$, the OLS solution for the regression coefficient is $b = \text{cov}(x,y)/\text{var}(x)$ or, in algebraic notation, $b = \mathbf{x}'\mathbf{y}/\mathbf{x}'\mathbf{x}$. The ridge regression solution for the same problem is given by $b = \mathbf{x}'\mathbf{y}/(\mathbf{x}'\mathbf{x} + \lambda)$, where $\lambda$ can be defined through cross-validation or by $\sigma^2_e/\sigma^2_b$, as previously shown. Thus, the role of $\lambda$ is regularization through shrinkage.

The LASSO univariate solution works slightly differently. It starts by finding the OLS solution $b_{ols} = \mathbf{x}'\mathbf{y}/(\mathbf{x}'\mathbf{x})$. When $b_{ols}$ is positive, we compute $b_{lasso} = b_{ols} - \lambda/(\mathbf{x}'\mathbf{x})$ and if this regression coefficient turns out to be negative, it is set at zero. When the $b_{ols}$ is negative, we compute $b_{lasso} = b_{ols} + \lambda/(\mathbf{x}'\mathbf{x})$ and if this regression coefficient turns out to be positive, it is set at zero. Thus, LASSO performs variable selection in addition to shrinkage, whereas the ridge is incapable of yielding null regression coefficients.

It is important to introduce the univariate solution of ridge and LASSO in order to understand how the multivariate problems are solved by coordinate descent. The idea of

coordinate descent is simple: to reduce the regression to a univariate version and solve one coefficient at a time until convergence. To do so, it is necessary to fit all but the one variable that is being updated. Thus the ridge solution becomes: $b_i = \mathbf{x_i}'(\mathbf{y} - \mathbf{X_{-i}b_{-i}})/(\mathbf{x_i}'\mathbf{x_i} + \lambda)$.

Legarra and Misztal (2008) provided a nice framework to prevent the recalculation of $\mathbf{X_{-i}b_{-i}}$ for every parameter, the Gauss-Seidel residual update (GSRU) algorithm. It starts by computing the residuals ($\mathbf{e} = \mathbf{y} - \mathbf{Xb}$). In order to update each coefficient $b_i$ from the iteration at a time $t$ to a time $t + 1$, the algorithm replaces the response variable ($\boldsymbol{y}$) with an adjusted residual term computed as $\tilde{\mathbf{e}} = \mathbf{e} + \mathbf{x}_i b_i^t$ and updates the coefficient as $b_i^{t+1} = \mathbf{x_i}'\tilde{\mathbf{e}}/(\mathbf{x}_i'\mathbf{x}_i + \lambda)$. The next step before moving on to the next coefficient $b_{i+1}$ is to update the residuals: $\mathbf{e} = \tilde{\mathbf{e}_i} - \mathbf{x}_i b_i^{t+1}$.

It is important to keep two particular characteristics about ridge regression and LASSO in mind: (1) Fixed effects and intercepts do not undergo regularization ($\lambda = 0$); and (2) it is highly recommended to centralize predictors that will undergo regularization.

The Bayesian counterpart of ridge regression (BRR) is a Gibbs sampler with closed form (de los Campos et al. 2013). Here, we will use a simple linear model $\mathbf{y} = \mu + \mathbf{Xb} + \mathbf{e}$ to illustrate how the algorithm of BRR, containing just the overall mean ($\mu$) and the genotypic information ($\mathbf{X}$). We want to estimate the marker effects ($\mathbf{b}$) and variance components ($\sigma_b^2$ and $\sigma_e^2$).

The intercept ($\mu$) is sampled from a normal distribution with mean $\sum(\mathbf{y} - \mathbf{Xb})/n$ and variance $\sigma_e^2/n$. The computation of the marker effects is analogous to the GSRU algorithm. Each $b_i$ is sampled from a normal distribution with mean $\mathbf{x}_i'\tilde{\mathbf{e}}/(\mathbf{x}_i'\mathbf{x}_i + \lambda)$ and variance $\sigma_e^2/(\mathbf{x}_i'\mathbf{x}_i + \lambda)$. Remember that $\tilde{\mathbf{e}}$ corresponds to the residual of all parameters except the

one being updated and λ is calculated as $\sigma_e^2/\sigma_b^2$. Variance components are sampled from $\chi_{\nu,S}^{-2}$, as $\sigma_a^2 = (\mathbf{b}'\mathbf{b} + S^*\nu^*)/(\chi^2_{n_b+\nu^*})$ and $\sigma_e^2 = (\mathbf{e}'\mathbf{e} + S^*\nu^*)/\chi^2_{n+\nu^*}$.

Pioneering in the use of regression models to generate breeding values, Meuwissen et al. (2001) proposed the use of a non-Gaussian process. They proposed a Bayesian shrinkage regression (BSR) in which each marker would have its own variance characterizing a t-process, so-called *BayesA*. The algorithm is almost identical to BRR described above, but each marker has a different λ for which the individual marker variance ($\sigma_{b_i}^2$) is computed as $(b_i^2 + S^*\nu^*)/(1 + \nu^*)$.

BayesA has some interesting characteristics. Marker effects follows a t distribution (tick tails) that allows SNPs to pursue large effect. Breeding values from BayesA are usually more accurate than BRR but they may be biased if allele coding is not centralized. Notice that the computation of variance components for each marker becomes sensitive to the prior specification (Lehermeier et al. 2013). To overcome this limitation, it is possible to conjugate the prior $S^*$ from a Gamma distribution (Gianola 2013).

Another BSR that has become very popular is the Bayesian LASSO proposed by Park and Casella (2008). It is a very consistent algorithm that assigns a double exponential distribution to marker effects (Fig10) in a fashion similar to the original LASSO (Tibshirani 1996). This causes a strong shrinkage (Gianola 2013) with low sensitivity to the prior specification (Lehermeier et al. 2013), but it does not perform variable selection as opposed to its non-Bayesian counterpart.

The Bayesian LASSO (BL) also assigns a variance to each marker, as does BayesA. However, BL computes $\sigma_{b_i}^2$ as a function of the residual variance and a scale parameter

($\tau_i$), thus: $\sigma^2_{\beta_i} = \sigma^2_e \tau^2_i$. The scale parameter $\tau_i^{-2}$ is sampled for each marker i from an inverse Gaussian distribution centered at $\sigma_e \phi / \beta_i$ and with a shape $\phi^2$. The smoothing parameter $\phi^2$ can be sampled from a gamma distribution (de los Campos et al. 2009) with rate $\sum \tau_i^2/2 + r^*$ and shape $p + s^*$, where $r^*$ and $s^*$ are the hyperpriors of rate and shape. Regression coefficients and residual variance are sampled as in BRR and BayesA.

Several algorithms estimate variance components and breeding values either by expressing the relationship among individuals through kinship or by directly regressing molecular markers; furthermore the accuracy of different algorithms changes according to the genetic architecture of the trait (de los Campos et al. 2013). The algorithm with the best learning properties provides the most accurate prediction, which may require breeders and geneticists to evaluate models through cross-validation for each trait.

One may believe that not all markers have a contribution to the trait of interest and that shrinkage does not eliminate markers from the model. In this case, some have proposed adding a variable selection term into the model, which would allow markers to pursue null effect. Indeed, each model presented earlier has an alternative version with variable selection: BayesA becomes BayesB (Meuwissen et al. 2001), BRR becomes BayesCπ (Habier et al. 2011), and BL has an expanded version proposed by Legarra et al. (2011b).

Meuwissen et al. (2001) proposed the first WGR with variable selection using the Metropolis-Hasting algorithm, which proposes that markers be included into the model at random. The proposed changes are accepted only if the model improves. Meuwissen's approach is robust at a high computational cost. Alternatively, there are the following

feasible variable selection algorithms that have been incorporated in the Gibbs sampler (O'Hara and Sillanpää 2009):

1. Stochastic search variable selection (George and McCulloch 1993);
2. Unconditional prior (Kuo and Mallick 1998);
3. Gibbs variable selection (Dellaportas et al. 2002).

We showed the computation of breeding values through kernel and regression methods for the purpose of selection, once these values were free of environmental noise. We also showed that the use of a Gaussian process to estimate breeding values fails to capture the effect of large effect QTL, as opposed to BayesA and BL.

The procedures of screening the whole-genome for large effect QTL by testing one marker at a time conditional to a polygenic term are called genome-wide association studies (GWAS). The polygenic term is used as an efficient way to avoid false-positives by controlling the population structure.

Non-Gaussian WGR methods are capable of capturing major effect alleles and, therefore, can be directly used to perform GWAS. LASSO and BayesCπ have been widely used for detecting QTLs (Colombiani et al. 2012, Fang et al. 2012, Li and Sillanpää 2012, Yi and Xu 2008). Furthermore, a comparison study performed by Legarra et al. (2015) pointed out the superiority of these methods over the traditional mixed models (ie. marker + polygene).

## 2.6. Data Quality Control and Association Analysis

Understanding the underlying genetics of quantitative traits provides basic knowledge for strategies of crop improvement (Sonah et al. 2014). The most common procedure to associate genetics and phenotypes with molecular tools is to find the markers associated

with phenotypes through either linkage or association mapping. Regardless of the genetic resource (ie. type of population), association studies have four fundamental steps: phenotyping, genotyping, mapping, and validation. Validation consists of performing the first three procedures of phenotyping, genotyping, and mapping upon an experimental population specially designed for this purpose (eg. near isogenic lines). Therefore, we will emphasize only the three initial steps.

### 2.6.1 Phenotyping

When traits are governed by many loci, sensitivity to environmental variation increases. It happens because the external stimuli affect the genetic expression of different loci at different levels. In soybeans some complex traits, like yield and drought tolerance, are highly variable across the genome regarding genetic expression (Guimarães-Dias et al. 2012, Le et al. 2011). In the context of minimizing environmental noise in phenotypes, research on field phenomics aims to generate or improve high-throughput and high-precision phenotyping techniques. This omic-integration has primarily helped to improve abiotic stress (Deshmukh et al. 2014).

It is possible to further reduce noise due to field variation through a Gaussian process using spatial statistics, such as kriging (Basso et al. 2000) that allows adjustment for spatial correlation among field trials (Banerjee et al. 2010, Zas 2006). Lado et al. (2013) was able to improve accuracy of genomic prediction in wheat by controlling field variation through spatial adjustments using a simple mixed model with a moving-mean covariate structure.

Kriging methods to control field variation can be used to compliment experimental design and unreplicated trials (Banerjee et al. 2010, Lado et al. 2013). Phenotypic data contains

the actual genetic information, the micro- and macro-environmental variation, and the interactions between environmental and genetic factors.

For this application of kriging, we can employ the following mixed effect model: $\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{Iv} + \mathbf{e}$, where the observed phenotype ($\mathbf{y}$) is a function of some fixed effect ($\mathbf{Xb}$), like block or environment, the genetic effect ($\mathbf{Zu}$) that allows specification of the association among individuals given $\mathbf{u}$~$\mathrm{N}(0, \mathbf{K}\sigma_a^2)$, the field variation ($\mathbf{Iv}$) term in which the spatial relationship (ie. distance between plots in the field) is defined by an exponential or Gaussian kernel ($\mathbf{S}$) such that $\mathbf{v}$~$\mathrm{N}(0, \mathbf{S}\sigma_s^2)$, and the residual term ($\mathbf{e}$) that contains random errors and higher-order interactions. The design matrix of the field variation is an identity matrix because each plot is observed once. According to Zas (2006), it is possible to obtain adjusted phenotypes ($\mathbf{y}^*$) by subtracting the field variation component from the observed phenotype: $\mathbf{y}^* = \mathbf{y} - \mathbf{v}$.

Adjusted phenotypic values provide robust results and many measures can help to evaluate such improvements (Table 4). With reduced environmental noise, genotypes tend to have a more stable performance across environments, which can be measured using a Pearson or Spearman correlation. Another measure of improvement is the increase in broad- and narrow-sense heritabilities, once that more variance is expected to be due to genetic factors.

### 2.6.2 Genotyping

High-throughput genotyping techniques have become very popular in plant breeding (Jarquín et al. 2014, Sohan et al. 2014), often with poor genotyping quality and a large amount of missing data (Halprin and Stephan 2009) that makes mapping and selection challenging (Jarquín et al. 2014, Poland and Rife 2012). Thus, the accurate imputation of

missing loci and good correction of SNP miscalls becomes essential for robust downstream analyses (Marchini and Howie 2010).

Two popular methods of genotypic imputation in plant breeding are random forest and hidden Markov models (HMM) (Swarts et al. 2014, Rutkoski et al. 2013). Random forest is a non-parametric method of prediction, classification, and imputation of mixed data types. It establishes a combination of decision-tree predictors, in which decision trees are bootstrapped to generate random independent vectors that constitute training forests. This is particularly useful for imputing unordered markers. Rutkoski et al. (2013) reported random forest as a promising method to impute genotyping-by-sequencing (GBS) data in wheat.

HMM are commonly employed in genetics and genomics for stochastic modeling of Markov processes, such as the computation of haplotypes. Assuming ordered markers, the HMM estimates the most likely path of states (ie. genotype) based on the transition probability of marker $m^t$ to change state given the previous marker $m^{t-1}$. In genetic terms, the three possible states for a diploid organism with two alleles for a given locus m are: $M_1M_1$, $M_1M_2$, and $M_2M_2$, disregarding linkage phase. HMM is the most common method for imputation of missing genotypes. In addition, Marchini and Howie (2010) showed that HMM can boost power and resolution of genome-wide association studies.

Other quality parameters with a major impact on analysis are the minor allele frequency (MAF) of molecular markers (Tabangin et al. 2009) and the marker ability of carrying a gene. The latter is estimated from the marker heritability (Forneris et al. 2015) when

markers are seen as molecular phenotypes and it used to identify markers that do not follow Mendelian segregation due to biased inheritance of alleles (Glémin 2010).

Minor alleles are very important for population stratification. Wen et al. (2008) found as many as nine subpopulations when evaluating the structure of 393 landraces and 196 native populations of soybeans in China. However, low MAF has two major drawbacks in association analysis: (1) it may increase the rate of false discoveries if one disregards the existence of subpopulation; and (2) even if an allele has major effect but it is only present in a low frequency (Fig11), this particular gene will become undetectable due to the lack of power associated with the low signal-to-noise ratio (Tabangin et al. 2009).

### 2.6.3 Gene Mapping

Recapitulating general ideas of association mapping previously discussed, the procedure starts with estimating the breeding values using a mixed model and testing the increase in likelihood that each marker provides when it is set as a covariate in the model.

Yu et al. (2006) proposed one of the first algorithms for GWAS in the mixed model framework: the unified mixed model (UMM) also known as the $K + Q$ method. The principle of UMM is to use some fixed effect that would contribute to control population structure (**Q**) besides the polygenic term. This usually entails a kernel method using pedigree, genomic data or both to estimate the kinship matrix (**K**). The fixed effect could be some principal components (Eigenvector of the kinship) or another set of categorical variables that indicates to which population individuals belong. However, solving the mixed model for every marker has a great computational burden.

Aulchenko et al. (2007) proposed an approximated method to avoid computing the mixed model every round, the genome-wide association using mixed model and regression (GRAMMAR) algorithm. The authors proposed to fit the animal model first and analyze the residual term as un-structured phenotypes, since the animal model is a Gaussian process incapable of capturing major genes. Although conveniently faster, the original GRAMMAR approach provides biased estimates of SNP effects. A modification of the GRAMMAR algorithm was proposed by Svishcheva et al. (2012) to address this limitation.

Kang et al. (2008) proposed the EMMA algorithm to provide a computational solution for the $K + Q$ model, finding the variance components as an optimization problem that maximizes the restricted log-likelihood (Dempster et al. 1981). EMMA includes some computing tricks, using the Eigen decomposition of the kinship matrix to speed up calculations and alternatives to classical kinship with reduced dimensions.

Even EMMA would be impractical for large datasets and teams have proposed two equivalent approximation methods that do not require the calculation of variance components every round in order to overcome this computational limitation: (1) Kang et al. (2010) proposed EMMA expedited (EMMAX). It generates an empirical relationship matrix to comprise multiple levels of relatedness with no need for principal components; and (2) Zhang et al. (2010) proposed the population parameter previously determined (P3D) algorithm that clusters individuals and estimates variance components first, then finds the optimal values for clusters, fixed effect, and marker for each locus under evaluation.

With efficient incorporation of Eigen terms for the optimization of the likelihood function and factorized markers in the kinship matrix, two newer implementations provide an even more efficient exact method to handle large data. Lippert et al. (2011) proposed the factored spectrally transformed (FaST) algorithm that factorizes markers and Zhou and Stephens (2012) proposed the genome-wide efficient mixed model association (GEMMA) algorithm.

In general, mixed models can increase power and prevent false positives at a reasonable cost, but this approach also presents some pitfalls, as summarized by Yang et al. (2014), such as the loss of power in case-control studies and double-fitting markers into the model. Double-fitting involves using markers both to build the kinship and as a covariate when the marker is being evaluated.

The use of WGR as a GWAS method could easily satisfy the limitation of double-fitting once each marker effect is inferred from a full conditional distribution that takes into account all other parameters. As shown in Figure 12, three other tricks were proposed by Wang (2015) and implemented by Xavier et al. (2015) to further increase power and resolution of GWAS: (1) Treat markers as a random effect (ie. empirical Bayes algorithm) to shrink the background noise to zero; (2) Use a sliding window to overcome double-fitting markers, removing the local markers from the polygenic term; (3) If any stratification factor is known a priori, then markers can be treated as the interaction *marker* × *subpopulation*.

## 2.7. Conclusions

The various models and algorithms all make important assumptions. Knowing how the computations work may help breeders to optimize statistical analysis and make better

decisions. Most statistical procedures in breeding theory are based on Gaussian process and can be computed through mixed models using kernels and regression models. We have presented here the flexibility possible by utilizing principles of machine learning and mixed models for selection, prediction, and mapping, as well as inferences of variance components.

References

Acquaah G (2009) Principles of plant genetics and breeding. John Wiley and Sons. The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK.

Aulchenko YS, De Koning DJ, Haley C (2007) Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. Genetics 177(1): 577-585.

Banerjee S, Finley AO, Waldmann P, Ericsson T (2010) Hierarchical spatial process models for multiple traits in large genetic trials. Journal of the American Statistical Association 105(490): 506-521.

Basso B, Ritchie JT, Pierce FJ, Braga RP, Jones JW (2001) Spatial validation of crop models for precision agriculture. Agricultural Systems 68(2): 97-112.

Beavis WD (1998) QTL analyses: power, precision, and accuracy. Molecular dissection of complex traits, 145-162.

Bernardo R, Nyquist WE (1998) Additive and testcross genetic variances in crosses among recombinant inbreds. Theoretical and applied genetics 97(1-2): 116-121.

Carvalho AD, Fritsche Neto R, Geraldi IO (2008) Estimation and prediction of parameters and breeding values in soybean using REML/BLUP and Least Squares. Crop Breeding and Applied Biotechnology 8(3): 219-224.

Cleveland DA, Soleri D (Eds.). (2002). Farmers, scientists, and plant breeding: integrating knowledge and practice. CABI.

Colombani C, Legarra A, Fritz S, Guillaume F, Croiseau P, Ducrocq V, Robert-Granié C (2013) Application of Bayesian least absolute shrinkage and selection operator (LASSO) and BayesCπ methods for genomic selection in French Holstein and Montbéliarde breeds. Journal of dairy science 96(1): 575-591.

Crow JF, Kimura M (1970) An introduction to population genetics theory. An introduction to population genetics theory.

Dardanelli JL, Balzarini M, Martínez MJ, Cuniberti M, Resnik S, Ramunda SF, et al. (2006) Soybean maturity groups, environments, and their interaction define mega-environments for seed composition in Argentina. Crop science 46(5): 1939-1947.

Dellaportas P, Forster JJ, Ntzoufras I. (2002) On Bayesian model and variable selection using MCMC. Statistics and Computing 12(1): 27-36.

de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MP (2013) Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193(2): 327-345.

de Los Campos G, Gianola D, Rosa GJ, Weigel KA, Crossa J (2010) Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genetics Research 92(04): 295-308.

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 1-38.

Dempster AP, Rubin DB, Tsutakawa RK (1981) Estimation in covariance components models. Journal of the American Statistical Association 76(374): 341-353.

Deshmukh RK, Sonah H, Patil G, Chen W, Prince S, Mutava R, et al. (2014) Integrating omic approaches for abiotic stress tolerance in soybean. Plant Genetics and Genomics, 5, 244.

Diffey S, Welsh A, Smith A, Cullis BR (2013) A faster and computationally more efficient REML (PX) EM algorithm for linear mixed models. Centre for Statistical and Survey Methodology, University of Wollongong, Working Paper 2-13, 8.

Egli DB (2008a) Soybean yield trends from 1972 to 2003 in mid-western USA. Field crops research 106(1): 53-59.

Egli DB (2008b) Comparison of corn and soybean yields in the United States: Historical trends and future prospects. Agronomy journal, 100(Supplement_3), S-79.

Fang M, Jiang D, Li D, Yang R, Fu W, Pu L, et al. (2012) Improved LASSO priors for shrinkage quantitative trait loci mapping. Theoretical and Applied Genetics 124(7): 1315-1324.

Farrall M (2004) Quantitative genetic variation: a post-modern view. Human molecular genetics, 13(suppl 1), R1-R7.

Fisher RA (1918). The correlation between relatives on the supposition of Mendelian inheritance. Transactions of the Royal Society of Edinburgh, 52: 399-433.

Forneris NS, Legarra A, Vitezica ZG, Tsuruta S, Aguilar I, Misztal I, Cantet RJ (2015) Quality Control of Genotypes Using Heritability Estimates of Gene Content at the Marker. Genetics 199(3): 675-681.

García-Cortés LA, Sorensen D (1996) On a multivariate implementation of the Gibbs sampler. Genetics Selection Evolution 28(1): 121-126.

Geman S, Geman D (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. Pattern Analysis and Machine Intelligence, IEEE Transactions on (6): 721-741.

Gengler N, Mayeres P, Szydlowski M (2007) A simple method to approximate gene content in large pedigree populations: application to the myostatin gene in dual-purpose Belgian Blue cattle. Animal 1(1): 21-28.

George EI, McCulloch RE (1993) Variable selection via Gibbs sampling. Journal of the American Statistical Association 88(423): 881-889.

Gianola D (2013) Priors in whole-genome regression: the Bayesian alphabet returns. Genetics 194(3): 573-596.

Gianola D, Fernando RL, Stella A (2006) Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics 173(3): 1761-1776.

Gianola D, Foulley JL, Fernando RL (1986) Prediction of breeding values when variances are not known. Genetics Selection Evolution 18(4): 485-498.

Gilmour AR, Gogel BJ, Cullis BR, Thompson R (2009) ASReml user guide release 3.0. VSN International Ltd, Hemel Hempstead, UK.

Gilmour AR, Thompson R, Cullis BR (1995) Average information REML: An efficient algorithm for variance parameter estimation in linear mixed models. Biometrics, 1440-1450.

Glémin S (2010) Surprising fitness consequences of GC-biased gene conversion: I. Mutation load and inbreeding depression. Genetics 185(3): 939-959.

González-Camacho JM, De Los Campos G, Pérez P, Gianola D, Cairns JE, Mahuku G, et al. (2012) Genome-enabled prediction of genetic values using radial basis function neural networks. Theoretical and Applied Genetics 125(4): 759-771.

Guimarães-Dias F, Neves-Borges AC, Viana AAB, Mesquita RO, Romano E, Grossi-de-Sa MDF, et al. (2012) Expression analysis in response to drought stress in soybean: Shedding light on the regulation of metabolic pathway genes. Genetics and molecular biology 35(1): 222-232.

Habier D, Fernando RL, Kizilkaya K, Garrick DJ (2011) Extension of the Bayesian alphabet for genomic selection. BMC bioinformatics 12(1): 186.

Halperin E, Stephan DA (2009) SNP imputation in association studies. Nature biotechnology 27(4): 349-351.

Henderson CR (1984) Applications of linear models in animal breeding. University of Guelph.

Henderson CR (1975) Best linear unbiased estimation and prediction under a selection model. Biometrics, 423-447.

Hofer A (1998) Variance component estimation in animal breeding: a review. Journal of Animal Breeding and Genetics 115(1|6): 247-265.

Imhof LA, Nowak MA (2006) Evolutionary game dynamics in a Wright-Fisher process. Journal of mathematical biology 52(5): 667-681.

Jarquín D, Kocak K, Posadas L, Hyma K, Jedlicka J, Graef G, Lorenz A (2014) Genotyping by sequencing for genomic prediction in a soybean breeding population. BMC genomics 15(1): 740.

Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, et al. (2010) Variance component model to account for sample structure in genome-wide association studies. Nature genetics 42(4): 348-354.

Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E (2008) Efficient control of population structure in model organism association mapping. Genetics 178(3): 1709-1723.

Kimura M, Crow JF (1964) The number of alleles that can be maintained in a finite population. Genetics 49(4): 725.

Kuo L, Mallick B (1998) Variable selection for regression models. Sankhya: The Indian Journal of Statistics, Series B, 65-81.

Lado B, Matus I, Rodríguez A, Inostroza L, Poland J, Belzile F, et al. (2013) Increased Genomic Prediction Accuracy in Wheat Breeding Through Spatial Adjustment of Field Trial Data. G3: Genes| Genomes| Genetics 3(12): 2105-2114.

Lander ES, Botstein D (1989) Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121(1): 185-199.

Le DT, Nishiyama R, Watanabe Y, Mochida K, Yamaguchi-Shinozaki K, Shinozaki K, Tran LSP (2011) Genome-wide survey and expression analysis of the plant-specific NAC transcription factor family in soybean during development and dehydration stress. DNA research, dsr015.

Legarra A, Croiseau P, Sanchez MP, Teyssèdre S, Sallé G, Allais S, et al. (2015) A comparison of methods for whole-genome QTL mapping using dense markers in four livestock species. Genetics Selection Evolution 47(1) 6.

Legarra A, Ricardi A, Filangi O (2011) GS3: Genomic Selection, Gibbs Sampling, Gauss-Seidel. snp.toulouse.inra.fr/~alegarra/.

Legarra A, Robert-Granié C, Croiseau P, Guillaume F, Fritz S (2011) Improved Lasso for genomic selection. Genetics research 93(01): 77-87.

Legarra A, Misztal I (2008) Technical note: Computing strategies in genome-wide selection. Journal of dairy science 91(1): 360-366.

Lehermeier C, Wimmer V, Albrecht T, Auinger HJ, Gianola D, Schmid VJ, Schön CC (2013) Sensitivity to prior specification in Bayesian genome-based prediction models. Statistical applications in genetics and molecular biology 12(3): 375-391.

Li Z, Sillanpää MJ (2012) Overview of LASSO-related penalized regression methods for quantitative trait mapping and genomic selection. Theoretical and Applied Genetics 125(3): 419-435.

Lim C (1997) An econometric classification and review of international tourism demand models. Tourism economics 3(1): 69-81.

Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D (2011) FaST linear mixed models for genome-wide association studies. Nature Methods 8(10): 833-835.

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

MacLeod IM, Hayes BJ, Goddard ME (2014) The Effects of Demography and Long-Term Selection on the Accuracy of Genomic Prediction with Sequence Data. Genetics 198(4):1671-1684.

Marchini J, Howie B (2010) Genotype imputation for genome-wide association studies. Nature Reviews Genetics 11(7): 499-511.

Matilainen K, Mäntysaari EA, Lidauer MH, Strandén I, Thompson R (2013) Employing a Monte Carlo Algorithm in Newton-Type Methods for Restricted Maximum Likelihood Estimation of Genetic Parameters. PloS One 8(12) e80821.

Meuwissen TMH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157(4): 1819-1829.

Meyer K (2007) WOMBAT: A tool for mixed model analyses in quantitative genetics by restricted maximum likelihood (REML). Journal of Zhejiang University Science B 8(11): 815-821.

Meyer K (1989) Restricted maximum likelihood to estimate variance components for animal models with several random effects using a derivative-free algorithm, Genetics Selection Evolution (21): 317-340.

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, Montpellier, France, August, 2002. Session 28. (pp. 1-2). Institut National de la Recherche Agronomique (INRA).

Myles S, Peiffer J, Brown PJ, Ersoz ES, Zhang Z, Costich DE, Buckler ES (2009) Association mapping: critical considerations shift from genotyping to experimental design. The Plant Cell Online 21(8): 2194-2202.

Nelder JA, Mead R (1965) A simplex method for function minimization. The computer journal 7(4): 308-313.

Nyquist WE, Baker RJ (1991) Estimation of heritability and prediction of selection response in plant populations. Critical reviews in plant sciences 10(3): 235-322.

O'Hara RB, Sillanpää MJ (2009) A review of Bayesian variable selection methods: what, how and which. Bayesian analysis 4(1): 85-117.

Orr HA (2005) The genetic theory of adaptation: a brief history. Nature Reviews Genetics 6(2): 119-127.

Park T, Casella G (2008) The bayesian lasso. Journal of the American Statistical Association 103(482): 681-686.

Patterson HD, Thompson R (1971) Recovery of inter-block information when block sizes are unequal. Biometrika 58(3): 545-554.

Pérez P, de los Campos G (2014) Genome-wide regression and prediction with the BGLR statistical package. Genetics, genetics-114.

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.

Poland JA, Rife TW (2012) Genotyping-by-sequencing for plant breeding and genetics. The Plant Genome 5(3): 92-102.

Rasmussen CE (2004) Gaussian processes in machine learning. In Advanced Lectures on Machine Learning, pp. 63-71. Springer Berlin Heidelberg.

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 Science 54(3): 939-943.

Rincker K, Nelson R, Specht J, Sleper D, Cary T, Cianzio SR, et al. (2014) Genetic improvement of US soybean in Maturity Groups II, III, and IV. Crop Science.

Robinson GK (1991) That BLUP is a good thing: The estimation of random effects. Statistical science, 15-32.

Rutkoski JE, Poland J, Jannink JL, Sorrells ME (2013) Imputation of unordered markers and the impact on genomic selection accuracy. G3: Genes| Genomes| Genetics 3(3): 427-439.

Searle SR (1979) Notes on variance component estimation: a detailed account of maximum likelihood and kindred methodology. Paper BU-673M, Biometrics Unit, Cornell University.

Sonah H, O'Donoughue L, Cober E, Rajcan I, Belzile F (2014) Identification of loci governing eight agronomic traits using a GBS|GWAS approach and validation by QTL mapping in soya bean. Plant biotechnology journal.

Sorensen D, Gianola D (2002) Likelihood, Bayesian, and MCMC methods in quantitative genetics. Springer.

Specht J E, Hume DJ, Kumudini SV (1999) Soybean yield potential-a genetic and physiological perspective. Crop Science 39(6): 1560-1570.

Strandén I, Christensen OF (2011) Allele coding in genomic evaluation. Genet Sel Evol 43(1).

St. Martin SK (1982) Effective population size for the soybean improvement program in maturity groups 00 to IV. Crop Science 22(1): 151-152.

Svishcheva GR, Axenovich TI, Belonogova NM, van Duijn CM, Aulchenko YS (2012) Rapid variance components-based method for whole-genome association analysis. Nature genetics 44(10): 1166-1170.

Swarts K, Li H, Romero Navarro JA, An D, Romay MC, Hearne S, et al. (2014) Novel Methods to Optimize Genotypic Imputation for Low-Coverage, Next-Generation Sequence Data in Crop Plants. The Plant Genome 7(3).

Tabangin ME, Woo JG, Martin LJ (2009, December) The effect of minor allele frequency on the likelihood of obtaining false positives. In BMC proceedings, Vol. 3, No. Suppl 7, p. S41. BioMed Central Ltd.

Tibshirani R (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological): 267-288.

VanRaden PM (2008) Efficient methods to compute genomic predictions. Journal of dairy science 91(11): 4414-4423.

Wang CS, Rutledge JJ, Gianola D (1993) Marginal inferences about variance components in a mixed linear model using Gibbs sampling. Genetics Selection Evolution 25:41-62.

Wang Q (2015) An Empirical Bayes Method for Genome-Wide Association Studies. W799/Statistical Genomics. In Plant and Animal Genome XXXII.

Wen ZX, Zhao TJ, Zheng YZ, Liu SH, Wang CE, Wang F, Gai JY (2008) Association analysis of agronomic and quality traits with SSR markers in Glycine max and Glycine soja in China: I. Population structure and associated markers. Acta Agronomica Sinica 34(7): 1169-1178.

Wricke G, Weber E (1986) Quantitative genetics and selection in plant breeding. Walter de Gruyter.

Wright S (1930) Evolution in Mendelian populations. Genetics 16(2): 97.

Wright S (1922) Coefficients of inbreeding and relationship. American Naturalist, 330-338.

Xavier A, Xu S, Muir WM, and Rainey KM (2015) NAM: Association Studies in Multiple Populations. Bioinformatics, btv448.

Xu S (2013) Mapping quantitative trait loci by controlling polygenic background effect. Genetics 195(4): 1209-1222.

Xu S (2003a) Theoretical basis of the Beavis effect. Genetics 165(4): 2259-2268.

Xu S (2003b) Estimating polygenic effects using markers of the entire genome. Genetics 163(2): 789-801.

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

Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL (2014) Advantages and pitfalls in the application of mixed-model association methods. Nature genetics 46(2): 100-106.

Yi N, and Xu S (2008) Bayesian LASSO for quantitative trait loci mapping. Genetics 179(2): 1045-1055.

Yu J, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF, et al. (2005) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature genetics 38(2): 203-208.

Zas R (2006) Iterative kriging for removing spatial autocorrelation in analysis of forest genetic trials. Tree genetics and genomes 2(4): 177-185.

Zeng ZB Hill WG (1986) The selection limit due to the conflict between truncation and stabilizing selection with mutation. Genetics 114(4): 1313-1328.

Zhang Z, Ersoz E, Lai CQ, Todhunter RJ, Tiwari HK, Gore MA, et al. (2010) Mixed linear model approach adapted for genome-wide association studies. Nature genetics 42(4): 355-360.

Zhang LX, Kyei-Boahen S, Zhang J, Zhang MH, Freeland TB, Watson CE, Liu X (2007) Modifications of optimum adaptation zones for soybean maturity groups in the USA. Crop Management, 6(1).

Zhou X, Stephens M (2012) Genome-wide efficient mixed-model analysis for association studies. Nature genetics 44(7): 821-824.

Zhou X, Stephens M (2014) Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nature methods 11(4): 407-409.

# CHAPTER 3: RELEVANT FACTORS FOR GENOMIC PREDICTION IN SOYBEANS

ABSTRACT

Economically relevant traits in plant breeding usually have complex genetic architectures. A large number of genes control the quantitative nature of these traits, each with a small contribution to the phenotype. For these traits, genomic selection seems to have attractive features and promises to boost genetic gains. Our goal was to evaluate genome-wide prediction of soybean (*Glycine max*) agronomic traits and yield components using machine learning approaches to evaluate different scenarios for implementing genomic selection. Novel multi-parent experimental populations known as next-generation populations have statistical and genetic properties ideal for association studies and prediction, which make these populations a great resource for supervised-learning experiments. We assessed a set of factors known to influence the accuracy of prediction using a nested association mapping population. These factors included training population size, genotyping density, prediction model, and phenotypic adjustment. Our overall model choice was a combination of the kernel and additive models, RKHS+BayesB. Higher genotyping density marginally improved prediction ability. Our study finds that breeding programs seeking efficient genomic selection would best allocate resources by increasing training-population size in combination with methods to improve quality of the phenotypic data.

## 3.1 Introduction

An increasing need for food quality and production requires fast and efficient genetic improvement of plants and animals. Nonetheless, traits that are relevant to meeting global food demands have complex genetic architecture, sensitive to environmental factors; in other words, low heritability. A large number of genes control the quantitative nature of these traits, each with a small contribution to the phenotype. Hence the use of genomic information for breeding purposes represents an important boost of genetic gains in low-heritability traits (Muir 2007).

Many breeding techniques designed for animal improvement have been successful for plants too (Cowling et al. 2015). Among those, the introduction of genomic selection (GS) into the plant breeding pipeline is promising and has attractive features (Heffner et al. 2009; Jannink et al. 2010; Nakaya and Isobe 2012). Plants provide an excellent framework for testing theory and applications related to GS because of the large number of offspring possible, their ability to be cloned and inbred easily, the short life-cycles of annuals, and their potential genomic properties favorable to GS, such as high levels of linkage disequilibrium (LD) (Hyten et al. 2006 2007). Yet GS must take many aspects into account to optimize genetic gains by using genomic data to best allocate resources (Meuwissen et al. 2001; Poland 2015).

In silico supervised machine learning experiments can determine which factors are relevant for this process using real and simulated data through cross-validation by testing different scenarios and letting the data "speak for itself", thereby indicating the combinations of methods and parameters that would provide the most satisfactory results. Credible inferences on complex traits require thorough evaluation of genetic architecture (Wimmer

et al. 2013). Consequently, designing a robust genome-wide prediction (GWP) system is a major concern. Most prediction models differ with respect to the assumptions they make over the behavior of marker effects (Kärkkäinen and Sillanpää 2012; Gianola 2013). The assumptions that best correspond to the real genetic architecture of the trait are likely to provide more reliable predictions (de los Campos et al. 2013). Without performing learning experiments to evaluate different assumptions, it is not possible to determine which model would offer the most consistent prediction (Habier et al. 2011; Okser et al. 2014).

Genomic enhanced breeding values (GEBVs), estimated through whole-genome regression, can help breeding programs to speed up the breeding process and save resources in multiple ways (Heffner et al. 2009; Endelman et al. 2014). Selection based on GEBVs is more reliable than phenotypes alone or the traditional Quantitative Trait Loci (QTL) pyramiding (Nakaya and Isobe 2012). Muir (2007) has shown in simulated studies that GEBVs also provide more genetic gains over the long term when compared to pedigree-based breeding values. In the plant breeding pipeline, GEBVs can help to: select un-phenotyped material (Heffner et al. 2008), which is particularly useful when the phenotyping process is somehow challenging; perform more accurate selection of advanced lines by adding the information of relatives (Endelman et al. 2014); identify and incorporate useful germplasm into the breeding pipeline (Chung et al. 2014); and elect parents for crosses with higher chances of transgressive segregation based on breeding values and genomic distance (Mohammadi et al. 2015). Yet studies of GWP are important because the methodology for GEBV estimation is not fully understood and the outputs may vary from trait to trait and crop to crop. According to Wimmer et al. (2013), the contribution to the prediction models of heritability, genotyping, and phenotyping when

applied to real data is not clear. In this study we attempt to evaluate the importance of a set of parameters that contribute to prediction of six complex traits in soybean. These parameters include genotyping density, training population size, phenotypic adjustment, environment, and combinations of prediction models.

Prediction studies often provide conflicting results that vary according to the genetic basis of the population under evaluation (de los Campos et al. 2013). Morrell et al. (2012) suggest using next-generation populations (NGPs) to maximize statistical properties of genomic studies, such as the power and resolution of genome-wide association mapping. NGPs are generated through controlled crosses to have reduced population structure and ascertainment bias. It is possible to further optimize genotypic information in NGPs by taking advantage of known haplotypes (Xu 2013b; Xavier et al. 2015). The two most common NGPs are nested association mapping (NAM) and multi-parent advanced generation intercross (MAGIC) populations. NAM is also seen as a subset of a MAGIC population in which multiple founders are crossed to a single standard parent as opposed to random inter-mating. Development of NAM panels seeks to capture "useful diversity" for the dissection of the genetic architecture of complex traits (Yu et al. 2008).

Guo et al. (2012) performed the first published study of GWP using a NAM population by analyzing three maize traits using individual bi-parental families as opposed to the NAM population as a whole. What NAM represents goes far beyond bi-parental populations (Hamblin et al. 2011) and thus, in this study we are treating NAM as a large population with complex genomic structure (Jannink et al. 2010), what provide an ideal scenario to study learning properties, in other words, how well statistical models learn from data and it affects prediction. The main objective of this study was to evaluate which factors have

the greatest impact on genomic prediction in soybeans using real data from a NAM population through supervised machine learning experiments.

## 3.2 Materials and Methods

### 3.2.1 Genetic material

To evaluate GWP we used SoyNAM, a soybean nested-association panel. The SoyNAM population (soynam.org) contains 5555 recombinant inbred lines (RIL) with maturity ranging from late maturity group II to early IV, derived from 40 biparental populations that share IA3023 as a common parent. Among the 40 founder parents, 17 lines are U.S. elite public germplasm, 15 have diverse ancestry, and eight are plant introductions. Lines were genotyped in the F5 generation with a 5k Single-Nucleotide Polymorphism (SNP) chip. The SNP chip was specially designed for this population, which called SNPs from the parental sequencing data to minimize the ascertainment bias associated to the nature of the genotyping technology (Daetwyler et al. 2013; Heslot et al. 2013).

After removing non-segregating SNPs, we coded alleles as 012 (Strandén and Christensen 2011) and imputed missing loci using random forest implemented in the R package missForest (Stekhoven and Buhlmann 2012). To reduce excess rare variants, we removed markers with a minor allele frequency (MAF) lower than 0.15 (Heslot et al. 2013). We also removed redundant markers so that the genotypic data would represent natural bins (Xu 2013b). The genotypic data contained 6.12% of heterozygous loci, slight lower than the expectation for an F5 generation (ie. 6.25%). Pairwise linkage disequilibrium between SNPs was phased via expectation-maximization (Asmussen et al. 1996) measured in terms of $r^2$ to illustrate the configuration of linkage blocks in this population, the LD heat map is shown in Figure 13.

We performed quality control using the NAM package by Xavier et al. (2015). To evaluate the impact of genotypic coverage on GWP, we tested subsets of the genotypic data as proposed by Meuwissen et al. (2001), with the whole panel, half panel, and quarter panel, corresponding to the 4077, 2039, and 1020 SNP markers respectively. The subsets containing half and a quarter of the whole panes were obtained by systematically picking one every two and four markers, respectively.

Afterwards, 196 lines had nearly identical genotypes (>95%) but remained in the prediction analysis. The relationship among lines in shown in Figure 14, where it is notable that the overall relationship within family is slightly higher than between family, since all individuals are either full- or half-siblings.

### 3.2.2 Phenotypes

Phenotypic data was collected from the SoyNAM population in 2013 and 2014 in West Lafayette, Indiana. In both years, lines were planted during the third week of May in two-row plots, 2.9m × 0.76m, at a density of approximately 36 plants/$m^2$.

Collection of phenotypic measurements proceeded as follows: Grain yield was measured in grams per plot adjusted to 13% of moisture. Days to maturity was collected three times a week, with back and forward scoring of plots that matured in the intervals. Number of reproductive nodes and pods in the main stem were counted in R7-R8, measuring 3 and 6 plants per plot for 2013 and 2014 respectively, with the count of pods per node (P/N) being the ratio of these data points.

### 3.2.3 Prediction Models

Two main types of prediction method are widely used in GWP; these are parametric and non-parametric prediction. Parametric methods are based on estimating the additive effect of allele substitution to molecular markers, and breeding values are computed as the sum of marker values of genotyped individual. Non-parametric methods work in non-linear fashion (Peréz-Rodríguez et al. 2012), which is particularly useful for the prediction of highly epistatic traits (Howard et al. 2014). Non-parametric methods include neural networks, random forest and kernel regressions.

Kernel regression is the most popular non-parametric method. Molecular markers are used to estimate genomic relationship among all genotypes, also known as kinship, and the breeding values are computed as the additive genetic-value that each individual contributes to its relatives. Kernel methods are Gaussian process that follow the Fisher's infinitesimal model, they do not assign values to markers and, therefore, are not capable of recognize large-effect QTLs (Sorensen and Gianola 2002). For this reason, genome-wide association studies use kernels to control the effect of genetic background (Bernardo 2013). Kernel methods were used in plant and animal breeding prior to the existence of molecular markers, applying pedigree information to generate the kinship among individuals (Bernardo 2010), the so-called animal model (Henderson 1984).

We tested the prediction performance of five additive models (parametric), two kernel models (non-parametric), and each combination of both on each of the six soybean traits. The combination of additive and kernel methods is a strategy of ensemble learning that seeks to use the kernel to account for polygenic background and the additive model to capture the marker effects (Kärkkäinen and Sillanpää 2012). This practice has commonly

been used to incorporate pedigree information into prediction models (Muir 2007, de los Campos et al. 2009, Heffner et al. 2009), but we used the molecular data to represent the relationship among genotypes instead (Howard et al. 2014).

The models we evaluated were BayesA, BayesB, BayesC, the Bayesian best linear unbiased predictor (BLUP), the Bayesian least absolute shrinkage and selection operator (BLASSO), and two kernel models, the reproducing kernel Hilbert spaces (RKHS) and the genomic best linear unbiased predictor (GBLUP). We represent the general model that describes the prediction employing both parametric and non-parametric terms in this study as

$$\boldsymbol{y} = \mu + \boldsymbol{X}\boldsymbol{\beta}\boldsymbol{\gamma} + \boldsymbol{\psi} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{y}$ is the response variable with $n$ observations, $\mu$ is the intercept, $\boldsymbol{X}$ is an $n \times p$ design matrix containing $p$ markers, $\boldsymbol{\beta}$ is the vector with length $p$ of marker effects identically distributes as normal, $t$ or double-exponential distribution according to the model's prior assumption, $\boldsymbol{\gamma}$ is a vector of zeros and ones binomially distributed that indicates which markers are included into the model, $\boldsymbol{\psi}$ is the polygenic term of the $n$ observations, assumed to be normally distributed as $\boldsymbol{\psi} \sim \mathrm{N}(0, \mathbf{K}\sigma_{\psi}^{2})$ where $\mathbf{K}$ represents the kinship among lines, and $\boldsymbol{\varepsilon}$ is the vector of residuals with length $n$, assumed to be normally independently distributed $\boldsymbol{\varepsilon} \sim \mathrm{N}(0, \mathrm{I}\sigma_{\varepsilon}^{2})$.

From the Bayesian standpoint, the parametric models BLUP, BayesA, BayesB, BayesC, and BLASSO (Meuwissen et al. 2001; Park and Casella 2008; Habier et al. 2011) differ in their assumptions over the prior distribution of marker effects ($\boldsymbol{\beta}\boldsymbol{\gamma}$). BLUP assumes that marker effects are normally distributed with the same variance, while BayesA assumes that

marker effects are *t* distributed as an infinite mixture of normals with independent variances. BayesB and BayesC, so-called slab priors, are equivalent to BayesA and BLUP with a variable selection (O'Hara and Sillanpää 2009) that allows markers to have zero effect with a probability of $1-\pi$, characterizing the prior distribution of marker effects as a mixture of binomial with *t* (BayesB) or normal (BayesC). BLASSO assigns a double-exponential density to marker effects that causes a strong shrinkage of effects toward zero but does not assign a zero effect, unlike the original LASSO (Tibshirani 1996).

Why does that matter? Double-exponential and *t* distributions have thick tails that allow markers to have large effect, which is a valid assumption for traits controlled by major genes (Kärkkäinen and Sillanpää 2012). BLUP and kernel-based procedures are Gaussian processes, meaning that they may not capture the existence of large-effect QTL (Sorensen and Gianola 2002). Due to the independent variance assigned to each marker by BayesA and BayesB, these models are sensitive to prior specification, and are considered weakly regularized and prone to overfit the data (Gianola 2013), however, to our knowledge, no literature have observed this trend.

With regard to the kernel models, we defined RKHS based on the kernel average model proposed by de los Campos et al. (2010). It utilizes three Gaussian kernels expressed as $\exp(-\mathbf{E}^2/\rho)$, where $\mathbf{E}$ represents the genetic-distance among genotypes computed as the Euclidean distance. The three kernels differ by the bandwidth parameter $\rho$, which represents three extreme values that the bandwidth could take, thus dismissing the need for calibration (González-Camacho et al. 2012). The GBLUP model is based on a single linear kernel (Xu 2013a) known as a realized genomic relationship matrix (GRM).

When the model included additive and polygenic term, both markers and kernels were fitted together. The regularization of markers and of each kernel occurs independently. The linear model was solved via Markov chains Monte Carlo (MCMC). The use of Gibbs sampling algorithm reduces the problem dimensionality by computing each term of the model, one at a time. Computing all parameters many times generates their distribution *a posteriori*, and the final estimator of each parameter is obtained by averaging out this distribution.

We used the R package BGLR to fit the genomic prediction models (Pérez and de los Campos 2014). The in-depth theoretical bases for the model building, algorithms and hyper-parameters are described elsewhere (Sorensen and Gianola 2002; Kärkkäinen and Sillanpää 2012; Gianola 2013; de los Campos et al. 2013; Pérez and de los Campos 2014).

### 3.2.4 Phenotypic Adjustment

Accounting for field variation in the phenotypic BLUPs can increase the genomic predictability and the response to selection (Lado et al. 2013). This pre-adjustment of phenotypic data is performed by the use of checks or blocks, or by removing the autocorrelation associated with plot-by-plot variation among field trials (Zas 2006). This study compared three scenarios, including no adjustment and two phenotype correction methods that use spatial statistics known as kriging. The general model computing spatial coefficients can be described as

$$\boldsymbol{y} = \mu + \boldsymbol{\psi} + f(\boldsymbol{x}) + \boldsymbol{\varepsilon}$$

where $\boldsymbol{y}$ is the observed phenotype, $\mu$ is the intercept, $\boldsymbol{\psi}$ is the polygenic term defining the genetic relationship among lines, $f(\boldsymbol{x})$ is a function that describes the microenvironmental

relationship among field trials and $\varepsilon$ is the residual term. A genetic term must be jointly fitted with the spatial variation term (Cappa and Cantet 2008) to avoid undesirable consequences such as bias and heterogeneous variance (de los Campos et al. 2013). We computed the adjusted phenotypic values as $\boldsymbol{y}^* = \boldsymbol{y} - f(\boldsymbol{x})$ (Zas 2006).

The kernel defining the field relationship among entries was based on the Euclidean distance $\mathbf{E}$ between plots in field, expressed as an exponential kernel $f(\boldsymbol{x}) = \exp(-\mathbf{E}/\rho)$ with a bandwidth parameter $\rho = 3.5$ found through cross-validation. For this given kernel, the relationship among plots is presented in Figure 15, where the horizontal correlation with neighbor plots is higher than vertical because field plots are rectangular.

The two models under evaluation differ by the polygenic term $\boldsymbol{\psi}$ that accounts for the genetic relationship among lines. Thus, we tested raw phenotypes with no adjustment (NO), the use of a linear kernel (LK) and the use of three Gaussian kernels (GK) to describe the kinship (Piepho 2009), the same kernels used for genomic prediction in the models GBLUP and RKHS. We hypothesized that estimation of the genetic term with multiple kernels would provide a more accurate distinction between the variation due to field and genetics than using regularized processes (Okser et al. 2014). We computed coefficients using the algorithm previously described by de los Campos et al. (2010) to solve kernel-based models.

Data adjusted by GK presented distribution nearly identical to the raw values. When the traits were adjusted with LK the distribution of phenotypes was observe slightly shrunken towards the mean, which could generate upward bias in subsequent prediction analysis.

3.2.5 Predictive Ability

Predictive ability (PA) is a standard measure to evaluate the robustness of a prediction. Lehermeier et al. (2013) defined PA as the correlations between predicted ($\hat{y}$) and observed values ($y$) and accuracy as PA divided square-root of heritability ($r_{y,\hat{y}}/h$). The prediction parameters are computed through *k*-fold cross validation.

To evaluate the effect of training population size, we sampled subsets of 250, 500, 1000, 2000, 3000, and 4000 lines at random as a training set to predict a validation set of 500 lines not included in the training set. This study, therefore, evaluated data with $k$-fold scheme where $k = \{0.5,1,3,5,9\}$. We performed 20 cross validations for each combination of the six population sizes (ie. value of $k$ above), six traits, two years, seventeen prediction models, three phenotypic adjustments, and three densities of marker coverage.

3.2.6. Trait Heritability

We estimated heritability ($h^2$) for each combination of trait, year, and phenotypic adjustment by restricted maximum log-likelihood (REML) using the EMMA algorithm (Kang et al. 2008) as implemented by Xavier et al. (2015) to solve a mixed model with a genomic covariance structure. The mixed model is defined in probabilistic terms as $y \sim N(\mu, ZKZ\sigma_a^2 + I\sigma_e^2)$, where $y$ is the phenotype of a given trait by year, $\mu$ is the overall mean, $Z$ is the incidence matrix of genotypes, $K$ is the GRM, $\sigma_a^2$ is the additive genetic variance, and $\sigma_e^2$ is the residual variance. We computed heritabilities as $h^2 = \sigma_a^2/(\sigma_a^2 + \sigma_e^2/r)$ with $r = 1$ replication.

We limit the scope of the study to the impact of multiple factor on heritability, PA and accuracy of GWP. Yet, we recognize that other suitable measures of prediction properties for comparison studies were suggested by Hastie et al. (2005) and Daetwyler et al. (2013)

could have been used to identify problems with model fit, including mean squared prediction error and prediction bias.

### 3.2.7 Statistical Inference

In a practical scenario, statistical significance does not always reflect relevance. Hence we are approaching the statistical analysis of GWP using principles of Bayesian decision theory. This method leads to a simple interpretation of the statistical inference, indicating the probability of a given level to be highest or overperform another level. The inferences on data were based on predictive ability using a hierarchical Bayesian model, one factor at a time, with a posterior distribution shaped as

$$\pi(\Theta, \Sigma \mid X) \propto f(X \mid \Theta, \Sigma)\pi(\Theta \mid \Sigma)\pi(\Sigma)$$

where $\Theta = (\theta_1, \theta_2, \ldots, \theta_p)$ and $\Sigma = (\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2)$ for a factor with $p$ levels. The distribution of the $i^{th}$ level is $x_{ij} = N(\theta_i, \sigma_i^2)$, in which the parameter $\theta_i$ is normally distributed as $N(\mu, \tau^2)$ and the variance $\sigma_i^2$ is inverse-Gamma distributed $IG(\alpha, \beta)$. We set the prior of $\theta$ as normal distribution with the mean and variance of the overall data ($\mu = 0.379$ and $\tau^2 = 0.016$), and the inverse-Gamma prior of each $\sigma^2$ had a rate $\alpha = 3$ and shape $\beta = 2$. Uninformative priors had little, if any, contribution to the posterior distribution of the parameters due to the large number of observations.

We computed statistical inferences based on the posterior distribution of $\Theta$. Comparison between two factor levels or the combinations of levels followed $P(\theta_\alpha > \theta_\beta \mid X)$, which computes as the proportion of Markov chains whose sample from $\theta_\alpha$ is greater than the sample from $\theta_\beta$. Comparison among all levels of a given factor had the following risk function computed in each Markov chain: 1 when the level represented the largest effect

and 0 otherwise, such that we were able to compute the posterior probability of each level to provide the highest predictive ability. The level of choice was, therefore, the one that minimized the expected risk *a posteriori.*

## 3.3 Results

### 3.3.1 Environmental factors

The environmental factors represented by field variation as the microenvironment (Fig16) and year as the macroenvironment (Fig17) affect the signaling of genetic effects. Different traits may not necessarily display the same sensitivity to environmental changes (Cappa and Cantet 2008). Consequences of the environmental noise are captured by changes in heritability, which is inversely proportional to the variance due to environmental factors. It is possible to notice the influence of microenvironmental variation in Figure 16 by comparing the results using no phenotypic adjustment (NO) and those of two different methods (LK and GK). Likewise, one notices the macroenvironmental variation in Figure 17.

Unreplicated field designs, like the one used in this study, often cause deflated heritability and predictive ability (Endelman et al. 2014), although unreplicated trials are still preferred in GWP and mapping studies (Jannink et al. 2010). According to the complexity of the population structure, genome-based heritability estimates can be lower than pedigree-based estimates (Dekkers 2012) and nevertheless, results indicate that even low heritable traits still provide reasonable accuracy. Muir (2007) pointed out that traits with low heritability display more potential to be exploited and, therefore, low heritability estimates do not always affect accuracy. On the other hand, the accuracy, as defined by Lehermeier et al. (2013), can be interpreted as the amount of genetic gains that genomic selection can exploit

and, consequently, less heritable traits may provide high accuracy by displaying a predictive ability comparable to more heritable traits.

Figures 16 and 17 also illustrate how the phenotypic adjustment of field trials and year affected heritability, PA, and accuracy in different soybean traits. The analysis of phenotypic adjustment indicates that the posterior probability of GK to provide the model with the highest PA across traits is 100%. In marginal terms, the posterior mean of PA increased by 18.89% (from 0.350 to 0.416) and mean heritability increased by 35.78% (from 0.341 to 0.603) when adjusting phenotypes with GK compared to no adjustments.

Yield was the trait most sensitive to phenotypic adjustments; the gains in PA reached 32.45% using GK (from 0.411 to 0.544) and heritability increased 42.03% (from 0.452 to 0.642). Maturity displayed the highest increase in heritability when adjusting phenotypes with GK (71.39%, from 0.346 to 0.593) and height was the only trait that adjustments using linear kernel provided the highest predictive ability. This last result indicates that adjustment of phenotypes can be sensitive to interaction between environment and genetics (de los Campos et al. 2013) and that not all quantitative traits are equally responsive to phenotypic adjustment. All three yield components displayed the highest PA and heritability under the GK approach. The control of environmental noise for yield components is critical. Previous studies summarized by Board and Kahlon (2011) show that these traits are very sensitive to various environmental stimuli.

The two environments, 2013 and 2014, showed similar results (Fig17) which indicates a stable level of genetic control across seasons, with the exception of height which showed a remarkable drop in PA and heritability in 2014. It is possible that when the field variation was calculated for height in 2014 using GK, the model was incapable of distinguishing

between field and genetic variation causing overfitting. According to Cappa and Cantet (2008), not fitting field and genomic covariance matrices jointly may harm the quality of breeding values. Nevertheless, our results indicate that model overfitting may occur even when employing multiple kernels. Pods, nodes, and pods per node (P/N) had a slight increase in PA from 2013 to 2014, averaging 5.07%. We attribute this increase in PA, heritability, and accuracy on yield components to the number of plants used to represent each field plot, which doubled from 2013 to 2014. Interestingly, doubling the observations per plot provided very little increase in the prediction parameters.

Most strategies that account for field variation include the use of checks, neighbor plots, and a well-planned experimental design (Heffner et al. 2009; Endelman et al. 2014; Lado et al. 2014). Our findings support that the use of sophisticated techniques based on multiple kernels effectively controls field variation. Likewise, improvements of phenotypic measures are not trivial to genome-wide prediction and field variation must not be ignored. Most traits showed similar values of heritability and PA across years, indicating some level of stability in the genetic control and predictability of traits under evaluation.

### 3.3.2 Training population size

Training population is the most impactful factor on PA (Fig20) and can define the success of GWP. Two main properties of the training set are critical to GWP, its relatedness to the validation set (Habier et al. 2007), and the population size (Nakaya and Isobe 2012). Good training sets must be somehow related to the germplasm under evaluation to capture the population structure and have a population size sufficient for an accurate estimation of allelic effects (Jannink et al. 2010). As with any real dataset, SoyNAM is a finite population with constrained structure. Thus the model calibration becomes more accurate as the

training set increases. The remaining question regards what population size is required for a sufficiently good prediction.

Quantitative traits are mostly controlled by alleles of small and medium effect, so that larger training sets will increase the signal-to-noise ratio (Muir 2007) and provide better learning properties (Okser et al. 2014), which potentially results in more accurate allelic effect estimates by minimizing the so-called Beavis effect at the whole-genome level (Xu 2003). Increasing the size of the training set can increase predictive ability as much as 80% (from 0.252 to 0.454) and accuracy 82% (from 0.404 to 0.734) across traits. The posterior mean of PA also increases across traits by 27.29% as the training set increases from 250 to 500 individuals, 18.49% from 500 to 1000 individuals, 12% from 1000 to 2000 individuals, 4.86% from 2000 to 3000 individuals, and 2.03% from 3000 to 4000 individuals. Our results indicate that a population containing between 1000 and 2000 would be an effective training set as gains become relatively marginal for populations greater than 2000 individuals (Fig18).

Besides the quantity of the training population, the quality also determines the success of prediction and long-term breeding (Bastiaansen et al. 2012). The quality of the training set with regard to its genetic variability depends on the effective population size ($N_e$), which is always smaller than the total number of genotypes. Soybean and other self-pollinated species often suffer from reduced effective population size because of their reproductive nature (Cowling et al. 2015; Hamblin et al. 2011). This issue is not as severe in this study due the variability of the NAM populations (Yu et al. 2008), but it must be considered in breeding populations restricted to the narrow bases of elite germplasm.

It is also necessary to point out that a minimal, and perhaps optimal, population size is required when the ultimate goal is to perform selection of unphenotyped material to save resources (Heffner 2009). On the other hand, when the training set is part of a breeding population that is being phenotyped and selected over generations, increasing the population size is always beneficial from the breeding perspective to increase genetic gains (Bastiaansen et al. 2012; Hamblin et al. 2011; Muir 2007).

Population size may be also critical for the choice of prediction model (Bastiaansen et al. 2012). For example, combined models (kernel+additive) keep improving the PA as the population size increases while other methods are more robust with smaller population sizes. The posterior probability of each model to provide the highest PA changed as the training population size increased (Table 6). In the next section, we discuss the how prediction models respond to various scenarios.

### 3.3.3 Prediction Model

The posterior distribution of PA among different models is shown in Figure 19 ranging from 0.376 to 0.384 and thus, it was possible to obtain an increase of 2.16% in PA by selecting an appropriate model. This is equivalent to increasing the population size from 3000 to 4000 individuals. Also, it must be kept in mind that we base these inferences on marginal terms, pooling all other variables, and increases in PA due to prediction model can be higher for specific combinations of trait, population size, marker density, and environment.

Combined methods, have a 92.8% posterior probability of displaying higher predictive ability than additive methods alone, while additive methods have a 100% probability of being better than kernel methods alone. Interestingly, BLUP and GBLUP model are two

model considered to be equivalent (Habier et al. 2007), but they did not appear to have the same learning properties. Here, the posterior probability of BLUP to overperform GBLUP was 83.8%, while the probability of the combination of both to overperform BLUP is 55.86%. According to Gianola et al. (2014), some weak learning properties of the GBLUP model can be overcome by resampling techniques such as bootstrapping aggregation.

The decision to include kernels (pedigree or genomic) in the prediction model depends on many factors, such as the marker density (Heffner et al. 2009), availability and complexity of pedigree data, and genetic architecture of the trait (de los Campos et al. 2013). Our results indicate that there is no advantage in utilizing RKHS or GBLUP alone (Tables 5 and 6) in contrast to reports from simulated studies of wheat and maize (González-Camacho et al. 2012; Pérez-Rodríguez et al. 2012; Howard et al. 2014). Bernardo (2014) suggests that kernel-based methods can be very effective when major QTLs exist, are known *a priori* and are included as fixed effect in the prediction model.

We observed the importance of kernel methods when combined with additive methods to boost the predictive ability. Results indicate that RHKS is a better complimentary method than GBLUP. Even though both kernel methods are somewhat additive, RKHS accounts for different levels or relationships among individuals through the use of non-linear kernels (de los Campos et al. 2010; González-Camacho et al. 2012). In addition, Habier et al. (2007) pointed out that markers can inform the relationship matrix and contribute to kernel methods regardless of actual linkage to any QTL, while this would harm any additive model unable to perform efficient variable selection.

Regarding the distribution of marker effects for the SoyNAM dataset, the posterior probability of *t* models (BayesA and BayesB) to display higher PA than Gaussian models

(BLUP and BayesC) was 99.4%, and there was a 77.3% probability of Gaussian models having higher PA than double-exponential models (BLASSO). These findings show non-regularized additive methods (BayesA and BayesB) displaying better predictive abilities than regularized additive methods (BLUP, BLASSO and BayesC). Nevertheless, the probability that BLASSO provides higher PA increases with the population size (Table 6) and it is possible that this could overcome the PA of BayesA and BayesB when larger training sets are available, in agreement with Wimmer et al. (2013) and Okser et al. (2014).

Efficient prediction models often rely on consistent variable selection (Okser et al. 2014) and the implementation of variable selection appears to be feasible strategy in soybeans. The posterior probability of variable selection models (BayesB and BayesC) to increase predictive ability was 86.4% when compared to the 'all-included' counterpart models (BayesA and BLUP). Cultivated soybeans have a small genome, large LD blocks, and restricted diversity (Hyten el al 2006 2007; Chung et al. 2014). These are genomic properties that would contribute to the efficient selection of markers linked to QTL, along with the genetic properties of the nested association panels in which all individuals are related. Our results are based on various scenarios and traits, with a diverging number of makers ($p$) and observations ($n$) that range from $n << p$ to $n >> p$. However, this result regarding variable selection may not extend to other plants. Wimmer et al. (2013) analyzed datasets of rice, wheat, and *Arabidopsis thaliana,* concluding that variable selection does improve plant breeding, even in the presence of major effect genes. To Wimmer et al. (2013), robust regularization and variable selection require a large population size, while our results indicate that the better performance of variable selection holds across traits (Table 5) and populations sizes (Table 6).

Pérez-Rodríguez et al. (2012) compared the performance of parametric and non-parametric genomic prediction models on two wheat traits across several environments, showing that each combination of trait and environment had an ideal model. Analyzing the data across environments, they found that the parametric model BayesB better predicted one trait while the non-parametric model RKHS better predicted another trait. Similarly, Zhong et al. (2009) also noticed that GBLUP and BayesB each predicted different barley traits better than the other. Our results show that the combination of both is beneficial. The posterior probability of the RKHS+BayesB model to show the highest PA across traits was 57.8%. Kärkkäinen and Sillanpää (2012) also report this synergy for a model of BayesB with the polygenic term expressed by kernels, perhaps because kernels account for structure while BayesB is relatively insensitive to the genetic relationship between the training and validation sets (Habier et al. 2007). But these properties are not always advantageous. In the absence of admixture, Guo et al. (2012) found that BLUP would be more suitable than BayesB for within-family selection in NAM populations. The higher performance of the combined RKHS+BayesB in our experiment can be viewed from a simple perspective of ensemble learning: While RKHS accounts for different degrees of relationship among individuals or "hidden heritability" (Okser et al. 2014), BayesB captures QTLs in disequilibrium with markers in an additive fashion.

Despite the marginal contribution of the choice of prediction model to the overall predictive ability (2.16%), the genetic architecture of a trait determines which prediction model works best (Bastiaansen et al. 2012; de los Campos et al. 2013). Conversely evaluating different prediction models provides insight into the true genetic architecture (Dekkers 2012). Nonetheless, from the perspective of model flexibility, we see that the combination of a

non-parametric term with an additive variable selection method can account for different genetic interactions. Kernel methods enable the model to capture some level of epistasis (González-Camacho et al. 2012; Howard et al. 2014) with no assumptions about additive inheritance (de los Campos et al. 2009; Gianola 2009) and BayesB allows markers to have large and/or null effect (Habier et al. 2011). However, BayesB is not always effective to learn the genetic architecture of traits (Gianola 2013; Wimmer et al. 2013). It will depend on the proportion of markers and observations. Dekkers (2012) suggested that, with sufficient data, BayesB could be used to fine map causative mutations and, in spite of having very influential priors and restricted Bayesian learning (Gianola 2009; Lehermeier et al. 2013), our results show BayesB to be an outstanding method with respect to its prediction ability in a variety of scenarios, particularly when combined with kernels.

### 3.3.4 Genotyping Density

The posterior probability that all SNPs would provide the best PA was 85.5%. However, the increase in the *posteriori* mean of PA associated with the number of SNPs was 0.64% (from 0.378 to 0.38). Higher genotyping density often does not provide a substantial increase in predictive properties (VanRaden et al. 2011) and subsets of the genotypic data sometimes overperform the entire dataset (Erbe et al. 2012). Xu (2013b) observed that artificial bins that compress genotypic information into fewer parameters could provide more accurate results than natural bins.

For the SoyNAM population, 1020 markers would be enough to provide a consistent prediction while higher density genotyping would provide only marginal gains in PA. This result is likely due to soybean's genomic properties, such as the existence of large

disequilibrium blocks presented in Figure 13 also reported by Hyten et al. (2007), and uneven distribution of SNPs in clusters reported in Li et al. (2014).

SoyNAM is a group of biparental populations without intercross generations comprising elite and non-elite germplasm; nevertheless the importance of larger SNP panels grows when the population structure is unknown, the number of generations increases and the LD between QTL and marker decays (Bastiaansen et al. 2012; Daetwyler et al. 2013). In agreement with VanRaden et al. (2011), our results support the preference for increased population size over higher genotyping density.

## 3.4 Conclusions

By comparing the gains associated with each factor across traits, we showed that training population size and phenotypic adjustments were the most relevant parameters with regard to predictive ability in the SoyNAM dataset (Fig20). Thus the resources that best optimize prediction are related to the size and quality of the training set. However, it is important to recall that the other factors in study also contribute to GWP and should be optimized as well.

The application of spatial statistics substantially improved the quality of our phenotypic data (Cappa and Cantet 2008), as reflected in higher estimates of heritability, predictive ability, and accuracy (Lado et al. 2014). Increasing the training population size also enhanced these prediction parameters. Nevertheless, the rate of improvement decreased rapidly above 2000 individuals, suggesting that an optimal population size exists and it was between 1000 and 2000 for the dataset in study. Yet, for the best allocation of resources, it is better to prioritize a larger population over high density genotyping (VanRaden et al. 2011; Bastiaansen et al. 2012).

We showed that comparison among prediction models plays two important roles: (1) it helps us to learn, understand, and quantify the genetic architecture (Bastiaansen et al. 2012; Dekkers 2012; Gianola 2013) and (2) it is necessary to decide which model or combination of models would provide the most reliable breeding values (Habier et al. 2007; Lehermeier et al. 2013). The best overall model choice was RKHS+BayesB, which combines methods to provide a more robust prediction (Kärkkäinen and Sillanpää 2012), but further research on variable selection, kernels, and regularization is necessary (Piepho 2009; de los Campos et al. 2010; Wimmer et al. 2013).

Reinforcing previous studies, we recognized the value of next-generation populations to exploit new genomic frontiers through machine learning procedures not limited to genome-wide associations (Guo et al. 2012; Gianola 2013; Okser et al. 2014; Poland 2015). NGPs have interesting statistical properties valuable for *in silico* experiments (Yu et al. 2008; Hamblin et al. 2011). Results from machine learning experiments based on real data, such as the present study, are fundamental for resource allocation, planning, and decision making in breeding programs that aim to optimize genetic gains (Muir 2007; Lehermeier et al. 2013; Endelman et al. 2014; Lado et al. 2014).

References

Asmussen S, Nerman O, Olsson M (1996) Fitting phase-type distributions via the EM algorithm. Scandinavian Journal of Statistics 23:419-441.

Bastiaansen JW, Coster A, Calus MP, van Arendonk JA, Bovenhuis H (2012) Long-term response to genomic selection: effects of estimation method and reference population structure for different genetic architectures. Genet. Sel. Evol. 44(3): 10-1186.

Bernardo R (2010) Breeding for quantitative traits in plants. 2nd ed. Stemma Press, Woodbury, MN.

Bernardo R (2013) Genome-wide markers for controlling background variation in association mapping. The Plant Genome 6(1).

Bernardo R (2014) Genome-wide selection when major genes are known. Crop Science 54(1): 68-75.

Board JE, Kahlon CS (2011) Soybean Yield Formation: What controls it and how it can be improved. Soybean Physiology and Biochemistry. INTECH Open Access Publisher.

Cappa EP, Cantet RJ (2008) Direct and competition additive effects in tree breeding: Bayesian estimation from an individual tree mixed model. Silvae Genetica 57(2): 45-55.

Cheng H, Qu L, Garrick DJ, Fernando RL (2015) A fast and efficient Gibbs sampler for BayesB in whole-genome analyses. Genet. Sel. Evol. 47(1): 1-7.

Chung WH, Jeong N, Kim J, Lee WK, Lee YG, et al. (2014) Population structure and domestication revealed by high-depth resequencing of Korean cultivated and wild soybean genomes. DNA research 21(2): 153-167.

Cowling WA, Stefanova KT, Beeck CP, Nelson MN, Hargreaves BL, et al. (2015) Using the Animal Model to Accelerate Response to Selection in a Self-Pollinating Crop. G3: Genes| Genomes| Genetics: g3-115.

Daetwyler HD, Calus MP, Pong-Wong R, de los Campos G, Hickey JM (2013) Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. Genetics 193(2): 347-365.

Dekkers JC (2012) Application of genomics tools to animal breeding. Current genomics 13(3): 207.

de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MP (2013) Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193(2): 327-345.

de los Campos G, Gianola D, Rosa GJ, Weigel KA, Crossa J (2010) Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genetics Research 92(04): 295-308.

de los Campos G, Naya H, Gianola D, Crossa J, Legarra A, et al. (2009) Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182(1): 375-385.

Endelman JB, Atlin GN, Beyene Y, Semagn K, Zhang X, et al. (2014) Optimal design of preliminary yield trials with genome-wide markers. Crop Science 54(1): 48-59.

Erbe M, Hayes BJ, Matukumalli LK, Goswami S, Bowman PJ, et al. (2012) Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. Journal of dairy science 95(7): 4114-4129.

Gianola D, Weigel KA, Krämer N, Stella A, Schön CC (2014) Enhancing genome-enabled prediction by bagging genomic BLUP.

Gianola D (2013) Priors in whole-genome regression: the Bayesian alphabet returns. Genetics 194(3): 573-596.

Gianola D, de los Campos G, Hill WG, Manfredi E, Fernando RL (2009) Additive genetic variability and the Bayesian alphabet. Genetics 183(1): 347-363.

González-Camacho JM, de los Campos G, Pérez P, Gianola D, Cairns JE, et al. (2012) Genome-enabled prediction of genetic values using radial basis function neural networks. Theoretical and Applied Genetics 125(4): 759-771.

Guo Z, Tucker DM, Lu J, Kishore V, Gay G (2012) Evaluation of genome-wide selection efficiency in maize nested association mapping populations. Theoretical and Applied Genetics 124(2): 261-275.

Habier D, Fernando RL, Kizilkaya K, Garrick DJ (2011) Extension of the Bayesian alphabet for genomic selection. BMC bioinformatics 12(1): 186.

Habier D, Fernando RL, Dekkers JC (2007) The impact of genetic relationship information on genome-assisted breeding values. Genetics 177(4): 2389-2397.

Hamblin MT, Buckler ES, Jannink JL (2011) Population genetics of genomics-based crop improvement methods. Trends in Genetics 27(3): 98-106.

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.

Heffner EL, Sorrells ME, Jannink JL (2009) Genomic selection for crop improvement. Crop Science 49(1): 1-12.

Henderson CR (1986) Estimation of variances in animal model and reduced animal model for single traits and single records. Journal of Dairy Science 69(5):1394-1402.

Heslot N, Rutkoski J, Poland J, Jannink JL, Sorrells ME (2013) Impact of marker ascertainment bias on genomic selection accuracy and estimates of genetic diversity. PLoS One 8(9): e74612.

Howard R, Carriquiry AL, Beavis WD (2014) Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. G3: Genes| Genomes| Genetics 4(6): 1027-1046.

Hyten DL, Choi IY, Song Q, Shoemaker RC, Nelson RL, et al. (2007) Highly variable patterns of linkage disequilibrium in multiple soybean populations. Genetics 175(4): 1937-1944.

Hyten DL, Song Q, Zhu Y, Choi IY, Nelson RL, et al. (2006) Impacts of genetic bottlenecks on soybean genome diversity. Proceedings of the National Academy of Sciences 103(45): 16666-16671.

Jannink JL, Lorenz AJ, Iwata H (2010) Genomic selection in plant breeding: from theory to practice. Briefings in functional genomics 9(2): 166-177.

Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, et al. (2008) Efficient control of population structure in model organism association mapping. Genetics 178(3): 1709-1723.

Kärkkäinen HP, Sillanpää MJ (2012) Back to basics for Bayesian model building in genomic selection. Genetics 191(3): 969-987.

Kuo L, Mallick B (1998) Variable selection for regression models. Sankhya: The Indian Journal of Statistics Series B: 65-81.

Lado B, Matus I, Rodríguez A, Inostroza L, Poland J, et al. (2013) Increased genomic prediction accuracy in wheat breeding through spatial adjustment of field trial data. G3: Genes| Genomes| Genetics 3(12): 2105-2114.

Legarra A, Misztal I (2008) Technical note: Computing strategies in genome-wide selection. Journal of dairy science 91(1): 360-366.

Lehermeier C, Wimmer V, Albrecht T, Auinger HJ, Gianola D, et al. (2013) Sensitivity to prior specification in Bayesian genome-based prediction models. Statistical applications in genetics and molecular biology 12(3): 375-391.

Li YH, Liu YL, Reif JC, Liu ZX, Liu B, et al. (2014) Biparental resequencing coupled with SNP genotyping of a segregating population offers insights into the landscape of recombination and fixed genomic regions in elite soybean. G3: Genes| Genomes| Genetics 4(4): 553-560.

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157(4): 1819-1829.

Mohammadi M, Tiede T, Smith KP (2015) PopVar: A genome-wide procedure for predicting genetic variance and correlated response in bi-parental breeding populations. Crop Sci. (5): 55(5): 2068-2077.

Morrell PL, Buckler ES, Ross-Ibarra J (2012) Crop genomics: advances and applications. Nature Reviews Genetics 13(2): 85-96.

Muir WM (2007) Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. Journal of Animal Breeding and Genetics 124(6): 342-355.

Nakaya A, Isobe SN (2012) Will genomic selection be a practical method for plant breeding? Annals of botany: mcs109.

O'Hara RB, Sillanpää MJ (2009) A review of Bayesian variable selection methods: what, how and which. Bayesian analysis 4(1): 85-117.

Okser S, Pahikkala T, Airola A, Salakoski T, Ripatti S, et al. (2014) Regularized machine learning in the genetic prediction of complex traits. PLOS Genetics 12(11): e1004754.

Park T, Casella G (2008) The Bayesian Lasso. Journal of the American Statistical Association 103(482): 681-686.

Pérez P, de los Campos G (2014) Genome-wide regression & prediction with the BGLR statistical package. Genetics: genetics-114.

Pérez-Rodríguez P, Gianola D, González-Camacho LM, Crossa J, Manès Y, et al. (2012) Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. G3: Genes| Genomes| Genetics 2(12): 1595-1605.

Piepho HP (2009) Ridge regression and extensions for genome-wide selection in maize. Crop Science 49(4): 1165-1176.

Poland J (2015) Breeding-assisted genomics. Current opinion in plant biology 24: 119-124.

Sorensen D, Gianola D (2002) Likelihood, Bayesian, and MCMC methods in quantitative genetics. Springer Science & Business Media.

Stekhoven DJ, Bühlmann P (2012) MissForest - nonparametric missing value imputation for mixed-type data. Bioinformatics 28(1): 112-118.

Strandén I, Christensen OF (2011) Allele coding in genomic evaluation. Genet. Sel. Evol. 43(1).

Tibshirani R (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B (Methodological): 267-288.

VanRaden PM, O'Connell JR, Wiggans GR, Weigel KA (2011) Genomic evaluations with many more genotypes. Genet. Sel. Evol. 43(10): 10-1186.

Wimmer V, Lehermeier C, Albrecht T, Auinger HJ, Wang Y, et al. (2013) Genome-wide prediction of traits with different genetic architecture through efficient variable selection. Genetics 195(2): 573-587.

Xavier A, Xu S, Muir WM, Rainey KM (2015) NAM: Association Studies in Multiple Populations. Bioinformatics. btv448.

Xu S (2013a) Mapping quantitative trait loci by controlling polygenic background effects. Genetics 195(4): 1209-1222.

Xu S (2013b) Genetic mapping and genomic selection using recombination breakpoint data. Genetics 195(3): 1103-1115.

Xu S (2003) Theoretical basis of the Beavis effect. Genetics 165(4): 2259-2268.

Yu J, Holland JB, McMullen MD, Buckler ES (2008) Genetic design and statistical power of nested association mapping in maize. Genetics 178(1): 539-551.

Zas R (2006) Iterative kriging for removing spatial autocorrelation in analysis of forest genetic trials. Tree genetics & genomes 2(4): 177-185.

Zhong S, Dekkers JC, Fernando RL, Jannink JL (2009) Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study. Genetics 182(1): 355-364.

**Table 1.** Number of times that each pairwise combination of traits was observed together. Main diagonal represent the total number of observation for each trait (bold).

| Trait† | Yld | Flo | Mat | Rep | Hgt | Ldg | Acc | Rcc | LSh | Node | Pod | P/N | SW | Int |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Yld | **15643** | 9992 | 15638 | 9990 | 15640 | 11082 | 11061 | 11059 | 11096 | 11331 | 11331 | 11331 | 10058 | 11331 |
| Flo | - | **10005** | 10000 | 10003 | 10002 | 9993 | 9970 | 9968 | 10005 | 10005 | 10005 | 10005 | 4426 | 10005 |
| Mat | - | - | **19012** | 10001 | 19009 | 14451 | 11070 | 11068 | 11105 | 14700 | 14700 | 14700 | 10063 | 14700 |
| Rep | - | - | - | **10004** | 10001 | 9994 | 9969 | 9967 | 10004 | 10004 | 10004 | 10004 | 4424 | 10004 |
| Hgt | - | - | - | - | **19014** | 14449 | 11072 | 11070 | 11107 | 14702 | 14702 | 14702 | 10065 | 14702 |
| Ldg | - | - | - | - | - | **14452** | 11060 | 11058 | 11095 | 14452 | 14452 | 14452 | 5518 | 14452 |
| Acc | - | - | - | - | - | - | **11075** | 11073 | 11075 | 11075 | 11075 | 11075 | 5529 | 11075 |
| Rcc | - | - | - | - | - | - | - | **11073** | 11073 | 11073 | 11073 | 11073 | 5528 | 11073 |
| LSh | - | - | - | - | - | - | - | - | **11110** | 11110 | 11110 | 11110 | 5529 | 11110 |
| Node | - | - | - | - | - | - | - | - | - | **14705** | 14705 | 14705 | 5762 | 14705 |
| Pod | - | - | - | - | - | - | - | - | - | - | **14705** | 14705 | 5762 | 14705 |
| P/N | - | - | - | - | - | - | - | - | - | - | - | **14705** | 5762 | 14705 |
| SW | - | - | - | - | - | - | - | - | - | - | - | - | **10065** | 5762 |
| Int | - | - | - | - | - | - | - | - | - | - | - | - | - | **14705** |

† 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; Pod, pods in the main stem; P/N, pods per node; SW, 100-seed weight; Int, internode length.

**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 | - | -0.059*** | 0.312*** | 0.313*** | 0.134*** | 0.013 | 0.311*** | 0.134*** | 0.12*** | 0.198*** | 0.177*** | 0.063*** | 0.072*** | -0.056*** |
| Flo | -0.048*** | - | 0.21*** | -0.533*** | 0.194*** | 0.07*** | -0.008 | 0.02* | -0.063*** | -0.001 | -0.038*** | -0.057*** | 0.046** | 0.128*** |
| Mat | 0.299*** | 0.302*** | - | 0.591*** | 0.418*** | 0.166*** | 0.179*** | 0.048*** | -0.003 | 0.205*** | 0.101*** | -0.072*** | 0.032** | 0.145*** |
| Rep | 0.405*** | -0.235*** | 0.747*** | - | 0.207*** | 0.095*** | 0.132*** | 0.034*** | 0.02* | 0.216*** | 0.133*** | -0.01 | -0.022 | 0.006 |
| Hgt | 0.123*** | 0.231*** | 0.399*** | 0.296*** | - | 0.352*** | 0.442*** | 0.249*** | -0.047*** | 0.337*** | 0.276*** | -0.012 | -0.024** | 0.417*** |
| Ldg | 0.03** | 0.051*** | 0.182*** | 0.133*** | 0.379*** | - | 0.302*** | 0.214*** | -0.134*** | 0.19*** | 0.193*** | 0.07*** | 0.002 | 0.114*** |
| Acc | 0.298*** | -0.005 | 0.175*** | 0.172*** | 0.426*** | 0.307*** | - | 0.533*** | -0.133*** | 0.303*** | 0.238*** | 0.06*** | 0.087*** | 0.094*** |
| Rcc | 0.121*** | 0.045*** | 0.059*** | 0.025** | 0.241*** | 0.225*** | 0.502*** | - | -0.049*** | 0.205*** | 0.139*** | 0.019* | 0.026* | 0.05*** |
| LSh | 0.151*** | -0.03** | -0.001 | -0.003 | -0.045*** | -0.141*** | -0.105*** | -0.042*** | - | -0.032*** | -0.029** | -0.014 | -0.028* | -0.028** |
| Node | 0.195*** | -0.011 | 0.229*** | 0.299*** | 0.387*** | 0.243*** | 0.29*** | 0.197*** | -0.049*** | - | 0.508*** | -0.033*** | -0.009 | -0.266*** |
| Pod | 0.177*** | -0.052*** | 0.104*** | 0.177*** | 0.276*** | 0.21*** | 0.232*** | 0.145*** | -0.013 | 0.597*** | - | 0.778*** | -0.056*** | -0.203*** |
| P/N | 0.07*** | -0.063*** | -0.059*** | -0.018* | 0.016* | 0.084*** | 0.06*** | 0.029** | 0.023** | 0.031*** | 0.768*** | - | -0.064*** | -0.042*** |
| SW | 0.075*** | 0.08*** | 0.054*** | 0.01 | -0.013 | 0.007 | 0.103*** | 0.023* | -0.05*** | -0.017 | -0.057*** | -0.059*** | - | 0.06*** |
| Int | -0.052*** | 0.159*** | 0.148*** | 0.007 | 0.429*** | 0.119*** | 0.095*** | 0.049*** | -0.027** | -0.315*** | -0.205*** | -0.04*** | 0.063*** | - |

* Significant at the 0.05 probability level.
** Significant at the 0.01 probability level.
*** Significant at the 0.001 probability level.
† 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; Pod, pods in the main stem; P/N, pods per node; SW, 100-seed weight; Int, internode length.

**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 | **0.632** | -0.291*** | 0.692*** | 0.798*** | 0.553*** | 0.503*** | 0.726*** | 0.53*** | 0.081*** | 0.58*** | 0.435*** | 0.153*** | 0.089*** | 0.08*** |
| Flo | -0.051*** | **0.7** | 0.205*** | -0.536*** | 0.385*** | 0.322*** | 0.038*** | -0.07*** | -0.326*** | -0.065*** | -0.122*** | -0.211*** | 0.127*** | 0.42*** |
| Mat | 0.344*** | 0.131*** | **0.822** | 0.714*** | 0.863*** | 0.71*** | 0.613*** | 0.29*** | -0.142*** | 0.465*** | 0.187*** | -0.17*** | 0.207*** | 0.487*** |
| Rep | 0.248*** | -0.64*** | 0.535*** | **0.716** | 0.454*** | 0.376*** | 0.496*** | 0.3*** | 0.102*** | 0.476*** | 0.256*** | -0.011 | 0.107*** | 0.084*** |
| Hgt | 0.269*** | 0.13*** | 0.469*** | 0.163*** | **0.881** | 0.891*** | 0.765*** | 0.522*** | -0.288*** | 0.394*** | 0.216*** | -0.07*** | 0.206*** | 0.666*** |
| Ldg | 0.088*** | 0.026** | 0.225*** | 0.108*** | 0.351*** | **0.658** | 0.831*** | 0.649*** | -0.424*** | 0.573*** | 0.454*** | 0.152*** | 0.068*** | 0.407*** |
| Acc | 0.355*** | -0.056*** | 0.177*** | 0.125*** | 0.459*** | 0.285*** | **0.729** | 0.896*** | -0.359*** | 0.536*** | 0.429*** | 0.165*** | 0.207*** | 0.312*** |
| Rcc | 0.197*** | -0.011 | 0.06*** | 0.046*** | 0.209*** | 0.143*** | 0.497*** | **0.604** | -0.319*** | 0.381*** | 0.303*** | 0.117*** | 0.163*** | 0.195*** |
| LSh | 0.096*** | -0.046*** | -0.018* | 0.009 | -0.058*** | -0.153*** | -0.147*** | -0.032*** | **0.594** | -0.024** | -0.039*** | -0.035*** | -0.077*** | -0.265*** |
| Node | 0.219*** | -0.003 | 0.224*** | 0.15*** | 0.361*** | 0.224*** | 0.309*** | 0.187*** | -0.049*** | **0.823** | 0.831*** | 0.382*** | -0.066*** | -0.422*** |
| Pod | 0.197*** | -0.043*** | 0.099*** | 0.096*** | 0.2*** | 0.192*** | 0.228*** | 0.09*** | -0.028** | 0.625*** | **0.837** | 0.83*** | -0.23*** | -0.478*** |
| P/N | 0.077*** | -0.062*** | -0.062*** | 0.002 | -0.047*** | 0.057*** | 0.045*** | -0.024** | 0.003 | 0 | 0.775*** | **0.746** | -0.321*** | -0.406*** |
| SW | 0.021* | -0.099*** | -0.071*** | -0.065*** | -0.005 | -0.031* | 0.044** | -0.009 | 0.001 | -0.045*** | -0.039** | -0.015 | **0.394** | 0.266*** |
| Int | 0.047*** | 0.119*** | 0.226*** | 0.033*** | 0.573*** | 0.117*** | 0.141*** | 0.04*** | -0.012 | -0.541*** | -0.377*** | -0.062*** | 0.028* | **0.854** |

* Significant at the 0.05 probability level.
** Significant at the 0.01 probability level.
*** Significant at the 0.001 probability level.
† 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; Pod, pods in the main stem; P/N, pods per node; SW, 100-seed weight; Int, internode length.

**Table 4.** Correlation between two years of SoyNAM phenotypic data (2013 and 2014) and narrow-sense heritability before kriging (BK) and after kriging (AK) for six soybean traits: plant height (Height), days to flowering (Flower), days to maturity (Mature), number of nodes (Nodes) and pods (Pods) and average canopy closure (ACC).

| Parameter | | Height | Flower | Mature | Nodes | Pods | ACC |
|---|---|---|---|---|---|---|---|
| Correlation | BK | 0.67 | 0.2 | 0.54 | 0.22 | 0.21 | 0.21 |
| | AK | 0.71 | 0.2 | 0.55 | 0.26 | 0.25 | 0.35 |
| Heritability | BK | 0.9 | 0.49 | 0.82 | 0.74 | 0.82 | 0.74 |
| | AK | 0.94 | 0.56 | 0.88 | 0.76 | 0.88 | 0.79 |

**Table 5.** Posterior probability of each model to provide the highest predictive ability of each trait and across traits (overall).

| | Height | Maturity | Nodes | Pods | P/N | Yield | Overall |
|---|---|---|---|---|---|---|---|
| BayesA | 0.041 | 0.11 | 0.042 | 0.04 | 0.03 | 0.06 | 0.015 |
| BayesB | 0.074 | 0.211 | 0.066 | 0.08 | 0.04 | 0.1 | 0.063 |
| BayesC | 0.011 | 0 | 0.046 | 0.03 | 0.05 | 0.06 | 0.001 |
| BLASSO | 0.003 | 0 | 0.001 | 0 | 0 | 0.02 | 0 |
| BLUP | 0.002 | 0 | 0.009 | 0.01 | 0.02 | 0.02 | 0 |
| GBLUP | 0 | 0 | 0.001 | 0 | 0 | 0.01 | 0 |
| GLUP+BayesA | 0.035 | 0.069 | 0.058 | 0.04 | 0.07 | 0.07 | 0.019 |
| GLUP+BayesB | 0.065 | 0.202 | 0.085 | 0.09 | 0.1 | 0.1 | 0.07 |
| GLUP+BayesC | 0.01 | 0 | 0.045 | 0.02 | 0.06 | 0.04 | 0 |
| GLUP+BLASSO | 0.004 | 0 | 0.004 | 0.01 | 0.01 | 0.02 | 0 |
| GLUP+BLUP | 0.003 | 0 | 0.016 | 0.01 | 0.02 | 0.02 | 0 |
| RKHS | 0.002 | 0 | 0 | 0 | 0 | 0 | 0 |
| RKHS+BayesA | 0.245 | 0.133 | 0.155 | 0.23 | 0.19 | 0.14 | 0.244 |
| RKHS+BayesB | 0.347 | 0.276 | 0.284 | 0.31 | 0.27 | 0.21 | 0.578 |
| RKHS+BayesC | 0.091 | 0 | 0.132 | 0.07 | 0.1 | 0.08 | 0.01 |
| RKHS+BLASSO | 0.036 | 0 | 0.012 | 0.02 | 0.01 | 0.02 | 0 |
| RKHS+BLUP | 0.031 | 0 | 0.045 | 0.03 | 0.04 | 0.03 | 0 |

**Table 6.** Posterior probability of each model to provide the highest predictive ability for different sizes of training population set.

| | 250 | 500 | 1000 | 2000 | 3000 | 4000 |
|---|---|---|---|---|---|---|
| BayesA | 0.04 | 0.03 | 0.08 | 0.09 | 0.07 | 0.05 |
| BayesB | 0.19 | 0.15 | 0.15 | 0.1 | 0.06 | 0.03 |
| BayesC | 0.03 | 0.02 | 0.02 | 0.01 | 0.01 | 0.01 |
| BLASSO | 0 | 0 | 0 | 0.02 | 0.03 | 0.03 |
| BLUP | 0 | 0 | 0 | 0 | 0 | 0 |
| GBLUP | 0 | 0 | 0 | 0 | 0 | 0 |
| GLUP+BayesA | 0.06 | 0.05 | 0.07 | 0.09 | 0.07 | 0.04 |
| GLUP+BayesB | 0.16 | 0.16 | 0.16 | 0.11 | 0.08 | 0.04 |
| GLUP+BayesC | 0.03 | 0.02 | 0.01 | 0.01 | 0.01 | 0.01 |
| GLUP+BLASSO | 0 | 0 | 0 | 0.01 | 0.02 | 0.02 |
| GLUP+BLUP | 0 | 0.01 | 0 | 0 | 0 | 0 |
| RKHS | 0 | 0 | 0 | 0 | 0 | 0 |
| RKHS+BayesA | 0.12 | 0.15 | 0.14 | 0.21 | 0.25 | 0.27 |
| RKHS+BayesB | 0.29 | 0.31 | 0.31 | 0.31 | 0.3 | 0.31 |
| RKHS+BayesC | 0.08 | 0.08 | 0.04 | 0.03 | 0.04 | 0.06 |
| RKHS+BLASSO | 0 | 0 | 0 | 0.01 | 0.04 | 0.13 |
| RKHS+BLUP | 0.01 | 0.02 | 0.01 | 0 | 0.01 | 0.03 |

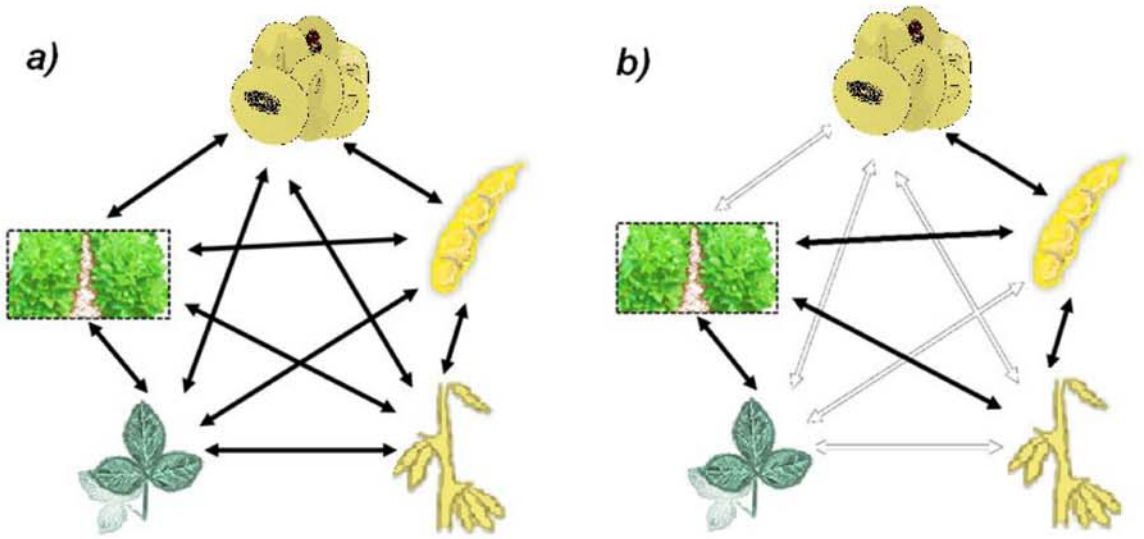


**Figure 1**. Representation of graphical modeling on soybean traits for causal structure studies. (**a**) Uninformative model and (**b**) sparse model. Black arrows represent direct associations and white arrows represent indirect associations.

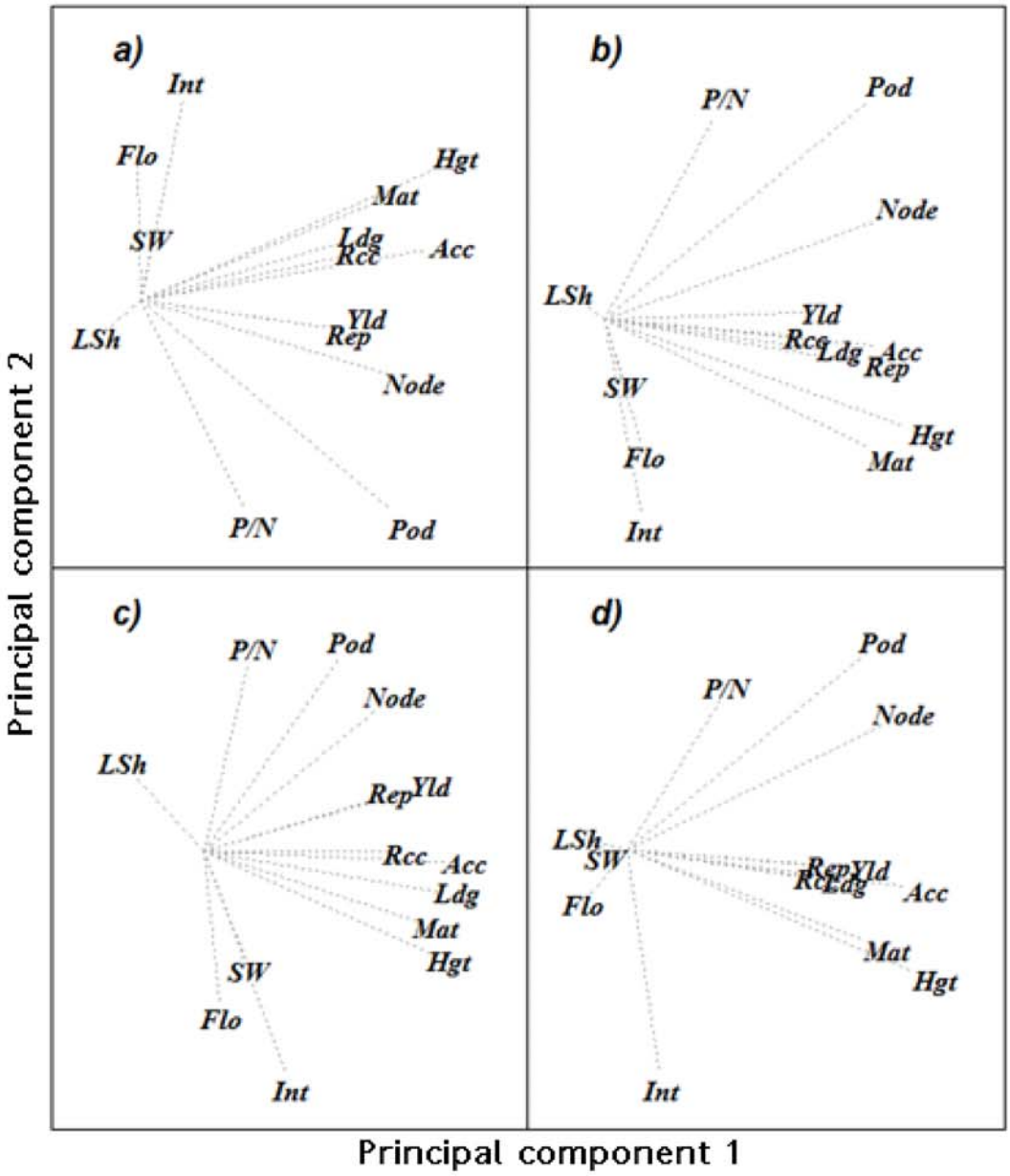


**Figure 2**. Principal component analysis of phenotypic Pearson (**a**), phenotypic Spearman (**b**), genetic (**c**) and environmental (**d**) correlations of soybean traits. Traits: grain yield (Yld), flowering (Flo), maturity (Mat), length of reproductive period (Rep), plant height (Hgt), lodging (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).

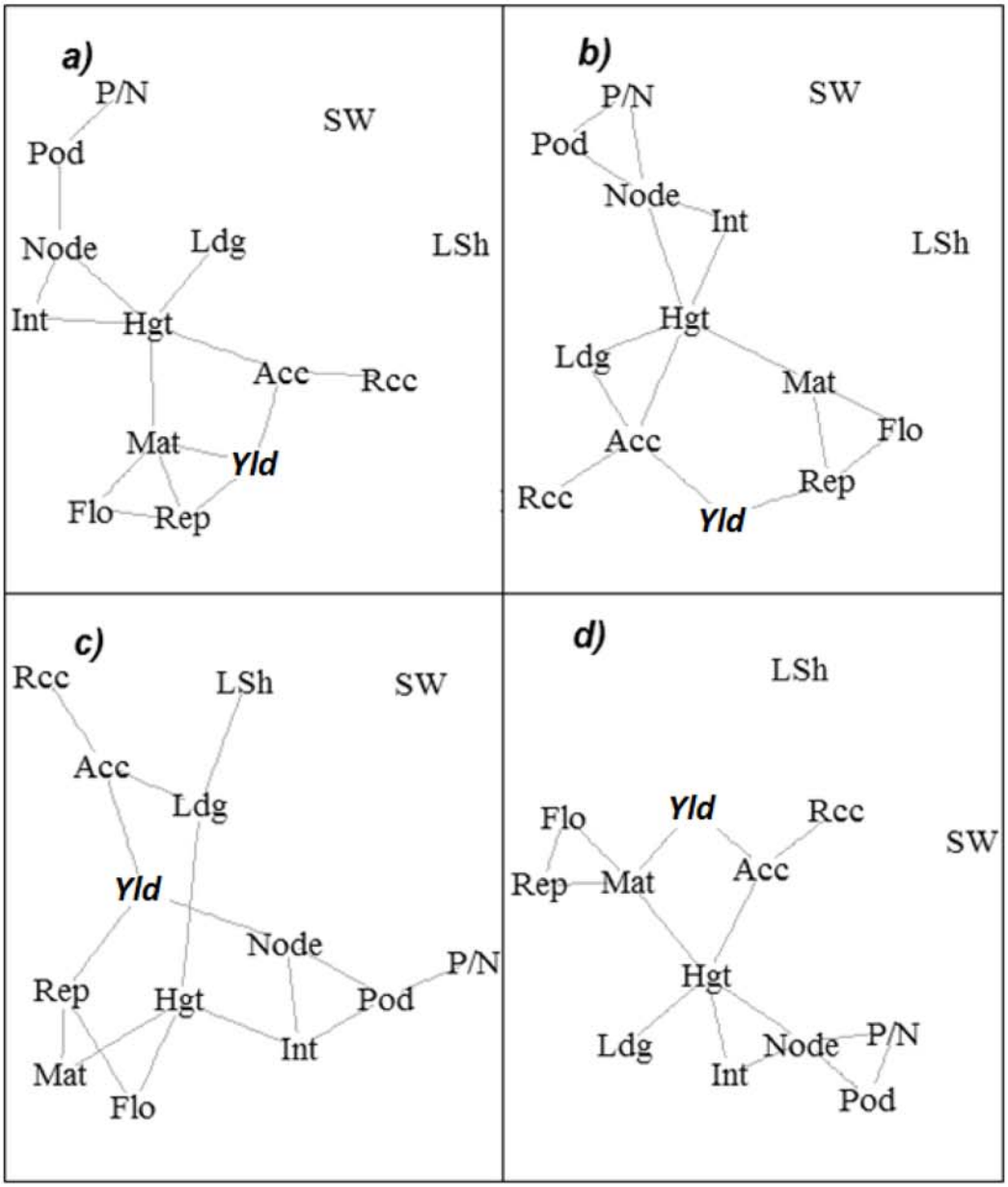


**Figure 3**. Graphical modeling of phenotypic Pearson (**a**), phenotypic Spearman (**b**), genetic (**c**) and environmental (**d**) correlations using the graphical LASSO of Meinshausen and Bühlmann (2006). Soybean traits include grain yield (Yld, bold), flowering (Flo), maturity (Mat), length of 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).

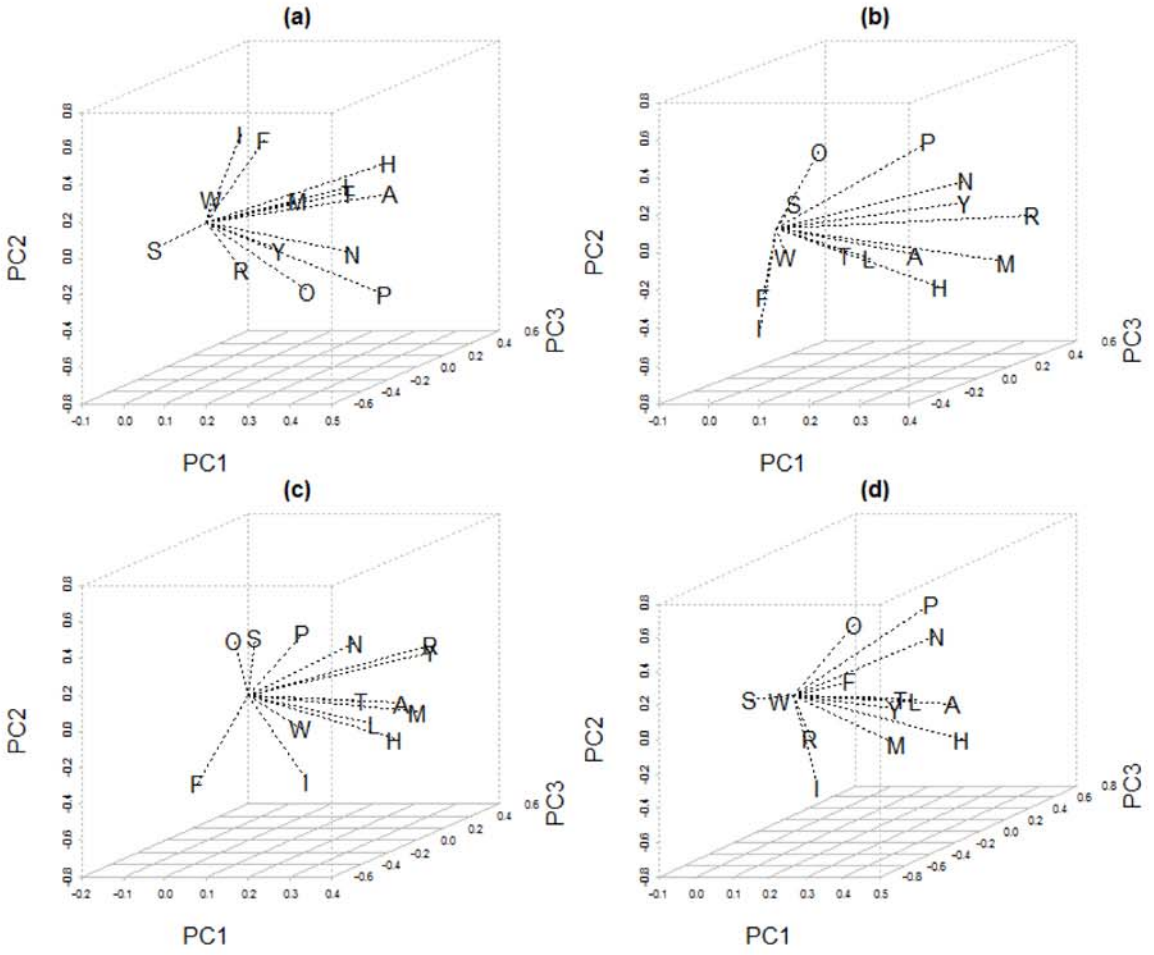


**Figure 4**. Principal component analysis of phenotypic Pearson (**a**), phenotypic Spearman (**b**), genetic (**c**) and environmental (**d**) correlations of soybean agronomic traits. Three principal components explain 47%, 48%, 76% and 49% of the total variance of a, b, c and d, respectively. Traits include grain yield (Y), flowering (F), maturity (M), reproductive period (R), plant height (H), lodging (L), average canopy closure (A), rate of canopy closure (T), leaflet shape (S), node number (N), pod number (P), pods per node (PN), seed weight (W) and internode length (I).

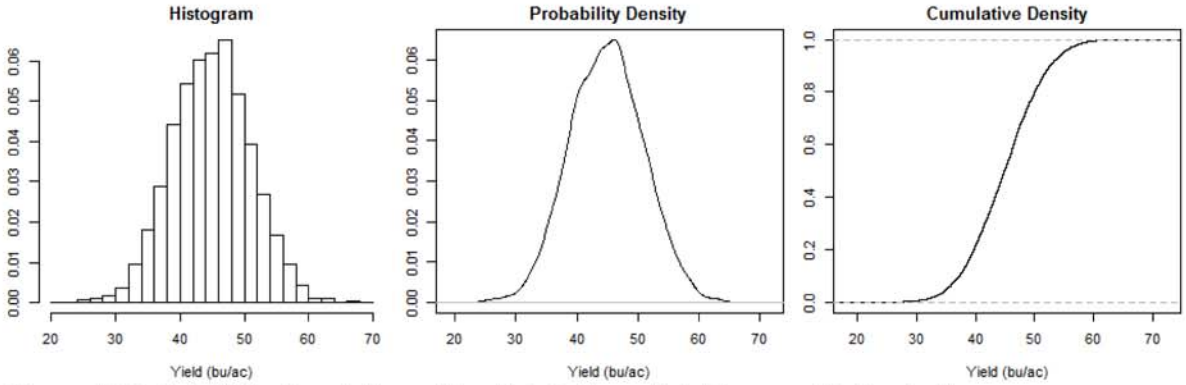


**Figure 5**. Probabilistic description of the distribution of yield, a quantitative trait.

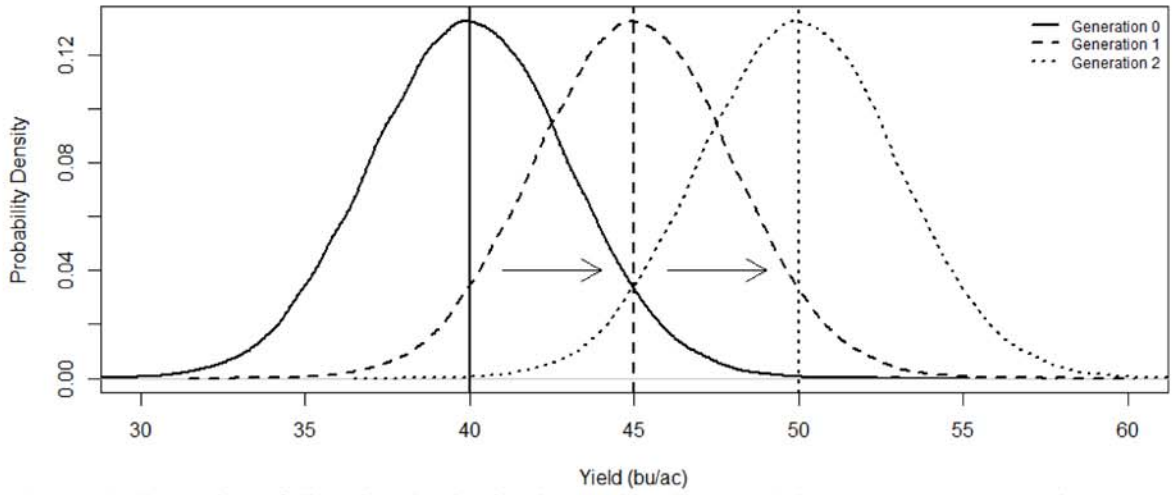


**Figure 6**. Illustration of directional selection increasing the population mean over generations.

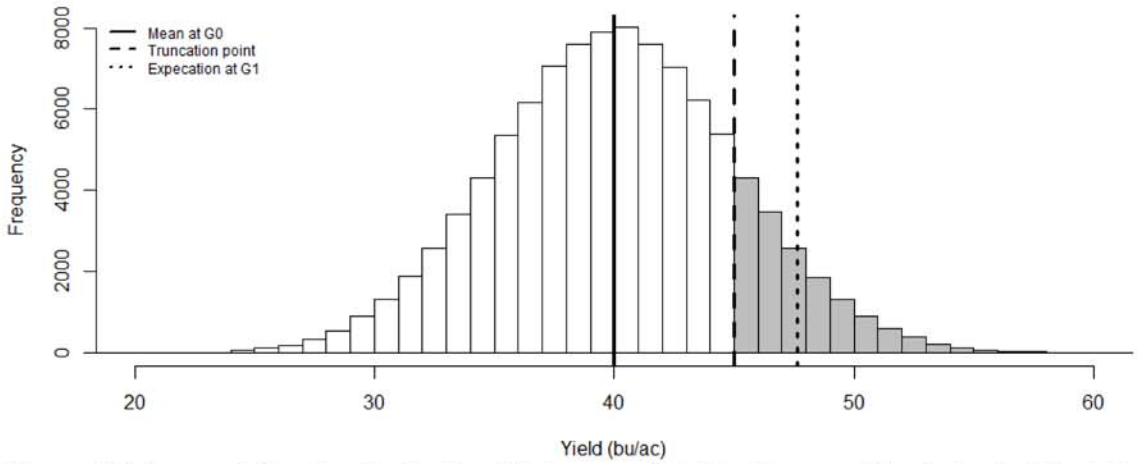


**Figure 7**. Scheme of directional selection: Histogram of yield with mean 40 and standard deviation 5, expected mean in the next generation 47.63 and truncation point 45 for selection intensity 1. Shaded bars represent the progenitors of the next generation.

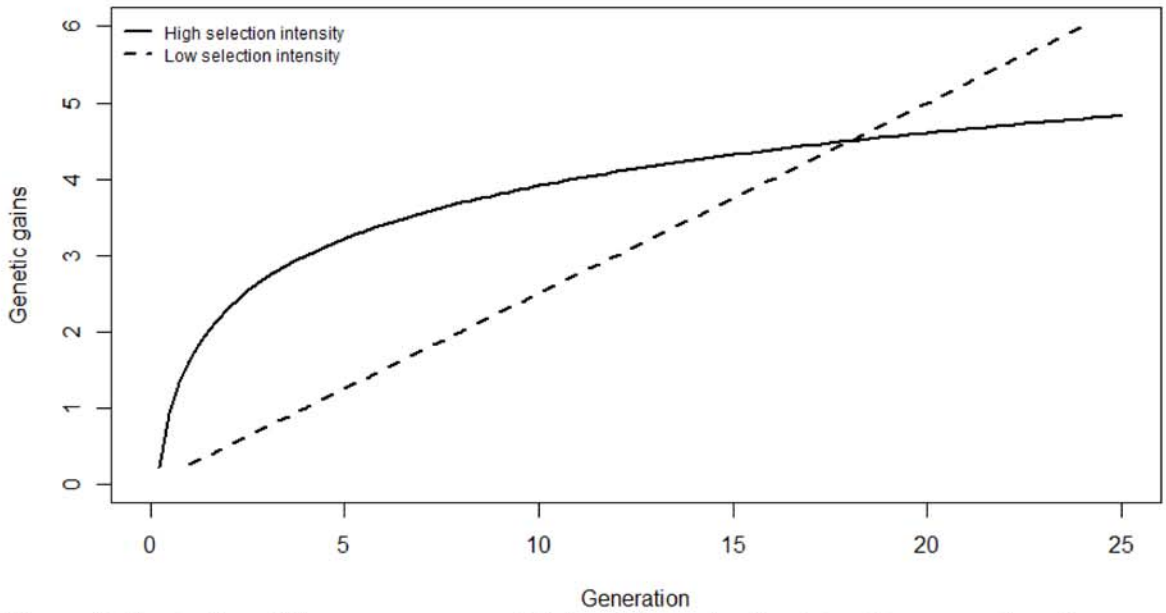


**Figure 8**. Illustration of the consequences of high and low selection intensity on genetic gains over generations of selection for a given quantitative trait.

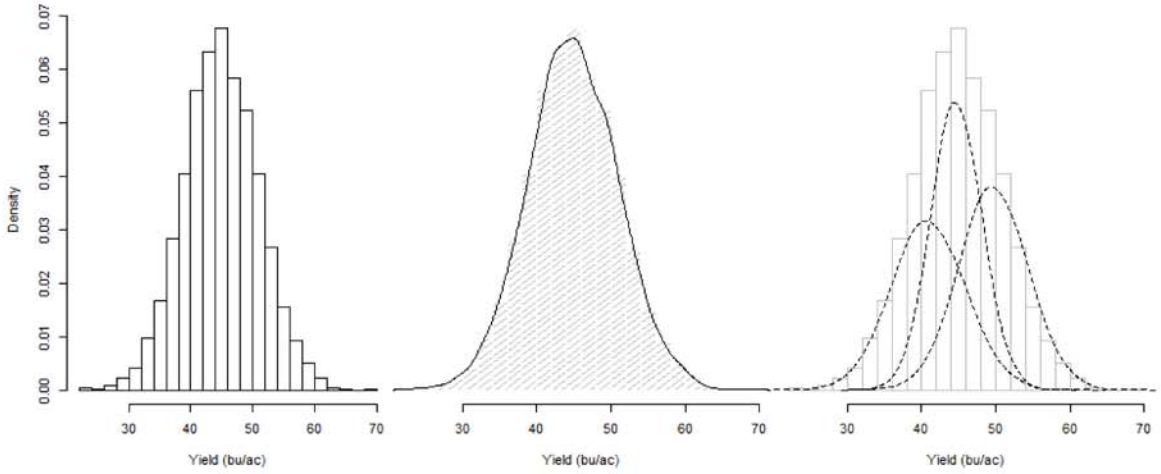


**Figure 9**. Histogram of yield (left) illustrating the distribution of a quantitative trait as single normal (center) and the distribution comprising a mixture of normal distributions (right) as proposed by Lander and Botstein (1989).

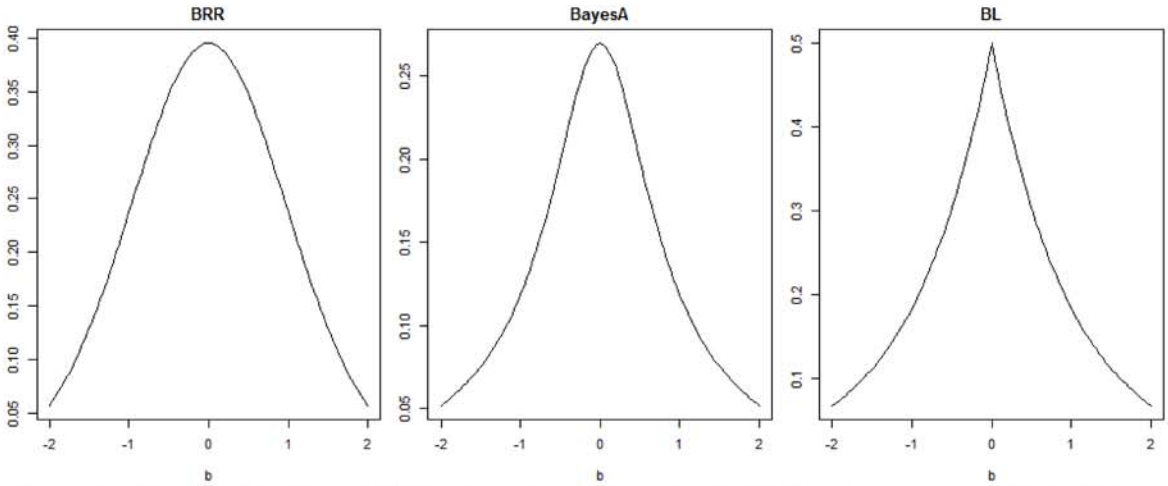


**Figure 10**. Density function of BRR, BayesA and Bayesian LASSO, where marker effects follow normal, t, and Laplace distributions, respectively.

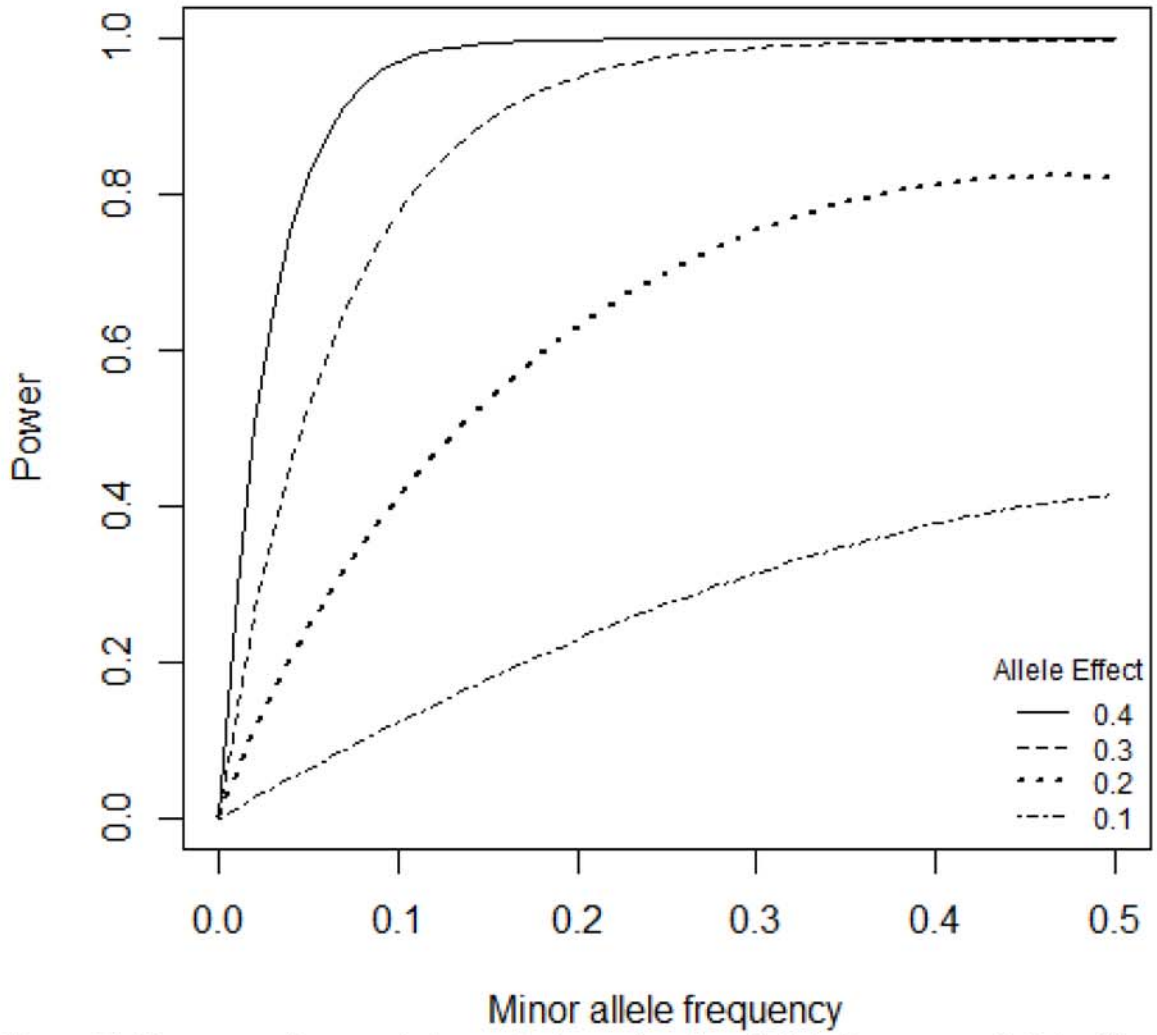


**Figure 11**. The power of an association analysis as a function of allele frequency and allele effect size, with a sample size of 1000. Adapted from Myles et al. (2009).

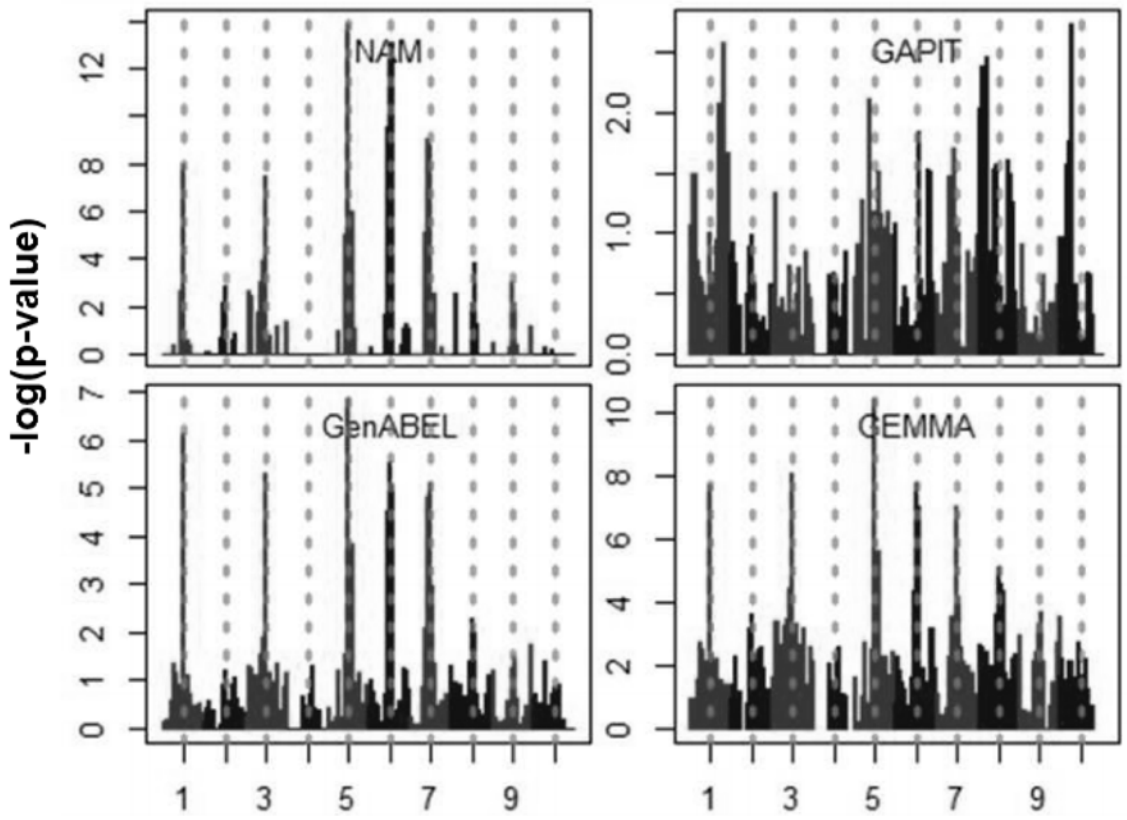


**Figure 12**. Manhattan plots of a simulated dataset with one QTL in the center of each chromosome using four different implementations of mixed models for GWAS.

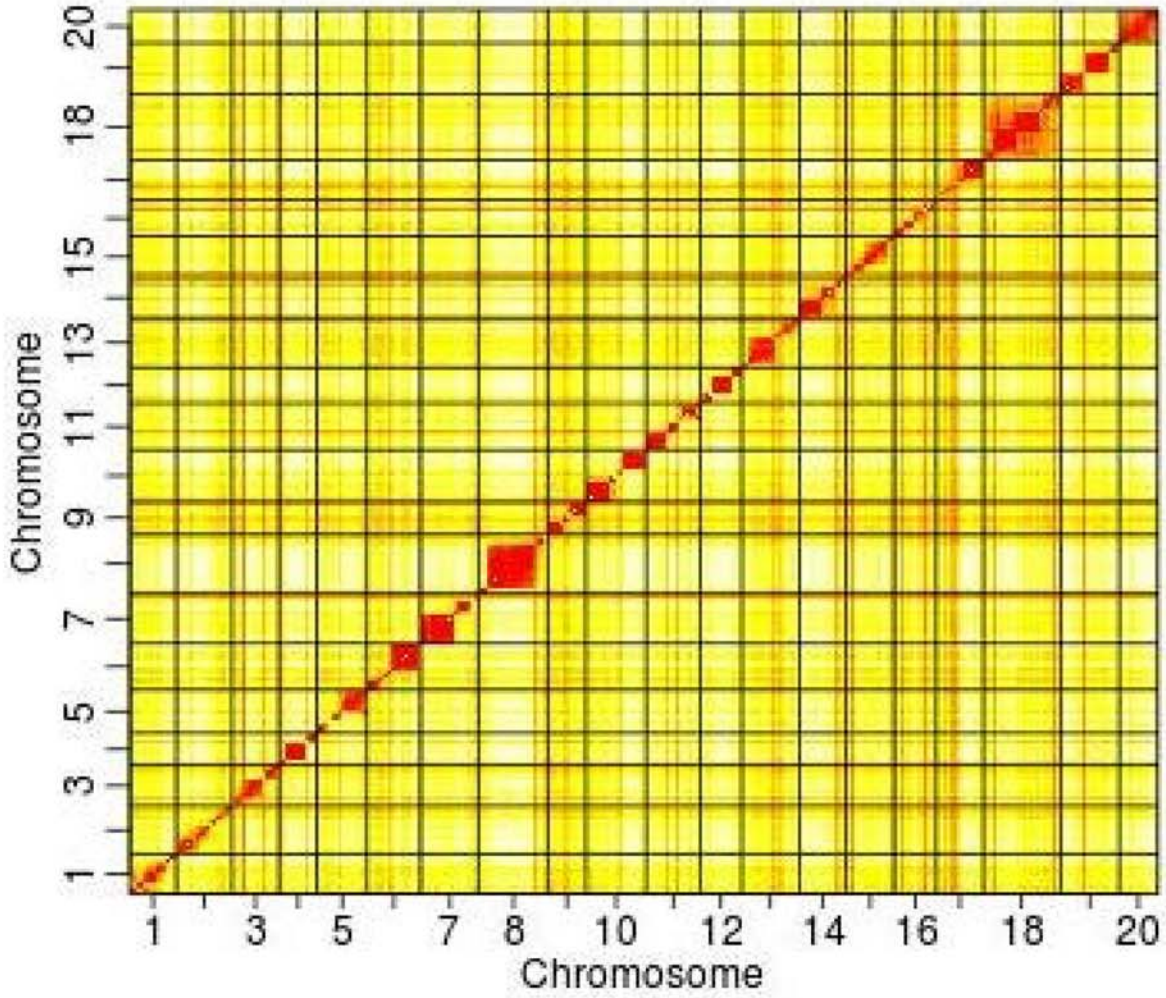


**Figure 13**. SNP panel pairwise linkage disequilibrium among the 4077 markers in terms of r^2 in the SoyNAM population.

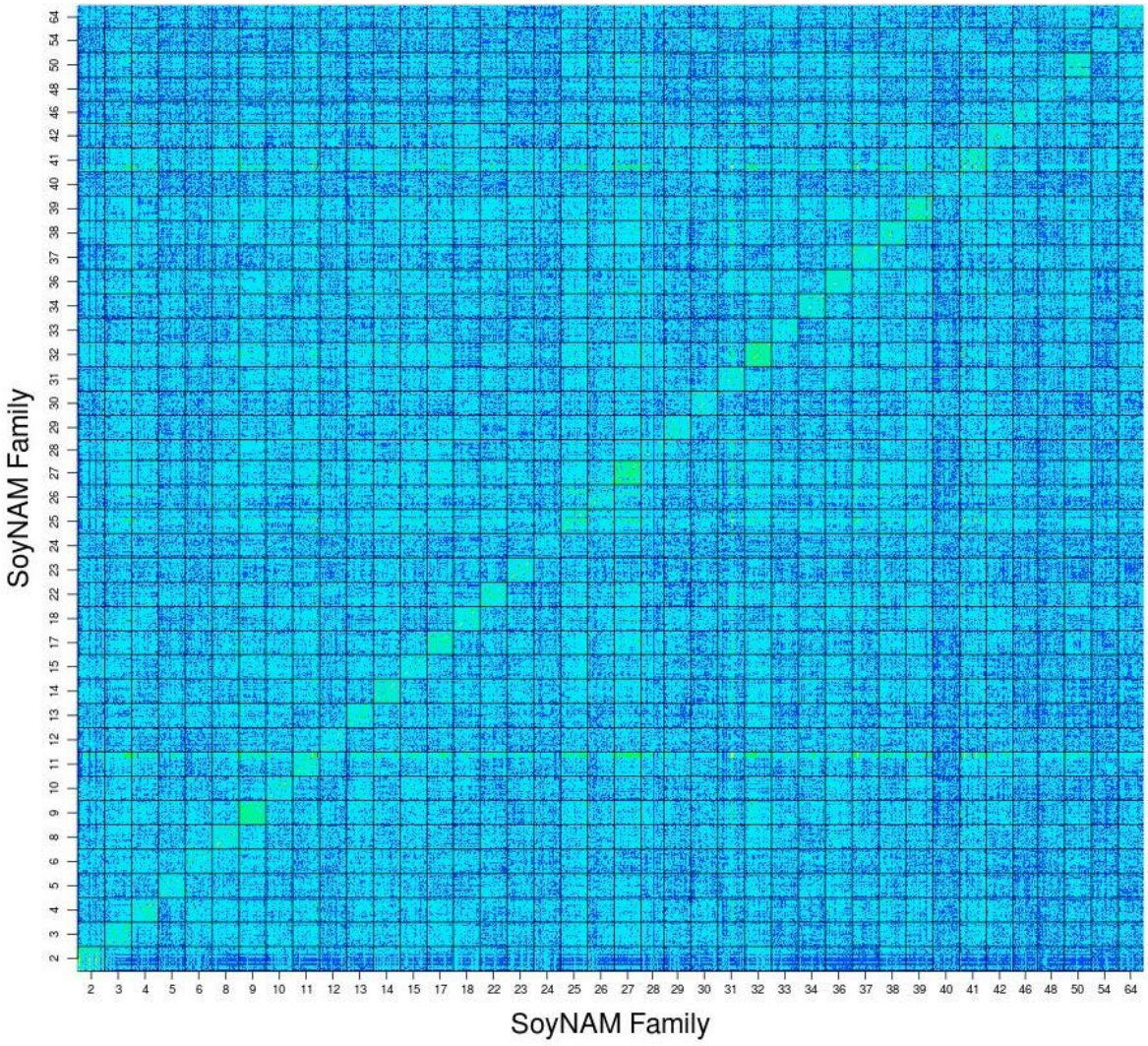


**Figure 14**. Heat map of the genomic relationship matrix of the 5555 individuals of the SoyNAM population with delimitations indicating family.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 0.05 | 0.05 | 0.05 | 0.06 | 0.06 | 0.06 | 0.05 | 0.05 | 0.05 |
| 0.09 | 0.1 | 0.11 | 0.12 | 0.12 | 0.12 | 0.11 | 0.1 | 0.09 |
| 0.16 | 0.19 | 0.21 | 0.23 | 0.24 | 0.23 | 0.21 | 0.19 | 0.16 |
| 0.26 | 0.33 | 0.4 | 0.46 | 0.49 | 0.46 | 0.4 | 0.33 | 0.26 |
| 0.32 | 0.42 | 0.56 | 0.75 | 1 | 0.75 | 0.56 | 0.42 | 0.32 |
| 0.26 | 0.33 | 0.4 | 0.46 | 0.49 | 0.46 | 0.4 | 0.33 | 0.26 |
| 0.16 | 0.19 | 0.21 | 0.23 | 0.24 | 0.23 | 0.21 | 0.19 | 0.16 |
| 0.09 | 0.1 | 0.11 | 0.12 | 0.12 | 0.12 | 0.11 | 0.1 | 0.09 |
| 0.05 | 0.05 | 0.05 | 0.06 | 0.06 | 0.06 | 0.05 | 0.05 | 0.05 |

**Figure 15**. Correlation between the central plot and neighbor plots using an Exponential kernel with bandwidth parameter $\rho$=3.5.

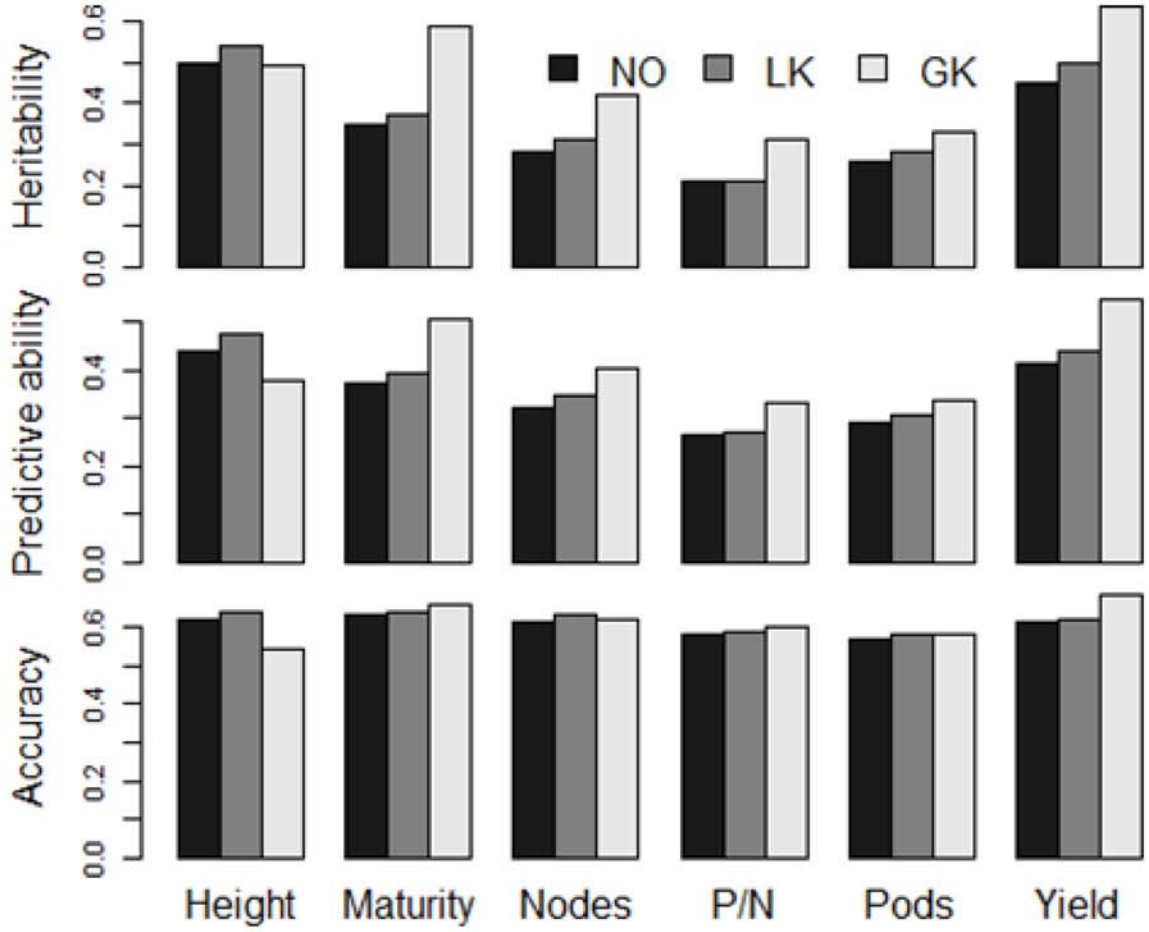


**Figure 16.** Heritability, predictive ability, and accuracy of six soybean traits with no phenotypic adjustment (NO), phenotypes corrected by field variation with a linear kernel (LK), and by three Gaussian kernels (GK).

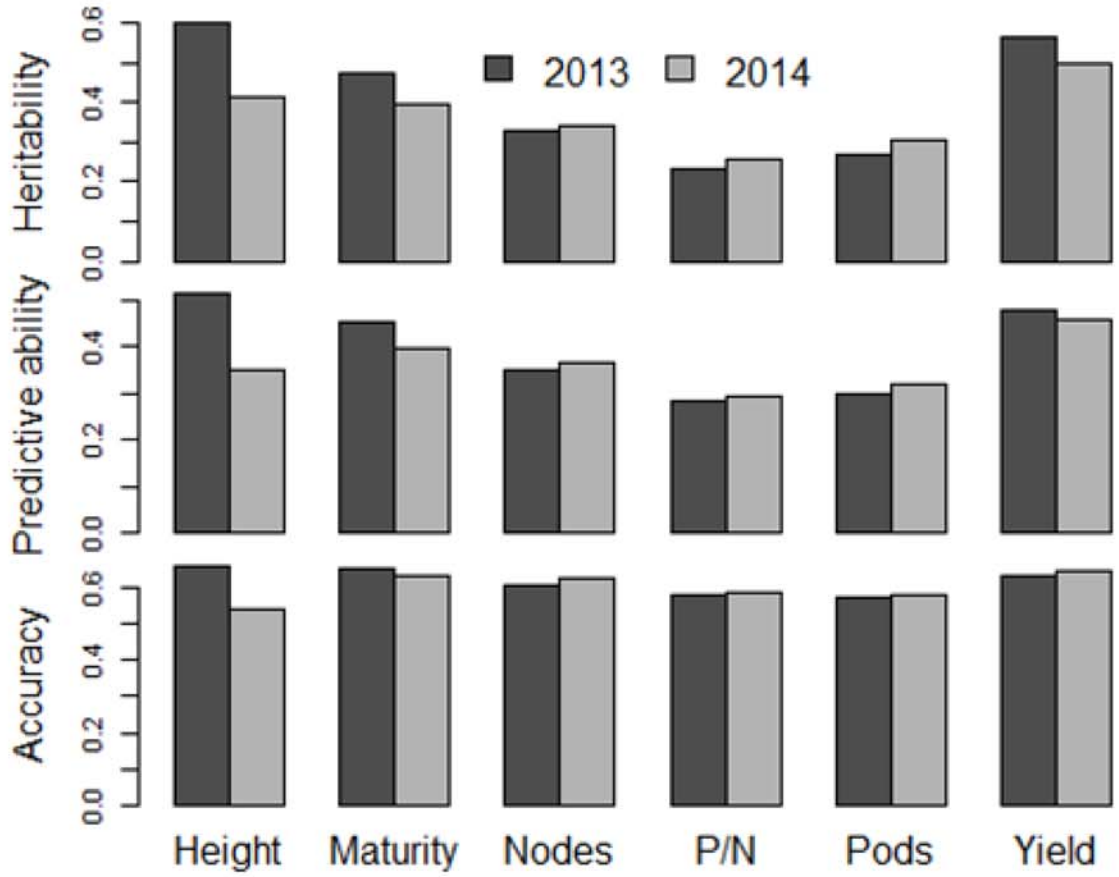


**Figure 17.** Heritability, predictive ability, and accuracy of six soybean traits in 2013 and 2014.

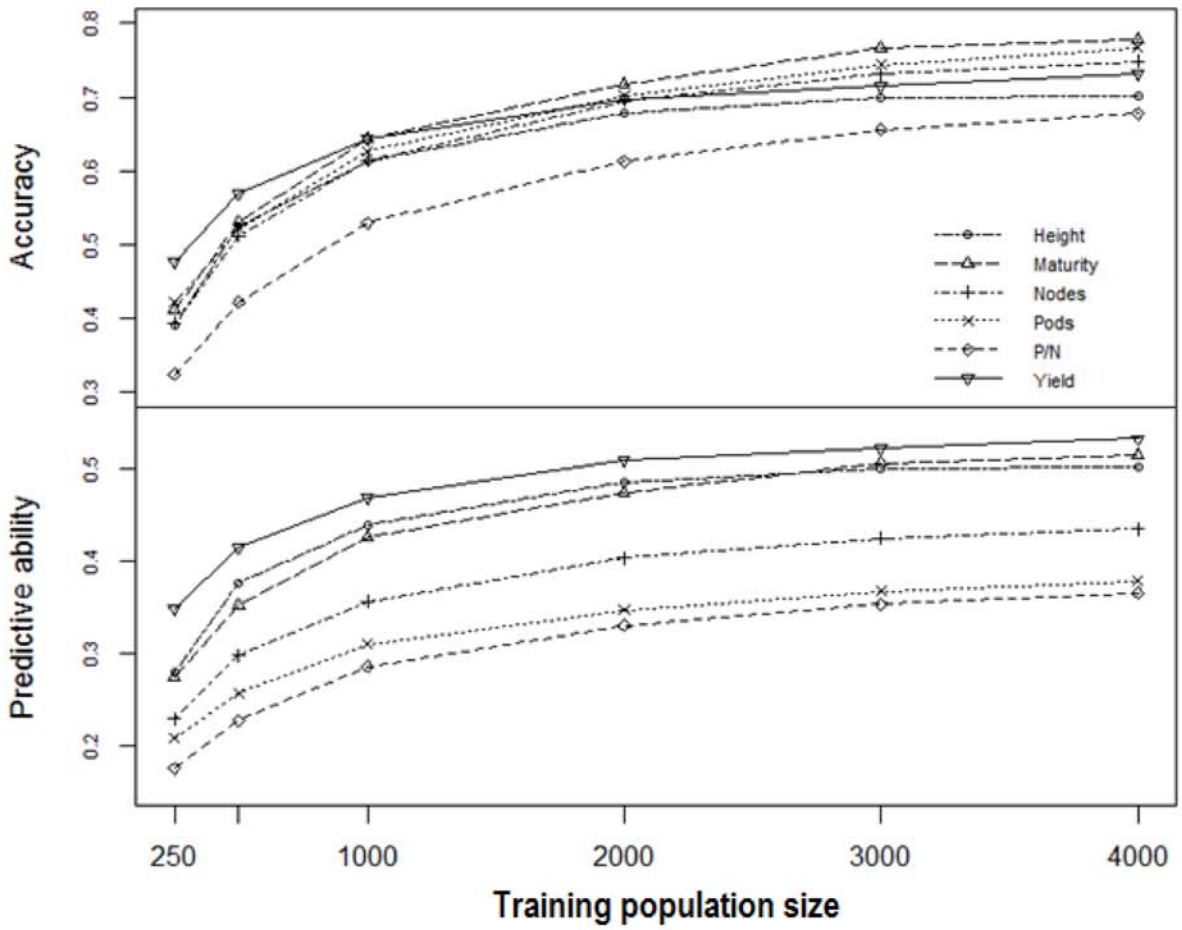


**Figure 18.** Learning curve: Effect of training population size in accuracy and predictive ability for different soybean traits.

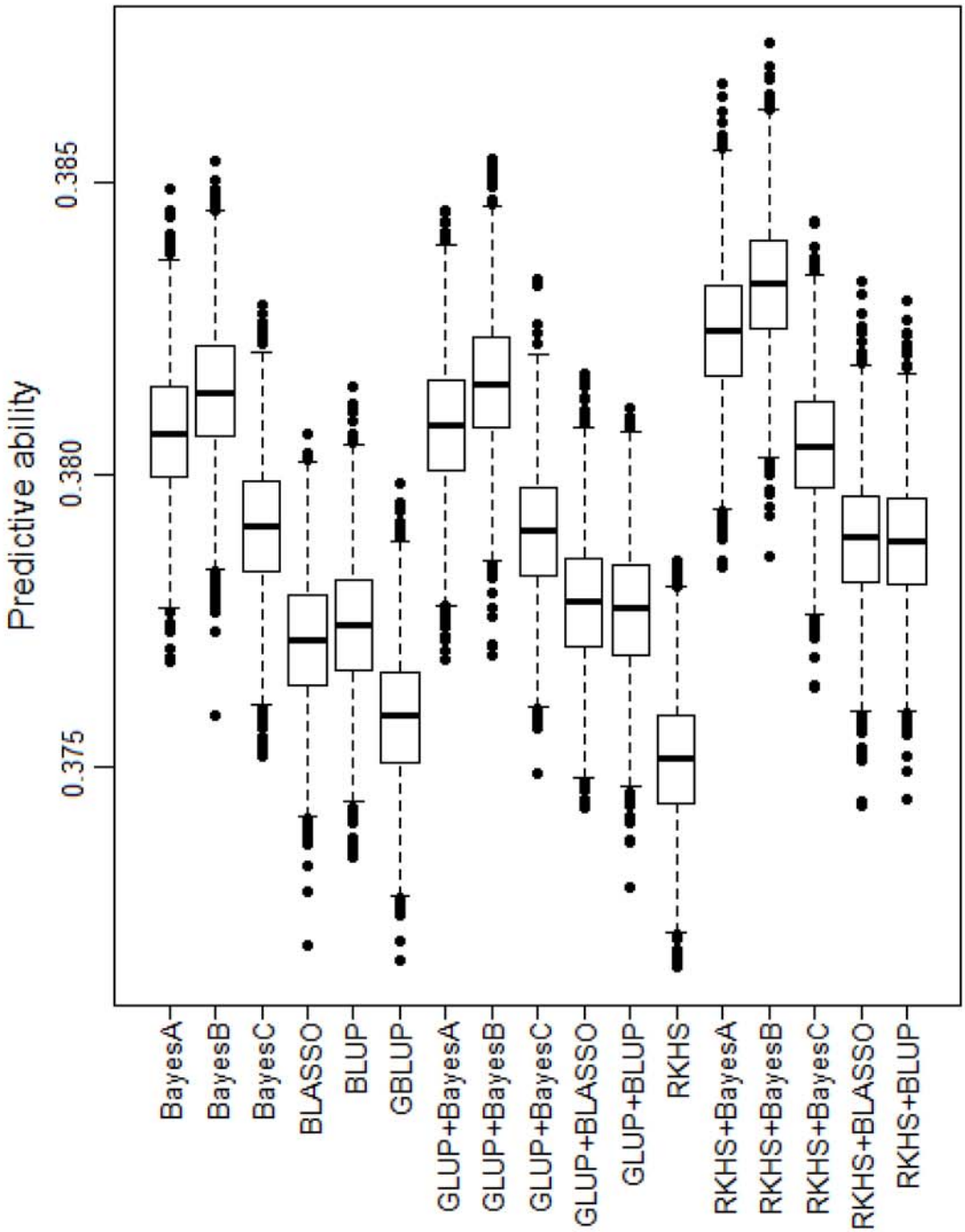


**Figure 19.** Posterior distribution of predictive ability for parametric models, non-parametric models and combination of each, across soybean traits.

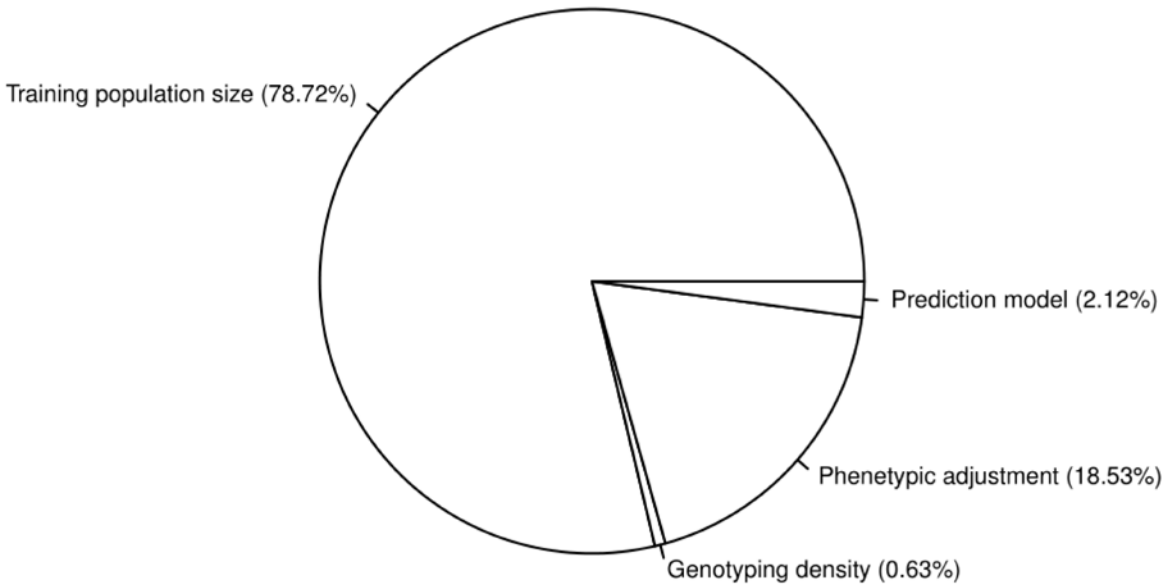


**Figure 20.** Percentage of attainable gains in predictive ability attributed to different parameters.

# VITA

Alencar began his professional formation attending a technical school for agriculture practices from 2004 to 2006. In 2011, He earned a B.Sc. degree in Agronomy at Federal University of St. Mary, granted with a research assistantship as an undergraduate student to work with soil physics (3 years) and potato breeding (2 years). His undergraduate research yielded three publications in peer-reviewed journals. After graduating, Alencar did an internship at KWS (7 months) and to attend classes at University of Minnesota (Spring 2012). He started graduate school at Purdue University with Dr. Katy Martin Rainey in soybean breeding in the spring semester of 2013, co-advised by Dr. William Muir. Alencar's research on yield components in the SoyNAM population called attention of Dow AgroSciences, which turned out funding his entire PhD research. With growing interest on statistical genetics, Alencar developed novel methods that would accommodate omic data of next-generation populations. Many statistical packages developed by Alencar were published on R and are being used worldwide for data analysis in the public and private sector. Alencar wrote eight manuscripts by the time of his graduation, having two published, three accepted, and three under submission. Alencar dedicates his research to improve the way breeding is done, trying to make it more data-driven and optimizing the use of novel technologies into breeding pipelines.