**NESTED ASSOCIATION MAPPING AND ACCURACY OF
PREDICTING GENOMIC BREEDING VALUE FOR AGRONOMIC AND SEED
COMPOSITION TRAITS IN THREE INTERSPECIFIC SOYBEAN
POPULATIONS**

A Dissertation
presented to
the Faculty of the Graduate School
at the University of Missouri-Columbia

In Partial Fulfillment
of the Requirements for the Degree
of Doctor of Philosophy in Plant Breeding, Genetics, and Genomics

---

by
EDUARDO BECHE
Dr. Andrew Scaboo, Dissertation Supervisor

FEBRUARY 2019

The undersigned, appointed by the Dean of the Graduate School,
have examined the Dissertation entitled

NESTED ASSOCIATION MAPPING AND ACCURACY OF PREDICTING
GENOMIC BREEDING VALUE FOR AGRONOMIC AND SEED COMPOSITION
TRAITS IN THREE INTERSPECIFIC SOYBEAN POPULATION

Presented by EDUARDO BECHE

A candidate for the degree of

Doctor of Philosophy in Plant Breeding, Genetics, and Genomics,
and hereby certify that, in their opinion, it is worthy of acceptance.

Dr. Andrew Scaboo, Chair

Dr. J. Grover Shannon, Committee member

Dr. Jason D. Gillman, Committee member

Dr. Jared E. Decker, Committee member

Dr. Timothy M. Beissinger, Committee member

ACKNOWLEDGEMENTS

First of all, I want to acknowledge the valuable guidance of my advisor Dr. Andrew M. Scaboo. His patient, supportive supervision, and financial support were essential to pursue my Ph.D. degree at Mizzou.

I would like to thank Dr. Grover Shannon for accepting me at Mizzou in the first place and Dr. Jason Gillman for his help during data analysis, discussion and writing. I also would like to thank all my committee members Drs. Grover Shannon, Jason Gillman, Jared Decker and Timothy Beissinger for the collaboration and dedication of time and effort to improve my research.

Thanks to the current and past member of the Northern Soybean Breeding group for assistance with field experiments and lab work. I know that I had a vast field experiment! None of this would be possible without everyone's help.

Also, I want to thank my funding agencies. The Brazilian agency Coordination for the Improvement of Higher Education Personnel (CAPES), for the scholarship during four years of my Ph.D. degree. For the United Soybean Board and Missouri Soybean Merchandising Council for the financial support of my research.

Finally, I must thank my family and friends that always support and encourage my journey far away from home. Last but not least, to my girlfriend, Monica Caldeira for her support, love, and partnership during grad school.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	ii
TABLE OF CONTENTS.....	iii
LIST OF ILLUSTRATIONS	vi
LIST OF TABLES	xi
ABSTRACT.....	xiii
Chapter I: INTRODUCTION AND LITERATURE REVIEW	15
Soybean production and use.....	15
Genetic diversity	16
Genetic Gain.....	18
Nested association mapping (NAM)	20
Genomic-based Predictions.....	22
Protein and oil content.....	24
Fatty acids	26
Amino acids.....	28
References	30
Chapter II: NESTED ASSOCIATION MAPPING OF GRAIN YIELD, PLANT MATURITY, PLANT HEIGHT AND LODGING IN THREE INTERSPECIFIC SOYBEAN POPULATIONS	46
Abstract	46
Introduction	47
Materials and methods	51
Germplasm development and phenotyping	51
Genotyping and SNP analysis	53
Nested association mapping analysis	54
Results	56
Phenotypic variation	56
Linkage Disequilibrium and Population Structure	57
Nested association mapping	58

Linkage Mapping.....	60
Discussion	62
References	68
2 Tables and Figures.....	80
Chapter III: NESTED ASSOCIATION MAPPING OF SEED COMPOSITION TRAITS IN THREE INTERSPECIFIC SOYBEAN POPULATIONS	113
Abstract	113
Introduction	115
Materials and methods	118
Germplasm and experimental design	118
Phenotypic data collection.....	120
Genotyping and SNP analysis	122
Nested association mapping analysis	123
Results	124
Linkage Disequilibrium and Population Structure	124
Phenotypic variation.....	125
Nested Association mapping	126
Linkage analysis	127
Discussion	129
References	135
3 Tables and Figures.....	146
Chapter IV: PROSPECTS OF GENOMIC PREDICTION USING TRAINING POPULATION DESIGN IN AN INTERSPECIFIC SOYBEAN NAM PANEL.....	197
Abstract	197
Introduction	199
Materials and methods	203
Germplasm development and phenotyping	203
Genotyping and SNP analysis	205
Statistical analysis and training population design.....	206

Results	207
Discussion	210
References	214
4 4 Tables and figures.....	224
Chapter V: OVERALL SUMMARY AND FUTURE PROSPECTS	229
References	231
VITA.....	233

LIST OF ILLUSTRATIONS

Figure 2-1. Histogram of grain yield (GY), plant maturity (PM), plant height (PH), and lodging (LOD) among the NAM population and the hub-parent (Williams 82) in 2016 and 2017..... 82

Figure 2-2. Linkage disequilibrium (LD) decay of the nested association mapping (NAM) population (A), Distribution of the NAM RILs under PC1 and PC2 (B), and fixation index of the NAM population over the entire genome where the red line represents the 99th percentile (0.29)..... 83

Figure 2-3. Manhattan plots of nested association mapping for grain yield, plant maturity, plant height, and lodging plotted against positions on each of the 20 chromosomes, across eight environments during 2016 (red dots) and 2017 (blue dots). The significant SNPs were distinguished by the FDR 0.05 (3.8×10^{-5}) threshold line.... 84

Figure 2-4. Manhattan plot of nested association mapping for grain yield (A), plant maturity (B), plant height (C), and lodging (D) plotted against positions on each of the 20 chromosomes in individual environments during 2016 (red dots) and 2017 (blue dots). The significant SNPs were distinguished by the FDR 0.05 (3.8×10^{-5}) threshold line..... 88

Figure 2-5. The difference in genotypic values (A, B, C) between lines segregating at SNP ss715593853 (A), ss715608829 (B), and ss715613171 (C) associated with plant maturity in the NAM population. The difference in genotypic values of plant height associated with SNP ss715616047 (D), and lodging associated with SNP ss715602804 (E), between lines in the NAM population across eight environments. The difference in genotypic values between lines segregating at SNP ss715625973 for grain

yield at Pop3 –Williams 82 x PI522226 across eight environments (F). The least square difference (LSD) and differences in mean (Δm), and the probability of a greater F statistic ($Pr > F$) and R^2 are in the figures. 89

Figure 2-6. Graphical display of significant QTL detected in three populations (Pop1- Williams 82 x PI464890B; Pop2 – Williams 82 x PI458536; Pop3 – Williams 82 x PI522226) for (A) grain yield, (B) plant maturity, (C) plant height and (D) lodging across eight environments. The dashed dark-red horizontal line indicates the threshold LOD score based on a 1000 permutation test. 90

Figure 2-7. Graphical display of significant QTL detected in three populations (Pop1- Williams 82 x PI464890B; Pop2 – Williams 82 x PI458536; Pop3 – Williams 82 x PI522226) for grain yield, plant maturity, plant height and, lodging in across eight environments in 2016 and 2017. The dashed dark-red horizontal line indicates the threshold LOD score based on a 1000 permutation test. 98

Figure 3-1. Quantitative trait loci associated with fatty acid and amino acid composition. Mean separation of black seed coat (Bl) and yellow (Y) seed coat for samples estimated using near-infrared reflectance (NIR) spectroscopy and samples estimated by analytical chemistry (Analytical) or wet lab for fatty acid (A) and amino acid composition (B). 16:0, palmitic acid; 18:0, stearic acid; 18:1, oleic acid; 18:2, linoleic acid; 18:3, linolenic acid; Cys, cysteine; Met, methionine. 146

Figure 3-2. Linkage disequilibrium (LD) decay of the nested association mapping (NAM) population (A), distribution of the NAM RILs under PC1 and PC2 (B). 146

Figure 3-3. Histogram of the frequency distribution for seed composition traits in the NAM panel. Shown is the average of the measurements across eight environments

during 2016 and 2017. Data presented as g kg^{-1} , amino acids concentration were adjusted for protein content..... 152

Figure 3-4. Pearson's phenotypic correlation coefficients between soybean seed protein, oil, fatty acid, and amino acids concentration adjusted by protein content. Correlation is significantly different from zero when $P\text{-value} > 5.14 \times 10^{-3}$. A cross sign marks non-significant correlations..... 153

Figure 3-5. Manhattan plot of SNPs and associated LOD score using the NAM panel for protein (A) and oil (B) across eight environments during 2016 (red dots) and 2017 (blue dots). The significant SNPs were distinguished by the FDR 0.05 (3.8×10^{-5}) threshold line. Mean separation of genotypic groups for significant SNP's associated with protein and oil on lead SNP ss715621669 (C), and ss715637176 (D) over eight environments. The least significance difference (LSD), differences in mean (Δm), the probability of the greater F statistic ($\text{Pr}>F$), and R^2 are in the figures..... 169

Figure 3-6. Manhattan plot of SNPs and associated LOD score using the NAM panel for cysteine (A), methionine (B), isoleucine (C), leucine (D), glycine (E), and alanine (F), across eight environments during 2016 (red dots) and 2017 (blue dots). The significant SNPs were distinguished by the FDR 0.05 (3.8×10^{-5}) threshold line. 170

Figure 3-7. Manhattan plot of SNPs and associated LOD score using the NAM panel for aspartic acid (A), glutamic acid (B), lysine (C), proline (D), threonine (E), and valine (F), across eight environments during 2016 (red dots) and 2017 (blue dots). The significant SNPs were distinguished by the FDR 0.05 (3.8×10^{-5}) threshold line. 171

Figure 3-8. Mean separation of genotypic groups of RILs for significant SNP's associated with cysteine (A), methionine (B), isoleucine (C), threonine (D), leucine (E),

alanine (F), glycine (G), and lysine (H) over eight environments. The least significance difference (LSD), differences in mean (Δm), the probability of the greater F statistic ($Pr > F$), and R^2 are in the figures. 172

Figure 3-9. Mean separation of genotypic groups of RILs for significant SNP's associated with leucine (A), aspartic acid (B), glutamic acid (C), and valine (D) over eight environments. The least significance difference (LSD), differences in mean (Δm), the probability of the greater F statistic ($Pr > F$), and R^2 are in the figures. 172

Figure 3-10. Manhattan plot of SNPs and associated LOD score using the NAM panel for stearic acid (A), palmitic acid (B), oleic acid (C), linoleic acid (D), and linolenic acid (E), across eight environments during 2016 (red dots) and 2017 (blue dots). The significant SNPs were distinguished by the FDR 0.05 (3.8×10^{-5}) threshold line. Mean separation of genotypic groups for significant SNPs associated with fatty acids composition (F, G, H, I, J) over eight environments. The least significance difference (LSD), differences in mean (Δm), the probability of the greater F statistic ($Pr > F$), and R^2 are in the figures. 173

Figure 3-11. Manhattan plot of SNPs and associated LOD score using NAM panel for protein (A), oil (B), palmitic acid (C), stearic acid (D), oleic acid (E), linoleic acid (F), linolenic acid (G), alanine (H), aspartic acid (I), cysteine (J), glutamic acid (K), glycine (L), isoleucine (M), leucine (N), lysine (O), methionine (P), proline (Q), threonine (R), and valine (S) in individuals environments (Albany, Columbia, Novelty and Rock Port) during 2016 (red dots) and 2017 (blue dots). The significant SNPs were distinguished by the FDR 0.05 (3.8×10^{-5}) threshold line. 192

Figure 4-1. Effect of training population size (A) and marker density (B) on the prediction accuracy of grain yield, maturity, protein and oil..... 225

Figure 4-2. Boxplots of the prediction accuracy of the genomic prediction models (RR-BLUP, BayesB, and BayesLASSO) in soybeans tested in three training populations designs for grain yield (A), maturity (B), protein (C), and oil (D). Training population designs included Across all families (Across/All), Leave one family out (Leave_one/Fam_out) and Within families (Within/Fam). 226

LIST OF TABLES

Table 2-1. Summary of mean square estimates from the analysis of variance for yield, plant maturity, plant height, and lodging during 2016 and 2017. ** and * Significant at the 0.001 and 0.01 probability level. h^2 , heritability on an entry-mean basis; W82, Williams 82 (hub-parent); SD, standard deviation 80

Table 2-2. Significant SNPs for grain yield, plant maturity, plant height, and lodging and their respective allelic effect in relation to the hub parent Williams 82 across eight environments during 2016 and 2017..... 105

Table 2-3. Significant QTL results in three populations (Pop1- Williams 82 x PI464890B; Pop2 – Williams 82 x PI458536; Pop3 – Williams 82 x PI522226) for four agronomic traits across eight environments in two years (2016 and 2017)..... 109

Table 2-4. Selected *G. soja*-derived RILs exhibiting significantly higher grain yield than the NAM hub-parent Williams 82 for test MG III and MG IV averaged over eight environments in 2016 and 2017..... 110

Table 3-1. Descriptive statistics for NIR calibrations used for estimating 19 seed composition traits of ground soybean samples (updated 2017, Perten Instruments)..... 147

Table 3-2. Mean square estimates of an analysis of variance for protein, oil, fatty acids, and amino acids during 2016 and 2017. ***, ** and * Significant at the 0.001, 0.01, and 0.05 probability level. h^2 , heritability on an entry-mean basis; SD, standard deviation; W82, mean of the hub parent (Williams 82)..... 148

Table 3-3. Significant SNPs for protein, oil, palmitic acid, stearic acid, linoleic acid, linolenic acid, oleic acid, alanine, aspartic acid, cysteine, glutamic acid, glycine, isoleucine, leucine, lysine, methionine, proline, threonine, and valine and their respective

allelic effect in relation to the hub parent Williams 82 across 8 environments during 2016 and 2017. Chr, chromosome; Position, in base pair; MAF, minimum allele frequency; 154

Table 3-4. Significant QTL results in three populations (Pop1- Williams 82 x PI464890B; Pop2 – Williams 82 x PI458536; Pop3 – Williams 82 x PI522226) for 19 seed composition traits across eight environments in two years (2016 and 2017). 193

Table 3-5. Major effect QTL and respective candidate genes associated with seed composition traits. 195

Table 4-1. Summary of descriptive statistics for yield, maturity, protein and oil across eight environments. 224

Table 4-2. Summary of phenotypic data analysis for grain yield, maturity, protein and oil. 224

Table 4-3. Prediction accuracy of three genomic prediction models including RR-BLUP, BayesB, and BayesLASSO, for grain yield, maturity, protein and oil using Within family training population design. 227

Table 4-4. Prediction accuracy of three genomic prediction models including RR-BLUP, BayesB, and Bayes LASSO, for grain yield, maturity, protein and oil in a leave one family out training population design. 228

ABSTRACT

Given the narrow genetic base of soybean, discovering useful traits in exotic germplasm could increase the diversity in the current elite gene pool. However, it is essential to characterize beneficial alleles from the wild soybean (*Glycine soja*) to enhance genetic gain. The objective of this study was to investigate grain yield, agronomic traits and seed composition traits using a soybean NAM panel containing crosses between Williams 82 (hub-parent) and three *Glycine soja* parents (PI464890B, PI458536, and PI522226). Field tests were conducted in Albany, Columbia, Novelty and Rock Port, Missouri for two years, 2016 and 2017, in an augmented incomplete block design with one replication in 2016 and two replications in 2017. The nested association mapping and linkage mapping could identify three major QTL for plant maturity from *Glycine soja* in Chromosomes 6, 11 and 12 presenting a significant increment in days to maturity. A major QTL for plant height was identified in Chromosome 13 and showed an increase in plant height for lines that carried the wild soybean allele. A significant QTL for grain yield from *Glycine soja* was detected in Chromosome 17 and showed a positive effect of 166.1 kg ha⁻¹ and yielded an average of 6% more than the *Glycine max* parent (Williams 82) across environments. Also, we identified 61 and 12 QTLs associated with seed composition traits in the NAM analysis and linkage analysis, respectively. Four QTLs showed pleiotropic effects with soybean seed composition traits. Two QTLs, one on Chromosome 5 and another on Chromosome 15 were associated with the fatty acid profile, explaining 3-18% of the phenotypic variance. The confirmed QTLs for protein and oil cqSeed protein-001 on Chromosome 15 and cqSeed protein-003 on Chromosome 20 were identified. Also, the QTL on chromosome 20 was associated with ten amino

acids. However, the allele associated with protein concentration was also responsible for a reduction in amino acid concentration. Another QTL on Chromosome 19 was associated with Cysteine, Methionine, and Leucine and explained 9-30% of the phenotypic variation. Our results reinforce that increasing protein may not increase amino acid concentrations and suggest independent genetic control for protein and sulfur-containing amino acids. In addition to the mapping study, we conducted a genomic prediction study in the NAM panel. Increasing the training population size from 50 to 300 individuals improved prediction accuracy from 0.49 to 0.77 (57% increase) across all traits, with little increment between 300 and 390 individuals (1%). Marker density had little impact on the prediction accuracy across traits, with a significant increment in prediction accuracy up to 1423 markers (18.5%). The training population design Across all families had higher prediction accuracies for all the traits compared with Leave one family out and Within families designs, with prediction accuracies ranged from moderate (0.55) to high (0.75). The NAM panel containing interspecific crosses, successfully predict polygenic traits. Our results showed encouraging prediction accuracies for grain yield (0.55-0.73), which is impressive from crosses originating from wild soybean. In conclusion, training population strategies where population size and multiple families were maximized (Across all families design) produce robust prediction accuracies for yield, maturity, protein, and oil. Genomic predictions might also accelerate genetic gain in pre-breeding efforts using wild soybeans.

Chapter I:

INTRODUCTION AND LITERATURE REVIEW

Soybean production and use

Soybean [*Glycine max* (L.) Meer] is the leading oilseed crop produced worldwide (FAOSTAT, 2018). World soybean production has tripled, between 1980 and 2017, with a total production of 240 million metric tons in 2016/2017 (FAOSTAT, 2018). The three top soybeans producers worldwide are the U.S.A., Brazil, and Argentina. These three countries are responsible for more than 80% of the total soybean production in the world (USDA - United department of Agriculture 2018).

The large demand for soybean products and the rise of international prices have increased the cultivated area in the U.S. to more than 36 million hectares, resulting in a record production of 119.5 million metric tons in 2017 (USDA - United department of Agriculture 2018). The total value of the U.S soybean crop was estimated at nearly \$41 billion in 2017, an increment of more than twofold when compared to the amount in 2006 (SoyStats, 2018).

Cultivated soybean seed typically has 20% oil and 40% protein (Wilson 2004) on a dry weight basis. The protein and oil are used in human diets or livestock feed and crucial in determining the value of the soybean crop globally. Soybean meal is an abundant protein supplement after the oil removal process and is mostly used in livestock feed and aquaculture, because of its highly digestible protein and complete amino acid profile (Steinfeld et al., 2006). Soybean oil can be used as cooking oil during food processing, as vegetable oil and as an ingredient in foods like salad dressing. Also, soybean oil is

converted into products like lubricants, solvents, cleaners, paints, and biodiesel (Johnson and Myers 1995; Lin et al. 2011)

Genetic diversity

Soybean was domesticated from the wild progenitor (*Glycine soja* Sieb. and Zucc.) 6,000 to 9,000 years ago, limiting the center of origin to East Asia, including, Korea, Japan and China (Carter et al. 2004; Doebley et al. 2006). Both *G. soja* and *G. max* have 20 chromosomes ($2n = 40$) and can be cross-pollinated, carrying out normal meiotic chromosome pairing, and producing fertile offspring (Carter et al. 2004).

During the domestication process more than half of the genetic diversity found in *G. soja* was lost, decreasing pairwise diversity (π) from 2.94×10^{-3} in *Glycine soja* to 1.05×10^{-3} in modern cultivars and increase linkage disequilibrium (Zhou et al. 2015b). Also, Hyten et al. (2006) compared modern varieties, Asian landraces, and *Glycine soja* accessions and observed a reduction in 81% of rare alleles in the modern soybean during domestication. Soybean has been bred in East Asia for thousands of years, but it was just introduced in the U.S. in 1765 (Hymowitz and Harlan 1983), although commercial expansion started in the 1940s.

North America soybean ancestors represent 80% of the genetic diversity from Asian landraces (Hyten et al. 2006). However, Gizlice et al. (1994) conducted a pedigree analysis of 258 public soybean varieties released in the United States between 1947 and 1988 and indicated that most modern public cultivars could be traced back to 35 founding plant introductions or landraces. Therefore, the lack of diversity in the current soybean elite germplasm creates a need for introgression of exotic germplasm to allow continuous

genetic gain into the future. Currently, it is estimated that ~8,500 unique accessions of *G. Soja* and 45,000 accessions of Asian landraces of *G. max* are maintained in germplasm banks around the world (Hyten et al. 2006; Wen et al. 2009).

Cultivated soybean has been selected for favorable traits useful in the modern agricultural system. The wild soybean traits include late flowering, excessive lodging, hard seed coat, small and black seeds, pod shattering, and a lack of uniformity in maturity (Liu et al. 2007). However, the *G. soja* gene pool represents a valuable resource for the improvement of the modern soybean, especially if more favorable alleles are identified for breeding purposes (Sedivy et al., 2017). Several genes and genomic regions have been identified in *G. soja* accessions for abiotic stresses such as exposure to salt and drought (Lee et al. 2009; Tuyen et al., 2010; Chen et al. 2013; Ning et al. 2017), biotic stresses such as soybean aphid (*Aphis glycines*) (Zhang et al. 2017c) and soybean cyst nematode (*Heterodera glycines*) (Wang et al. 2001; Winter et al. 2007; Zhang et al. 2017b) improved seed composition such as increased protein content and improved fatty acid and amino acid profiles (Ha et al. 2014; Yan et al. 2014; Leamy et al. 2017; La et al. 2019), and potentially domestication traits (Zhou et al. 2015a; Bandillo et al. 2017).

However, little research has been done to investigate the potential for wild soybean for improving grain yield. In general, when crosses using *G. soja* were made, multiple backcrossing is required to eliminate undesirable traits from the wild parent (Carpenter and Fehr 1986). Akpertey et al. (2014) evaluated multiple backcross populations from *G. max* x *G. soja* and observed no significant yield increment from *G. soja*-derived lines. Although, QTL for grain yield from *G. soja* have been mapped. Concibido et al. (2003) performed a two-year multi-environment yield trial with 256 BC₂

individuals and identified a QTL for grain yield from a *G. soja* accession and observed a 9% increase in yield from lines that were carrying the wild soybean allele. Also, Li et al. (2008) used a similar approach and identified a QTL associated with grain yield on chromosome 5 (Satt511). The QTL had an additive yield effect of 191 kg ha⁻¹ for the marker regression method and 235 kg ha⁻¹ for the interval mapping method. Across environments in the validation population, lines carrying the *G. soja* allele demonstrated a 6.3% higher yield than lines carrying the recurrent parent allele.

The potential use of exotic germplasm for improving agriculture productivity is enormous, yet only a tiny fraction of the exotic germplasm available is used in commercial breeding aimed at improved elite variety development.

Genetic Gain

Genetic improvements in the yield potential of modern soybean varieties and the implementation of improved management practices have enabled an enormous increase in productivity, from 700 kg ha⁻¹ in 1924 to more than 3000 kg ha⁻¹ in 2017 (USDA - United department of Agriculture 2018). In the last century, recent studies in the United States demonstrate genetic gains in yield have ranged from 12 to 27 kg ha⁻¹ year⁻¹ or close to 1% per year (Rowntree et al. 2013; Rincker et al. 2014; Rogers et al. 2015). The genetic gain in grain yield is ~1.3% per year worldwide (Ray et al. 2013). However, to keep up the new demand for soybeans Ray et al. (2013) estimated a genetic gain of 2.4% per year needed well into the future.

To meet this challenge and to increase genetic gain in crops, plant breeders have to use genomic technologies currently available to increase selection intensity and genetic

variance as well as decrease/eliminate breeding cycles and non-genetic variance. Genomic selection (GS), a concept proposed by Meuwissen et al. (2001), has emerged as a solution for the deficiencies of marker-assisted selection (MAS) and has recently proven to be a valuable approach for plant breeding (Nakaya and Isobe 2012; Desta and Ortiz 2014).

The genetic gain from selection (ΔG) in the GS model theoretically can be expressed as, $\Delta G = \frac{i r_A \delta_A^2}{\Delta t}$, where i , is the selection intensity; r_A , is the prediction accuracy; δ_A^2 , is the square root of additive genetic variance; and Δt , is the breeding cycle time (Heffner et al. 2010; Desta and Ortiz 2014). Plant breeders can use genomic selection to increase genetic gain by improving their selection intensity, accuracy, and reducing the breeding cycle (Heffner et al. 2010). Also, GS can allow breeders to accomplish selection in non-adaptive environments such as winter nurseries or greenhouses, making possible at least two to three generations of selection and recombination per year in annual crops.

Using GS for oil palm (*Elaeis guineensis*), Wong and Bernardo (2008) found a 4-25% larger response, 26-57% lower cost per unit gain, and a significant reduction in the breeding cycle comparing GS with phenotypic selection. Also, it has been shown that GS can outperform MAS and phenotypic selection in several crop species, such as soybean (*Glycine max*) (Zhang et al. 2016b), corn (*Zea mays*) (Cerrudo et al. 2018), rice (*Oryza sativa*) (Spindel et al. 2015), and wheat (*Triticum aestivum*) (Battenfield et al. 2016).

Increasing genetic diversity is essential to continually achieving genetic gain (Daetwyler et al. 2015). This is particularly true during multiple cycles of selection when low-frequency alleles can be lost (Jannink 2010). GS strategies to preserve levels of

genetic diversity have been tested. Daetwyler et al. (2015), utilized optimal haploid breeding value (OHV) simulations in double haploid corn populations and observed up to 0.6 standard deviations more genetic gain than GS while preserving a greater amount of genetic diversity. Müller et al. (2018), used a similar strategy and observed that OHV maintain greater genetic diversity while increasing genetic gain. GS should be expanded on the more diverse panel and, such as Multi-parent Advanced Generation Inter Crosses (MAGIC) and Nested Association Mapping (NAM) populations (Yu et al. 2008; Kover et al. 2009).

Nested association mapping (NAM)

Quantitative trait loci (QTL) mapping and Genome-wide association studies (GWAS) have been used extensively to predict the genetic basis of polygenic traits (Holland 2007). One drawback of QTL mapping is that the estimated QTL interval is usually wide, due to the extensive linkage disequilibrium (LD) in the population, and the traditionally low marker density throughout the genome. QTLs are often population or environment specific (Bernardo 2010), which does not generate reproducible results in non-related environments, populations, and individuals.

As an alternative, GWAS relies on historical LD to increase the analysis power (Zhu et al. 2008), without the time and effort associated with population development (Korte and Farlow 2013). GWAS studies have identified many nucleotide variants affecting complex traits in soybeans, such as oil and protein (Hwang et al. 2014; Cao et al. 2017), flowering time and maturity (Zhang et al. 2015; Mao et al. 2017), soybean cyst nematode (*Heterodera glycines*) (Zhang et al. 2016). However, this method has low

power to detect rare variants and could be confounded by the population structure, since unbalanced populations can lead to false marker-trait associations (Price et al. 2010; Auer and Lettre 2015).

Maize geneticists developed a new approach to overcome some pitfalls of conventional GWAS and linkage mapping techniques (Yu et al. 2008). The design consisted of multi-cross recombinant inbred lines (RILs) where each family has a unique parent crossed with a common “hub” parent, (or “founder” line). The NAM design combines aspects of two genetic approaches: linkage mapping through genetic recombination and association mapping through historical linkage disequilibrium presented in the populations, improving the resolution and the power for detection of rare alleles in the analysis (Yu et al. 2008). NAM populations have also been used successfully to dissect complex traits in several crop species, such as rice (*Oryza sativa*) (Fragoso et al., 2017), barley (*Hordeum vulgare*) (Sharma et al. 2018), and wheat (*Triticum aestivum*) (Bajgain et al. 2016). In soybean, Li et al. (2017) identified 139 QTLs related to flowering time, explaining more than 80% of the phenotypic variation presented in that population, demonstrating the power of marker-trait detection in the NAM design.

The SoyNAM panel (<https://www.soybase.org/SoyNAM/>), was derived from crosses between IA3023 (hub parent) and 40 different *G. max* elite cultivars, plant introductions and breeding lines with exotic ancestry. Xavier et al. (2018) utilized the SoyNAM panel and identified a genomic region positively associated with grain yield stability across a range of environments in chromosome 18, giving insights about the QTL x environment interaction presented in the trait. In another study using the same

NAM panel, Xavier et al. (2017) identified a QTL for canopy coverage on Chromosome 19 provides an increase in grain yield of $\sim 50 \text{ kg ha}^{-1}$, due to the high correlation (0.87) between canopy coverage and grain yield, indirect selection for yield in early season thought canopy coverage may have a great application in soybean breeding programs. More recently using the same SoyNAM panel tested in 22 environments across the United States Midwest, Diers et al. (2018) identified markers associated with yield, plant maturity, plant height, plant lodging, and seed weight. The allelic effects ranged between families, with a high frequency of positive yield alleles from elite founders, yet favorable yield alleles from exotic germplasm were identified.

Genomic-based Predictions

Selection of critical quantitative traits in plants is traditionally based on the phenotypic evaluation of the individuals or population. However, with the new advances in high-throughput genotyping the cost per data point of molecular markers has drastically reduced and the use of marker-assisted selection (MAS) in plant breeding has increased dramatically in the last decade (Heffner et al. 2009).

Traditional MAS is a multiple step technique that requires the identification and validation of quantitative trait loci (QTLs), which have made a considerable contribution to understanding the genetic architecture of complex traits. However, the results have been inconsistent, and the statistical methods used to identify QTLs were inadequate to detect small effects QTLs (Bernardo 2008). With complex polygenic traits, such as grain yield, these small effect genes play an essential role in the trait genetic architecture (Wang et al. 2016). These issues explain why the implementation of MAS has been

limited in breeding programs, and genetic gain has been relatively poor (Dekkers 2004; Bernardo 2008)

An alternative to traditional MAS method was proposed by Meuwissen et al. (2001), the genomic prediction (GP) approach, can predict breeding value by estimating the genetic variance attributed to each locus within a dense marker panel. This analysis attempts to capture the total additive variance from all markers available instead of just significant makers (Hayes et al. 2009). Using genome-wide markers, it is expected that every trait locus to be in linkage disequilibrium (LD) with at least one marker locus in the population, improving the prediction of complex quantitative traits. GP involves two phases; first it uses a “training population” of individuals that have been genotyped and phenotyped to fit the model, and second, a “testing set” of individuals that only have been genotyped is utilized (Jannink et al. 2010). Genomic estimated breeding values (GEBV) of individuals are calculated, and these GEBVs summarize the presence of favorable alleles from a genome-wide markers analysis.

Several statistical methods have been developed for GP: Ridge regression BLUP (RR-BLUP) (Endelman 2011), genomic best linear unbiased prediction (GBLUP) (VanRaden 2008), and Bayesian methods (BayesA, BayesB, BayesC π) (Meuwissen et al. 2001; Habier et al. 2011). Daetwyler et al. (2013) studied wheat and pine datasets and observed no significant differences between models when the polygenic model was used, whereas Bayesian models had better predictions when fewer QTLs are evaluated and concluded that no single model could be used as a standard for predictions. Also, new studies are allowing to incorporate GxE interactions (Crossa et al. 2017), epistatic effects

(Duhnen et al. 2017), fixed effect markers (Spindel et al. 2016), and crop growth modeling (Technow et al. 2015; Cooper et al. 2016).

A large amount of GP studies have been reported in crop species such as wheat (*Triticum aestivum*) (Bassi et al. 2015); corn (*Zea mays*) (Zhang et al. 2017d), rice (*Oryza sativa*) (Spindel et al. 2015), and soybean (*Glycine max*) (Jarquín et al. 2014; Zhang et al. 2016b; Jarquin et al. 2016). Duhnen et al. (2017) evaluated soybean breeding lines from MGIII and IV and reported accuracies for grain yield ranging from 0.47 to 0.63, depending on the training population and genomic prediction model.

The prediction accuracy of GP is affected by various factors related to; the genetic architecture of the trait under study (Jannink 2010; Lorenz 2013). Highly heritable traits were expected to have higher prediction accuracies, based on the relatedness of the training set and testing set (Habier et al. 2007, 2010). More related individuals produce higher prediction accuracies; based on the marker density used in the prediction, generally prediction accuracy increase with more markers used in the model. However, this is not always the case (Heffner et al. 2011; VanRaden et al. 2011). For example, in the SoyNAM panel (Xavier et al. 2016) observed marginal gain in prediction accuracy after using 1020 markers, based on the size of the training population, with larger training sets tending to increase accuracy (Isidro et al. 2015; Xavier et al. 2016).

Protein and oil content

Cultivated soybean seed typically has 20% oil and 40% protein on a dry weight basis (Wilson 2004). Protein meal is a major product of soybean processing and can be used in human diet or livestock feed, whereas oil can be used as cooking oil and has a

variety of industrial applications. Understanding the genetic architecture of seed protein and oil is essential to improve seed quality, and also the competitiveness of the crop. Wilson (2004), reported a variation of 34 to 56% protein concentration and 8 to 27% oil content in the USDA soybean germplasm collection. La et al. (2019) evaluated a core collection of *G. soja* accessions and reported a variation of 36 to 40.2% for the *G. max* checks and 39.2 to 48.1% for the wild accessions. There have been many studies to date that show a relatively wide range of protein and oil concentration across germplasm collections, yet there is still relatively little information on the complete genetic architecture of these traits.

Rincker et al. (2014) evaluated a set of soybean cultivars released between 1923-2008 and observed a decrease in protein concentration at a rate of $0.22 \text{ g kg}^{-1} \text{ yr}^{-1}$ (MG II and III), and $0.16 \text{ g kg}^{-1} \text{ yr}^{-1}$ (MG IV). Otherwise, the oil content increased in the same period. This reduction in protein concentration is largely due to the improvement of grain yield across the decades of breeding, and the generally significant negative correlation observed between yield and protein and protein and oil (Chung et al. 2003; La et al. 2019).

Protein and oil content in soybean is genetically complex and highly polygenic, and also shows significant interaction with the environment (Chung et al. 2003; Lee et al. 2007; Warrington et al. 2015). Several QTLs have been identified for protein and oil (Diers et al. 1992; Brummer et al. 1997; Panthee et al. 2005; Hwang et al. 2014a). Phansak et al. (2016) evaluated 48 biparental populations and detected eight QTLs associated with protein and ten with oil content.

Among 241 QTLs reported for seed protein content and 315 for oil content (<http://www.soybase.org>, “SoyBase browser,” verified 01/20/2019), two confirmed QTLs cqSeed protein-001 (Chr 15) and cqSeed protein-003 (Chr 20), that were mapped in multiple populations and sources (Diers et al. 1992; Brummer et al. 1997; Chung et al. 2003; Warrington et al. 2015; Phansak et al. 2016), and were positively associated with protein and negative with oil content. In general, the cqSeed protein-003 increase seed protein content by 2% and typically resulted in a simultaneous 1% decrease in seed oil content. This negative relationship is due to pleiotropic effects between high protein alleles (s) and low oil, or high protein allele(s) in repulsion phase with high oil allele(s) (Chung et al. 2003). In the studies of Sebolt et al. (2000) and Nichols et al. (2006), the cqSeed protein-003 from the *G. soja* PI468916 resulted in a significant reduction of grain yield. However, recently Mian et al. (2017) released “Highpro1” a high protein and yield MG III conventional soybean cultivar. Highpro1 cultivar carries the cqSeed protein-003 from Danbaekdong source (Warrington et al. 2015). Suggesting that increase protein without significant yield losses can be achieved.

Fatty acids

Five fatty acids are found in soybean oil. They include two saturated fatty acids determine the oil quality in soybean, 10% palmitic (16:0) and 4% stearic (18:0), and three unsaturated fatty acids, 22% oleic (18:1), 55% linoleic (18:2), and 10% linolenic (18:3) (Wilson 2004; Lee et al. 2007). Based on the number of double-bonds, fatty acids are categorized into saturated acids (no double bonds – 16:0 and 18:0), monosaturated acid

(one double bond – 18:0) and, polyunsaturated acid (more than one double bond – 18:2 and 18:3),

Polyunsaturated fatty acids are more susceptible to oil oxidation, which deteriorates the oil properties and reduces shelf life (Mounts et al. 1988; Kamal-Eldin 2006). Hydrogenation is used to reduce the unstable polyunsaturated portion in the soybean oil. However, during the process trans fats are generated. Trans fats acids have been associated with negative health risks, such as coronary heart disease (Kromhout 1999), obesity, diabetes and elevated cholesterol (Dhaka et al. 2011; Takeuchi and Sugano 2017).

The genetic control of fatty acid composition in soybean is complex and governed by both major and minor genes (Wang et al. 2014a). Many loci have been identified to be associated with the fatty acid profile (Priolli et al. 2015; Fang et al. 2017). Li et al. (2015) identified eight loci associated with linoleic acid, 12 with linolenic acid, 5 with oleic acid, 7 with palmitic acid and 5 with stearic acid. Recently, Leamy et al. (2017) evaluated a set of wild soybeans and identified novel QTLs associated with fatty acid composition.

Soybean geneticists and breeders have been manipulating levels of individual fatty acids and changing the oil profile of the seed for decades. Recent advances in oil quality have been achieved using a combination of mutated fatty acid desaturase genes, loss of function for *fad2-1A*, *fad2-1B*, and *fad3*, results in a high oleic (>80%) low linolenic (<3%) soybean oil (Pham et al. 2010, 2012). Transgenic and genome editing approaches have been successfully implemented to increase oleic acid content in soybeans seeds (Hoshino et al. 2010; Haun et al. 2014; Sweeney et al. 2017). Oleic acid increases stability and expands the utility of soybean oil. Oil with higher content of oleic

acid would have greater applications in food processing and industrial products (Lee et al. 2007; Clemente and Cahoon 2009).

Amino acids

Amino acids are divided into, essential means they cannot be synthesized by animals, and non-essential can be synthesized by animals. Among 20 amino acids were found in soybean seeds, cysteine (Cys), methionine (Met), lysine (Lys), tryptophan (Try) and, threonine (Thr) have been identified as the most important for the monogastric animals, such as poultry and swine (Boisen et al. 2000).

Soybean meal is deficient in sulfur-containing amino acids Met, Thr, Cys, and Lys (Warrington et al. 2015). Met and Cys are essential amino acids and cannot be synthesized by monogastric animals (Wu 2014). Therefore, supplementation with natural sources or synthetic amino acids is used by the industry to overcome the deficiency in soybean meal. The deficiency in amino acids costs animal producers and the industry has been reported in 2001 to be approximately 100 million dollars annually (Imsande 2001).

The value of the soybean crop can be enhanced through genetic improvement of the amino acids limited in soybean meal, such as methionine and cysteine (Pantalone 2012; Zhang et al. 2018a). QTLs have previously been reported for essential amino acids such as methionine and cysteine (Panthee et al. 2006). Wang et al. (2015) identified eight QTLs associated with cysteine and methionine content, and three of them were confirmed in another genetic background.

Warrington et al. (2015) evaluated 140 F₅-RILs from Benning x Danbaekkong cross and identified QTLs associated with Lys (Chr 08, 20), Thr (Chr 01, 09, 17), Met

(Chr 06, 09, 10), Cys (Chr 10) and, a pleiotropic QTL on Chr 20 associated with protein, Lys, Cys, Thr, Met, and Met+Cys. Also, they observed that the allele from Danbaekong on Chr 20 increased protein, but reduced Met, Cys and Lys levels. Recently, Li et al. (2018) using 345 RILs from Zheng 92116 (46% protein) x Liaodou 14 (38% protein) observed twelve cluster QTLs on Chrs 6, 9, 13, 17 and 20 for 17 amino acids evaluated, the QTL on Chr 17 was associated with Gly, Ala, Ile, Leu, Val, Met, Asp, Glu, Lys and Phe and explained 8.8 to 16.8% of the phenotypic variation for the related amino acids. Also, La et al. (2019) evaluated *Glycine soja* accessions and identified a QTL on Chr 13 associated with Asp and Glu content. Also, a higher amino acid range was observed in the wild accessions for both Cys (14.9-19.2 g kg⁻¹) and Met (14-16.9 g kg⁻¹) compared to *Glycine max* checks cultivars (Cys, 15.3-16.5 g kg⁻¹; Met, 15.1-16.5 g kg⁻¹).

The challenge is improving amino acid levels while maintaining acceptable levels of both protein and oil in the seed. Soybean meal with higher levels of essential amino acids would generate an increment in the profitability of the animal industry due to less need for amino acid supplementation.

References

- Akperthey, A., M. Belaffif, G.L. Graef, M.A. Rouf Mian, J. Grover Shannon, et al. 2014. Effects of selective genetic introgression from wild soybean to soybean. *Crop Sci.* 54(6): 2683–2695. doi: 10.2135/cropsci2014.03.0189.
- Auer, P.L., and G. Lettre. 2015. Rare variant association studies: considerations, challenges and opportunities. *Genome Med.* 7(1): 16. doi: 10.1186/s13073-015-0138-2.
- Bajgain, P., M.N. Rouse, T.J. Tsilo, G.K. Macharia, S. Bhavani, et al. 2016. Nested Association Mapping of Stem Rust Resistance in Wheat Using Genotyping by Sequencing (G. Bai, editor). *PLoS One* 11(5): e0155760. doi: 10.1371/journal.pone.0155760.
- Bassi, F.M., A.R. Bentley, G. Charmet, R. Ortiz, and J. Crossa. 2015. Breeding schemes for the implementation of genomic selection in wheat (*Triticum* spp.). *Plant Sci.* 242: 23–36. doi: 10.1016/j.plantsci.2015.08.021.
- Battenfield, S.D., C. Guzmán, R.C. Gaynor, R.P. Singh, R.J. Peña, et al. 2016. Genomic Selection for Processing and End-Use Quality Traits in the CIMMYT Spring Bread Wheat Breeding Program. *Plant Genome* 9(2): 0. doi: 10.3835/plantgenome2016.01.0005.
- Bernardo, R. 2008. Molecular markers and selection for complex traits in plants: Learning from the last 20 years. *Crop Sci.* 48(5): 1649–1664. doi: 10.2135/cropsci2008.03.0131.
- Bernardo, R. 2010. *Breeding for Quantitative Traits in Plants*. Second. Stemma Press, Woodbury.

Boisen, S., T. Hvelplund, and M.R. Weisbjerg. 2000. Ideal amino acid profiles as a basis for feed protein evaluation. *Livest. Prod. Sci.* 64(2–3): 239–251. doi: 10.1016/S0301-6226(99)00146-3.

Brummer, E.C., G.L. Graef, J. Orf, J.R. Wilcox, and R.C. Shoemaker. 1997. Mapping QTL for Seed Protein and Oil Content in Eight Soybean Populations. *Crop Sci.* 37(2): 370. doi: 10.2135/cropsci1997.0011183X003700020011x.

Carpenter, J.A., and W.R. Fehr. 1986. Genetic variability for desirable agronomic traits in populations containing Glycine soja germplasm. *Crop Sci.* 26(4): 681–686. doi: 10.2135/cropsci1986.0011183X002600040008x.

Carter, T.E., T. Hymowitz, and R.L. Nelson. 2004. Biogeography, Local Adaptation, Vavilov, and Genetic Diversity in Soybean. *Biological Resources and Migration*. Springer Berlin Heidelberg, Berlin, Heidelberg. p. 47–59

Cerrudo, D., S. Cao, Y. Yuan, C. Martinez, E.A. Suarez, et al. 2018. Genomic Selection Outperforms Marker Assisted Selection for Grain Yield and Physiological Traits in a Maize Doubled Haploid Population Across Water Treatments. *Front. Plant Sci.* 9: 366. doi: 10.3389/fpls.2018.00366.

Chung, J., H.L. Babka, G.L. Graef, P.E. Staswick, D.J. Lee, et al. 2003. The seed protein, oil, and yield QTL on soybean linkage group I. *Crop Sci.* 43(3): 1053–1067. doi: 10.2135/cropsci2003.1053.

Clemente, T.E., and E.B. Cahoon. 2009. Soybean oil: genetic approaches for modification of functionality and total content. *Plant Physiol.* 151(3): 1030–40. doi: 10.1104/pp.109.146282.

Concibido, V., B. La Vallee, P. McIaird, N. Pineda, J. Meyer, et al. 2003. Introgression of a quantitative trait locus for yield from Glycine soja into commercial soybean cultivars. *Theor. Appl. Genet.* 106(4): 575–582. doi: 10.1007/s00122-002-1071-5.

Cooper, M., F. Technow, C. Messina, C. Gho, and L. Radu Totir. 2016. Use of crop growth models with whole-genome prediction: Application to a maize multienvironment trial. *Crop Sci.* 56(5): 2141–2156. doi: 10.2135/cropsci2015.08.0512.

Crossa, J., P. Pérez-Rodríguez, J. Cuevas, O. Montesinos-López, D. Jarquín, et al. 2017. Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. *Trends Plant Sci.* 22(11): 961–975. doi: 10.1016/j.tplants.2017.08.011.

Daetwyler, H.D., M.P.L. Calus, R. Pong-Wong, G. de los Campos, and J.M. Hickey. 2013. Genomic prediction in animals and plants: Simulation of data, validation, reporting, and benchmarking. *Genetics* 193(2): 347–365. doi: 10.1534/genetics.112.147983.

Daetwyler, H.D., M.J. Hayden, G.C. Spangenberg, and B.J. Hayes. 2015. Selection on optimal haploid value increases genetic gain and preserves more genetic diversity relative to genomic selection. *Genetics* 200(4): 1341–1348. doi: 10.1534/genetics.115.178038.

Dekkers, J.C.M. 2004. Commercial application of marker- and gene-assisted selection in livestock: Strategies and lessons. *J. Anim. Sci.* 82: E313–328. doi: 82/13_suppl/E313 [pii].

Desta, Z.A., and R. Ortiz. 2014. Genomic selection: Genome-wide prediction in plant improvement. *Trends Plant Sci.* 19(9): 592–601. doi: 10.1016/j.tplants.2014.05.006.

Dhaka, V., N. Gulia, K.S. Ahlawat, and B.S. Khatkar. 2011. Trans fats-sources, health risks and alternative approach - A review. *J. Food Sci. Technol.* 48(5): 534–41. doi: 10.1007/s13197-010-0225-8.

Diers, B.W., W.R. Fehr, and R.C. Shoemaker. 1992. RFLP analysis of soybean seed protein and oil content. *Theor. Appl. Genet.* 83: 608–612. doi: 10.1007/BF00226905 (accessed 15 September 2018).

Diers, B.W., J. Specht, K.M. Rainey, P. Cregan, Q. Song, et al. 2018. Genetic Architecture of Soybean Yield and Agronomic Traits. *G3.* 8(10): 3367–3375. doi: 10.1534/g3.118.200332.

Doebley, J.F., B.S. Gaut, and B.D. Smith. 2006. The molecular genetics of crop domestication. *Cell* 127(7): 1309–21. doi: 10.1016/j.cell.2006.12.006.

Duhnen, A., A. Gras, S. Teyssèdre, M. Romestant, B. Claustres, et al. 2017. Genomic selection for yield and seed protein content in Soybean: A study of breeding program data and assessment of prediction accuracy. *Crop Sci.* 57(3): 1325–1337. doi: 10.2135/cropsci2016.06.0496.

Endelman, J.B. 2011. Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *Plant Genome J.* 4(3): 250. doi: 10.3835/plantgenome2011.08.0024.

Fang, C., Y. Ma, S. Wu, Z. Liu, Z. Wang, et al. 2017. Genome-wide association studies dissect the genetic networks underlying agronomical traits in soybean. *Genome Biol.* 18(1): 161. doi: 10.1186/s13059-017-1289-9.

Gizlice, Z., T.E. Carter, and J.W. Burton. 1994. Genetic Base for North American Public Soybean Cultivars Released between 1947 and 1988. *Crop Sci.* 34(5): 1143. doi: 10.2135/cropsci1994.0011183X003400050001x.

Habier, D., R.L. Fernando, and J.C.M. Dekkers. 2007. The impact of genetic relationship information on genome-assisted breeding values. *Genetics* 177(4): 2389–2397. doi: 10.1534/genetics.107.081190.

Habier, D., R.L. Fernando, K. Kizilkaya, and D.J. Garrick. 2011. Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics* 12(1): 186. doi: 10.1186/1471-2105-12-186.

Habier, D., J. Tetens, F.R. Seefried, P. Lichtner, and G. Thaller. 2010. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. *Genet. Sel. Evol.* 42(1): 5. doi: 10.1186/1297-9686-42-5.

Haun, W., A. Coffman, B.M. Clasen, Z.L. Demorest, A. Lowy, et al. 2014. Improved soybean oil quality by targeted mutagenesis of the fatty acid desaturase 2 gene family. *Plant Biotechnol. J.* 12(7): 934–940. doi: 10.1111/pbi.12201.

Hayes, B.J., P.J. Bowman, A.J. Chamberlain, and M.E. Goddard. 2009. Invited review: Genomic selection in dairy cattle: Progress and challenges. *J. Dairy Sci.* 92(2): 433–443. doi: 10.3168/jds.2008-1646.

Heffner, E.L., J.L. Jannink, H. Iwata, E. Souza, and M.E. Sorrells. 2011. Genomic selection accuracy for grain quality traits in biparental wheat populations. *Crop Sci.* 51(6): 2597–2606. doi: 10.2135/cropsci2011.05.0253.

Heffner, E.L., A.J. Lorenz, J.L. Jannink, and M.E. Sorrells. 2010. Plant breeding with Genomic selection: Gain per unit time and cost. *Crop Sci.* 50(5): 1681–1690. doi: 10.2135/cropsci2009.11.0662.

Heffner, E.L., M.E. Sorrells, and J.L. Jannink. 2009. Genomic selection for crop improvement. *Crop Sci.* 49(1): 1–12. doi: 10.2135/cropsci2008.08.0512.

Holland, J. 2007. Genetic architecture of complex traits in plants. *Curr. Opin. Plant Biol.* 10(2): 156–161. doi: 10.1016/j.pbi.2007.01.003.

Hoshino, T., Y. Takagi, and T. Anai. 2010. Novel GmFAD2-1b mutant alleles created by reverse genetics induce marked elevation of oleic acid content in soybean seeds in combination with GmFAD2-1a mutant alleles. *Breed. Sci.* 60(4): 419–425. doi: 10.1270/jsbbs.60.419.

Hwang, E.-Y., Q. Song, G. Jia, J.E. Specht, D.L. Hyten, et al. 2014. A genome-wide association study of seed protein and oil content in soybean. *BMC Genomics* 15(1): 1. doi: 10.1186/1471-2164-15-1.

Hymowitz, T., and J.R. Harlan. 1983. Introduction of soybean to North America by Samuel Bowen in 1765. *Econ. Bot.* 37(4): 371–379. doi: 10.1007/BF02904196.

Hyten, D.L., Q. Song, Y. Zhu, I.-Y. Choi, R.L. Nelson, et al. 2006. Impacts of genetic bottlenecks on soybean genome diversity. *Proc. Natl. Acad. Sci. U. S. A.* 103(45): 16666–71. doi: 10.1073/pnas.0604379103.

Imssande, J. 2001. Selection of Soybean Mutants with Increased Concentrations of Seed Methionine and Cysteine. *Crop Sci.* 41(2): 510. doi: 10.2135/cropsci2001.412510x.

Isidro, J., J.L. Jannink, D. Akdemir, J. Poland, N. Heslot, et al. 2015. Training set optimization under population structure in genomic selection. *TAG. Theor. Appl. Genet.* 128(1): 145–158. doi: 10.1007/s00122-014-2418-4.

Jannink, J.L. 2010. Dynamics of long-term genomic selection. *Genet. Sel. Evol.* 42(1): 35. doi: 10.1186/1297-9686-42-35.

Jannink, J.L., A.J. Lorenz, and H. Iwata. 2010. Genomic selection in plant breeding: From theory to practice. *Briefings Funct. Genomics Proteomics* 9(2): 166–177. doi: 10.1093/bfgp/elq001.

Jarquín, D., K. Kocak, L. Posadas, K. Hyma, J. Jedlicka, et al. 2014. Genotyping by sequencing for genomic prediction in a soybean breeding population. *BMC Genomics* 15(1): 740. doi: 10.1186/1471-2164-15-740.

Jarquín, D., J. Specht, and A. Lorenz. 2016. Prospects of Genomic Prediction in the USDA Soybean Germplasm Collection: Historical Data Creates Robust Models for Enhancing Selection of Accessions. *G3*. 6(8): 2329–2341. doi: 10.1534/g3.116.031443.

Johnson, L.A., and D.J. Myers. 1995. Industrial Uses for Soybeans. In: Erickson, D.R., editor, *Practical Handbook of Soybean Processing and Utilization*. AOCS Press. p. 380–427

Kamal-Eldin, A. 2006. Effect of fatty acids and tocopherols on the oxidative stability of vegetable oils. *Eur. J. Lipid Sci. Technol.* 108(12): 1051–1061. doi: 10.1002/ejlt.200600090.

Korte, A., and A. Farlow. 2013. The advantages and limitations of trait analysis with GWAS: A review. *Plant Methods* 9(1): 29. doi: 10.1186/1746-4811-9-29.

Kover, P.X., W. Valdar, J. Trakalo, N. Scarcelli, I.M. Ehrenreich, et al. 2009. A multiparent advanced generation inter-cross to fine-map quantitative traits in *Arabidopsis thaliana* (R. Mauricio, editor). *PLoS Genet.* 5(7): e1000551. doi: 10.1371/journal.pgen.1000551.

Kromhout, D. 1999. Fatty acids, antioxidants, and coronary heart disease from an epidemiological perspective. *Lipids* 34(S1): S27–S31. doi: 10.1007/BF02562225.

La, T., E. Large, E. Taliercio, Q. Song, J.D. Gillman, et al. 2019. Characterization of Select Wild Soybean Accessions in the USDA Germplasm Collection for Seed Composition and Agronomic Traits. *Crop Sci.* 59(0): 1–19. doi: 10.2135/cropsci2017.08.0514.

Leamy, L.J., H. Zhang, C. Li, C.Y. Chen, and B.H. Song. 2017. A genome-wide association study of seed composition traits in wild soybean (*Glycine soja*). *BMC Genomics* 18(1): 18. doi: 10.1186/s12864-016-3397-4.

Lee, J.-D., K.D. Bilyeu, and J.G. Shannon. 2007. Genetics and Breeding for Modified Fatty Acid Profile in Soybean Seed Oil. *J. Crop Sci. Biotechnol.* 10(4): 201–210. <http://www.soystats.com>.

Li, S., Y. Cao, J. He, T. Zhao, and J. Gai. 2017. Detecting the QTL-allele system conferring flowering date in a nested association mapping population of soybean using a novel procedure. *Theor. Appl. Genet.* 130(11): 2297–2314. doi: 10.1007/s00122-017-2960-y.

Li, D., T.W. Pfeiffer, and P.L. Cornelius. 2008. Soybean QTL for Yield and Yield Components Associated with Alleles. *Crop Sci.* 48(2): 571. doi: 10.2135/cropsci2007.06.0361.

Li, Y., J.C. Reif, Y. Ma, H. Hong, Z. Liu, et al. 2015. Targeted association mapping demonstrating the complex molecular genetics of fatty acid formation in soybean. *BMC Genomics* 16(1): 841. doi: 10.1186/s12864-015-2049-4.

Li, X., R. Tian, S. Kamala, H. Du, W. Li, et al. 2018. Identification and verification of pleiotropic QTL controlling multiple amino acid contents in soybean seed. *Euphytica* 214(6): 93. doi: 10.1007/s10681-018-2170-y.

Lin, L., Z. Cunshan, S. Vittayapadung, S. Xiangqian, and D. Mingdong. 2011. Opportunities and challenges for biodiesel fuel. *Appl. Energy* 88(4): 1020–1031. doi: 10.1016/j.apenergy.2010.09.029.

Liu, B., T. Fujita, Z.-H. Yan, S. Sakamoto, D. Xu, et al. 2007. QTL mapping of domestication-related traits in soybean (*Glycine max*). *Ann. Bot.* 100(5): 1027–38. doi: 10.1093/aob/mcm149.

Lorenz, A.J. 2013. Resource Allocation for Maximizing Prediction Accuracy and Genetic Gain of Genomic Selection in Plant Breeding: A Simulation Experiment. *G3*. 3(3): 481–491. doi: 10.1534/g3.112.004911.

Meuwissen, T.H.E., B.J. Hayes, and M.E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157: 1819–1829. doi: 11290733.

Mian, M.A.R., L. McHale, Z. Li, and A.E. Dorrance. 2017. Registration of ‘Highpro1’ Soybean with High Protein and High Yield Developed from a North × South Cross. *J. Plant Regist.* 11(1): 51–54. doi: 10.3198/jpr2016.03.0013crc.

Mounts, T.L., K. Warner, G.R. List, R. Kleiman, W.R. Fehr, et al. 1988. Effect of altered fatty acid composition on soybean oil stability. *J. Am. Oil Chem. Soc.* 65(4): 624–628. doi: 10.1007/BF02540691.

Müller, D., P. Schopp, and A.E. Melchinger. 2018. Selection on Expected Maximum Haploid Breeding Values Can Increase Genetic Gain in Recurrent Genomic Selection. *G3*, 8(4): 1173–1181. doi: 10.1534/g3.118.200091.

Nakaya, A., and S.N. Isobe. 2012. Will genomic selection be a practical method for plant breeding? *Ann. Bot.* 110(6): 1303–16. doi: 10.1093/aob/mcs109.

Nichols, D.M., K.D. Glover, S.R. Carlson, J.E. Specht, and B.W. Diers. 2006. Fine Mapping of a Seed Protein QTL on Soybean Linkage Group I and Its Correlated Effects on Agronomic Traits. *Crop Sci.* 46(2): 834. doi: 10.2135/cropsci2005.05-0168.

Pantalone, V. 2012. Designing Soybeans for 21st Century Markets. In: WILSON, R.F., editor, *Designing Soybeans for 21st Century Markets*. 1st ed. Academic Press and AOCS Press, Urbana, IL. p. 189–218

Panthee, D.R., V.R. Pantalone, C.E. Sams, A.M. Saxton, D.R. West, et al. 2006. Quantitative trait loci controlling sulfur containing amino acids, methionine and cysteine, in soybean seeds. *Theor. Appl. Genet.* 112(3): 546–553. doi: 10.1007/s00122-005-0161-6.

Panthee, D.R., V.R. Pantalone, D.R. West, A.M. Saxton, and C.E. Sams. 2005. Quantitative Trait Loci for Seed Protein and Oil Concentration, and Seed Size in Soybean. *Crop Sci.* 45(5): 2015. doi: 10.2135/cropsci2004.0720.

Pham, A.T., J.D. Lee, J.G. Shannon, and K.D. Bilyeu. 2010. Mutant alleles of FAD2-1A and FAD2-1B combine to produce soybeans with the high oleic acid seed oil trait. *BMC Plant Biol.* 10(1): 195. doi: 10.1186/1471-2229-10-195.

Pham, A.T., J.G. Shannon, and K.D. Bilyeu. 2012. Combinations of mutant FAD2 and FAD3 genes to produce high oleic acid and low linolenic acid soybean oil. *Theor. Appl. Genet.* 125(3): 503–515. doi: 10.1007/s00122-012-1849-z.

Phansak, P., W. Soonsuwon, D.L. Hyten, Q. Song, P.B. Cregan, et al. 2016. Multi-Population Selective Genotyping to Identify Soybean [*Glycine max* (L.) Merr.] Seed Protein and Oil QTLs. *G3.* 6(6): 1635–1648. doi: 10.1534/g3.116.027656.

Price, A.L., N.A. Zaitlen, D. Reich, and N. Patterson. 2010. New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.* 11(7): 459–463. doi: 10.1038/nrg2813.

Priolli, R.H.G., J.B. Campos, N.S. Stabellini, J.B. Pinheiro, and N.A. Vello. 2015. Association mapping of oil content and fatty acid components in soybean. *Euphytica* 203(1): 83–96. doi: 10.1007/s10681-014-1264-4.

Ray, D.K., N.D. Mueller, P.C. West, and J.A. Foley. 2013. Yield Trends Are Insufficient to Double Global Crop Production by 2050 (J.P. Hart, editor). *PLoS One* 8(6): e66428. doi: 10.1371/journal.pone.0066428.

Rincker, K., R. Nelson, J. Specht, D. Sleper, T. Cary, et al. 2014. Genetic Improvement of U.S. Soybean in Maturity Groups II, III, and IV. *Crop Sci.* 54(0): 1419–1432. doi: 10.2135/cropsci2013.10.0665.

Rogers, J., P. Chen, A. Shi, B. Zhang, A. Scaboo, et al. 2015. Agronomic performance and genetic progress of selected historical soybean varieties in the southern USA (R. Singh, editor). *Plant Breed.* 134(1): 85–93. doi: 10.1111/pbr.12222.

Rowntree, S.C., J.J. Suhre, N.H. Weidenbenner, E.W. Wilson, V.M. Davis, et al. 2013. Genetic Gain $\times$ Management Interactions in Soybean: I. Planting Date. *Crop Sci.* 53(3): 1128. doi: 10.2135/cropsci2012.03.0157.

Sebolt, A.M., R.C. Shoemaker, and B.W. Diers. 2000. Analysis of a Quantitative Trait Locus Allele from Wild Soybean That Increases Seed Protein Concentration in Soybean. *Crop Sci.* 40(5): 1438–1444. doi: 10.2135/cropsci2000.4051438x.

Sharma, R., F. Draicchio, H. Bull, P. Herzig, A. Maurer, et al. 2018. Genome-wide association of yield traits in a nested association mapping population of barley reveals new gene diversity for future breeding. *J. Exp. Bot.* 69(16): 3811–3822. doi: 10.1093/jxb/ery178.

Spindel, J.E., H. Begum, D. Akdemir, B. Collard, E. Redoña, et al. 2016. Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. *Heredity.* 116(4): 395–408. doi: 10.1038/hdy.2015.113.

Spindel, J., H. Begum, D. Akdemir, P. Virk, B. Collard, et al. 2015. Genomic Selection and Association Mapping in Rice (*Oryza sativa*): Effect of Trait Genetic Architecture, Training Population Composition, Marker Number and Statistical Model on Accuracy of Rice Genomic Selection in Elite, Tropical Rice Breeding Lines (R. Mauricio, editor). *PLoS Genet.* 11(2): 1–25. doi: 10.1371/journal.pgen.1004982.

Sweeney, D.W., M. Carrero-Colón, and K.A. Hudson. 2017. Characterization of New Allelic Combinations for High-Oleic Soybeans. *Crop Sci.* 57(2): 611. doi: 10.2135/cropsci2015.09.0596.

Takeuchi, H., and M. Sugano. 2017. Industrial Trans Fatty Acid and Serum Cholesterol: The Allowable Dietary Level. *J. Lipids* 2017: 1–10. doi: 10.1155/2017/9751756.

Technow, F., C.D. Messina, L.R. Totir, and M. Cooper. 2015. Integrating crop growth models with whole genome prediction through approximate Bayesian computation (I. De Smet, editor). *PLoS One* 10(6): e0130855. doi: 10.1371/journal.pone.0130855.

USDA - United department of Agriculture. 2018. National agriculture statistics service. <https://quickstats.nass.usda.gov/results> (accessed 23 August 2018).

VanRaden, P.M., J.R. O'Connell, G.R. Wiggans, and K.A. Weigel. 2011. Genomic evaluations with many more genotypes. *Genet. Sel. Evol.* 43(1): 10. doi: 10.1186/1297-9686-43-10.

VanRaden, P.M. 2008. Efficient Methods to Compute Genomic Predictions. *J. Dairy Sci.* 91(11): 4414–4423. doi: 10.3168/jds.2007-0980.

Wang, D., B.W. Diers, P.R. Arelli, and R.C. Shoemaker. 2001. Loci underlying resistance to Race 3 of soybean cyst nematode in *Glycine soja* plant introduction 468916. *Theor. Appl. Genet.* 103(4): 561–566. doi: 10.1007/PL00002910.

Wang, X., G.-L. Jiang, M. Green, R.A. Scott, D.L. Hyten, et al. 2014. Quantitative trait locus analysis of unsaturated fatty acids in a recombinant inbred population of soybean. *Mol. Breed.* 33(2): 281–296. doi: 10.1007/s11032-013-9948-3.

Wang, X., G.-L. Jiang, Q. Song, P.B. Cregan, R.A. Scott, et al. 2015. Quantitative trait locus analysis of seed sulfur-containing amino acids in two recombinant inbred line populations of soybean. *Euphytica* 201(2): 293–305. doi: 10.1007/s10681-014-1223-0.

Wang, S.B., Y.J. Wen, W.L. Ren, Y.L. Ni, J. Zhang, et al. 2016. Mapping small-effect and linked quantitative trait loci for complex traits in backcross or DH populations via a multi-locus GWAS methodology. *Sci. Rep.* 6(1): 29951. doi: 10.1038/srep29951.

Warrington, C. V., H. Abdel-Haleem, D.L. Hyten, P.B. Cregan, J.H. Orf, et al. 2015. QTL for seed protein and amino acids in the Benning $\times$ Danbaekkong soybean population. *Theor. Appl. Genet.* 128(5): 839–850. doi: 10.1007/s00122-015-2474-4.

Wen, Z., Y. Ding, T.-J. Zhao, and J. Gai. 2009. Genetic diversity and peculiarity of annual wild soybean (*G. soja* Sieb. et Zucc.) from various eco-regions in China. *Theor. Appl. Genet.* 119(2): 371–381. doi: 10.1007/s00122-009-1045-y.

Wilson, R.F. 2004. Soybeans: Improvement, Production, and Uses. In: Boerma, H.R. and Specht, J.E., editors. 3rd ed. ASA, CSSA, and SSSA, Madison, WI. p. 621–677

Winter, S.M.J., B.J. Shelp, T.R. Anderson, T.W. Welacky, and I. Rajcan. 2007. QTL associated with horizontal resistance to soybean cyst nematode in *Glycine soja* PI464925B. *Theor. Appl. Genet.* 114(3): 461–472. doi: 10.1007/s00122-006-0446-4.

Wu, G. 2014. Dietary requirements of synthesizable amino acids by animals: a paradigm shift in protein nutrition. *J. Anim. Sci. Biotechnol.* 5(1): 34. doi: 10.1186/2049-1891-5-34.

Xavier, A., B. Hall, A.A. Hearst, K.A. Cherkauer, and K.M. Rainey. 2017. Genetic Architecture of Phenomic-Enabled Canopy Coverage in *Glycine max*. *Genetics* 206(2): 1081–1089. doi: 10.1534/genetics.116.198713.

Xavier, A., D. Jarquin, R. Howard, V. Ramasubramanian, J.E. Specht, et al. 2018. Genome-Wide Analysis of Grain Yield Stability and Environmental Interactions in a Multiparental Soybean Population. *G3*. 8(2): 519–529. doi: 10.1534/g3.117.300300.

Xavier, A., W.M. Muir, and K.M. Rainey. 2016. Assessing Predictive Properties of Genome-Wide Selection in Soybeans. *G3*. 6(8): 2611–6. doi: 10.1534/g3.116.032268.

Yu, J., J.B. Holland, M.D. McMullen, and E.S. Buckler. 2008. Genetic design and statistical power of nested association mapping in maize. *Genetics* 178(1): 539–51. doi: 10.1534/genetics.107.074245.

Zhang, X., P. Pérez-Rodríguez, J. Burgueño, M. Olsen, E. Buckler, et al. 2017a. Rapid Cycling Genomic Selection in a Multiparental Tropical Maize Population. *G3*. 7(7): 2315–2326. doi: 10.1534/g3.117.043141.

Zhang, J., Q. Song, P.B. Cregan, and G.L. Jiang. 2016. Genome-wide association study, genomic prediction and marker-assisted selection for seed weight in soybean (*Glycine max*). *Theor. Appl. Genet.* 129(1): 117–130. doi: 10.1007/s00122-015-2614-x.

Zhang, H., Q. Song, J.D. Griffin, and B.-H. Song. 2017b. Genetic architecture of wild soybean (*Glycine soja*) response to soybean cyst nematode (*Heterodera glycines*). *Mol. Genet. Genomics* 292(6): 1257–1265. doi: 10.1007/s00438-017-1345-x.

Zhang, J., X. Wang, Y. Lu, S.J. Bhusal, Q. Song, et al. 2018. Genome-wide Scan for Seed Composition Provides Insights into Soybean Quality Improvement and the Impacts of Domestication and Breeding. *Mol. Plant* 11(3): 460–472. doi: 10.1016/J.MOLP.2017.12.016.

Zhang, S., Z. Zhang, C. Bales, C. Gu, C. DiFonzo, et al. 2017c. Mapping novel aphid resistance QTL from wild soybean, *Glycine soja* 85-32. *Theor. Appl. Genet.* 130(9): 1941–1952. doi: 10.1007/s00122-017-2935-z.

Zhou, Z., Y. Jiang, Z. Wang, Z. Gou, J. Lyu, et al. 2015. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat. Biotechnol.* 33(4): 408–414. doi: 10.1038/nbt.3096.

Chapter II:

NESTED ASSOCIATION MAPPING OF GRAIN YIELD, PLANT MATURITY, PLANT HEIGHT AND LODGING IN THREE INTERSPECIFIC SOYBEAN POPULATIONS

Abstract

Given the narrow genetic base of domesticated soybean, it is essential to characterize beneficial alleles from novel and exotic germplasm, such as wild soybean to enhance genetic gains for favorable traits in the future. Nested association mapping (NAM) is a population development approach for mapping studies that can compare allele effects from the same QTL between multiple parents. The objective of this study was to investigate marker associations with complex agronomic traits in a NAM panel consisting of 392 recombinant inbred lines derived from three unique bi-parental interspecific soybean populations. We conducted multi-environment field experiment during 2016 and 2017 in eight environments and measured seed yield, plant maturity, plant height, and plant lodging. The nested association mapping and linkage mapping identified three major QTL for plant maturity from *G. soja* presented on chromosomes 6, 11 and 12 which *Glycine soja* alleles were associated with a significant increment in days to maturity. A major QTL for plant height was identified on chromosome 13 and was responsible for increased plant height for lines that carried the *G. soja* allele. A significant QTL for grain yield from *G. soja* was detected on chromosome 17 and showed a positive effect of 166.1 kg ha⁻¹, RIL's with this allele on average yielded 6% more than the *Glycine max* parent. These findings demonstrate the usefulness and

potential to enhance important agronomic traits using alleles from wild soybean germplasm.

Introduction

Soybean [*Glycine max* (L.) Meer] is one of the most prominent oilseed crops worldwide (FAOSTAT, 2018). Over the past three decades, the total global production of soybean has tripled to more than 300 million metric tons in 2016/17 (FAOSTAT, 2018). Genetic improvements in the yield potential of modern soybean varieties and the implementation of improved management practices have enabled an enormous increase in productivity; from 700 kg ha⁻¹ in 1924 to more than 3000 kg ha⁻¹ in 2017 (USDA, 2018). Recent studies in the United States demonstrate genetic gains in yield have ranged from 12 to 27 kg ha⁻¹ year⁻¹ in the last century, or close to 1% year⁻¹ (Rowntree et al. 2013; Rincker et al. 2014; Rogers et al. 2015), while average yield increases worldwide were 1.3% per year (Ray et al. 2013). However, a pedigree analysis indicated that most modern public cultivars could be traced back to 35 founding plant introductions or landraces (Gizlice et al. 1994). Therefore, the lack of genetic diversity in modern public and private soybean breeding programs is of primary concern to ensure continued genetic gains in soybean grain yield well into the future.

Soybean is believed to have been domesticated from the wild progenitor (*Glycine soja* Sieb. and Zucc.) 6,000 to 9,000 years ago in East Asia, including, Korea, Japan, and China (Carter et al. 2004; Doebley et al. 2006). During the domestication process, more than half of the genetic diversity found in *G. soja* was lost, decreasing pairwise diversity (π) from 2.94×10^{-3} in *Glycine soja* to 1.05×10^{-3} in modern cultivars (Hyten et al. 2006;

Zhou et al. 2015b). Currently, it is estimated that ~8,500 unique accessions of *G. soja* and 45,000 accessions of Asian landraces of *G. max* are maintained in germplasm banks around the world (Hyten et al. 2006; Wen et al. 2009). The collection from Germplasm Resources Information Network (GRIN) – USDA contains ~1,100 *G. soja* accessions (Song et al. 2015). However, only a tiny fraction of the publicly available *G. soja* germplasm has been used in commercial breeding aimed at improving elite soybean variety development.

Domesticated soybean has been selected for favorable traits that are useful in an industrialized production system. In contrast, wild soybeans exhibit undesirable agronomic traits such as excessive lodging, late flowering, small black seeds, hard seed coat reducing uniform germination and emergence, lack of plant uniformity of maturity within lines, and excessive pod shattering (Liu et al. 2007). Despite the many unfavorable traits found in *G. soja* germplasm, several important plant introductions, genes, and genomic regions have been reported to be presented within *G. soja* germplasm, including those associated with tolerance to biotic stresses such as soybean aphid (*Aphis glycines*) (Zhang et al. 2017c) and soybean cyst nematode (*Heterodera glycines*) (Wang et al. 2001; Winter et al. 2007; Zhang et al. 2017b), tolerance to abiotic stresses such as exposure to salt and drought (Lee et al. 2009; Chen et al. 2013; Ning et al. 2017; Nisa et al. 2017), and improved seed composition such as increased protein content, improved fatty acid and amino acid profiles (Ha et al. 2014; Yan et al. 2014; Leamy et al. 2017; La et al. 2019).

Although much work has been done recently to understand the implications of domestication as well as the genetic architecture of wild soybean for a variety of different

traits, little research has been done to investigate the potential for wild soybean alleles to be used for improving grain yield. Ertl and Fehr (1985), evaluated two populations derived from *G. soja* x *G. max* backcrosses across two years and four environments and observed no transgressive segregation for grain yield comparing the *G. max* recurrent parent. In another study using a similar approach, Concibido et al. (2003) mapped QTL associated with grain yield from *G. soja* on chromosome 14. The yield QTL significantly increase yield by 9.4% when averaged across all test environments. Also, Li et al. (2008) utilized two backcrossed populations from *G. soja* x *G. max* and identified a QTL associated with grain yield on chromosome 5 (Satt511). Grain yield of the lines carrying the *G. soja* allele was averaged across all environments, and those lines demonstrated a 6.3% yield increase over lines carrying the recurrent parent allele. Thus, there is some evidence that *G. soja* germplasm may harbor useful genes for improving the agronomic performance of modern day soybean cultivars

Quantitative trait loci (QTL) mapping and genome-wide association studies (GWAS) have been used extensively to dissect the genetic architecture of polygenic traits (Holland 2007). One drawback of QTL mapping is that the estimated QTL interval is usually relatively large, due to the extensive linkage disequilibrium (LD), and the low marker density throughout the genome (Xu et al. 2017). QTLs are often population-specific, which does not generate reproducible results in non-related environments, populations, and individuals (Kang et al. 2009; Bernardo 2010). As an alternative, GWAS relies on historical LD to increase the analysis power (Zhu et al. 2008) of detecting common marker associations. GWAS studies have identified many nucleotide variants affecting complex traits in soybeans including seed oil and protein content

(Hwang et al. 2014b; Cao et al. 2017), flowering time and maturity (Zhang et al. 2015; Mao et al. 2017), and soybean cyst nematode (*Heterodera glycine*) (Zhang et al. 2016a). However, this method has low power to detect associations with rare alleles and could be confounded by population structure (Price et al. 2010; Auer and Lettre 2015). Maize geneticists developed a new approach to overcome some pitfalls of conventional GWAS and linkage mapping techniques (Yu et al. 2008), which consists of multi-cross recombinant inbred lines (RILs); each family has a unique parent crossed with a common “hub” parent, (or “founder”). The NAM design combines aspects of two genetic approaches: linkage mapping through genetic recombination and association mapping through historical linkage disequilibrium in the populations, to improve both the resolution and the power for detection of rare alleles (Yu et al. 2008).

The SoyNAM project (<http://www.soybase.org/SoyNAM>) had the objective to understand the genetic architecture of seed yield in soybeans using NAM RILs derived from crosses between IA3023 (hub parent) and 40 different *G. max* adapted cultivars/varieties selected for improving grain yield. Results from SoyNAM are being published. A significant QTL for canopy coverage was identified on chromosome 19, which provides an increase in grain yield of $\sim 50 \text{ kg ha}^{-1}$ (Xavier et al. 2017a). In another study with the same NAM panel, Xavier et al. (2018) identified a genomic region positively associated with grain yield stability across a range of environments on chromosome 18, giving insights about the QTL x environment interaction prevalent in most reported QTL for grain yield. Again using the same soybean NAM panel, Diers et al. (2018) identified 23 significant regions associated with yield and 29 associated with seed weight, demonstrating the complexity of grain yield-related traits. NAM populations

have also been used to dissect complex traits in other important crop species such as rice (*Oryza sativa*) (Fragoso et al., 2017), barley (*Hordeum vulgare*) (Sharma et al. 2018), and wheat (*Triticum aestivum*) (Bajgain et al. 2016). Although the large *G. max* NAM panel has been successful in increasing our understanding of the genetics controlling different traits, there continues to be a need also to investigate the genetics of wild soybean utilizing modern population structures.

The objectives of this study were to identify genomic regions associated with complex agronomic traits in wild soybean germplasm and to identify alleles from *G. soja* accessions that could be used to improve agronomic traits in domesticated soybean. This was accomplished by using a nested association mapping panel with 392 RILs derived from three bi-parental *G. max* x *G. soja* cross combinations assessed across eight environments in Missouri.

Materials and methods

Germplasm development and phenotyping

Our NAM panel consists of 392 F₄-derived RILs, developed from three unique bi-parental cross combinations between the *G. max* hub parent Williams 82 (Bernard and Cremeens 1988) and three *G. soja* plant introductions (PI464890B, PI458536, and PI522226) maintained by the United States Department of Agriculture (USDA) in the Germplasm Resources Information Network – GRIN NPGS (<https://www.ars-grin.gov/>). Cross-pollination of the hub parent Williams 82 and each of the three *G. soja* plant introductions were conducted in Champaign-Urbana, IL during the summer of 2010, and the true F₁ hybrid plants were grown at the same environment in 2011 and bulk

harvested. Approximately 10,000 F₂ plants were grown at the Bradford Research Center in Columbia, MO during the summer of 2012 for each of the three populations (30,000 total plants), and F₂ plants were bulk harvested. In 2013, approximately 100,000 F₃ plants were grown at the Bradford Research Center in Columbia, MO for each of the three populations (300,000 plants total), and approximately 500 single plants were harvested separately from each F₃ population (1500 total plants), the F₃ plants were selected based on agronomic traits, such as maturity, lodging, and shattering. The F_{3:4} plants rows were grown at the Bay Farm Research Facility in Columbia, MO during 2014, and a single plant was harvested separately from each row. The F_{4:5} plant rows were grown at the Bay Farm Research Facility in Columbia, MO during 2015, and each plant row was bulk harvested for development of the F₄-derived RILs. Only lines with sufficient seed for broad testing were carried forward, and RILs were randomly chosen to represent an appropriate equal number from each of the three bi-parental populations to reduce the NAM panel size to a manageable field experiment.

During the growing seasons of 2016 and 2017, RILs were planted in four-row research plots with a row spacing of 76 cm, a row length of 3.6 m, and a 1.2 m fallow alley. The environments included the Hundley-Whaley Research Center in Albany, MO (40°14'53"N 94°19'51"W), the Bay Farm Research Facility in Columbia, MO (38°57'50"N 92°19'42"W), the Greenley Research Center in Novelty, MO (39°57'28.23"N 92°10'40.45"W) and the Graves-Chapple Research Center in Rock Port, MO (40° 16'18.09" N 95°28' 26.38"W). The field research plots were arranged in an augmented incomplete block experimental design with one replication in 2016 and two replications in 2017, with two check cultivars per block in both years. Genotypes were

randomly assigned to a block and replication, and blocks were randomly ordered in the field across all environments. The experiment was divided into two field blocks at each environment for group MG III and MG IV RILs to accommodate combine harvesting across all environments.

Grain yield was measured by harvesting the two center rows of each four-row plot with an Almaco R1 or an Almaco SPC-40 plot combine (Nevada, Iowa). The total seed weight and seed moisture were measured in real time on the combine, and the final grain yield was calculated as kg ha^{-1} on a 13% moisture basis. Plant maturity was recorded as the day when 95% of the pods in the center two rows of each plot were mature; R8 stage (Fehr et al., 1971), indicated by the development of brown pod color and considering September 1st as day one. Lodging was rated for the center two rows of each plot at R8 stage according to the following score: 1 – all plants erect; 2 – 5% to 25% of the plants prostrate; 3 - 25% to 50% of the plants prostrate; 4 – 50% to 80% of the plants prostrate; 5 – all plants prostrate. Plant height was recorded as the average distance (cm) between the apical meristem and the soil surface of plants in the center two rows of each plot.

Genotyping and SNP analysis

Genomic DNA was extracted from young trifoliate leaf tissue of 10 random plants for each RIL and the respective parents during the V4/V5 growth stage. A Qiagen Dneasy Plant 96 kit (QIAGEN, Valencia, CA) was used for high-quality DNA extraction, and the protocol given for the kit was followed. DNA samples were submitted to the Soybean Genomics and Improvement Laboratory, USDA-ARS, for genotyping using the Illumina Infinium BARCSoySNP6K BeadChip (Song et al. 2013). Genotypes were called using the software Genome Studio (Illumina, San Diego, California USA). SNPs

with less than 90% call rate were excluded, and SNPs were imputed using the LD-KNNi method implemented in version 5 of Tassel software (Bradbury et al. 2007). A total of 5,786 SNPs remained for subsequent quality control, association mapping, and linkage analysis. The R package 'ParentOffspring' (Abdel-Haleem et al. 2013) with a threshold of 80% of similarity between parents, and a principal component analysis using version 5 of the TASSEL software (Bradbury et al. 2007) was used to eliminate RILs that exhibited genotypes inconsistent with expected genetic similarity to the respective parents.

Only the filtered SNP marker data were used to measure the linkage disequilibrium (LD) in the NAM panel using the sliding window method implemented in version 5 of TASSEL software. The LD SNPs were estimated using the pairwise squared correlation (r^2), then they were plotted against SNP pair distances to obtain the rate of the LD decay in the heterochromatin and euchromatin region (Song et al. 2013). We calculated the fixation index (F_{ST}) among the three populations in the NAM panel using the unbiased estimator (Weir and Cockerham 1984) implemented in the NAM package (Xavier et al. 2015).

Nested association mapping analysis

Best linear unbiased estimators (BLUE) were calculated to reduce non-genetic and error variance associated with environment and field variation. BLUE values were then used in the nested association mapping according to the model adopted by Jarquín et al. (2014): $y_{ijkl} = \mu + g_i + cb_l + l_j + gl_{ij} + r_{k(j)} + b_{l(k)} + e_{ijk}$, where μ is the trait mean, g_i is the genetic effect of the genotypes, cb_k is the interaction effect between the checks and incomplete blocks, l_j is the effect of environments, gl_{ij} is the interaction

effect between the genotypes and environments, $r_{k(j)}$ is the effect of the replications nested in environment, $b_{l(k)}$ is the incomplete block effect nested within replicate, and e_{ijkl} is the residual. Broad-sense heritability (h^2) on an entry-mean basis was calculated as follows: $h^2 = \sigma_g^2 / (\sigma_g^2 + \sigma_{ge}^2/l + \sigma_e^2/rl)$, where σ_g^2 , σ_{ge}^2 and σ_e^2 are the genetic variance, genotype x environment variance and error variance; l is the number of environments and r is the number of replications (Fehr 1991).

The nested association mapping was conducted using the package NAM (Xavier et al. 2015) developed for analysis of SoyNAM project. The mixed linear model designed for multiple parent intercross populations was used for the SNP and haplotype-based association (Wei and Xu 2016). The mixed linear model is described as follows: $y = \mu + X\alpha + g + cov + e$, where μ is the intercept, X is the allele matrix from SNP/haplotype data and family information, α is the SNP/haplotypes effects, g is the population structure effect, cov is the covariate, and e is the residual effect. The R8 date was used as a covariate in the NAM analysis to account for maturity effects on grain yield, plant height and lodging. A false discovery rate (FDR) threshold at $\alpha \leq 0.05$ level was used to declare SNP significant in the association mapping.

Linkage mapping analysis

We used the consensus genetic linkage map constructed by Song et al. (2016) based on 21,478 SNPs mapped in the Williams 82 x PI479752 (*G. soja*) population to define SNP positions. QTL mapping was performed using the composite interval mapping (CIM) functionality qtl package in R (Broman and Sen 2009). To fit multiple QTL models, we used the function fitqtl which drops one QTL at a time. The likelihood

of determination (LOD) for defining QTL significance was determined by a 1000 permutation test analysis at $\alpha \leq 0.05$ significant level and estimated confidence intervals of the environment of QTL via used 1.5-LOD support intervals.

We searched for candidate genes, published genes, and published QTLs using the soybean reference genome available at SoyBase (www.soybase.org) in the Wm82.a2.v1 assembly. The SNPs and QTL intervals that were highly associated with traits were considered as a potential region for candidate genes. SoyBase was also used to search for QTLs previously found to affect the traits to date.

Results

Phenotypic variation

The analysis of variance (ANOVA) for all traits evaluated in this study is presented in Table 2-1. The results indicate significant genetic differences among RILs for all measured traits, and both environmental effects and RIL x environment interaction were significant ($p < 0.01$) for all the traits (Table 2-1). The entry-mean based estimates of heritability (h^2) were 0.51-0.64 for grain yield (GY), 0.87-0.93 for plant maturity (PM), 0.82-0.93 for plant height (PH), and 0.85-0.91 for lodging score (LOD).

Continuous and normal distribution were observed for all phenotypic traits (Table 2-1 and Figure 2-1) measured in this study. Transgressive segregation was observed in both directions when comparing the phenotypic values of the hub-parent (Williams 82) and estimated values for each *G. soja* parent suggesting the effect of multiple genes controlling the traits and appropriate distribution of alleles among hub-parent and *G. soja* parents. The normal frequency distributions for each trait also suggest a random

population structure (Figure 2-2B) to move forward with subsequent mapping efforts to identify marker associations for major QTL.

Linkage Disequilibrium and Population Structure

Linkage disequilibrium (LD) decay varied among euchromatin and heterochromatin regions (Figure 2-2A). The LD decayed to half its maximum r^2 value (0.44) at an average 619.5.1 kb in euchromatin and 892.6kb in heterochromatin regions. Since we estimated LD in experimental bi-parental populations we developed, a greater extent of LD is expected when comparing expected LD to natural soybean populations (Zhou et al. 2015b) due to the limited amount of recombination and selection in the experimental populations (Hyten et al. 2007). Principal component analysis (PCA) was used to evaluate the variation in genotypic data of the NAM population (Figure 2-2B). The PCA could not distinguish clearly the sub-population structures likely due to the half-sib families relationship among the three subpopulations, yet PC1 explained 10.1% of the variation, while PC2 explained 6.3% of the genotypic variation for makers.

More than 70% of the SNPs had an estimated F_{ST} value smaller than 0.1 (Figure 2-2C). Just 0.02% of the SNPs had an F_{ST} value larger than the 99th percentile (red line). To distinguish the SNPs under selection and potential genetic drift in the NAM population, we utilized the smoothing kernel (blue lines) procedure implemented in the NAM package (Flori et al. 2009; Xavier et al. 2015). The results revealed significant F_{ST} values on chromosomes 3, 5, 6, 9, 11 and 20. The SNPs ss715594391, ss715593858, and ss715594468 on chromosome 6 had the highest F_{ST} values 0.67, 0.66, 0.63, respectively. ss715593858 is ~400kb from the region of *E1* (*Glyma.06g207800*) (Xia et al. 2012),

well-known photoperiod response/plant maturity gene. This is likely a result of fixing the maturity date during the inbreeding process to evaluate the RILs in central and northern Missouri.

Nested association mapping

A nested association mapping (NAM) analysis was employed to conduct the association analysis between SNP markers and phenotypes. Four SNPs on Chr 16 were significantly associated with grain yield (GY) in 2017 across the average of four environments (Figure 2-3), showing allelic effects ranging from -116.2 to 59.4 kg ha⁻¹. This marker and associated locus has been identified to be in the genomic region for the gene regulating pod shattering, *Pdhl* (Funatsuki et al. 2014). Hence, the relatively large negative allelic effects on grain yield.

Forty-nine SNPs were significantly associated with plant maturity (PM) on chromosomes 6, 11 and 12 in 2016 and 2017 (Figure 2-3). The regions were consistently associated with the trait across environments, and the QTL on Chr 6 and 12 were observed in all the eight environments, Chr 11 QTL was found in Albany 2016 (ALB2016), Columbia 2016 (CLM2016), and Novelty (NOV2016) (Figure 2-4). The allelic effects for PM ranged from -2.02 to 5.73 days depending on the SNP associated and population (Table 2-2), with a higher frequency of estimated positive PM alleles, was evident from the *G. soja* founder parents. The three SNPs most associated with PM on Chr 6 were ss715594029, ss715593853, and ss715594119. These SNPs were co-located in the same region with the known *E1* photoperiod response/plant maturity gene (Xia et al. 2012). Also, the SNPs associated with PM on Chr 11 and 12 were reported to be

associated with flowering/maturity periods (Zhang et al. 2015; Fang et al. 2017; Mao et al. 2017).

The three QTL identified by ss715593853 on Chr 6, ss715608829 on Chr 11 and ss715613171 on Chr 12, and explained 36%, 16% and 18% of the genetic variation for PM, respectively (Figure 2-5A, B, C). RILs carrying the genotype CC (*G. soja*) at ss715593853 exhibited a positive increment of 4.3 days in PM compared to RILs with the TT (*G. max*) genotype (Figure 2-5A). A similar trend was observed for ss715608829, and ss715613171, where lines carrying the *G. soja* allele were 3.6 and 3.9 days later PM than lines carrying *G. max* allele, respectively (Figure 2-5B, C). The allelic effect for the QTLs on Chr 6, 11 and 12 varied between *G. soja* parent, with a higher frequency of estimated alleles that effected PM from *G. soja*.

Seven SNPs were associated with plant height (PH) across four environments in 2016 and 2017 (Figure 2-3), included one SNP on Chr 10 and six on Chr 13. The QTL on Chr 13 was identified in five of the eight environments across years (Figure 2-4), with an allelic effect ranging from -2.86 to 13.07 cm (Table 2-2). The most significant SNP for the trait was ss715616047 on Chr 13, as it explained 12% of the genetic variation and exhibited an increment in PH of 19.6 cm when comparing the lines carrying the AA (*G. soja*) genotype to lines carrying the CC (*G. max*) genotype. The region of the QTL on Chr 13 was coincident with qPH13.1, a *G. soja* confirmed QTL, which in previous studies significant increased total plant height (Zhang et al. 2018b).

A region on Chr 8 (ss715602804 and ss715602662) was significantly associated with lodging (LOD) in 2016. Four SNPs in 2017 were associated with LOD, ss715579524 on Chr 1, ss715602804 on Chr 8, ss715630642 on Chr 18 and ss715633186

on Chr 19 (Table 2-2). The QTL region associated with ss715602804 was most associated with PM across environments and identified in Novelty 2016 and Columbia 2017. The QTL on Chr 8 was responsible for explaining 17% of the phenotypic variation, and lines carrying the TT allele (*G. soja*) exhibited a significant increment of 0.46 in LOD score when compared to the lines with the CC allele (*G. max*) (Figure 2-5E).

Linkage Mapping

Twelve QTLs in three bi-parental populations were significantly associated with GY, PM, PH, and LOD on eleven chromosomes in 2016 and 2017 (Figure 2-6). Three QTL were detected for GY on Chr 17, 18 and 19; four QTL with PM on Chr 6, 11, 12 and 20; three QTL with PH on Chr 1, 2, and 13, and one QTL associated with LOD on Chr 9.

The QTL associated with GY on chromosomes 17, 18 and 20 explained 24.4%, 21.2% and 17.5% of the total phenotypic variance and additive effects ranged from -297.2 to 166.1 kg ha⁻¹ (Table 2-3). Only the qGY-17 had a positive impact on GY from *G. soja*. The QTL was identified across two environments in 2016, Columbia and Albany. Lines carrying the *G. soja* allele (BB) averaged 197.1 kg ha⁻¹ more or had a 6% increase in grain yield when compared to the lines carrying the *G. max allele* (AA) across environments.

Four QTL were detected for PM in multiple environments in 2016 and 2017 (Figure 2-6 and Figure 2-7). A major effect QTL on Chr 6 was detected in all of the three populations, and the phenotypic variance among individuals ranged from 28.3 to 52.6%. The additive effect of qPM-6-1, qPM-6-2, and qPM-6-3 ranged from 2.34 to 5.5 days

(Table 2-3). Three minor QTLs were detected in specific populations; qPM-11 on Chr 11 in Williams 82 x PI464890B (Pop1), qPM-12 on Chr 12 in Williams 82 x PI522226 (Pop3) and qPM-20 on Chr 20 in Pop1. The confidence interval for the QTL qPM-6-1, qPM-6-2, qPM-6-3, qPM-11, and qPM-12 overlapped with genomic regions previously identified with the NAM analysis (Table 2-2). The QTL qPM-20 was responsible for 12.5% of the phenotypic variation with a negative effect of -2.87 days on PM, and qPM-20 was located close to the *E4* gene (Liu et al. 2008). Also, an epistatic interaction was significant between qPM-6-1 x qPM-20 in 2017, accounting for more than 5.6% of the phenotypic variation for PM.

Three QTL were identified for plant height (PH) in multiple environments on Chr 1, 2, and 13 (Figure 2-6). qPH-1 on Chr 1 was observed in the population Williams 82 x PI458536 (Pop2), qPH-2 on chromosome 2 was identified in the population Williams 82 x PI464890B (Pop2), qPH-1 accounted for 18.7% of the variation, while qPH-2 accounted for 15.2% to 20.1% of the phenotypic variation (Table 2-3). Both qPH-1 and qPH-2 showed a negative effect on PH, from -8.9 to -14.2 cm. qPH-13 was identified in the population Williams 82 x PI522226 (Pop3), and the phenotypic variance explained was 23.3%, with a positive effect of 11.87 cm. The qPH-13 interval overlapped with the region previously identified on chromosome 13 in the NAM analysis (Table 2-2), which was associated with greater plant height (Figure 2-5D).

One QTL on chromosome 9 was associated with lodging score (LOD) in the population Williams 82 x PI458536 (Pop2) (Figure 2-6 and Table 2-3). The phenotypic variation explained for the qLOD-9 was 14.3%, and the additive effect was 0.24 on the 1-

5 scale. The qLOD-9 interval region overlaps with Lodging 5-9 QTL reported by Lee et al. (1996).

Discussion

In this study, we evaluated a nested association mapping panel by using both linkage mapping in three bi-parental populations, as well as association mapping with the entire panel of RILs. This approach has been successful in understanding the genetic architecture for several soybean traits, as well as other crops (Yu et al. 2008; Fragoso et al. 2017; Li et al. 2017; Diers; et al. 2018; Xavier et al. 2018). We identified in the NAM mapping the shattering gene *Pdh1* (Funatsuki et al. 2014) associated with the marker ss715624484 on Chromosome 16. *Pdh1* increases pod dehiscence by promoting the torsion of the pod (Funatsuki et al. 2014). Breeding for shattering resistance (*pdh1*) allowed soybean production to expand to more arid conditions (Bandillo et al. 2017) and the selection for *pdh1* soybean varieties was crucial for expansion and the success of industrialized soybean system (Funatsuki et al. 2014).

Four QTL associated with PM in the NAM panel were identified. The region associated with PM on Chr 6 (*E1* gene - ss715593853/qPM-6-1, 6-2, 6-3) (Xia et al. 2012) was the most consistent QTL identified across environments and showed the largest allelic effects on PM (Figures 2-4 and 2-7; Table 2-2). When comparing the allelic effects of ss715593853 the difference between *G. soja* allele (*E1*) and *G. max* (*e1-as*) was 4.3 days in PM. The minor allele frequency (MAF) for the significant SNPs was relatively high, ranging from 0.15 to 0.50. The NAM design can increase MAF of rare alleles in the population improving the power to detect associations. Also, the *E1* gene is

under strong selection in the NAM population, revealed by the highest F_{ST} values by SNPs surrounding the gene locus (Figure 2-2C). In 1971, *E1/e1* and *E2/e2* two major genes that affect time to flowering and maturity were identified (Bernard 1971). Both dominant genes *E1* and *E2* caused delays on flowering, although with a stronger effect from *E1* (Bernard 1971). Later, in 2005 *E1* was fine mapped with two tightly linked DNA markers, Satt365 (0.1 cM) and GM169 (0.4 cM) (Yamanaka et al. 2005). Xia et al. (2012), through positional cloning located the position of *E1* (*Glyma.06g207800*). *E1* is part of the phytochrome A signaling pathway and down-regulated *GmFT2a* and *GmFT5a*, orthologs of *Arabidopsis* *FLOWERING LOCUS T* responsible for early flowering (Kong et al. 2010). Xia et al. (2012), characterize *E1* as functional, and two non-functional alleles *e1-fs* and *e1-nl*, and one not fully functional *e1-as*.

These results impact breeding strategies for maturity and soybean adaptation in different regions in the US. Langewisch et al. (2017) evaluated a series of combinations of *E1* and *e1-as* from maturity group (MG) 000 to X and showed a higher frequency of *e1-as* allele in MG IV and below and a higher allele frequency of *E1* in MG V and above. *E1* promotes late flowering may be used in southern US and *e1-as* promotes earlier flowering may be used in the northern US in MG 0 to IV (Wolfgang and An 2017; Langewisch et al. 2017).

The QTL on Chr 12 (ss715613171) was identified in seven of eight environments, from both mapping methods. The allelic effects were higher in the Pop3, ranging from 0.62 to 1.30 days. In the linkage analysis, qPM-12 was significant for Pop3 in 2016 and 2017 and explained ~22% of the phenotypic variation in the population for the trait. SNPs associated with the first flower were reported in the same region by Mao et al.

(2017) and Fang et al. (2017). The QTL on Chr 11 (ss715608829/qPM-11) was detected in three environments in the NAM analysis and linkage mapping analysis. The largest allelic effects were observed on Pop1 in 2016 (1.14-1.35 days – Table 2). Li et al. (2017) reported a QTL for days to flowering (qDTF-11-4) using a NAM panel in the same region with an additive effect of 1.87 days. They reported in qDTF-11-4 interval *Glyma11g15580* as a possible candidate gene, which is involved in long-day photoperiodism and flowering processes.

A QTL on chromosome 20 (qPM-20) was significantly associated with PM in the linkage mapping analysis for Pop2 with a negative effect of -2.9 days, explaining 12.5% of the PM variation in that specific population. QTLs are often population-specific which makes the results less reproducible in diverse genetic background (Bernardo 2010). The NAM design can compare allelic effects from the same SNP from multiple parents, enabling the distinction of allelic effects for each *G. soja* plant introduction relative to the Williams 82 hub parent allele(s) (Xavier et al., 2015). Pinpointing candidate genes for the QTLs identified on Chr 11, 12 and 20 are uncertain due to the extensive LD presented in the NAM population and the wide interval obtained in the linkage mapping analysis. The qPM-20 region is ~3.7Mb apart from the gene *E4*. The gene encodes the phytochrome protein GmphyA2 (*Glyma.20g090000*) (Liu et al. 2008), and recessive alleles of *E4* provide earlier flowering (Xu et al. 2013), which is the phenotype observed in RIL which inherited the *G. soja* form of this QTL.

Wild soybeans are well known for excessive plant growth and extreme lodging, or prostrate growth habit. Until now, more than 239 and 87 QTLs have been reported to be associated with plant height (PH) and lodging (LOD) in soybean, respectively

(<http://www.soybase.org>, “SoyBase browser”, verified 01/20/2019). We identified a significant QTL for PH on Chr 13 in both mapping methods (ss715616047/qPH-13). The lines carrying the *G. soja* allele showed a significant increase in PH compared to the lines carrying the *G. max* allele (Figure 2-5D). Our findings were supported by Zhang et al. (2018b) since they fine mapped a QTL from *G. soja* in the same region of ss715616047/qPH-13, which exhibited a greater PH for the lines that carried the wild allele. RT-PCR showed that *Glyma.13g249400* demonstrated higher expression in genotypes that expressed higher PH; also sequence data identified a difference in six amino acids from *Glyma.13g249400* between parents (Zhang et al. 2018b).

Two SNPs were associated with LOD on Chr 8. The SNP ss715602804 showed a significant increment of 0.46 units in the LOD score in the lines that carried the *G. soja* allele (Figure 3E). Another QTL was identified in the linkage analysis on Chr 9 (Figure 4D and Table 4), qLOD-9 expressed a positive allelic effect of 0.24 units in the LOD score. Kabelka et al. (2006) identified significant gains up to 0.4 units in lodging score in the lines that carried the SCN resistant gene from *G. soja*. Introgression of wild soybean alleles in currently elite germplasm may result in linkage drag of unfavorable traits, such as excessive plant growth and lodging (Zhang et al. 2017a). However, this issue could be fixed with backcrossing with more adaptable germplasm and more cycles of selection (Ertl and Fehr 1985; Concibido et al. 2003).

The qGY-17 exhibited an increment in GY of 6% across environments, showing a benefit from *G. soja* allele compared to the *G. max* allele, although the QTL was only significant in Pop3 (Williams 82 x PI522226). Similar results were found by Concibido et al. (2003) using backcross populations from *G. soja* x *G. max*, where it was identified

positive effects of *G. soja* alleles in two of six genetic backgrounds. Also, a QTL (qYIE-D2-1) in the same region was detected by Wang et al. (2014), and SNP markers were identified for pod number and seed weight (Hu et al. 2014; Yan et al. 2017). Further evaluation using different genetic backgrounds and environments will be necessary to better understand the impact of the qGY-17. Assessment of large yield testing in *G. soja* x *G. max* populations is challenging, due to the presence of deleterious agronomic traits such as shattering, lodging, viny growth and difficulties related to mechanical harvest of plots.

An impressive number of *G. soja* x *G. max* RILs yielded significantly more than the *G. max* parent (Williams 82) (Table 2-4). Among those SA15-109, SA15-66476, SA15-63423, and SA15-66563 yielded between 5 to 17% more than the Williams 82 depending on the environment. The high yielding line SA15-109, from Williams 82 x PI 458536, yielded 400 kg ha⁻¹ more and matured one day later than Williams 82 across four environments in 2017. In Rock Port 2017, this line yielded 725 kg ha⁻¹ more and matured the same day as Williams 82. Another high yielding line SA15-66476, from Williams 82 x PI 522226, yielded 583.4 kg ha⁻¹ more and matured one day earlier than Williams 82. Ertl and Fehr (1985) and Wang et al. (2004) were unsuccessful in increasing yield potential in interspecific crosses using *G. soja*. On the other hand, Concibido et al. (2003) and Li et al. (2008) identified positive grain yield QTLs from *G. soja* crosses. A major outcome of this study was positive transgressive segregates for grain yield using *Glycine soja* in the pedigree, and the best line significantly outperformed the *Glycine max* parent (Williams 82) by 400 kg ha⁻¹ across environments. Similar results are expected and often

achieved using *G. max* x *G. max* crosses, however, yield gain using single crosses of *G. max* x *G. soja* is surprising.

In summary, 16 QTLs from *Glycine soja* associated with increased grain yield, extended plant maturity, increased plant height, and increased lodging were identified by nested association mapping and linkage mapping. A novel QTL for grain yield on chromosome 17 from *G. soja* was detected in the linkage analysis, and lines carrying the wild soybean allele was associated with an increase of 6% in grain yield. There have been very few reports in the long history of soybean genetics of QTL associated with increasing grain yield derived from interspecific crosses with *G. soja*, and our results suggest further investigation is needed to understand the potential use of this or others QTL. Novel regions associated with plant maturity were identified on Chr 11, and 12 and the photoperiod response/plant maturity gene *El* was confirmed on Chr 6. A major QTL associated with PH was identified in Chr 13, exhibiting increased plant height. The use of wild soybean germplasm for commercial breeding purposes is still quite rare for breeders, mainly because of the difficulties to work with undomesticated lines. These results indicate breeding with *G. soja* germplasm can successfully introduce new alleles with the potential to add favorable alleles to the current elite soybean gene pool.

References

- Abdel-Haleem, H., P. Ji, H.R. Boerma, and Z. Li. 2013. An R package for SNP marker-based parent-offspring tests. *Plant Meth.* 9(1): 44. doi: 10.1186/1746-4811-9-44.
- Auer, P.L., and G. Lettre. 2015. Rare variant association studies: considerations, challenges and opportunities. *Genome Med.* 7(1): 16. doi: 10.1186/s13073-015-0138-2.
- Bajgain, P., M.N. Rouse, T.J. Tsilo, G.K. Macharia, S. Bhavani, et al. 2016. Nested Association Mapping of Stem Rust Resistance in Wheat Using Genotyping by Sequencing. *PLoS One* 11(5): e0155760. doi: 10.1371/journal.pone.0155760.
- Bandillo, N.B., J.E. Anderson, M.B. Kantar, R.M. Stupar, J.E. Specht, et al. 2017. Dissecting the Genetic Basis of Local Adaptation in Soybean. *Sci. Rep.* 7(1): 17195. doi: 10.1038/s41598-017-17342-w.
- Bernard, R.L. 1971. Two Major Genes for Time of Flowering and Maturity in Soybeans1. *Crop Sci.* 11(2): 242. doi: 10.2135/cropsci1971.0011183X001100020022x.
- Bernard, R.L., and C.R. Cremeens. 1988. Registration of ‘Williams 82’ Soybean. *Crop Sci.* 28(6): 1027–1028. doi: 10.2135/cropsci1988.0011183X002800060049x.
- Bernardo, R. 2010. *Breeding for Quantitative Traits in Plants*. Second. Stemma Press, Woodbury.
- Bradbury, P.J., Z. Zhang, D.E. Kroon, T.M. Casstevens, Y. Ramdoss, et al. 2007. TASSEL: software for association mapping of complex traits in diverse samples.
- Broman, K.W., and S. Sen. 2009. *A Guide to QTL Mapping with R/qtl*. Springer, New York, NY.
- Cao, Y., S. Li, Z. Wang, F. Chang, J. Kong, et al. 2017. Identification of Major Quantitative Trait Loci for Seed Oil Content in Soybeans by Combining Linkage and

Genome-Wide Association Mapping. *Front. Plant Sci.* 8(1222): 1–10. doi: 10.3389/fpls.2017.01222.

Carter, T.E., T. Hymowitz, and R.L. Nelson. 2004. Biogeography, Local Adaptation, Vavilov, and Genetic Diversity in Soybean. *Biological Resources and Migration*. Springer Berlin Heidelberg, Berlin, Heidelberg. p. 47–59

Chen, P., K. Yan, H. Shao, and S. Zhao. 2013. Physiological Mechanisms for High Salt Tolerance in Wild Soybean (*Glycine soja*) from Yellow River Delta, China: Photosynthesis, Osmotic Regulation, Ion Flux and antioxidant Capacity (P.C.G. Ferreira, editor). *PLoS One* 8(12): e83227. doi: 10.1371/journal.pone.0083227.

Concibido, V., B. La Vallee, P. McIaird, N. Pineda, J. Meyer, et al. 2003. Introgression of a quantitative trait locus for yield from *Glycine soja* into commercial soybean cultivars. *Theor. Appl. Genet.* 106(4): 575–582. doi: 10.1007/s00122-002-1071-5.

Diers, B.W., J. Specht, K.M. Rainey, P. Cregan, Q. Song, et al. 2018. Genetic Architecture of Soybean Yield and Agronomic Traits. *G3*. 8(10): 3367–3375. doi: 10.1534/g3.118.200332.

Doebley, J.F., B.S. Gaut, and B.D. Smith. 2006. The molecular genetics of crop domestication. *Cell* 127(7): 1309–21. doi: 10.1016/j.cell.2006.12.006.

Ertl, D.S., and W.R. Fehr. 1985. Agronomic Performance of Soybean Genotypes from *Glycine max* x *Glycine soja* Crosses1. *Crop Sci.* 25(4): 589–592. doi: 10.2135/cropsci1985.0011183X002500040003x.

Fang, C., Y. Ma, S. Wu, Z. Liu, Z. Wang, et al. 2017. Genome-wide association studies dissect the genetic networks underlying agronomical traits in soybean. *Genome Biol.* 18(1): 161. doi: 10.1186/s13059-017-1289-9.

Fehr, W.R. 1991. Principles of cultivar development. Vol. 1. Theory and technique. Macmillan, New York.

Flori, L., S. Fritz, F. Jaffrézic, M. Boussaha, I. Gut, et al. 2009. The genome response to artificial selection: A case study in dairy cattle. *PLoS One* 4(8): e6595. doi: 10.1371/journal.pone.0006595.

Fragoso, C.A., M. Moreno, Z. Wang, C. Heffelfinger, L.J. Arbelaez, et al. 2017. Genetic Architecture of a Rice Nested Association Mapping Population. *G3*. 7(6): 1913–1926. doi: 10.1534/g3.117.041608.

Funatsuki, H., M. Suzuki, A. Hirose, H. Inaba, T. Yamada, et al. 2014. Molecular basis of a shattering resistance boosting global dissemination of soybean. *Proc. Natl. Acad. Sci. U. S. A.* 111(50): 17797–802. doi: 10.1073/pnas.1417282111.

Gizlice, Z., T.E. Carter, and J.W. Burton. 1994. Genetic Base for North American Public Soybean Cultivars Released between 1947 and 1988. *Crop Sci.* 34(5): 1143. doi: 10.2135/cropsci1994.0011183X003400050001x.

Ha, B.-K., H.-J. Kim, V. Velusamy, T.D. Vuong, H.T. Nguyen, et al. 2014. Identification of quantitative trait loci controlling linolenic acid concentration in PI483463 (Glycine soja). *Theor. Appl. Genet.* 127(7): 1501–1512. doi: 10.1007/s00122-014-2314-y.

Holland, J. 2007. Genetic architecture of complex traits in plants. *Curr. Opin. Plant Biol.* 10(2): 156–161. doi: 10.1016/j.pbi.2007.01.003.

Hu, Z., G. Kan, G. Zhang, D. Zhang, D. Hao, et al. 2014. Genetic diversity analysis using simple sequence repeat markers in soybean. *Plant Genet. Resour.* 12(S1): S87–S90. doi: 10.1017/S1479262114000331.

Hwang, E.-Y., Q. Song, G. Jia, J.E. Specht, D.L. Hyten, et al. 2014. A genome-wide association study of seed protein and oil content in soybean. *BMC Genomics* 15(1): 1–12. doi: 10.1186/1471-2164-15-1.

Hyten, D.L., I.Y. Choi, Q. Song, R.C. Shoemaker, R.L. Nelson, et al. 2007. Highly variable patterns of linkage disequilibrium in multiple soybean populations. *Genetics* 175(4): 1937–1944. doi: 10.1534/genetics.106.069740.

Hyten, D.L., Q. Song, Y. Zhu, I.-Y. Choi, R.L. Nelson, et al. 2006. Impacts of genetic bottlenecks on soybean genome diversity. *Proc. Natl. Acad. Sci. U. S. A.* 103(45): 16666–71. doi: 10.1073/pnas.0604379103.

Jarquín, D., K. Kocak, L. Posadas, K. Hyma, J. Jedlicka, et al. 2014. Genotyping by sequencing for genomic prediction in a soybean breeding population. *BMC Genomics* 15(1): 740. doi: 10.1186/1471-2164-15-740.

Kabelka, E.A., S.R. Carlson, and B.W. Diers. 2006. Glycine soja PI 468916 SCN resistance loci's associated effects on soybean seed yield and other agronomic traits. *Crop Sci.* 46(2): 622–629. doi: 10.2135/cropsci2005.06-0131.

Kang, S.-T., M. Kwak, H.-K. Kim, M.-G. Choung, W.-Y. Han, et al. 2009. Population-specific QTLs and their different epistatic interactions for pod dehiscence in soybean [*Glycine max* (L.) Merr.]. *Euphytica* 166(1): 15–24. doi: 10.1007/s10681-008-9810-6.

Kong, F., B. Liu, Z. Xia, S. Sato, B.M. Kim, et al. 2010. Two Coordinately Regulated Homologs of FLOWERING LOCUS T Are Involved in the Control of Photoperiodic Flowering in Soybean. *PLANT Physiol.* 154(3): 1220–1231. doi: 10.1104/pp.110.160796.

La, T., E. Large, E. Taliercio, Q. Song, J.D. Gillman, et al. 2019. Characterization of Select Wild Soybean Accessions in the USDA Germplasm Collection for Seed Composition and Agronomic Traits. *Crop Sci.* 59(0): 1–19. doi: 10.2135/cropsci2017.08.0514.

Langewisch, T., J. Lenis, G.L. Jiang, D. Wang, V. Pantalone, et al. 2017. The development and use of a molecular model for soybean maturity groups. *BMC Plant Biol.* 17(1): 91. doi: 10.1186/s12870-017-1040-4.

Leamy, L.J., H. Zhang, C. Li, C.Y. Chen, and B.H. Song. 2017. A genome-wide association study of seed composition traits in wild soybean (*Glycine soja*). *BMC Genomics* 18(1): 18. doi: 10.1186/s12864-016-3397-4.

Lee, S.H., M.A. Bailey, M.A.R. Mian, T.E. Carter, D.A. Ashley, et al. 1996. Molecular markers associated with soybean plant height, lodging, and maturity across locations. *Crop Sci.* 36(3): 728–735. doi: 10.2135/cropsci1996.0011183X003600030035x.

Lee, J.-D., J.G. Shannon, T.D. Vuong, and H.T. Nguyen. 2009. Inheritance of Salt Tolerance in Wild Soybean (*Glycine soja* Sieb. and Zucc.) Accession PI483463. *J. Hered.* 100(6): 798–801. doi: 10.1093/jhered/esp027.

Li, S., Y. Cao, J. He, T. Zhao, and J. Gai. 2017. Detecting the QTL-allele system conferring flowering date in a nested association mapping population of soybean using a

novel procedure. *Theor. Appl. Genet.* 130(11): 2297–2314. doi: 10.1007/s00122-017-2960-y.

Li, D., T.W. Pfeiffer, and P.L. Cornelius. 2008. Soybean QTL for Yield and Yield Components Associated with Alleles. *Crop Sci.* 48(2): 571. doi: 10.2135/cropsci2007.06.0361.

Liu, B., T. Fujita, Z.-H. Yan, S. Sakamoto, D. Xu, et al. 2007. QTL mapping of domestication-related traits in soybean (*Glycine max*). *Ann. Bot.* 100(5): 1027–38. doi: 10.1093/aob/mcm149.

Liu, B., A. Kanazawa, H. Matsumura, R. Takahashi, K. Harada, et al. 2008. Genetic Redundancy in Soybean Photoresponses Associated with Duplication of the Phytochrome A Gene. *Genetics* 180: 995–1007. doi: 10.1534/genetics.108.092742.

Mao, T., J. Li, Z. Wen, T. Wu, C. Wu, et al. 2017. Association mapping of loci controlling genetic and environmental interaction of soybean flowering time under various photo-thermal conditions. *BMC Genomics* 18(1): 415. doi: 10.1186/s12864-017-3778-3.

Ning, W., H. Zhai, J. Yu, S. Liang, X. Yang, et al. 2017. Overexpression of Glycine soja WRKY20 enhances drought tolerance and improves plant yields under drought stress in transgenic soybean. *Mol. Breed.* 37(2): 19. doi: 10.1007/s11032-016-0614-4.

Nisa, Z. un, A.I. Mallano, Y. Yu, C. Chen, X. Duan, et al. 2017. GsSNAP33, a novel Glycine soja SNAP25-type protein gene: Improvement of plant salt and drought tolerances in transgenic *Arabidopsis thaliana*. *Plant Physiol. Biochem.* 119: 9–20. doi: 10.1016/j.plaphy.2017.07.029.

Price, A.L., N.A. Zaitlen, D. Reich, and N. Patterson. 2010. New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.* 11(7): 459–463. doi: 10.1038/nrg2813.

Ray, D.K., N.D. Mueller, P.C. West, and J.A. Foley. 2013. Yield Trends Are Insufficient to Double Global Crop Production by 2050 (J.P. Hart, editor). *PLoS One* 8(6): e66428. doi: 10.1371/journal.pone.0066428.

Reinprecht, Y., V.W. Poysa, K. Yu, I. Rajcan, G.R. Ablett, et al. 2006. Seed and agronomic QTL in low linolenic acid, lipoxygenase-free soybean (*Glycine max* (L.) Merrill) germplasm. *Genome* 49(12): 1510–1527. doi: 10.1139/g06-112.

Rincker, K., R. Nelson, J. Specht, D. Sleper, T. Cary, et al. 2014. Genetic Improvement of U.S. Soybean in Maturity Groups II, III, and IV. *Crop Sci.* 54(0): 1419–1432. doi: 10.2135/cropsci2013.10.0665.

Rogers, J., P. Chen, A. Shi, B. Zhang, A. Scaboo, et al. 2015. Agronomic performance and genetic progress of selected historical soybean varieties in the southern USA. *Plant Breed.* 134(1): 85–93. doi: 10.1111/pbr.12222.

Rowntree, S.C., J.J. Suhre, N.H. Weidenbenner, E.W. Wilson, V.M. Davis, et al. 2013. Genetic Gain $\times$ Management Interactions in Soybean: I. Planting Date. *Crop Sci.* 53(3): 1128. doi: 10.2135/cropsci2012.03.0157.

Sharma, R., F. Draicchio, H. Bull, P. Herzig, A. Maurer, et al. 2018. Genome-wide association of yield traits in a nested association mapping population of barley reveals new gene diversity for future breeding. *J. Exp. Bot.* 69(16): 3811–3822. doi: 10.1093/jxb/ery178.

Song, Q., D.L. Hyten, G. Jia, C. V. Quigley, E.W. Fickus, et al. 2013. Development and Evaluation of SoySNP50K, a High-Density Genotyping Array for Soybean. *PLoS One* 8(1): e54985. doi: 10.1371/journal.pone.0054985.

Song, Q., D.L. Hyten, G. Jia, C. V. Quigley, E.W. Fickus, et al. 2015. Fingerprinting Soybean Germplasm and Its Utility in Genomic Research. *G3*. 5(10): 1999–2006. doi: 10.1534/g3.115.019000.

Song, Q., J. Jenkins, G. Jia, D.L. Hyten, V. Pantalone, et al. 2016. Construction of high resolution genetic linkage maps to improve the soybean genome sequence assembly Glyma1.01. *BMC Genomics* 17(1): 33. doi: 10.1186/s12864-015-2344-0.

Specht, J.E., K. Chase, M. Macrander, G.L. Graef, J. Chung, et al. 2001. Soybean response to water: A QTL analysis of drought tolerance. *Crop Sci.* 41(2): 493–509. doi: 10.2135/cropsci2001.412493x.

Wang, D., B.W. Diers, P.R. Arelli, and R.C. Shoemaker. 2001. Loci underlying resistance to Race 3 of soybean cyst nematode in *Glycine soja* plant introduction 468916. *Theor. Appl. Genet.* 103(4): 561–566. doi: 10.1007/PL00002910.

Wang, X., G.-L. Jiang, M. Green, R.A. Scott, Q. Song, et al. 2014. Identification and validation of quantitative trait loci for seed yield, oil and protein contents in two recombinant inbred line populations of soybean. *Mol. Genet. Genomics* 289(5): 935–949. doi: 10.1007/s00438-014-0865-x.

Wang, D., A.M. Procopiuk, B.W. Diers, and G.L. Graef. 2004. Identification of putative QTL that underlie yield in interspecific soybean backcross populations. *TAG Theor. Appl. Genet.* 108(3): 458–467. doi: 10.1007/s00122-003-1449-z.

Wei, J., and S. Xu. 2016. A Random-Model Approach to QTL Mapping in Multiparent Advanced Generation Intercross (MAGIC) Populations. *Genetics* 202(2): 471–486. doi: 10.1534/genetics.115.179945.

Weir, B.S., and C.C. Cockerham. 1984. Estimating F-Statistics for the Analysis of Population Structure. *Evolution*. 38(6): 1358. doi: 10.2307/2408641.

Wen, Z., Y. Ding, T.-J. Zhao, and J. Gai. 2009. Genetic diversity and peculiarity of annual wild soybean (*G. soja* Sieb. et Zucc.) from various eco-regions in China. *Theor. Appl. Genet.* 119(2): 371–381. doi: 10.1007/s00122-009-1045-y.

Winter, S.M.J., B.J. Shelp, T.R. Anderson, T.W. Welacky, and I. Rajcan. 2007. QTL associated with horizontal resistance to soybean cyst nematode in *Glycine soja* PI464925B. *Theor. Appl. Genet.* 114(3): 461–472. doi: 10.1007/s00122-006-0446-4.

Wolfgang, G., and Y. Qiang C. 2017. Genetic separation of southern and northern soybean breeding programs in North America and their associated allelic variation at four maturity loci. *Mol. Breed.* 37(1): 8. doi: 10.1007/s11032-016-0611-7.

Xavier, A., B. Hall, A.A. Hearst, K.A. Cherkauer, and K.M. Rainey. 2017. Genetic Architecture of Phenomic-Enabled Canopy Coverage in *Glycine max*. *Genetics* 206(2): 1081–1089. doi: 10.1534/genetics.116.198713.

Xavier, A., D. Jarquin, R. Howard, V. Ramasubramanian, J.E. Specht, et al. 2018. Genome-Wide Analysis of Grain Yield Stability and Environmental Interactions in a Multiparental Soybean Population. *G3*. 8(2): 519–529. doi: 10.1534/g3.117.300300.

Xavier, A., S. Xu, W.M. Muir, and K.M. Rainey. 2015. NAM: association studies in multiple populations. *Bioinformatics* 31(23): btv448. doi: 10.1093/bioinformatics/btv448.

Xia, Z., S. Watanabe, T. Yamada, Y. Tsubokura, H. Nakashima, et al. 2012. Positional cloning and characterization reveal the molecular basis for soybean maturity locus E1 that regulates photoperiodic flowering. *Proc. Natl. Acad. Sci.* 109(32): E2155–E2164. doi: 10.1073/pnas.1117982109.

Xu, Y., P. Li, Z. Yang, and C. Xu. 2017. Genetic mapping of quantitative trait loci in crops. *Crop J.* 5(2): 175–184. doi: 10.1016/j.cj.2016.06.003.

Xu, M., Z. Xu, B. Liu, F. Kong, Y. Tsubokura, et al. 2013. Genetic variation in four maturity genes affects photoperiod insensitivity and PHYA-regulated post-flowering responses of soybean. *BMC Plant Biol.* 13(1): 91. doi: 10.1186/1471-2229-13-91.

Yamanaka, N., S. Watanabe, K. Toda, M. Hayashi, H. Fuchigami, et al. 2005. Fine mapping of the FT1 locus for soybean flowering time using a residual heterozygous line derived from a recombinant inbred line. *Theor. Appl. Genet.* 110(4): 634–639. doi: 10.1007/s00122-004-1886-3.

Yan, L., N. Hofmann, S. Li, M.E. Ferreira, B. Song, et al. 2017. Identification of QTL with large effect on seed weight in a selective population of soybean with genome-wide association and fixation index analyses. *BMC Genomics* 18(529): 1–11. doi: 10.1186/s12864-017-3922-0.

Yan, L., L.-L. Xing, C.-Y. Yang, R.-Z. Chang, M.-C. Zhang, et al. 2014. Identification of quantitative trait loci associated with soybean seed protein content using two populations derived from crosses between *Glycine max* and *Glycine soja*. *Plant Genet. Resour.* 12(S1): S104–S108. doi: 10.1017/S1479262114000379.

Yu, J., J.B. Holland, M.D. McMullen, and E.S. Buckler. 2008. Genetic design and statistical power of nested association mapping in maize. *Genetics* 178(1): 539–51. doi: 10.1534/genetics.107.074245.

Zhang, H., C. Li, E.L. Davis, J. Wang, J.D. Griffin, et al. 2016. Genome-Wide Association Study of Resistance to Soybean Cyst Nematode (*Heterodera glycines*) HG Type 2.5.7 in Wild Soybean (*Glycine soja*). *Front. Plant Sci.* 7(1214): 1–11. doi: 10.3389/fpls.2016.01214.

Zhang, H., N. Mittal, L.J. Leamy, O. Barazani, and B.H. Song. 2017a. Back into the wild—Apply untapped genetic diversity of wild relatives for crop improvement. *Evol. Appl.* 10(1): 5–24. doi: 10.1111/eva.12434.

Zhang, J., Q. Song, P.B. Cregan, R.L. Nelson, X. Wang, et al. 2015. Genome-wide association study for flowering time, maturity dates and plant height in early maturing soybean (*Glycine max*) germplasm. *BMC Genomics* 16(1): 217. doi: 10.1186/s12864-015-1441-4.

Zhang, H., Q. Song, J.D. Griffin, and B.-H. Song. 2017b. Genetic architecture of wild soybean (*Glycine soja*) response to soybean cyst nematode (*Heterodera glycines*). *Mol. Genet. Genomics* 292(6): 1257–1265. doi: 10.1007/s00438-017-1345-x.

Zhang, X., W. Wang, N. Guo, Y. Zhang, Y. Bu, et al. 2018. Combining QTL-seq and linkage mapping to fine map a wild soybean allele characteristic of greater plant height. *BMC Genomics* 19(1): 226. doi: 10.1186/s12864-018-4582-4.

Zhang, S., Z. Zhang, C. Bales, C. Gu, C. DiFonzo, et al. 2017c. Mapping novel aphid resistance QTL from wild soybean, *Glycine soja* 85-32. *Theor. Appl. Genet.* 130(9): 1941–1952. doi: 10.1007/s00122-017-2935-z.

Zhou, Z., Y. Jiang, Z. Wang, Z. Gou, J. Lyu, et al. 2015. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat. Biotechnol.* 33(4): 408–414. doi: 10.1038/nbt.3096.

Zhu, C., M. Gore, E.S. Buckler, and J. Yu. 2008. Status and Prospects of Association Mapping in Plants. *Plant Genome J.* 1(1): 5. doi: 10.3835/plantgenome2008.02.0089.

Tables and Figures

Table 2-1. Summary of mean square estimates from the analysis of variance for yield, plant maturity, plant height, and lodging during 2016 and 2017. ** and * Significant at the 0.001 and 0.01 probability level. h^2 , heritability on an entry-mean basis; W82, Williams 82 (hub-parent); SD, standard deviation

Source of Variance		Mean Square			
		Grain Yield (kg ha ⁻¹)	Plant Maturity (days)	Plant Height (cm)	Lodging (1-5)
2016	Genotype (G)	53632**	222*	1179**	1.65**
	Environment (E)	1204950**	90516**	32287**	55.2**
	Check x Block	35146**	25	585**	0.71**
	G x E	11616*	59*	299**	0.26**
	Residual	8146	14.1	59	0.13**
	h^2	0.52	0.93	0.91	0.89
	Mean RILs	3596.6	31.3	123.7	3.6
	Mean W82	3613.2	27.3	113.2	2.9
	SD	287	4.2	17.7	0.6
	Range	1945 - 5978	18 - 42	67.3 - 179.3	1.5 - 5
2017	Genotype (G)	3193873**	134.3**	2353**	5.51**
	Environment (E)	119326886**	3321.6**	220030**	115.96**
	Check x Block	1622614**	13.2	1417**	2.49**
	E(rep)	139479	1.3	423	0.196
	Block(rep)	296786*	24.2**	372	0.6
	G x E	238206**	14.1**	286**	0.49**
	Residual	125579	8.9	222	0.29
	h^2	0.51	0.91	0.92	0.91
	Mean RILs	3401.6	30.8	119.3	4.1
	Mean W82	3558.2	26.5	111.8	3.1
	SD	241.9	4.1	16.9	0.77

		Range	1803 - 4559	14 - 43	69.9 - 162.8	1.5 - 5
Maturity Group IV	2016	Genotype (G)	1493147**	89.4**	2310.4**	1.40**
		Environment (E)	85201558**	19903.2**	17178**	10.67**
		Check x Block	2621268**	7.3*	666.3**	1.8**
		G x E	372522**	20.6**	433.1**	0.35**
		Residual	141238	3.2	14.1	0.13
		h^2	0.54	0.87	0.82	0.87
		Mean RILs	3443.7	38.3	134.7	3.92
		Mean W82	3653.2	29.1	116.2	2.83
		SD	233.4	4.7	25.5	0.56
	2017	Range	1653 - 5356	26 - 53	44 - 226.3	1.5 - 5
		Genotype (G)	2067726**	458.1**	4029**	4.07**
		Environment (E)	30046942**	6656.8**	141020**	33.52**
		Check x Block	1965437**	435.4**	1723**	3.04**
		E(rep)	3873504**	858.2**	1092	1.11**
		Block(rep)	333184*	73.8**	753	0.14
		G x E	210934**	46.7**	917**	0.53**
		Residual	93632	20.7	780	0.24
		h^2	0.64	0.93	0.9	0.85
		Mean RILs	3382.1	37.2	124.5	4.1
		Mean W82	3484.4	30.1	113.6	3.1
		SD	210.6	5.9	22.2	0.61
		Range	1350 - 4743	21 - 52	56.7 - 179.5	1.5 - 5

Figure 2-1. Histogram of grain yield (GY), plant maturity (PM), plant height (PH), and lodging (LOD) among the NAM population and the hub-parent (Williams 82) in 2016 and 2017.

Figure 2-2. Linkage disequilibrium (LD) decay of the nested association mapping (NAM) population (A), Distribution of the NAM RILs under PC1 and PC2 (B), and fixation index of the NAM population over the entire genome where the red line represents the 99th percentile (0.29).

Figure 2-3. Manhattan plots of nested association mapping for grain yield, plant maturity, plant height, and lodging plotted against positions on each of the 20 chromosomes, across eight environments during 2016 (red dots) and 2017 (blue dots). The significant SNPs were distinguished by the FDR 0.05 (3.8×10^{-5}) threshold line.

c)

PLANT HEIGHT - ALBANY

PLANT HEIGHT - COLUMBIA

PLANT HEIGHT - NOVELTY

PLANT HEIGHT - ROCK PORT

Figure 2-4. Manhattan plot of nested association mapping for grain yield (A), plant maturity (B), plant height (C), and lodging (D) plotted against positions on each of the 20 chromosomes in individual environments during 2016 (red dots) and 2017 (blue dots). The significant SNPs were distinguished by the FDR 0.05 (3.8×10^{-5}) threshold line.

Figure 2-5. The difference in genotypic values (A, B, C) between lines segregating at SNP ss715593853 (A), ss715608829 (B), and ss715613171 (C) associated with plant maturity in the NAM population. The difference in genotypic values of plant height associated with SNP ss715616047 (D), and lodging associated with SNP ss715602804 (E), between lines in the NAM population across eight environments. The difference in genotypic values between lines segregating at SNP ss715625973 for grain yield at Pop3 – Williams 82 x PI522226 across eight environments (F). The least square difference (LSD) and differences in mean (Δm), and the probability of a greater F statistic ($Pr > F$) and R^2 are in the figures.

Figure 2-6. Graphical display of significant QTL detected in three populations (Pop1-Williams 82 x PI464890B; Pop2 – Williams 82 x PI458536; Pop3 – Williams 82 x PI522226) for (A) grain yield, (B) plant maturity, (C) plant height and (D) lodging across eight environments. The dashed dark-red horizontal line indicates the threshold LOD score based on a 1000 permutation test.

Figure 2-7. Graphical display of significant QTL detected in three populations (Pop1-Williams 82 x PI464890B; Pop2 – Williams 82 x PI458536; Pop3 – Williams 82 x PI522226) for grain yield, plant maturity, plant height and, lodging in across eight environments in 2016 and 2017. The dashed dark-red horizontal line indicates the threshold LOD score based on a 1000 permutation test.

Table 2-2. Significant SNPs for grain yield, plant maturity, plant height, and lodging and their respective allelic effect in relation to the hub parent Williams 82 across eight environments during 2016 and 2017.

Allelic Effect											Genes/QTLs	
Year	Chr ^{\$}	SNP ID	Position ^{\$}	<i>Fst</i> [*]	MAF [#]	Allele	-log(10)P	PI464890B	PI458536	PI522226		
Grain Yield (kg ha ⁻¹)												
2017	16	ss715624484	31787658	0.00	0.23	T/C	5.07	59.4	-116.2	-23.3	<i>Pdh1</i> (Funatsuki et al., 2014)	
	16	ss715624210	29720274	0.06	0.09	A/G	4.55	4.3	-115.7	0.0		
	16	ss715624192	29528259	0.02	0.16	T/C	4.41	4.6	-76.0	-32.4		
	16	ss715624203	29681065	0.02	0.17	T/G	4.15	1.2	-98.7	3.6		
Plant Maturity (days)												
2016	6	ss715594002	25799930	0.13	0.49	C/T	9.84	1.18	0.54	0.08	<i>El/e1-as</i> (Xia et al., 2012)	
	6	ss715593853	20370075	0.14	0.43	T/C	9.25	0.72	0.85	0.17		
	6	ss715593820	19282687	0.02	0.40	A/G	8.3	0.92	0.51	0.16		
	6	ss715594079	28816949	0.12	0.43	A/G	8.29	1.14	0.58	-0.03		
	6	ss715594119	30534657	0.13	0.50	G/A	7.95	1.09	0.60	-0.04		
	6	ss715594029	26981990	0.12	0.47	G/T	7.83	1.11	0.63	-0.11		
	6	ss715594336	38703918	0.15	0.49	T/C	6.62	1.10	-0.05	0.41		
	6	ss715594135	31288514	0.06	0.43	A/C	6.54	1.23	0.77	-0.64		
	6	ss715594041	27322640	0.10	0.47	A/G	6.08	0.95	0.44	-0.05		
	6	ss715594312	37732009	0.15	0.38	A/G	5.29	0.92	0.03	0.36		
	6	ss715593843	20152926	0.44	0.38	T/C	4.85	-0.09	0.90	0.40		
	6	ss715593873	20943239	0.44	0.40	T/C	4.82	-0.09	0.92	0.37		
	6	ss715594302	37364558	0.13	0.43	A/C	4.71	0.79	0.09	0.32		
	6	ss715593866	20739900	0.47	0.22	T/C	4.21	-0.11	1.23	0.00		
	6	ss715594345	38999119	0.16	0.47	T/C	4	0.64	0.15	0.30		
	6	ss715594028	26933523	0.10	0.15	C/T	3.78	2.88	-0.84	-0.81		
	11	ss715608829	11269310	0.02	0.29	G/A	7.78	1.35	-0.08	0.07		First flower (Fang et al., 2017) (Mao et al., 2017) (Li et al., 2017) (Zhang et al., 2015)
	11	ss715608825	11131574	0.06	0.24	C/T	6.6	1.30	-0.29	0.26		
	11	ss715608847	11450492	0.04	0.29	A/G	6.57	1.14	-0.04	0.19		
	11	ss715608984	13075838	0.03	0.28	T/C	6	0.98	0.21	0.01		
	11	ss715608810	10999596	0.09	0.31	T/C	5.92	1.26	-0.18	0.04		

	11	ss715608803	10926986	0.09	0.29	T/C	5.52	1.30	-0.33	0.08	
	11	ss715609066	14571958	0.04	0.27	T/C	5.51	1.01	0.15	-0.01	
	11	ss715608919	11932224	0.03	0.27	G/A	5.49	1.00	0.10	0.06	
	11	ss715608797	10892564	0.10	0.28	G/A	5.38	1.26	-0.35	0.12	
	11	ss715608857	11633102	0.11	0.22	G/A	5.26	1.05	0.15	-0.07	
	11	ss715608784	10802698	0.15	0.27	A/G	5.07	1.48	-0.72	0.07	
	11	ss715609103	14896456	0.03	0.26	G/A	5	0.87	0.22	0.00	
	11	ss715608840	11383598	0.28	0.15	C/T	4.72	1.34	-0.54	0.15	
	11	ss715609099	14829557	0.05	0.23	A/C	4.28	0.90	0.11	-0.03	
	11	ss715609111	15000789	0.26	0.12	T/C	3.83	1.53	-0.46	-0.52	
	11	ss715608852	11572077	0.18	0.28	A/C	3.74	0.86	0.19	-0.04	
	12	ss715613171	5392662	0.06	0.26	C/T	4.42	0.19	0.17	0.68	First flower
	12	ss715613192	5610878	0.08	0.24	T/C	3.94	0.18	0.19	0.62	(Fang et al., 2017)
	12	ss715613162	5309573	0.22	0.20	G/A	3.86	0.27	0.00	0.72	(Mao et al., 2017)
	12	ss715613203	5706745	0.08	0.26	G/A	3.73	0.16	0.19	0.59	
2017	6	ss715594029	26981990	0.12	0.47	G/T	16.77	1.94	1.66	-0.11	E1/e1-as
	6	ss715594119	30534657	0.13	0.50	G/A	16.76	1.91	1.55	0.04	(Xia et al., 2012)
	6	ss715593853	20370075	0.14	0.43	T/C	16.38	0.92	1.94	0.50	
	6	ss715594079	28816949	0.12	0.43	A/G	15.88	2.05	1.34	0.04	
	6	ss715594002	25799930	0.13	0.49	C/T	15.43	1.80	1.21	0.38	
	6	ss715594135	31288514	0.06	0.43	A/C	13.1	2.08	1.96	-1.18	
	6	ss715593820	19282687	0.02	0.40	A/G	12.29	1.57	0.69	0.65	
	6	ss715594041	27322640	0.10	0.47	A/G	10.23	1.29	1.43	-0.12	
	6	ss715594336	38703918	0.15	0.49	T/C	10.18	2.11	0.00	0.62	
	6	ss715594302	37364558	0.13	0.43	A/C	10	1.88	0.10	0.72	
	6	ss715593843	20152926	0.44	0.38	T/C	9.36	-0.74	2.14	0.94	
	6	ss715593873	20943239	0.44	0.40	T/C	9.35	-0.75	2.14	0.94	
	6	ss715594312	37732009	0.15	0.38	A/G	9.27	1.81	0.28	0.55	
	6	ss715593889	21363243	0.38	0.38	A/G	8.64	-1.33	2.08	1.36	
	6	ss715594345	38999119	0.16	0.47	T/C	8.33	1.42	0.37	0.70	

	6	ss715593866	20739900	0.47	0.22	T/C	7.87	-0.86	2.85	0.00	
	6	ss715594403	41268682	0.16	0.49	G/T	7.79	1.37	0.51	0.49	
	6	ss715594028	26933523	0.10	0.15	C/T	7.41	5.73	0.00	-2.02	
	6	ss715594474	43688393	0.19	0.45	A/G	6.17	0.91	0.19	0.98	
	6	ss715593740	17356033	0.00	0.45	G/A	6.07	1.34	0.23	0.60	
	6	ss715594428	42308631	0.17	0.49	A/C	5.52	1.06	0.16	0.70	
	6	ss715594445	42883965	0.18	0.46	T/C	4.99	0.60	0.32	0.92	
	6	ss715594354	39421942	0.03	0.41	C/T	4.96	1.87	0.59	-0.72	
	6	ss715593703	17188046	0.00	0.40	A/G	4.44	1.13	-0.09	0.73	
	6	ss715593520	16372276	0.02	0.31	T/C	4.41	0.59	0.48	0.51	
	12	ss715613171	5392662	0.06	0.26	C/T	6.29	0.03	0.55	1.30	First flower
	12	ss715613192	5610878	0.08	0.24	T/C	6.02	0.12	0.54	1.20	(Fang et al., 2017)
	12	ss715613203	5706745	0.08	0.26	G/A	5.79	-0.06	0.59	1.26	(Mao et al., 2017)
	12	ss715613110	4780837	0.04	0.26	A/G	4.9	0.05	0.60	1.05	
	12	ss715613162	5309573	0.22	0.20	G/A	4.89	0.03	0.00	1.55	
	12	ss715613311	6663270	0.10	0.25	C/T	4.22	0.08	0.36	0.94	
	12	ss715613207	5786241	0.02	0.20	A/G	4.18	-0.04	0.50	1.21	
	12	ss715613125	4925203	0.08	0.24	T/C	4.15	-0.11	0.62	1.02	
				Plant Height (cm)							
2016	10	ss715606293	3426604	0.06	0.09	G/A	4.98	0.23	0.57	-2.20	
	13	ss715616047	37463730	0.05	0.12	C/A	5.51	2.31	-1.98	-2.26	PH 1-g11, 1-g12
	13	ss715615814	35688167	0.07	0.09	G/T	4.8	2.86	4.39	8.29	(Zhang et al., 2015)
	13	ss715615796	35515983	0.02	0.16	A/G	4.43	2.41	1.86	-1.87	qPH13.1
	13	ss715615876	36073020	0.02	0.07	C/A	4.15	4.99	8.86	13.07	(Zhang et al., 2018)
	13	ss715616199	39307253	0.00	0.18	G/A	3.89	0.96	-0.39	0.03	
2017	13	ss715616047	37463730	0.05	0.118	C/A	5.86	1.74	0.76	0.37	
	13	ss715616074	38032737	0.10	0.249	G/A	4.92	1.31	2.13	-1.36	qPH13.1
	13	ss715616026	37144714	0.04	0.11	T/C	4.81	-1.53	0.68	2.21	(Zhang et al., 2018)
	13	ss715616134	38620130	0.10	0.172	A/G	4.52	-1.70	2.45	0.78	
	13	ss715616131	38554316	0.16	0.341	C/T	3.86	-0.56	1.91	0.23	
				Lodging (1-5 scale)							

2016	8	ss715602804	8836971	0.21	0.14	C/T	4.19	0.01	0.03	0.01	
	8	ss715602662	7070019	0.11	0.18	G/A	3.88	-0.04	0.04	0.03	
2017	1	ss715579524	45064727	0.06	0.08	A/G	3.87	-0.02	0.06	0.06	
	8	ss715602804	8836971	0.21	0.14	C/T	4.2	-0.03	0.04	0.06	
	18	ss715630642	458242	0.01	0.10	T/G	4.42	-0.78	0.00	0.00	Lod20-7 (Reinprecht et al. 2006)
	19	ss715633186	1489018	0.09	0.25	G/A	4.38	0.03	-0.53	-0.09	Lod9-3 (Specht et al. 2001)

---

* Fst, fixation index

MAF, minimum allele frequency

§ Chr, chromosome

§ Position in base pair

Table 2-3. Significant QTL results in three populations (Pop1- Williams 82 x PI464890B; Pop2 – Williams 82 x PI458536; Pop3 – Williams 82 x PI522226) for four agronomic traits across eight environments in two years (2016 and 2017).

TRAIT - GRAIN YIELD (Kg ha ⁻¹)										
	Year	QTL	Chr	Peak marker	Position (cM)	1.5 Interval	Marker Interval	LOD	R ² (%)	Add effect*
Pop3	2016	qGY-17	17	Gm17:13125043	73.39	72.12-90.40	Gm17:12210561..37989202	6.6	24.4	166.1
Pop2	2016	qGY-18	18	Gm18:60456270	100	99.13-102	Gm18:60278667..60938849	9.3	21.2	-397.2
Pop1	2017	qGY-20	20	Gm20:2432223	24.9	18.89-27.89	Gm20:1789933..2804571	6.4	17.5	-182.3
TRAIT - PLANT MATURITY (days)										
Pop1	2017	qPM-6-1	6	Gm06:28816949	111.77	111.44-112.34	Gm6:21363243..30534657	21.4	48.1	4.2
Pop2	2016	qPM-6-2	6	Gm06:19282687	110	32.4-112.02	Gm6:5493202..30534657	11.7	34.8	2.3
Pop2	2017	qPM-6-2	6	Gm06:19282687	110	109.72-111	Gm:619282687..21363243	23.3	52.6	5.5
Pop3	2017	qPM-6-3	6	Gm06:25799930	111.52	110.94-116.94	Gm:621363243..43980786	11.4	28.3	3.7
Pop1	2016	qPM-11	11	Gm11:17462922	76.3	69.3-93.68	Gm11:10446473..37237023	6.3	10.3	1.7
Pop3	2016	qPM-12	12	Gm12:5309573	29.74	25.02-35	Gm12:4592200..6023395	6.4	21.7	2.3
Pop3	2017	qPM-12	12	Gm12:5309573	29.74	23.6-30.74	Gm12:4268262..5475411	12.1	22.1	3.9
Pop1	2017	qPM-20	20	Gm20:37466130	66.89	60.31-67.89	Gm20:36153048..37551598	7.8	12.5	-2.9
TRAIT - PLANT HEIGHT (cm)										
Pop2	2017	qPH-1	1	Gm01:4726775	38.71	35-40	Gm1:3905200..4912239	6.5	18.7	-14.2
Pop1	2016	qPH-2	2	Gm02:1398489	5	4.0-6.0	Gm2:1033642..1461422	11.3	15.2	-8.9
Pop1	2017	qPH-2	2	Gm02:1398489	5	4.0-7.0	Gm2:1033642..1830094	10.1	20.1	-12.3
Pop3	2017	qPH-13	13	Gm13:37144714	95.55	94.17-98.17	Gm13:37018225..38032737	8.5	23.3	11.8
TRAIT - LODGING (1 to 5 scale)										
Pop2	2016	qLOD-9	9	Gm09:38507112	85.25	77.21-90	Gm9:36946741..39067581	8.1	14.3	0.24
Epistatic Interactions between QTL										
	Trait	QTL	Chr	Marker 1	Marker 2			LOD	R ² (%)	P-value
Pop1	MAT 17	qPM-6-1 x qPM-20	6 x 20	28816949	37466130			5.8	5.6	0.012

*Additive effect represents the impact of converting homozygous *G. max* allele to homozygous *G. soja* allele.

Table 2-4. Selected *G. soja*-derived RILs exhibiting significantly higher grain yield than the NAM hub-parent Williams 82 for test MG III and MG IV averaged over eight environments in 2016 and 2017.

Environment	Entry	Pedigree	Yield	Yield*	Maturity
			Kg ha ⁻¹	%	days after Sep 1
Combined 2016					
	Williams 82	<i>G. max</i> parent	3995.7	100	28
	SA15-63645	Williams 82 x PI 458536	4588.0	115	29
	SA15-64483	Williams 82 x PI 458536	4559.7	114	32
	SA15-66082	Williams 82 x PI 458536	4518.9	113	25
	SA15-64246	Williams 82 x PI 522226	4469.3	112	33
	SA15-66599	Williams 82 x PI 522226	4460.2	112	25
	SA15-64031	Williams 82 x PI 464890B	4383.9	110	32
	SA15-66476	Williams 82 x PI 522226	4379.5	110	26
	SA15-63639	Williams 82 x PI 458536	4350.7	109	28
	SA15-63899	Williams 82 x PI 458536	4298.0	108	32
	SA15-64507	Williams 82 x PI 458536	4273.4	107	25
	SA15-63650	Williams 82 x PI 458536	4272.4	107	33
	SA15-66563	Williams 82 x PI 522226	4265.0	107	26
	SA15-65988	Williams 82 x PI 458536	4250.5	106	36
	SA15-64255	Williams 82 x PI 522226	4230.8	106	30
	SA15-63648	Williams 82 x PI 458536	4216.0	106	33
	SA15-63663	Williams 82 x PI 458536	4187.2	105	29
	SA15-63634	Williams 82 x PI 458536	4182.0	105	36
		LSD [§] (P<0.05)	183		
Albany 2017					
	Williams 82	<i>G. max</i> parent	3658.3	100	30
	SA15-66563	Williams 82 x PI 522226	4243.2	116	26
	SA15-63423	Williams 82 x PI 464890B	4080.0	112	28
	SA15-66476	Williams 82 x PI 522226	4025.4	110	27
	SA15-66599	Williams 82 x PI 522226	3967.0	108	27
	SA15-109	Williams 82 x PI 458536	3879.1	106	31

		LSD (P<0.05)	173		
Columbia 2017					
Williams 82	<i>G. max</i> parent	4099.2	100	24	
SA15-66476	Williams 82 x PI 522226	4682.6	114	23	
SA15-109	Williams 82 x PI 458536	4611.6	112	25	
SA15-66456	Williams 82 x PI 522226	4565.8	111	19	
SA15-63422	Williams 82 x PI 464890B	4548.5	111	20	
SA15-63844	Williams 82 x PI 458536	4465.7	109	25	
SA15-64031	Williams 82 x PI 464890B	4453.4	109	24	
SA15-66563	Williams 82 x PI 522226	4449.0	109	22	
SA15-66389	Williams 82 x PI 522226	4398.1	107	32	
SA15-66418	Williams 82 x PI 522226	4385.0	107	20	
SA15-63157	Williams 82 x PI 464890B	4378.8	107	21	
SA15-66599	Williams 82 x PI 522226	4359.2	106	19	
SA15-63798	Williams 82 x PI 458536	4334.9	106	19	
		LSD (P<0.05)	218		
Novelty 2017					
Williams 82	<i>G. max</i> parent	3731.3	100	25	
SA15-66389	Williams 82 x PI 522226	4196.1	112	32	
SA15-63423	Williams 82 x PI 464890B	4137.0	111	21	
SA15-63650	Williams 82 x PI 458536	4134.5	111	21	
SA15-65891	Williams 82 x PI 464890B	4089.3	110	35	
SA15-109	Williams 82 x PI 458536	3992.2	107	24	
		LSD (P<0.05)	225		
Rock Port 2017					
Williams 82	<i>G. max</i> parent	4230.8	100	32	
SA15-109	Williams 82 x PI 458536	4956.6	117	32	
SA15-63645	Williams 82 x PI 458536	4919.7	116	30	
SA15-63422	Williams 82 x PI 464890B	4912.5	116	29	
SA15-63650	Williams 82 x PI 458536	4904.7	116	30	
SA15-63182	Williams 82 x PI 464890B	4817.3	114	46	

SA15-63662	Williams 82 x PI 458536	4776.2	113	27
SA15-64246	Williams 82 x PI 522226	4724.1	112	44
SA15-66082	Williams 82 x PI 458536	4698.1	111	27
SA15-63862	Williams 82 x PI 458536	4692.1	111	47
SA15-63976	Williams 82 x PI 464890B	4644.5	110	29
SA15-63423	Williams 82 x PI 464890B	4640.5	110	30
SA15-64507	Williams 82 x PI 458536	4600.0	109	25
SA15-66056	Williams 82 x PI 458536	4572.1	108	46
SA15-63836	Williams 82 x PI 458536	4565.0	108	34
SA15-64256	Williams 82 x PI 522226	4543.0	107	35
SA15-63157	Williams 82 x PI 464890B	4537.7	107	30
SA15-64465	Williams 82 x PI 458536	4512.7	107	28
SA15-64348	Williams 82 x PI 522226	4482.4	106	41
SA15-63663	Williams 82 x PI 458536	4470.5	106	30
LSD (P<0.05)		234		

Combined 2017				
Williams 82	<i>G. max</i> parent	3929.8	100	27
SA15-109	Williams 82 x PI 458536	4329.2	110	28
SA15-66476	Williams 82 x PI 522226	4312.0	110	24
SA15-63423	Williams 82 x PI 464890B	4239.4	108	25
SA15-66563	Williams 82 x PI 522226	4129.0	105	24
SA15-66599	Williams 82 x PI 522226	4109.7	105	24
SA15-63650	Williams 82 x PI 458536	4095.2	104	24
SA15-63422	Williams 82 x PI 464890B	4081.6	104	25
LSD (P<0.05)		149		

§ LSD, least significant difference

* Yield as a percentage of the *G. max* parent Williams 82

Chapter III:
NESTED ASSOCIATION MAPPING OF SEED COMPOSITION
TRAITS IN THREE INTERSPECIFIC SOYBEAN
POPULATIONS

Abstract

Cultivated soybean seed typically has 20% oil and 40% protein on a dry weight basis. Soybean protein meal and oil can be used in human diets or livestock feed and are crucial in determining the value of the soybean crop. More recently, advances in manipulating fatty acid profile to achieve better quality oil for a wide range of applications has been largely successful. In contrast, manipulating protein and amino acid profiles to improve soybean meal value has been challenging. The objective of this study was to identify quantitative trait loci (QTL) associated with protein, oil, fatty acids and amino acid profile in an interspecific nested association mapping population developed from cross-pollination between “Williams 82” (hub-parent) and three *Glycine soja* parents (PI464890B, PI458536, and PI522226). Field tests with 392 recombinant inbred lines (RIL) were carried out in Missouri across eight environments during 2016 and 2017, in an augmented incomplete block design with two checks cultivars per block. The NAM population was phenotyped for seed composition traits using wet chemistry and near-infrared spectroscopy. We identified 61 and 12 QTLs associated with seed composition traits via the NAM analysis and linkage analysis, respectively. Four QTLs showed pleiotropic effects with other soybean seed composition traits. Two QTLs, on Chr 5 and another QTL on chromosome 15 was associated oleic, linoleic and linolenic acid, and

explained 3-18% of the phenotypic variance. We confirmed QTLs for protein and oil, cqSeed protein-001 on Chr 15 and cqSeed protein-003 on Chr 20 were identified and, the QTL on chromosome 20 was associated with ten amino acid. However, the allele associated with protein concentration was also responsible for a reduction in certain amino acids. Another QTL on Chr 19 was associated with Cysteine, Methionine, and Leucine and explained 9-30% of the phenotypic variation. Our results reinforce that increasing protein may not lead to increased amino acid concentrations and suggest independent genetic control for protein and sulfur-containing amino acids. The markers detected in this study provide information to facilitate further studies to improve soybean protein, oil, fatty acids, and amino acid composition and are expected to be useful for plant breeders to enhance seed composition traits.

Introduction

World soybean production has increased more than 3-fold, between 1980 and 2017, with a total production of 240 million metric tons in 2016/2017 (FAOSTAT, 2018). Cultivated soybean typically has 20% oil and 40% protein on a dry weight basis (Wilson 2004). Protein meal and oil can be used in human diets or livestock feed and are crucial to the value of the soybean crop. A better understanding of the genetic basis of seed protein and oil variation is essential to improve seed quality and market value of the crop.

Protein and oil content in soybean is genetically complex and highly polygenic and has been shown to have significant interaction with the environment (Chung et al. 2003; Lee et al. 2007). Several QTLs have been identified for protein and oil (Brummer et al. 1997; Panthee et al. 2005; Hwang et al. 2014a). Among 241 QTLs reported for seed protein content (<http://www.soybase.org>, “SoyBase browser”, verified 01/20/2019), two confirmed QTLs cqSeed protein-001 (Chr 15) and cqSeed protein-003 (Chr 20), were mapped in multiple populations (Diers et al. 1992; Brummer et al. 1997; Bandillo et al. 2015; Phansak et al. 2016), and were positively associated with protein and negative with oil content. Increasing seed protein content by 2% typically results in a simultaneous 1% decrease in seed oil content (Chung et al. 2003). This negative relationship is either due to pleiotropic effects between high protein alleles (s) on lower oil, or high protein allele(s) in repulsion phase with higher oil allele(s) (Chung et al. 2003). The cqSeed protein-003 on Chr 20 was identified in an 8.4 Mb region (25-33 Mb) (Bolon et al. 2010). Vaughn et al. (2014) and Hwang et al. (2014c) narrowed down the region to 2.4 and 900 kb, respectively. Bandillo et al. (2015) identified five haplotype blocks between Vaughn et al. (2014), and Hwang et al. (2014c) interval regions. Two haplotypes between 30.38-

32.05 Mb showed the highest significant SNPs associated with the protein content. The most plausible candidate genes identified by Bandillo et al. (2015) were *Glyma20g85000*, *Glyma20g86000*, and *Glyma20g87000*.

The oil quality in soybean is determined by two saturated fatty acids, palmitic (16:0), stearic (18:0), and three unsaturated fatty acids, oleic (18:1), linoleic (18:2), and linolenic (18:3) (Lee et al. 2007). Polyunsaturated fatty acids (linoleic and linolenic acid) are more susceptible to oil oxidation, deteriorating the oil properties and shelf life (Mounts et al. 1988; Kamal-Eldin 2006). Hydrogenation is used to increase monosaturated fatty acids profile in the soybean oil. However, during the process *trans* fats are generated, which are associated with health issues, such as coronary heart disease (Kromhout 1999).

Recent advances in oil quality have been achieved using a combination of mutant fatty acid desaturase mutant genes, *fad2-1A*, *fad2-1B*, and *fad3* genes, resulting in a high oleic (>80%) low linolenic (<3%) soybean oil (Pham et al. 2012). Oleic acid increases oil stability and extends the utility of soybean oil for human and industrial uses. Many loci have been identified to be associated with fatty acid profile (Priolli et al. 2015; Li et al. 2015; Fang et al. 2017) demonstrating the complexity of the fatty acid control in soybean. Recently (Leamy et al. 2017) using the *G. soja* plant introductions (PIs) and (Zhang et al. 2018a) using *G. max* PIs identified novel regions associated with fatty acid composition.

Amino acid composition also plays an essential role in soybean meal quality particularly for animal nutrition, especially in poultry and swine. The value of the soybean meal can be enhanced through genetic improvement of limited amino acids in soybean meal, such as methionine and cysteine (Pantalone 2012; Zhang et al. 2018a).

Supplementation with natural sources or synthetic amino acids is used by industry to overcome amino acid deficiency in soybean meal which increases the feed cost (Imsande, 2001). QTLs have been reported for essential amino acids such as methionine, cysteine, lysine, and threonine (Panthee et al. 2006; Wang et al. 2015; Warrington et al. 2015). Recently, (Li et al. 2018) using 345 RILs observed twelve cluster QTLs on chromosomes 6, 9, 13, 17 and 20 affecting and explaining from 2.4-16.8% the phenotypic variation of 17 amino acids. Despite these studies, little improvement in soybean amino acid profiles in modern soybean cultivars has been made to improve animal feed value.

Most of the genetic architecture of the seed composition traits have been dissected using either traditional linkage analysis or genome-wide association analysis (GWAS). Nested association mapping (NAM), consisting of a multi-cross design with several families, where each family shares a common “hub” parent, improves the resolution and detection of rare alleles in the analysis (Yu et al. 2008). The NAM design has been used successfully to dissect complex traits in soybean such as flowering time (Li et al. 2017), canopy coverage (Xavier et al. 2017a), yield stability (Xavier et al. 2018), and grain yield, lodging and seed mass (Diers et al. 2018). All of these studies have largely used *G. max* germplasm for developing their respective populations and association mapping panels. Therefore, there is a need to understand the genetic architecture of value-added seed composition traits derived from wild soybean germplasm. Understanding the genetic inheritance behind seed composition traits and their effects are crucial for enhancing soybean value and market demands. The goal of this study was to identify QTLs associated with protein concentration, oil content, five fatty acids, and twelve amino

acids in a soybean NAM population from three *G. max* x *G. soja* crosses evaluate were evaluated across multi-environments.

Materials and methods

Germplasm and experimental design

Our NAM panel contains 392 F₄-derived RILs, developed from three unique biparental cross combinations between the *G. max* hub parent Williams 82 (Bernard and Cromeens 1988) and three *G. soja* plant introductions PI464890B, PI458536, and PI522226 maintained by the United States Department of Agriculture (USDA) in the Germplasm Resources Information Network – GRIN NPGS (<https://www.ars-grin.gov/>). Cross-pollination of the hub parent Williams 82 and each of the three *G. soja* plant introductions was conducted in Champaign-Urbana, IL during the summer of 2010, and the true F₁ hybrid plants were grown at the same environment in 2011 and bulk harvested. Approximately 10,000 F₂ plants were grown at the Bradford Research Center in Columbia, MO during the summer of 2012 for each of the three populations (30,000 total plants), and F₂ plants were bulk harvested. In 2013, approximately 100,000 F₃ plants were grown at the Bradford Research Center in Columbia, MO for each of the three populations (300,000 plants total), and approximately 500 single plants were harvested separately from each F₃ population (1500 total plants), the F₃ plants were selected based on agronomic traits, such as maturity, lodging, and shattering. The F_{3:4} plants rows were grown at the Bay Farm Research Facility in Columbia, MO during 2014, and a single plant was harvested separately from each row. The F_{4:5} plant rows were grown at the Bay Farm Research Facility in Columbia, MO during 2015, and each

plant row was bulk harvested for development of the F₄-derived RILs. Only lines with sufficient seed for broad testing were carried forward, and RILs were randomly chosen to represent an appropriate equal number from each of the three bi-parental to reduce the NAM panel size to a manageable field experiment.

During the growing seasons of 2016 and 2017, RILs were planted in four-row research plots with a row spacing of 76 cm, a row length of 3.6 m, and a 1.2 m fallow alley. The environments included the Hundley-Whaley Research Center in Albany, MO (40°14'53"N 94°19'51"W), the Bay Farm Research Facility in Columbia, MO (38°57'50"N 92°19'42"W), the Greenley Research Center in Novelty, MO (39°57'28.23"N 92°10'40.45"W) and the Graves-Chapple Research Center in Rock Port, MO (40° 16'18.09" N 95°28' 26.38"W). The field research plots were arranged in an augmented incomplete block experimental design with one replication in 2016 and two replications in 2017, with two check cultivars per block in both years. Genotypes were randomly assigned to a block and replication, and blocks were randomly ordered in the field across all environments. The experiment was divided into two field blocks at each environment for group MG III and MG IV RILs to accommodate combine harvesting across all environments.

Best linear unbiased estimators (BLUE) were used in the nested association mapping and linkage mapping according to the model adopted by (Jarquín et al. 2014): $y_{ijkl} = \mu + g_i + cb_l + l_j + gl_{ij} + r_{k(j)} + b_{l(k)} + e_{ijk}$, where μ is the trait mean, g_i is the genetic effect of the genotypes, cb_k is the interaction effect between the checks and incomplete blocks, l_j is the effect of environments, gl_{ij} is the interaction effect between the genotypes and environments, $r_{k(j)}$ is the effect of the replications nested in

environment, $b_{l(k)}$ is the incomplete block effect nested within replicate, and e_{ijkl} is the residual. Genotypes and environments were treated as fixed effects, the analysis was performed using the package lme4 in R (Bates et al. 2015). Broad-sense heritability (h^2) on an entry-mean basis was calculated as follows: $h^2 = \sigma_g^2 / (\sigma_g^2 + \sigma_{ge}^2/l + \sigma_e^2/r)$, where σ_g^2 , σ_{ge}^2 and σ_e^2 are the genetic variance, genotype x environment variance and error variance; l is the number of environments and r is the number of replications (Fehr 1991).

Phenotypic data collection

The two middle rows of the yield plots were harvested mechanically with an Almaco R1 or Almaco SPC40 plot combine (Nevada, Iowa). 50 mL of soybean seeds were randomly sampled from all the yield plots and were milled using a Perten laboratory Mill 3600 (Perten Instruments, Hägersten Sweden). In 2016, 3523 samples were analyzed via near-infrared spectroscopy using a Perten DA 7250 (Perten Instruments, Hägersten Sweden). Two hundred fifty samples that captured the broadest range in reflectance spectrum presented in the 2016 seed composition dataset (n= 3523 samples) were analyzed for seed composition traits via wet chemistry to help increase the accuracy and repeatability of the NIR equations used for prediction.

Analytical chemistry of crude protein, oil, fatty acids, and twelve amino acids was measured to include in the Perten calibrations. Samples were sent to the University of Missouri Agricultural Experiment Station Chemical Laboratory, University of Missouri (Columbia, MO) to measure amino acids and crude protein content. Crude protein was determined by combustion analysis (LECO) (AOAC 2006), and twelve amino acids were

determined by the hydrolysis method (Gehrke et al. 1987). The fatty acid profile of total oil, palmitic acid (C16:0), stearic acid (C18:0), oleic acid (C18:1), linoleic acid (C18:2), and linolenic acid (C18:3), were analyzed at the Bay Farm Research Facility in Columbia, MO following the procedure described by (Han et al. 2014), using an Agilent series 6890 capillary gas chromatograph (Palo Alto, CA). Perten Instruments and University of Minnesota technical staff helped to develop the NIR calibrations. The statistics of NIR calibrations are presented in Table 3-1. The coefficient of determination of cross-validation (R^2CV) ranged from 0.69 for methionine to 0.96 for oleic acid, with a standard error of prediction for the 250 Missouri samples (SEP-MO), ranged from 0.02 for methionine and 3.3 for linoleic acid.

Seed protein, oil, fatty acid, and amino acid concentrations were then estimated in all ground samples using near-infrared reflectance (NIR) spectroscopy Perten DA 7250 (Perten Instruments, Hägersten Sweden). Oil, protein, and amino acids were estimated on a dry weight basis, and fatty acids were determined as a percentage relative to the total oil content in the sample. The amino acids evaluated were, lysine (Lys), cysteine (Cys), methionine (Met), threonine (Thre), isoleucine (Ile), leucine (Leu), valine (Val), alanine (Ala), glutamic acid (Glu), glycine (Gly), and proline (Pro). The amino acid profiles were adjusted to the total protein concentration of the sample.

In our preliminary analysis was identified a QTL on chromosome 8, tagged by the SNP ss715602749 associated with most fatty acids and amino acids measured. (Figure 3-1), the QTL was right on top of the major locus responsible for black/yellow seed coat with pigmented hila (*I* locus) (Sonah et al. 2015; Zhou et al. 2015b; Fang et al. 2017). The *I* allele produces a yellow seed coat with clear hila; *i*ⁱ, yellow seed coat with

pigmented hila, and *i*, black seed coat with pigmented hila. The data supports that the QTL on Chromosome 8 was a NIR artifact of the anthocyanins presented in lines with black coat seeds. The NIR equations, in general, underestimated the fatty acid and amino acid data from black seed coat samples and exhibited a higher deviation when comparing yellow seed coat samples (Figure 3-1), probably due to the lack of black seed coat samples incorporated in the current equations. To eliminate the artifact QTL from the mapping analysis, all the 98 black seed coat recombinant inbred lines were removed, utilizing a total of 294 RILs.

Genotyping and SNP analysis

Genomic DNA was extracted from young trifoliate leaf tissue of 10 random plants for each RIL and the respective parents during the V4/V5 growth stage. A Qiagen Dneasy Plant 96 kit (QIAGEN, Valencia, CA) was used for high-quality DNA extraction. DNA samples were submitted to the Soybean Genomics and Improvement Laboratory, USDA-ARS, for genotyping using the Illumina Infinium BARCSoySNP6K BeadChip (Song et al. 2013). Genotypes were called using the software Genome Studio (Illumina, San Diego, California USA). SNPs with less than 90% call rate being excluded and SNPs were imputed using the LD-KNNi method implemented in version 5 of Tassel software (Bradbury et al. 2007). A total of 5,786 SNPs remained for subsequent quality control, association mapping, and linkage analysis. The R package 'ParentOffspring' (Abdel-Haleem et al. 2013) with a threshold of 80% of similarity between parents, and a principal component analysis using version 5 of the TASSEL software (Bradbury et al.

2007) was used to eliminate RILs that exhibited genotypes inconsistent with expected genetic similarity to the respective parents.

Only the filtered SNP marker data were used to measure the linkage disequilibrium (LD) in the NAM panel, using the sliding window method implemented in version 5 of TASSEL software. SNPs in LD were estimated using the pairwise squared correlation (r^2) and were plotted against SNP pair distances to obtain the rate of the LD decay in the heterochromatin and euchromatin region (Song et al. 2013).

Nested association mapping analysis

The nested association mapping was conducted using the package NAM (Xavier et al. 2015) developed from the SoyNAM project. The mixed linear model designed for multiple parent intercross populations was used for the SNP association (Wei and Xu 2016). The mixed linear model is described as follows: $y = \mu + X\alpha + g + e$, where μ is the intercept, X is the allele matrix from SNP data and family information, α is the SNP effects, g is the population structure effect, and e is the residual effect. A false discovery rate (FDR) threshold at $\alpha \leq 0.05$ level was used to declare SNP significant in the association mapping.

Linkage mapping analysis

We used the consensus genetic linkage map constructed by Song et al. (2016) based on 21,478 SNPs mapped in the Williams 82 x PI479752 (*G. soja*) population to define SNP positions. QTL mapping was performed using the composite interval mapping (CIM) functionality qtl package in R (Broman and Sen 2009). To fit multiple

QTL models, we used the function `fitqtl` which drops one QTL at a time. The likelihood of determination (LOD) for defining QTL significance was determined by a 1000 permutation test analysis at $\alpha \leq 0.05$ significant level and estimated confidence intervals of the environment of QTL via 1.5-LOD support intervals.

We searched for candidate genes, published genes, and published QTLs using the soybean reference genome available at SoyBase (<http://www.soybase.org>) in the Wm82.a2.v1 assembly. The SNPs and QTL intervals that were highly associated with traits were considered as a potential region for candidate genes. SoyBase was also used to search for QTLs previously found to affect the traits to date.

Results

Linkage Disequilibrium and Population Structure

Linkage disequilibrium (LD) decay varied among euchromatin and heterochromatin regions (Figure 3-2A). The LD decayed to half its maximum r^2 value (0.44) at an average 619.5 kb in euchromatin and 892.6kb in heterochromatin region. Since we estimated LD in experimental crosses, a greater LD is expected when comparing natural soybean populations (Zhou et al. 2015b), due to the limited amount of recombination and selection in the experimental crosses (Hyten et al. 2007). Principal component (PC) analysis was used to evaluate the variation in genotypic data of the NAM population (Figure 3-2B). The PC analysis could not distinguish the NAM population in sub-groups. PC1 explained 11.2% of the variation, while PC2 explained 8.4% of the genotypic variation.

Phenotypic variation

There were significant differences ($P < 0.001$) among the recombinant inbred lines for all seed composition traits evaluated (Table 3-2). Also, environmental effects and genotype x environment interaction were significant ($p < 0.05$) for protein, oil, fatty acids and the majority of amino acids. The entry-mean heritability (h^2) for protein and oil ranged from 0.87 to 0.97, 0.78 to 0.97 for fatty acids, and 0.69 to 0.89 for amino acids. These moderate to high heritability estimates suggest that the genetic variation accounted for a majority of the phenotypic variance for the seed composition traits

The histogram for all traits evaluated is presented in Figure 3-3. A normal frequency distribution was observed for all seed composition traits was observed. These results suggested the distribution of both positive and negative alleles among *Glycine soja* parents and Williams 82.

The well-known negative relationship between oil and protein was observed in this study (-0.90 , $p < 0.01$) (Figure 3-4). Oil was significantly correlated with linolenic acid (-0.76 , $p < 0.01$), palmitic acid (-0.50 , $p < 0.01$), and stearic acid (0.26 , $p < 0.05$). Oleic acid was significantly associated with palmitic acid, linoleic acid and linolenic acid (-0.53 , $p < 0.01$; -0.98 , $p < 0.01$; -0.57 , $p < 0.01$). Palmitic acid was significantly correlated with stearic acid, linoleic acid and linolenic acid (-0.33 , $p < 0.01$; 0.38 , $p < 0.01$; 0.62 , $p < 0.01$, respectively). Linoleic acid was significantly correlated with linolenic acid (0.42 , $p < 0.01$).

Except for the positive correlation between protein and glutamic acid (Glu) (0.21 , $p < 0.05$), seed protein was negatively associated with all the amino acids evaluated, ranging from -0.12 , $p < 0.05$ to -0.95 , $p < 0.01$ (Figure 3-4). However, a weak negative

correlation was detected between aspartic acid (Asp) and cysteine (Cys) with protein. All studied amino acids were positively correlated between themselves (0.12, $p < 0.05$ to 0.92, $p < 0.01$), except threonine (Thre) and glutamine (Glu) which were negatively correlated (-0.16, $p < 0.05$). Non-significant associations were observed between, lysine (Lys) and Glu, Lys and Glu, Cys and isoleucine (Ile), Cys and leucine (Leu), Cys and valine (Val), Ile and Glu, alanine (Ala) and Glu, and Glu and proline (Pro).

Nested Association mapping

The nested association mapping (NAM), revealed 61 genomic regions associated with seed composition traits (Table 3-3), and a total of seven regions showed pleiotropic effects (Table 1). Three QTLs were identified for protein, four QTLs for oil, six for fatty acids and 17 for amino acids.

A major pleiotropic QTL impacting protein, oil, Ala, Gly, Ile, Leu, Thr, Asp, Glu, Met, Val, and Lys was identified on chromosome 20 (ss715637176 and ss715637437) (Figures 3-5A-B, 3-6 and 3-7). It explained 28% of the variation for protein and oil (Figure 3-5C) and between 1-21% for amino acid composition (Figures 3-8 and 3-9). RILs with the *G. soja* allele had 28.5 g kg⁻¹ higher protein and 18.5 g kg⁻¹ lower oil compared to lines carrying *G. max* allele. However, the *G. soja* allele was associated with a reduction of 0.14, 0.41, 0.74, 0.60, 0.39, 1.77, 0.25 g kg⁻¹ for Thre, Leu, Ala, Gly, Lys, and Val respectively, and an increment of 0.11 g kg⁻¹ for Asp and 0.87 g kg⁻¹ for Glu.

The QTL on chromosome 15 (ss715621669) was also associated with protein and oil and explained 8% of the phenotypic variation. Lines that carried the *G. soja* allele exhibited 13.9 g kg⁻¹ higher protein and 9.4 g kg⁻¹ lower oil than the lines carrying *G.*

max allele. Both QTLs were consistently associated with the traits across environments in 2016 and 2017 (Table 3-3). SNP ss715637176 resulted in a protein ratio of 1.54 and ss715621669 of 1.47. A major QTL on Chr 19 tagged by the SNP ss715634618, was associated with Cys, Met, and Leu (Figure 3-6A, B, and C). This QTL accounted for 9-30% of the phenotypic variation for the amino acids profile. The lines carrying the *G. soja* allele (AA), had a significant reduction of 0.45 g kg⁻¹ in Cys and 0.34 g kg⁻¹ in Met, and an increment of 0.60 g kg⁻¹ in Leu comparing lines carrying the *G. max* allele (AA) (Figure 3-8A-B and Figure 3-9A).

Five QTLs on chromosomes 5, 11 and 15 were associated with the fatty acid profile (Figure 3-10A-E). The QTLs on Chr 5 (ss715591621) explained between 3-5% of phenotypic variation for palmitic acid, oleic acid, and linoleic acid. The lines with *G. soja* allele tagged for ss715591621 showed a reduction of 1.32 and 21.7 g kg⁻¹ in palmitic acid and linoleic acid, and 23.7 g kg⁻¹ higher oleic acid compared to lines with the *G. max* allele (Figures 3-10G, H, I). The QTL on chromosome 11 was associated with stearic acid and explained 11% of the phenotypic variation, and lines carrying the *G. soja* allele had a 4.05 g kg⁻¹ reduction in stearic acid (Figure 3-10F). A QTL for linolenic acid was identified on chromosome 15. Explaining 12% the phenotypic variation and lines carrying the *G. soja* allele had 9.35 g kg⁻¹ higher linolenic acid compared to lines with the *G. max* allele.

Linkage analysis

Twelve QTLs were associated with seed composition traits in the three biparental interspecific soybean populations (Table 3-4). Among them, two QTLs had pleiotropic

effects on multiple traits. A QTL on chromosome 20 was associated with protein and oil, and the LOD-1.5 interval overlapped QTLs associated with Thre, Ile, Leu, Ala, Gly. These QTLs explained 6.4 to 40.2% of the phenotypic variance, and their LOD score ranged from 5.89 to 14.6 (Table 3-4). In general, the QTL from *G. soja* showed a positive effect on protein and negative effect on oil and amino acids. QTL on chromosome 19 was associated with Cys and Met on multiple populations and explained between 12 and 15.6% of the phenotypic variance. The *G. soja* allele decreased Cys and Met compared with *G. max* allele.

Another pleiotropic QTL on chromosome 5 was associated with oleic and linolenic acid and showed increased oleic acid and decreased linoleic acid in multiple populations. These QTLs explained 7.3-18.4% of the phenotypic variance in these traits. QTL for linolenic acid was identified on chromosome 15 with positive effects on the traits from *G. soja* allele and explained 7.6-10% of the phenotypic variation for the trait. A QTL affecting stearic acid was observed in chromosome 11 for population 1 and explained 10.2% of the phenotypic variation; the *G. soja* allele had negative effects on stearic acid.

In general, linkage analysis confirmed the major QTLs identified in the NAM analysis, e.g., QTL on chromosome 20 was associated with protein, oil, and amino acids, QTL on chromosome 19 associated with amino acid composition, QTLs on chromosomes 5, 11 and 15 were associated with fatty acid profile.

Discussion

We detected two major QTLs with pleiotropic effects for protein and oil. One QTL located on chromosome 20 (ss715637176) and another on chromosome 15 (ss715621669), lines carrying the version of each *G. soja* allele had higher protein and lower oil content. The QTLs explain between 8 and 40% of the phenotypic variance depending on the mapping method and population. QTLs on chromosomes 15 and 20 were in the same region of confirmed QTLs cqSeed protein-001 on Chr 15 and cqSeed protein-003 on Chr 20 (Fasoula et al. 2004; Nichols et al. 2006). Also, they were mapped in multiple populations and sources in the same region (Diers et al. 1992; Brummer et al. 1997; Bandillo et al. 2015; Warrington et al. 2015; Phansak et al. 2016). These results verify previously reported QTL and provide confidence in our methods of population development, phenotypic data collection, and subsequent analysis for identifying QTL associated with protein and oil, amino acids and fatty acids.

We used three *G. soja* accessions for developing this interspecific NAM population and found positive allelic effects for protein and negative for oil for the two QTLs in all the sources, suggesting these alleles may be spread out across most *G. soja* accessions, or each locus could have several allele variants. Also, we observed a strong negative correlation between protein and oil ($r=-0.90$, $P<0.0001$), similar results were found in numerous studies (Chung et al. 2003; Hwang et al. 2014a; Phansak et al. 2016). These results suggested that breeding for oil and protein improvement simultaneously using QTLs on chromosomes 15 and 20 would be challenging due to the linkage between oil and protein QTLs.

The QTL region on chromosome 20 overlapped with QTLs associated with nine seed amino acids, explaining 4-21% of the phenotypic variation depending on the amino acid. The *G. soja* allele that conferred high protein (+28.5 g kg⁻¹) affected Ile, Leu, Thre, Ala, Gly, Lys, and Val negatively, and Asp and Glu positively. Similar results were found by (Warrington et al. 2015), where they observed an adverse effect from the protein QTL on Chr 20 (cqSeed protein-003) from Danbaekkong source of 1.1, 1.2, and 0.3 g kg⁻¹ in Cys, Thre, and Cys+Met, respectively. The negative effect of increased protein on essential amino acid composition was also observed in the phenotypic correlation values, where the protein was negatively correlated with all amino acids, except Glu. Similar results were observed by Warrington et al. (2015) and La et al. (2019). Pfarr et al. (2018) evaluated the protein effect on the amino acid profile and found that Lys, Met, Cys, and Thre were reduced with increased protein concentration, only Glu had increased. These results indicate that increasing protein content using this QTL is detrimental to the overall essential amino acid composition.

In general, amino acids showed a strong and positive correlation between themselves, likely because these groups of amino acids share the same biosynthesis pathway. For example, Leu, Ile, and Val have a common biosynthesis pathway (Singh 1999). Asp, Ile, Met, Lys, and Thre were in the aspartate metabolic pathway (Azevedo et al. 1997, 2006), and Cys is the sulfur donor in the biosynthesis pathway of Met (Saito 1999; Azevedo et al. 2006). The understanding and manipulation of these biosynthetic pathways is the key to improving amino acid composition in soybean.

Another QTL on chromosome 19 was associated with a reduction of Cys and Met in lines carrying the *G. soja* allele (GG), although, the GG allele increased Leu. The QTL

on Chr 19 overlapped with a region of a previously QTL identified by (Panthee et al. 2006), tagged by Satt313 and Satt076, and it was associated with Glu, Leu, Ile, and Tyr. The QTL on chromosome 19 was identified by correcting the amino acid profile by the protein concentration, thus, eliminating the protein impact on the discovery of new alleles for specific amino acid composition. Zhang et al. (2018), identified only four of the 92 amino acid QTLs using amino acids in dry weight and protein correction amino acids. La et al. (2019) evaluated a core collection of wild soybeans observed when amino acids are calculated based on the total seed dry weight; the amino acid contents increased as protein increased. However, the opposite occurs when amino acid content was measured based on the total protein content. Our results suggested that sulfur-containing and other essential amino acids concentrations can be enhanced independent from protein content.

Among several candidate genes associated with sulfur-containing amino acids (Cys and Met) on Chr 19, four candidate genes 5-20kb apart the most significant SNP (Table 3-5). Glyma.19g119200 encodes Cysteine synthase 26 (CS26), Glyma.19g120400 and Glyma.19g120600 encode an isopropylmalate synthase (IPMS2) involved in leucine biosynthesis, and Glyma.19g132000 encodes Cystathionine beta-lyase (CBL) involved in Met biosynthetic pathway. In *Arabidopsis*, IPMS1 and IPMS2 regulated the biosynthesis of leucine (de Kraker et al. 2006). IPMS1 mutant plants in rice, showed a reduction of 10%-50% in Leu, Ile, Phe, Thr, Cys, Gly, Val, and Pro during germination compared to the wild-type plants (He et al. 2018). Transgenic potato plants with 50% reduced CBL levels had a decrease in Met levels in leaves (Maimann et al. 2000). It is clear that IPMS1 and IMPS2 play an essential role in the biosynthesis of branched-chain amino acids, Leu, Ile and Val (Xing and Last 2017).

There is an increase in demand for low-cost plant protein sources, as protein consumption, especially animal-derived protein is on the rise worldwide (Henchion et al. 2017). Soybean can produce a meal with high protein concentration (50%) (Fischer et al. 2001), but the amino acid profile needs improvement to increase soybean meal value (Lagos and Stein 2017). Soybean is deficient in Met and Cys content which are required by monogastric animals (Krishnan et al. 2005; Wu 2014; Warrington et al. 2015). To overcome the deficiency, supplementation with synthetic amino acids is used by the animal industry, which increased the feed costs (Imsande 2001). Therefore, the improvement of amino acid balance in soybean cultivars would improve the economic value of the soybean and reduce the cost associated with supplementation. From a practical standpoint, if the objective was to increase absolute content of amino acids (dry basis), improving total soybean protein content is possible, but will decrease essential amino acid profile of soybean seeds. Our results suggest that sulfur-containing and other essential amino acids can be enhanced independent from protein content.

Five QTLs on chromosomes 5, 11 and 15 were associated with the fatty acid profile. A QTL on chromosome 5 was tagged by SNP ss715591621 and explained 8-13% of the phenotypic variance. It was associated with oil content, palmitic acid, oleic acid, and linoleic acid (Table 3-5), resulting in a reduction in palmitic and linoleic acid, and an increase in oleic acid for the lines carrying the *G. soja* allele. This QTL overlapped with a region previously associated with oil content, oleic acid and linolenic acid (Bandillo et al. 2015; Li et al. 2015). Among the candidate gene possibilities in this region is Glyma.05g218600 that encodes beta-ketoacyl-(acyl carrier protein) synthase I (KASI), associated with the novo fatty acid (FA) synthesis in *Arabidopsis* (Wu and Xue 2010).

KASI extends to C16 and KASII catalyzes from C16 to C18. We found a QTL associated with linolenic acid on chromosome 15, an area where previous QTLs affecting linolenic and oil have been reported (Bandillo et al. 2015; Zhang et al. 2018a). In this region 20kb from the most significant SNP is located Glyma.15g046300 that encodes a KCS7 (3-ketoacyl-CoA) involved in the biosynthesis of very long chain fatty acids. In *Arabidopsis* (Todd et al. 1999; Kim et al. 2013) and rice (Wang et al. 2017), 3-ketoacyl-CoA synthase family proteins have been reported to be involved in wax synthesis. Also, KCS enzymes have been characterized in *Arabidopsis* and were responsible for the extension of C18 to C20 or C22 in storage lipids (Kunst et al. 1992).

Linoleic and linolenic acid are more susceptible to oil oxidation compared to oleic acid causing a reduced shelf life (Kamal-Eldin 2006). Hydrogenation has been used to increase monosaturated fatty acids profile to reduce the unstable polyunsaturated portion in the soybean oil. However, during the process *trans* fatty are generated. Trans fatty acids have been associated with health risks, such as coronary heart disease (Kromhout 1999). High oleic acid (>80%) low linolenic acid (<3%) soybean oil has been successfully achieved by using a combination of fatty acid desaturase genes (Pham et al. 2010, 2012), and by transgenic or genomic editing approaches (Sweeney et al. 2017). We identified QTLs associated with fatty acid composition from *G. soja* on different genomic regions than generally reported in the literature (<http://www.soybase.org>, “SoyBase browser”, verified 01/20/2019). In similar studies (Li et al. 2015) using a mini-core collection of Chinese landraces and (Leamy et al. 2017) using *G. soja* accessions identified new regions associated with oleic acid, linoleic acid, and linolenic acid. These

results indicate that *G. soja* likely has several genes conferring unique fatty acid profiles that are yet to be utilized in a commercial breeding program

In this study, we identified 61 and 12 QTLs associated with seed composition traits by NAM analysis and linkage analysis, respectively. Most of the QTLs identified were reported in different studies and will be useful for a better understanding of the genetic architecture of seed composition traits, especially amino acids. Also, our results reinforce that increasing protein will not also increase the essential amino acid concentration and suggest that the sulfur-containing amino acids may be enhanced independently from protein. The identification of these markers provides information to facilitate further studies to improve protein, oil, fatty acids and amino acid composition in soybean seed as well as identify genes controlling these traits in wild soybean germplasm.

References

- Abdel-Haleem, H., P. Ji, H.R. Boerma, and Z. Li. 2013. An R package for SNP marker-based parent-offspring tests. *Plant Methods* 9(1): 44. doi: 10.1186/1746-4811-9-44.
- AOAC. 2006. AOAC official method 990.03, protein (crude) in animal feed, combustion method. *Official methods of analysis of AOAC International*. p. 30–31
- Azevedo, R.A., P. Arruda, W.L. Turner, and P.J. Lea. 1997. The biosynthesis and metabolism of the aspartate derived amino acids in higher plants. *Phytochemistry* 46(3): 395–419. <http://www.ncbi.nlm.nih.gov/pubmed/9332022>.
- Azevedo, R.A., M. Lancien, and P.J. Lea. 2006. The aspartic acid metabolic pathway, an exciting and essential pathway in plants. *Amino Acids* 30(2): 143–162. doi: 10.1007/s00726-005-0245-2.
- Bandillo, N., D. Jarquin, Q. Song, R. Nelson, P. Cregan, et al. 2015. A Population Structure and Genome-Wide Association Analysis on the USDA Soybean Germplasm Collection. *Plant Genome* 8(3). doi: 10.3835/plantgenome2015.04.0024.
- Bates, D., M. Mächler, B. Bolker, and S. Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Softw.* 67(1): 1–48. doi: 10.18637/jss.v067.i01.
- Bernard, R.L., and C.R. Cromeens. 1988. Registration of ‘Williams 82’ Soybean. *Crop Sci.* 28(6): 1027–1028. doi: 10.2135/cropsci1988.0011183X002800060049x.
- Bolon, Y.T., B. Joseph, S.B. Cannon, M.A. Graham, B.W. Diers, et al. 2010. Complementary genetic and genomic approaches help characterize the linkage group I seed protein QTL in soybean. *BMC Plant Biol.* 10(1): 41. doi: 10.1186/1471-2229-10-41.

- Bradbury, P.J., Z. Zhang, D.E. Kroon, T.M. Casstevens, Y. Ramdoss, et al. 2007. TASSEL: software for association mapping of complex traits in diverse samples.
- Broman, K.W., and S. Sen. 2009. A Guide to QTL Mapping with R/qtl. Springer, New York, NY.
- Brummer, E.C., G.L. Graef, J. Orf, J.R. Wilcox, and R.C. Shoemaker. 1997. Mapping QTL for Seed Protein and Oil Content in Eight Soybean Populations. *Crop Sci.* 37(2): 370. doi: 10.2135/cropsci1997.0011183X003700020011x.
- Chung, J., H.L. Babka, G.L. Graef, P.E. Staswick, D.J. Lee, et al. 2003. The seed protein, oil, and yield QTL on soybean linkage group I. *Crop Sci.* 43(3): 1053–1067. doi: 10.2135/cropsci2003.1053.
- Diers, B.W., W.R. Fehr, and R.C. Shoemaker. 1992. RFLP analysis of soybean seed protein and oil content. *Theor. Appl. Genet.* 83: 608–612. doi: 10.1007/BF00226905.
- Diers, B.W., J. Specht, K.M. Rainey, P. Cregan, Q. Song, et al. 2018. Genetic Architecture of Soybean Yield and Agronomic Traits. *G3.* 8(10): 3367–3375. doi: 10.1534/g3.118.200332.
- Fang, C., Y. Ma, S. Wu, Z. Liu, Z. Wang, et al. 2017. Genome-wide association studies dissect the genetic networks underlying agronomical traits in soybean. *Genome Biol.* 18(1): 161. doi: 10.1186/s13059-017-1289-9.
- Fasoula, V.A., D.K. Harris, and H.R. Boerma. 2004. Validation and designation of quantitative trait loci for seed protein, seed oil, and seed weight from two soybean populations. *Crop Sci.* 44(4): 1218–1226. doi:10.2135/cropsci2004.1218.

Fehr, W.R. 1991. Principles of cultivar development. Vol. 1. Theory and technique. Macmillan, New York.

Fischer, M., L. V Kofod, H.A. Schols, S.R. Piersma, H. Gruppen, et al. 2001. Enzymatic extractability of soybean meal proteins and carbohydrates: heat and humidity effects. *J. Agric. Food Chem.* 49(9): 4463–9.

Gehrke, C.W., P.R. Rexroad, R.M. Schisla, J.S. Absheer, and R.W. Zumwalt. 1987. Quantitative analysis of cystine, methionine, lysine, and nine other amino acids by a single oxidation--4 hour hydrolysis method. *J. Assoc. Off. Anal. Chem.* 70(1): 171–4.

Han, S.I., J.-H. Chae, K. Bilyeu, J.G. Shannon, and J.-D. Lee. 2014. Non-destructive Determination of High Oleic Acid Content in Single Soybean Seeds by Near Infrared Reflectance Spectroscopy. *J. Am. Oil Chem. Soc.* 91(2): 229–234. doi: 10.1007/s11746-013-2369-y.

He, Y., J. Cheng, Y. He, B. Yang, Y. Cheng, et al. 2018. Influence of isopropylmalate synthase OsIPMS1 on seed vigour associated with amino acid and energy metabolism in rice. *Plant Biotechnol. J.*: 1–16. doi: 10.1111/pbi.12979.

Henchion, M., M. Hayes, A. Mullen, M. Fenelon, and B. Tiwari. 2017. Future Protein Supply and Demand: Strategies and Factors Influencing a Sustainable Equilibrium. *Foods* 6(7): 53. doi: 10.3390/foods6070053.

Hwang, E.-Y., Q. Song, G. Jia, J.E. Specht, D.L. Hyten, et al. 2014a. A genome-wide association study of seed protein and oil content in soybean. *BMC Genomics* 15(1): 1. doi: 10.1186/1471-2164-15-1.

Hwang, E.-Y., Q. Song, G. Jia, J.E. Specht, D.L. Hyten, et al. 2014b. A genome-wide association study of seed protein and oil content in soybean. *BMC Genomics* 15(1): 1. doi: 10.1186/1471-2164-15-1.

Hyten, D.L., I.Y. Choi, Q. Song, R.C. Shoemaker, R.L. Nelson, et al. 2007. Highly variable patterns of linkage disequilibrium in multiple soybean populations. *Genetics* 175(4): 1937–1944. doi: 10.1534/genetics.106.069740.

Imssande, J. 2001. Selection of Soybean Mutants with Increased Concentrations of Seed Methionine and Cysteine. *Crop Sci.* 41(2): 510. doi: 10.2135/cropsci2001.412510x.

Jarquín, D., K. Kocak, L. Posadas, K. Hyma, J. Jedlicka, et al. 2014. Genotyping by sequencing for genomic prediction in a soybean breeding population. *BMC Genomics* 15(1): 740. doi: 10.1186/1471-2164-15-740.

Kamal-Eldin, A. 2006. Effect of fatty acids and tocopherols on the oxidative stability of vegetable oils. *Eur. J. Lipid Sci. Technol.* 108(12): 1051–1061. doi: 10.1002/ejlt.200600090.

Kim, J., J.H. Jung, S.B. Lee, Y.S. Go, H.J. Kim, et al. 2013. Arabidopsis 3-ketoacyl-coenzyme A synthase9 is involved in the synthesis of tetracosanoic acids as precursors of cuticular waxes, suberins, sphingolipids, and phospholipids. *Plant Physiol.* 162(2): 567–80. doi: 10.1104/pp.112.210450.

de Kraker, J.W., K. Luck, S. Textor, J.G. Tokuhisa, and J. Gershenzon. 2006. Two Arabidopsis Genes (IPMS1 and IPMS2) Encode Isopropylmalate Synthase, the Branchpoint Step in the Biosynthesis of Leucine. *PLANT Physiol.* 143(2): 970–986. doi: 10.1104/pp.106.085555.

Krishnan, H.B., J.O. Bennett, W.S. Kim, A.H. Krishnan, and T.P. Mawhinney. 2005. Nitrogen lowers the sulfur amino acid content of soybean (*Glycine max* [L.] Merr.) by regulating the accumulation of Bowman-Birk protease inhibitor. *J. Agric. Food Chem.* 53(16): 6347–6354. doi: 10.1021/jf050510i.

Kromhout, D. 1999. Fatty acids, antioxidants, and coronary heart disease from an epidemiological perspective. *Lipids* 34(S1): S27–S31. doi: 10.1007/BF02562225.

Kunst L., Taylor D. C., U.E.W. 1992. Fatty acid elongation in developing seeds of *Arabidopsis thaliana*. *Plant Physiol. Biochem.* 30(4): 425–434.

La, T., E. Large, E. Taliercio, Q. Song, J.D. Gillman, et al. 2019. Characterization of Select Wild Soybean Accessions in the USDA Germplasm Collection for Seed Composition and Agronomic Traits. *Crop Sci.* 59(0): 1–19. doi: 10.2135/cropsci2017.08.0514.

Lagos, L. V., and H.H. Stein. 2017. Chemical composition and amino acid digestibility of soybean meal produced in the United States, China, Argentina, Brazil, or India. *J. Anim. Sci.* 95(4): 1626–1636. doi: 10.2527/jas2017.1440.

Leamy, L.J., H. Zhang, C. Li, C.Y. Chen, and B.H. Song. 2017. A genome-wide association study of seed composition traits in wild soybean (*Glycine soja*). *BMC Genomics* 18(1): 18. doi: 10.1186/s12864-016-3397-4.

Lee, J.-D., K.D. Bilyeu, and J.G. Shannon. 2007. Genetics and Breeding for Modified Fatty Acid Profile in Soybean Seed Oil. *J. Crop Sci. Biotechnol.* 10(4): 201–210.

Li, S., Y. Cao, J. He, T. Zhao, and J. Gai. 2017. Detecting the QTL-allele system conferring flowering date in a nested association mapping population of soybean using a

novel procedure. *Theor. Appl. Genet.* 130(11): 2297–2314. doi: 10.1007/s00122-017-2960-y.

Li, Y., J.C. Reif, Y. Ma, H. Hong, Z. Liu, et al. 2015. Targeted association mapping demonstrating the complex molecular genetics of fatty acid formation in soybean. *BMC Genomics* 16(1): 841. doi: 10.1186/s12864-015-2049-4.

Li, X., R. Tian, S. Kamala, H. Du, W. Li, et al. 2018. Identification and verification of pleiotropic QTL controlling multiple amino acid contents in soybean seed. *Euphytica* 214(6): 93. doi: 10.1007/s10681-018-2170-y.

Maimann, S., C. Wagner, O. Kreft, M. Zeh, L. Willmitzer, et al. 2000. Transgenic potato plants reveal the indispensable role of cystathione β -lyase in plant growth and development. *Plant J.* 23(6): 747–758. doi: 10.1046/j.1365-313X.2000.00842.x.

Mounts, T.L., K. Warner, G.R. List, R. Kleiman, W.R. Fehr, et al. 1988. Effect of altered fatty acid composition on soybean oil stability. *J. Am. Oil Chem. Soc.* 65(4): 624–628. doi: 10.1007/BF02540691.

Nichols, D.M., K.D. Glover, S.R. Carlson, J.E. Specht, and B.W. Diers. 2006. Fine Mapping of a Seed Protein QTL on Soybean Linkage Group I and Its Correlated Effects on Agronomic Traits. *Crop Sci.* 46(2): 834. doi: 10.2135/cropsci2005.05-0168.

Pantalone, V. 2012. Designing Soybeans for 21st Century Markets. In: WILSON, R.F., editor, *Designing Soybeans for 21st Century Markets*. 1st ed. Academic Press and AOCS Press, Urbana, IL. p. 189–218

Panthee, D.R., V.R. Pantalone, C.E. Sams, A.M. Saxton, D.R. West, et al. 2006. Quantitative trait loci controlling sulfur containing amino acids, methionine and cysteine,

in soybean seeds. *Theor. Appl. Genet.* 112(3): 546–553. doi: 10.1007/s00122-005-0161-6.

Panthee, D.R., V.R. Pantalone, D.R. West, A.M. Saxton, and C.E. Sams. 2005. Quantitative Trait Loci for Seed Protein and Oil Concentration, and Seed Size in Soybean. *Crop Sci.* 45(5): 2015. doi: 10.2135/cropsci2004.0720.

Pfarr, M.D., M.J. Kazula, J.E. Miller-Garvin, and S.L. Naeve. 2018. Amino acid balance is affected by protein concentration in soybean. *Crop Sci.* 58(5): 2050–2062. doi: 10.2135/cropsci2017.11.0703.

Pham, A.T., J.D. Lee, J.G. Shannon, and K.D. Bilyeu. 2010. Mutant alleles of FAD2-1A and FAD2-1B combine to produce soybeans with the high oleic acid seed oil trait. *BMC Plant Biol.* 10(1): 195. doi: 10.1186/1471-2229-10-195.

Pham, A.T., J.G. Shannon, and K.D. Bilyeu. 2012. Combinations of mutant FAD2 and FAD3 genes to produce high oleic acid and low linolenic acid soybean oil. *Theor. Appl. Genet.* 125(3): 503–515. doi: 10.1007/s00122-012-1849-z.

Phansak, P., W. Soonsuwon, D.L. Hyten, Q. Song, P.B. Cregan, et al. 2016. Multi-Population Selective Genotyping to Identify Soybean [*Glycine max* (L.) Merr.] Seed Protein and Oil QTLs. *G3.* 6(6): 1635–1648. doi: 10.1534/g3.116.027656.

Priolli, R.H.G., J.B. Campos, N.S. Stabellini, J.B. Pinheiro, and N.A. Vello. 2015. Association mapping of oil content and fatty acid components in soybean. *Euphytica* 203(1): 83–96. doi: 10.1007/s10681-014-1264-4.

Rowntree, S.C., J.J. Suhre, N.H. Weidenbenner, E.W. Wilson, V.M. Davis, et al. 2013. Genetic Gain $\times$ Management Interactions in Soybean: I. Planting Date. *Crop Sci.* 53(3): 1128. doi: 10.2135/cropsci2012.03.0157.

Saito, K. 1999. Biosynthesis of cysteine. In: Singh B.K, editor, Plant Amino Acids: Biochemistry and Biotechnology. 1st ed. Marcel Dekker Inc, New York. p. 267–291

Singh, B.K. 1999. Biosynthesis of valine, leucine, and isoleucine. In: Singh, B.K., editor, Plant Amino Acids: Biochemistry and Biotechnology. Marcel Dekker Inc, New York. p. 227–247

Sonah, H., L. O'Donoghue, E. Cober, I. Rajcan, and F. Belzile. 2015. Identification of loci governing eight agronomic traits using a GBS-GWAS approach and validation by QTL mapping in soya bean. *Plant Biotechnol. J.* 13(2): 211–221. doi: 10.1111/pbi.12249.

Song, Q., D.L. Hyten, G. Jia, C. V. Quigley, E.W. Fickus, et al. 2013. Development and Evaluation of SoySNP50K, a High-Density Genotyping Array for Soybean (T. Zhang, editor). *PLoS One* 8(1): e54985. doi: 10.1371/journal.pone.0054985.

Song, Q., J. Jenkins, G. Jia, D.L. Hyten, V. Pantalone, et al. 2016. Construction of high resolution genetic linkage maps to improve the soybean genome sequence assembly Glyma1.01. *BMC Genomics* 17(1): 33. doi: 10.1186/s12864-015-2344-0.

Sweeney, D.W., M. Carrero-Colón, and K.A. Hudson. 2017. Characterization of New Allelic Combinations for High-Oleic Soybeans. *Crop Sci.* 57(2): 611. doi: 10.2135/cropsci2015.09.0596.

Todd, J., D. Post-Beittenmiller, and J.G. Jaworski. 1999. KCS1 encodes a fatty acid elongase 3-ketoacyl-CoA synthase affecting wax biosynthesis in *Arabidopsis thaliana*. *Plant J.* 17(2): 119–130. doi: 10.1046/j.1365-313X.1999.00352.x.

Vaughn, J.N., R.L. Nelson, Q. Song, P.B. Cregan, and Z. Li. 2014. The Genetic Architecture of Seed Composition in Soybean Is Refined by Genome-Wide Association Scans Across Multiple Populations. *G3*. 4(11): 2283–2294. doi: 10.1534/g3.114.013433.

Wang, X., Y. Guan, D. Zhang, X. Dong, L. Tian, et al. 2017. A β -Ketoacyl-CoA Synthase Is Involved in Rice Leaf Cuticular Wax Synthesis and Requires a CER2-LIKE Protein as a Cofactor. *Plant Physiol.* 173(2): 944–955. doi: 10.1104/pp.16.01527.

Wang, X., G.-L. Jiang, Q. Song, P.B. Cregan, R.A. Scott, et al. 2015. Quantitative trait locus analysis of seed sulfur-containing amino acids in two recombinant inbred line populations of soybean. *Euphytica* 201(2): 293–305. doi: 10.1007/s10681-014-1223-0.

Warrington, C. V., H. Abdel-Haleem, D.L. Hyten, P.B. Cregan, J.H. Orf, et al. 2015a. QTL for seed protein and amino acids in the Benning $\times$ Danbaekkong soybean population. *Theor. Appl. Genet.* 128(5): 839–850. doi: 10.1007/s00122-015-2474-4.

Warrington, C. V., H. Abdel-Haleem, D.L. Hyten, P.B. Cregan, J.H. Orf, et al. 2015b. QTL for seed protein and amino acids in the Benning $\times$ Danbaekkong soybean population. *Theor. Appl. Genet.* 128(5): 839–850. doi: 10.1007/s00122-015-2474-4.

Wei, J., and S. Xu. 2016. A Random-Model Approach to QTL Mapping in Multiparent Advanced Generation Intercross (MAGIC) Populations. *Genetics* 202(2): 471–486. doi: 10.1534/genetics.115.179945.

Weir, B.S., and C.C. Cockerham. 1984. Estimating F-Statistics for the Analysis of Population Structure. *Evolution* (N. Y). 38(6): 1358. doi: 10.2307/2408641.

Wilson, R.F. 2004. Soybeans: Improvement, Production, and Uses. In: Boerma, H.R. and Specht, J.E., editors. 3rd ed. ASA, CSSA, and SSSA, Madison, WI. p. 621–677

Wu, G. 2014. Dietary requirements of synthesizable amino acids by animals: a paradigm shift in protein nutrition. *J. Anim. Sci. Biotechnol.* 5(1): 34. doi: 10.1186/2049-1891-5-34.

Wu, G.-Z., and H.-W. Xue. 2010. Arabidopsis β -ketoacyl-[acyl carrier protein] synthase i is crucial for fatty acid synthesis and plays a role in chloroplast division and embryo development. *Plant Cell* 22(11): 3726–44. doi: 10.1105/tpc.110.075564.

Xavier, A., B. Hall, A.A. Hearst, K.A. Cherkauer, and K.M. Rainey. 2017. Genetic Architecture of Phenomic-Enabled Canopy Coverage in Glycine max. *Genetics* 206(2): 1081–1089. doi: 10.1534/genetics.116.198713.

Xavier, A., D. Jarquin, R. Howard, V. Ramasubramanian, J.E. Specht, et al. 2018. Genome-Wide Analysis of Grain Yield Stability and Environmental Interactions in a Multiparental Soybean Population. *G3*. 8(2): 519–529. doi: 10.1534/g3.117.300300.

Xavier, A., S. Xu, W.M. Muir, and K.M. Rainey. 2015. NAM: association studies in multiple populations. *Bioinformatics* 31(23): btv448. doi: 10.1093/bioinformatics/btv448.

Xing, A., and R.L. Last. 2017. A Regulatory Hierarchy of the Arabidopsis Branched-Chain Amino Acid Metabolic Network. *Plant Cell* 29(6): 1480–1499. doi: 10.1105/tpc.17.00186.

Yu, J., J.B. Holland, M.D. McMullen, and E.S. Buckler. 2008. Genetic design and statistical power of nested association mapping in maize. *Genetics* 178(1): 539–51. doi: 10.1534/genetics.107.074245.

Zhang, J., X. Wang, Y. Lu, S.J. Bhusal, Q. Song, et al. 2018. Genome-wide Scan for Seed Composition Provides Insights into Soybean Quality Improvement and the

Impacts of Domestication and Breeding. *Mol. Plant* 11(3): 460–472. doi: 10.1016/J.MOLP.2017.12.016.

Zhou, Z., Y. Jiang, Z. Wang, Z. Gou, J. Lyu, et al. 2015. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat. Biotechnol.* 33(4): 408–414. doi: 10.1038/nbt.3096.

Tables and Figures

SNP	Chr	Position	MAF	Allele	Traits	Genes.QTLs
ss715602749	8	8268861	0.27	A/C	Palm, Ste,Oleic,Met,Lys,Cys,Val,Pro,Asp,Glu	Seed_coat_color_/Hilum_color

Figure 3-1. Quantitative trait loci associated with fatty acid and amino acid composition. Mean separation of black seed coat (BI) and yellow (Y) seed coat for samples estimated using near-infrared reflectance (NIR) spectroscopy and samples estimated by analytical chemistry (Analytical) or wet lab for fatty acid (A) and amino acid composition (B). 16:0, palmitic acid; 18:0, stearic acid; 18:1, oleic acid; 18:2, linoleic acid; 18:3, linolenic acid; Cys, cysteine; Met, methionine.

Figure 3-2. Linkage disequilibrium (LD) decay of the nested association mapping (NAM) population (A), distribution of the NAM RILs under PC1 and PC2 (B).

Table 3-1. Descriptive statistics for NIR calibrations used for estimating 19 seed composition traits of ground soybean samples (updated 2017, Perten Instruments)

Parameter	Protein	Oil	Palmitic acid	Stearic acid	Oleic acid	Linoleic acid	Linolenic acid
N of samples	3790	3537	4341	3823	3930	3874	3942
Min [*]	33.44	9.92	4.36	2.07	13.27	0.40	0.80
Max [§]	53.55	26.47	18.94	6.33	89.59	66.70	15.37
SECV [#]	0.86	0.84	0.93	0.43	3.99	3.52	1.37
SEP [§]	1.25	1.00	0.80	0.36	2.51	3.30	0.65
R ² CV [±]	0.94	0.93	0.83	0.82	0.96	0.96	0.82
Parameter	Alanine	Aspartic Acid	Cysteine	Glutamic Acid	Glycine	Isoleucine	
N of samples	2557	2388	2565	2380	2388	2605	
Min	1.47	3.66	0.42	5.46	1.42	1.52	
Max	2.19	6.30	0.88	9.87	2.21	2.51	
SECV	0.04	0.12	0.04	0.22	0.04	0.07	
SEP	0.06	0.18	0.04	0.34	0.04	0.06	
R ² CV	0.90	0.93	0.75	0.92	0.92	0.83	
Parameter	Leucine	Lysine	Methionine	Proline	Threonine	Valine	
N of samples	2569	2515	2612	2384	2610	2574	
Min	2.54	2.19	0.40	1.54	1.25	1.57	
Max	4.18	3.40	0.72	2.66	2.01	2.72	
SECV	0.08	0.06	0.03	0.08	0.05	0.08	
SEP	0.11	0.09	0.02	0.09	0.05	0.09	
R ² CV	0.92	0.90	0.69	0.86	0.84	0.85	

* Min, minimum values of the prediction range; [§]Max, the maximum value of the prediction range; [#]SECV, Standard Error of Cross Validation; [§]SEP, Standard Error of Prediction University of Missouri samples; [±]R²CV, the coefficient of determination of cross-validation.
Data presented as percent of dry seed weight except for fatty acids as a percent of total oil.

Table 3-2. Mean square estimates of an analysis of variance for protein, oil, fatty acids, and amino acids during 2016 and 2017. ***, ** and * Significant at the 0.001, 0.01, and 0.05 probability level. h^2 , heritability on an entry-mean basis; SD, standard deviation; W82, mean of the hub parent (Williams 82).

2016									
Maturity group III							g kg ⁻¹		
	Genotype (G)	Environment (E)	Check x Block	G x E	Residual	h^2	Mean	W82	SD
Protein	1825.7***	12492***	616.7***	77.7*	59.7	0.95	438.4	418.9	23.4
Oil	1010.6***	333.08***	42.49***	22.77**	14.03	0.93	187.2	207.3	16.5
Palmitic Acid	62.3***	144.15***	12.41*	17.20***	6.02	0.72	113.8	110.4	5.3
Stearic Acid	208.8***	2193.7***	89.36***	12.38***	5.29	0.94	39.0	37.3	8.1
Oleic Acid	7857.6***	10570.8***	450.1***	365.5***	156.1	0.91	189.3	196.1	47.6
Linoleic Acid	5305.1***	11235.8***	225.6*	317.5***	143.6	0.88	555.7	561.9	39.9
Linolenic Acid	609.7***	1544.07***	40.9***	27.8***	8.27	0.83	92.8	79.5	13.3
Alanine	1.40***	4.37***	0.28**	0.38***	0.12	0.69	42.3	42.9	0.7
Aspartic Acid	2.9***	9.12***	0.62*	0.71*	0.47	0.72	109.4	109.7	1.1
Cysteine	0.1***	0.60***	0.18**	0.12**	0.07	0.87	11.6	12.0	0.6
Glutamic Acid	11.4***	47.85***	3.95**	1.9*	1.53	0.8	178.5	179.1	2.1
Glycine	1.05***	1.64***	0.07	0.09	0.07	0.74	41.9	42.5	0.6
Isoleucine	1.3***	2.53***	0.22	0.41**	0.16	0.68	48.1	48.6	0.7
Leucine	2.86***	1.86**	0.28	0.77**	0.33	0.72	77.0	77.9	1.0
Lysine	20.58***	340.82***	3.62	3.97*	2.18	0.79	71.2	74.3	2.7
Methionine	0.40***	2.53***	0.14***	0.09**	0.03	0.8	12.8	13.0	0.4
Proline	1.4***	5.38***	0.32*	0.17	0.17	0.83	50.4	51.0	0.7
Threonine	0.85***	12.3***	0.40***	0.1	0.13	0.81	36.7	37.2	0.6
Valine	0.98***	8.89***	0.53**	0.18	0.2	0.78	50.4	50.3	0.6
Maturity group IV							g kg ⁻¹		

	Genotype (G)	Environment (E)	Check x Block	G x E	Residual	h^2	Mean	W82	SD
Protein	1526.7***	2770.3***	1198.8***	339.4***	65.6	0.87	444.0	421.7	24.7
Oil	894.8***	738.4***	307.6***	134.6***	45.6	0.92	181.4	207.0	17.0
Palmitic Acid	105.1***	210.7***	34.8***	24.48	17.2	0.87	114.0	111.3	8.9
Stearic Acid	183.3***	549.1***	104.3***	24.3***	11.5	0.93	37.9	36.1	7.8
Oleic Acid	3175***	2112***	453***	539***	207	0.91	191.9	195.2	40.2
Linoleic Acid	3569.6***	31009.9***	518.8***	218.5	152.3	0.97	546.3	559.8	52.2
Linolenic Acid	613.5***	2081.5***	117.2***	37.53	25.84	0.97	98.7	80.1	14.4
Alanine	7.8***	17.8***	3.3	3.2	2.5	0.74	42.2	42.8	3.4
Aspartic Acid	7.1***	27.1***	2.41**	2.16***	1.2	0.82	108.9	109.7	1.5
Cysteine	11.9***	32.4***	3.3	3.22*	2.15	0.84	15.9	12.0	54.9
Glutamic Acid	20.4***	41.5***	5.1	6.5***	3.4	0.81	177.3	179.8	5.2
Glycine	11.6***	142.6***	3.5**	3.9***	2.1	0.80	41.9	42.5	4.7
Isoleucine	15.7***	59.9***	3.6	5.68***	2.5	0.78	48.4	48.5	5.9
Leucine	10.4***	86.1***	4**	3.5**	2.2	0.80	76.8	77.9	4.0
Lysine	16.3***	124.7***	2.4	3.8	3.1	0.87	76.1	73.4	54.4
Methionine	6.2***	30.6***	2.4**	2.17**	1.49	0.79	13.9	13.0	40.9
Proline	6.7***	57.1***	2.4*	2.12	1.2	0.81	50.4	50.8	2.9
Threonine	7.4***	61.6***	2.2	2.9	2.1	0.76	38.6	37.3	25.2
Valine	5.6***	22.9***	1.9**	1.8***	0.9	0.81	50.6	50.4	2.9

2017

	Maturity Group III					g kg ⁻¹			
	Genotype (G)	Environment (E)	Check x Block	G x E	Residual	h^2	Mean	W82	SD
Protein	4193***	67216***	1320***	145***	111.15	0.97	427.9	413.4	27.4
Oil	1617***	36595***	106***	44***	31.74	0.96	192.1	208.9	16.5
Palmitic Acid	114.4***	76.6***	38.9***	24.8**	19.5	0.86	112.1	109.4	5.8
Stearic Acid	796***	1325***	195***	18***	14.2	0.97	43.5	44.4	11.3

Oleic Acid	17558***	44971***	834***	524***	264	0.93	203.1	213.7	50.7
Linoleic Acid	12301***	28101***	974***	508***	289.66	0.85	543.4	541.9	44.1
Linolenic Acid	957***	1208***	86***	46***	33	0.89	92.0	82.1	12.5
Alanine	2.63***	4.88***	1.1***	0.14**	0.09	0.8	42.5	42.8	0.7
Aspartic Acid	3.83***	9.01***	2.76***	0.87**	0.74	0.77	108.3	108.9	1.1
Cysteine	2.42***	3.25***	0.47***	0.15**	0.12	0.89	11.7	12.2	0.7
Glutamic Acid	16.7***	76.2***	8.8***	2.1**	1.76	0.84	176.4	176.9	2.1
Glycine	1.7***	6.66***	0.56***	0.1**	0.07	0.86	42.0	42.4	0.6
Isoleucine	2.66***	1.53***	0.79***	0.23**	0.19	0.87	47.9	48.5	0.7
Leucine	3.87***	9.25***	0.95***	0.34***	0.26	0.89	77.0	77.5	0.9
Lysine	63.1***	80.9***	17.5***	3.7**	2.1	0.85	72.0	74.4	3.5
Methionine	1.05**	0.21**	0.19**	0.05*	0.04	0.84	12.8	13.1	0.4
Proline	4.4***	2.5***	0.78***	0.23	0.21	0.87	50.6	51.0	0.9
Threonine	1.23***	4.78***	0.81***	0.14***	0.09	0.86	36.8	37.1	0.6
Valine	1.57***	3.02***	1.03***	0.30*	0.27	0.85	50.1	50.2	0.7
----- Maturity Group IV ----- g kg ⁻¹ -----									
	Genotype (G)	Environment (E)	Check x Block	G x E	Residual	<i>h</i> ²	Mean	W82	SD
Protein	3006***	8858***	266***	71***	41.1	0.97	440.8	414.6	22.5
Oil	1540***	1223***	118***	27***	18.7	0.96	183.7	207.5	15.1
Palmitic Acid	101.1***	170.9***	37.2***	22.8**	20.2	0.87	113.2	109.6	5.7
Stearic Acid	699***	674***	92***	17***	13.8	0.97	39.6	44.0	10.4
Oleic Acid	6547***	4973***	11035***	491***	321.34	0.91	198.2	215.1	36.1
Linoleic Acid	3492***	9698***	6839***	440***	316.7	0.87	545.7	544.0	29.3
Linolenic Acid	966***	1685***	561***	45***	34.16	0.89	97.0	81.9	12.7
Alanine	1.98**	1.21**	0.46**	0.12*	0.11	0.81	42.1	42.8	0.6
Aspartic Acid	3.58**	2.98**	2.19**	1.01**	0.82	0.83	108.4	109.1	1.1

Cysteine	2.21***	1.02***	0.61***	0.15	0.14	0.89	11.6	12.0	0.6
Glutamic Acid	19***	66.6***	6.4***	2.6**	2.06	0.82	176.5	177.5	2.3
Glycine	1.79***	0.58***	0.28***	0.11**	0.09	0.78	41.7	42.4	0.6
Isoleucine	1.68***	1.93***	0.82***	0.24***	0.19	0.86	47.7	48.4	0.7
Leucine	3.44***	2.66***	1.68***	0.32*	0.29	0.88	76.7	77.6	0.8
Lysine	26***	515***	41***	3***	1.69	0.86	70.6	74.0	2.5
Methionine	0.81***	0.69***	0.07**	0.04	0.04	0.83	12.7	13.1	0.4
Proline	3.76***	10.92***	0.5***	0.26**	0.22	0.85	50.1	51.0	0.8
Threonine	0.82***	27.67***	0.26***	0.1**	0.08	0.87	36.7	37.1	0.5
Valine	1.22***	0.14***	1.48***	0.33	0.31	0.84	50.0	50.4	0.7

Figure 3-3. Histogram of the frequency distribution for seed composition traits in the NAM panel. Shown is the average of the measurements across eight environments during 2016 and 2017. Data presented as g kg^{-1} , amino acids concentration were adjusted for protein content.

Figure 3-4. Pearson's phenotypic correlation coefficients between soybean seed protein, oil, fatty acid, and amino acids concentration adjusted by protein content. Correlation is significantly different from zero when $P\text{-value} > 5.14 \times 10^{-3}$. A cross sign marks non-significant correlations.

Table 3-3. Significant SNPs for protein, oil, palmitic acid, stearic acid, linoleic acid, linolenic acid, oleic acid, alanine, aspartic acid, cysteine, glutamic acid, glycine, isoleucine, leucine, lysine, methionine, proline, threonine, and valine and their respective allelic effect in relation to the hub parent Williams 82 across 8 environments during 2016 and 2017. Chr, chromosome; Position, in base pair; MAF, minimum allele frequency;

Chr	SNP ID	Position	MAF	Allele	-log(10)P	PI464890B	PI458536	PI522226
PROTEIN								
2016								
15	ss715621669	3468596	0.19	G/T	4.4	0.20	-0.01	0.20
20	ss715637176	28389137	0.20	C/T	11.7	0.07	0.53	0.17
20	ss715637225	29594697	0.20	A/G	9.7	0.25	0.59	-0.17
20	ss715637220	29512680	0.19	A/G	9.6	0.17	0.38	0.15
20	ss715637033	23971966	0.21	A/C	9.3	0.13	0.38	0.18
20	ss715637052	24202089	0.20	A/G	8.6	0.17	0.32	0.20
20	ss715637150	27578013	0.21	T/C	8.2	0.05	0.41	0.19
20	ss715637169	28070832	0.15	A/G	8.2	0.17	0.58	-0.07
20	ss715637167	28017701	0.19	T/C	7.1	0.19	0.42	-0.03
20	ss715639021	6831960	0.29	T/C	7.1	0.17	0.58	-0.18
20	ss715637358	32603292	0.20	A/G	7.0	0.03	0.39	0.10
20	ss715636783	16909668	0.24	A/G	6.8	0.27	0.45	-0.16
20	ss715637229	29711410	0.11	T/C	6.7	0.00	0.63	-0.04
20	ss715636880	19781743	0.23	T/C	6.6	0.27	0.46	-0.17
20	ss715638954	4730717	0.24	T/G	6.6	0.05	0.34	0.21
20	ss715638934	4660681	0.23	C/T	6.6	0.06	0.30	0.23
20	ss715637171	28174755	0.12	C/A	6.5	0.00	0.63	-0.04
20	ss715637240	29983050	0.15	A/G	6.5	0.00	0.42	0.19
20	ss715637433	33215713	0.30	T/C	6.5	0.13	0.22	0.13
20	ss715636898	20207615	0.23	A/G	6.4	0.27	0.46	-0.17
20	ss715636906	20401020	0.24	T/C	6.4	0.27	0.46	-0.17
20	ss715637157	2773655	0.23	A/G	6.4	0.17	-0.02	0.32
20	ss715639074	8984320	0.24	T/G	6.3	0.28	0.47	-0.21
20	ss715637015	23666474	0.24	A/G	6.2	0.28	0.40	-0.15
20	ss715637127	26836040	0.14	G/T	6.1	0.32	0.00	0.24
20	ss715637552	3420379	0.17	A/G	6.0	0.08	0.19	0.31
20	ss715638583	4378776	0.15	T/C	5.9	0.17	0.13	0.25
20	ss715639080	9255340	0.22	C/T	5.9	0.29	0.40	-0.17
20	ss715637004	23452366	0.23	A/G	5.8	0.25	0.47	-0.20
20	ss715636534	10095001	0.23	T/C	5.8	0.29	0.41	-0.20
20	ss715636835	18531300	0.23	T/C	5.8	0.33	0.51	-0.37
20	ss715636844	18829763	0.23	G/A	5.8	0.33	0.52	-0.40
20	ss715638157	4018421	0.16	A/G	5.4	0.17	0.14	0.22
20	ss715639009	6499121	0.22	T/C	5.3	0.34	0.44	-0.35
20	ss715637319	3166447	0.16	A/C	5.3	0.18	0.07	0.26

20	ss715636976	22480124	0.25	T/C	5.3	0.30	0.33	-0.16
20	ss715636739	15908501	0.31	A/G	5.2	0.22	0.40	-0.14
20	ss715637222	2954372	0.25	G/T	5.1	0.15	0.15	0.12
20	ss715636989	22995309	0.23	A/G	4.7	0.28	0.34	-0.18
20	ss715639046	7863544	0.13	A/G	4.6	0.00	0.82	-0.42
20	ss715637185	28550287	0.37	G/A	4.6	0.05	0.21	0.12
20	ss715639057	8185857	0.34	T/C	4.6	0.23	0.27	-0.07
20	ss715637163	27884457	0.12	A/G	4.4	0.00	0.46	0.03
20	ss715637140	2724895	0.19	A/G	4.1	0.23	0.02	0.13
2017								
13	ss715615423	32254260	0.09	T/C	4.1	-0.15	0.65	-0.08
15	ss715621781	3845971	0.38	T/C	4.5	-0.14	0.23	0.22
20	ss715637176	28389137	0.20	C/T	9.7	0.11	0.58	0.06
20	ss715637225	29594697	0.20	A/G	7.5	0.28	0.72	-0.44
20	ss715637358	32603292	0.20	A/G	7.5	0.17	0.47	-0.07
20	ss715637169	28070832	0.15	A/G	7.5	0.23	0.80	-0.39
20	ss715637033	23971966	0.21	A/C	7.1	0.22	0.36	0.07
20	ss715637220	29512680	0.19	A/G	6.9	0.22	0.36	0.05
20	ss715637150	27578013	0.21	T/C	6.6	0.18	0.37	0.07
20	ss715637052	24202089	0.20	A/G	6.6	0.21	0.33	0.09
20	ss715639021	6831960	0.29	T/C	6.5	0.18	0.64	-0.23
20	ss715638954	4730717	0.24	T/G	6.5	0.00	0.38	0.26
20	ss715637229	29711410	0.11	T/C	6.3	0.00	0.81	-0.31
20	ss715638934	4660681	0.23	C/T	6.1	-0.01	0.28	0.37
20	ss715638583	4378776	0.15	T/C	5.9	0.14	0.11	0.37
20	ss715637171	28174755	0.12	C/A	5.8	0.00	0.79	-0.30
20	ss715637552	3420379	0.17	A/G	5.7	0.01	0.17	0.45
20	ss715637185	28550287	0.37	G/A	5.5	0.11	0.31	0.01
20	ss715639046	7863544	0.13	A/G	5.2	0.00	0.95	-0.56
20	ss715638157	4018421	0.16	A/G	5.0	0.09	0.13	0.34
20	ss715637319	3166447	0.16	A/C	5.0	0.10	0.07	0.38
20	ss715637167	28017701	0.19	T/C	4.9	0.27	0.46	-0.27
20	ss715636783	16909668	0.24	A/G	4.7	0.22	0.44	-0.14
20	ss715637222	2954372	0.25	G/T	4.6	0.14	0.17	0.12
20	ss715636880	19781743	0.23	T/C	4.6	0.22	0.44	-0.16
20	ss715637240	29983050	0.15	A/G	4.5	0.00	0.41	0.12
20	ss715636898	20207615	0.23	A/G	4.5	0.22	0.44	-0.16
20	ss715636906	20401020	0.24	T/C	4.5	0.22	0.44	-0.16
20	ss715637004	23452366	0.23	A/G	4.4	0.25	0.45	-0.22
20	ss715637380	32784352	0.23	A/G	4.4	0.27	0.22	-0.06
20	ss715637015	23666474	0.24	A/G	4.4	0.22	0.40	-0.15
20	ss715639074	8984320	0.24	T/G	4.3	0.23	0.45	-0.20
20	ss715636739	15908501	0.31	A/G	4.2	0.23	0.36	-0.13
20	ss715639080	9255340	0.22	C/T	4.0	0.23	0.37	-0.15

20	ss715636844	18829763	0.23	G/A	4.0	0.30	0.53	-0.45
20	ss715637157	2773655	0.23	A/G	4.0	0.16	0.03	0.22
20	ss715636989	22995309	0.23	A/G	4.0	0.26	0.37	-0.19
20	ss715636835	18531300	0.23	T/C	4.0	0.29	0.52	-0.42
20	ss715637126	26785339	0.30	T/C	4.0	0.14	0.19	0.03

OIL
2016

15	ss715621669	3468596	0.19	G/T	4.8	-0.28	0.09	-0.10
20	ss715637176	28389137	0.20	C/T	11.4	-0.12	-0.33	-0.08
20	ss715637169	28070832	0.15	A/G	9.8	-0.20	-0.48	0.19
20	ss715637033	23971966	0.21	A/C	8.8	-0.24	-0.17	-0.05
20	ss715637220	29512680	0.19	A/G	8.6	-0.24	-0.17	-0.05
20	ss715637225	29594697	0.20	A/G	8.6	-0.21	-0.40	0.19
20	ss715637358	32603292	0.20	A/G	6.8	-0.07	-0.32	0.03
20	ss715637150	27578013	0.21	T/C	6.5	-0.12	-0.19	-0.08
20	ss715637052	24202089	0.20	A/G	6.4	-0.20	-0.11	-0.09
20	ss715637167	28017701	0.19	T/C	6.4	-0.25	-0.21	0.09
20	ss715637229	29711410	0.11	T/C	5.9	0.00	-0.47	0.11
20	ss715637433	33215713	0.30	T/C	5.7	-0.16	-0.13	-0.01
20	ss715637171	28174755	0.12	C/A	5.6	0.00	-0.46	0.11
20	ss715637185	28550287	0.37	G/A	5.5	-0.10	-0.15	-0.04
20	ss715638954	4730717	0.24	T/G	5.2	-0.08	-0.14	-0.14
20	ss715637004	23452366	0.23	A/G	5.2	-0.27	-0.21	0.15
20	ss715639021	6831960	0.29	T/C	5.0	-0.13	-0.35	0.15
20	ss715638583	4378776	0.15	T/C	4.9	-0.12	-0.07	-0.15
20	ss715636898	20207615	0.23	A/G	4.9	-0.23	-0.20	0.10
20	ss715636906	20401020	0.24	T/C	4.9	-0.23	-0.20	0.10
20	ss715637157	2773655	0.23	A/G	4.9	-0.12	-0.02	-0.14
20	ss715636989	22995309	0.23	A/G	4.9	-0.28	-0.16	0.14
20	ss715637552	3420379	0.17	A/G	4.8	-0.03	-0.12	-0.22
20	ss715636783	16909668	0.24	A/G	4.7	-0.20	-0.20	0.08
20	ss715639074	8984320	0.24	T/G	4.6	-0.24	-0.21	0.14
20	ss715636880	19781743	0.23	T/C	4.6	-0.20	-0.20	0.09
20	ss715637015	23666474	0.24	A/G	4.5	-0.23	-0.15	0.08
20	ss715636976	22480124	0.25	T/C	4.5	-0.24	-0.16	0.11
20	ss715638934	4660681	0.23	C/T	4.4	-0.08	-0.11	-0.15
20	ss715639046	7863544	0.13	A/G	4.3	0.00	-0.57	0.33
20	ss715639080	9255340	0.22	C/T	4.2	-0.24	-0.15	0.11
20	ss715636534	10095001	0.23	T/C	4.2	-0.25	-0.16	0.13
20	ss715637127	26836040	0.14	G/T	4.1	-0.18	0.00	-0.12
20	ss715637126	26785339	0.30	T/C	4.0	-0.12	-0.06	-0.06

2017

5	ss715591873	39924642	0.18	C/T	5.4	-0.10	-0.04	-0.14
5	ss715591621	38231403	0.32	G/A	4.7	-0.14	-0.09	0.01

5	ss715591872	39858447	0.24	T/C	4.6	-0.10	-0.07	-0.06
5	ss715591623	38259743	0.31	A/G	4.1	-0.13	-0.07	-0.01
12	ss715612118	3175818	0.31	A/G	4.3	0.06	-0.23	-0.05
15	ss715621669	3468596	0.19	G/T	5.2	-0.25	0.05	-0.11
15	ss715621781	3845971	0.38	T/C	4.7	-0.01	-0.10	-0.10
15	ss715621715	3621773	0.27	C/T	4.2	0.00	-0.09	-0.13
20	ss715637176	28389137	0.20	C/T	7.9	0.03	-0.38	-0.05
20	ss715637169	28070832	0.15	A/G	7.1	-0.02	-0.60	0.27
20	ss715637358	32603292	0.20	A/G	7.0	-0.05	-0.38	0.10
20	ss715637229	29711410	0.11	T/C	6.0	0.00	-0.50	0.22
20	ss715637225	29594697	0.20	A/G	6.0	-0.07	-0.46	0.23
20	ss715638954	4730717	0.24	T/G	5.8	0.08	-0.27	-0.19
20	ss715637171	28174755	0.12	C/A	5.7	0.00	-0.51	0.22
20	ss715639046	7863544	0.13	A/G	5.7	0.00	-0.62	0.35
20	ss715637552	3420379	0.17	A/G	5.4	0.08	-0.15	-0.32
20	ss715638934	4660681	0.23	C/T	5.3	0.08	-0.21	-0.23
20	ss715638583	4378776	0.15	T/C	5.2	-0.03	-0.10	-0.22
20	ss715639021	6831960	0.29	T/C	4.9	0.00	-0.43	0.12
20	ss715637033	23971966	0.21	A/C	4.9	-0.06	-0.22	-0.05
20	ss715637220	29512680	0.19	A/G	4.7	-0.05	-0.22	-0.04
20	ss715637185	28550287	0.37	G/A	4.7	0.00	-0.22	-0.01
20	ss715637150	27578013	0.21	T/C	4.5	-0.03	-0.23	-0.05
20	ss715637052	24202089	0.20	A/G	4.3	-0.05	-0.21	-0.05
20	ss715638157	4018421	0.16	A/G	4.0	0.02	-0.12	-0.20

PALMITIC ACID

2017

5	ss715591651	38596037	0.19	C/A	4.85	0.04	-0.09	-0.01
5	ss715591623	38259743	0.31	A/G	4.68	0.03	-0.08	-0.01
5	ss715591621	38231403	0.32	G/A	3.91	0.03	-0.07	-0.01
20	ss715637319	3166447	0.16	A/C	4.1	0.02	0.03	0.04
20	ss715637052	24202089	0.20	A/G	3.99	0.05	0.02	0.02

STEARIC ACID

2017

8	ss715602862	9436305	0.08	A/G	6.38	0.00	-0.29	-0.09
8	ss715602663	7103134	0.18	G/A	6.16	-0.03	-0.32	0.07
8	ss715602667	7180829	0.19	T/G	6.16	-0.03	-0.32	0.07
8	ss715602749	8268861	0.27	A/C	5.84	-0.10	-0.19	0.00
8	ss715602778	8577294	0.18	A/G	5.31	-0.12	-0.17	0.05
8	ss715602662	7070019	0.18	G/A	5.25	-0.04	-0.29	0.06
8	ss715602804	8836971	0.14	C/T	4.61	-0.01	-0.35	0.16
8	ss715602765	8483831	0.15	T/C	4.6	-0.10	-0.18	0.04
8	ss715602796	8695745	0.17	A/C	4.43	0.06	-0.34	0.08
8	ss715602547	4959623	0.39	T/C	4.41	-0.02	-0.10	-0.04
8	ss715602691	7567751	0.14	G/A	4.33	0.00	-0.17	-0.09

8	ss715602729	7955652	0.19	G/A	4.27	-0.13	-0.22	0.14
8	ss715602756	8305965	0.27	T/C	4.17	-0.14	-0.12	0.04
8	ss715602760	8396392	0.19	A/G	4.15	-0.12	-0.11	0.01
8	ss715602635	6371819	0.07	C/T	4.05	0.00	-0.39	0.18
11	ss715608857	11633102	0.22	G/A	3.99	0.02	-0.11	-0.07

OLEIC ACID

2016

5	ss715591621	38231403	0.32	G/A	19.77	0.19	1.08	-0.01
5	ss715591623	38259743	0.31	A/G	17.95	0.22	0.98	0.02
5	ss715591651	38596037	0.19	C/A	15.47	0.14	0.96	0.08
5	ss715591873	39924642	0.18	C/T	8.33	0.00	0.82	0.04
5	ss715591872	39858447	0.24	T/C	7.91	0.16	1.02	-0.45
5	ss715591655	38636402	0.43	G/A	7.86	0.25	1.08	-0.55
5	ss715591714	39054172	0.50	G/A	6.8	0.00	-0.96	0.28
5	ss715591883	39995603	0.11	C/T	6.69	-0.20	0.95	0.00
5	ss715592199	41481303	0.44	T/C	5.77	0.00	0.49	0.15
5	ss715591636	38457343	0.12	G/T	5.21	0.49	0.00	0.36
5	ss715591659	38674350	0.27	C/T	4.79	-0.11	-1.45	1.08

2017

5	ss715591651	38596037	0.19	C/A	17.72	-0.05	0.95	0.19
5	ss715591621	38231403	0.32	G/A	16.72	0.12	0.88	0.02
5	ss715591623	38259743	0.31	A/G	15.47	0.12	0.84	0.03
5	ss715591714	39054172	0.50	G/A	10.34	0.14	-1.10	0.23
5	ss715591872	39858447	0.24	T/C	7.9	0.18	0.79	-0.30
5	ss715591873	39924642	0.18	C/T	7.61	0.07	0.63	0.04
5	ss715591655	38636402	0.43	G/A	5.85	0.40	0.60	-0.39
5	ss715591883	39995603	0.11	C/T	5.41	-0.40	0.87	0.00
5	ss715591636	38457343	0.12	G/T	5.21	0.30	0.00	0.45
5	ss715591954	40452052	0.40	G/T	4.54	-0.09	-0.76	0.38
5	ss715591728	39083568	0.21	A/G	4.13	0.45	-0.53	-0.47
5	ss715591767	39273308	0.26	T/G	4	0.53	-0.53	-0.43

LINOLEIC ACID

2016

5	ss715591621	38231403	0.32	G/A	17.85	-0.31	-0.83	0.02
5	ss715591623	38259743	0.31	A/G	16.11	-0.32	-0.74	-0.02
5	ss715591651	38596037	0.19	C/A	12.56	-0.28	-0.72	0.02
5	ss715591714	39054172	0.50	G/A	6.96	0.14	0.85	-0.38
5	ss715591873	39924642	0.18	C/T	6.58	-0.10	-0.60	-0.01
5	ss715591872	39858447	0.24	T/C	6.15	-0.17	-0.81	0.39
5	ss715591655	38636402	0.43	G/A	5.9	-0.21	-0.86	0.45
5	ss715591636	38457343	0.12	G/T	5.54	-0.62	0.00	-0.15
5	ss715591659	38674350	0.27	C/T	5.36	0.16	1.38	-1.07
5	ss715591883	39995603	0.11	C/T	5.16	0.06	-0.70	0.00
5	ss715592199	41481303	0.44	T/C	4.36	-0.06	-0.33	-0.14

2017								
5	ss715591621	38231403	0.32	G/A	17.68	-0.23	-0.67	-0.05
5	ss715591651	38596037	0.19	C/A	17.13	-0.06	-0.65	-0.28
5	ss715591623	38259743	0.31	A/G	15.21	-0.24	-0.59	-0.07
5	ss715591714	39054172	0.50	G/A	10.97	-0.04	0.94	-0.22
5	ss715591872	39858447	0.24	T/C	8.42	-0.23	-0.55	0.11
5	ss715591873	39924642	0.18	C/T	8.1	-0.18	-0.38	-0.15
5	ss715591636	38457343	0.12	G/T	7.23	-0.33	0.00	-0.46
5	ss715591728	39083568	0.21	A/G	4.95	-0.33	0.43	0.47
5	ss715591655	38636402	0.43	G/A	4.55	-0.36	-0.39	0.27
5	ss715591659	38674350	0.27	C/T	4.42	0.27	0.88	-0.81
5	ss715591883	39995603	0.11	C/T	4.39	0.17	-0.61	0.00
5	ss715591954	40452052	0.40	G/T	4.25	0.08	0.60	-0.25
5	ss715591767	39273308	0.26	T/G	4.2	-0.30	0.39	0.36

LINOLENIC ACID

2016								
3	ss715584618	18316	0.23	G/A	4.58	0.26	-0.14	0.12
15	ss715621669	3468596	0.19	G/T	7.9	0.29	-0.07	0.12
15	ss715622817	5097389	0.07	T/C	4.27	0.36	0.00	-0.10

2017								
15	ss715621669	3468596	0.19	G/T	8.81	0.36	-0.10	0.06
15	ss715622817	5097389	0.07	T/C	6.06	0.41	0.00	-0.15
15	ss715622843	5522858	0.29	T/C	5.49	0.27	0.01	-0.07
15	ss715621715	3621773	0.27	C/T	5.35	0.03	0.01	0.15
15	ss715622832	5312718	0.23	C/T	5.11	0.26	0.00	-0.06
15	ss715621402	2565153	0.30	C/T	4.89	0.07	-0.09	0.19
15	ss715622847	5638918	0.25	T/G	4.59	0.30	0.02	-0.14
15	ss715621420	2624165	0.31	G/T	4.48	0.07	-0.08	0.18
15	ss715621748	3727108	0.22	G/A	4.47	0.08	-0.15	0.21
15	ss715620312	1128092	0.12	G/A	4.33	0.12	0.08	0.03
15	ss715621452	2729636	0.36	A/C	4.3	0.02	-0.04	0.17
15	ss715622873	5924951	0.24	A/G	4.11	0.25	0.07	-0.15
20	ss715636662	135948	0.31	G/A	4.7	0.08	0.13	-0.07

ALANINE

2016								
5	ss715590982	33929183	0.32	C/T	4.48	0.00	-0.01	0.00
20	ss715638667	44728794	0.26	A/C	5.53	0.00	0.03	-0.02
20	ss715638703	44936870	0.43	T/C	5.27	0.00	0.03	-0.02
20	ss715638761	45394375	0.50	T/G	4.79	0.00	0.02	-0.02
20	ss715638718	45116741	0.24	C/A	4.28	-0.01	0.03	0.00
20	ss715638812	45724030	0.50	T/C	4.07	0.00	0.02	-0.01
20	ss715638700	44914896	0.41	G/A	3.98	0.00	0.02	-0.02

2017								
20	ss715637176	28389137	0.20	C/T	7.07	-0.03	-0.13	-0.01

20	ss715637033	23971966	0.21	A/C	6.86	-0.04	-0.13	0.01
20	ss715637169	28070832	0.15	A/G	6.8	-0.03	-0.26	0.12
20	ss715637220	29512680	0.19	A/G	6.55	-0.04	-0.14	0.02
20	ss715637052	24202089	0.20	A/G	6.48	-0.04	-0.14	0.01
20	ss715637552	3420379	0.17	A/G	6.41	0.03	-0.09	-0.12
20	ss715638062	3880320	0.06	T/C	6.38	0.00	-0.11	-0.10
20	ss715637150	27578013	0.21	T/C	6.3	-0.03	-0.14	0.02
20	ss715637225	29594697	0.20	A/G	6.05	-0.08	-0.17	0.12
20	ss715638583	4378776	0.15	T/C	5.95	-0.01	-0.10	-0.05
20	ss715638157	4018421	0.16	A/G	5.88	0.01	-0.10	-0.07
20	ss715637015	23666474	0.24	A/G	5.35	-0.05	-0.16	0.08
20	ss715637167	28017701	0.19	T/C	5.35	-0.05	-0.16	0.09
20	ss715636783	16909668	0.24	A/G	5.31	-0.04	-0.17	0.08
20	ss715636898	20207615	0.23	A/G	5.31	-0.05	-0.18	0.09
20	ss715636906	20401020	0.24	T/C	5.31	-0.05	-0.18	0.09
20	ss715636880	19781743	0.23	T/C	5.2	-0.05	-0.18	0.08
20	ss715639074	8984320	0.24	T/G	5.14	-0.05	-0.18	0.10
20	ss715637004	23452366	0.23	A/G	5.08	-0.05	-0.18	0.10
20	ss715636844	18829763	0.23	G/A	5.05	-0.06	-0.20	0.15
20	ss715636835	18531300	0.23	T/C	4.98	-0.06	-0.19	0.14
20	ss715638934	4660681	0.23	C/T	4.94	0.01	-0.11	-0.06
20	ss715637358	32603292	0.20	A/G	4.92	-0.04	-0.12	0.04
20	ss715638954	4730717	0.24	T/G	4.78	0.01	-0.09	-0.05
20	ss715637319	3166447	0.16	A/C	4.77	-0.01	-0.07	-0.06
20	ss715637157	2773655	0.23	A/G	4.71	0.00	-0.10	-0.03
20	ss715636739	15908501	0.31	A/G	4.67	-0.05	-0.14	0.07
20	ss715639021	6831960	0.29	T/C	4.58	-0.05	-0.15	0.07
20	ss715637240	29983050	0.15	A/G	4.58	0.00	-0.15	0.01
20	ss715639080	9255340	0.22	C/T	4.36	-0.06	-0.13	0.08
20	ss715636534	10095001	0.23	T/C	4.21	-0.06	-0.14	0.09
20	ss715639009	6499121	0.22	T/C	4.18	-0.07	-0.15	0.12
20	ss715637171	28174755	0.12	C/A	4.05	0.00	-0.19	0.07
20	ss715636976	22480124	0.25	T/C	4	-0.07	-0.12	0.07

ASPARTIC ACID

2017

1	ss715579021	3281127	0.44	A/G	4.35	0.06	-0.17	0.10
6	ss715594860	47631544	0.11	T/C	4.21	0.05	-0.28	0.00
8	ss715602862	9436305	0.08	A/G	5.45	0.00	-0.22	0.02
8	ss715602826	9111316	0.25	G/A	4.06	0.06	-0.23	0.08
13	ss715616840	5872811	0.09	G/A	4.38	0.00	-0.30	0.00
13	ss715614448	26198265	0.05	C/T	4.31	0.00	-0.44	0.31
14	ss715619471	48703687	0.01	T/G	4.38	0.00	-0.30	0.00
18	ss715629442	1848675	0.01	G/A	4.08	0.00	-0.27	0.00
18	ss715629620	1957770	0.07	T/C	4	0.18	-0.34	0.01

18	ss715630257	3219211	0.03	C/T	3.98	0.26	-0.42	0.00
19	ss715636383	8391354	0.12	C/A	4.38	0.00	-0.30	0.00
20	ss715637485	33676755	0.07	C/T	4.38	0.00	-0.30	0.00
20	ss715637704	35432652	0.03	C/A	4.38	0.00	-0.30	0.00
20	ss715637437	33275560	0.08	C/T	4.18	0.00	-0.31	0.04
20	ss715637450	33405047	0.10	C/T	4.11	-0.04	-0.31	0.12
20	ss715637459	33482879	0.11	C/T	4.05	0.00	-0.36	0.14

CYSTEINE

2017

8	ss715602662	7070019	0.18	G/A	6.23	0.03	-0.17	-0.01
8	ss715602663	7103134	0.18	G/A	6.07	0.03	-0.16	-0.02
8	ss715602667	7180829	0.19	T/G	6.07	0.03	-0.16	-0.02
8	ss715602749	8268861	0.27	A/C	5.95	-0.02	-0.07	-0.06
8	ss715602778	8577294	0.18	A/G	5.33	-0.02	-0.08	-0.03
8	ss715602862	9436305	0.08	A/G	5.27	0.00	-0.10	-0.11
8	ss715602786	8627848	0.15	A/G	5.21	0.00	-0.11	-0.04
8	ss715602796	8695745	0.17	A/C	4.88	-0.05	-0.13	0.04
8	ss715602729	7955652	0.19	G/A	4.83	-0.04	-0.12	0.03
8	ss715602635	6371819	0.07	C/T	4.81	0.00	-0.21	0.08
8	ss715602760	8396392	0.19	A/G	4.81	-0.02	-0.07	-0.04
8	ss715602804	8836971	0.14	C/T	4.76	-0.01	-0.13	0.01
8	ss715602756	8305965	0.27	T/C	4.52	-0.01	-0.07	-0.05
8	ss715602765	8483831	0.15	T/C	4.5	-0.01	-0.09	-0.03
8	ss715602773	8541523	0.15	G/T	4.47	0.02	-0.09	-0.04
8	ss715602691	7567751	0.14	G/A	4.11	0.00	-0.10	-0.02
19	ss715634600	37323698	0.18	A/G	11.98	-0.09	0.04	-0.10
19	ss715634557	37020042	0.18	A/G	11.36	-0.09	0.04	-0.08
19	ss715634591	37222096	0.15	A/C	11	-0.07	0.00	-0.09
19	ss715634618	37463936	0.16	A/G	10.52	-0.07	0.03	-0.10
19	ss715634623	37530198	0.17	A/G	10.51	-0.07	0.03	-0.10
19	ss715634642	37657631	0.26	T/C	9.81	-0.06	0.04	-0.11
19	ss715634541	36875523	0.20	A/G	8.53	-0.09	0.02	-0.05
19	ss715634697	37932358	0.29	C/T	8.25	-0.06	0.06	-0.11
19	ss715634738	38402678	0.22	A/G	7.79	-0.07	0.07	-0.11
19	ss715634685	37893995	0.27	G/A	7.68	-0.05	0.05	-0.11
19	ss715634777	38814028	0.21	T/C	7.28	-0.08	0.05	-0.07
19	ss715634728	38315356	0.15	A/G	7.26	-0.08	0.09	-0.10
19	ss715634477	36276581	0.20	G/A	7.24	-0.07	0.02	-0.06
19	ss715634720	38159634	0.29	G/A	6.96	-0.03	0.06	-0.12
19	ss715634789	38870732	0.13	T/G	6.85	-0.07	0.04	-0.07
19	ss715634465	36191998	0.21	G/T	6.75	-0.09	0.02	-0.04
19	ss715634792	38917571	0.15	A/G	6.7	-0.07	0.03	-0.07
19	ss715634768	38692980	0.12	A/G	6.63	-0.07	0.05	-0.08
19	ss715634765	38659685	0.14	C/T	6.33	-0.06	0.03	-0.08

19	ss715634802	39114852	0.23	G/A	6.12	-0.07	0.06	-0.08
19	ss715634723	38239219	0.36	A/C	6.06	-0.05	0.11	-0.12
19	ss715634452	36034992	0.10	A/G	5.38	-0.08	0.03	-0.07
19	ss715634485	36378440	0.27	T/C	5.29	-0.05	0.01	-0.05
19	ss715634808	39151328	0.12	C/T	5.1	-0.07	0.04	-0.06
19	ss715634285	35398444	0.09	T/G	4.64	-0.09	0.02	-0.04
19	ss715634759	38606452	0.10	C/A	4.58	-0.08	0.06	-0.08
19	ss715634198	34687566	0.31	A/G	4.05	-0.03	0.00	-0.04
19	ss715634366	35744912	0.17	C/T	3.99	-0.07	0.01	-0.02
19	ss715634496	36505181	0.13	A/G	3.93	-0.10	0.01	0.00

GLUTAMIC ACID

2016

1	ss715578859	27463902	0.14	G/T	4.01	-0.01	0.13	-0.06
7	ss715597259	34580678	0.09	G/A	6.44	-0.08	0.14	0.01
7	ss715597008	27998252	0.03	G/A	5.3	-0.10	0.16	0.00
19	ss715635181	42089062	0.23	C/T	4.58	-0.04	0.12	0.02
20	ss715638576	43713385	0.03	G/A	4.16	0.00	0.07	0.00

2017

5	ss715591623	38259743	0.31	A/G	4.13	-0.01	-0.11	-0.08
5	ss715591651	38596037	0.19	C/A	3.99	-0.01	-0.16	-0.04
5	ss715591621	38231403	0.32	G/A	3.94	-0.01	-0.11	-0.08
8	ss715602862	9436305	0.08	A/G	5.18	0.00	-0.46	-0.02
8	ss715602826	9111316	0.25	G/A	4.15	0.10	-0.41	0.03
13	ss715616840	5872811	0.09	G/A	4.18	0.00	-0.84	0.00
14	ss715619471	48703687	0.01	T/G	4.18	0.00	-0.84	0.00
19	ss715636383	8391354	0.12	C/A	4.18	0.00	-0.84	0.00
20	ss715637437	33275560	0.08	C/T	4.22	0.00	-0.80	0.02
20	ss715637485	33676755	0.07	C/T	4.18	0.00	-0.84	0.00
20	ss715637704	35432652	0.03	C/A	4.18	0.00	-0.84	0.00
20	ss715637459	33482879	0.11	C/T	3.97	0.00	-0.85	0.11

GLYCINE

2016

7	ss715597008	27998252	0.03	G/A	3.77	-0.02	0.04	0.00
20	ss715638667	44728794	0.26	A/C	3.76	0.00	0.03	-0.02
20	ss715638599	43967040	0.30	A/G	3.46	-0.01	0.02	0.00

2017

20	ss715638062	3880320	0.06	T/C	4.38	0.00	-0.05	-0.09
20	ss715637176	28389137	0.20	C/T	4.34	0.00	-0.09	-0.01
20	ss715637033	23971966	0.21	A/C	4.16	-0.01	-0.10	0.00
20	ss715637220	29512680	0.19	A/G	3.93	-0.01	-0.10	0.01
20	ss715637169	28070832	0.15	A/G	3.86	0.00	-0.18	0.07

ISOLEUCINE

2016

7	ss715597259	34580678	0.09	G/A	4.24	-0.03	0.04	0.01

17	ss715625733	10876251	0.18	A/C	3.92	0.01	-0.02	0.01
20	ss715638667	44728794	0.26	A/C	4.19	-0.01	0.03	-0.02
20	ss715638718	45116741	0.24	C/A	4.13	-0.01	0.03	-0.01
20	ss715638703	44936870	0.43	T/C	4.04	-0.01	0.03	-0.01
2017								
20	ss715637552	3420379	0.17	A/G	4.97	0.04	-0.06	-0.13
20	ss715637033	23971966	0.21	A/C	4.92	-0.04	-0.10	0.00
20	ss715637176	28389137	0.20	C/T	4.73	-0.02	-0.10	-0.01
20	ss715638062	3880320	0.06	T/C	4.66	0.00	-0.05	-0.13
20	ss715637052	24202089	0.20	A/G	4.65	-0.04	-0.08	-0.02
20	ss715637220	29512680	0.19	A/G	4.62	-0.04	-0.10	0.01
20	ss715637150	27578013	0.21	T/C	4.34	-0.03	-0.10	0.00
20	ss715638583	4378776	0.15	T/C	4.19	0.00	-0.06	-0.07
20	ss715638157	4018421	0.16	A/G	4.11	0.01	-0.06	-0.08
20	ss715637157	2773655	0.23	A/G	4.07	0.01	-0.06	-0.07
LEUCINE								
2016								
20	ss715637157	2773655	0.23	A/G	4.62	0.00	0.01	-0.03
20	ss715637240	29983050	0.15	A/G	3.49	0.00	0.00	-0.02
2017								
19	ss715634642	37657631	0.26	T/C	4.9	-0.01	0.05	0.11
19	ss715634618	37463936	0.16	A/G	4.5	0.00	0.03	0.12
19	ss715634623	37530198	0.17	A/G	4.48	-0.01	0.03	0.12
19	ss715634697	37932358	0.29	C/T	4	0.00	0.00	0.12
20	ss715637176	28389137	0.20	C/T	6.24	-0.05	-0.15	-0.01
20	ss715637052	24202089	0.20	A/G	6.21	-0.05	-0.14	-0.02
20	ss715637220	29512680	0.19	A/G	6.11	-0.06	-0.15	0.01
20	ss715637033	23971966	0.21	A/C	5.95	-0.05	-0.15	0.00
20	ss715637150	27578013	0.21	T/C	5.86	-0.05	-0.15	0.01
20	ss715637552	3420379	0.17	A/G	5.38	0.06	-0.10	-0.17
20	ss715638157	4018421	0.16	A/G	5.05	0.02	-0.10	-0.12
20	ss715638583	4378776	0.15	T/C	4.81	-0.02	-0.11	-0.07
20	ss715637169	28070832	0.15	A/G	4.8	-0.07	-0.22	0.10
20	ss715637167	28017701	0.19	T/C	4.68	-0.08	-0.18	0.12
20	ss715638062	3880320	0.06	T/C	4.4	0.00	-0.07	-0.16
20	ss715638934	4660681	0.23	C/T	4.38	0.00	-0.11	-0.07
20	ss715638954	4730717	0.24	T/G	4.37	0.00	-0.12	-0.06
20	ss715637240	29983050	0.15	A/G	4.36	0.00	-0.16	-0.02
20	ss715637225	29594697	0.20	A/G	4.31	-0.08	-0.21	0.16
20	ss715636751	1621036	0.26	T/C	4.12	-0.02	-0.22	0.20
20	ss715637358	32603292	0.20	A/G	4.1	-0.03	-0.13	0.02
20	ss715637157	2773655	0.23	A/G	4.05	0.00	-0.07	-0.08
20	ss715636739	15908501	0.31	A/G	3.91	-0.04	-0.19	0.07

LYSINE

2016								
13	ss715617177	7733811	0.15	C/T	5.49	0.00	0.00	0.06
13	ss715617244	8232824	0.15	G/T	4.49	-0.01	0.00	0.07
13	ss715617240	8168052	0.19	C/T	4.28	0.00	0.00	0.06
2017								
2	ss715581101	12244605	0.28	A/G	5.04	0.77	-2.38	1.04
2	ss715581097	12168684	0.24	T/G	4.89	0.76	-2.39	1.05
4	ss715588903	47718975	0.38	A/C	4.34	0.45	-1.26	0.58
13	ss715616840	5872811	0.09	G/A	10.82	0.00	-3.17	0.00
13	ss715613769	11787225	0.22	C/T	10.65	1.30	-4.83	2.27
13	ss715614448	26198265	0.05	C/T	9.65	0.00	-4.10	1.99
13	ss715615458	32712123	0.05	C/A	9.62	0.00	-2.79	0.74
13	ss715613775	12045535	0.16	T/C	9.46	2.06	-4.14	0.00
13	ss715613795	12724562	0.11	T/C	9.46	4.36	-8.39	0.00
13	ss715615399	32011502	0.05	T/C	9.34	3.47	-8.80	1.95
13	ss715615057	297095	0.17	A/G	8.77	1.49	-4.19	1.74
13	ss715616465	42150006	0.15	A/G	4.76	1.59	-3.92	1.15
13	ss715613823	13613605	0.08	C/T	4.69	0.27	-1.58	0.00
13	ss715613854	14588417	0.08	G/A	4.69	0.27	-1.58	0.00
13	ss715613889	16365275	0.09	C/T	4.65	0.39	-1.64	0.00
13	ss715613721	10330320	0.19	T/G	4.38	0.50	-1.89	1.20
13	ss715613882	16096547	0.05	C/T	4.24	1.47	-3.59	0.00
13	ss715613987	19040172	0.06	G/A	4.24	1.47	-3.59	0.00
13	ss715616094	38249824	0.32	T/C	4.22	0.09	-0.42	-0.19
13	ss715615381	31860090	0.02	G/A	4.17	1.44	-2.68	0.00
15	ss715620529	13014539	0.11	T/C	5.2	1.45	-3.06	0.84
15	ss715623122	8263547	0.23	G/T	4.34	0.26	-1.15	0.80
20	ss715637459	33482879	0.11	C/T	11.02	0.00	-5.06	3.82
20	ss715637485	33676755	0.07	C/T	10.82	0.00	-3.17	0.00
20	ss715637704	35432652	0.03	C/A	10.82	0.00	-3.17	0.00
20	ss715637437	33275560	0.08	C/T	10.78	0.00	-4.77	3.29
20	ss715637964	37410040	0.12	T/C	9.96	0.00	-4.00	1.78
20	ss715638506	43146832	0.12	A/G	9.72	1.24	-4.88	2.40
20	ss715638568	43631496	0.13	T/C	9.72	1.24	-4.88	2.40
20	ss715638352	41722165	0.07	C/T	9.55	1.97	-4.10	0.00
20	ss715637982	37746786	0.12	C/T	9.35	2.10	-4.16	0.00
20	ss715637450	33405047	0.10	C/T	8.97	1.84	-4.59	1.24
20	ss715638769	45460065	0.23	T/C	8.51	1.63	-4.45	1.22
20	ss715637606	34474878	0.24	T/C	6.21	1.29	-3.37	0.77
20	ss715637735	35625615	0.28	C/T	5.43	0.95	-1.88	-0.02
20	ss715637480	33634187	0.15	G/A	4.75	1.22	-3.70	1.08
20	ss715638154	40166186	0.08	C/A	4.25	0.91	-1.80	0.00
20	ss715638583	4378776	0.15	T/C	4.22	-0.04	-0.51	-0.10

METHIONINE

2016								
6	ss715594079	28816949	0.43	A/G	4.42	0.00	0.00	0.01
6	ss715594302	37364558	0.43	A/C	4.32	0.00	0.00	0.01
6	ss715594119	30534657	0.50	G/A	4.31	0.00	0.00	0.01
6	ss715594029	26981990	0.47	G/T	4.25	0.00	0.00	0.01
6	ss715593853	20370075	0.43	T/C	3.94	0.00	0.00	0.00
7	ss715597259	34580678	0.09	G/A	4.75	-0.01	0.02	0.00
2017								
19	ss715634618	37463936	0.16	A/G	16.32	-0.09	0.03	-0.06
19	ss715634623	37530198	0.17	A/G	16.07	-0.09	0.03	-0.06
19	ss715634557	37020042	0.18	A/G	15.8	-0.08	0.02	-0.06
19	ss715634600	37323698	0.18	A/G	15.67	-0.08	0.02	-0.06
19	ss715634642	37657631	0.26	T/C	14.96	-0.08	0.05	-0.07
19	ss715634591	37222096	0.15	A/C	14.46	-0.07	0.00	-0.05
19	ss715634697	37932358	0.29	C/T	13.56	-0.08	0.07	-0.09
19	ss715634541	36875523	0.20	A/G	12.72	-0.09	0.02	-0.04
19	ss715634477	36276581	0.20	G/A	12.62	-0.07	0.04	-0.06
19	ss715634738	38402678	0.22	A/G	12.29	-0.09	0.06	-0.07
19	ss715634728	38315356	0.15	A/G	12.27	-0.08	0.05	-0.06
19	ss715634685	37893995	0.27	G/A	11.76	-0.07	0.05	-0.08
19	ss715634768	38692980	0.12	A/G	11.52	-0.08	0.04	-0.05
19	ss715634792	38917571	0.15	A/G	10.81	-0.09	0.03	-0.04
19	ss715634789	38870732	0.13	T/G	10.67	-0.09	0.03	-0.04
19	ss715634777	38814028	0.21	T/C	10.46	-0.10	0.05	-0.04
19	ss715634720	38159634	0.29	G/A	10.41	-0.06	0.06	-0.08
19	ss715634723	38239219	0.36	A/C	10.3	-0.07	0.09	-0.09
19	ss715634765	38659685	0.14	C/T	10.25	-0.07	0.02	-0.05
19	ss715634802	39114852	0.23	G/A	8.82	-0.10	0.05	-0.04
19	ss715634465	36191998	0.21	G/T	8.79	-0.10	0.04	-0.03
19	ss715634485	36378440	0.27	T/C	8.46	-0.06	0.03	-0.04
19	ss715634759	38606452	0.10	C/A	8.46	-0.09	0.04	-0.05
19	ss715634452	36034992	0.10	A/G	8.08	-0.08	0.03	-0.06
19	ss715634808	39151328	0.12	C/T	8.05	-0.09	0.03	-0.02
19	ss715634774	38732345	0.08	T/G	7.52	-0.10	0.04	-0.03
19	ss715634366	35744912	0.17	C/T	6.91	-0.07	0.02	-0.03
19	ss715634285	35398444	0.09	T/G	6.78	-0.09	0.02	-0.03
19	ss715634173	34472334	0.12	A/C	5.68	-0.08	0.03	-0.03
19	ss715634509	36641660	0.14	G/A	5.36	-0.12	0.04	0.01
19	ss715634523	36691284	0.10	C/T	5.28	-0.16	0.06	0.09
19	ss715634310	35549926	0.16	A/G	5.27	-0.08	0.03	-0.01
19	ss715634479	36313784	0.17	T/C	5.24	0.01	0.01	-0.07
19	ss715634496	36505181	0.13	A/G	4.98	-0.10	0.03	0.01
19	ss715634144	34243974	0.11	T/C	4.96	-0.07	0.02	-0.02
19	ss715634505	36588644	0.09	A/G	4.67	-0.15	0.06	0.08

19	ss715634198	34687566	0.31	A/G	4.66	-0.04	0.01	-0.03
19	ss715634152	34339992	0.14	A/G	4.2	-0.06	0.02	-0.03
19	ss715634117	34006819	0.14	T/G	4.18	-0.07	0.01	-0.01
19	ss715634225	34924899	0.15	G/A	4.15	-0.08	0.03	-0.01
19	ss715634184	34600226	0.18	T/C	4.04	-0.08	0.03	-0.01
19	ss715634713	38121212	0.19	G/A	4.04	-0.12	0.08	0.00
19	ss715634147	34320762	0.13	A/C	3.92	-0.05	0.01	-0.03
PROLINE								
2016								
13	ss715616332	40623113	0.28	T/C	4.08	0.01	0.01	0.00
2017								
8	ss715602804	8836971	0.14	C/T	4.77	0.00	-0.20	0.03
8	ss715602796	8695745	0.17	A/C	4.46	0.00	-0.22	0.06
8	ss715602773	8541523	0.15	G/T	4.36	-0.01	-0.12	-0.03
8	ss715602635	6371819	0.07	C/T	3.99	0.00	-0.27	0.12
THREONINE								
2016								
6	ss715594148	31863080	0.10	C/T	4.05	0.00	0.00	0.01
6	ss715594302	37364558	0.43	A/C	4.02	0.00	0.00	0.01
11	ss715610530	38261784	0.48	G/A	4.58	-0.01	0.00	0.02
13	ss715617177	7733811	0.15	C/T	4.12	0.00	0.00	0.01
17	ss715626052	13438444	0.09	G/A	4.35	0.00	0.00	0.01
20	ss715636783	16909668	0.24	A/G	4.65	0.00	-0.01	0.00
20	ss715637052	24202089	0.20	A/G	4.64	0.00	0.00	-0.01
20	ss715639074	8984320	0.24	T/G	4.63	0.00	-0.01	0.00
20	ss715636739	15908501	0.31	A/G	4.57	0.00	0.00	0.00
20	ss715639021	6831960	0.29	T/C	4.43	0.00	0.00	0.00
20	ss715636534	10095001	0.23	T/C	4.41	0.00	0.00	0.00
20	ss715639080	9255340	0.22	C/T	4.22	0.00	0.00	0.00
20	ss715637004	23452366	0.23	A/G	4.15	0.00	-0.01	0.00
20	ss715636880	19781743	0.23	T/C	4.03	0.00	-0.01	0.00
20	ss715636898	20207615	0.23	A/G	4.01	0.00	-0.01	0.00
20	ss715636906	20401020	0.24	T/C	4.01	0.00	-0.01	0.00
20	ss715639057	8185857	0.34	T/C	3.95	0.00	0.00	0.00
20	ss715637033	23971966	0.21	A/C	3.92	0.00	0.00	0.00
20	ss715637015	23666474	0.24	A/G	3.91	0.00	0.00	0.00
2017								
20	ss715637176	28389137	0.20	C/T	7.85	-0.03	-0.10	0.02
20	ss715637220	29512680	0.19	A/G	7.71	-0.04	-0.11	0.02
20	ss715637033	23971966	0.21	A/C	7.59	-0.03	-0.11	0.01
20	ss715637150	27578013	0.21	T/C	7.58	-0.03	-0.11	0.02
20	ss715637052	24202089	0.20	A/G	7.14	-0.03	-0.10	0.01
20	ss715638157	4018421	0.16	A/G	6.78	0.00	-0.07	-0.04
20	ss715637225	29594697	0.20	A/G	6.7	-0.05	-0.13	0.09

20	ss715637167	28017701	0.19	T/C	6.59	-0.05	-0.13	0.08
20	ss715638583	4378776	0.15	T/C	6.51	-0.02	-0.08	-0.02
20	ss715637552	3420379	0.17	A/G	6.5	0.01	-0.07	-0.06
20	ss715638934	4660681	0.23	C/T	5.92	-0.01	-0.08	-0.03
20	ss715637319	3166447	0.16	A/C	5.84	-0.02	-0.06	-0.03
20	ss715637169	28070832	0.15	A/G	5.74	-0.04	-0.14	0.08
20	ss715638954	4730717	0.24	T/G	5.56	-0.01	-0.08	-0.02
20	ss715636844	18829763	0.23	G/A	5.24	-0.05	-0.15	0.13
20	ss715637240	29983050	0.15	A/G	5.17	0.00	-0.12	0.01
20	ss715636835	18531300	0.23	T/C	5.1	-0.04	-0.14	0.12
20	ss715637222	2954372	0.25	G/T	4.97	-0.03	-0.07	0.02
20	ss715639021	6831960	0.29	T/C	4.96	-0.03	-0.12	0.07
20	ss715636739	15908501	0.31	A/G	4.95	-0.03	-0.12	0.07
20	ss715637358	32603292	0.20	A/G	4.85	-0.03	-0.08	0.03
20	ss715637015	23666474	0.24	A/G	4.83	-0.03	-0.12	0.07
20	ss715638062	3880320	0.06	T/C	4.82	0.00	-0.07	-0.05
20	ss715636783	16909668	0.24	A/G	4.79	-0.03	-0.13	0.07
20	ss715637157	2773655	0.23	A/G	4.77	-0.02	-0.05	-0.02
20	ss715636880	19781743	0.23	T/C	4.76	-0.03	-0.13	0.07
20	ss715637229	29711410	0.11	T/C	4.7	0.00	-0.14	0.06
20	ss715637241	30048849	0.08	G/T	4.68	0.00	-0.08	-0.03
20	ss715639009	6499121	0.22	T/C	4.67	-0.05	-0.12	0.11
20	ss715636898	20207615	0.23	A/G	4.59	-0.03	-0.13	0.07
20	ss715636906	20401020	0.24	T/C	4.59	-0.03	-0.13	0.07
20	ss715639074	8984320	0.24	T/G	4.56	-0.03	-0.13	0.08
20	ss715637004	23452366	0.23	A/G	4.51	-0.03	-0.13	0.07
20	ss715636534	10095001	0.23	T/C	4.38	-0.04	-0.11	0.07
20	ss715637171	28174755	0.12	C/A	4.28	0.00	-0.14	0.06
20	ss715637163	27884457	0.12	A/G	4.27	0.00	-0.14	0.06
20	ss715639080	9255340	0.22	C/T	4.15	-0.04	-0.11	0.07
20	ss715639057	8185857	0.34	T/C	4.14	-0.04	-0.09	0.06

VALINE

2016

5	ss715590584	30966423	0.11	G/T	4.12	0.00	0.06	-0.02
8	ss715602602	5701842	0.19	C/T	4.86	0.00	0.03	0.00
17	ss715628156	7513295	0.05	G/A	4.26	0.00	0.04	0.01
20	ss715638703	44936870	0.43	T/C	6.17	-0.01	0.05	-0.02
20	ss715638718	45116741	0.24	C/A	6.01	-0.02	0.05	-0.02
20	ss715638667	44728794	0.26	A/C	5.86	-0.01	0.05	-0.02
20	ss715638576	43713385	0.03	G/A	5.74	0.00	0.03	0.00
20	ss715638682	44830703	0.40	G/T	5.22	-0.01	0.05	-0.02
20	ss715638812	45724030	0.50	T/C	5.21	-0.02	0.05	-0.02
20	ss715638903	46314790	0.25	C/T	4.99	-0.01	0.05	-0.02
20	ss715638700	44914896	0.41	G/A	4.61	-0.01	0.04	-0.02

20	ss715638795	45603015	0.18	C/A	4.52	-0.01	0.05	-0.02
20	ss715638175	40486756	0.34	A/G	4.23	-0.01	0.03	0.00
20	ss715638624	44446670	0.11	G/A	3.97	0.00	0.03	-0.01
2017								
2	ss715582348	42073855	0.16	G/A	4.7	-0.06	-0.27	0.15
8	ss715602862	9436305	0.08	A/G	6.21	0.00	-0.18	0.02
10	ss715607096	4103498	0.07	G/T	6.12	0.00	-0.45	0.15
13	ss715615458	32712123	0.05	C/A	7.64	0.00	-0.34	0.07
13	ss715616840	5872811	0.09	G/A	7.62	0.00	-0.38	0.00
13	ss715615399	32011502	0.05	T/C	6.65	0.45	-1.01	0.15
13	ss715614448	26198265	0.05	C/T	6.57	0.00	-0.48	0.23
13	ss715613775	12045535	0.16	T/C	6.42	0.23	-0.48	0.00
13	ss715613795	12724562	0.11	T/C	6.29	0.50	-0.98	0.00
13	ss715613769	11787225	0.22	C/T	5.86	0.15	-0.55	0.24
13	ss715615057	297095	0.17	A/G	5.48	0.13	-0.48	0.21
16	ss715624888	36218124	0.23	A/G	5.34	0.15	-0.52	0.18
18	ss715630257	3219211	0.03	C/T	7.07	0.28	-0.51	0.00
18	ss715630025	2493998	0.03	C/T	5.75	0.32	-1.05	0.36
18	ss715629620	1957770	0.07	T/C	5.45	0.17	-0.36	0.03
18	ss715629442	1848675	0.01	G/A	5.35	0.00	-0.28	0.00
20	ss715637485	33676755	0.07	C/T	7.62	0.00	-0.38	0.00
20	ss715637704	35432652	0.03	C/A	7.62	0.00	-0.38	0.00
20	ss715637459	33482879	0.11	C/T	7.07	0.00	-0.49	0.25
20	ss715637437	33275560	0.08	C/T	6.96	0.00	-0.47	0.21
20	ss715637964	37410040	0.12	T/C	6.83	0.00	-0.47	0.20
20	ss715638352	41722165	0.07	C/T	6.71	0.20	-0.47	0.00
20	ss715637982	37746786	0.12	C/T	6.35	0.26	-0.49	0.00
20	ss715638506	43146832	0.12	A/G	6.11	0.15	-0.57	0.27
20	ss715638568	43631496	0.13	T/C	6.11	0.15	-0.57	0.27
20	ss715638769	45460065	0.23	T/C	5.3	0.16	-0.51	0.16
20	ss715638954	4730717	0.24	T/G	4.4	0.03	-0.12	-0.01
20	ss715637606	34474878	0.24	T/C	4.38	0.16	-0.42	0.10
20	ss715638157	4018421	0.16	A/G	4.28	0.03	-0.10	-0.03
20	ss715636739	15908501	0.31	A/G	4.08	0.01	-0.17	0.08
20	ss715639021	6831960	0.29	T/C	4.07	0.01	-0.17	0.08
20	ss715637735	35625615	0.28	C/T	4.06	0.11	-0.25	0.01
20	ss715637552	3420379	0.17	A/G	3.98	0.03	-0.10	-0.03

Figure 3-5. Manhattan plot of SNPs and associated LOD score using the NAM panel for protein (A) and oil (B) across eight environments during 2016 (red dots) and 2017 (blue dots). The significant SNPs were distinguished by the FDR 0.05 (3.8×10^{-5}) threshold line. Mean separation of genotypic groups for significant SNP's associated with protein and oil on lead SNP ss715621669 (C), and ss715637176 (D) over eight environments. The least significance difference (LSD), differences in mean (Δm), the probability of the greater F statistic ($Pr > F$), and R^2 are in the figures.

Figure 3-6. Manhattan plot of SNPs and associated LOD score using the NAM panel for cysteine (A), methionine (B), isoleucine (C), leucine (D), glycine (E), and alanine (F), across eight environments during 2016 (red dots) and 2017 (blue dots). The significant SNPs were distinguished by the FDR 0.05 (3.8×10^{-5}) threshold line.

Figure 3-7. Manhattan plot of SNPs and associated LOD score using the NAM panel for aspartic acid (A), glutamic acid (B), lysine (C), proline (D), threonine (E), and valine (F), across eight environments during 2016 (red dots) and 2017 (blue dots). The significant SNPs were distinguished by the FDR 0.05 (3.8×10^{-5}) threshold line.

Figure 3-8. Mean separation of genotypic groups of RILs for significant SNP's associated with cysteine (A), methionine (B), isoleucine (C), threonine (D), leucine (E), alanine (F), glycine (G), and lysine (H) over eight environments. The least significance difference (LSD), differences in mean (Δm), the probability of the greater F statistic ($Pr>F$), and R^2 are in the figures.

Figure 3-9. Mean separation of genotypic groups of RILs for significant SNP's associated with leucine (A), aspartic acid (B), glutamic acid (C), and valine (D) over eight environments. The least significance difference (LSD), differences in mean (Δm), the probability of the greater F statistic ($Pr>F$), and R^2 are in the figures.

Figure 3-10. Manhattan plot of SNPs and associated LOD score using the NAM panel for stearic acid (A), palmitic acid (B), oleic acid (C), linoleic acid (D), and linolenic acid (E), across eight environments during 2016 (red dots) and 2017 (blue dots). The significant SNPs were distinguished by the FDR 0.05 (3.8×10^{-5}) threshold line. Mean separation of genotypic groups for significant SNPs associated with fatty acids composition (F, G, H, I, J) over eight environments. The least significance difference (LSD), differences in mean (Δm), the probability of the greater F statistic ($Pr > F$), and R^2 are in the figures.

A)

PROTEIN - ALBANY

PROTEIN - COLUMBIA

PROTEIN - NOVELTY

PROTEIN - ROCK PORT

B)

OIL - ALBANY

OIL - COLUMBIA

OIL - NOVELTY

OIL - ROCK PORT

c)

PALMITIC ACID - ALBANY

PALMITIC ACID - COLUMBIA

PALMITIC ACID - NOVELTY

PALMITIC ACID - ROCK PORT

D)

STEARIC ACID - ALBANY

STEARIC ACID - COLUMBIA

STEARIC ACID - NOVELTY

STEARIC ACID - ROCK PORT

E)

OLEIC ACID - ALBANY

OLEIC ACID - COLUMBIA

OLEIC ACID - NOVELTY

OLEIC ACID - ROCK PORT

F)

LINOLEIC ACID - ALBANY

LINOLEIC ACID - COLUMBIA

LINOLEIC ACID - NOVELTY

LINOLEIC ACID - ROCK PORT

G)

LINOLENIC ACID - ALBANY

LINOLENIC ACID - COLUMBIA

LINOLENIC ACID - NOVELTY

LINOLENIC ACID - ROCK PORT

H)

ALANINE - ALBANY

ALANINE - COLUMBIA

ALANINE - NOVELTY

ALANINE - ROCK PORT

I)

ASPARTIC ACID - ALBANY

ASPARTIC ACID - COLUMBIA

ASPARTIC ACID - NOVELTY

ASPARTIC ACID - ROCK PORT

J)

CYSTEINE - ALBANY

CYSTEINE - COLUMBIA

CYSTEINE - NOVELTY

CYSTEINE - ROCK PORT

K)

GLUTAMIC ACID - ALBANY

GLUTAMIC ACID - COLUMBIA

GLUTAMIC ACID - NOVELTY

GLUTAMIC ACID - ROCK PORT

L)

GLYCINE - ALBANY

GLYCINE - COLUMBIA

GLYCINE - NOVELTY

GLYCINE - ROCK PORT

M)

ISOLEUCINE - ALBANY

ISOLEUCINE - COLUMBIA

ISOLEUCINE - NOVELTY

ISOLEUCINE - ROCK PORT

N)

LEUCINE - ALBANY

LEUCINE - COLUMBIA

LEUCINE - NOVELTY

LEUCINE - ROCK PORT

o)

LYSINE - ALBANY

LYSINE - COLUMBIA

LYSINE - NOVELTY

LYSINE - ROCK PORT

P)

METHIONINE - ALBANY

METHIONINE - COLUMBIA

METHIONINE - NOVELTY

METHIONINE - ROCK PORT

Q)

PROLINE - ALBANY

PROLINE - COLUMBIA

PROLINE - NOVELTY

PROLINE - ROCK PORT

R)

THREONINE - ALBANY

THREONINE - COLUMBIA

THREONINE - NOVELTY

THREONINE - ROCK PORT

Figure 3-11. Manhattan plot of SNPs and associated LOD score using NAM panel for protein (A), oil (B), palmitic acid (C), stearic acid (D), oleic acid (E), linoleic acid (F), linolenic acid (G), alanine (H), aspartic acid (I), cysteine (J), glutamic acid (K), glycine (L), isoleucine (M), leucine (N), lysine (O), methionine (P), proline (Q), threonine (R), and valine (S) in individuals environments (Albany, Columbia, Novelty and Rock Port) during 2016 (red dots) and 2017 (blue dots). The significant SNPs were distinguished by the FDR 0.05 (3.8×10^{-5}) threshold line.

Table 3-4. Significant QTL results in three populations (Pop1- Williams 82 x PI464890B; Pop2 – Williams 82 x PI458536; Pop3 – Williams 82 x PI522226) for 19 seed composition traits across eight environments in two years (2016 and 2017).

	Year	QTL	Chr	Peak marker	Position (cM)	1.5 Interval	LOD	R ² (%)	Add effect*
PROTEIN									
Pop1	2016	qPRO-20	20	20207615	32.87	30.1-41.8	7.8	14.7	0.85
	2017		20	20401020	32.87	29.6-37.9	8.1	12.3	0.78
Pop2	2016		20	29594697	33.752	0-37	13.3	40.2	1.96
	2017		20	20207615	32.87	0-37.72	13.4	37.8	1.86
Pop3	2016		20	26500747	33.52	26.3-51.3	10.5	33.9	1.12
	2017		20	26500747	33.52	6.3-35.3	7.6	16.9	1.09
OIL									
Pop1	2016	qOIL-20	20	20207615	37.89	24.8-42.5	6.5	28.3	-0.34
Pop2	2016		20	29594697	33.752	0-38	14.6	38.7	-1.49
	2017		20	20207615	32.87	0-37	13.3	35.4	-1.16
Pop3	2016		20	26500747	33.52	26.3-34.3	7.8	37.9	-0.87
	2017		20	26500747	33.52	26.3-34.3	7.5	26.9	-0.62
STEARIC ACID									
Pop2	2017	qSTE-11	11	2625440	12.9	3.1-25.5	7.1	10.20%	-0.13
OLEIC ACID									
Pop1	2016	qOLE-5	5	38596037	119.6	101.4-121.45	7.2	18.4	0.24
Pop2	2016		5	40044590	112.1	98.2-119.37	12.6	18.2	0.19
Pop2	2017		5	40044590	112.1	98.2-119.37	11.2	18.2	0.2
Pop3	2016		5	35152904	81.09	75.6-102	6.3	8.8	0.08
LINOLEIC ACID									
Pop1	2017	qLIN-5	5	38596037	119.6	102.3-121.45	8.1	18.4	-0.22
Pop2	2016		5	40044590	112.1	98.2-119.37	9.9	17.4	-0.2
Pop2	2017		5	40044590	112.1	98.2-119.37	12	18.2	-0.2
Pop3	2016		5	35152904	81.09	75.6-102	8.7	8.8	-0.07
Pop3	2017		5	35152904	81.09	75.6-102	6.9	7.3	-0.08

LINOLENIC ACID									
Pop1	2017		15	3621773	16.4	9.4-33.8	6.9	10.2	0.09
Pop3	2016	qLINL-15	15	4315169	19.4	4.5-32.2	9	7.6	0.08
Pop3	2017		15	4406751	20	4.5-32.2	8.3	7.7	0.1
CYSTEINE									
Pop1	2017	qCYS-19	19	37222096	45.7	37.3-54.9	11	7.8	-0.15
Pop3	2017		19	37932358	50	35.6-52.6	9.9	6.5	-0.04
METHIONINE									
Pop1	2017		19	37463936	47.6	39.3-52.9	12.8	7.8	0.15
Pop2	2017	qMET-19	19	30103637	31.4	17.5-39.1	12.4	15.8	-0.27
Pop3	2017		19	37932358	50.9	38.754.9	15.6	6.5	0.04
THREONINE									
Pop1	2017		20	12318232	32.2	29.6-37.9	7.1	8.7	-0.08
Pop2	2016	qTHR-20	20	35378782	56	33.4-62.4	7.5	6.4	-0.06
Pop2	2017		20	20207615	32.2	15.3-38.7	7.2	22.1	-0.25
ISOLEUCINE									
Pop1	2017	qILE-20	20	12318232	32.2	25.6-37.9	6	8.7	-0.05
LEUCINE									
Pop3	2017	qLEU-20	20	20884990	29.07	12.1-43.4	5.9	6.5	-0.1
ALANINE									
Pop2	2017	qALA-20	20	19781743	32.7	0-33.4	11.8	22	-0.25
Pop3	2017		20	14721991	32.4	0-37.3	7.2	6.5	-0.1
GLYCINE									
Pop2	2017	qGLY-20	20	19781743	32.7	0-33.7	8.4	22.1	-0.23

*Additive effect represents the impact of converting homozygous *G. max* allele to homozygous *G. soja* allele.

Table 3-5. Major effect QTL and respective candidate genes associated with seed composition traits.

SNP	Chr ^s	Position [#]	MAF [*]	Allele	Traits	QTLs	Candidate Gene	Annotation
ss715591621	5	38231403	0.32	G/A	Palmitic, Oleic, Linoleic, Oil	Linolenic 4-g5, oleic 3-g1, oil 4-g14-g18(Bandillo et al. 2015; Li et al. 2015; Zhang et al. 2018a)	Glyma.05g218600	3-ketoacyl-acyl carrier protein synthase I (KASI) fatty acid synthesis
ss715608857	11	11633102	0.22	G/A	Stearic	Seed long-chain fatty acid 1-g5.1-g5.3(Fang et al. 2017)		
ss715621669	15	3468596	0.19	G/T	Linolenic , Oil, Protein	Linolenic 3g2-g3, protein 3-g12-g15, oil 4-g9-g12 (Bandillo et al. 2015; Zhang et al. 2018a)	Glyma.15g046300	3-ketoacyl-CoA synthase 7 (KCS7) fatty acid biosynthetic process
							Glyma.15g052500	fatty acid biosynthetic process
ss715634618	19	37463936	0.16	A/G	Cys, Met, Leu	Satt313, Satt076 - Glu1-5, Leu1-5, Tyr1-5, Ile1-5 (Panthee et al. 2006)	Glyma.19g119200	Cysteine synthase 26 (CS26) - cysteine synthase isopropylmalate synthase (IPMS2) - leucine biosynthesis
							Glyma.19g120400	IPMS2 - leucine biosynthesis
							Glyma.19g120600	Cystathionine beta-lyase (CBL) - methionine biosynthetic pathway
							Glyma.19g132000	
ss715637176	20	28389137	0.20	C/T	Protein, Oil, Ala,	Trp1-6, Aps1-3, protein 3g2-g11, oil 4g1-g8		

ss715637437	20	33275560	0.08	C/T	Gly, Iso, Leu, Thre Protein, Oil, Asp, Glu, Met, Val, Lys	(Panthee et al. 2006; Bandillo et al. 2015) protein 3g2-g11, oil 4g1- g8 (Bandillo et al. 2015)

---

*MAF, minimum allele frequency

Position in base pair

§Chr, chromosome

Chapter IV:

**PROSPECTS OF GENOMIC PREDICTION USING TRAINING
POPULATION DESIGN IN AN INTERSPECIFIC SOYBEAN
NAM PANEL**

Abstract

Agronomically important traits generally have complex genetic architecture, where many genes, each one with a small additive effect controls the trait. Genomic prediction has been designed to increase genetic gain and efficiency in plant breeding programs. The objective of this study was to evaluate the training population design on prediction accuracy of genomic prediction models in wild soybean-derived materials. We used a soybean NAM panel containing three *G. max* x *Glycine soja* families evaluated in multi-environment. Three training population designs were used, a) Within family: predictions were made exclusively within each family. b) Leave one family out: each family is predicted using a training set that contains the other two families. c) Across all families: all the individuals from the three families were randomly assigned to either the training or validation set. As in previous studies, training population size showed a strong relationship with accuracy; Increasing the training population size from 50 to 300 individuals improved prediction accuracy from 0.49 to 0.77 (57% increase) across all traits, with little increment in average between 300 and 390 individuals (1%). Marker density had little impact on the prediction accuracy across traits, with a significant increment in prediction accuracy up to 1423 markers (18.5%). Across all families design had higher prediction accuracies for all the traits compared with Leave one family out

and Within family design, with prediction accuracies ranged from moderate (0.55) to high (0.75) across traits. The NAM panel containing interspecific crosses, successfully predict grain yield, maturity, oil and protein. Our results showed encouraging prediction accuracies for grain yield (0.55-0.73), which is impressive from crosses originated from half-wild soybean materials. In conclusion, training population strategies where population size and multiple families were maximized (Across all design) produce robust prediction accuracies for yield, maturity, protein, and oil. Genomic predictions might also accelerate genetic gain in pre-breeding efforts using wild soybeans.

Introduction

Soybean [*Glycine max* (L.) Merr.] is an important cash crop that serves as a major source of oil and protein for human and animal consumption. Soybean crop value exceeded \$41 billion in the United States during 2017 (Soystats, 2018). Soybean planted area and total production has increased since 2015 (USDA - United department of Agriculture 2018).

Most of the success of modern soybean cultivation in North America is due to breeding efforts to significantly increase grain yield during the last 80 years (Rincker et al. 2014). However, the increasing need for food and limited arable land requires more efficient genetic improvement of crop productivity. Agronomic important traits generally have complex genetic architecture, where a large number of genes each one with a small additive effect on the phenotype control the trait (Hill 2010) thus, decreasing selection accuracy, especially in grain yield that is highly correlated with environmental factors (Yan and Rajcan 2003; Xavier et al. 2018).

Genomic prediction (GP), was first proposed by Meuwissen et al. (2001) for predicting phenotypic performance by estimating the genomic breeding value attributed to each locus using a dense marker panel. Then, genomic estimated breeding values (GEBV) for individuals are calculated, and the GEBVs summarize the presence of favorable alleles from a genome-wide standpoint. It is expected that every trait locus will be in linkage disequilibrium (LD) with at least one marker locus in the population,

improving the prediction of complex quantitative traits (Jannink 2010). In general, genomic prediction outperforms marker-assisted selection and phenotypic selection by ~30% (Bernardo and Yu 2007; Heffner et al. 2010; Albrecht et al. 2011). Also, it can reduce the breeding cycle (Heffner et al. 2010), because breeding values can be estimated for individuals without phenotyping.

The prediction accuracy of GP is affected by various factors related to the genetic architecture of the trait under study (Jannink 2010; Lorenz 2013). Highly heritable traits are expected to have higher prediction accuracies as training set and testing set are more closely related (Habier et al. 2007, 2010). Closely related individuals produce higher prediction accuracies due to the marker density used in the prediction. Prediction accuracy generally increases when more markers are used in the model. However, this is not always the case (Heffner et al. 2011; VanRaden et al. 2011).

Recent results indicate that larger training populations (TP) increase accuracy (Isidro et al. 2015; Xavier et al. 2016). However, smaller TP can be as accurate as larger TP (Habier et al. 2009) because higher accuracy estimates are also related to the marker diversity present in the individuals (Calus 2010). Designing appropriate training populations are the key to achieve high accuracy in prediction modeling. Several studies have tested different TP designs to optimize GP in highly diverse panels and populations (Rincent et al. 2012, 2017). Edwards et al. (2018) evaluated 44 bi-parental, and triparental wheat crosses. They observed that TP with unrelated crosses resulted in lower prediction accuracies. However, no differences in prediction accuracies were shown between TP with exclusively related crosses or unrelated x related crosses (50% related and 50% unrelated). Crossa et al. (2016), tested different TP designs using diverse wheat

lines and observed that TP with 20% diverse lines had similar accuracies as those TP from random sampling.

Several statistical models have been tested in genomic prediction. Models that treat marker effects as random across the genome, such as RR-BLUP and G-BLUP (Meuwissen et al. 2001; VanRaden 2008), models allow markers effects to have different prior distributions, such as BayesB, BayesA, and Bayesian LASSO (Meuwissen et al. 2001; De Los Campos et al. 2009). Non-parametric models, such as random forest (RF), reproducing kernel Hilbert spaces regressions (RKHS), and support vector regression (SVR) (Breiman 2001; Gianola and Van Kaam 2008; Moser et al. 2009). More recently, with the higher volume and complexity of data generated, neural networks and machine learning have been used in genomic prediction (Gianola et al. 2011; González-Camacho et al. 2012; Ornella et al. 2014).

Generally, nonlinear and nonparametric models outperform linear (parametric) models in prediction accuracy. Pérez-Rodríguez et al. (2012) observed a consistent superiority of RKHS over BayesA, BayesB, Bayesian LASSO, and RR-BLUP for days to heading and grain yield in wheat. However, Roorkiwal et al. (2016) observed no significant difference between RF, BayesA, BayesB, Bayesian LASSO, RR-BLUP, and BayesC π for days to flowering and maturity, 100-seed weight and grain yield in chickpea. Similar results were observed in soybean, with little variation of accuracy between models (Xavier et al. 2016). Although, when gene action is due to non-additive effects (epistasis and dominance), nonparametric models provided more accurate predictions, and parametric models were more accurate in additive genetic architecture (Howard et al. 2014; Momen et al. 2018)

Genomic prediction has been used successfully to predict important agronomic traits in soybean, such as yield, protein, seed weight, and maturity (Jarquin et al. 2014, 2016; Zhang et al. 2016b; Duhnen et al. 2017). Jarquin et al. (2016) demonstrated that GP models could be used to select accessions from the USDA Soybean Germplasm Collection. Also, Duhnen et al. (2017) showed the potential use of GP to predict protein, grain yield, and protein yield.

The objective of this study was to evaluate training populations design, size, and genetic structure on the prediction accuracy of GEBV prediction models using a diverse interspecific soybean NAM panel, containing three *G. max* x *Glycine soja* populations evaluated across eight environments.

Materials and methods

Germplasm development and phenotyping

Our NAM panel contains 392 F₄-derived RILs, developed from three unique biparental cross combinations between the *G. max* hub parent Williams 82 (Bernard and Cromeens 1988) and three *G. soja* plant introductions PI464890B, PI458536, and PI522226 which are maintained by the United States Department of Agriculture (USDA) in the Germplasm Resources Information Network – GRIN NPGS (<https://www.ars-grin.gov/>). Cross-pollination of the hub parent Williams 82 and each of the three *G. soja* plant introductions was conducted in Champaign-Urbana, IL during the summer of 2010, and the true F₁ hybrid plants were grown at the same environment in 2011 and bulk harvested. Approximately 10,000 F₂ plants were grown at the Bradford Research Center in Columbia, MO during the summer of 2012 for each of the three populations (30,000 total plants), and F₂ plants were bulk harvested. In 2013, approximately 100,000 F₃ plants were grown at the Bradford Research Center in Columbia, MO for each of the three populations (300,000 plants total), and approximately 500 single plants were harvested separately from each F₃ population (1500 total plants), the F₃ plants were selected based on agronomic traits, such as maturity, lodging, and shattering. The F_{3:4} plant rows were grown at the Bay Farm Research Facility in Columbia, MO during 2014, and a single plant was harvested separately from each row. The F_{4:5} plant rows were grown at the Bay Farm Research Facility in Columbia, MO during 2015, and each plant row was bulk harvested for development of the F₄-derived RILs. Only lines with sufficient seed for broad testing were carried forward, and RILs were randomly chosen to

represent an appropriate equal number from each of the three bi-parental populations to reduce the NAM panel size to a manageable field experiment.

During the growing seasons of 2016 and 2017, RILs were planted in four-row research plots with a row spacing of 76 cm, a row length of 3.6 m, and a 1.2 m fallow alley. The environments included the Hundley-Whaley Research Center in Albany, MO (40°14'53"N 94°19'51"W), the Bay Farm Research Facility in Columbia, MO (38°57'50"N 92°19'42"W), the Greenley Research Center in Novelty, MO (39°57'28.23"N 92°10'40.45"W) and the Graves-Chapple Research Center in Rock Port, MO (40° 16'18.09" N 95°28' 26.38"W). The field research plots were arranged in an augmented incomplete block experimental design with one replication in 2016 and two replications in 2017, with two check cultivars per block in both years. Genotypes were randomly assigned to a block and replication, and blocks were randomly ordered in the field across all environments. The experiment was divided into two field blocks at each environment for group MG III and MG IV RILs to accommodate combine harvesting across all environments.

Grain yield was measured by harvesting the two center rows of each four-row plot with an Almaco R1 or an Almaco SPC-40 plot combine (Nevada, Iowa). The total seed weight and seed moisture were measured in real time on the combine, and the final grain yield was calculated as kg ha⁻¹ on a 13% moisture basis. Plant maturity was recorded as the day when 95% of the pods in the center two rows of each plot were mature; R8 stage (Fehr et al., 1971), indicated by the development of brown pod color and considering September 1st as day one. 50 mL of soybean seeds were randomly sampled from all the yield plots and were milled using a Perten laboratory Mill 3600 (Perten Instruments,

Hägersten Sweden). Protein and oil were estimated in the ground sample via near-infrared spectroscopy using a Perten DA 7250 (Perten Instruments, Hägersten Sweden). Oil and protein were estimated on a dry weight basis.

Genotyping and SNP analysis

About 50 mg of leaf tissue from 10 random plants from each RIL and the parents were used to isolate DNA by using DNeasy Plant 96kit (QIAGEN, Valencia, CA) following the instructions of the manufacturer. DNA samples were submitted to the Soybean Genomics and Improvement Laboratory, USDA-ARS, for genotyping using the Illumina Infinium BARCSoySNP6K BeadChip (Song et al. 2013). Genotypes were called using the software Genome Studio (Illumina, San Diego, California USA). SNPs with less than 90% call rate were excluded from the analysis. SNPs were imputed using the LD-KNNi method implemented in version 5 of Tassel (Bradbury et al. 2007). The R package 'ParentOffspring' (Abdel-Haleem et al. 2013) with a threshold of 80% of similarity between parents, and a principal component analysis using version 5 of the TASSEL software (Bradbury et al. 2007) was used to eliminate RILs that exhibited genotypes inconsistent with expected genetic similarity to the respective parents.

Statistical analysis and training population design

Best linear unbiased estimators (BLUE) were used as an entry to perform genomic prediction analysis according to the model adopted by (Jarquín et al. 2014): $y_{ijkl} = \mu + g_i + cb_l + l_j + gl_{ij} + r_{k(j)} + b_{l(k)} + e_{ijkl}$, where μ is the trait mean, g_i is the genetic effect of the genotypes, cb_k is the interaction effect between the checks and incomplete blocks, l_j is the effect of environments, gl_{ij} is the interaction effect between the genotypes and environments, $r_{k(j)}$ is the effect of the replications nested in environment, $b_{l(k)}$ is the incomplete block effect nested within replicate, and e_{ijkl} is the residual. Genotypes and environments were treated as fixed effects, the analysis was performed using the package lme4 in R (Bates et al. 2015). Broad-sense heritability (h^2) on an entry-mean basis was calculated as follows: $h^2 = \sigma_g^2 / (\sigma_g^2 + \sigma_{ge}^2/l + \sigma_e^2/r)$, where σ_g^2 , σ_{ge}^2 and σ_e^2 are the genetic variance, genotype x environment variance and error variance; l is the number of environments and r is the number of replications (Fehr 1991).

Three models were used to fit the genomic prediction; the first model was the ridge regression best linear unbiased predictor (RR-BLUP) obtained using the R package ‘rrBLUP’ (Endelman 2011). In this method, markers are assumed to be random. Two Bayesian models with variable selection for markers effects were used, BayesB (Meuwissen et al. 2001) and Bayesian least absolute shrinkage and selection operator (Bayesian LASSO) (De Los Campos et al. 2009). Both methods were implemented in the R package ‘bWGR’ (Xavier et al. 2017b).

Three training populations designs were used to assess the effect of training population design on prediction accuracy in interspecific soybean crosses. a) Within

family: predictions were made exclusively within families, each family is predicted using training, and validation sets contain individuals from the same family than predicted using 5-fold cross-validation repeated 20 times. b) Leave one family out: families were assigned to either training or validation set, each family is predicted using a training set that contains the other two families, which means the training set does not share any individual from the validation set, the cross-validation was repeated 20 times in each possible combination of families. c) Across all families: all individuals from the three families were randomly assigned to either the training or validation set, without considering a family relationship, than predicted using 5-fold cross-validation repeated 20 times. The prediction accuracies were calculated as the Pearson correlation coefficient between the corrected phenotypes (BLUEs) and genomic estimated breeding values (GEBVs).

Results

The NAM panel was evaluated for their prediction performances for yield, maturity, protein, and oil in multi-environment trials (Table 4-1). The overall yield mean for the NAM trials was 3225.5 kg ha⁻¹ for grain yield, 32.3 days for maturity, 427 g kg⁻¹ for protein and 192 g kg⁻¹ for oil. Over 4500 plots were harvested across eight environments. Heritability on an entry-mean basis (Table 4-2) were 0.65, 0.92, 0.89, and 0.87 for yield, maturity, protein, and oil, respectively. These moderate to high heritability for maturity, protein, and oil suggested that the genetic variation accounted for a major part of the phenotypic variance.

The training population (TP) size showed a strong relationship with accuracy (Figure 4-1A). Increasing the training population size from 50 to 300 individuals improved prediction accuracy from 0.49 to 0.77 (57% increase) across all traits. However, between 300 and 390 individual the prediction accuracy only increased 0.019 (1%) across traits. Marker density was tested by dividing the total markers set into subsets of SNP markers, using the whole panel, half panel, a quarter panel, and a half of a quarter panel (Meuwissen et al. 2001), corresponding to the 5695, 2847, 1423 and 711 SNP markers, respectively. Marker density had a smaller impact on the prediction accuracy across traits (Figure 4-1B). The only significant increment in prediction accuracy was observed up to 1423 SNP markers, resulting in an improvement of 18.5%. No significant improvement was observed using more than 1423 SNP markers (1.5%). However, all traits had the highest prediction accuracies when all the SNP markers were used.

Figure 4-2 illustrates the performance of three prediction models under the training populations (TP) designs and assessing the impact of population structure on prediction accuracy. Accuracies were higher Across all families for all the traits compared to Leave one family out and Within families designs, with prediction accuracies ranging from moderate (0.55) to high (0.75) across traits. These results illustrate the improvement in prediction accuracy when combining multiple populations to produce a robust training set. In general, prediction accuracies from the Bayesian LASSO model outperformed RR-BLUP and were slightly higher than BayesB across training population designs and traits. The Bayesian LASSO model generated a

prediction accuracy of 0.72 for grain yield, 0.67 for maturity, 0.75 for protein and 0.77 for oil in Across all families TP design.

Within family design tended to produce slightly higher prediction accuracy than a Leave one family out design, across traits and models (Figure 4-2). However, prediction accuracy for grain yield, maturity, protein, and oil was highly variables across families (Table 4-3). The accuracy reached high values of 0.77 within Fam1 for protein and moderate values as 0.51 within Fam1 for grain yield. Prediction accuracy of 0.59, 0.62, and 0.60 for Fam1, Fam2, and Fam3 were observed for grain yield, respectively. The highest accuracies observed in Within family was observed for protein and oil (0.61-0.72). This could be explained by the smaller population size in this training population and the genetic architecture of each trait, as larger training sets can better estimate small effect QTLs (Beavis 1998; Xu 2003).

Overall Leave one family out design, showed no significant differences for prediction accuracy between models for maturity, protein, and oil, with higher accuracy for BayesB in grain yield (0.65) (Figure 4-1). The poor performance of this TP design comparing Across all families was due to the high variation of predictions across the three possible cross validations for the design (Table 4-4). Prediction accuracies were as high as 0.76 for oil when Fam1+Fam3 was used as TP and Fam2 as validation population (VP). However, lower prediction accuracy was observed for grain yield (0.42) when used Fam2+Fam3 as TP and Fam1 as VP. This pattern was also observed in other traits and training population designs (Table 4-4). The highest prediction accuracy for grain yield was observed in Fam1+Fam2 as TP and Fam3 as VP (0.69). Relatedness between training and validation population is required to acquire high prediction accuracy (Habier

et al. 2007). It was likely some training population designs caused different levels of related lines across training and validation population.

Discussion

Prediction accuracy can be affected by several factors such as genetic architecture of the trait (Lorenz 2013), size of the training population (Xavier et al. 2016), marker density (Norman et al. 2018), relatedness between training and testing sets (Habier et al. 2010), statistical models (Heslot et al. 2012), and population structure (Wray et al. 2013). From a practical breeding standpoint, all this information is essential for the implementation of genomic selection strategies in a breeding program and should guide breeders in decision-making. Research surrounding genomic selection implementation and viability has been a popular topic in animal and plant breeding (Calus 2010; Nakaya and Isobe 2012; Wolc et al. 2016; Wallén et al. 2017; Crossa et al. 2017).

Our results help to understand some of the issues of genomic selection implementation in plant breeding programs. Training population size had a high impact on prediction accuracy (Figure 4-1). However, results have been controversial about how many individuals are ideal in a training set, Xavier et al. (2016) evaluated SoyNAM panel and found the maximum values for accuracy at 4000 individuals in the training set, yet with little improvement after 2000 individuals. On the other hand, studies with much smaller training sets (150 to 300 individuals) generate high prediction accuracies as well (Isidro et al. 2015; Akdemir et al. 2015). Our results showed a higher increment in prediction accuracy until 300 individuals and only a slight increase from 300 to 400. Norman et al. (2018) utilized a panel with more than 10,000 wheat lines genotyped across

18,000 SNPs, observed the increase in prediction accuracy declined above 2,000 individuals, suggesting an accuracy plateau will be achieved in the training population. It will depend on the criteria of the breeder to decide when genotype more individuals will be cost-effective to efficiently improve prediction accuracy and selection.

Marker density is another critical point for the accuracy of genomic prediction. Our results were assessed utilizing four markers densities (5695, 2847, 1423 and 711 SNP markers) in five-fold cross-validation with 20 replications. All traits showed a higher response to marker density up to 1423 SNPs, which was sufficient to generate an accurate calibration within the NAM panel. This reduced marker density in soybean has been previously reported (Xavier et al. 2016; Peixoto et al. 2017) due to the extensive linkage disequilibrium (LD) in the soybean genome (Hyten et al. 2007). An increment in the marker density can be beneficial when large training sets are being tested combined with fast LD decay between markers and the QTLs (Huang et al. 2012). Norman et al. (2018) showed significant gains in prediction accuracy up to 5,000 markers using a wheat panel containing ~10,000 individuals.

Genetic diversity is another critical point in designing training populations. Recently Crossa et al. (2016) evaluated diverse wheat accessions from Mexico and Iran from the wheat gene bank, and Jarquin et al. (2016) evaluated diverse accessions from the USDA soybean germplasm collection. Both studies agreed that an effective training population is possibly designing the training set with high diversity accessions without any harm on prediction accuracy of the validation set. Crossa et al. (2016), tested training sets with 10% and 20% of diverse lines and observed that training sets using 20% diverse lines achieved similar prediction accuracies as random cross-validation. Our results

demonstrated significant increments in prediction accuracy when more individuals from different populations were included in the training set (Across all families TP design). One drawback of genomic selection is that it can lead to loss of genetic variance after a few generations of selection (Rutkoski et al. 2015). It is theorized that the use of high genetically diverse individuals (landraces and wild species) in the training set would allow maintaining genetic variance after a couple of generations of selection.

The lack of genetic diversity present in elite soybean varieties is well known (Hyten et al. 2006), and *Glycine soja* has the potential to introduce new allelic diversity to improve genetic gain for desired traits. However, most of the efforts have been focused on mapping for biotic and abiotic stresses (Winter et al. 2007; Chen et al. 2013), and seed composition (Ha et al. 2014; Leamy et al. 2017). Few attempts for mapping yield QTLs have been done (Concibido et al. 2003; Li et al. 2008). Our results revealed encouraging prediction accuracies for grain yield (0.55-0.73), which is impressive from crosses originating from wild soybean. Individuals can be selected based on their genetic estimated breeding value (GEBV), targeting favorable genetic variations genome-wide, which has immense value on predicting polygenic traits, such as grain yield. In a breeding perspective, genomic selection could be used to predict and improve pre-breeding populations, and enhance the use of diverse germplasm in soybean commercial breeding programs (Crossa et al. 2016; Jarquin et al. 2016). It will also allow selection of parents from pre-breeding populations with a higher frequency of favorable alleles (Gorjanc et al. 2016). Developing interspecific populations is challenging due to the difficulties related to work with undomesticated species. This is particularly true in

evaluating grain yield from crosses with *G. soja*, because of undesirable traits such as excessive lodging, viny grown, shattering and difficulties to mechanically harvest plots.

Our report for prediction accuracy estimates were similar to those found in soybeans for yield, maturity, oil and protein (Jarquin et al. 2016), and superior to those reported for protein and yield (Duhnen et al. 2017). Our prediction estimates were similar to those reported for protein and protein quality traits in wheat (Haile et al. 2018) and maturity and grain yield in chickpea (Roorkiwal et al. 2016). Also, Jarquin et al. (2016), evaluated similar training population designs as we used in this study, and observed One/All (similar to Across all families) and One/Group (Within family) were superior to the training population design Group/All (Leave one family out).

In conclusion, our results showed that using training population strategies, where population size and multiple families were maximized (Across all families) produce robust prediction accuracies for yield, protein, and oil. Genomic predictions might also accelerate genetic gain and selection on pre-breeding efforts. To the best of our knowledge, this is the first report of using genomic prediction strategies in a pre-breeding effort using wild soybean populations.

References

- Abdel-Haleem, H., P. Ji, H.R. Boerma, and Z. Li. 2013. An R package for SNP marker-based parent-offspring tests. *Plant Methods* 9(1): 44. doi: 10.1186/1746-4811-9-44.
- Akdemir, D., J.I. Sanchez, and J.L. Jannink. 2015. Optimization of genomic selection training populations with a genetic algorithm. *Genet. Sel. Evol.* 47(1): 38. doi: 10.1186/s12711-015-0116-6.
- Albrecht, T., V. Wimmer, H.J. Auinger, M. Erbe, C. Knaak, et al. 2011. Genome-based prediction of testcross values in maize. *Theor. Appl. Genet.* 123(2): 339–350. doi: 10.1007/s00122-011-1587-7.
- Bates, D., M. Mächler, B. Bolker, and S. Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Softw.* 67(1): 1–48. doi: 10.18637/jss.v067.i01.
- Beavis, W.D. 1998. QTL analyses: power, precision, and accuracy. In: Paterson, A.H., editor, *Molecular Dissection of Complex Traits*. CRC Press, New York, NY. p. 145–162
- Bernard, R.L., and C.R. Cromeens. 1988. Registration of ‘Williams 82’ Soybean. *Crop Sci.* 28(6): 1027–1028. doi: 10.2135/cropsci1988.0011183X002800060049x.
- Bernardo, R., and J. Yu. 2007. Prospects for genomewide selection for quantitative traits in maize. *Crop Sci.* 47(3): 1082–1090. doi: 10.2135/cropsci2006.11.0690.

- Bradbury, P.J., Z. Zhang, D.E. Kroon, T.M. Casstevens, Y. Ramdoss, et al. 2007. TASSEL: software for association mapping of complex traits in diverse samples.
- Breiman, L. 2001. Random Forests. *Mach. Learn.* 45(1): 5–32. doi: 10.1023/A:1010933404324.
- Calus, M.P.L. 2010. Genomic breeding value prediction: methods and procedures. *animal* 4(02): 157–164. doi: 10.1017/S1751731109991352.
- Chen, P., K. Yan, H. Shao, and S. Zhao. 2013. Physiological Mechanisms for High Salt Tolerance in Wild Soybean (*Glycine soja*) from Yellow River Delta, China: Photosynthesis, Osmotic Regulation, Ion Flux and antioxidant Capacity (P.C.G. Ferreira, editor). *PLoS One* 8(12): e83227. doi: 10.1371/journal.pone.0083227.
- Concibido, V., B. La Vallee, P. McIaird, N. Pineda, J. Meyer, et al. 2003. Introgression of a quantitative trait locus for yield from *Glycine soja* into commercial soybean cultivars. *Theor. Appl. Genet.* 106(4): 575–582. doi: 10.1007/s00122-002-1071-5.
- Crossa, J., D. Jarquín, J. Franco, P. Pérez-Rodríguez, J. Burgueño, et al. 2016. Genomic Prediction of Gene Bank Wheat Landraces. *G3*. 6(7): 1819–1834. doi: 10.1534/g3.116.029637.
- Crossa, J., P. Pérez-Rodríguez, J. Cuevas, O. Montesinos-López, D. Jarquín, et al. 2017. Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. *Trends Plant Sci.* 22(11): 961–975. doi: 10.1016/j.tplants.2017.08.011.
- Duhnen, A., A. Gras, S. Teyssèdre, M. Romestant, B. Claustres, et al. 2017. Genomic selection for yield and seed protein content in Soybean: A study of breeding

program data and assessment of prediction accuracy. *Crop Sci.* 57(3): 1325–1337. doi: 10.2135/cropsci2016.06.0496.

Edwards, S.M., J.B. Buntjer, R. Jackson, A.R. Bentley, J. Lage, et al. 2018. The effects of training population design on genomic prediction accuracy in wheat. *bioRxiv*: 443267. doi: 10.1101/443267.

Endelman, J.B. 2011. Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *Plant Genome J.* 4(3): 250. doi: 10.3835/plantgenome2011.08.0024.

Fehr, W.R. 1991. Principles of cultivar development. Vol. 1. Theory and technique. Macmillan, New York.

Gianola, D., and J.B. Van Kaam. 2008. Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics* 178(4): 2289–2303. doi: 10.1534/genetics.107.084285.

Gianola, D., H. Okut, K.A. Weigel, and G.J. Rosa. 2011. Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. *BMC Genet.* 12(1): 87. doi: 10.1186/1471-2156-12-87.

González-Camacho, J.M., G. de los Campos, P. Pérez, D. Gianola, J.E. Cairns, et al. 2012. Genome-enabled prediction of genetic values using radial basis function neural networks. *Theor. Appl. Genet.* 125(4): 759–771. doi: 10.1007/s00122-012-1868-9.

Gorjanc, G., J. Jenko, S.J. Hearne, and J.M. Hickey. 2016. Initiating maize pre-breeding programs using genomic selection to harness polygenic variation from landrace populations. *BMC Genomics* 17(1): 30. doi: 10.1186/s12864-015-2345-z.

Ha, B.-K., H.-J. Kim, V. Velusamy, T.D. Vuong, H.T. Nguyen, et al. 2014. Identification of quantitative trait loci controlling linolenic acid concentration in PI483463 (Glycine soja). *Theor. Appl. Genet.* 127(7): 1501–1512. doi: 10.1007/s00122-014-2314-y.

Habier, D., R.L. Fernando, and J.C.M. Dekkers. 2007. The impact of genetic relationship information on genome-assisted breeding values. *Genetics* 177(4): 2389–2397. doi: 10.1534/genetics.107.081190.

Habier, D., R.L. Fernando, and J.C.M. Dekkers. 2009. Genomic selection using low-density marker panels. *Genetics* 182(1): 343–353. doi: 10.1534/genetics.108.100289.

Habier, D., R.L. Fernando, K. Kizilkaya, and D.J. Garrick. 2011. Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics* 12(1): 186. doi: 10.1186/1471-2105-12-186.

Habier, D., J. Tetens, F.R. Seefried, P. Lichtner, and G. Thaller. 2010. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. *Genet. Sel. Evol.* 42(1): 5. doi: 10.1186/1297-9686-42-5.

Haile, J.K., A. N'Diaye, F. Clarke, J. Clarke, R. Knox, et al. 2018. Genomic selection for grain yield and quality traits in durum wheat. *Mol. Breed.* 38(6): 75. doi: 10.1007/s11032-018-0818-x.

Heffner, E.L., J.L. Jannink, H. Iwata, E. Souza, and M.E. Sorrells. 2011. Genomic selection accuracy for grain quality traits in biparental wheat populations. *Crop Sci.* 51(6): 2597–2606. doi: 10.2135/cropsci2011.05.0253.

Heffner, E.L., A.J. Lorenz, J.L. Jannink, and M.E. Sorrells. 2010. Plant breeding with Genomic selection: Gain per unit time and cost. *Crop Sci.* 50(5): 1681–1690. doi: 10.2135/cropsci2009.11.0662.

Heslot, N., H.P. Yang, M.E. Sorrells, and J.L. Jannink. 2012. Genomic selection in plant breeding: A comparison of models. *Crop Sci.* 52(1): 146–160. doi: 10.2135/cropsci2011.06.0297.

Hill, W.G. 2010. Understanding and using quantitative genetic variation. *Philos. Trans. R. Soc. B Biol. Sci.* 365(1537): 73–85. doi: 10.1098/rstb.2009.0203.

Howard, R., A.L. Carriquiry, and W.D. Beavis. 2014. Parametric and Nonparametric Statistical Methods for Genomic Selection of Traits with Additive and Epistatic Genetic Architectures. *G3.* 4(6): 1027–1046. doi: 10.1534/g3.114.010298.

Huang, B.E., A.W. George, K.L. Forrest, A. Kilian, M.J. Hayden, et al. 2012. A multiparent advanced generation inter-cross population for genetic analysis in wheat. *Plant Biotechnol. J.* 10(7): 826–839. doi: 10.1111/j.1467-7652.2012.00702.x.

Hyten, D.L., I.Y. Choi, Q. Song, R.C. Shoemaker, R.L. Nelson, et al. 2007. Highly variable patterns of linkage disequilibrium in multiple soybean populations. *Genetics* 175(4): 1937–1944. doi: 10.1534/genetics.106.069740.

Hyten, D.L., Q. Song, Y. Zhu, I.-Y. Choi, R.L. Nelson, et al. 2006. Impacts of genetic bottlenecks on soybean genome diversity. *Proc. Natl. Acad. Sci. U. S. A.* 103(45): 16666–71. doi: 10.1073/pnas.0604379103.

Isidro, J., J.L. Jannink, D. Akdemir, J. Poland, N. Heslot, et al. 2015. Training set optimization under population structure in genomic selection. *TAG. Theor. Appl. Genet.* 128(1): 145–158. doi: 10.1007/s00122-014-2418-4.

Jannink, J.L. 2010. Dynamics of long-term genomic selection. *Genet. Sel. Evol.* 42(1): 35. doi: 10.1186/1297-9686-42-35.

Jarquín, D., K. Kocak, L. Posadas, K. Hyma, J. Jedlicka, et al. 2014. Genotyping by sequencing for genomic prediction in a soybean breeding population. *BMC Genomics* 15(1): 740. doi: 10.1186/1471-2164-15-740.

Jarquín, D., J. Specht, and A. Lorenz. 2016. Prospects of Genomic Prediction in the USDA Soybean Germplasm Collection: Historical Data Creates Robust Models for Enhancing Selection of Accessions. *G3*. 6(8): 2329–2341. doi: 10.1534/g3.116.031443.

Leamy, L.J., H. Zhang, C. Li, C.Y. Chen, and B.H. Song. 2017. A genome-wide association study of seed composition traits in wild soybean (*Glycine soja*). *BMC Genomics* 18(1): 18. doi: 10.1186/s12864-016-3397-4.

Li, D., T.W. Pfeiffer, and P.L. Cornelius. 2008. Soybean QTL for Yield and Yield Components Associated with Alleles. *Crop Sci.* 48(2): 571. doi: 10.2135/cropsci2007.06.0361.

Lorenz, A.J. 2013. Resource Allocation for Maximizing Prediction Accuracy and Genetic Gain of Genomic Selection in Plant Breeding: A Simulation Experiment. *G3*. 3(3): 481–491. doi: 10.1534/g3.112.004911.

De Los Campos, G., H. Naya, D. Gianola, J. Crossa, A. Legarra, et al. 2009. Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics* 182(1): 375–385. doi: 10.1534/genetics.109.101501.

Meuwissen, T.H.E., B.J. Hayes, and M.E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157: 1819–1829. doi: 11290733.

Momen, M., A.A. Mehrgardi, A. Sheikhi, A. Kranis, L. Tusell, et al. 2018. Predictive ability of genome-assisted statistical models under various forms of gene action. *Sci. Rep.* 8(1): 12309. doi: 10.1038/s41598-018-30089-2.

Moser, G., B. Tier, R. Crump, M. Khatkar, and H. Raadsma. 2009. A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. *Genet. Sel. Evol.* 41(1): 56. doi: 10.1186/1297-9686-41-56.

Nakaya, A., and S.N. Isobe. 2012. Will genomic selection be a practical method for plant breeding? *Ann. Bot.* 110(6): 1303–16. doi: 10.1093/aob/mcs109.

Norman, A., J. Taylor, J. Edwards, and H. Kuchel. 2018. Optimising Genomic Selection in Wheat: Effect of Marker Density, Population Size and Population Structure on Prediction Accuracy. *G3.* 8(9): g3.200311.2018. doi: 10.1534/g3.118.200311.

Ornella, L., P. Pérez, E. Tapia, J.M. González-Camacho, J. Burgueño, et al. 2014. Genomic-enabled prediction with classification algorithms. *Heredity.* 112(6): 616–626. doi: 10.1038/hdy.2013.144.

Peixoto, L.D.A., T.C. Moellers, J. Zhang, A.J. Lorenz, L.L. Bhering, et al. 2017. Leveraging genomic prediction to scan germplasm collection for crop improvement. *PLoS One* 12(6). doi: 10.1371/journal.pone.0179191.

Pérez-Rodríguez, P., D. Gianola, J.M. González-Camacho, J. Crossa, Y. Manès, et al. 2012. Comparison Between Linear and Non-parametric Regression Models for Genome-Enabled Prediction in Wheat. *G3.* 2(12): 1595–1605. doi: 10.1534/g3.112.003665.

Rincent, R., A. Charcosset, and L. Moreau. 2017. Predicting genomic selection efficiency to optimize calibration set and to assess prediction accuracy in highly

structured populations. *Theor. Appl. Genet.* 130(11): 2231–2247. doi: 10.1007/s00122-017-2956-7.

Rincent, R., D. Laloë, S. Nicolas, T. Altmann, D. Brunel, et al. 2012. Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: Comparison of methods in two diverse groups of maize inbreds (*Zea mays* L.). *Genetics* 192(2): 715–728. doi: 10.1534/genetics.112.141473.

Rincker, K., R. Nelson, J. Specht, D. Sleper, T. Cary, et al. 2014. Genetic Improvement of U.S. Soybean in Maturity Groups II, III, and IV. *Crop Sci.* 54(0): 1419–1432. doi: 10.2135/cropsci2013.10.0665.

Roorkiwal, M., A. Rathore, R.R. Das, M.K. Singh, A. Jain, et al. 2016. Genome-Enabled Prediction Models for Yield Related Traits in Chickpea. *Front. Plant Sci.* 7: 1666. doi: 10.3389/fpls.2016.01666.

Rutkoski, J., R.P. Singh, J. Huerta-Espino, S. Bhavani, J. Poland, et al. 2015. Genetic Gain from Phenotypic and Genomic Selection for Quantitative Resistance to Stem Rust of Wheat. *Plant Genome* 8(2): 0. doi: 10.3835/plantgenome2014.10.0074.

Song, Q., D.L. Hyten, G. Jia, C. V. Quigley, E.W. Fickus, et al. 2013. Development and Evaluation of SoySNP50K, a High-Density Genotyping Array for Soybean (T. Zhang, editor). *PLoS One* 8(1): e54985. doi: 10.1371/journal.pone.0054985.

USDA - United department of Agriculture. 2018. National agriculture statistics service. <https://quickstats.nass.usda.gov/results> (accessed 23 August 2018).

VanRaden, P.M., J.R. O'Connell, G.R. Wiggans, and K.A. Weigel. 2011. Genomic evaluations with many more genotypes. *Genet. Sel. Evol.* 43(1): 10. doi: 10.1186/1297-9686-43-10.

VanRaden, P.M. 2008. Efficient Methods to Compute Genomic Predictions. *J. Dairy Sci.* 91(11): 4414–4423. doi: 10.3168/jds.2007-0980.

Wallén, S.E., M. Lillehammer, and T.H.E. Meuwissen. 2017. Strategies for implementing genomic selection for feed efficiency in dairy cattle breeding schemes. *J. Dairy Sci.* 100(8): 6327–6336. doi: 10.3168/jds.2016-11458.

Winter, S.M.J., B.J. Shelp, T.R. Anderson, T.W. Welacky, and I. Rajcan. 2007. QTL associated with horizontal resistance to soybean cyst nematode in Glycine soja PI464925B. *Theor. Appl. Genet.* 114(3): 461–472. doi: 10.1007/s00122-006-0446-4.

Wolc, A., A. Kranis, J. Arango, P. Settar, J.E. Fulton, et al. 2016. Implementation of genomic selection in the poultry industry. *Anim. Front.* 6(1): 23. doi: 10.2527/af.2016-0004.

Wray, N.R., J. Yang, B.J. Hayes, A.L. Price, M.E. Goddard, et al. 2013. Pitfalls of predicting complex traits from SNPs. *Nat. Rev. Genet.* 14(7): 507–515. doi: 10.1038/nrg3457.

Xavier, A., D. Jarquin, R. Howard, V. Ramasubramanian, J.E. Specht, et al. 2018. Genome-Wide Analysis of Grain Yield Stability and Environmental Interactions in a Multiparental Soybean Population. *G3.* 8(2): 519–529. doi: 10.1534/g3.117.300300.

Xavier, A., W.M. Muir, and K.M. Rainey. 2016. Assessing Predictive Properties of Genome-Wide Selection in Soybeans. *G3.* 6(8): 2611–6. doi: 10.1534/g3.116.032268.

Xavier, A., S. Xu, W. Muir, and K.M. Rainey. 2017. Genomic prediction using subsampling. *BMC Bioinformatics* 18(1): 191. doi: 10.1186/s12859-017-1582-3.

Xu, S. 2003. Theoretical basis of the Beavis effect. *Genetics* 165(4): 2259–68. <http://www.ncbi.nlm.nih.gov/pubmed/14704201>.

Yan, W., and I. Rajcan. 2003. Prediction of cultivar performance based on single-versus multiple-year tests in soybean. *Crop Sci.* 43(2): 549–555. doi: 10.2135/cropsci2003.0549.

Zhang, J., Q. Song, P.B. Cregan, and G.L. Jiang. 2016. Genome-wide association study, genomic prediction and marker-assisted selection for seed weight in soybean (*Glycine max*). *Theor. Appl. Genet.* 129(1): 117–130. doi: 10.1007/s00122-015-2614-x.

Tables and figures

Table 4-1. Summary of descriptive statistics for yield, maturity, protein and oil across eight environments.

Trait	Mean	SD*	Coef. Variation %	Environments	Total data points (plots)
Grain yield - kg ha ⁻¹	3225.2	812.3	19.2	8	4504
Maturity - days	32.3	6.1	11.3	8	4836
Protein - g kg ⁻¹	427.8	27.4	7.4	8	4456
Oil - g kg ⁻¹	192.1	16.5	8.5	8	4456

*SD – Standard deviation

Table 4-2. Summary of phenotypic data analysis for grain yield, maturity, protein and oil.

Trait	σ_g^2 *	σ_{ge}^2 †	σ_e^2 §	h^2 #
Grain yield	10024.2	27488.3	16565.4	0.65
Maturity	146.3	63.4	44.3	0.92
Protein	158.6	48.7	109.3	0.89
Oil	102.2	39.8	86.2	0.87

* σ_g^2 - Genotypic variance

† σ_{ge}^2 - Variance of genotype x environment interaction

§ σ_e^2 - Residual variance

h^2 - Heritability on an entry-mean basis

Figure 4-1. Effect of training population size (A) and marker density (B) on the prediction accuracy of grain yield, maturity, protein and oil.

Figure 4-2. Boxplots of the prediction accuracy of the genomic prediction models (RR-BLUP, BayesB, and BayesLASSO) in soybeans tested in three training populations designs for grain yield (A), maturity (B), protein (C), and oil (D). Training population designs included Across all families (Across/All), Leave one family out (Leave_one/Fam_out) and Within families (Within/Fam).

Table 4-3. Prediction accuracy of three genomic prediction models including RR-BLUP, BayesB, and BayesLASSO, for grain yield, maturity, protein and oil using Within family training population design.

	RR-BLUP		BayesB		BayesLASSO	
	GRAIN YIELD					
	Accuracy	SD [*]	Accuracy	SD	Accuracy	SD
Fam1 [‡]	0.51	0.10	0.60	0.03	0.67	0.03
Fam2 [§]	0.55	0.09	0.62	0.05	0.68	0.04
Fam3 [#]	0.62	0.08	0.57	0.10	0.62	0.06
	MATURITY					
	Accuracy	SD	Accuracy	SD	Accuracy	SD
Fam1	0.59	0.11	0.63	0.09	0.67	0.08
Fam2	0.62	0.22	0.55	0.12	0.63	0.09
Fam3	0.63	0.08	0.61	0.09	0.67	0.07
	PROTEIN					
	Accuracy	SD	Accuracy	SD	Accuracy	SD
Fam1	0.67	0.10	0.72	0.07	0.77	0.07
Fam2	0.65	0.11	0.58	0.16	0.60	0.15
Fam3	0.62	0.17	0.67	0.07	0.72	0.08
	OIL					
	Accuracy	SD	Accuracy	SD	Accuracy	SD
Fam1	0.66	0.06	0.72	0.09	0.70	0.09
Fam2	0.62	0.19	0.68	0.11	0.61	0.16
Fam3	0.66	0.11	0.62	0.09	0.65	0.06

* SD - standard deviation

[‡] Fam1 - Williams 82 x PI464890B

[§] Fam2 - Williams 82 x PI458536

[#] Fam3 - Williams 82 x PI522226

Table 4-4. Prediction accuracy of three genomic prediction models including RR-BLUP, BayesB, and Bayes LASSO, for grain yield, maturity, protein and oil in a leave one family out training population design.

	RR-BLUP		BayesB		BayesLASSO	
	GRAIN YIELD					
	Accuracy	SD	Accuracy	SD	Accuracy	SD
TP Fam2+3/VP Fam1	0.35	0.03	0.45	0.02	0.45	0.04
TP Fam1+3/VP Fam2	0.56	0.05	0.64	0.06	0.6	0.07
TP Fam1+2/VP Fam3	0.67	0.02	0.7	0.04	0.7	0.01
	MATURITY					
	Accuracy	SD	Accuracy	SD	Accuracy	SD
	TP Fam2+3/VP Fam1	0.56	0.02	0.52	0.02	0.54
TP Fam1+3/VP Fam2	0.59	0.03	0.57	0.05	0.51	0.02
TP Fam1+2/VP Fam3	0.44	0.02	0.54	0.06	0.53	0.04
	PROTEIN					
	Accuracy	SD	Accuracy	SD	Accuracy	SD
	TP Fam2+3/VP Fam1	0.68	0.04	0.67	0.03	0.69
TP Fam1+3/VP Fam2	0.72	0.01	0.73	0.04	0.71	0.03
TP Fam1+2/VP Fam3	0.55	0.06	0.58	0.01	0.58	0.06
	OIL					
	Accuracy	SD	Accuracy	SD	Accuracy	SD
	TP Fam2+3/VP Fam1	0.64	0.08	0.68	0.09	0.66
TP Fam1+3/VP Fam2	0.75	0.1	0.76	0.11	0.76	0.03
TP Fam1+2/VP Fam3	0.62	0.04	0.5	0.05	0.56	0.04

SD, standard deviation; TP, training population; VP, validation population; Fam1 - Williams 82 x PI464890B; Fam2 – Williams 82 x PI458536; Fam3 – Williams 82 x PI522226.

Chapter V: OVERALL SUMMARY AND FUTURE PROSPECTS

Discovering, characterization and evaluation of new genetic variants from wild soybean has led to valuable information for soybean breeders to address the problem of low genetic diversity, and to boost genetic gain. Our results provide valuable resources for breeding material, especially selecting inbred lines with wild soybean alleles already integrated into more adaptable material. The nested association mapping for agronomic and seed composition traits identified significant and novel associations between markers and traits such as grain yield, plant maturity, plant height, lodging, sulfur-containing amino acids, which are deficient in soybean seed protein, and fatty acid composition. One QTL from *Glycine soja* was associated with grain yield on chromosome 17, yielded 6% more than the lines carrying the allele from the *Glycine max* parent. A reasonable number of *G. soja* x *G. max* lines yielded more than the *G. max* parent (Williams 82). Among those SA15-109, SA15-66476, SA15-63423, and SA15-66563 yielded between 5 and 17% more than the Williams 82 depending on the environment.

Further studies to identify candidate genes associated with protein oil, fatty acid, and amino acid are necessary to understand the genetic mechanisms behind. The QTLs should be tested in different backgrounds to be confirmed, and near-isogenic lines (NILs) can be used to validate most likely candidate loci. The well-reported protein/oil QTLs cqSeed protein-001 on Chr 15 and cqSeed protein-003 on Chr 20 (Fasoula et al. 2004; Nichols et al. 2006), were confirmed in the NAM panel. Both QTLs showed positive effects on protein concentration, whereas reducing the oil content in a proportion of 1:1.5. Also, they have been mapped in multiple populations from different sources (Diers et al.

1992; Brummer et al. 1997; Bandillo et al. 2015; Warrington et al. 2015; Phansak et al. 2016).

Genomic selection applied to plant breeding can increase the prediction accuracy of complex traits and improve genetic gain over time. Our results showed that prediction accuracy improved up to 300 individuals in the training population and a slight improvement from 300 to 400. Marker density is another critical point to consider during genomic prediction. Our results were assessed utilizing four markers densities 5695, 2847, 1423 and 711 SNP markers in five-fold cross-validation with 20 replications. All traits experienced a higher response to marker density up to 1423 SNPs, showing that this was adequate to generate an accurate prediction calibration within the NAM panel.

Regarding prediction accuracy, our reports were similar to previous studies in soybeans for yield and maturity (Jarquin et al. 2016) and superior other studies on protein and yield (Duhnen et al. 2017). Our NAM panel containing *Glycine soja*, successfully predicted grain yield, oil, protein, and maturity. Our results showed encouraging prediction accuracies for grain yield (0.55-0.73), from crosses originating from wild soybean. In conclusion, training population strategies where population size and multiple families were maximized produce robust prediction accuracies for yield, protein, and oil. Genomic predictions might also accelerate genetic gain in pre-breeding efforts using wild soybeans.

References

- Bandillo, N., D. Jarquin, Q. Song, R. Nelson, P. Cregan, et al. 2015. A Population Structure and Genome-Wide Association Analysis on the USDA Soybean Germplasm Collection. *Plant Genome* 8(3). doi: 10.3835/plantgenome2015.04.0024.
- Brummer, E.C., G.L. Graef, J. Orf, J.R. Wilcox, and R.C. Shoemaker. 1997. Mapping QTL for Seed Protein and Oil Content in Eight Soybean Populations. *Crop Sci.* 37(2): 370. doi: 10.2135/cropsci1997.0011183X003700020011x.
- Diers, B.W., W.R. Fehr, and R.C. Shoemaker. 1992. RFLP analysis of soybean seed protein and oil content. *Theor. Appl. Genet.* 83: 608–612. doi: 10.1007/BF00226905.
- Duhnen, A., A. Gras, S. Teyssèdre, M. Romestant, B. Claustres, et al. 2017. Genomic selection for yield and seed protein content in Soybean: A study of breeding program data and assessment of prediction accuracy. *Crop Sci.* 57(3): 1325–1337. doi: 10.2135/cropsci2016.06.0496.
- Fasoula, V.A., D.K. Harris, and H.R. Boerma. 2004. Validation and designation of quantitative trait loci for seed protein, seed oil, and seed weight from two soybean populations. *Crop Sci.* 44(4): 1218–1226. doi:10.2135/cropsci2004.1218.
- Jarquin, D., J. Specht, and A. Lorenz. 2016. Prospects of Genomic Prediction in the USDA Soybean Germplasm Collection: Historical Data Creates Robust Models for Enhancing Selection of Accessions. *G3.* 6(8): 2329–2341. doi: 10.1534/g3.116.031443.
- Nichols, D.M., K.D. Glover, S.R. Carlson, J.E. Specht, and B.W. Diers. 2006. Fine Mapping of a Seed Protein QTL on Soybean Linkage Group I and Its Correlated Effects on Agronomic Traits. *Crop Sci.* 46(2): 834. doi: 10.2135/cropsci2005.05-0168.

Phansak, P., W. Soonsuwon, D.L. Hyten, Q. Song, P.B. Cregan, et al. 2016. Multi-Population Selective Genotyping to Identify Soybean [*Glycine max* (L.) Merr.] Seed Protein and Oil QTLs. *G3*. 6(6): 1635–1648. doi: 10.1534/g3.116.027656.

Warrington, C. V., H. Abdel-Haleem, D.L. Hyten, P.B. Cregan, J.H. Orf, et al. 2015. QTL for seed protein and amino acids in the Benning $\times$ Danbaekkong soybean population. *Theor. Appl. Genet.* 128(5): 839–850. doi: 10.1007/s00122-015-2474-4.

VITA

Eduardo Beche was born June 22, 1988, in Vere, a small town in southern Brazil. He studied Agronomy at Technological Federal University of Parana and graduated in December of 2010 with a Bachelor of Science degree. Between 2011 and 2013 he pursued his Master of Science degree in Crop Science at the same University, studying the causes of genetic gain in wheat over the last 70 years of breeding in southern Brazil. In 2014, he was awarded a Brazilian Government Scholarship to pursue his Ph.D. degree in plant breeding, genetics, and genomics at the University of Missouri, Columbia, under the advisement of Dr. Andrew Scaboo and Dr. Grover Shannon.