5-2023

Genomic Selection for Yield and Seed Composition Stability in an Applied Soybean Breeding Program

Benjamin Harms

University of Nebraska - Lincoln

GENOMIC SELECTION FOR YIELD AND SEED COMPOSITION STABILITY

IN AN APPLIED SOYBEAN BREEDING PROGRAM

by

Benjamin Mark Harms

A THESIS

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Master of Science

Major: Agronomy

Under the Supervision of Professor David L. Hyten

Lincoln, Nebraska

May, 2023

GENOMIC SELECTION FOR YIELD AND SEED COMPOSITION STABILITY

IN AN APPLIED SOYBEAN BREEDING PROGRAM

Benjamin Mark Harms, M.S.

University of Nebraska, 2023

Advisor: David L. Hyten

Stability traits are of primary importance in plant breeding to ensure consistency in phenotype across a range of environments. However, selection efficiency and accuracy for stability traits can be hindered due to the requirement of obtaining phenotype data across multiple years and environments for proper stability analysis. Genomic selection is a method that allows prediction of a phenotype prior to observation in the field using genome-wide marker data and phenotype data from a training population. To assess prediction of stability traits, two elite-yielding soybean populations developed three years apart in the same breeding program were used. The individuals in each population were tested across three years and seven or more environments, allowing for calculation of observed stability and assessment of prediction accuracy. The primary goal of this research was to provide an overview of genomic selection for yield stability, protein content stability, and oil content stability in an applied soybean breeding program. Factors affecting prediction accuracy were assessed, including SNP density, SNP marker type, and stability measure. Briefly, predictive abilities were low across all stability traits and stability measures for prediction across populations, ranging from -0.01 to 0.37. During applied prediction of non-parametric measures for yield stability, we obtained rank coincidence of roughly 0.65. When individuals in the top half of predicted stability

are selected, roughly 65% of those individuals are expected to be in the top half of observed stability. For prediction of protein and oil content stability for static environmental variance stability, we obtained rank coincidence of 0.59 and 0.58. While predictive abilities were too low for use in a breeding program, rank coincidence gave more promising results for applied genomic selection for stability traits. With improvement in methods such as prediction model, SNP type, and greater training and validation phenotype environments, there is potential for genomic selection to effectively improve stability in a breeding program by implementing selection at an early stage when phenotype data are insufficient to select for stability.

Author's Acknowledgements

There are so many people to thank for their help and support on this thesis. I am extremely indebted to my advisor, Dr. David Hyten, who granted me the opportunity to pursue this project and provided expertise, guidance, support, and the freedom to develop throughout these three years. This endeavor would not have been possible without Dr. George Graef for contributing his expertise in developing this project, and I am very appreciative of the experience I have gained from being included in the soybean breeding group. I am also grateful to Dr. Réka Howard, who provided statistical expertise and support throughout the preparation of this thesis. Additionally, I thank the Nebraska Soybean Board for providing funding for this project.

I have been lucky to be part of the UNL soybean genomics lab and thank my lab colleagues for their support, their eagerness to offer help, and their patience with my inexperience at times while working on this thesis. The same goes for the UNL soybean breeding group, whom the phenotyping of this project would have been impossible without. I am very appreciative of their invaluable help and support with this thesis. Additionally, I thank the numerous faculty and graduate students who have provided support and valuable feedback on this thesis, and all other department staff who have made this possible.

Lastly, I thank my family and friends for their unwavering support.

Table of Contents

Figures

Supplementary Figures Continued

Tables

CHAPTER I

LITERATURE REVIEW

Stability Traits

Yield Stability Background

Yield stability is a trait of primary importance in the advancement and release of experimental soybean lines. Stability is the measure of phenotypic consistency of a particular genotype grown across a set of environments. Breeders strive to advance and release lines that display the most reliable yield possible within a specific geographic range. From afar, this goal seems relatively simple. Execution is not as straightforward. This review will cover the different categories of stability measures, when they best suit a breeder's goals, and identify the challenges and limitations of assessing yield stability in a breeding program.

Yield stability can be measured and categorized in a variety of ways. There are two primary concepts of phenotypic stability: static stability and dynamic stability. A soybean line with high static yield stability would display a consistent, unchanging yield across environments (Becker & Léon, 1988). Conversely, dynamic stability is measured as the sum of differences between a genotype's yield response and overall mean yield in a specific environment. Thus, dynamic stability accounts for changes in environment mean yield, and an expectation of higher yield in favorable environments and lower yield in environments that are generally less productive. With a wide assortment of stability measures available, a breeder must have a clear goal for selection before choosing a specific stability measure for evaluation of experimental lines.

From Becker and Leon's 1988 review paper, *Stability Analysis in Plant Breeding*, a seminal work summarizing the early developments of stability analysis, the static stability concept is not often used for yield stability, but more so for "traits that must be maintained at all cost" like disease resistance and quality traits (Becker & Léon, 1988). Though the ideal breeding line would display consistently high yield across all environments and a high degree of static stability, this is not a realistic expectation due to genotype by environment interaction (GxE) between lines and environments. While repeatability is lower across environments, the use of dynamic stability as selection criteria can ensure soybean lines have higher yield potential in favorable environments (Lin & Binns, 1991; Sneller & Dombek, 1997). Dynamic stability measures are more useful for breeders and farmers in areas across the United States where growing conditions are typically favorable. In United States cropping systems, environments are improved by inputs such as irrigation, fertilizer, herbicides, and pesticides. Breeding programs that use dynamic yield stability for a selection index can ensure their varieties have high yields in favorable environments and take advantage of additional inputs.

Dynamic Stability Measures

Wricke's ecovalence is a conventional univariate dynamic stability estimate which uses the partitioning of genotype by environment interaction across all environments to assess stability, with $W_i = 0$ being the most stable genotype (Becker & Léon, 1988; Wricke, 1962). Due to its simplicity, Wricke's ecovalence is a widely used stability measure for assessing phenotypic dynamic stability, and gives the same rankings as another popular dynamic stability measure - Shukla's stability variance (Shukla, 1972). The formula for Wricke's ecovalence is as follows:

$$W_i = \sum_j(X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}_{...})^2$$

where X_{ij} = Mean of genotype i in environment j, $\bar{X}_{i.}$ = Mean yield of genotype i across

environments, $\bar{X}_{j.}$ = Mean of environment j across genotypes, $\bar{X}_{...}$ = Overall mean.

Essentially, this stability measure calculates the deviation between a genotype's mean

yield in an environment and the population's overall environmental mean. This deviation

is summed across all environments. The most stable line exhibits the lowest Wricke's

ecovalence value. In addition to partitioning GxE interactions, dynamic stability can also

be calculated using a regression method (Becker & Léon, 1988). Deviation mean squares,

also known as deviation from regression, is a dynamic stability measure using the

regression approach. Its formula is shown below:

$$s^2{}_{di} = \frac{1}{E-2}[\sum_j(X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}_{..})^2 - (b_i - 1)^2 \sum_j(\bar{X}_{.j} - \bar{X}_{..})^2]$$

where E is the number of environments, and b_i is Finlay and Wilkinson's coefficient of

regression, often used as a static stability measure (Finlay & Wilkinson, 1963):

$$b_i = 1 + \frac{\sum_j(X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}_{..})(\bar{X}_{.j} - \bar{X}_{..})}{\sum_j(\bar{X}_{.j} - \bar{X}_{..})2}$$

Non-parametric stability measures are based on rank differences across

environments. These measures are an interesting alternative to parametric measures that

quantify GxE, as no assumption is needed of the distribution of phenotypic values

(Nassar & Hühn, 1987). Four of the primary non-parametric statistics are known as

Nassar and Huehn's statistics (Nassar & Hühn, 1987; Pour-Aboughadareh et al., 2022).

The first statistic ($S^{(1)}$) the mean of the absolute rank differences of a genotype across all

environments, calculated as:

$$S^{(1)} = 2\sum_j^{n-1}\frac{\sum_{j'=j+1}^n |r_{ij} - r'_{ij}|}{[N(N-1)]}$$

where r_{ij} is the rank of the *i*th genotype in the *j*th environment, and N is the number of

test environments. Calculation of the other three statistics is similar to $S^{(1)}$. The second

statistic ($S^{(2)}$) is the variance among the ranks over tested environments. The third statistic

($S^{(3)}$) is the sum of absolute deviations for each genotype relative to the mean of ranks.

And finally, $S^{(6)}$ is the sum of squares of rank for each genotype relative to the mean of

ranks. The lowest value of S-statistic is most desirable. Another non-parametric stability

measure is Kang's rank-sum (Kang, 1988). This method ranks lines based on both mean

yield and stability according to Shukla's stability variance, a dynamic univariate measure

(Shukla, 1972). Essentially, each genotype is assigned a whole number starting at one, as

the highest yielding line with lowest Shukla's stability variance. Kang's rank-sum assigns

the same weight to both mean yield and stability. The lowest value of rank sum is most

desirable.

Another common method of assessing stability is the additive main effects and

multiplicative interaction (AMMI) model. The AMMI model uses ANOVA to account

for main effects of genotype and environment, as most traditional stability measures do.

However, AMMI is unique in that PCA is included to partition residual variance and

quantify genotype by environment interaction. AMMI is thought to be more suitable for

partitioning multiplicative genotype by environment interactions via PCA, thereby

identifying interaction patterns in a way not possible with only main effects ANOVA

analysis (Zobel et al., 1988). Although calculation is much different, AMMI is thought to

be a dynamic stability measure in most cases (Pour-Aboughadareh et al., 2022). The

AMMI model is defined as:

$$Y_{ij} = \mu + g_i + e_j + \sum_{n=1}^{N} \lambda_k \gamma_{ik} \delta_{jk} + \rho_{ij} + \varepsilon_{ij}$$

where Υ_{ij} is the response variable of the ith genotype in the jth environment, μ is the overall mean, g_i is the fixed genotype effect, e_j is the fixed environmental effect, N is the number of principal components in the AMMI model, λ_k is a value of the K axis in principal component analysis, γ_{ik} is the genotype factor of the vector associated with λ_k from the interaction matrix, δ_{jk} is the environmental factor of the vector associated with λ_k from the interaction matrix, ρ_{ij} is the residual GxE interaction, and ε_{ij} is the random, independent error term with mean and distribution $\varepsilon_{ij} \sim N(0,\sigma^2)$.

Biology of Stability

Previous studies have aimed to increase understanding of how stability is controlled from a biological perspective. Among these are genome-wide association studies to identify genomic regions and genes influencing phenotypic consistency across environments, along with hormones and epistatic interactions that play a role in stability traits. Here, a few studies are highlighted and assess their possible impact on stability improvement in breeding programs.

In a study conducted by Xavier and colleagues (2018), a genome-wide association study was conducted for both yield genotype-environment interaction (GxE) and one stability measure to identify areas of the genome that are associated with stability. They found six significant QTL for GxE (20-35% of variance explained), and one significant QTL associated with the FW stability measure (1% of variance explained). This discrepancy in the number of significant loci and their variance explained between GxE and the mathematical stability parameter may show in part how the stability measure lacks to explain stability as compared to quantifying GxE explicitely, and these results were validated more recently (Happ et al., 2021). Also, the majority of stability QTL

discovered contained nearby genes generally known to play a role in stress response like water, salinity, flooding, photo-oxidative, and development under stress. This is expected because the ability for a genotype to maintain a consistent phenotype across environments can depend, at least in part, on its response in low-yielding environments. Often, these environments present some form of stress for the plant to overcome.

A recent review paper covered the plant hormone abscisic acid (ABA) and its role when plants are under stress conditions (Kavi Kishor et al., 2022). This review has a nice diagram showing that under stress environments such as drought, salinity, and temperature, ABA levels are elevated. Under increased ABA, stomata are closed, tillering is promoted (in grasses), and seed filling is decreased. Under basal level ABA when stresses are not present, the plant allocates more resources to seed filling, causing increased seed yields. It is mentioned that keeping ABA at a relatively basal level could help the plant to carry out seed filling under stress, making the control of ABA a possibly valuable way to maintain yield in stress environments. This, in turn, could help improve the stability of yield and seed components across environments.

Another study assessed overexpression of a MADS-box gene targeted during maize domestication, zmm28, for possible improvement of yield and stability across environments (Wu et al., 2019). This large-scale study found that overexpression of this single gene increased early plant vigor (plant height and leaf biomass) and total leaf area. These changes resulted in significantly positive and consistent yield increases across both low-yielding and high-yielding environments when compared to the control without overexpression of zmm28. This single gene could possibly contribute to maintaining consistently higher yields across both low and high-yielding environments. It may be

worth assessing other MADS-box genes or domestication genes that could have this same effect.

The effect of epistasis on yield stability was assessed in a study published in 2017 (Sehgal et al., 2017). They identified a combination of four markers that explained 20% of variance for yield stability coefficient in wheat. QTL identified for yield stability coefficient shared the same genomic regions as those reported for grain yield in drought and heat stress environments, similar to some of the stress response genes found in studies previously mentioned. In terms of epistasis, half of the markers associated with yield stability coefficient were involved in significant epistatic effects, meaning that the marker-marker interaction between these QTL caused significant differences in stability. There were also cases where the epistatic interactions between QTL were strong in most environments but not in irrigated environments, pointing again to the complexity of stress response in stability. These epistatic interactions were generally weaker for grain yield than for yield stability. More studies are needed to truly understand the full scope of how epistasis affects stability and its possible application in plant breeding.

To wrap up discussion on the possible ways that yield stability is controlled biologically, a review paper was published that covers many different types of stress, and how selection for specific genetic elements conferring adaptation to environmental stresses could improve stability across changing environments (Mickelbart et al., 2015). The stresses covered included flooding, drought, salinity, nutrient deficiencies, and temperature. Most examples in this review were assessing how genetic variation in major crop species or wild relatives of major crop species could be utilized for improvement of response in stress environments. One example is given of specific wild relative species to

wheat and maize with resistance to flooding, and these species display genetic variation with potential to be integrated via transgenic methods into primary crop species. Another example is in drought stress, where there are multiple ways to improve drought tolerance. One way to improve drought tolerance is through deeper rooting for enhanced water acquisition, which has known QTL in a variety of crops. The other way of improving drought tolerance is through suppressing ABA for improved yield, which may be possible by altering ABA receptors. In the end, applications of improving stability generally involve selection for a desired gene from diverse germplasm or transgenic integration of genes from similar species adapted to a specific stress. This review shows potential for improvement of stability by focusing on specific stresses that are common within a specific geographic region, and the importance of maintaining diversity within a breeding population to identify germplasm that are more tolerant to those specific stresses.

From the literature outlined, there are a variety of ways that stability of yield is biologically determined in plants. Many of these are related to biological mechanisms when plants are under stress environments including genes involved in stress response and development, hormone levels, and epistatic interactions between markers. Woven into this, there is also a level of gene by environment interaction which results in differential expression of genes, causing the up or downregulation of genes involved in transcription factors, kinases, and hormones to help the plant survive under stress (Le et al., 2012). From this, we can deduce that stability is controlled by genetic factors, environmental factors, and the interaction between genotype and environment. This

complex web of biological mechanisms, particularly under stress environments, may give rise to methods for improving stability traits in plant breeding programs.

Limitations

There are many limitations involved with assessing and interpreting yield stability values. GxE is a term used to describe the difference in response when a particular genotype is exposed to different environments (Comstock & Moll, 1963). As previously mentioned, the fact that yield stability is essentially the sum of GxE across environments is a substantial limitation. This means that it is difficult to partition variance for yield stability and determine heritability. Trait heritability is a fundamental factor in plant breeding, as it determines the amount of phenotypic variance for a trait due to genetic effects that can be passed to the next generation (Falconer & Mackay, 1996). A trait with low heritability is expected to be more difficult to improve and maintain in a breeding population, and more difficult to predict using genomic tools (H. Zhang et al., 2019). Also, the specificity of heritability values can add confusion to their interpretation for stability traits. Heritability is specific to populations and environments (Falconer & Mackay, 1996), and is also specific to the type of stability measure being used. It's been shown that static stability parameters are more heritable, or repeatable, than dynamic (Lin & Binns, 1991). Further, stability parameters within the same category (static or dynamic) are sometimes highly correlated, but otherwise not related (Sneller & Dombek, 1997). Because of this, the field of inference is dramatically reduced for calculated heritability values for stability traits, and values are less meaningful than for traits such as yield and plant height.

Yield stability measurements display low repeatability which is another limitation when researching stability traits. It has been shown that a yield stability measurement in one set of environments is not representative of a genotype's stability in a different set of environments (Pham & Kang, 1988; Sneller & Dombek, 1997). Measurement reliability is critical for any trait being used for selection in a breeding program. A lack of measurement repeatability increases the chance of inaccurate predictions and selections for yield stability in breeding programs, particularly if environmental conditions vary widely across years at a particular location. When advancing and releasing lines, breeders must choose a specific geographic region to test a population in. Data generated from this target region will ensure repeatability and predictability of performance. Even then, there are unavoidable differences across years in environmental conditions at specific geographic locations, which continuously cloud reliability of stability values.

The genetic complexity of yield stability as previously mentioned presents challenges when working with the trait. Yield stability, as previously mentioned, is essentially a measurement of GxE summed across a set of environments. Currently, there is little known about the genetic determinants of GxE in terms of stability. It has been shown that stability traits have complex genetic architecture due to a lack of stable QTL (Y. Wang et al., 2015). A 2016 study by Huang and colleagues reported evidence to suggest that GxE pattern variance can partially be explained by genetic elements when assessing AMMI stability measures (Huang et al., 2016). The multitude of stability parameters available only adds to the complexity of understanding the genetic control of stability traits. QTL discovery has been shown to differ between stability parameters (Happ et al., 2021). For example, AMMI and traditional stability estimates tend to rank

individuals differently because they measure stability differently. This results in different marker effects and significant QTL from GWAS analysis. Much like stability measurements, which have low repeatability and are environment-specific, discovered stability QTL are a direct result of the specific environments sampled, the quantity of environments, and the stability parameter used. Due to the lack of one agreed-upon method for stability analysis, confusion about the genetic control for stability traits remains, and tools such as marker assisted selection have not yet been applied for improvement of yield stability in breeding programs.

One of the most fundamental limitations when assessing stability traits is the amount of phenotypic data needed across environments and years for each genotype. Data from only one year, or only a few environments will not likely render a useful stability measure (Becker & Léon, 1988). The effect of this limitation on breeding programs is twofold. First, the amount of resources needed to assess stability is increased with more plots, field inputs, and labor for planting, maintenance, and harvest across multiple locations and years. According to Piepho, standard errors should be used to decide if the number of environments provides an adequate parameter estimate (which, to get a coefficient of variation of 20%, requires at least 50 environments) (Piepho, 1998). While lines advanced to the late stages of yield trials may eventually be tested in around 50 environments, this is an unrealistic standard to uphold for phenotypic stability analysis in the early to middle stages of the breeding pipeline. Even when 15 to 20 locations are used, low estimates of single-year stability will be observed (Becker & Léon, 1988). Generally, 10 to 15 environments across multiple years is an acceptable sample for calculating stability values, though a consensus has not been agreed upon. The large

amount of phenotypic data required for stability analysis increases the monetary cost to a breeding program. Another effect on breeding programs is the amount of time required to obtain valid stability data. Breeders must wait years before using stability as selection criteria in an experimental population. This could have detrimental effects on selection accuracy in earlier stages of yield trials when data from few locations are available.

Soybean seed composition

Soybean seeds are made up of two valuable components: protein and oil. Soybean is a premier source of vegetable protein primarily used for animal feed, and oil used mainly for human consumption and biofuels. The concentration of soybean seed protein and oil follow some general trends. An early study from Shannon and colleagues reported that protein and oil concentrations are negatively correlated (Shannon et al., 1972). From the same study, it was found that protein concentration and yield are negatively correlated, highlighting the difficulty of producing high protein soybean varieties while maintaining high oil content and/or yield. The results from this early study have been substantiated many times over the past 50 years. Also, heritability for protein and oil tends to be higher than yield heritability, meaning that genetic elements make up more phenotypic variance for seed composition than for yield (Assefa et al., 2019; Liu et al., 2019; X. Wang et al., 2014). Traits with higher heritability tend to be more accurately predicted using genomic selection (H. Zhang et al., 2019), and more effectively bred and selected for in breeding programs. Another interesting characteristic to note is that seed composition is impacted by maturity group. According to a recent meta-analysis of 21 studies across 11 states, varieties with higher maturity group (e.g., V, VI, VII) tend to have higher protein concentration (Assefa et al., 2019).

Seed composition stability is a measure of the consistency of seed protein and oil concentration across environments, years, and locations. Stability for soybean seed composition is perceived differently than yield stability. As previously mentioned, dynamic stability is normally used to measure yield stability because it specifically accounts for each environment based on population mean yield. This results in selected lines with increased ability to take advantage of favorable environments. For seed composition, static stability is likely the more suitable measurement, as breeders are trying to uphold a minimum concentration in order to develop the most profitable and useful product for farmers. For farmers selling the crop, it is important to meet a specific minimum concentration in both protein and oil for the seed to be economically viable. Thus, stability measures such as environmental variance (S^2) and regression coefficient (b_i) are of interest for seed composition (Becker & Léon, 1988; Finlay & Wilkinson, 1963; Roemer, 1917). The FW regression coefficient is shown above in calculation of dynamic deviation from regression. While regression coefficient can be used to identify dynamic stability, a genotype stable according to the static concept has $b_i = 0$. Environmental variance, one of the earliest-developed stability measures, is calculated as:

$$S^2 = \sum \frac{(R_{ij} - m_i)^2}{e - 1}$$

where R_{ij} is the yield of the ith genotype in the jth environment, m_i is the grand mean of genotype i across all environments, and e is the number of environments. Minimum values of S^2 show the greatest stability.

GxE interaction effects are generally smaller for seed composition than yield (Whaley & Eskandari, 2019). A high level of static stability is more attainable with seed composition than yield, as the trait is less genetically complex, has less GxE across

environments, and higher heritability. In a study assessing seed quality trait stability, it was observed that dynamic stability could not adequately account for the GxE pattern of seed quality traits in wheat (Huang et al., 2016). This indicates that static stability may be more suitable for measuring seed composition stability.

Genomic Selection

Background

Traditional breeding methods, such as recurrent phenotypic selection, have been used for thousands of years to advance lines with high genetic potential for human use. Over the course of many generations, recurrent phenotypic selection increases the frequency of favorable alleles in a population by the intermating and selection of lines that display a desired phenotype. While this method is proven effective, it is quite time consuming as years of selection and intermating are needed to obtain a population of desired genotypes. Today, the world population continues to grow which continually increases the demand for crops. With challenges to food production mounting, such as the diminishing area of agricultural land, climate change, and water scarcity, genetic gains from plant breeding must be at the forefront of an effort to increase yield in primary crop species.

With recent developments in genome sequencing technology and subsequent decrease in cost of sequencing, methods that make use of sequence data for use in marker assisted selection and genomic selection have become a focal point in the goal to accelerate genetic gain in plant breeding programs (Bernardo, 2008; Theo Meuwissen, 2007). One of the most popular methods utilizing molecular data in breeding is genomic prediction. Genomic prediction is a method in which genome-wide markers are used to

predict the phenotype of untested individuals, allowing for timely selection and advancement of superior lines (Meuwissen et al., 2001). The effectiveness of genomic prediction can be attributed to one overarching concept: that individuals with similar genotype will share a similar phenotypic resemblance. The potential benefits of using genomic prediction in plant breeding are vast, and it has been shown to increase the rate of genetic gain in breeding programs (Bernardo & Yu, 2007; Heffner et al., 2010). Once the foundation of genomic prediction is built by gathering appropriate populations with high quality phenotype and genotype data, this method can be a powerful tool for predicting and selecting for complex traits in lines that have not yet been planted in the field.

Genomic selection process

The process of genomic prediction starts with choosing a training population. This is a group of individuals that become the basis for predictions using their genomic data and observed phenotype data for a trait of interest. Training populations are critical to the success of genomic prediction. When the training population is more closely related to the selection population, a higher prediction accuracy can be expected than if the populations share less relation (Theo Meuwissen, 2007). Generally, an ideal training population should be carefully chosen so that it is closely related to the selection population, while having a high degree of genetic diversity (Werner et al., 2020). Training population size increase can also benefit prediction accuracy, though not to the degree that relatedness and diversity can affect prediction accuracy (Edwards et al., 2019). For a breeder, this means ensuring that training population individuals share some

relatedness with the selection population via pedigree while having as much genetic

diversity as possible.

After a breeder chooses a training population specific to a selection population,

the appropriate data must be acquired for each individual line. In plant breeding,

phenotype data for traits of interest are often collected from field observations. These

traits include measurements such as yield, days to maturity, and plant height. Typically,

traits are measured in multiple environments with two or more replicates per

environment. A two-stage phenotypic analysis, common in modern breeding programs,

splits phenotypic analysis of multi-environment trials into two parts (Möhring & Piepho,

2009). The first part gives an adjustment of a genotype's response in each environment

considering experimental design (Best Linear Unbiased Estimator or BLUE), and the

second uses environmental BLUEs from stage one to predict each line's performance

across all environments (Best Linear Unbiased Predictor or BLUP; Henderson, 1984).

These BLUP values are directly comparable even in unbalanced data sets, making BLUP

analysis a critical tool in plant breeding where unbalanced data and incomplete blocks are

common. If an experimental design includes diagonal checks, spatial adjustments can be

made in the calculation of BLUEs. Spatial adjustments account for field heterogeneity

but add an extra degree of model complexity and computational load without providing

noticeable benefits to prediction accuracy (Bernal-Vasquez et al., 2014).

When analyzing phenotypic stability, this process is slightly different. One

common way to calculate stability is to obtain the average of all reps in each environment

for each genotype (Pour-Aboughadareh et al., 2019). Similar to the two-stage analysis,

stability can also be assessed by estimating BLUEs for each line within each

environment. In this approach, a mixed model is fitted with genotype as a fixed effect, and experimental design factors including block and rep as random effects. An estimate of genotype performance adjusted for experimental design in each environment is acquired through this analysis. So, stability values can be calculated either by using environmental averages or environmental BLUEs for each genotype as input, before calculation using a specific stability formula.

In addition to phenotypic data for the training population, genomic data must be obtained for each line in both the training and selection populations. Genomic data are typically acquired through DNA sequencing in the form of single nucleotide polymorphisms (SNPs). In whole-genome sequencing, millions of SNPs are acquired in positions scattered across the entire genome. If avoiding ascertainment bias, the small subset of SNPs used for genomic prediction should be a random sample of the polymorphisms present within a population. When the marker set for genomic prediction is sequenced as a predetermined set of markers (SNP array), or marker selection from whole genome sequence data is not random, the markers have an ascertainment bias and are not a representative sample of the polymorphisms in a population. Though predetermined SNP arrays are a common way to conduct genomic selection, it induces ascertainment bias. With increased levels of ascertainment bias, any inference made about allele frequencies in a population, such as diversity or differentiation, will have increased levels of inaccuracy (Heslot et al., 2013; Moragues et al., 2010). In a comparison of one representative marker set and one displaying ascertainment bias, genomic prediction accuracies were not significantly different for three of four traits in winter wheat (Heslot et al., 2013). When using whole genome sequence data, sometimes

SNPs with greater effect on a trait from GWAS (genome-wide association study) are used for genomic prediction. Consistently, the prediction accuracy when using these SNPs has been shown to be higher than random SNPs (Brøndum et al., 2015; Luo et al., 2021; van den Berg et al., 2016).

After genomic data are obtained, a genomic relationship matrix of all training and selection population individuals must be assembled. GBLUP is a method of genomic prediction which utilizes a genomic relationship matrix. The genomic relationship matrix uses marker data to give a measure of similarity between each pair of individuals at the genomic level. This similarity is then used to make predictions about an untested individual's phenotype and breeding value as a parent. One of the most popular genomic relationship matrices was proposed by VanRaden in 2008. The VanRaden matrix uses observed genome-wide marker data and minor allele frequencies to assign weights to individual marker relationship and ultimately genomic relationship between two individuals (VanRaden, 2008). The relationship matrix is what makes GBLUP more computationally efficient than other genomic prediction methods such as RR-BLUP, which estimates the effect of each individual marker, and uses the sum of all marker effects to predict a phenotype.

Prediction Models

After all the preparation work is completed, a genomic prediction model must be trained. This step is relatively simple to conduct. GBLUP, an additive linear model, is one of the most widely used prediction models. The standard additive GBLUP prediction model equation is (Henderson, 1984; T. Meuwissen et al., 2001):

$$y = Xb + Zu + e$$

where **y** is the vector of phenotypes from the training population, **X** is the design matrix for fixed effects, **b** is the vector of estimated fixed effects, **Z** is the design matrix for random effects, **u** is the vector of predicted random effects, and **e** is the vector of residuals. Often, fixed effects are estimated apart from genomic prediction, further simplifying the model by removing the fixed effects term. The resulting model consists of observed phenotypes as the response, the genomic relationship matrix with genotype as a random effect, and an intercept of 1 for fixed effects. Genotype is fit as a random term because the goal is to predict genotype performance in the form of BLUPs. The resulting GBLUPs are predictors of performance for each genotype in the same form as the model response vector. BLUPs account for unknown variation by shrinking predicted values closer to the overall mean. These values can then be ranked, and inferences can be made about future performance of individuals.

There are other types of genomic prediction models. Ridge regression BLUP (rrBLUP) is equivalent to GBLUP. While GBLUP makes predictions based on genomic relationship, rrBLUP estimates marker effects for the trait of interest and sums an individual's markers to predict a phenotype. The standard additive GBLUP model has been shown to give similar accuracies to non-parametric Bayesian and semi-parametric RKHS (Random kernel Hilbert spaces) (Huang et al., 2016; Jarquín, Kocak, et al., 2014; Meher et al., 2022). When accounting for environment and GxE in the GBLUP model, prediction accuracies have been shown to improve upon the standard additive GBLUP model with no additional terms, and artificial intelligence models for the prediction of yield in wheat and soybean, protein and oil content in soybean, and drought tolerance in

maize (Burgueño et al., 2012; Dias et al., 2018; Jarquín, Crossa, et al., 2014; Lopez-Cruz et al., 2015; Ray et al., 2022; X. Zhang et al., 2015).

Model performance assessment

After prediction is conducted, results need to be interpreted in the context of model performance and its application. The standard method for model performance assessment is cross validation. Cross validation is the repeated prediction of different iterations of training and validation sets in a population with phenotype data available for all individuals. The most widely used cross validation method in genomic prediction studies is k-fold cross validation. With k-fold cross validation, a population is randomly split into k groups, indicating the number of iterations of training and test sets. With each iteration, a different group is predicted with the remaining k-1 groups as the training set. After all iterations have been run, prediction accuracies are calculated.

Prediction accuracy and predictive ability are two terms that refer to different measures of performance. Predictive ability is commonly used to evaluate accuracy of predictions by calculating the correlation between predicted values and observed phenotype. This is a simple method to assess the effectiveness of prediction when phenotype data are available for predicted individuals, as often the case in cross-validation within a single population. Prediction accuracy is the correlation between predicted breeding values and true breeding values, which are commonly not known in genomic prediction scenarios (Ould Estaghvirou et al., 2013). An estimate of true prediction accuracy can be calculated by dividing predictive ability by the square root of broad-sense heritability for the trait. Prediction accuracy estimates can result in values

greater than 1, which is why predictive ability estimates are much more commonly used, ranging from -1 to 1.

Cross validation is a common method for comparing the performance of two models, making it a useful tool in the process of model development and validation. The shortcoming of cross validation within a closely related population is the lack of inference about model performance in a real-life scenario. In an applied breeding program, genomic prediction is most valuable when the phenotype of an untested individual is predicted. Normally, the training population in this scenario is not in the same family as the individual being predicted. Cross validation is generally conducted within a single population of related individuals, resulting in inflated prediction accuracies due to population structure within the test population (Werner et al., 2020). Studies with the goal of assessing prediction accuracy in a real-life scenario are rare because genotyping and phenotyping a separate population doubles cost and increases the length of project completion.

Factors affecting prediction accuracy

It has been shown that marker density, statistical method, minor allele frequency, heritability, and genetic architecture are five influential factors affecting prediction accuracy (H. Zhang et al., 2019). It is also known that when training population and selection population individuals are not closely related, a lower prediction accuracy can be expected (Meuwissen, 2007). Generally, a training population should be carefully chosen so that it is closely related to the selection population, while having a high degree of genetic diversity (Werner et al., 2020). These two criteria are crucial to obtaining the full prediction potential from a genomic selection model. One of the fundamental

principles of genomic prediction is that two closely related individuals share a similar phenotype. If two individuals are phenotyped in different environments, have vastly different environmental sensitivity (Falconer & Mackay, 1996), or the trait of interest has low heritability, reduced prediction accuracy is expected. Generally, traits with low heritability are more difficult to predict because phenotypes are less reflective of their genetic content (Resende et al., 2012). Also mentioned is SNP marker density. As the number of SNP markers used in genomic prediction increases, prediction accuracy is increased until a threshold is reached where increasing SNP number no longer benefits prediction accuracy (H. Zhang et al., 2019). Traits with complex genetic architecture tend to exhibit a greater response to increasing marker density, which is expected of the yield stability and seed composition stability traits. These are among the components that can largely be controlled by the researcher and must be considered when conducting genomic selection.

Genomic Prediction for Stability Traits

Few studies have been published attempting genomic prediction for stability traits. These studies can be examined to compare methodology and results and provide valuable insight on ways to improve methods in the future. The processes in both stability estimation and genomic prediction involve many decisions and possible directions.

One recent study regarding genomic selection for stability traits grew 273 elite soft winter wheat breeding lines across two years and 12 total environments, primarily with one rep per environment (Huang et al., 2016). After adjustment of yield phenotype values via BLUPs, stability was assessed for four agronomic and seven seed quality traits using AMMI stability (SIPC – sum of IPC scores) and conventional Eberhart and Russel

regression (ERR). The VanRaden genomic relationship matrix was constructed using 3,919 SNPs (VanRaden, 2008). Four different prediction models were used, including RR-BLUP, Bayesian ridge regression (BRR), reproducing kernel Hilbert spaces (RKHS), and elastic net (EN). Prediction accuracy was calculated as the correlation between predicted GEBVs and observed phenotype BLUPs. Principal Component analysis (PCA) was also conducted to define population structure within the 273 wheat lines. Two populations were determined, one large population of 235 lines (LG), and one small population of 38 lines (SG). These populations were used in two scenarios to observe the effect of population structure on prediction accuracy for stability traits.

The prediction accuracy of the four models demonstrated that rrBLUP performed best overall with the first or second highest accuracy for 21 of 22 traits. In addition, prediction with a larger set of 13,198 SNPs had equal or lower accuracy for most traits compared to a set of 3,919 SNPs. Due to the increase in computational efficiency, the smaller SNP set was used for all subsequent genomic predictions. Since double shrinkage from BLUPs to GEBVs could be problematic, the effect on accuracy between using BLUEs and BLUPs for the training phenotype vector was assessed and showed identical prediction accuracy when using either one. This identical prediction accuracy was likely due to the high heritability of the traits. Interestingly, they found that prediction accuracy was higher for yield stability (0.44) than yield (0.33).

The highest accuracy for SIPC-AMMI stability was achieved when only the first two principal components were used for yield stability calculation. In the discussion, this was equated to the later IPCs being composed of more background GEI noise (Gauch Jr., 1988). The first few IPCs give a better estimate of GEI pattern, and since this resulted in

higher prediction accuracies, the idea was reinforced that stability estimates based on GEI pattern are under genetic control (Sneller & Dombek, 1997). SIPC-AMMI also had higher prediction accuracy than ERR stability in every prediction scenario. The two dynamic stability measures used in this study were also used for seed quality traits and resulting prediction accuracies were generally lower than yield stability predictions. Other studies have reported higher prediction accuracies for static stability in seed quality traits, so it is possible that a static measure can better capture the GxE pattern across environments for seed composition stability (Y. Wang et al., 2015). Overall, this study raises interesting points about stability measures capturing different patterns of GEI that may be under genetic control and, therefore has potential to be predicted with moderate accuracy through genomic selection. While moderate prediction accuracies were obtained in cross validation among genetically similar individuals, prediction accuracy was inadequate across less-related populations.

Another study, published in 2021 by O'Connor and colleagues, assessed genomic prediction accuracy for yield and yield stability in Australian macadamia tree production (O'Connor et al., 2021). One of the primary goals of this study was to increase the rate of genetic gain in a macadamia nut breeding program, and is the first genomic prediction study in macadamia nut. The study consisted of 295 unreplicated macadamia tree lines from 32 families. The trees were grown across 4 sites over 8 years, with phenotype data collected from years 5 to 8. From the four years of yield data, BLUPs were calculated for individual tree performance across all four years (one BLUP value per tree) and within each year (four BLUP values per tree). Instead of using traditional or AMMI stability measures, yield stability was calculated as the standard deviation of the four year-specific

BLUPs for each tree. The output standard deviations were then used as the phenotype

basis for the genomic prediction model.

Predictive ability was determined by comparing the resulting GEBVs to the

previously calculated yield SD values. Cross validation was used within the population in

two ways: 1) randomization of all individuals into five sets and 2) individuals grouped by

family and related families into five sets. Scenario 2 represents more of the expected

situation when predicting in a real-world breeding program. Predictive ability was

reported as ranging from 0.03 to 0.29 among random grouping of CV sets, and -0.09 to

0.12 among family grouping of CV sets. These results are expected because in random

grouping, each individual in the CV set will likely be a sibling or share a parent with an

individual in the training population (TP). This high degree of genetic relatedness, as

previously stated, is known to increase prediction accuracy and predictive ability. As a

result, they estimate that genomic selection would double genetic gains compared to

traditional selection methods in a macadamia nut breeding program despite low

prediction accuracy.

There are multiple reasons why prediction accuracies were low in this study,

including the genetic complexity of yield and yield stability, the species tested, the

unreplicated experimental design, and the unconventional calculation of yield stability.

This paper further proved the broad potential and applicability of genomic prediction to

increase the rate of genetic gain, even for a crop species lagging in development of

genomic tools. Ultimately, this study provides an example of translating genomic

prediction methods to a less-established species, a different method of predicting yield

stability, and interpretation of results using a cross validation scheme with similarity to a real-world breeding scenario.

A 2020 study assessed genomic prediction and GWAS for the stability of yield, heading date, and plant height in a population of 456 diverse winter wheat lines evaluated in 9 environments (Lozada & Carter, 2020). The genomic prediction model used was RR-BLUP five-fold cross validation, and three stability parameters were predicted, including Findlay-Wilkinson (FW, traditional static), ASI (AMMI stability index), and ASV (AMMI stability value). FW is a static stability parameter, which is not commonly used for yield stability calculation.

The mean prediction accuracies for yield stability were 0.25, 0.28, and 0.49 for ASI, ASV, and FW respectively. They found that high FW stability was often found with higher yielding lines, indicating that this stability index had likely been selected simultaneously with high yielding genotypes. Thus, FW stability has a more robust genetic basis in this population than other stability measures. The study used four IPC to calculate AMMI stability, which resulted low prediction accuracies. While all IPC were responsible for significant effects, the combination did not capture enough GxE pattern under genetic control to give high prediction accuracies. Other studies, such as Huang et al., 2016 mentioned earlier, demonstrate the advantage of using fewer IPCs to reduce background noise and separate the GxE under genetic control. Overall, it is valuable to know the degree to which accuracy decreases when too many IPC are used, and that traditional stability measures can be under greater genetic control when under selection with mean yield in a breeding program.

Summary

There are important results from these studies that can be translated to new methods to predict stability traits. Experimental design and methods are crucial to agronomic trials, and specifically those with a focus on stability. We've found from these studies that lower prediction accuracy can result from experimental design flaws (fewer environments sampled, no replication), the method of calculating stability (AMMI vs. traditional), and the prediction model used. This collection of work can inform future studies looking to improve prediction for stability traits and develop methods for effective use in real-world breeding programs.

References

Assefa, Y., Purcell, L. C., Salmeron, M., Naeve, S., Casteel, S. N., Kovács, P., Archontoulis, S., Licht, M., Below, F., Kandel, H., Lindsey, L. E., Gaska, J., Conley, S., Shapiro, C., Orlowski, J. M., Golden, B. R., Kaur, G., Singh, M., Thelen, K., … Ciampitti, I. A. (2019). Assessing variation in us soybean seed composition (protein and oil). *Frontiers in Plant Science*, *10*(March). https://doi.org/10.3389/fpls.2019.00298

Becker, H. C., & Léon, J. (1988). Stability Analysis in Plant Breeding. *Plant Breeding*, *101*(1), 1–23. https://doi.org/10.1111/j.1439-0523.1988.tb00261.x

Bernal-Vasquez, A. M., Möhring, J., Schmidt, M., Schönleben, M., Schön, C. C., & Piepho, H. P. (2014). The importance of phenotypic data analysis for genomic prediction - a case study comparing different spatial models in rye. *BMC Genomics*, *15*(1), 1–17. https://doi.org/10.1186/1471-2164-15-646

Bernardo, R. (2008). Molecular markers and selection for complex traits in plants: Learning from the last 20 years. *Crop Science*, *48*(5), 1649–1664.

https://doi.org/10.2135/cropsci2008.03.0131

Bernardo, R., & Yu, J. (2007). Prospects for genomewide selection for quantitative traits in maize. *Crop Science*, *47*(3), 1082–1090. https://doi.org/10.2135/cropsci2006.11.0690

Brøndum, R. F., Su, G., Janss, L., Sahana, G., Guldbrandtsen, B., Boichard, D., & Lund, M. S. (2015). Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction. *Journal of Dairy Science*, *98*(6), 4107–4116. https://doi.org/10.3168/jds.2014-9005

Burgueño, J., de los Campos, G., Weigel, K., & Crossa, J. (2012). Genomic prediction of breeding values when modeling genotype × environment interaction using pedigree and dense molecular markers. *Crop Science*, *52*(2), 707–719. https://doi.org/10.2135/cropsci2011.06.0299

Comstock, R. E., & Moll, R. H. (1963). Genotype environment interactions. *Statistical Genetics and Plant Breeding*, *982*, 164–196.

Dias, K. O. D. G., Gezan, S. A., Guimarães, C. T., Nazarian, A., Da Costa E Silva, L., Parentoni, S. N., De Oliveira Guimarães, P. E., De Oliveira Anoni, C., Pádua, J. M. V., De Oliveira Pinto, M., Noda, R. W., Ribeiro, C. A. G., De Magalhães, J. V., Garcia, A. A. F., De Souza, J. C., Guimarães, L. J. M., & Pastina, M. M. (2018). Improving accuracies of genomic predictions for drought tolerance in maize by joint modeling of additive and dominance effects in multi-environment trials. *Heredity*, *121*(1), 24–37. https://doi.org/10.1038/s41437-018-0053-6

Edwards, S. M. K., Buntjer, J. B., Jackson, R., Bentley, A. R., Lage, J., Byrne, E., Burt, C., Jack, P., Berry, S., Flatman, E., Poupard, B., Smith, S., Hayes, C., Gaynor, R. C.,

Gorjanc, G., Howell, P., Ober, E., Mackay, I. J., & Hickey, J. M. (2019). The effects of training population design on genomic prediction accuracy in wheat. *Theoretical and Applied Genetics*, *132*(7), 1943–1952. https://doi.org/10.1007/s00122-019-03327-y

Falconer, D. S., & Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics* (Fourth). Longman.

Finlay, K. W., & Wilkinson, G. N. (1963). The analysis of adaptation in a plant-breeding programme. *Australian Journal of Agricultural Research*, *14*(6), 742–754. https://doi.org/10.1071/AR9630742

Gauch Jr., H. G. (1988). Model Selection and Validation for Yield Trials with Interaction. *International Biometric Society*, *44*(3), 705–715.

Happ, M. M., Graef, G. L., Wang, H., Howard, R., Posadas, L., & Hyten, D. L. (2021). Comparing a Mixed Model Approach to Traditional Stability Estimators for Mapping Genotype by Environment Interactions and Yield Stability in Soybean [Glycine max (L.) Merr.]. *Frontiers in Plant Science*, *12*(March), 1–12. https://doi.org/10.3389/fpls.2021.630175

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

Henderson, C. R. (1984). *Applications of Linear Models in Animal Breeding*. University of Guelph, Guelph, Ontario, Canada.

Heslot, N., Rutkoski, J., Poland, J., Jannink, J. L., & Sorrells, M. E. (2013). Impact of Marker Ascertainment Bias on Genomic Selection Accuracy and Estimates of

Genetic Diversity. *PLoS ONE*, *8*(9). https://doi.org/10.1371/journal.pone.0074612

Huang, M., Cabrera, A., Hoffstetter, A., Griffey, C., Van Sanford, D., Costa, J., McKendry, A., Chao, S., & Sneller, C. (2016). Genomic selection for wheat traits and trait stability. *Theoretical and Applied Genetics*, *129*(9), 1697–1710. https://doi.org/10.1007/s00122-016-2733-z

Jarquín, D., Crossa, J., Lacaze, X., Du Cheyron, P., Daucourt, J., Lorgeou, J., Piraux, F., Guerreiro, L., Pérez, P., Calus, M., Burgueño, J., & de los Campos, G. (2014). A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theoretical and Applied Genetics*, *127*(3), 595–607. https://doi.org/10.1007/s00122-013-2243-1

Jarquín, D., Kocak, K., Posadas, L., Hyma, K., Jedlicka, J., Graef, G., & Lorenz, A. (2014). Genotyping by sequencing for genomic prediction in a soybean breeding population. *BMC Genomics*, *15*(1), 1–10. https://doi.org/10.1186/1471-2164-15-740

Kang, M. S. (1988). A RANK-SUM METHOD FOR SELECTING HIGH-YIELDING , STABLE CORN GENOTYPES Author (s): M . S . Kang Published by : Akadémiai Kiadó Stable URL : http://www.jstor.org/stable/23782771 Accessed : 28-06-2016 09 : 15 UTC. *Cereal Research Communications*, *16*(1–2), 113–115.

Kavi Kishor, P. B., Tiozon, R. N., Fernie, A. R., & Sreenivasulu, N. (2022). Abscisic acid and its role in the modulation of plant growth, development, and yield stability. *Trends in Plant Science*, *27*(12), 1283–1295. https://doi.org/10.1016/j.tplants.2022.08.013

Le, D. T., Nishiyama, R., Watanabe, Y., Tanaka, M., Seki, M., Ham, L. H., Yamaguchi-Shinozaki, K., Shinozaki, K., & Tran, L. S. P. (2012). Differential Gene Expression

in Soybean Leaf Tissues at Late Developmental Stages under Drought Stress

Revealed by Genome-Wide Transcriptome Analysis. *PLoS ONE*, *7*(11).

https://doi.org/10.1371/journal.pone.0049522

Lin, C. S., & Binns, M. R. (1991). Genetic properties of four types of stability parameter.

Theoretical and Applied Genetics, *82*(4), 505–509.

Liu, L., Prenger, E., Zhang, J., Little, B., Mian, M. A. R., & Li, Z. (2019). Impact of

genotype, seed composition, agronomic trait, and environment on soybean test

weight. *Journal of Crop Improvement*, *33*(6), 711–729.

https://doi.org/10.1080/15427528.2019.1659205

Lopez-Cruz, M., Crossa, J., Bonnett, D., Dreisigacker, S., Poland, J., Jannink, J. L.,

Singh, R. P., Autrique, E., & de los Campos, G. (2015). Increased prediction

accuracy in wheat breeding trials using a marker × environment interaction genomic

selection model. *G3: Genes, Genomes, Genetics*, *5*(4), 569–582.

https://doi.org/10.1534/g3.114.016097

Lozada, D. N., & Carter, A. H. (2020). Insights into the genetic architecture of

phenotypic stability traits in winter wheat. *Agronomy*, *10*(3).

https://doi.org/10.3390/agronomy10030368

Luo, Z., Yu, Y., Xiang, J., & Li, F. (2021). Genomic selection using a subset of SNPs

identified by genome-wide association analysis for disease resistance traits in

aquaculture species. *Aquaculture*, *539*(December 2020), 736620.

https://doi.org/10.1016/j.aquaculture.2021.736620

Meher, P. K., Rustgi, S., & Kumar, A. (2022). Performance of Bayesian and BLUP

alphabets for genomic prediction: analysis, comparison and results. *Heredity*,

128(6), 519–530. https://doi.org/10.1038/s41437-022-00539-9

Meuwissen, T., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*(4), 1819–1829.

Meuwissen, Theo. (2007). Genomic selection : Marker assisted selection on a genome wide scale. *Journal of Animal Breeding and Genetics*, *124*(6), 321–322. https://doi.org/10.1111/j.1439-0388.2007.00708.x

Mickelbart, M. V., Hasegawa, P. M., & Bailey-Serres, J. (2015). Genetic mechanisms of abiotic stress tolerance that translate to crop yield stability. *Nature Reviews Genetics*, *16*(4), 237–251. https://doi.org/10.1038/nrg3901

Möhring, J., & Piepho, H. P. (2009). Comparison of weighting in two-stage analysis of plant breeding trials. *Crop Science*, *49*(6), 1977–1988. https://doi.org/10.2135/cropsci2009.02.0083

Moragues, M., Comadran, J., Waugh, R., Milne, I., Flavell, A. J., & Russell, J. R. (2010). Effects of ascertainment bias and marker number on estimations of barley diversity from high-throughput SNP genotype data. *Theoretical and Applied Genetics*, *120*(8), 1525–1534. https://doi.org/10.1007/s00122-010-1273-1

Nassar, R., & Hühn, M. (1987). Studies on Estimation of Phenotypic Stability : Tests of Significance for Nonparametric Measures of Phenotypic Stability Author (s): R . Nassar and M . Hühn Published by : International Biometric Society Stable URL : http://www.jstor.org/stable/2531947. *Biometrics*, *43*(1), 45–53.

O'Connor, K. M., Hayes, B. J., Hardner, C. M., Alam, M., Henry, R. J., & Topp, B. L. (2021). Genomic selection and genetic gain for nut yield in an Australian macadamia breeding population. *BMC Genomics*, *22*(1), 1–12.

https://doi.org/10.1186/s12864-021-07694-z

Ould Estaghvirou, S. B., Ogutu, J. O., Schulz-Streeck, T., Knaak, C., Ouzunova, M., Gordillo, A., & Piepho, H. P. (2013). Evaluation of approaches for estimating the accuracy of genomic prediction in plant breeding. *BMC Genomics*, *14*(1). https://doi.org/10.1186/1471-2164-14-860

Pham, H. N., & Kang, M. S. (1988). Interrelationships among and Repeatability of Several Stability Statistics Estimated from International Maize Trials. *Crop Science*, *28*(6), 925–928. https://doi.org/10.2135/cropsci1988.0011183x002800060010x

Piepho, H. P. (1998). Methods for comparing the yield stability of cropping systems - A review. *Journal of Agronomy and Crop Science*, *180*(4), 193–213. https://doi.org/10.1111/j.1439-037X.1998.tb00526.x

Pour-Aboughadareh, A., Khalili, M., Poczai, P., & Olivoto, T. (2022). Stability Indices to Deciphering the Genotype-by-Environment Interaction (GEI) Effect: An Applicable Review for Use in Plant Breeding Programs. *Plants*, *11*(3). https://doi.org/10.3390/plants11030414

Pour-Aboughadareh, A., Yousefian, M., Moradkhani, H., Poczai, P., & Siddique, K. H. M. (2019). STABILITYSOFT: A new online program to calculate parametric and non-parametric stability statistics for crop traits. *Applications in Plant Sciences*, *7*(1), 1–6. https://doi.org/10.1002/aps3.1211

Ray, S., Jarquin, D., & Howard, R. (2022). Comparing artificial-intelligence techniques with state-of-the-art parametric prediction models for predicting soybean traits. *Plant Genome*, *August 2021*, 1–19. https://doi.org/10.1002/tpg2.20263

Resende, J. F. R., Muñoz, P., Resende, M. D. V., Garrick, D. J., Fernando, R. L., Davis,

J. M., Jokela, E. J., Martin, T. A., Peter, G. F., & Kirst, M. (2012). Accuracy of genomic selection methods in a standard data set of loblolly pine (Pinus taeda L.). *Genetics*, *190*(4), 1503–1510. https://doi.org/10.1534/genetics.111.137026

Roemer, J. (1917). Sinde die ertagdreichen Sorten ertagissicherer. *DLG-Mitt*, *32*(1), 87–89.

Sehgal, D., Autrique, E., Singh, R., Ellis, M., Singh, S., & Dreisigacker, S. (2017). Identification of genomic regions for grain yield and yield stability and their epistatic interactions. *Scientific Reports*, *7*(November 2015), 1–12. https://doi.org/10.1038/srep41578

Shannon, G., Wilcox, J., & Probst, A. (1972). Estimated Gains from Selection for Protein and Yield in the F4 Generation of Six Soybean Populations. *Crop Science*, *12*(6), 824–826.

Shukla, G. K. (1972). Some statistical aspects of partitioning genotype-environmental components of variability. *Heredity*, *29*(2), 237–245. https://doi.org/10.1038/hdy.1972.87

Sneller, C. H., & Dombek, D. (1997). Repeatability of Yield Stability Statistics in Soybean. *Crop Science*, *34*, 383–390.

van den Berg, I., Boichard, D., Guldbrandtsen, B., & Lund, M. S. (2016). Using sequence variants in linkage disequilibrium with causative mutations to improve across-breed prediction in dairy cattle: A simulation study. *G3: Genes, Genomes, Genetics*, *6*(8), 2553–2561. https://doi.org/10.1534/g3.116.027730

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, *91*(11), 4414–4423. https://doi.org/10.3168/jds.2007-0980

Wang, X., Jiang, G. L., Green, M., Scott, R. A., Song, Q., Hyten, D. L., & Cregan, P. B. (2014). Identification and validation of quantitative trait loci for seed yield, oil and protein contents in two recombinant inbred line populations of soybean. *Molecular Genetics and Genomics : MGG*, *289*(5), 935–949. https://doi.org/10.1007/s00438-014-0865-x

Wang, Y., Mette, M. F., Miedaner, T., Wilde, P., Reif, J. C., & Zhao, Y. (2015). First insights into the genotype-phenotype map of phenotypic stability in rye. *Journal of Experimental Botany*, *66*(11), 3275–3284. https://doi.org/10.1093/jxb/erv145

Werner, C. R., Gaynor, R. C., Gorjanc, G., Hickey, J. M., Kox, T., Abbadi, A., Leckband, G., Snowdon, R. J., & Stahl, A. (2020). How Population Structure Impacts Genomic Selection Accuracy in Cross-Validation: Implications for Practical Breeding. *Frontiers in Plant Science*, *11*(December), 1–14. https://doi.org/10.3389/fpls.2020.592977

Whaley, R., & Eskandari, M. (2019). Genotypic main effect and genotype-by-environment interaction effect on seed protein concentration and yield in food-grade soybeans (Glycine max (L.) Merrill). *Euphytica*, *215*(2), 1–17. https://doi.org/10.1007/s10681-019-2344-2

Wricke, G. (1962). Evaluation Method for Recording Ecological Differences in Field Trials. *Z Pflanzenzücht*, *47*, 92–96.

Wu, J., Lawit, S. J., Weers, B., Sun, J., Mongar, N., van Hemert, J., Melo, R., Meng, X., Rupe, M., Clapp, J., Collet, K. H., Trecker, L., Roesler, K., Peddicord, L., Thomas, J., Hunt, J., Zhou, W., Hou, Z., Wimmer, M., … Habben, J. E. (2019). Overexpression of zmm28 increases maize grain yield in the field. *Proceedings of*

the National Academy of Sciences of the United States of America, *116*(47), 23850–23858. https://doi.org/10.1073/pnas.1902593116

Xavier, A., Jarquin, D., Howard, R., Ramasubramanian, V., Specht, J. E., Graef, G. L., Beavis, W. D., Diers, B. W., Song, Q., Cregan, P. B., Nelson, R., Mian, R., Grover Shannon, J., McHale, L., Wang, D., Schapaugh, W., Lorenz, A. J., Xu, S., Muir, W. M., & Rainey, K. M. (2018). Genome-wide analysis of grain yield stability and environmental interactions in a multiparental soybean population. *G3: Genes, Genomes, Genetics*, *8*(2), 519–529. https://doi.org/10.1534/g3.117.300300

Zhang, H., Yin, L., Wang, M., Yuan, X., & Liu, X. (2019). Factors affecting the accuracy of genomic selection for agricultural economic traits in maize, cattle, and pig populations. *Frontiers in Genetics*, *10*(MAR), 1–10. https://doi.org/10.3389/fgene.2019.00189

Zhang, X., Pérez-Rodríguez, P., Semagn, K., Beyene, Y., Babu, R., López-Cruz, M. A., San Vicente, F., Olsen, M., Buckler, E., Jannink, J. L., Prasanna, B. M., & Crossa, J. (2015). Genomic prediction in biparental tropical maize populations in water-stressed and well-watered environments using low-density and GBS SNPs. *Heredity*, *114*(3), 291–299. https://doi.org/10.1038/hdy.2014.99

Zobel, R. W., Wright, M. J., & Gauch Jr., H. G. (1988). Statistical Analysis of a Yield Trial. *Agronomy Journal*, *80*(3), 388–393.

CHAPTER II

GENOMIC SELECTION FOR YIELD STABILITY

Abstract

Yield stability is of primary importance to breeding programs, but selection accuracy can be hindered in early stages of yield trials due to the years of yield data needed for phenotypic observation of the trait. The accuracy of genomic selection was assessed for three types of stability measures (AMMI, conventional dynamic, non-parametric), in three different prediction scenarios (across-population prediction and cross-validation within each population), using two populations separated by three years in a public same breeding program. Predictions were conducted using whole genome sequence data and yield stability values calculated from yield data across three years and 7-11 environments per line. Our results show that while overall predictive abilities (r_{MP}) were low (0.06 – 0.37), across-population prediction accuracies were similar to cross-validation accuracies. The accuracy and applicability of genomic selection for yield stability was enhanced by including a 'rank coincidence' measure of accuracy. Rank coincidence gave the expected accuracy for selection of the top half of a population, where selection accuracies ranged from 0.45 to 0.65. Furthermore, rank-based non-parametric stability measures have consistently higher prediction accuracies than AMMI (additive main effects and multiplicative interaction) and conventional dynamic stability. Prediction of yield stability across separate breeding populations is feasible in an applied breeding scenario with the correct stability measure. These methods have shown potential to provide increased selection accuracy for stability at early stages in a breeding pipeline when insufficient phenotypic information are available for yield stability estimation.

Introduction

Yield is often the paramount trait for improvement in plant breeding programs during selection and advancement of experimental lines. The importance of yield is so great that through the first handful of selection cycles, traits that improve or protect yield are often the only trait being selected. In addition to superior yield, selected lines must exhibit reliable yield across multiple environments within a target geographic region; referred to as yield stability. Typically, the breeding pipeline starts with a large number of lines tested in few environments over the first couple of years. In later years, population size is much smaller as superior lines are advanced and inferior lines selected out. At this point, the few remaining advanced lines are grown across numerous environments allowing for assessment of stability. The logistical limitations of phenotyping for stability induce high financial cost and labor to breeding programs and delay selection of stability to the later stages, likely affecting final selection accuracy for stability. This bottleneck of phenotype data for stability assessment makes yield stability a prime target for genomic selection at early stages of the breeding pipeline.

Stability is a measure of phenotypic consistency across a range of environments (Becker & Léon, 1988). Numerous traditional ANOVA-based measures have been developed to calculate stability in two primary categories: static and dynamic (Becker & Léon, 1988; Lin et al., 1986). Static stability measures the degree to which a line's performance remains unchanged across environments. Dynamic stability is a measure of yield consistency relative to the overall mean yield of an environment. This unequal yield response across different environments, also known as GxE (genotype-by-environment interaction), is quantified when assessing dynamic stability. While traditional static and

dynamic stability are univariate ANOVA-based measures, another group of stability measures has been introduced more recently. AMMI (Additive main effects and multiplicative interaction) is a model allowing for simultaneous estimation of additive main effects using ANOVA, and complex interaction using PCA (Zobel et al., 1988). This approach is often considered most appropriate for partitioning complex GxE interactions and estimating stability of yield in agronomic field trials.

With multiple methods for estimating yield stability, the genetic control of yield stability is not straightforward and can lead to challenges when utilizing genomic tools for selection of stability in a breeding program. It has been shown that stability traits have complex genetic architecture due to lack of stable, large-effect QTLs (Lozada & Carter, 2020; Y. Wang et al., 2015). The contrast in stability measure categories adds to the complexity of understanding the genetic control of stability traits. A recent study demonstrated this complexity by finding different QTL when using GWAS to map regions of the genome contributing to AMMI stability, conventional stability, and explicit modelling of GxE (Happ et al., 2021). Very few QTL were shared across stability categories. This inconsistency in genetic control likely causes the difference in observed rankings generated from varying stability measures. For trait improvement, the absence of stable QTL across models and stability measures makes marker-assisted selection (MAS) ineffective for improving stability in a breeding program. Genomic selection can be used to overcome the genetic complexity of stability traits and accelerate the rate of genetic gain in traits where MAS is incapable.

Over the past couple of decades, advancement in genotyping and sequencing technology has allowed genomic selection to become a prominent tool in plant and

animal breeding for predicting complex traits (Bernardo, 2008; Meuwissen, 2007; H. Wang et al., 2022). In genomic selection, genome-wide SNP (single nucleotide polymorphism) markers and phenotype data from a population are used to train a prediction model (Meuwissen et al., 2001). The phenotype of an individual with available SNP data can then be predicted before it is observed in the field. This method has been shown to increase the rate of genetic gain for complex traits in breeding programs (Bernardo & Yu, 2007; Heffner et al., 2010). One of the most common genomic prediction models used is the standard additive GBLUP model. This model has been shown to yield sufficient accuracy when compared to Bayesian, semiparametric (RKHS), and more complex additive models, for predicting traits with many small-effect QTL (Huang et al., 2016; Jarquín et al., 2014; Meher et al., 2022). Overall, the power of genomic selection to predict complex traits and accelerate selection makes method an intuitive choice to overcome the limitations of selecting phenotypic stability in plant breeding.

Genomic prediction for stability traits is relatively unexplored. One study assessed prediction for stability of multiple traits such as yield, test weight, plant height, and heading date, in wheat (Huang et al., 2016). They obtained low to moderate cross-validation prediction accuracies for two stability measures: AMMI stability index and the Eberhart and Russel regression. In comparing four different prediction models they found ridge regression BLUP (rrBLUP, equivalent to GBLUP) had highest prediction accuracy. In calculating AMMI stability, they found when using only the first few PCs, prediction accuracy increased. They hypothesized that the primary PCs explain more of the GxE pattern under genetic control (Sneller & Dombek, 1997), and later PCs being made up

primarily of GxE noise (Gauch Jr., 1988). Since this study was published, the results of AMMI stability prediction have not been repeated. An assessment of the predictability of a wide range of stability measures could help determine the most effective method for predicting stability and provide further evidence of the genetic control of stability traits.

In addition to exploring different stability measurements, genomic selection accuracy can be influenced by multiple elements. Training population size, composition, and relation to selection population, are important factors for prediction accuracy (Edwards et al., 2019; Werner et al., 2020). Other factors impacting selection accuracy are marker density, statistical method, heritability, and genetic architecture of the trait (H. Zhang et al., 2019). While heritability and genetic architecture are intrinsic to the trait of interest; training population design, marker density, and statistical methods are elements that can be optimized through experimental design to increase prediction accuracy. Furthermore, an understanding of the SNP marker density needed to obtain adequate prediction accuracy for stability, and the type of SNP marker that gives the highest prediction accuracy, are crucial pieces for the application of such methods in a real-world breeding program.

In this study, genomic prediction for yield stability is examined, specifically targeting how controllable elements of genomic prediction can be optimized for application in a real-world scenario. Two populations of experimental soybean lines from the University of Nebraska soybean breeding program were used. These populations are intended to represent a real-world application of genomic prediction for yield stability, with the training population developed three years prior to the validation population in the breeding pipeline. Whole genome sequence data and yield data from 8-11

environments across 3 years were collected for each line in both populations. Genomic predictions were conducted in an applied scenario using the earlier-derived population as the training population (TP) for prediction of the later population (validation population, or VP), and assessment of predicted value accuracy in comparison to observed stability for the VP. The accuracies from this scenario were also compared to TP cross-validation and VP cross-validation. Also assessed were the effects of SNP marker density, SNP marker type (high-effect vs. random), stability measure (conventional vs. AMMI), and prediction scenario, on prediction accuracy for yield stability. The primary goal was to assess the practicality of genomic prediction for yield stability as it could be applied in a breeding program. The ability to select for stability traits in earlier stages of the breeding pipeline would help enrich lines for greater yield stability in later stages of selection in the breeding program, ultimately improving soybean cultivars.

Materials and Methods

Plant Material

The populations in this study were acquired from the University of Nebraska Soybean Breeding program. Both the training population (TP) and validation population (VP) were made up of F_4-derived lines from biparental crosses of elite-yielding lines and three generations of selfing. TP lines underwent two to three years of yield trials before being selected to represent a variety of lines with low to high yield stability. The training population, developed three years prior to the VP, consisted of 197 experimental F_4-derived lines from maturity group I, II, III, and IV. VP lines underwent one to two years of yield trials prior to selection for this population, and these lines were randomly selected from a larger population of 1,500 experimental lines with a range of low to high

yield and yield stability. The validation population consisted of 200 F_4-derived lines belonging to maturity groups II and III. Before analysis, TP lines would be filtered to 190 lines to avoid identical lines in both populations, and VP lines would be filtered to 165 total lines as 35 lines had sequence coverage less than 0.3X as explained in the 'Genomic Data' section.

Field Tests

Field trials for this study were conducted over the course of six years. The training population, as described by Happ et al., was tested from 2017 to 2019 in 13 total environments (Happ et al., 2021). For a short description, yield tests were primarily located in the eastern half of Nebraska. Maturity groups I and II were grown at combinations of Cotesfield, Phillips, Mead, Lincoln, and Wymore across three years. Maturity group III and IV lines were grown at combinations of Phillips, Lincoln, and Wymore across three years. In the end, each line was grown in seven or eight environments with three replicates at each environment. An augmented incomplete randomized block design was used for training population yield trials. Seed was treated with CruiserMaxx before planting. For both populations, grain weight and moisture were taken at harvest and adjusted to 13% moisture for final yield calculation.

The validation population yield trials took place over three years from 2020 to 2022, across 18 total environments. VP environments were primarily located in the eastern half of Nebraska, along with locations in Iowa and Missouri. Maturity group II was grown at combinations of Cotesfield, Lincoln, Mead, Phillips, and Arcadia, IA across three years. Group III lines were grown at combinations of Cook, Lincoln, Mead, Phillips, Albany and Novelty, MO, and Winterset, IA across three years. All lines were

grown in 11 total environments across 3 years. An augmented incomplete randomized

block design was used for validation population field tests. Blocks contained 40 lines for

MGII and 28 lines for MGIII including maturity checks. Two replicates of each line were

grown per site. Plots were planted in four-row plots of 6 meters in length with 0.76-meter

row spacing. Seed was treated with CruiserMaxx before planting.

Yield Data Analysis

Variance components for yield trials were tested using the following model

$$(1) \quad Y = g_i + e_j + ge_{ij} + b_k(r_l) + \varepsilon$$

where Y is the vector of raw yield values, g_i is the effect of genotype i, e_j is the

environmental effect, ge_{ij} is the interaction between genotype and environment, $b_k(r_l)$ is

replicate nested within block in each environment, and ε is the residual error. Each term

was fit as random. The model was run in ASReml-R and resulting BLUPs were used as

observed values for yield prediction accuracy calculation. Broad-sense heritability was

calculated from ANOVA variance components as $H^2 = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_{GE/er}^2 + \sigma_{\varepsilon/r}^2}$ where σ_G^2 is the

genotypic variance, σ_{GE}^2 is genotype-by-environment interaction variance, σ_ε^2 is the

residual variance, e is the number of environments, and r is the number of replications

per environment.

Prior to yield prediction, BLUEs were calculated in ASReml-R for each line

within each environment to correct for experimental design:

$$(2) \quad Y = g_i + b_k(r_l) + \varepsilon$$

where Y is the vector of raw yield values, g_i is the fixed effect of genotype i, $b_k(r_l)$ is

replicate nested within block in each environment as a random effect, and ε is the random

residual error. The resulting BLUEs were used as training values for genomic prediction

of yield BLUPs across all environments using the GBLUP model in linear form:

$$(3) \quad Y = g_i + e_j + ge_{ij} + \varepsilon$$

Predicted GEBVs were then compared to observed BLUPs from the first model to assess

prediction accuracy for yield BLUPs.

Stability Measures

Yield data from field tests were used to calculate stability parameters. Traditional

stability parameters were calculated using 'stabilitysoft', an R package allowing the

calculation of over a dozen static and dynamic stability parameters (Pour-Aboughadareh

et al., 2019). For stabilitysoft calculation, the reps for each genotype in each environment

were averaged. For yield the dynamic stability concept was used, which has two

approaches: partitioning GxE interactions and regression approach. Wricke's ecovalence

is a measure for partitioning GxE interactions. The formula for Wricke's ecovalence is as

follows:

$$(4) \quad W_i = \sum_j (X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}_{..})^2$$

where X_{ij} = Mean of genotype i in environment j, $\bar{X}_{i.}$ = Mean yield of genotype i across

environments, $\bar{X}_{j.}$ = Mean of environment j across genotypes, $\bar{X}_{...}$ = Overall mean

(Wricke, 1962). As a dynamic stability parameter, Wricke's calculates the deviation of a

genotype's mean from the population's overall environmental mean and sums this

deviation across all environments. Deviation mean squares, also known as deviation from

regression, is a dynamic stability measure using the regression approach. Its formula is

shown below:

$$(5) \quad s^2_{di} = \frac{1}{E-2} [\sum_j (X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}_{..})^2 - (b_i - 1)^2 \sum_j (\bar{X}_{.j} - \bar{X}_{..})^2]$$

where E is the number of environments, and b_i is Finlay and Wilkinson's coefficient of regression as follows (Finlay & Wilkinson, 1963):

$$(6) \quad b_i = 1 + \frac{\sum_j (X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}_{..})(\bar{X}_{.j} - \bar{X}_{..})}{\sum_j (\bar{X}_{.j} - \bar{X}_{..})^2}$$

Lastly, for conventional dynamic stability, Coefficient of Variation was used (Francis & Kannenberg, 1978). This statistic is often implemented in conjunction with Environmental Variance and mean yield.

Non-parametric stability measures are based on rank differences across environments. These measures are an interesting alternative to parametric measures that quantify GxE, as no assumption is needed of the distribution of phenotypic values (Nassar & Hühn, 1987). Four of the primary non-parametric statistics are known as Nassar and Huehn's statistics (Nassar & Hühn, 1987; Pour-Aboughadareh et al., 2022). The first statistic ($S^{(1)}$) the mean of the absolute rank differences of a genotype across all environments. The second statistic ($S^{(2)}$) is the variance among the ranks over tested environments. The third statistic ($S^{(3)}$) is the sum of absolute deviations for each genotype relative to the mean of ranks. And finally, $S^{(6)}$ is the sum of squares of rank for each genotype relative to the mean of ranks. The lowest value of S-statistic is most desirable. Another non-parametric stability measure is Kang's rank-sum (Kang, 1988). This method ranks lines based on both mean yield and stability according to Shukla's stability variance, a dynamic univariate measure (Shukla, 1972). Essentially, each genotype is assigned a whole number starting at one, as the highest yielding line with lowest Shukla's stability variance. Kang's rank-sum assigns the same weight to both mean yield and stability. The lowest value of rank sum is most desirable.

In addition to conventional parametric and non-parametric stability measures, AMMI measures were also calculated. AMMI is unique in that it utilizes both ANOVA and principal component analysis (PCA) to partition residual variance and quantify genotype-by-environment interaction (Zobel et al., 1988). AMMI is thought to be more suitable for partitioning multiplicative genotype by environment interactions via PCA, thereby identifying interaction patterns in a way not possible with only main effects ANOVA analysis (Zobel et al., 1988). The AMMI model is defined as:

$$(7) \quad Y_{ij} = \mu + g_i + e_j + \sum_{n=1}^{N} \lambda_k \gamma_{ik} \delta_{jk} + \rho_{ij} + \varepsilon_{ij}$$

where Y_{ij} is the response variable of the ith genotype in the jth environment, μ is the overall mean, g_i is the fixed genotype effect, e_j is the fixed environmental effect, N is the number of principal components in the AMMI model, λ_k is a value of the K axis in principal component analysis, γ_{ik} is the genotype factor of the vector associated with λ_k from the interaction matrix, δ_{jk} is the environmental factor of the vector associated with λ_k from the interaction matrix, ρ_{ij} is the residual GxE interaction, and ε_{ij} is the random, independent error term with mean and distribution $\varepsilon_{ij} \sim N(0, \sigma^2)$. AMMI stability parameters were calculated using the R package 'ammistability' with environment, genotype, and rep as model terms (Ajay et al., 2019).

The conventional univariate stability measures used in this study were Wricke's ecovalence, deviation from regression, and coefficient of variation. The non-parametric measures used were Nassar and Huehn's S^1, S^2, S^3, S^6, and Kang's rank-sum. These eight stability measures were calculated using the stabilitysoft R package (Pour-Aboughadareh et al., 2019). For AMMI, the measures used were Sum across environments of GEI

modelled by AMMI (AMGE), AMMI Based Stability Parameter (ASTAB), Averages of the squared eigenvector values (EV), Stability measure based on fitted AMMI Model (FA), and Absolute value of the relative contribution of IPCs to the interaction (ZA). The calculation of AMMI stability measures for each population were conducted using the ammistability R package and principal components from the AMMI model. For each stability measure in this study, smaller values are considered more stable.

Genomic Data

Whole genome skim sequence data of the training population was obtained from Happ et al., (2021). NovaSeq data were acquired for TP lines after this study, giving an average of 10X coverage (Illumina, Hayward, CA). The methods to perform whole genome skim sequencing on the validation population were conducted as described by Happ et al., (2019). Briefly, DNA was extracted using a CTAB method (Keim, 1988). DNA samples were normalized and pooled by plate before library preparation was conducted using the IGENOMIX Riptide High-Throughput Rapid Library Prep kit and protocol (IGENOMIX, Doral, FL). Whole genome sequence data was acquired using the Illumina NextSeq 500 (Illumina, Hayward, CA). An average of < 1x coverage was obtained for each of the 200 lines (Supplementary Table 1). The reference panel for SNP genotyping and imputation was generated as described by Happ et al., (2019). SNP genotyping was conducted on the previously called SNPs using GATK4 version 4.1 HaplotypeCaller (Poplin et al., 2018). Imputation was conducted using Beagle v5.2 and the aforementioned reference panel (Browning et al., 2018). After imputation, 10,286,817 genome-wide SNPs were called for each individual in both populations. Genomic data for the training and validation populations were filtered for MAF (minor allele frequency)

< 0.05 and linkage disequilibrium $r^2 > 0.4$. MAF filtering discarded 8,849,209 SNPs in the training population to leave 1,437,578 SNPs. Of those SNPs, 865,499 were discarded due to LD filtering, leaving 572,079 TP SNP positions. For the validation population, MAF filtered out 8,561,523 SNPs to leave 1,725,264 positions. Of those, LD filtering removed another 916,959 positions to result in 808,305 SNPs. Both MAF and LD filtering steps were conducted using Plink1.9 (Purcell et al., 2007). After separate filtering of both populations, TP and VP data sets were reduced to common SNP positions resulting in a total of 373,348 SNPs across all 20 chromosomes. 35 of the 200 VP lines had < 0.3X sequence coverage and were filtered out. The threshold for imputation accuracy of 0.3X had been identified previously (Happ et al., 2019). At 0.3X, imputation accuracy of 93% is comparable to 94% accuracy when sequence coverage is 1X. Below 0.3X, imputation accuracy decreases quickly to roughly 90% at 0.1X coverage. With the goal of as little imputation error as possible, 0.3X was the coverage threshold for this study.

SNP Selection for Prediction Model

SNPs were selected for inclusion in the genomic selection models by comparing two different methods. The first method is selection of SNPs with highest effect on stability from a simple GWAS analysis. GWAS was conducted using the QK MLM method with its equation as follows (Price et al., 2006; Yu et al., 2006):

$$(8) \quad y = X\beta + S\alpha + Qv + Zu + e$$

where y is the vector of phenotypic observations; X is the incidence matrix for fixed effects; b is the vector of fixed effects other than the SNP being tested and population structure; S is the incidence matrix of SNP effects; α is the vector of SNP effects; Q is a

matrix relating y to v; v is a vector of population effects; Z is the incidence matrix for random effects; u is the vector of background SNP effects; and e is residual error. GWAS was conducted for both populations using the R package "R-MVP", using PCA to correct for population structure (Yin et al., 2021). Principal components for population structure correction and reduction of false positives were obtained using Plink1.9 (Supplementary figures 4 & 5) (Purcell et al., 2007). For yield prediction, BLUPs were calculated from a model similar to (2), but the block/rep nested term was excluded and replaced with block and rep as random terms. GWAS was conducted separately for both populations to simulate a real-world scenario where phenotype data is not available for a population, and high-effect SNPs from another population must be used. One thousand five hundred high-effect SNPs from TP analysis were used for across-population prediction of VP and VP cross-validation. Similarly, 1,500 high-effect VP SNPs were used for TP cross-validation. For yield, marker densities of 1,500, 3,000, and 5,000 were tested. Prediction accuracy was highest for 1,500 markers. The precise methods used when selecting high-effect SNPs are outlined below.

Due to the considerable difference between values and rankings of AMMI and conventional dynamic stability, separate GWAS analyses were conducted for each category using one stability measure with moderate to high correlation with the other measures in that category, ASTAB AMMI and Wricke's Ecovalence. To avoid redundant markers in close proximity, the highest-effect markers within each 10,000 base pair window were kept. For each SNP density, an equal number of the highest-effect markers were selected from each chromosome. A separate GWAS was conducted for both populations and stability types. High-effect SNPs from TP GWAS were used for TP – VP

across-population prediction and VP cross validation. High-effect SNPs from VP GWAS were used for TP cross validation. SNP densities of 500, 600, 700, 800, 900, 1,000, 1,500, 2,000, 2,500, 3,000, 3,500, 4,000, 4,500, 5,000, 7,500, 10,000, 12,500, and 15,000 were tested for the ASTAB.AMMI stability measure.

Random SNP selection was done by randomly choosing an equal number of SNP positions from each chromosome across the genome. In analyzing the effect of SNP density on prediction accuracy, as SNP sets increased in number, new random SNPs from across all chromosomes were added to the smaller SNP set. This was done in R. All random SNP prediction scenarios used the same random SNP sets.

Genomic Prediction

Genomic prediction was conducted using the GBLUP model. GBLUP requires a genomic relationship matrix to quantify relationship between all individuals of the populations involved. The VanRaden relationship matrix was used in this study (VanRaden, 2008). The GBLUP mixed linear model equation is as follows (Henderson, 1984; Meuwissen et al., 2001):

$$(9) \quad Y = Xb + Zu + e$$

where Y is the response vector, X is the design matrix for fixed effects, b is the vector of fixed effects, Z is the design matrix for random effects, u is the vector of predicted random effects, and e is residual error. Genomic predictions were conducted using ASReml-R (Butler et al., 2018). The three GBLUP models used are shown in linear form:

$$(10) \quad y = \mu + g_i + \varepsilon_i$$

$$(11) \quad y = \mu + g_i + E_j + \varepsilon_i$$

$$(12) \quad y = \mu + g_i + E_j + gE_{ij} + \varepsilon_i$$

where y is the vector of response, μ is the intercept or mean to be added back to BLUP

values, g_i is the random additive VanRaden genomic relationship between individuals in

both populations with $g_i \sim N(0,\sigma^2{}_g)$, E_j is the random environmental effect as a binary

incidence matrix identifying lines grown in each environment with $E_i \sim N(0,\sigma^2{}_E)$, gE_{ij} is

the random interaction between genotype and environment with $gE_{ij} \sim N(0,\bigoplus_{j=1}^{J}\sigma^2{}_{gE})$,

and ε_{ij} is the residual with $\varepsilon_{ij} \sim N(0,\sigma^2{}_\varepsilon)$. The terms $\sigma^2{}_g$, $\sigma^2{}_E$, and $\sigma^2{}_{gE}$, and $\sigma^2{}_\varepsilon$ are the

variance components for genotype, environment, GxE, and residuals. Equation (10) is the

standard additive GBLUP model utilizing only genomic relationship as a random model

term. Model (11) adds a random environmental term as an incidence matrix to specify

which environments each line was grown in, and (12) adds the random interaction

between genotype and environment. Model comparison of prediction accuracy was

conducted to observe effectiveness for predicting yield stability across populations.

For yield prediction, training values were BLUEs from the previously stated

model, where each genotype has one BLUE per environment. In linear form, the GBLUP

model used is as follows:

$$(13) \quad y = g_i + e_j + ge_{ij} + \varepsilon$$

with model terms accounting for genotype (g_i), environment (e_j), genotype by

environment interaction (ge_{ij}), and residuals (ε). As a result, one BLUP value was

predicted for each line. This value is a prediction of overall yield performance across all

environments. Genomic prediction was done in three scenarios: across-population

prediction, TP cross-validation, and VP cross-validation. Across-population prediction

was done by training the model with the training population BLUEs to predict the

validation population. Cross-validation was conducted in five repetitions of random five-

fold cross-validation. Within each rep, the population was randomly split into five groups. For each fold, four groups would be the training population and one group the validation population. This was repeated until all five groups had been predicted. The correlation between predicted and calculated BLUPs was calculated for each fold and the average accuracy across all five folds in each repetition was reported. Standard deviation between fold accuracy values was also reported. Across-population prediction accuracy for yield was the correlation between VP BLUP values from (1), and predicted VP GBLUP values.

The same three prediction scenarios were assessed for prediction of yield stability. The primary scenario was the training population used to train the model and predict the validation population, which is meant to simulate a real-world genomic prediction scenario. Prediction accuracy for this scenario was examined in how it varied with changing SNP density, SNP selection method (GWAS vs random), and stability measure. Two other scenarios assessed were TP 5-fold cross-validation, and VP 5-fold cross-validation. Each 5-fold cross validation was repeated 5 times, with prediction accuracies averaged across the 5 folds within each repetition.

Three types of prediction accuracy were assessed in this study. Each were calculated in R version 4.2 using correlation functions or by basic calculation (R Core Team, 2021). Typical predictive ability (r_{MP}) was measured as the correlation between predicted stability values and observed stability values. Rankings tend to be more useful for plant breeding purposes, so two measures of predicted ranking accuracy were also assessed. Spearman's rank correlation coefficient is the correlation between predicted rankings and observed rankings (Best & Roberts, 1975). Spearman's rank correlation

coefficient was calculated in R using the set of predicted stability ranks and observed stability ranks. Lastly, an unofficial accuracy estimate that hereafter will be referred to as "rank coincidence" measures the proportion of lines in the same half (top or bottom) of predicted stability rankings and observed stability rankings. This was calculated by hand in R using observed and predicted stability rankings.

Results

Phenotype and Genotype Data

The 213 training population lines were tested in 13 total environments across three years in Eastern Nebraska (Figure S2.1 & S2.2). The mean yield across all environments and years was 5072.4 kg/ha with the highest average yield at Phillips 2018 with 6008.3 kg/ha. The highest yield observation for the lines was 8396.8 kg/ha, and the lowest observation was 1121.1 kg/ha. The 200 validation population lines were tested in 18 total environments across three years in Eastern Nebraska, Western Iowa, and Northern Missouri (Figure S2.1 & S2.3). The mean yield across all environments and years was 4940.5 kg/ha, while the highest yielding environment was at Cotesfield, NE 2021 with an average yield of 6868.9 kg/ha. The highest yield observation was 8578.2 kg/ha and the lowest observation was 1293.9 kg/ha. Heritability for yield was vastly different between the two populations, with 0.56 for TP and 0.22 for VP (Table 2.1). This is likely due to the larger number of environments and environmental variance in VP yield trials as compared to TP. Stability values across all environments were calculated for five AMMI stability measures and three conventional dynamic stability measures for both populations. Within each population, stability measure distributions were typically skewed (Figures S2.4-7).

Table 2.1. Variance component estimates of yield trial data for both TP and VP. For heritability calculation, TP included 12 total environments and 3 replicates per environment, and VP included 19 total environments and 2 replicates per environment.

Population	Block	block:rep(env)	Env	Genotype	GxE	Residuals	Heritability
TP	0.00	38.35	61.46	27.01	16.54	61.68	0.56
VP	0.00	8.17	248.32	7.10	7.51	49.88	0.22

Whole genome sequence coverage for the 190 TP lines was generally high with an average of 9.8X, ranging from 1X to 33X (Table S2.1). VP coverage for the 165 lines used in this study averaged 0.86X and ranged between 0.31X and 7.7X. After filtering MAF and LD, 373,348 genome-wide SNP markers were left for genomic prediction. The positions of these markers across chromosomes can be found in Figure S2.13. Q-Q plots and Manhattan plots for each GWAS (TP and VP: AMMI stability, conventional stability, yield BLUPs) are shown in Figures S2.14-19.

Yield Prediction

Table 2.2. Cross-validation and across-population accuracies for yield BLUP prediction. Mean predictive ability (r_{MP}) and standard deviation between r values across 5 folds within each repetition are reported. TP cross-validation used 1,500 high-effect VP SNPs . Across-population prediction and VP cross-validation used 1,500 high-effect TP SNPs.

| TP cross-validation			VP cross-validation		
Rep	mean r	sd	Rep	mean r	SD
1	0.26	0.11	1	0.32	0.26
2	0.34	0.12	2	0.35	0.15
3	0.28	0.19	3	0.40	0.06
4	0.36	0.10	4	0.36	0.18
5	0.29	0.13	5	0.45	0.05
Average	**0.31**	0.13	Average	**0.38**	0.14
across-population r: **-0.11**					

First, yield was predicted in three scenarios for comparison with other studies (Table 2.2). These three scenarios were across-population prediction, TP cross-validation, and VP cross-validation. For five repetitions of five-fold cross-validation, predictive

ability (r_{MP}) ranged from 0.26 to 0.36 for TP cross-validation, and 0.32 to 0.45 for VP cross-validation. The overall average r_{MP} for each population was 0.31 and 0.38 for TP and VP cross-validation, respectively. Across-population prediction of yield BLUPs resulted in a much lower r_{MP} of -0.11. The relationship between populations based on shared parents was also visualized (Figure S2.20). Within each population over 75% of individuals share a full sibling and roughly 20% share a half sibling, while only 3% have no siblings. Between populations, there are no parents shared by lines. Our results highlight the disparity between cross-validation within a population of closer-related individuals tested in the same environments, and prediction across two separate breeding populations tested in different environments for yield BLUPs.

***Yield Stability Prediction Across Populations***

SNP Density and Type
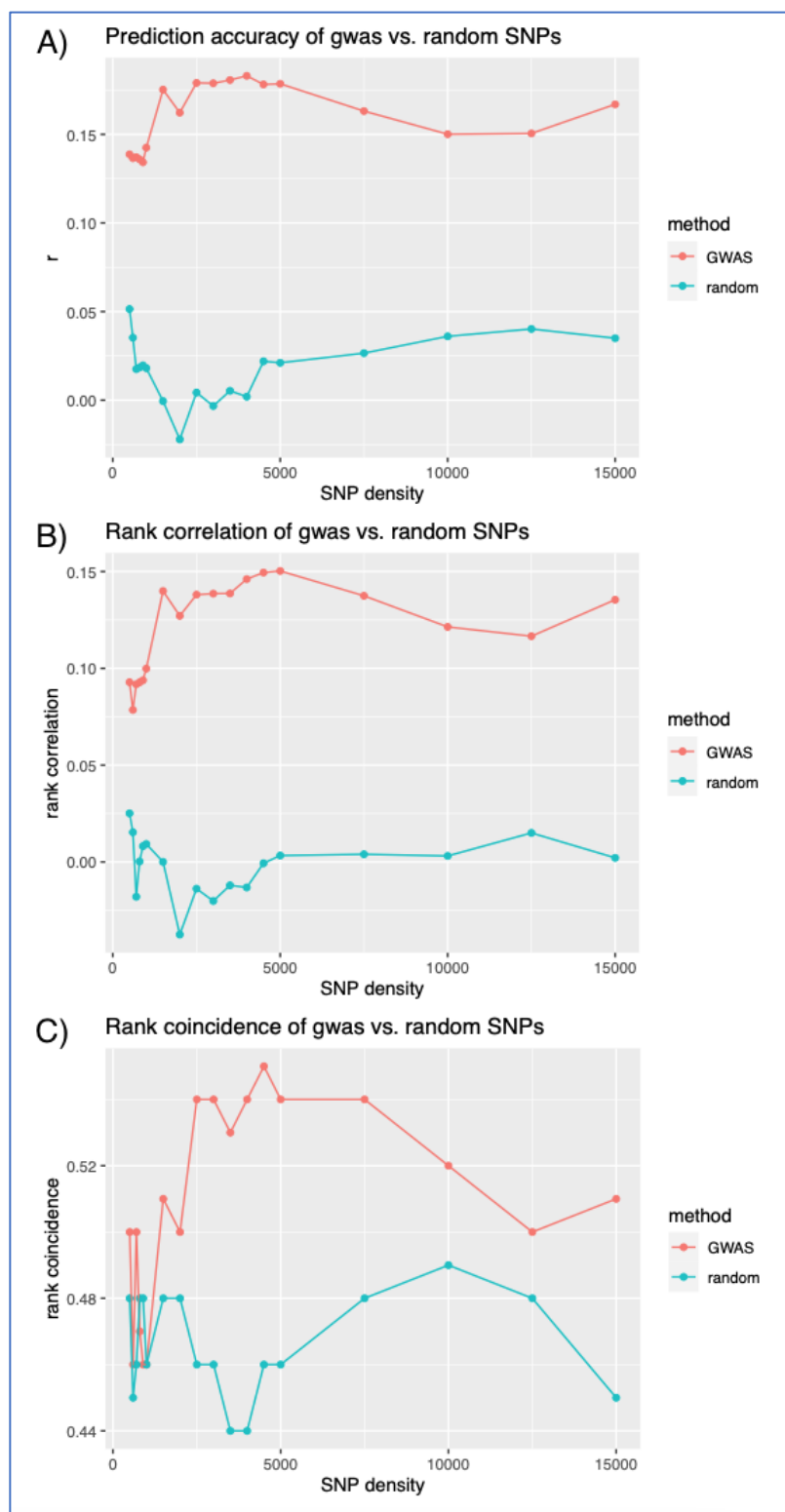
Figure 2.1. SNP density vs. three different types of prediction accuracy: A) predictive ability (r_{MP}), B) rank correlation, and C) rank coincidence, for across population prediction of ASTAB AMMI stability. The red line indicates GWAS SNPs, and the blue-green line represents random SNPs. SNP densities ranged from 500 to 15,000.

The first prediction scenario assessed was the training population predicting the validation population for ASTAB stability. The effect of SNP density and SNP selection method on three estimates of prediction accuracy was observed (Figure 2.1). In all three prediction accuracy scenarios, high effect SNPs consistently performed better than the randomly selected SNPs. Around a marker density of 1,500, a plateau in predictive ability begins with an r_{MP} of 0.17 and rank correlation of 0.14. The change in r_{MP} from 1,500 to 15,000 markers was negligible. Rank coincidence values have more fluctuation than the other two prediction accuracy estimates. This is likely due to the amount of fluctuation that can come with predicted value rankings at low SNP densities.
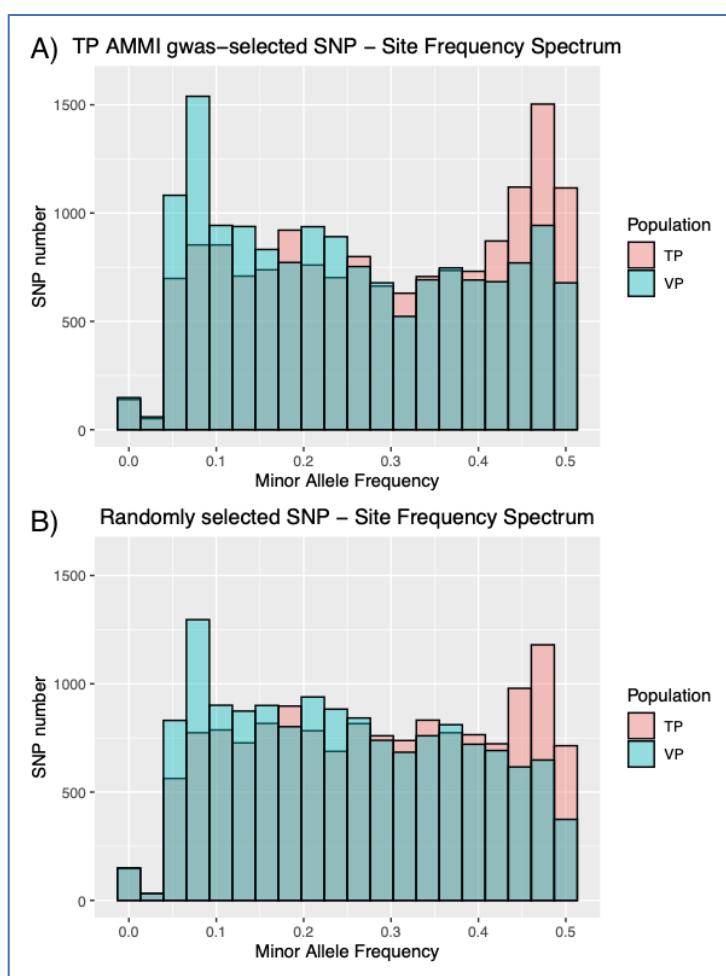
Figure 2.2. Minor allele frequencies of SNP sets for both populations. A) 15,300 TP GWAS-selected SNPs for AMMI stability and B) 15,000 randomly selected SNPs.

Distributions for the unfiltered SNP set and MAF/LD filtered set are reported in supplementary materials.

In order to further elucidate the difference in prediction accuracy between the highest-effect SNPs from GWAS and random SNPs, the site frequency spectrum was plotted for each set of 15,000 SNPs used in the first prediction scenario (Figure 2.2). The MAF distributions for both SNP sets were very similar. Though after observing the SFS distribution for the set of 373,348 SNPs after MAF and LD filtering, from which the randomly selected SNPs were sampled, this becomes less surprising (Figure S2.21). Thus, the difference in prediction accuracy between the two sets isn't due to differing MAF of markers. The difference in prediction accuracy is likely explained by higher-effect SNPs providing better predictive ability.

Stability Measure

Table 2.3. Prediction accuracy comparison between 13 stability measures using the same set of 1,500 GWAS-selected SNP markers. Across-population prediction was conducted using TP to predict VP. Prediction accuracies include r_{MP}, rank correlation, and rank coincidence. Genomic heritability for each stability measure was calculated from model variance components.

Stability measure	r	rank corr.	rank coincidence	Heritability
AMGE.AMMI	0.096	0.053	0.53	0.066
ASTAB.AMMI	0.175	0.140	0.51	0.484
EV.AMMI	0.176	0.154	0.53	0.509
FA.AMMI	0.157	0.100	0.51	0.412
ZA.AMMI	0.058	-0.019	0.45	0.213
Coefficient of variation	0.073	0.126	0.55	0.462
Deviation from regression	0.022	0.058	0.56	0.521
Wricke's ecovalence	0.071	0.106	0.57	0.624
S1	0.367	0.380	0.65	0.303
S2	0.363	0.386	0.65	0.405
S3	0.292	0.289	0.62	0.552
S6	0.224	0.223	0.62	0.498
Kang's rank sum	0.194	0.212	0.61	0.386

To observe if certain stability measures are more predictable than others, prediction accuracy was assessed for five AMMI, three conventional univariate, and five non-parametric stability measures (Table 2.3). Each prediction used the same set of 1,500 GWAS selected SNP markers. Non-parametric stability measures displayed consistently higher r_{MP} (0.19 – 0.37) than AMMI and dynamic parametric measures (0.02 – 0.18). This was also consistent across rank correlation (NP: 0.21 – 0.39; other: -0.02 – 0.15) and rank coincidence (NP: 0.61 – 0.65; other: 0.45 – 0.57). Of the other stability measures, Wricke's ecovalence had the highest rank coincidence with 0.57. Generally, AMMI stability measures had higher r_{MP} and rank correlation, but lower rank coincidence than parametric dynamic measures. r_{MP} values were generally low, which was expected due to prediction of a complex trait across populations.

Model Comparison

Three different GBLUP models were assessed for the prediction of yield stability across populations for each stability measure (Table S2.4). The three models include the standard additive GBLUP with no other terms, GBLUP model with environment (E), and GBLUP with environment plus genotype-by-environment interaction (GxE). Generally, the models accounting for E and E + GxE did not improve prediction accuracy across populations compared to the standard additive GBLUP model. Often, the ASReml log likelihood for model fit would improve (closer to zero) with the more complex models, but this improvement translated to decreased prediction accuracies. Only two stability measures, Coefficient of Variation and Nassar and Huhn's S2 statistic, showed noticeable

increases in r_{MP}, rank correlation, and rank coincidence with E and E + GxE models. For

the eleven other measures, the standard additive GBLUP model performed best.

Yield Stability Cross-Validation

Table 2.4. Average prediction accuracies for 5 repetitions of 5-fold cross-validation within the training population. Thirteen stability measures were predicted with 1,500 GWAS-selected SNPs from VP GWAS. Predictive ability (r_{MP}), standard deviations (SD), rank correlation, and rank coincidence were averaged across five folds of CV.

Stability measure	r (SD)	rank corr	rank coincidince
AMGE.AMMI	-	-	-
ASTAB.AMMI	0.071 (0.09)	0.026	0.516
EV.AMMI	0.063 (0.10)	0.046	0.516
FA.AMMI	0.075 (0.09)	0.003	0.51
ZA.AMMI	0.089 (0.08)	-0.012	0.508
Coefficient Of variation	0.210 (0.11)	0.162	0.55
Deviation from regression	0.153 (0.13)	0.075	0.506
Wricke's ecovalence	0.133 (0.12)	0.098	0.534
S1	0.266 (0.10)	0.243	0.622
S2	0.327 (0.09)	0.316	0.628
S3	0.255 (0.11)	0.179	0.55
S6	0.211 (0.13)	0.121	0.566
Kang's rank sum	0.221 (0.11)	0.187	0.582

Table 2.5. Average prediction accuracies for 5 repetitions of 5-fold cross-validation within the validation population. Thirteen stability measures were predicted with 1,500 GWAS-selected SNPs from TP GWAS. Predictive ability (r_{MP}), standard deviations (SD), rank correlation, and rank coincidence were averaged across five folds of CV.

Stability measure	r (SD)	rank corr	rank coincidince
AMGE.AMMI	-	-	-
ASTAB.AMMI	0.326 (0.10)	0.152	0.54
EV.AMMI	0.230 (0.11)	0.163	0.526
FA.AMMI	0.206 (0.10)	0.133	0.532
ZA.AMMI	0.059 (0.14)	-0.046	0.468
Coefficient of variance	0.106 (0.15)	-0.054	0.504
Deviation from regression	0.109 (0.13)	-0.040	0.49
Wricke's ecovalence	0.004 (0.14)	-0.108	0.45
S1	0.419 (0.11)	0.409	0.674
S2	0.392 (0.11)	0.397	0.664
S3	0.278 (0.13)	0.244	0.582
S6	0.261 (0.16)	0.193	0.568
Kang's rank sum	0.170 (0.20)	0.072	0.542

Lastly, cross validation within the training population and within the validation population was asessed (Tables 2.4 and 2.5). Each cross validation used the set of 1,500 GWAS SNPs from the opposite population's GWAS. Predicted values for five AMMI, three conventional dynamic univariate stability, and 5 dynamic non-parametric measures were predicted. Prediction accuracies were averaged across five repetitions of five-fold cross-validation. For TP cross-validation, r_{MP} ranged from 0.07 to 0.33 across stability measures (Table 2.4). Rank correlation was between 0.03 and 0.32, while rank coincidence was measured from 0.51 to 0.63. r_{MP} for VP cross-validation were between 0 and 0.42; rank correlations ranged from -0.11 to 0.41; rank coincidence values were between 0.47 to 0.67 (Table 2.5). Accuracy values for both cross-validation scenarios were similar to across population prediction of yield stability. Population structure was

visualized using the first two principal components from Plink1.9 PCA (Figure 2.3, Figure S2.22). Overall, the similarity between across-population and cross-validation accuracies is explained by their overlap across the first two PCs. In addition, the proportion of lines with at least one full-sibling, half-sibling, or no siblings within each population and between them are shown in Figure S2.20. Within each population, the proportion of lines with at least one full-sibling, at least one half-sibling, and no siblings are very similar. This explains the similarity in TP and VP cross-validation accuracy for yield stability. However, the similarity in accuracy across populations is surprising given that no lines share parents between the populations, and stability values exhibit a low degree of repeatability across different sets of environments.
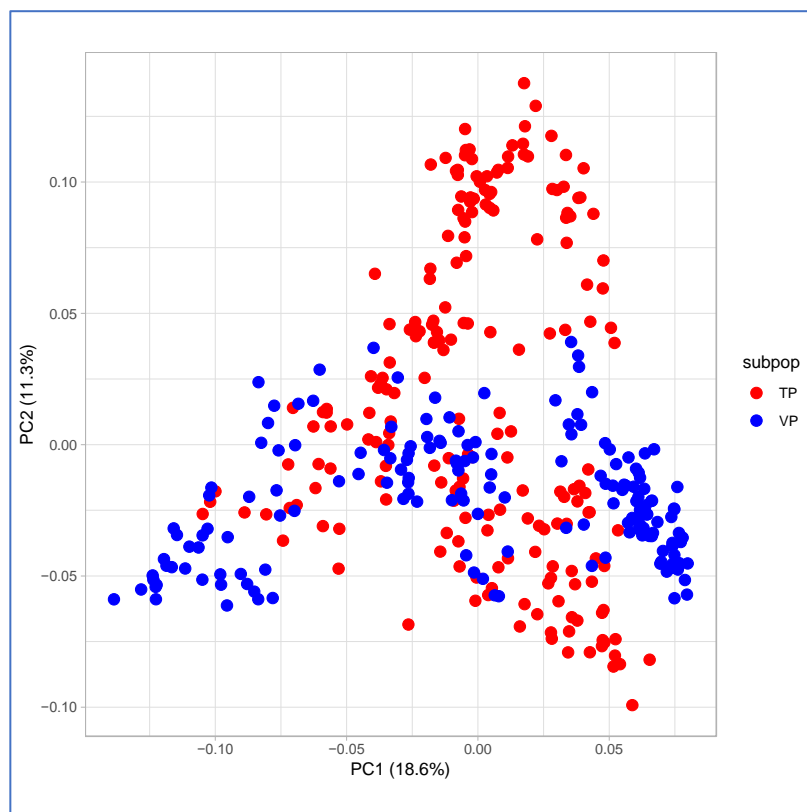
Figure 2.3. Visualization of population structure using the first two principal components from Plink for 190 TP individuals (red) and 165 VP individuals (blue).

Discussion

Selection and improvement of stability traits in breeding programs is often challenging due to the amount of phenotype data needed to accurately assess phenotypic stability across environments. While genomic selection is an intuitive tool to attempt for alleviating this limitation, few studies have focused on genomic prediction for stability traits. If genomic selection for yield stability is to be applied in a breeding program, there are certain elements of genomic selection methods that must be assessed and optimized. In this study, the efficacy of applied genomic prediction for yield stability in soybean was assessed. This was done by examining the effect of factors such as SNP density, SNP type, and stability measure using two soybean populations from a public breeding program in two different prediction scenarios.

There are a variety of different types of stability measures to choose from when conducting stability analysis. These measures are calculated in different ways, and their ability to be predicted using genomic selection is variable. Typical AMMI and dynamic univariate stability measures quantify and sum GxE interaction across environments (Becker & Léon, 1988; Zobel et al., 1988). Stability measures that quantify GxE were more difficult to predict than rank-based non-parametric stability measures (Tables 2.3-2.5). While non-parametric stability measures are still considered dynamic in nature, they are based on deviations in yield rank across environments rather than quantifying and summing GxE across environments (Nassar & Hühn, 1987; Pour-Aboughadareh et al., 2022). Across all three prediction scenarios, non-parametric measures consistently displayed the highest r_{MP}, rank correlation, and rank coincidence. The rank-based

function of non-parametric measures likely allows them to be more predictable within and across populations when compared to measures that partition GxE interaction.

Genomic prediction for stability traits would be most advantageous for application in the early yield trial phase of the breeding pipeline at a stage when data from testing environments are insufficient to select for observed phenotypic stability. In this situation, rank coincidence would be the most useful accuracy estimate, where selections can be made on individuals in the top half of predicted value ranks. AMMI and conventional univariate measures displayed relatively low rank coincidence values, often not improving on expected random selection accuracy of 0.5 (Tables 2.3-2.5). With a rank coincidence accuracy around 0.65 across populations for the S^1 non-parametric measure, when a breeder advances the top half of individuals in predicted stability ranks, roughly 65% of those individuals will be in the top half of true yield stability in a population. This method has potential to increase selection accuracy by 30% compared to random selection (assuming random selection accuracy is 50% when advancing half of a population). Furthermore, rank coincidence provides a powerful tool for breeders to predict and select for complex traits with increased efficiency, even when r_{MP} values are not high enough to be of use.

One of the primary goals of this study, and an aspect rarely assessed in genomic prediction studies, is the efficacy of predicting yield stability across populations as would be typical in an applied breeding scenario. Ultimately, genomic prediction of yield stability in an applied scenario does not yield sufficient accuracy for AMMI and conventional dynamic measures using the methodologies and approach presented (Figure 2.1, Table 2.3). Predictive abilities were low to moderate across stability measures, with

non-parametric measures having the highest predictive ability (r_{MP}), rank correlation, and rank coincidence. Although r_{MP} values were not of use, rank coincidence values (0.61 – 0.65) approached a level high enough for use in a real-world breeding program for non-parametric stability. This demonstrates the difficulty of predicting complex traits across populations, even when those populations are three years apart within the same breeding program and share ancestors. There is a disparity in direct relationship between populations where between populations, there are no lines that share even one parent (Figure S2.20). We hypothesized that cross-validation would yield higher prediction accuracy due to the full-sibling and half-sibling content within populations. However, cross validation within the training population and validation population generally displayed accuracies similar to across population prediction (Tables 2.2 and 2.3). Due to the relative success of predicting non-parametric stability measures, it may be possible to obtain adequate accuracies in an applied breeding scenario. This will likely require more rigorous testing of methods such as training population development and assessment of other types of prediction models.

Low predictive abilities were expected in this study when comparing to the few existing past studies. One 2016 study found cross-validation predictive abilities for yield stability in wheat to be 0.44 for ASI AMMI stability and 0.36 for Eberhart and Russell Regression (Huang et al., 2016). Another 2016 study in winter wheat had cross-validation accuracies of 0.25 for ASI, 0.28 for ASV AMMI, and 0.49 for Finlay-Wilkinson regression (Lozada & Carter, 2020). It is difficult to compare predictive abilities between studies of this nature, as all studies have slightly experimental design, statistical methods, and plant material. For across population prediction, it is no surprise that our accuracies

were generally much lower than these for AMMI and conventional stability measures. These studies did not assess prediction for non-parametric stability measures, which was the category of stability most easily predicted in this study.

A primary factor affecting prediction accuracy is SNP density and type of genomic data used (H. Zhang et al., 2019). Prediction accuracy generally plateaued at 1,500 markers (Figure 2.1). High-effect SNPs from GWAS also gave considerably higher r_{MP} than random SNPs. This is consistent with previous data that have shown the increased effectiveness of SNPs from GWAS in genomic prediction (Brøndum et al., 2015; Luo et al., 2021; van den Berg et al., 2016). The question was then raised of whether the difference in prediction accuracy is due to SNP effect, or due to the ascertainment bias of selecting positions with more common polymorphism in a population. We hypothesized that the increase in predictive ability from using higher-effect SNPs may be due to increased MAF as compared to random SNPs. This is because SNP arrays that explain large amounts of genetic variance target positions with high MAF (Matukumalli et al., 2009), and SNP sets with high MAF have been identified to increase prediction accuracy for some traits in soybean (Jarquín et al., 2014). However, when comparing site frequency spectra (SFS) of GWAS and random markers in our circumstance, there is little noticeable difference between them (Figure 2.2). The overall structure of SFS for GWAS and randomly selected SNPs in each population was consistent with the total set of 373,348 MAF and LD-filtered SNPs (Figure S2.21). Due to the similarity of MAF distribution in high effect GWAS and random positions, the difference in prediction accuracy may due to the increased effect of GWAS SNPs when compared to random SNPs.

Another important factor affecting genomic prediction accuracy is the choice of statistical method, or prediction model (H. Zhang et al., 2019). The standard additive GBLUP prediction model was used for all main text predictions in this study. When observing the ability of the standard additive GBLUP with no additional terms, GBLUP + E, and GBLUP + E + GxE, the simplest model excluding E and GxE yielded the highest prediction accuracies for eleven of thirteen stability measures (Table S2.4). Coefficient of Variation and Nassar and Huhn's S2 statistic were the two stability measures with increased prediction accuracy from including E and GxE in the prediction model. The equal performance from the more complex models may be due to the stability measures accounting for the GxE of each line, as stability is essentially a measurement of GxE (Becker & Léon, 1988). Also, the addition of E as an incidence matrix identifying the specific environments where each line was grown likely had no impact on performance due to the low number of environments in this study. Aside from these reasons, the results of lowered prediction accuracy with the E + GxE model are somewhat surprising given the number of previous studies demonstrating increased accuracy with such a model (Burgueño et al., 2012; Jarquín, Crossa, et al., 2014; Ray et al., 2022).

Prediction of across-environment yield BLUPs was conducted so that methods from this study could be compared to other studies. Yield BLUP prediction accuracies for TP and VP cross-validation are comparable to other studies, with average cross-validation r_{MP} values of 0.31 for TP and 0.38 for VP (Table 2.2). Stewart-Brown and colleagues assessed prediction of yield BLUPs in a similar manner within an applied soybean breeding program (Stewart-Brown et al., 2019). They achieved yield predictive

abilities of roughly 0.26. For across-population prediction, we achieved an r_{MP} of -0.11 while the aforementioned study had an r_{MP} of 0.1. The literature regarding yield prediction across populations of this nature using the BLUP approach is sparse. We hypothesized that there would be decreased predictive abilities across populations compared to within-population cross-validation, but the magnitude of this change was not expected. The negative predictive ability across-populations may have been caused primarily by prediction across less-related individuals. There are no full or half-siblings shared between populations (Figure S2.20). Another factor that likely affected predictive ability was the vast difference in environments between populations, as they were tested across different years. Some of this difference is shown by differentiation in variance components and heritability between TP and VP (Table 2.1), while the contrast in environment mean yield and ranking (PC1) are also evident between populations (Figures S2.8 - S2.11). The negative predictive ability was certainly surprising given the ability to predict yield stability across populations.

The assessment of stability traits comes with multiple limitations. One primary limitation to all stability studies is acquiring sufficient phenotype data in many environments across multiple years. A greater environment number would have allowed us to look at other factors influencing prediction accuracy for yield stability, such as the number of environments needed to closely resemble true stability within a region, and how environment number affects prediction accuracy for stability. Possibly the most influential limitation of this study is the low repeatability of stability values across different sets of environments (Pham & Kang, 1988; Sneller & Dombek, 1997). One point that should be further explored is the aspect of stability being predicted. When

conducting across-population genomic prediction with training and validation lines grown in different sets of environments, what exactly is being predicted? Is it the stability of VP lines in training set environments, or an aspect of stability that is under greater genetic control and consistent across different sets of environments? These are important questions to answer in order to further understand the possible role and utility of genomic prediction for stability traits.

Conclusion

The goal of this study was to lay the groundwork for genomic prediction of yield stability and other stability traits for application in a breeding program. Across-population prediction accuracies were generally too low for applied use when predicting AMMI and conventional dynamic stability measures. On the other hand, non-parametric measures displayed improved ability to be predicted, and displayed rank coincidence values approaching an acceptable level for applied use. The use of high-effect SNPs from GWAS with a density of 1,500 markers, and a standard additive GBLUP model without inclusion of E and GxE terms, resulted in the highest prediction accuracies for yield stability across two breeding populations. These results will inform future studies looking to improve on our methods, possibly by assessing predictability with a greater number of training and validation environments, implementing more complex prediction models, and assessing other types of genomic data such as tagSNPs. With improvements to methods and experimental design, it seems possible to achieve sufficient prediction accuracy for application of genomic prediction for yield stability in an applied scenario across breeding populations, specifically for non-parametric stability measures. The effectiveness of improving yield stability using genomic selection must also be compared

to traditional phenotypic selection and genomic selection for yield in order to determine if this method is worthwhile in an applied scenario. In turn, these methods have potential to give breeders greater selection accuracy at an earlier stage and increase the rate of genetic gain for stability traits in breeding programs.

References

Ajay, B. C., Aravind, J., & Abdul Fiyaz, R. (2019). Ammistability: R package for ranking genotypes based on stability parameters derived from AMMI model. *Indian Journal of Genetics and Plant Breeding*, *79*(2), 460–466. https://doi.org/10.31742/IJGPB.79.2.10

Becker, H. C., & Léon, J. (1988). Stability Analysis in Plant Breeding. *Plant Breeding*, *101*(1), 1–23. https://doi.org/10.1111/j.1439-0523.1988.tb00261.x

Bernardo, R. (2008). Molecular markers and selection for complex traits in plants: Learning from the last 20 years. *Crop Science*, *48*(5), 1649–1664. https://doi.org/10.2135/cropsci2008.03.0131

Bernardo, R., & Yu, J. (2007). Prospects for genomewide selection for quantitative traits in maize. *Crop Science*, *47*(3), 1082–1090. https://doi.org/10.2135/cropsci2006.11.0690

Best, D. J., & Roberts, D. E. (1975). The Upper Tail Probabilities of Spearman ' s Rho. *Journal of the Royal Statistical Society*, *24*(3), 377–379.

Brøndum, R. F., Su, G., Janss, L., Sahana, G., Guldbrandtsen, B., Boichard, D., & Lund, M. S. (2015). Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction. *Journal of Dairy Science*, *98*(6), 4107–4116. https://doi.org/10.3168/jds.2014-9005

Browning, B. L., Zhou, Y., & Browning, S. R. (2018). A One-Penny Imputed Genome from Next-Generation Reference Panels. *American Journal of Human Genetics*, *103*(3), 338–348. https://doi.org/10.1016/j.ajhg.2018.07.015

Burgueño, J., de los Campos, G., Weigel, K., & Crossa, J. (2012). Genomic prediction of breeding values when modeling genotype × environment interaction using pedigree and dense molecular markers. *Crop Science*, *52*(2), 707–719. https://doi.org/10.2135/cropsci2011.06.0299

Butler, D. G., Cullis, B. R., Gilmour, A. R., Gogel, B. J., & Thompson, R. (2018). *ASReml-R Reference Manual Version 4*. http://www.homepages.ed.ac.uk/iwhite/asreml/uop.

Edwards, S. M. K., Buntjer, J. B., Jackson, R., Bentley, A. R., Lage, J., Byrne, E., Burt, C., Jack, P., Berry, S., Flatman, E., Poupard, B., Smith, S., Hayes, C., Gaynor, R. C., Gorjanc, G., Howell, P., Ober, E., Mackay, I. J., & Hickey, J. M. (2019). The effects of training population design on genomic prediction accuracy in wheat. *Theoretical and Applied Genetics*, *132*(7), 1943–1952. https://doi.org/10.1007/s00122-019-03327-y

Finlay, K. W., & Wilkinson, G. N. (1963). The analysis of adaptation in a plant-breeding programme. *Australian Journal of Agricultural Research*, *14*(6), 742–754. https://doi.org/10.1071/AR9630742

Francis, T. R., & Kannenberg, L. W. (1978). YIELD STABILITY STUDIES IN SHORT-SEASON MAIZE. A DESCRIPTIVE METHOD FOR GROUPING GENOTYPES. *Canadian Journal of Plant Science*, *58*(4), 1029–1034.

Gauch Jr., H. G. (1988). Model Selection and Validation for Yield Trials with

Interaction. *International Biometric Society*, *44*(3), 705–715.

Happ, M. M., Graef, G. L., Wang, H., Howard, R., Posadas, L., & Hyten, D. L. (2021). Comparing a Mixed Model Approach to Traditional Stability Estimators for Mapping Genotype by Environment Interactions and Yield Stability in Soybean [Glycine max (L.) Merr.]. *Frontiers in Plant Science*, *12*(March), 1–12. https://doi.org/10.3389/fpls.2021.630175

Happ, M. M., Wang, H., Graef, G. L., & Hyten, D. L. (2019). Generating high density, low cost genotype data in Soybean [Glycine max (L.) Merr.]. *G3: Genes, Genomes, Genetics*, *9*(7), 2153–2160. https://doi.org/10.1534/g3.119.400093

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

Henderson, C. R. (1984). *Applications of Linear Models in Animal Breeding*. University of Guelph, Guelph, Ontario, Canada.

Huang, M., Cabrera, A., Hoffstetter, A., Griffey, C., Van Sanford, D., Costa, J., McKendry, A., Chao, S., & Sneller, C. (2016). Genomic selection for wheat traits and trait stability. *Theoretical and Applied Genetics*, *129*(9), 1697–1710. https://doi.org/10.1007/s00122-016-2733-z

Jarquín, D., Crossa, J., Lacaze, X., Du Cheyron, P., Daucourt, J., Lorgeou, J., Piraux, F., Guerreiro, L., Pérez, P., Calus, M., Burgueño, J., & de los Campos, G. (2014). A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theoretical and Applied Genetics*, *127*(3), 595–607. https://doi.org/10.1007/s00122-013-2243-1

Jarquín, D., Kocak, K., Posadas, L., Hyma, K., Jedlicka, J., Graef, G., & Lorenz, A. (2014). Genotyping by sequencing for genomic prediction in a soybean breeding population. *BMC Genomics*, *15*(1), 1–10. https://doi.org/10.1186/1471-2164-15-740

Kang, M. S. (1988). A RANK-SUM METHOD FOR SELECTING HIGH-YIELDING , STABLE CORN GENOTYPES Author (s): M . S . Kang Published by : Akadémiai Kiadó Stable URL : http://www.jstor.org/stable/23782771 Accessed : 28-06-2016 09 : 15 UTC. *Cereal Research Communications*, *16*(1–2), 113–115.

Keim, P. (1988). A rapid protocol for isolating soybean DNA. *Soybean Genetics Newsletter*, *15*, 150–152.

Lin, C. S., Binns, M. R., & Lefkovitch, L. P. (1986). Stability Analysis: Where Do We Stand? *Crop Science*, *26*(5), 894–900.

Lozada, D. N., & Carter, A. H. (2020). Insights into the genetic architecture of phenotypic stability traits in winter wheat. *Agronomy*, *10*(3). https://doi.org/10.3390/agronomy10030368

Luo, Z., Yu, Y., Xiang, J., & Li, F. (2021). Genomic selection using a subset of SNPs identified by genome-wide association analysis for disease resistance traits in aquaculture species. *Aquaculture*, *539*(December 2020), 736620. https://doi.org/10.1016/j.aquaculture.2021.736620

Matukumalli, L. K., Lawley, C. T., Schnabel, R. D., Taylor, J. F., Allan, M. F., Heaton, M. P., O'Connell, J., Moore, S. S., Smith, T. P. L., Sonstegard, T. S., & Van Tassell, C. P. (2009). Development and Characterization of a High Density SNP Genotyping Assay for Cattle. *PLoS ONE*, *4*(4). https://doi.org/10.1371/journal.pone.0005350

Meher, P. K., Rustgi, S., & Kumar, A. (2022). Performance of Bayesian and BLUP

alphabets for genomic prediction: analysis, comparison and results. *Heredity*, *128*(6), 519–530. https://doi.org/10.1038/s41437-022-00539-9

Meuwissen, T., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*(4), 1819–1829.

Meuwissen, Theo. (2007). Genomic selection : Marker assisted selection on a genome wide scale. *Journal of Animal Breeding and Genetics*, *124*(6), 321–322. https://doi.org/10.1111/j.1439-0388.2007.00708.x

Nassar, R., & Hühn, M. (1987). Studies on Estimation of Phenotypic Stability : Tests of Significance for Nonparametric Measures of Phenotypic Stability Author (s): R . Nassar and M . Hühn Published by : International Biometric Society Stable URL : http://www.jstor.org/stable/2531947. *Biometrics*, *43*(1), 45–53.

Pham, H. N., & Kang, M. S. (1988). Interrelationships among and Repeatability of Several Stability Statistics Estimated from International Maize Trials. *Crop Science*, *28*(6), 925–928. https://doi.org/10.2135/cropsci1988.0011183x002800060010x

Poplin, R., Ruano-Rubio, Valentin DePristo, M. A., Fennell, T. J., O Carneiro, M., Van der Auwera, G. A., Kling, D. E., Gauthier, L. D., Levy-Moonshine, A., Roazen, D., Shakir, K., Joel, T., Sheila, C., Chris, W., Monkol, L., Stacey, G., Mark, J. D., Ben, N., Daniel, G. M., & Eric, B. (2018). Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxIV*, 1–22.

Pour-Aboughadareh, A., Khalili, M., Poczai, P., & Olivoto, T. (2022). Stability Indices to Deciphering the Genotype-by-Environment Interaction (GEI) Effect: An Applicable Review for Use in Plant Breeding Programs. *Plants*, *11*(3). https://doi.org/10.3390/plants11030414

Pour-Aboughadareh, A., Yousefian, M., Moradkhani, H., Poczai, P., & Siddique, K. H. M. (2019). STABILITYSOFT: A new online program to calculate parametric and non-parametric stability statistics for crop traits. *Applications in Plant Sciences*, *7*(1), 1–6. https://doi.org/10.1002/aps3.1211

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*(8), 904–909. https://doi.org/10.1038/ng1847

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., De Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

Ray, S., Jarquin, D., & Howard, R. (2022). Comparing artificial-intelligence techniques with state-of-the-art parametric prediction models for predicting soybean traits. *Plant Genome*, *August 2021*, 1–19. https://doi.org/10.1002/tpg2.20263

Shukla, G. K. (1972). Some statistical aspects of partitioning genotype-environmental components of variability. *Heredity*, *29*(2), 237–245. https://doi.org/10.1038/hdy.1972.87

Sneller, C. H., & Dombek, D. (1997). Repeatability of Yield Stability Statistics in Soybean. *Crop Science*, *34*, 383–390.

Stewart-Brown, B. B., Song, Q., Vaughn, J. N., & Li, Z. (2019). Genomic selection for yield and seed composition traits within an applied soybean breeding program. *G3:*

Genes, Genomes, Genetics, *9*(7), 2253–2265. https://doi.org/10.1534/g3.118.200917

Team, R. C. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.r-project.org/

van den Berg, I., Boichard, D., Guldbrandtsen, B., & Lund, M. S. (2016). Using sequence variants in linkage disequilibrium with causative mutations to improve across-breed prediction in dairy cattle: A simulation study. *G3: Genes, Genomes, Genetics*, *6*(8), 2553–2561. https://doi.org/10.1534/g3.116.027730

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, *91*(11), 4414–4423. https://doi.org/10.3168/jds.2007-0980

Wang, H., Campbell, B., Happ, M., McConaughy, S., Lorenz, A., Amundsen, K., Song, Q., Pantalone, V., & Hyten, D. (2022). Development of molecular inversion probes for soybean progeny genomic selection genotyping. *Plant Genome*, *August 2022*, 1–11. https://doi.org/10.1002/tpg2.20270

Wang, Y., Mette, M. F., Miedaner, T., Wilde, P., Reif, J. C., & Zhao, Y. (2015). First insights into the genotype-phenotype map of phenotypic stability in rye. *Journal of Experimental Botany*, *66*(11), 3275–3284. https://doi.org/10.1093/jxb/erv145

Werner, C. R., Gaynor, R. C., Gorjanc, G., Hickey, J. M., Kox, T., Abbadi, A., Leckband, G., Snowdon, R. J., & Stahl, A. (2020). How Population Structure Impacts Genomic Selection Accuracy in Cross-Validation: Implications for Practical Breeding. *Frontiers in Plant Science*, *11*(December), 1–14. https://doi.org/10.3389/fpls.2020.592977

Wricke, G. (1962). Evaluation Method for Recording Ecological Differences in Field Trials. *Z Pflanzenzücht*, *47*, 92–96.

Yin, L., Zhang, H., Tang, Z., Xu, J., Yin, D., Zhang, Z., Yuan, X., Zhu, M., Zhao, S., Li, X., & Liu, X. (2021). rMVP: A Memory-efficient, Visualization-enhanced, and Parallel-accelerated Tool for Genome-wide Association Study. *Genomics, Proteomics and Bioinformatics*, *19*(4), 619–628. https://doi.org/10.1016/j.gpb.2020.10.007

Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., Kresovich, S., & Buckler, E. S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, *38*(2), 203–208. https://doi.org/10.1038/ng1702

Zhang, H., Yin, L., Wang, M., Yuan, X., & Liu, X. (2019). Factors affecting the accuracy of genomic selection for agricultural economic traits in maize, cattle, and pig populations. *Frontiers in Genetics*, *10*(MAR), 1–10. https://doi.org/10.3389/fgene.2019.00189

Zobel, R. W., Wright, M. J., & Gauch Jr., H. G. (1988). Statistical Analysis of a Yield Trial. *Agronomy Journal*, *80*(3), 388–393.

CHAPTER III

GENOMIC SELECTION FOR SEED COMPOSITION STABILITY

Abstract

The composition of soybean seeds, particularly protein and oil content, are critical to the commodity value and end use of soybeans after harvest. It is important for breeders to ensure that the lines released for public use have consistent seed composition across a range of environments. However, phenotypic selection for seed composition stability cannot occur until data are collected from multiple years and locations. The effectiveness and accuracy of genomic prediction for seed composition stability was assessed, most notably by using a real-world genomic prediction between two populations (training and validation) separated by three years in the same breeding program. The data used included whole genome sequence data and three static stability measures (environmental variance, regression coefficient, and Wricke's ecovalence) for protein and oil content across three years and 7-9 environments per line. Our results show that predictive abilities (r_{MP}) in cross-validation and across populations were low for both protein stability (-0.01 – 0.17) and oil stability (-0.02 – 0.41), and there was little difference in accuracy between across population prediction and cross-validation. Validation population oil stability cross-validation resulted in the highest predictive ability of 0.41 and rank coincidence of 0.66. In predicting stability, the standard additive GBLUP model was compared to a GBLUP model accounting for environment and genotype-by-environment interaction. The complex interaction model did not yield increased prediction accuracies, likely because stability values already account for GxE across environments. Overall, the methods highlighted for prediction of seed composition

stability across breeding populations require improvement before use in a real-world scenario. This improvement could come through the use of different prediction models, SNP selection methods, and more robust training and validation phenotype data.

Introduction

Soybean production provides a leading source of protein for livestock feed and oil for human consumption, and biofuels. Seed composition determines the commodity value and end use of soybean seeds after harvest. Soybeans with high protein and high oil are desirable, though the two components are negatively correlated, increasing the difficulty of maximizing both protein and oil at once (Shannon et al., 1972). Protein content and yield are also negatively correlated (Shannon et al., 1972), and farmers are paid for upholding a minimum protein content with no extra profit for exceeding this level. Due to the lack of profit and lower yield from increased protein content, it is desirable for soybeans to maintain the minimum protein content without exceeding by a large amount. Soybean seed composition and nutritional value are known to be largely controlled by genotype, but also vary due to environmental conditions at the time of seed development (Whaley & Eskandari, 2019; Wijewardana et al., 2019). This variability and uncertainty of seed composition across environments is of interest to breeders, as it affects the end use of harvested soybeans and the profitability of farmers. Phenotypic stability is a measure of consistency of a phenotype across a range of environments (Becker & Léon, 1988). There are a variety of ways to measure stability, with the two primary categories being static and dynamic stability (Becker & Léon, 1988; Lin et al., 1986). Static stability is a measurement of among-environment variance, or the degree to which a genotype's performance remains unchanged across environments. This is often used for disease

resistance traits, or traits that must be maintained "at all cost" (Becker & Léon, 1988). Dynamic stability is the degree to which a genotype's response is consistent with the population's mean phenotype in each environment, ensuring greater performance in more favorable environments. The choice of stability category depends on the goal for selection. For seed components, the goal is often to select for elite levels of one seed component (protein or oil) and maintain a minimum level of the other to maximize profitability and end-use of soybean seeds. As mentioned earlier, it is beneficial to maintain a protein content close to the specified minimum level in order to maximize yield. In this case, static stability provides a measurement of consistency to ensure that a minimum seed composition is upheld irrespective of the environment the soybeans are grown in.

Soybean seed composition stability has been assessed using the weighted average of absolute scores biplot (WAASB) for seven seed quality traits, with use of the multi-trait stability index (MTSI) to select the most stable lines for all components simultaneously (Abdelghany et al., 2021; Olivoto et al., 2019). However, WAASB is a stability measure based on AMMI and BLUP models that utilizes a mixed linear model to partition genotype-by-environment interactions (GxE), lending itself to more of a dynamic stability estimate (Olivoto et al., 2019). Conversely, to measure static stability, environmental variance (EV) and Finlay and Wilkinson's regression coefficient (b_i) are commonly used methods. Low values of EV are more stable for environmental variance (Becker & Léon, 1988; Roemer, 1917). The regression coefficient is less straightforward in Finlay and Wilkinson's method.. Finlay and Wilkinson defined b_i closest to zero is the most stable according to the static concept, but Eberhart and Russell defined $b_i = 1$ to be

stable in accordance to the dynamic concept (Becker & Léon, 1988; Eberhart & Russell, 1966; Finlay & Wilkinson, 1963). The regression coefficient (b_i) had been used previously to assess soybean seed composition stability (Balešević-Tubić et al., 2011). Due to the variety of opinions surrounding b_i, EV is the preferred and more true static stability measurement.

Another commonly used stability measure is Wricke's ecovalence. Generally, Wricke's ecovalence is thought to be a stability measure under the dynamic concept because it calculates phenotypic consistency based on the mean of each environment a line is tested in (Becker & Léon, 1988; Wricke, 1962). This also means that Wricke's can be used in a static concept if environmental means are equal or relatively similar. Pour-Aboughadareh and colleagues have noted that "dynamic and static stability concepts depend on the nature of the data" (Pour-Aboughadareh et al., 2022). If environments are similar, dynamic measures such as Wricke's ecovalence can be used to quantify static stability.

The biggest limitation to phenotypic selection of stability traits is the amount of time and resources needed to obtain sufficient data. There currently is not a consensus number of environments and years of data needed to accurately assess stability. The regression coefficient (b_i) has been shown to display moderate repeatability across only two years, but most other stability measures require more robust datasets (Sneller & Dombek, 1997). For stability analysis, the more environments and years of data, the more accurate the resulting stability value will be (Becker & Léon, 1988). When more environments are available, calculated stability is closer to its "true" stability value. The act of growing the same lines across numerous environments and years is not only costly

in terms of monetary expense and labor, but the time needed to complete this process can be detrimental to selection accuracy for stability traits. Genomic prediction has been proven a viable method for overcoming these types of challenges in plant breeding programs.

The prediction of complex traits has become more accessible over the past couple of decades as sequencing technology has advanced and cost of obtaining genome sequence data has declined (Bernardo, 2008; Meuwissen, 2007). Genomic selection involves developing a population to train a prediction model with genome-wide SNP (single nucleotide polymorphism) markers and phenotype data from the trait of interest (Meuwissen et al., 2001). Then, the phenotype of an individual with available sequence data can be predicted without being observed in the field. This method provides a powerful tool to expedite the selection process, and has been shown to increase the rate of genetic gain in breeding programs (Bernardo & Yu, 2007; Heffner et al., 2010). Genomic prediction accuracy has been shown to be influenced by marker density, statistical method, heritability, and genetic architecture of the trait (H. Zhang et al., 2019). It has also been shown that training population size, composition, and relation to selection population are key factors affecting prediction accuracy (Edwards et al., 2019; Werner et al., 2020). Elements such as heritability and genetic architecture are intrinsic to a trait and cannot be changed, but other factors mentioned such as training population design, statistical method, and marker density can be modified to increase accuracy for a particular trait. In terms of statistical method, the standard additive GBLUP model has proven efficient in predicting complex traits with many small-effect QTL when compared to other models such as Bayesian, semiparametric (RKHS), and more complex additive

models (Huang et al., 2016; Jarquín, Kocak, et al., 2014; Meher et al., 2022). Genomic

selection may be the best available option to overcome the limitations of assessing seed

composition stability early in the breeding process and aid in the timely advancement of

stable genotypes in a breeding program.

Currently, little is known about the predictability of stability traits, including seed

composition stability in soybean. Static stability measures are more repeatable across

years and environments than dynamic stability (Lin & Binns, 1991). This may mean that

static stability is more predictable than dynamic stability across populations and different

sets of environments. Due to the negative correlation between protein and oil content,

there are also questions about the relationship between protein stability and oil stability

that have yet to be answered. In addition are protein and oil stability correlated, and are

they equally predictable using genomic selection, or is one component is more

predictable than the other. These are key details for breeders to understand when

developing methods for using genomic prediction for selection of seed composition

stability in a breeding program.

In this study, genomic prediction for soybean seed composition stability is

examined. The primary objective was to assess the practicality of applying genomic

prediction of seed composition stability traits in a real-world breeding program. The

independent training and validation populations used in this study were derived from the

University of Nebraska soybean breeding program. They were developed three years

apart from each other in the same breeding program and field tested in independent years.

The effects of SNP marker density, selection method of SNP markers for including in the

genomic selection model, stability measure, and prediction scenario on prediction

accuracy for stability of soybean protein and oil concentration were evaluated separately. The correlation between protein stability and oil stability was also investigated to give insight on the best practices of selecting for these traits.

Materials and Methods

Plant Material

The populations in this study were acquired from the University of Nebraska-Lincoln (UNL) Soybean Breeding program. The training population from the soybean breeding program consists of 197 experimental F_4-derived lines from elite yield crosses, ranging from maturity group I to IV. This population was originally yield tested for three years, 2017 through 2019, for a yield stability GWAS study (Happ et al., 2021). The validation population was also selected from the UNL soybean breeding program and developed three years after the TP. The VP consists of 200 lines that range in maturity from groups II and III. Before analysis, TP lines were be filtered to 190 lines to avoid identical lines in both populations, and VP lines were filtered to 165 total lines as 35 lines had sequence coverage less than 0.3X which was the cutoff for the sequence coverage needed for accurate SNP imputation.

Field Tests

Field trials for this study were conducted over the course of six years and ten locations (Figure S2.1). The training population, as described by Happ et al., (2021) was tested from 2017 to 2019 in thirteen total environments. Field tests were primarily located in the eastern half of Nebraska. Field trial location was determined based on maturity group. Maturity groups I and II were grown at combinations of Cotesfield, Phillips, Mead, Lincoln, and Wymore across three years. Maturity group III and IV lines were

grown at combinations of Phillips, Lincoln, and Wymore across three years. Overall, each line was grown in at least 7-8 environments over three years. An augmented incomplete randomized block design was used for the UNL training population yield trials with three replicates grown at each site. Two-row plots were planted in 2017, and four-row plots grown in 2018 and 2019 with the center two rows harvested. Prior to planting, seed was treated with CruiserMaxx before planting and grain weight and moisture were taken at harvest and adjusted to 13% moisture for final yield calculation.

The validation population yield trials took place over three years from 2020 to 2022, across 18 total environments. VP environments were primarily located in the eastern half of Nebraska, along with locations in Iowa and Missouri. Maturity group II was grown at combinations of Cotesfield, Lincoln, Mead, Phillips, and Arcadia, IA across three years. Group III lines were grown at combinations of Cook, Lincoln, Mead, Phillips, Albany and Novelty, MO, and Winterset, IA across three years. All lines were grown in 11 total environments across 3 years. An augmented incomplete randomized block design was used for validation population field tests. Blocks contained 40 lines for MGII and 28 lines for MGIII including maturity checks. Two replicates of each line were grown per site. Plots were planted in four-row plots of 6 meters in length with 0.76-meter row spacing. The center two rows were harvested for each plot. Seed was treated with CruiserMaxx before planting, with grain weight and moisture collected at harvest and adjusted to 13% for final yield calculation.

Seed composition measurement

Seed samples from field tests were used to measure seed composition. The Foss Infratec 1241 grain analyzer and the Foss Infratec Nova NIR analyzer (FOSS, Hillerød,

Denmark) were both used to measure seed composition. In total, TP lines in MGI and

MGII had 8 environments with three repetitions for seed composition. MGIII and MGIV

had 7 environments with three repetitions for seed composition. All VP lines had 9

environments of seed composition data, with two replications per environment.

Protein and Oil Content Analysis

Variance components for protein and oil were tested using the following model

$$(1) \quad Y = g_i + e_j + ge_{ij} + b_k(r_l) + \varepsilon$$

where Y is the vector of protein or oil values, g_i is the effect of genotype i, e_j is the

environmental effect, ge_{ij} is the interaction between genotype and environment, $b_k(r_l)$ is

replicate nested within block in each environment, and ε is the residual error. Each term

was fit as random. The model was run in ASReml-R and resulting BLUPs were used as

observed values for yield prediction accuracy calculation (Butler et al., 2018). Broad-

sense heritability was calculated from ANOVA variance components as $H^2 =$

$\frac{\sigma_G^2}{\sigma_G^2 + \sigma_{GE/er}^2 + \sigma_{\varepsilon/r}^2}$ where σ_G^2 is the genotypic variance, σ_{GE}^2 is genotype-by-environment

interaction variance, σ_ε^2 is the residual variance, e is the number of environments, and r

is the number of replications per environment.

Prior to protein and oil prediction, BLUEs were calculated in ASReml-R for each

line within each environment to correct for experimental design:

$$(2) \quad Y = g_i + b_k(r_l) + \varepsilon$$

Where Y is the vector of raw yield values, g_i is the fixed effect of genotype i, $b_k(r_l)$ is

replicate nested within block in each environment as a random effect, and ε is the random

residual error. The resulting BLUEs were used as training values for genomic prediction

of yield BLUPs across all environments using the GBLUP model in linear form:

$$(3) \quad Y = g_i + e_j + ge_{ij} + \varepsilon$$

Predicted GEBVs were then compared to observed BLUPs from the first model to assess

prediction accuracy for protein BLUPs and oil BLUPs.

Seed Composition Stability

Seed compositions were then used to calculate stability parameters. The

regression coefficient was calculated using 'stabilitysoft', an R package allowing the

calculation of over a dozen static and dynamic stability parameters (Pour-Aboughadareh

et al., 2019). Environmental variance was calculated by writing a script in R (R Core

Team, 2021). For stabilitysoft calculation, the reps for each genotype in each

environment were averaged. For this study, the static stability concept is used. There are

two primary static stability measures. The first is environmental variance (S^2), which is

essentially a measure of deviation in phenotype across environments, relative to a

genotype's mean (Becker & Léon, 1988; Roemer, 1917). The equation for environmental

variance is:

$$(4) \quad S^2 = \sum \frac{(R_{ij} - m_i)^2}{e - 1}$$

where R_{ij} is the yield of the ith genotype in the jth environment, m_i is the grand mean of

genotype i across all environments, and e is the number of environments. Minimum

values of S^2 show the greatest stability. Finlay and Wilkinson's coefficient of regression

(b$_i$) is calculated as follows (Finlay & Wilkinson, 1963):

$$(5) \quad b_i = 1 + \frac{\sum_j (X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}_{..})(\bar{X}_{.j} - \bar{X}_{..})}{\sum_j (\bar{X}_{.j} - \bar{X}_{..})^2}$$

where X$_{ij}$ is the observed phenotypic mean value of genotype i in environment j, $\bar{X}_{i.}$ is the

mean of genotype i across all environments, $\bar{X}_{.j}$ is the overall mean of environment j, and

$\bar{X}_{..}$ is the grand mean. While regression coefficient can be used to identify dynamic stability, a genotype stable according to the static concept has $b_i = 0$.

Wricke's ecovalence is a measure for partitioning GxE interactions. Wricke's calculates the deviation of a genotype's mean from the population's overall environmental mean and sums this deviation across all environments. Wricke's was tested in terms of static stability in this study. The formula for Wricke's ecovalence is as follows:

$$(6) \quad W_i = \sum_j (X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}_{..})^2$$

where X_{ij} = Mean of genotype i in environment j, $\bar{X}_{i.}$ = Mean yield of genotype i across environments, $\bar{X}_{j.}$ = Mean of environment j across genotypes, $\bar{X}_{...}$ = Overall mean (Wricke, 1962).

Genomic Data

Whole genome skim sequence methods for the training population can be found in Happ et al., (2021). A more complete NovaSeq whole genome sequence dataset was acquired for TP lines after the 2021 study, giving an average of 10X coverage (Illumina, Hayward CA, USA). Briefly, DNA extraction was performed using a CTAB method (Keim, 1988). DNA samples were normalized and pooled by plate before library preparation was conducted using the IGENOMIX Riptide High-Throughput Rapid Library Prep kit and protocol (IGENOMIX, Doral, FL). Whole genome sequence data was acquired using the Illumina NextSeq 500 (Illumina, Hayward CA, USA). For the validation population an average of < 1x coverage was obtained for each of the 200 lines. The reference panel for SNP genotyping and imputation was generated as described by Happ et al., (2019). SNP genotyping was conducted on the previously called SNPs using

GATK4 version 4.1 HaplotypeCaller (Poplin et al., 2018). Imputation was conducted using Beagle v5.2 and the aforementioned reference panel (Browning et al., 2018). After imputation, 10,286,817 genome-wide SNPs were available for each individual in both populations. Genomic data for the training and validation populations were filtered for MAF (minor allele frequency) < 0.05 and linkage disequilibrium r^2 > 0.4 using Plink1.9 (Purcell et al., 2007). MAF filtering discarded 8,849,209 SNPs in the training population to leave 1,437,578 SNPs. Of those SNPs, 865,499 were discarded due to LD filtering, leaving 572,079 TP SNP positions. For the validation population, MAF filtered out 8,561,523 SNPs leaving 1,725,264 positions. Of those, LD filtering removed another 916,959 positions to result in 808,305 SNPs. After separate filtering of both populations, TP and VP data sets were reduced to common SNP positions resulting in a total of 373,348 SNPs across all 20 chromosomes. Thirty-five of the 200 VP lines had < 0.3X sequence coverage and were filtered out. The threshold for imputation accuracy of 0.3X had been identified previously (Happ et al., 2019). At 0.3X, imputation accuracy of 93% is comparable to 94% accuracy when sequence coverage is 1X. Below 0.3X, imputation accuracy decreases quickly to roughly 90% at 0.1X coverage. With the goal of as little imputation error as possible, 0.3X was the coverage threshold for this study.

SNP selection

Two types of methods for selecting SNP markers to include in the genomic selection model were compared in this study. The first method was selecting SNPs with the highest effect on stability from a simple GWAS analysis. GWAS was conducted using the QK MLM method with equation is as follows (Price et al., 2006; Yu et al., 2006):

$$(7) \quad y = X\beta + S\alpha + Qv + Zu + e$$

where y is the vector of phenotypic observations; X is the incidence matrix for fixed

effects; b is the vector of fixed effects other than the SNP being tested and population

structure; S is the incidence matrix of SNP effects; α is the vector of SNP effects; Q is a

matrix relating y to v; v is a vector of population effects; Z is the incidence matrix for

random effects; u is the vector of background SNP effects; and e is residual error. GWAS

was conducted for both populations using the R package "R-MVP", using PCA to correct

for population structure (Yin et al., 2021). Principal components for population structure

correction were obtained using Plink1.9 (Purcell et al., 2007). Nine PCs were used for TP

correction of population structure, and six were used for VP structure correction. GLM

and MLM were compared, with MLM giving QQ-plot with observed quantiles closer to

expected. To avoid redundant markers in high linkage disequilibrium, the highest-effect

marker within each 10,000 base pair window was filtered into a list of 20,391 SNPs. For

each SNP density, an equal number of top effect markers from each chromosome were

selected. GWAS was conducted separately for both populations to simulate a real-world

scenario where phenotype data is not available for a population, and high-effect SNPs

from another population must be used. For this reason, high-effect SNPs from TP GWAS

were used for TP – VP prediction and VP cross validation. High-effect SNPs from VP

GWAS were used for TP cross validation. Manhattan and Q-Q plots for each scenario are

included in supplementary materials.

Random SNP selection was done by randomly choosing an equal number of SNP

positions from each chromosome across the genome. In analyzing the effect of SNP

density on prediction accuracy, as SNP sets increased in number, new random SNPs from

across all chromosomes were added to the smaller SNP set. This was done in R (R Core

Team, 2021). All random SNP prediction scenarios used the same random SNP sets.

Genomic Prediction

Genomic prediction was conducted using the GBLUP model. GBLUP requires a

genomic relationship matrix to quantify relationship between all individuals of the

populations involved. The VanRaden relationship matrix was used in this study

(VanRaden, 2008). The GBLUP equation is as follows:

$$(8) \quad Y = Xb + Zu + e$$

where Y is the response vector, X is the design matrix for fixed effects, b is the vector of

fixed effects, Z is the design matrix for random effects, u is the vector of predicted

random effects, and e is residual error. Genomic predictions were conducted using

ASReml-R (Butler et al., 2018). The three GBLUP models compared are shown in linear

form as (Henderson, 1984; Meuwissen et al., 2001):

$$(9) \quad y = \mu + g_i + \varepsilon_i$$

$$(10) \quad y = \mu + g_i + E_j + \varepsilon_i$$

$$(11) \quad y = \mu + g_i + E_j + gE_{ij} + \varepsilon_i$$

where y is the vector of response, μ is the intercept or mean to be added back to BLUP

values, g_i is the random additive VanRaden genomic relationship between individuals in

both populations with $g_i \sim N(0,\sigma^2{}_g)$, E_j is the random environmental effect as a binary

incidence matrix identifying lines grown in each environment with $E_i \sim N(0,\sigma^2{}_E)$, gE_{ij} is

the random interaction between genotype and environment with $gE_{ij} \sim N(0,\bigoplus_{j=1}^{J}\sigma^2{}_{gE})$,

and ε_{ij} is the residual with $\varepsilon_{ij} \sim N(0,\sigma^2{}_\varepsilon)$. The terms $\sigma^2{}_g$, $\sigma^2{}_E$, and $\sigma^2{}_{gE}$, and $\sigma^2{}_\varepsilon$ are the

variance components for genotype, environment, GxE, and residuals. Equation (10) is the

standard additive GBLUP model utilizing only genomic relationship as a random model term. Model (11) adds a random environmental term as an incidence matrix to specify which environments each line was grown in, and (12) adds the random interaction between genotype and environment. Model comparison of prediction accuracy was conducted to observe effectiveness for predicting yield stability across populations. Model comparison was conducted to observe effectiveness for predicting yield stability across populations.

Three different prediction scenarios were examined. The primary scenario was the training population used to train the model and predict the validation population, which is meant to simulate a real-world genomic prediction scenario. The prediction accuracy for this scenario was assessed when multiple SNP densities, SNP selection methods (GWAS vs random), and stability measures are used. Two other scenarios assessed were TP 5-fold cross-validation, and VP 5-fold cross-validation. Each 5-fold cross validation was repeated 5 times, with prediction accuracies averaged across the 5 folds within each repetition.

Three types of prediction accuracy were assessed in this study. Each were calculated in R using a correlation function or by calculation using basic functions (R Core Team, 2021). Typical predictive ability (r_{MP}) was measured as the correlation between predicted stability values and observed stability values. Rankings tend to be more useful for plant breeding purposes, so two measures of predicted ranking accuracy were assessed. Spearman's rank correlation coefficient is the correlation between predicted rankings and observed rankings (Best & Roberts, 1975). Spearman's rank correlation coefficient was calculated in R using the set of predicted stability ranks and

observed stability ranks. Lastly, an unofficial accuracy estimate referred to as "rank coincidence" measures the proportion of lines in the same half (top or bottom) of predicted stability rankings and observed stability rankings. This was calculated in R using observed and predicted stability rankings.

Results

Phenotype and Genotype data

The training population was tested in 14 total environments across three years in Central and Eastern Nebraska (Supplementary Figures 1 and 3). Mean protein across environments for the training population ranged from 32.3% to 35.3%, and mean oil ranged from 17.7% to 19.7%. The environment with highest average protein content was Phillips West 2018 with a mean of 35.3%, and the lowest average protein environment was Wymore 2017 with a mean of 32.3%. The environment with highest average oil content was Phillips North 2017 with 19.7%, and lowest average oil content environment was Phillips West 2018 with an average of 17.7%. The validation population was tested in 15 total environments across three years with locations in Nebraska, Iowa, and Missouri (Supplementary figures 2 and 4). Mean protein ranged from 31.1% to 34.0% across environments, and mean oil ranged from 19.2% to 20.8%. The highest average protein environment for the VP was Cotesfield, NE 2021 with 34.0%, and the lowest average protein environment was Lincoln, NE 2022 with 31.1%. The VP environments with highest and lowest mean oil were Mead Drip 2021/Cook 2022 and Phillips 2022 with 20.8% and 19.2% oil, respectively.

Table 3.1. Variance component estimates of protein and oil content data for both TP and VP. For heritability calculation, TP included 14 total environments and 3 replicates per environment, and the VP included 15 total environments and 2 replicates per environment.

Trait	Population	Block	block:rep(env)	Environment	Genotype	GxE	Residuals	Heritability
Protein	TP	0.069	0.072	0.495	0.052	0.250	0.847	0.15
Protein	VP	0.028	0.024	0.747	0.257	0.070	0.326	0.61
Oil	TP	0.054	0.048	0.121	0.022	0.092	0.241	0.21
Oil	VP	0.042	0.013	0.288	0.121	0.041	0.077	0.75

Each line was grown in a subset of its population's environments depending on maturity group. Heritability for protein and oil were higher in the validation population than the training population (Table 3.1). The variance component for genotype was much smaller, and genotype-by-environment much larger proportionally for TP protein and oil than VP. This may be due to the fact that each line was tested in fewer environments in the TP when compared to the VP. Stability for protein and oil content was calculated for each line in both populations for all environments they were tested in. The distribution of stability values within each population were generally skewed (Figures S3.7 – S3.10). Environments also varied slightly in mean protein and oil, as well as PC1 from AMMI analysis, indicating a difference in rankings (Figures S3.11 – S3.14). For each line, protein stability was plotted against oil stability for environmental variance, regression coefficient, and Wricke's ecovalence, to identify if any correlation is present between them (Figure 3.1). For EV and RC stability, no correlation was observed with r^2 values of 0.047 and 0.004, respectively. Wricke's ecovalence showed slightly greater correlation between protein and oil stability with $r^2 = 0.27$.
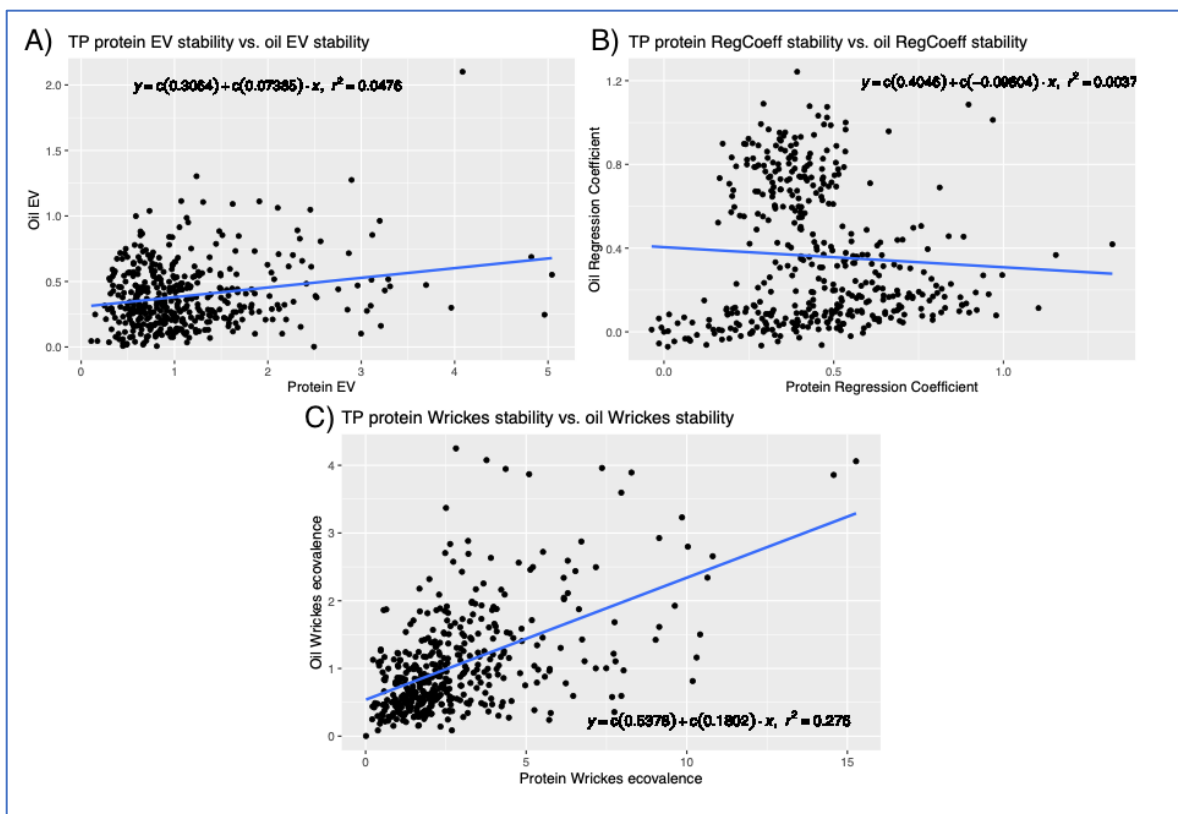
Figure 3.1. Correlation between protein stability and oil stability for three different stability measures in TP and VP combined. Stability measures include A) environmental variance (EV), B) regression coefficient, and C) Wricke's ecovalence.

Whole genome sequence coverage for the 190 TP lines was generally high with an average of 9.8X, ranging from 1X to 33X (Table S2.1). VP coverage for the 165 lines used in this study averaged 0.86X and ranged between 0.31X and 7.7X. After filtering MAF and LD, 373,348 genome-wide SNP markers were left for genomic prediction. The positions of these markers across chromosomes can be found in Supplementary Figure 18. For each population, GWAS analysis yielded effect estimates of each SNP marker for regression coefficient protein stability, regression coefficient oil stability, protein BLUPs across environments, and oil BLUPs across environments (Figures S3.15 – S3.22).

Protein and oil content prediction

Table 3.2. Cross-validation (CV) and across-population accuracies for protein content BLUP prediction. Mean predictive ability (r_{MP}) and standard deviation between r values across 5 folds within each repetition are reported. TP CV used 5,000 high-effect VP SNPs. Across-population prediction and VP CV used 5,000 high-effect TP SNPs.

TP cross-validation protein			VP cross-validation protein		
Rep	mean r	SD	Rep	mean r	SD
1	0.28	0.14	1	0.50	0.13
2	0.32	0.09	2	0.53	0.07
3	0.36	0.16	3	0.51	0.09
4	0.32	0.16	4	0.50	0.10
5	0.38	0.10	5	0.52	0.07
Average	**0.33**	0.13	Average	**0.51**	0.09
across-population r: **0.45**					

Table 3.3. Cross-validation (CV) and across-population accuracies for oil content BLUP prediction. Mean predictive ability (r_{MP}) and standard deviation between r values across 5 folds within each repetition are reported. TP CV used 5,000 high-effect VP SNPs. Across-population prediction and VP CV used 5,000 high-effect TP SNPs.

TP cross-validation oil			VP cross-validation oil		
Rep	mean r	SD	Rep	mean r	SD
1	0.23	0.05	1	0.47	0.11
2	0.25	0.08	2	0.49	0.17
3	0.23	0.11	3	0.53	0.15
4	0.17	0.18	4	0.52	0.10
5	0.23	0.13	5	0.50	0.10
Average	**0.22**	0.11	Average	**0.50**	0.13
across-population r: **0.21**					

In order to compare the genomic prediction with our data to other studies, the

prediction of protein and oil content BLUPs across all environments was evaluated

(Tables 3.2 and 3.3). For each trait, across-population TP prediction of VP, along with 5

repetitions of 5-fold cross-validation within TP and VP sets were conducted. For

prediction of protein content BLUPs, the TP and VP cross-validation averaged 0.33 and

0.51, respectively, for predictive abilities (Table 3.2). Across-population protein BLUP

prediction resulted in a predictive ability of 0.45. For oil content BLUP cross-validation,

average predictive abilities across repetitions were 0.22 and 0.50 for TP and VP cross-

validation (Table 3.3). Across-population prediction of oil BLUPs yielded a predictive

ability of 0.21. In cross-validation scenarios, the validation population had consistently

higher average predictive ability. This may be related to the heritability issue for protein

and oil in the training population. Across-population predictive ability was higher for

protein prediction than oil prediction.

Stability prediction across populations

Table 3.4. Prediction accuracy comparison between 3 stability measures (environmental variance, regression coefficient, Wricke's ecovalence) for protein and oil using the same set of 5,000 GWAS-selected SNP markers. Across-population prediction was conducted using TP to predict VP. Prediction accuracies include r_{MP} , rank correlation, and rank coincidence. Genomic heritability for each stability measure was calculated from prediction model variance components.

Stability measure	r	rank correlation	rank coincidince	Heritability
EV protein	0.164	0.145	0.59	0.439
Regression Coefficient protein	-0.012	-0.034	0.51	0.699
Wricke's ecovalence protein	0.041	-0.001	0.49	0.082
EV oil	0.154	0.144	0.58	0.198
Regression Coefficient oil	-0.014	-0.033	0.48	0.666
Wricke's ecovalence oil	0.097	0.021	0.53	0.047

Accuracy was assessed for prediction of protein and oil stability across

populations using TP genotype and phenotype data to predict VP stability (Table 3.4).

This scenario is similar to a real-world breeding scenario, where a validation population

is predicted using only data from a training population developed three years earlier in

the same program. Environmental variance, regression coefficient, and Wricke's

ecovalence were predicted for both protein and oil stability. Across both traits,

environmental variance was the stability measure with highest predictability across

populations with predictive ability (r_{MP}) of 0.16 and 0.15, rank correlation of 0.14 and

0.14, and rank coincidence of 0.59 and 0.58 for protein and oil stability, respectively.

Regression coefficient had r_{MP} of -0.01 and rank correlation of -0.03 for both traits, with

rank coincidence of 0.51 for protein stability and 0.48 for oil stability. Wricke's

ecovalence values were slightly higher with r_{MP} of 0.04 and 0.10, rank correlation of 0.0

and 0.02, and rank coincidence of 0.49 and 0.53 for protein stability and oil stability,

respectively. Genomic heritability was also assessed. Since the SNPs used were highest

effect GWAS-selected SNPs from TP regression coefficient stability, RC had the highest

heritability. EV had the second highest heritability across both traits, and Wricke's

ecovalence with consistently the lowest heritability. With rank coincidence of 0.59 and

0.58 for EV protein stability and oil stability prediction, a selection accuracy of 59% was

obtained when selecting the top half of lines when using prediction for selection.

Assuming 50% accuracy is random, our results indicate that genomic selection can

increase selection accuracy by approximately 9% compared to random selection for

protein and oil stability in an applied prediction scenario.
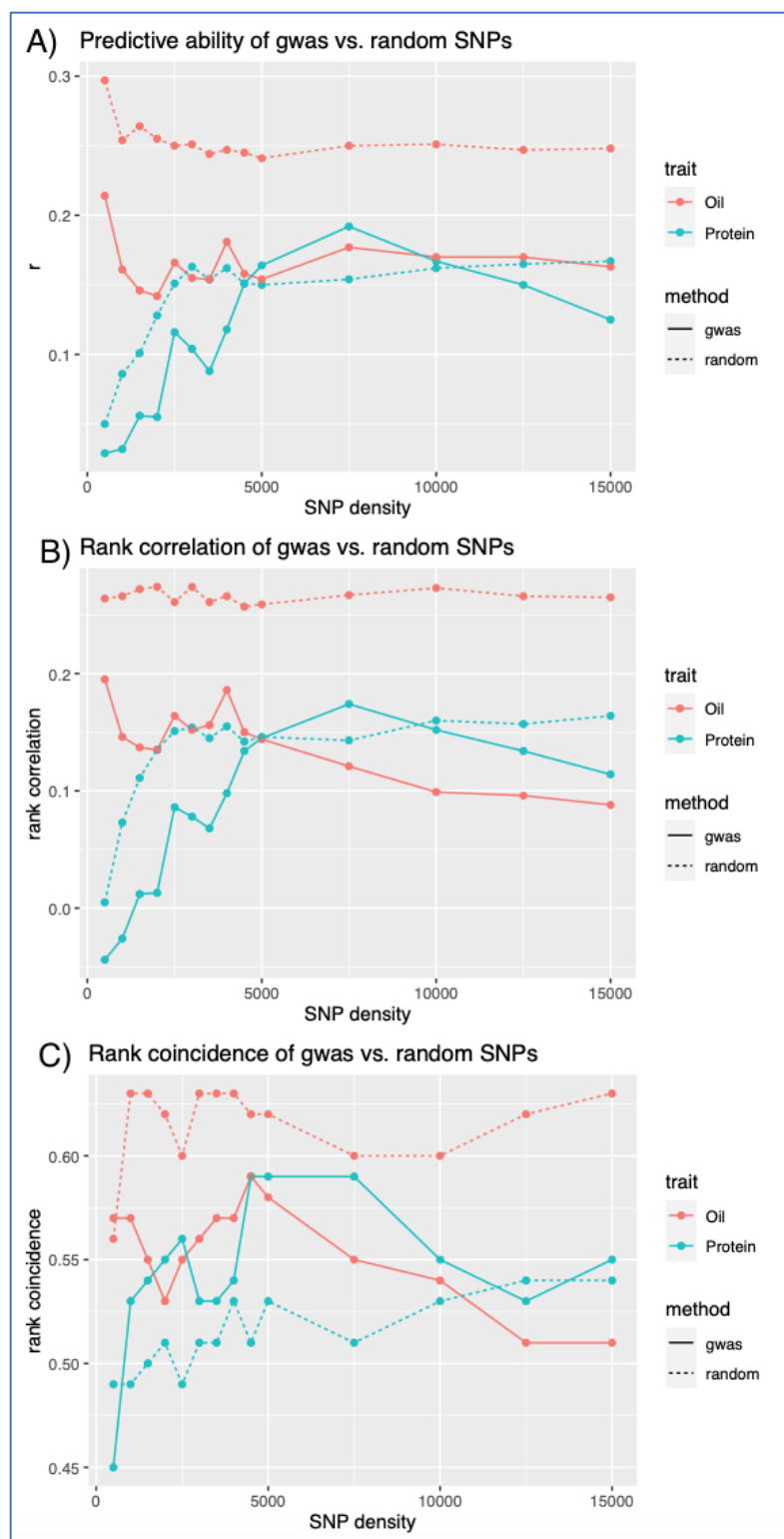
Figure 3.2. SNP density vs. three different types of prediction accuracy: A) predictive ability (r_{MP}), B) rank correlation, and C) rank coincidence, for across population prediction of protein (blue-green lines) and oil (red lines) EV stability. Solid lines represent GWAS SNPs, and dashed lines represent random SNPs. SNP densities ranged from 500 to 15,000.

There are many controllable factors that influence genomic selection accuracy. In order to inform the best practices for genomic prediction of protein and oil stability for our data, the effect of SNP marker density and method of SNP selection was tested on three types of genomic prediction accuracy (Figure 3.2). In the across-population prediction of environmental variance stability, a SNP density of 5,000 gives the highest prediction accuracy results. However, it was unexpected that in many scenarios, random SNPs performed better than GWAS-selected SNPs. In particular for oil stability, GWAS SNPs had lower associated prediction accuracy across nearly every SNP density. To visualize the effect of TP GWAS and random SNPs in terms on VP stability, SNP effect was plotted of 15,000 TP GWAS SNPs and 15,000 random SNPs on TP and VP protein and oil stability for the first five chromosomes (Figures S3.24 and S3.25). The high effect SNPs for TP protein and oil stability do not share the same effect for VP protein and oil stability. Generally, the highest effect sizes from GWAS analysis were low, often between +- 0.1. Furthermore during GWAS analysis, there were no significant markers identified (Figures S3.15 – S3.22).

Stability cross-validation within populations

Table 3.5. Average prediction accuracies for 5 repetitions of 5-fold cross-validation within the training population. Three stability measures for protein and oil were predicted with 5,000 VP GWAS SNPs for regression coefficient stability in protein and oil separately. Predictive ability (mean r), standard deviations (SD), rank correlation, and rank coincidence were averaged across all 5 folds of cross-validation.

Stability measure	mean r (SD)	rank correlation	rank coincidence
EV protein	0.277 (0.142)	0.140	0.526
Regression Coefficient protein	0.195 (0.156)	0.161	0.55
Wricke's ecovalence protein	0.187 (0.151)	0.066	0.468
EV oil	0.115 (0.114)	0.066	0.534
Regression Coefficient oil	0.195 (0.121)	0.135	0.56
Wricke's ecovalence oil	-0.021 (0.165)	-0.155	0.436

Table 3.6. Average prediction accuracies for 5 repetitions of 5-fold cross-validation within the validation population. Three stability measures for protein and oil were predicted with 5,000 VP GWAS SNPs for regression coefficient stability in protein and oil separately. Predictive ability (mean r), standard deviations (SD), rank correlation, and rank coincidence were averaged across all 5 folds of cross-validation.

Stability measure	mean r (SD)	rank correlation	rank coincidence
EV protein	0.197 (0.135)	0.166	0.566
Regression Coefficient protein	0.162 (0.146)	0.143	0.548
Wricke's ecovalence protein	0.178 (0.150)	0.056	0.496
EV oil	0.406 (0.096)	0.391	0.662
Regression Coefficient oil	0.353 (0.098)	0.335	0.648
Wricke's ecovalence oil	0.033 (0.103)	-0.073	0.458

Cross-validation within the TP and VP were conducted for protein and oil stability (Table 3.5 & 3.6). With protein stability, the EV measure had the highest mean predictive ability with 0.27 and 0.20 for TP and VP cross-validation, respectively. Rank coincidence values ranged from 0.47 to 0.57 for protein stability. Both populations were similar in their cross-validation predictive abilities, rank correlation, and rank coincidence measures. They CV protein stability accuracies were also similar to accuracy values from prediction across populations. For oil stability, EV and regression coefficient had the highest mean r_{MP} with 0.12 and 0.20 for TP cross-validation, and 0.46 and 0.35 for VP cross-validation. The highest oil stability rank coincidence for TP and VP cross-validation were 0.56 and 0.66, respectively. The least-predictable oil stability measure was Wricke's ecovalence across all three accuracy measurements in both populations. The validation population had higher cross-validation accuracies than the training population for oil stability. This may be due to the different environments and years the populations were tested in, along with the TP tested in fewer environments than VP. Also, TP oil stability values were generally skewed to a higher degree than VP oil stability values, which goes against the normal distribution assumption of GBLUP. When viewing the population structure figure (Figure S3.23), the two populations largely overlap, with little difference in population structure. Also, within each population, they have a very similar proportion of individuals with at least one other full-sibling or half-sibling (Figure S2.20). These items show that deviation in cross-validation accuracy is likely not caused by population structure, but more so the phenotype data available in this study.

Discussion

Genomic selection has potential to aid in the improvement of stability traits by allowing for prediction and selection of stability at a stage before adequate data are available for stability analysis. In this study, a basic foundation for prediction of stability traits in an applied breeding program has been provided. While genomic selection studies typically conduct cross-validation within a single population, a valuable method for model validation, this study assessed prediction across two separate populations from the same breeding program, allowing for assessment of expected prediction accuracy in an applied scenario.

Prediction accuracies for protein and oil stability for three stability measures were low across populations (Table 3.4). Of the three stability measures, environmental variance displayed the greatest ability to be predicted for both traits. While r_{MP} and rank coincidence remained lower than what would be considered "useful" in a plant breeding setting, EV stability was noticeably more predictable than regression coefficient and Wricke's ecovalence across populations. This could mean EV stability is more consistent across populations and different sets of environments, or that EV may have a greater level of genetic control. While there are no past studies to directly compare our results to, the two static stability measures, EV and regression coefficient, were expected to outperform the traditionally dynamic Wricke's ecovalence in this study. The reason for this is static stability has been shown to be more repeatable across different sets of environments (Lin & Binns, 1991). Dynamic stability measures such as Wricke's ecovalence measure the deviation between a line's phenotype in an environment and the population mean in that environment (Becker & Léon, 1988). Static stability measures do not rely on population values, and measure consistency by calculating the sum of

deviations between a genotype's yield in each environment and its mean yield across environments. In the context of this study, Wricke's ecovalence was used in a static concept due to the general similarity between environments with the same maturity groups for protein and oil content. Though not all environments had equal population means for protein and oil content (Figure S3.1 – S3.4), so Wricke's may have been calculating a slightly dynamic response. Overall, environmental variance, the preferred static stability measure (Becker & Léon, 1988), was most easily predicted in terms of all three prediction accuracy measures in an applied breeding setting.

SNP density has been identified as a primary factor affecting genomic prediction accuracy (H. Zhang et al., 2019). Across-population prediction results of environmental variance protein and oil stability were quite variable across SNP densities (Figure 3.2). A SNP density of 5,000 markers was generally most effective in predicting both protein and oil stability, though these trends varied slightly by trait and SNP type. While fluctuations in rank coincidence occurred when using greater than 5,000 SNPs, predictive ability and rank correlation generally remained consistent or slightly declined after this point.

Markers selected with highest effect size from GWAS analysis were expected to have higher prediction accuracy than random SNPs based on past studies (Brøndum et al., 2015; Luo et al., 2021; van den Berg et al., 2016). Surprisingly, random SNPs generally performed similar or better than GWAS SNPs across densities ranging from 500 to 15,000 for predictive ability, rank correlation, and rank coincidence. The reason for this unexpected result may be inconsistency in effect of TP GWAS SNPs in the VP. The difference in unbalanced experimental design between the two populations may be another factor that led to this unexpected result. For example, TP lines were tested in

seven or eight environments, while VP lines were tested in nine across different years. Second, the TP was made up of lines from maturity groups I, II, III, and IV, while the VP was made up of lines from only maturity groups II and III. Maturity group has been shown to affect seed composition, as lines with higher maturity group tend to have higher protein and lower maturity groups have higher oil content (Assefa et al., 2019). The difference in environments and maturity group may have created a disconnect between populations to cause low accuracy using GWAS SNPs from only one population. To look further into this, the effect of TP GWAS-selected regression coefficient SNPs and random SNPs was plotted on VP GWAS regression coefficient effects (Figures S3.24 and S3.25). The plotted GWAS SNPs do not share similar effect across populations, and for VP stability, random SNPs share similar effect to SNPs found to have a high effect on TP stability. Furthermore, from the GWAS analysis in TP and VP for protein and oil stability, no significant markers were identified across all four scenarios (Figures S3.17, S3.18, S3.21, S3.22). This inconsistency in marker effect leads us to believe that protein and oil stability has a complex genetic architecture as found in other studies for yield stability (Happ et al., 2021; Huang et al., 2016; Lozada & Carter, 2020), and further explains the ineffectiveness of GWAS SNPs for genomic prediction.

Conducting cross-validation within populations can provide an idea of the impact on prediction accuracy for stability when training and validation sets are closely related and grown in the same environments. With cross-validation, a similar mean r_{MP}, rank correlation, and rank coincidence between TP and VP for protein stability was observed (Tables 3.5 and 3.6). Cross-validation generally gave higher r_{MP}, but had a similar rank coincidence when compared to across-population prediction for protein stability.

Therefore, protein stability prediction did not benefit from greater relatedness between populations and stability values from the same environments. For oil stability, VP cross-validation displayed abnormally high r_{MP}, rank correlation, and rank coincidence for EV and RC measures. These values are closer to what would be expected for cross-validation accuracies for protein and oil stability. Rank coincidence can be a valuable tool for breeders in genomic selection for complex traits. With a rank coincidence of 0.66, 66% of individuals in the top half of predicted oil stability ranks are expected to be in the top half of observed oil stability ranks. This indicates that when advancing the top half of individuals based on predicted oil stability rank, two thirds of advanced lines are expected to truly be in the top half of stability. This level of accuracy is not far from being useful to breeders in an applied genomic selection setting, depending on the goals for selection. Aside from VP oil stability cross-validation, greater relatedness between training and validation sets and identical growing environments did not noticeably improve prediction of stability when compared to across-population prediction.

Multiple studies have shown the effectiveness of including genotype-by-environment interaction (GxE) in prediction modeling of complex traits (Burgueño et al., 2012; Jarquín, Crossa, et al., 2014; Ray et al., 2022). The more complex prediction model which accounts for environment (E) and GxE was tested to determine if it would result in higher prediction accuracy for stability (Table S3.1). The prediction models accounting for E and E + GxE did not increase prediction accuracy in any scenario for protein stability and oil stability across all stability measures. The standard additive GBLUP model performed equal to or better than these models in all cases, which suggests the stability values sufficiently account for GxE in each prediction scenario. In addition, the

small number of environments in this study likely contributed to the ineffectiveness of including E alone in the prediction model. If more environment combinations were present, E may help to further improve prediction accuracy.

The impact on prediction accuracy of conducting a separate GWAS for each stability measure during SNP selection was assessed (Table S3.2). Regression coefficient GWAS SNPs were used for all predictions thus far due to its correlation with Wricke's ecovalence (Figures S3.5 and S3.6). Separate GWAS SNPs led to higher genomic heritability for each stability measure in protein and oil stability. The prediction accuracies slightly increased for Wricke's ecovalence and decreased for EV stability, the measure with least correlation to regression coefficient, when predicting across populations. This leads to the conclusion that GWAS SNPs from only regression coefficient stability can be used with relatively equal effectiveness to separate GWAS SNPs for each stability measure. This provides a simpler pipeline for prediction of multiple stability measures at once.

The relationship between protein and oil content in soybean has previously been identified as a negative correlation (Shannon et al., 1972). However, the relationship between protein stability and oil stability had not been identified. In assessing the relationship between protein stability and oil stability for each stability measure, a low correlation was observed between the two traits with r^2 ranging from 0.004 to 0.27 (Figure 3.1). This low correlation indicates that selection for the most stable lines for protein content cannot be assumed to have high oil stability as well, and vice versa. This suggests that breeding programs will need to select for protein and oil stability independently.

In order to compare our data and methods to similar prediction studies, protein and oil BLUPs across all environments were predicted (Tables 3.2 and 3.3). A study from Stewart-Brown and colleagues predicted soybean protein and oil content BLUPs using multiple populations with similar prediction scenarios to this study (Stewart-Brown et al., 2019). Using four biparental populations, they tested within-population cross-validation and across-population prediction of protein and oil BLUPs in soybean. Predictive abilities of 0.33 and 0.51 were observed for TP and VP cross-validation of protein BLUPs. The VP r_{MP} of 0.51 was similar to the aforementioned study's average of 0.60. For across population protein BLUP prediction, an r_{MP} of 0.45 was obtained, compared to the 2019 study's across population predictive ability of 0.50. For cross-validation of oil content BLUPs, predictive abilities of 0.22 and 0.50 were obtained for TP and VP, respectively. Across-population predictive ability for oil content BLUPs was 0.21 in our study. The previously mentioned study averaged 0.52 for predictive ability of oil BLUP cross-validation, and 0.27 for across population prediction of BLUPs. Overall, our VP cross-validation and across-population results were similar to Stewart-Brown and colleagues, indicating the validity of the data and prediction methods used. We suspect the TP cross-validation was less effective due to the substantially lower heritability observed for protein and oil according to variance components (Table 3.1). We also expect our cross-validation accuracies to be lower because the 2019 study used biparental populations with greater population structure and relatedness between individuals than our breeding populations. Overall, our predictive abilities were as expected when compared to a study with similar populations and prediction scenarios.

A limitation to this study, and all other stability studies, is the lack of phenotype data from a greater number of environments. While sufficient environments for proper stability analysis were obtained, a greater number of environments would have improved stability estimation and possibly prediction accuracy. Regardless, all stability measures have a level of uncertainty across different sets of environments. While less prevalent in static stability measures, the issue of repeatability is one of the greatest limitations to stability analysis (Pham & Kang, 1988; Sneller & Dombek, 1997). Between any two sets of environments, there will undoubtedly be some deviation in environmental conditions, and the stability of a line will not be equal when tested in two different geographical regions. In terms of genomic prediction for stability in a specific breeding program, the best that can be done is training the prediction model with stability values from environments within the program's geographic range of improvement. Despite these limitations, there is still value in the ability to accurately predict stability for protein and oil, thereby increasing selection accuracy for lines with the greatest consistency of seed composition across environments.

Conclusion

In this study a general set of basic practices has been outlined for genomic selection of seed composition stability across populations from the same breeding program. The most useful application of genomic prediction for stability traits would be within the first or second year of field trials to improve selection accuracy for stability at a stage when phenotype data are insufficient for proper stability analysis. Promising results were obtained for oil stability prediction through both applied prediction and cross-validation, and EV stability was generally most predictable across components.

Rank coincidence improved the applicability of predicting seed composition across populations, where selection accuracy values better than random were obtained when r_{MP} was very low and not useful. Translating these results to effective across population prediction of seed composition stability will require further work. An assessment of non-parametric and semi-parametric prediction models, along with the exploration of other SNP types such as tagSNPs, could potentially increase prediction accuracy across populations. An assessment of the genetic control of seed composition stability and its consistency across populations grown in different environments would also be helpful for understanding the future potential and application of genomic prediction for seed composition stability.

References

Abdelghany, A. M., Zhang, S., Azam, M., Shaibu, A. S., Feng, Y., Qi, J., Li, J., Li, Y., Tian, Y., Hong, H., Lamlom, S. F., Li, B., & Sun, J. (2021). Exploring the phenotypic stability of soybean seed compositions using multi-trait stability index approach. *Agronomy*, *11*(11), 1–13. https://doi.org/10.3390/agronomy11112200

Assefa, Y., Purcell, L. C., Salmeron, M., Naeve, S., Casteel, S. N., Kovács, P., Archontoulis, S., Licht, M., Below, F., Kandel, H., Lindsey, L. E., Gaska, J., Conley, S., Shapiro, C., Orlowski, J. M., Golden, B. R., Kaur, G., Singh, M., Thelen, K., … Ciampitti, I. A. (2019). Assessing variation in us soybean seed composition (protein and oil). *Frontiers in Plant Science*, *10*(March). https://doi.org/10.3389/fpls.2019.00298

Balešević-Tubić, S., Dordević, V., Miladinović, J., Dukić, V., & Tatić, M. (2011). Stability of soybean seed composition. *Genetika*, *43*(2), 217–227.

https://doi.org/10.2298/GENSR1102217B

Becker, H. C., & Léon, J. (1988). Stability Analysis in Plant Breeding. *Plant Breeding*, *101*(1), 1–23. https://doi.org/10.1111/j.1439-0523.1988.tb00261.x

Bernardo, R. (2008). Molecular markers and selection for complex traits in plants: Learning from the last 20 years. *Crop Science*, *48*(5), 1649–1664. https://doi.org/10.2135/cropsci2008.03.0131

Bernardo, R., & Yu, J. (2007). Prospects for genomewide selection for quantitative traits in maize. *Crop Science*, *47*(3), 1082–1090. https://doi.org/10.2135/cropsci2006.11.0690

Best, D. J., & Roberts, D. E. (1975). The Upper Tail Probabilities of Spearman ' s Rho. *Journal of the Royal Statistical Society*, *24*(3), 377–379.

Brøndum, R. F., Su, G., Janss, L., Sahana, G., Guldbrandtsen, B., Boichard, D., & Lund, M. S. (2015). Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction. *Journal of Dairy Science*, *98*(6), 4107–4116. https://doi.org/10.3168/jds.2014-9005

Browning, B. L., Zhou, Y., & Browning, S. R. (2018). A One-Penny Imputed Genome from Next-Generation Reference Panels. *American Journal of Human Genetics*, *103*(3), 338–348. https://doi.org/10.1016/j.ajhg.2018.07.015

Burgueño, J., de los Campos, G., Weigel, K., & Crossa, J. (2012). Genomic prediction of breeding values when modeling genotype × environment interaction using pedigree and dense molecular markers. *Crop Science*, *52*(2), 707–719. https://doi.org/10.2135/cropsci2011.06.0299

Butler, D. G., Cullis, B. R., Gilmour, A. R., Gogel, B. J., & Thompson, R. (2018).

ASReml-R Reference Manual Version 4.

http://www.homepages.ed.ac.uk/iwhite/asreml/uop.

Eberhart, S. A., & Russell, W. A. (1966). Stability Parameters for Comparing Varieties. *Crop Science*, *6*(1), 36–40.

Edwards, S. M. K., Buntjer, J. B., Jackson, R., Bentley, A. R., Lage, J., Byrne, E., Burt, C., Jack, P., Berry, S., Flatman, E., Poupard, B., Smith, S., Hayes, C., Gaynor, R. C., Gorjanc, G., Howell, P., Ober, E., Mackay, I. J., & Hickey, J. M. (2019). The effects of training population design on genomic prediction accuracy in wheat. *Theoretical and Applied Genetics*, *132*(7), 1943–1952. https://doi.org/10.1007/s00122-019-03327-y

Finlay, K. W., & Wilkinson, G. N. (1963). The analysis of adaptation in a plant-breeding programme. *Australian Journal of Agricultural Research*, *14*(6), 742–754. https://doi.org/10.1071/AR9630742

Happ, M. M., Graef, G. L., Wang, H., Howard, R., Posadas, L., & Hyten, D. L. (2021). Comparing a Mixed Model Approach to Traditional Stability Estimators for Mapping Genotype by Environment Interactions and Yield Stability in Soybean [Glycine max (L.) Merr.]. *Frontiers in Plant Science*, *12*(March), 1–12. https://doi.org/10.3389/fpls.2021.630175

Happ, M. M., Wang, H., Graef, G. L., & Hyten, D. L. (2019). Generating high density, low cost genotype data in Soybean [Glycine max (L.) Merr.]. *G3: Genes, Genomes, Genetics*, *9*(7), 2153–2160. https://doi.org/10.1534/g3.119.400093

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

https://doi.org/10.2135/cropsci2009.11.0662

Henderson, C. R. (1984). *Applications of Linear Models in Animal Breeding*. University of Guelph, Guelph, Ontario, Canada.

Huang, M., Cabrera, A., Hoffstetter, A., Griffey, C., Van Sanford, D., Costa, J., McKendry, A., Chao, S., & Sneller, C. (2016). Genomic selection for wheat traits and trait stability. *Theoretical and Applied Genetics*, *129*(9), 1697–1710. https://doi.org/10.1007/s00122-016-2733-z

Jarquín, D., Crossa, J., Lacaze, X., Du Cheyron, P., Daucourt, J., Lorgeou, J., Piraux, F., Guerreiro, L., Pérez, P., Calus, M., Burgueño, J., & de los Campos, G. (2014). A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theoretical and Applied Genetics*, *127*(3), 595–607. https://doi.org/10.1007/s00122-013-2243-1

Jarquín, D., Kocak, K., Posadas, L., Hyma, K., Jedlicka, J., Graef, G., & Lorenz, A. (2014). Genotyping by sequencing for genomic prediction in a soybean breeding population. *BMC Genomics*, *15*(1), 1–10. https://doi.org/10.1186/1471-2164-15-740

Keim, P. (1988). A rapid protocol for isolating soybean DNA. *Soybean Genetics Newsletter*, *15*, 150–152.

Lin, C. S., & Binns, M. R. (1991). Genetic properties of four types of stability parameter. *Theoretical and Applied Genetics*, *82*(4), 505–509.

Lin, C. S., Binns, M. R., & Lefkovitch, L. P. (1986). Stability Analysis: Where Do We Stand? *Crop Science*, *26*(5), 894–900.

Lozada, D. N., & Carter, A. H. (2020). Insights into the genetic architecture of phenotypic stability traits in winter wheat. *Agronomy*, *10*(3).

https://doi.org/10.3390/agronomy10030368

Luo, Z., Yu, Y., Xiang, J., & Li, F. (2021). Genomic selection using a subset of SNPs identified by genome-wide association analysis for disease resistance traits in aquaculture species. *Aquaculture*, *539*(December 2020), 736620. https://doi.org/10.1016/j.aquaculture.2021.736620

Meher, P. K., Rustgi, S., & Kumar, A. (2022). Performance of Bayesian and BLUP alphabets for genomic prediction: analysis, comparison and results. *Heredity*, *128*(6), 519–530. https://doi.org/10.1038/s41437-022-00539-9

Meuwissen, T., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*(4), 1819–1829.

Meuwissen, Theo. (2007). Genomic selection : Marker assisted selection on a genome wide scale. *Journal of Animal Breeding and Genetics*, *124*(6), 321–322. https://doi.org/10.1111/j.1439-0388.2007.00708.x

Olivoto, T., Lúcio, A. D. C., da Silva, J. A. G., Marchioro, V. S., de Souza, V. Q., & Jost, E. (2019). Mean performance and stability in multi-environment trials i: Combining features of AMMI and BLUP techniques. *Agronomy Journal*, *111*(6), 2949–2960. https://doi.org/10.2134/agronj2019.03.0220

Olivoto, T., Lúcio, A. D. C., da Silva, J. A. G., Sari, B. G., & Diel, M. I. (2019). Mean performance and stability in multi-environment trials II: Selection based on multiple traits. *Agronomy Journal*, *111*(6), 2961–2969. https://doi.org/10.2134/agronj2019.03.0221

Pham, H. N., & Kang, M. S. (1988). Interrelationships among and Repeatability of Several Stability Statistics Estimated from International Maize Trials. *Crop Science*,

28(6), 925–928. https://doi.org/10.2135/cropsci1988.0011183x002800060010x

Poplin, R., Ruano-Rubio, Valentin DePristo, M. A., Fennell, T. J., O Carneiro, M., Van der Auwera, G. A., Kling, D. E., Gauthier, L. D., Levy-Moonshine, A., Roazen, D., Shakir, K., Joel, T., Sheila, C., Chris, W., Monkol, L., Stacey, G., Mark, J. D., Ben, N., Daniel, G. M., & Eric, B. (2018). Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxIV*, 1–22.

Pour-Aboughadareh, A., Khalili, M., Poczai, P., & Olivoto, T. (2022). Stability Indices to Deciphering the Genotype-by-Environment Interaction (GEI) Effect: An Applicable Review for Use in Plant Breeding Programs. *Plants*, *11*(3). https://doi.org/10.3390/plants11030414

Pour-Aboughadareh, A., Yousefian, M., Moradkhani, H., Poczai, P., & Siddique, K. H. M. (2019). STABILITYSOFT: A new online program to calculate parametric and non-parametric stability statistics for crop traits. *Applications in Plant Sciences*, *7*(1), 1–6. https://doi.org/10.1002/aps3.1211

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*(8), 904–909. https://doi.org/10.1038/ng1847

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., De Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

Ray, S., Jarquin, D., & Howard, R. (2022). Comparing artificial-intelligence techniques with state-of-the-art parametric prediction models for predicting soybean traits. *Plant Genome*, *August 2021*, 1–19. https://doi.org/10.1002/tpg2.20263

Roemer, J. (1917). Sinde die ertagdreichen Sorten ertagissicherer. *DLG-Mitt*, *32*(1), 87–89.

Shannon, G., Wilcox, J., & Probst, A. (1972). Estimated Gains from Selection for Protein and Yield in the F4 Generation of Six Soybean Populations. *Crop Science*, *12*(6), 824–826.

Sneller, C. H., & Dombek, D. (1997). Repeatability of Yield Stability Statistics in Soybean. *Crop Science*, *34*, 383–390.

Stewart-Brown, B. B., Song, Q., Vaughn, J. N., & Li, Z. (2019). Genomic selection for yield and seed composition traits within an applied soybean breeding program. *G3: Genes, Genomes, Genetics*, *9*(7), 2253–2265. https://doi.org/10.1534/g3.118.200917

Team, R. C. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.r-project.org/

van den Berg, I., Boichard, D., Guldbrandtsen, B., & Lund, M. S. (2016). Using sequence variants in linkage disequilibrium with causative mutations to improve across-breed prediction in dairy cattle: A simulation study. *G3: Genes, Genomes, Genetics*, *6*(8), 2553–2561. https://doi.org/10.1534/g3.116.027730

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, *91*(11), 4414–4423. https://doi.org/10.3168/jds.2007-0980

Werner, C. R., Gaynor, R. C., Gorjanc, G., Hickey, J. M., Kox, T., Abbadi, A., Leckband, G., Snowdon, R. J., & Stahl, A. (2020). How Population Structure Impacts Genomic

Selection Accuracy in Cross-Validation: Implications for Practical Breeding. *Frontiers in Plant Science*, *11*(December), 1–14. https://doi.org/10.3389/fpls.2020.592977

Whaley, R., & Eskandari, M. (2019). Genotypic main effect and genotype-by-environment interaction effect on seed protein concentration and yield in food-grade soybeans (Glycine max (L.) Merrill). *Euphytica*, *215*(2), 1–17. https://doi.org/10.1007/s10681-019-2344-2

Wijewardana, C., Reddy, K. R., & Bellaloui, N. (2019). Soybean seed physiology, quality, and chemical composition under soil moisture stress. *Food Chemistry*, *278*(November 2018), 92–100. https://doi.org/10.1016/j.foodchem.2018.11.035

Wricke, G. (1962). Evaluation Method for Recording Ecological Differences in Field Trials. *Z Pflanzenzücht*, *47*, 92–96.

Yin, L., Zhang, H., Tang, Z., Xu, J., Yin, D., Zhang, Z., Yuan, X., Zhu, M., Zhao, S., Li, X., & Liu, X. (2021). rMVP: A Memory-efficient, Visualization-enhanced, and Parallel-accelerated Tool for Genome-wide Association Study. *Genomics, Proteomics and Bioinformatics*, *19*(4), 619–628. https://doi.org/10.1016/j.gpb.2020.10.007

Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., Kresovich, S., & Buckler, E. S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, *38*(2), 203–208. https://doi.org/10.1038/ng1702

Zhang, H., Yin, L., Wang, M., Yuan, X., & Liu, X. (2019). Factors affecting the accuracy

of genomic selection for agricultural economic traits in maize, cattle, and pig

populations. *Frontiers in Genetics*, *10*(MAR), 1–10.

https://doi.org/10.3389/fgene.2019.00189

Supplementary Figures

Figures S2.1 – S2.22
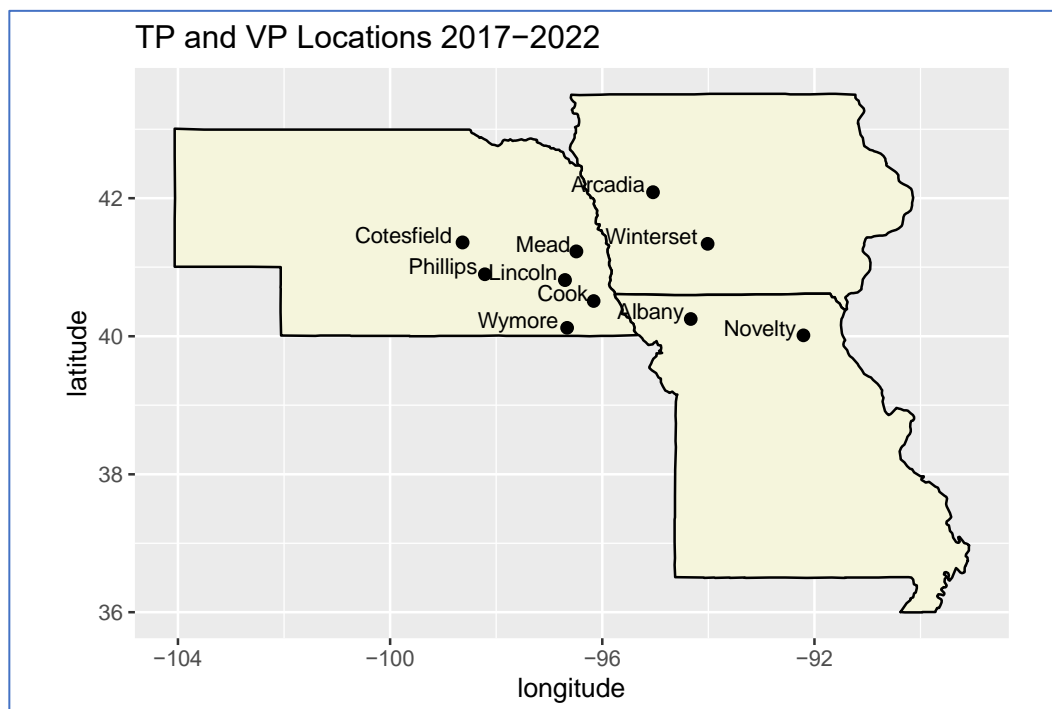
Figure S2.1. Map of geographic locations for TP and VP yield tests.
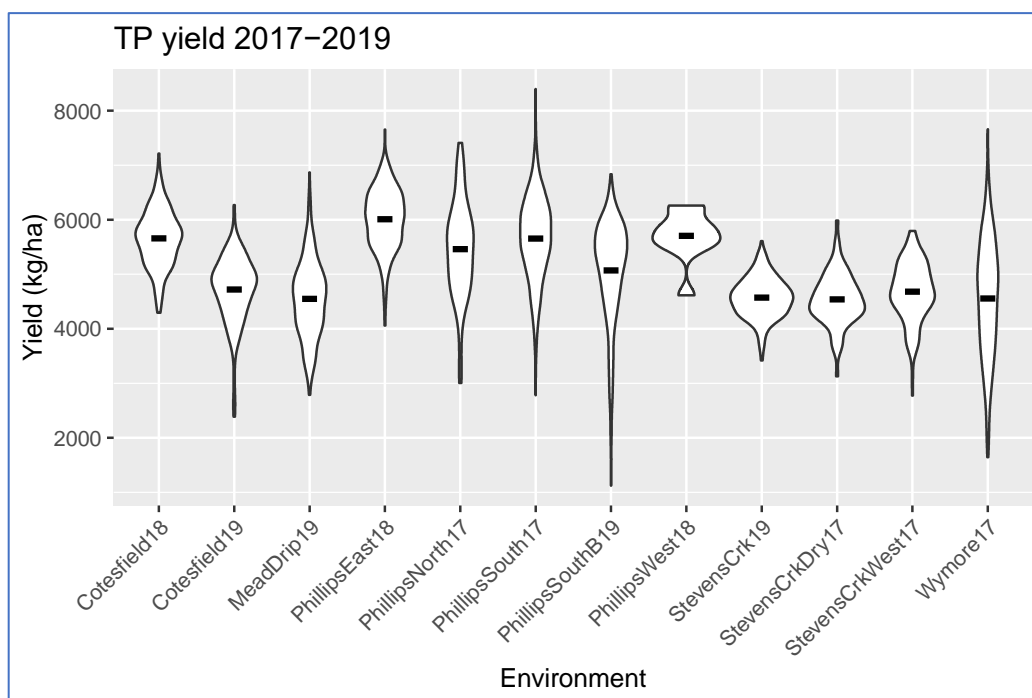
Figure S2.2. Training population yield distributions for all environments from 2017 to 2019. All environments were in Nebraska.
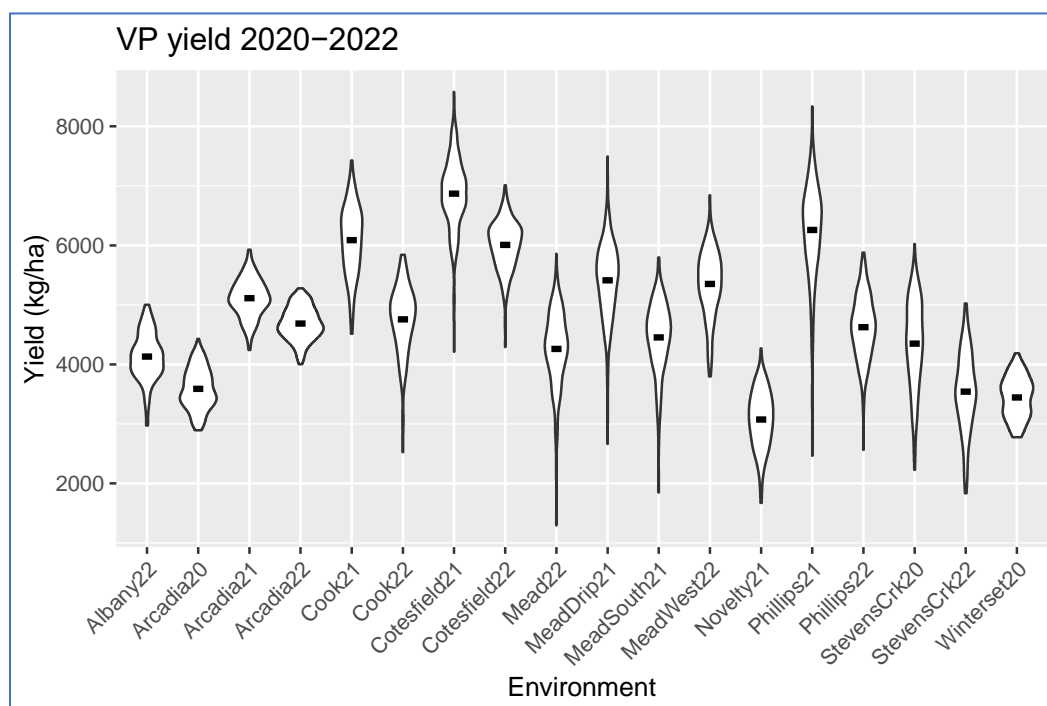
Figure S2.3. Validation population yield distributions for all environments from 2020-2022. All environments were in Nebraska, aside from Arcadia and Winterset, IA, and Albany and Novelty, MO.
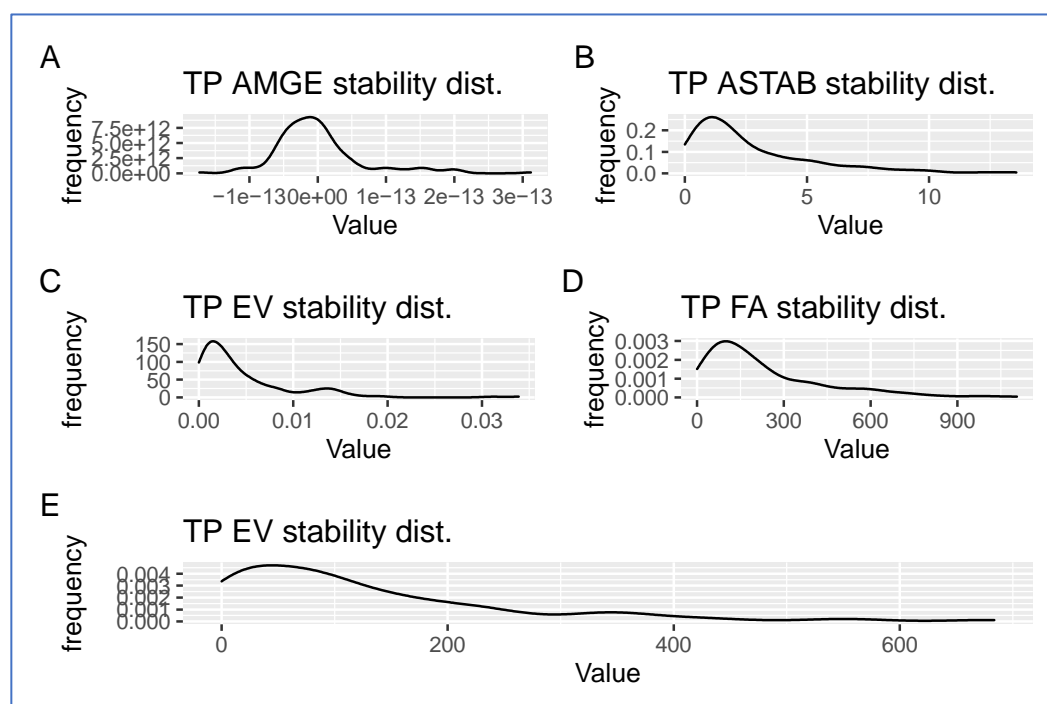
Figure S2.4. Distribution of training population AMMI stability values.
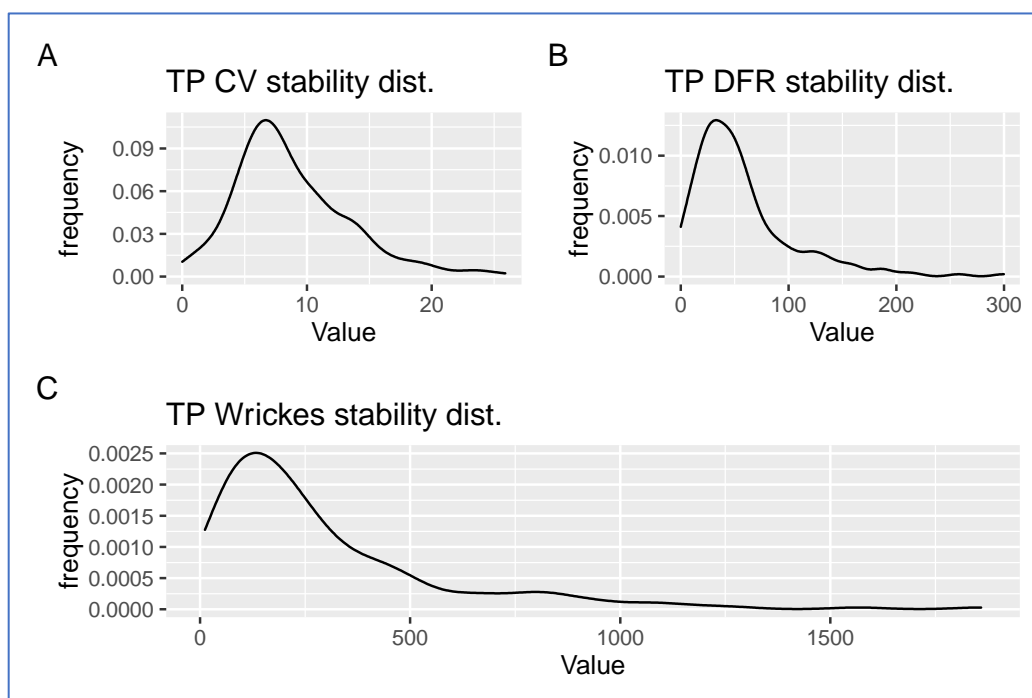
Figure S2.5. Distribution of training population conventional univariate stability values.
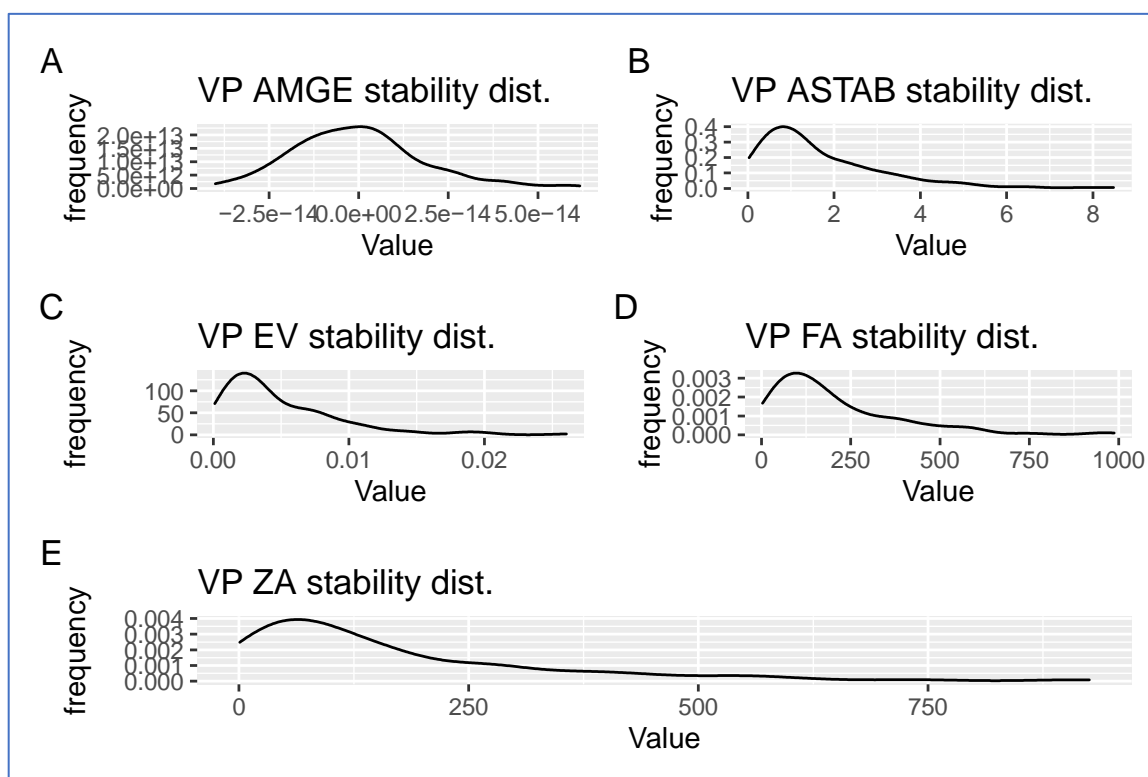
Figure S2.6. Distribution of validation population AMMI stability values.
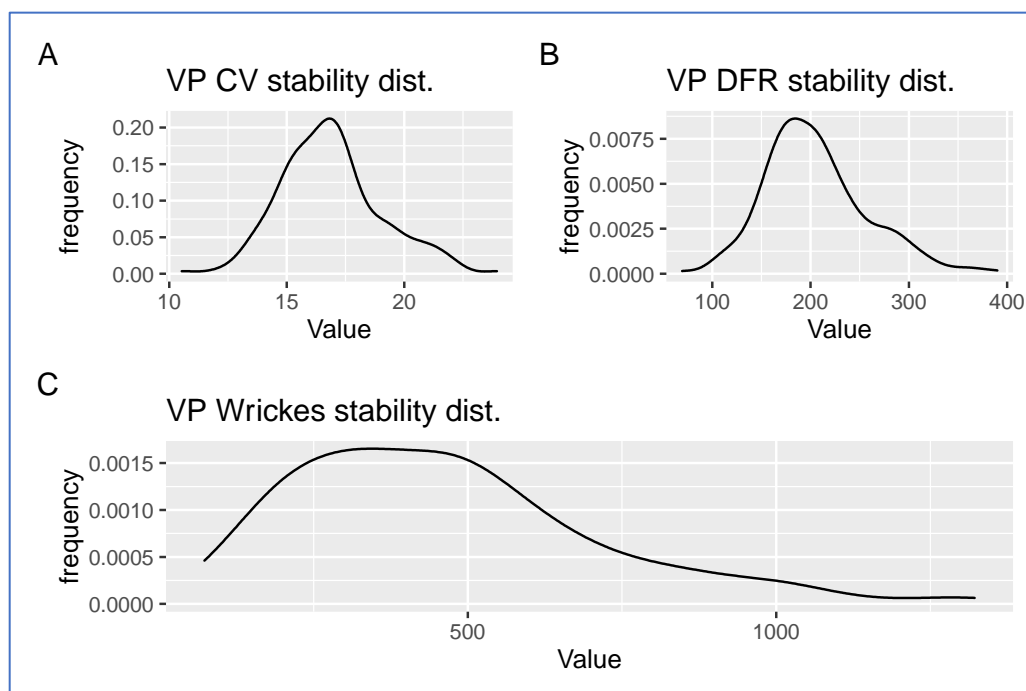
Figure S2.7. Distribution of validation population conventional univariate stability values.
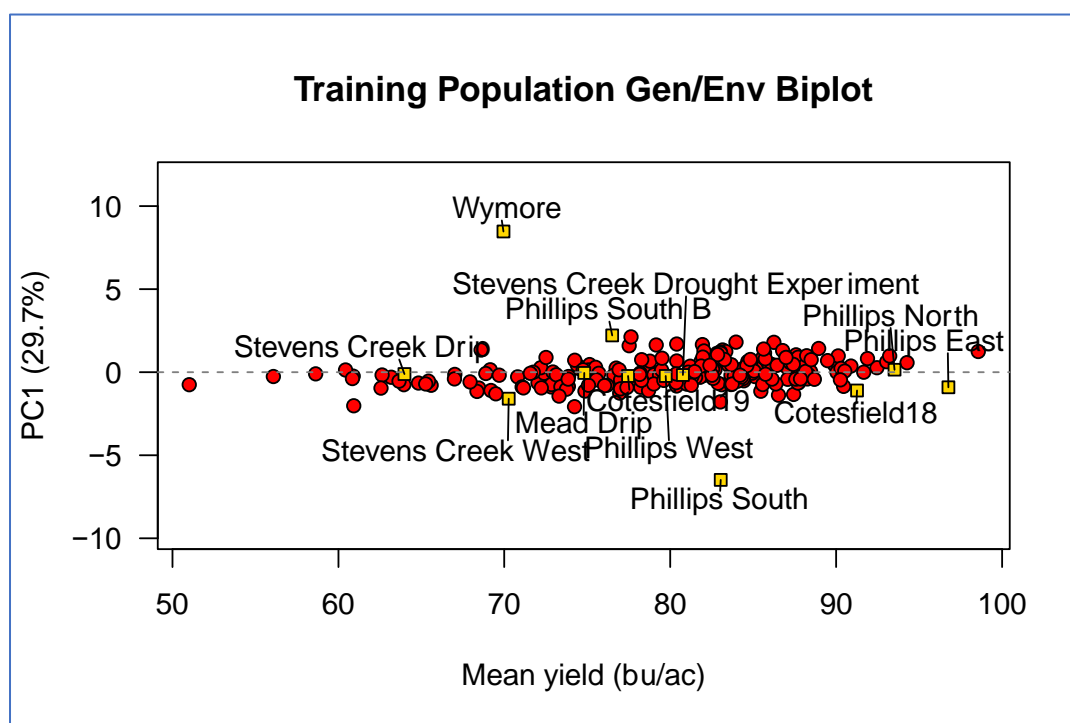
Figure S2.8. Biplot of mean yield vs. AMMI PC1 (29.7%) for training population genotypes and environments. Genotypes are in red, environments labelled and in yellow.
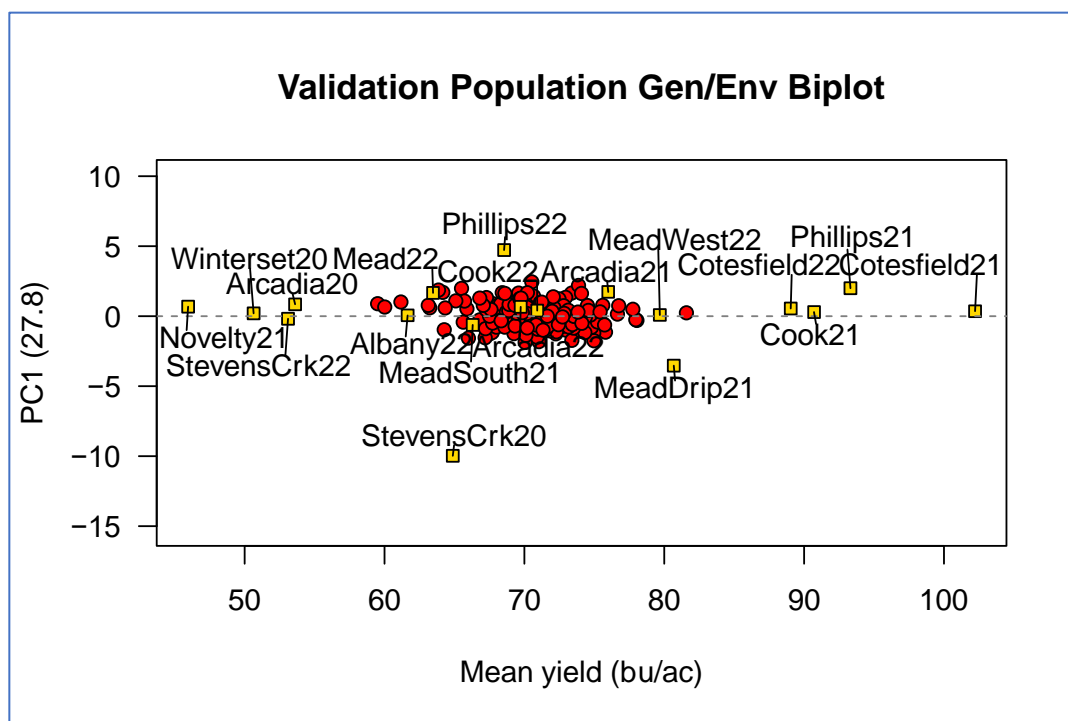
Figure S2.9. Biplot of mean yield vs. AMMI PC1 (27.8%) for validation population genotypes and environments. Genotypes are in red, environments labelled and in yellow.
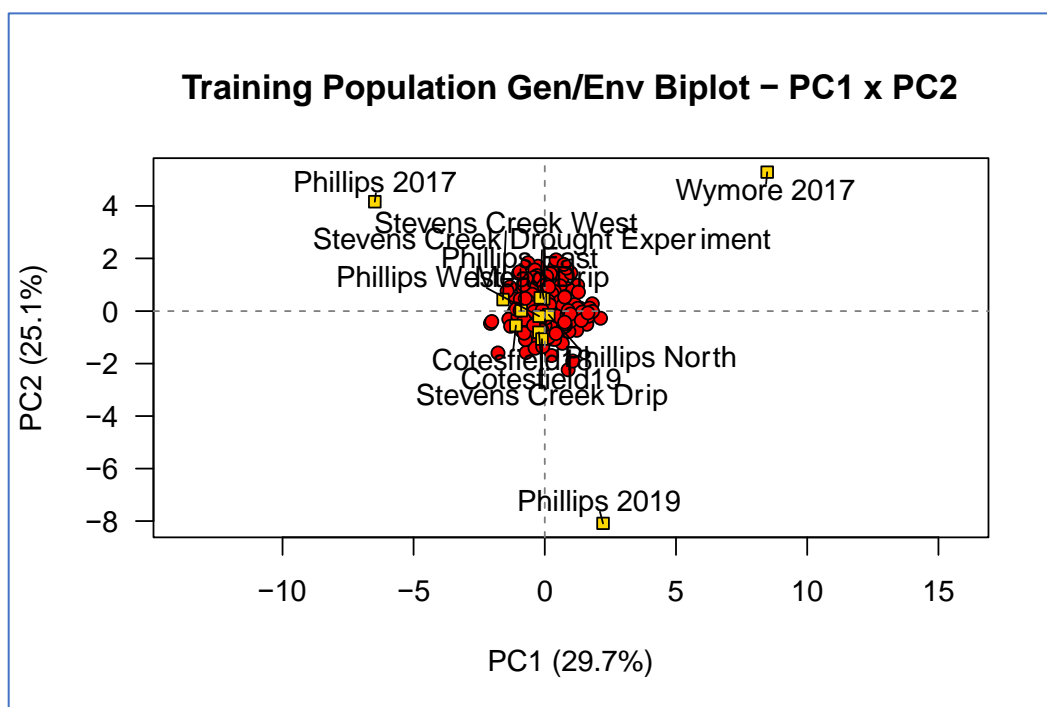
Figure S2.10. Biplot of AMMI PC1 (29.7%) vs PC2 (25.1%) for training population genotypes and environments. Genotypes are in red, environments labelled and in yellow.
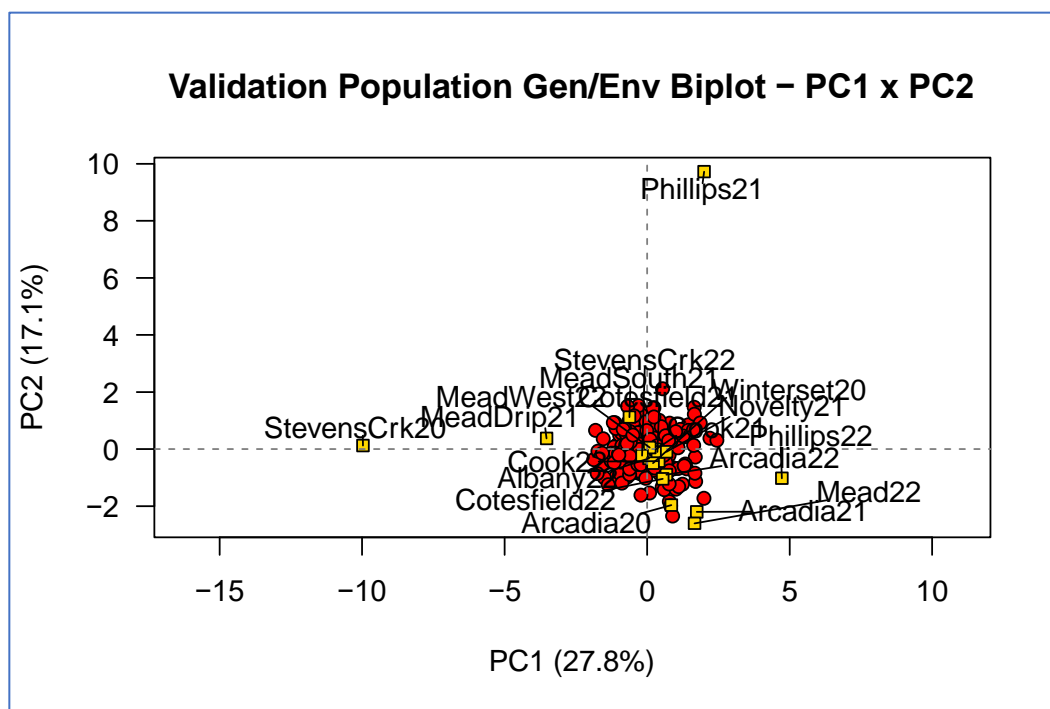
Figure S2.11. Biplot of AMMI PC1 (27.8%) vs PC2 (17.1%) for training population genotypes and environments. Genotypes are in red, environments labelled and in yellow.
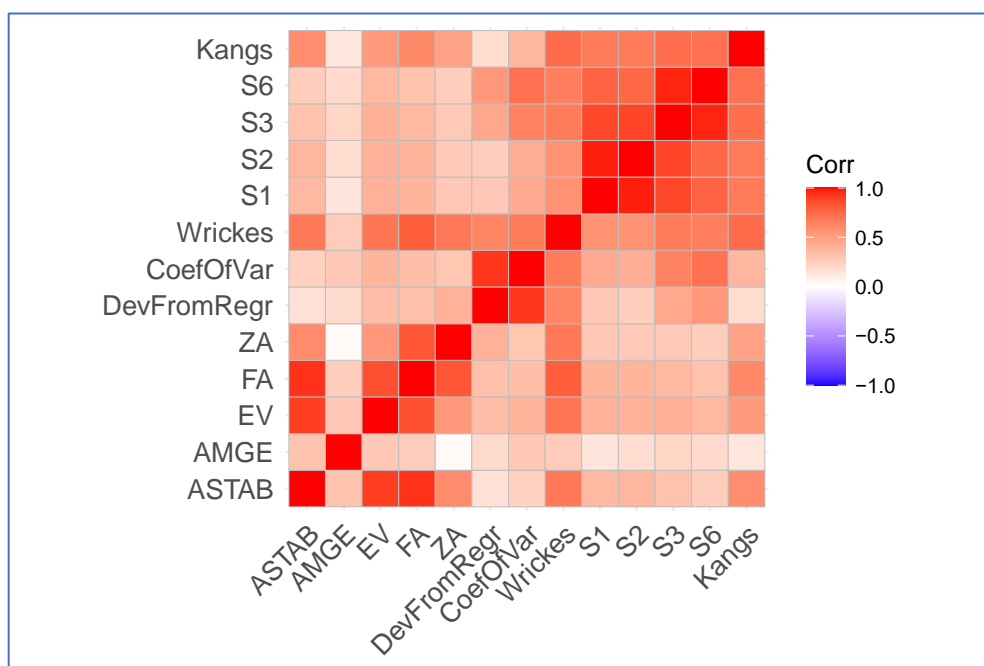
Figure S2.12. Correlation between five AMMI stability measures, three conventional univariate dynamic stability measures, and five non-parametric dynamic stability measures.
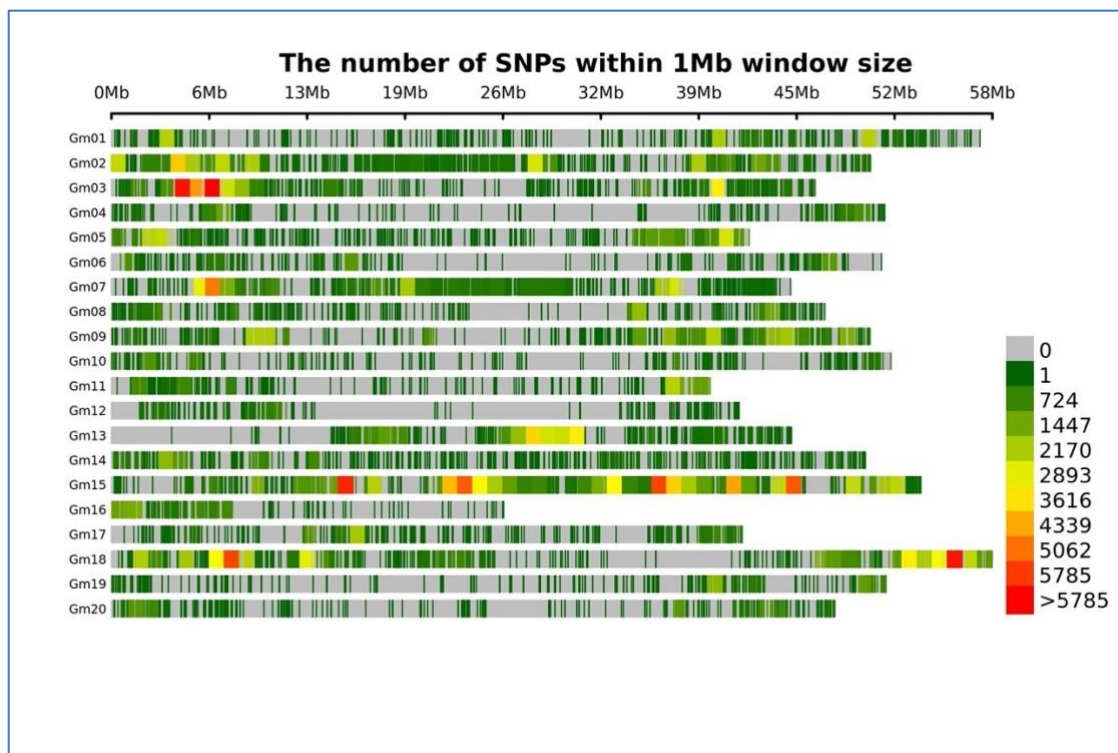
Figure S2.13. Distribution of 373,348 SNP markers used for GWAS and random SNP selection, visualized across all 20 soybean chromosomes (rMVP).
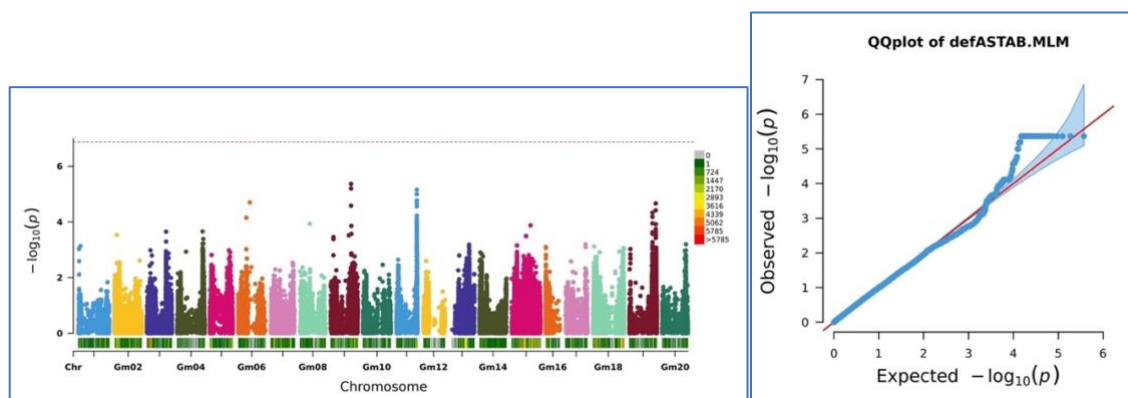
Figure S2.14. Manhattan plot (left) for TP GWAS of AMMI-based stability parameter (ASTAB.AMMI). QQ plot (right) for TP GWAS of AMMI-based stability parameter (ASTAB.AMMI). The most significant marker from each 10,000 base pair window was used for genomic prediction across populations and VP cross-validation of conventional stability measures (rMVP).
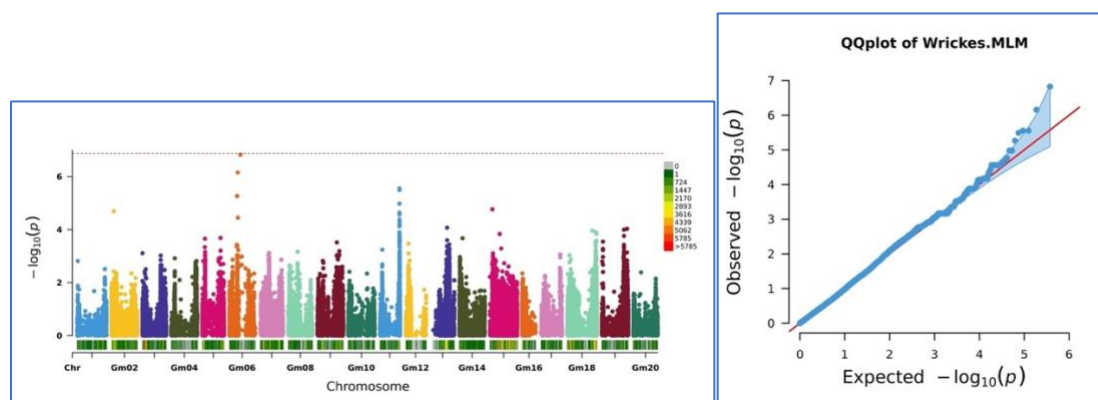
Figure S2.15. Manhattan plot for TP GWAS of Wricke's ecovalence univariate dynamic stability parameter (left), and QQ-plot for TP GWAS of Wricke's ecovalence stability (right). The most significant marker from each 10,000 base pair window was used for genomic prediction across populations and VP cross-validation of conventional stability measures (rMVP).
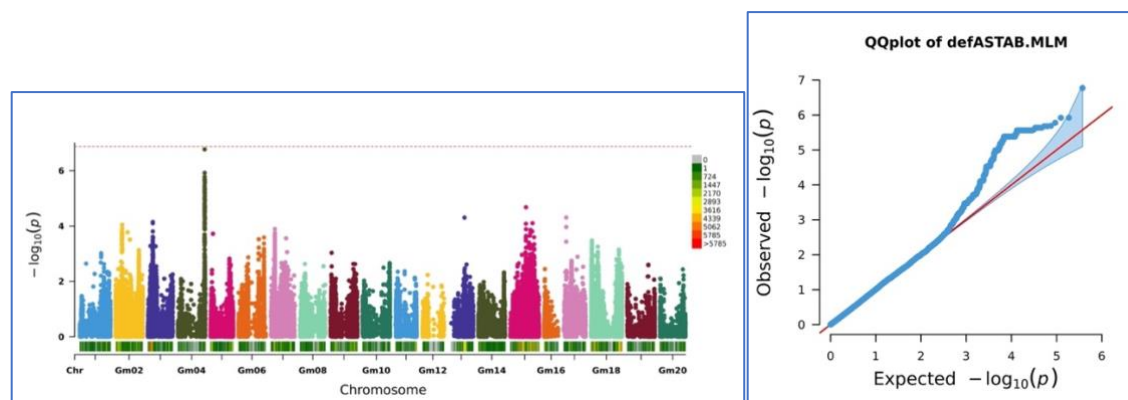
Figure S2.16. Manhattan plot (left) for VP GWAS of AMMI-based stability parameter (ASTAB.AMMI). QQ plot (right) for VP GWAS of AMMI-based stability parameter (ASTAB.AMMI). The most significant marker from each 10,000 base pair window was used for TP cross-validation of AMMI stability measures (rMVP).
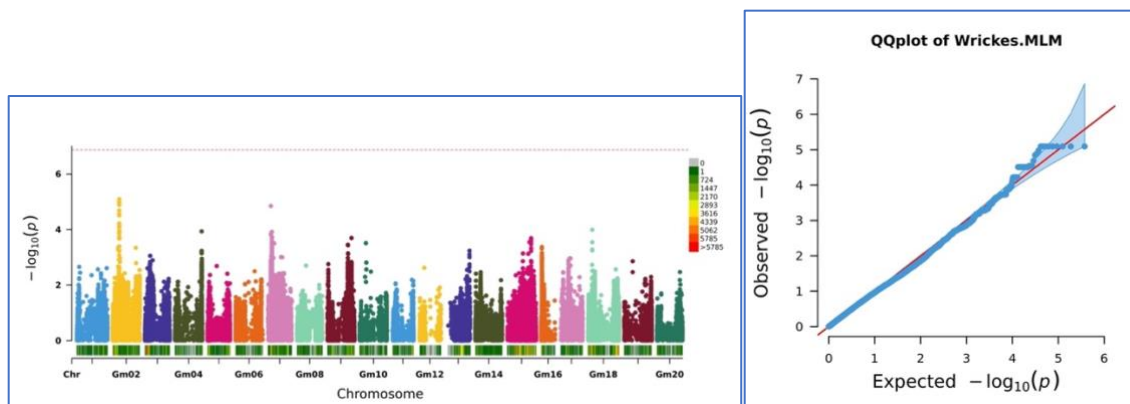
Figure S2.17. Manhattan plot for VP GWAS of Wricke's ecovalence univariate dynamic stability parameter (left), and QQ-plot for VP GWAS of Wricke's ecovalence stability (right). The most significant marker from each 10,000 base pair window was used for TP cross-validation of conventional stability measures (rMVP).
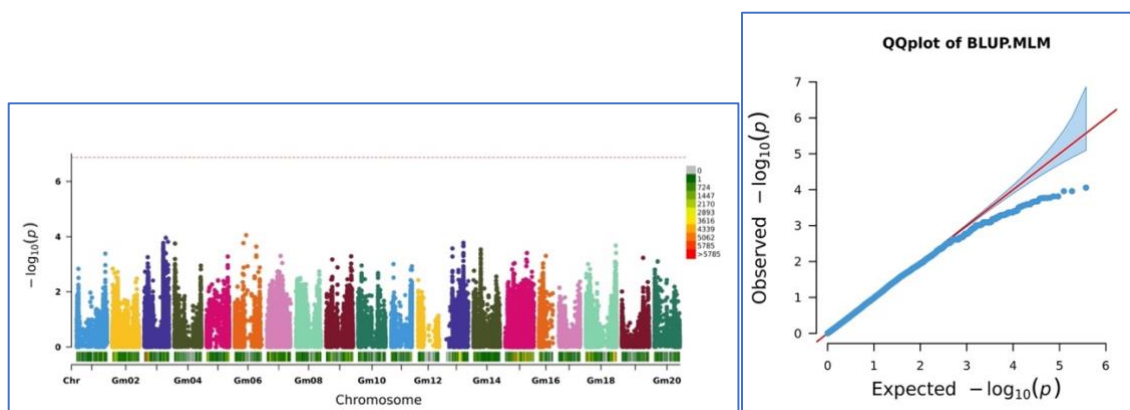
Figure S2.18. Manhattan plot (left) for TP GWAS of yield BLUPs across all environments (left). QQ plot (right) for TP GWAS of yield BLUPs. The most significant marker from each 10,000 base pair window was used for genomic prediction of yield for across-population prediction and VP cross-validation (rMVP).
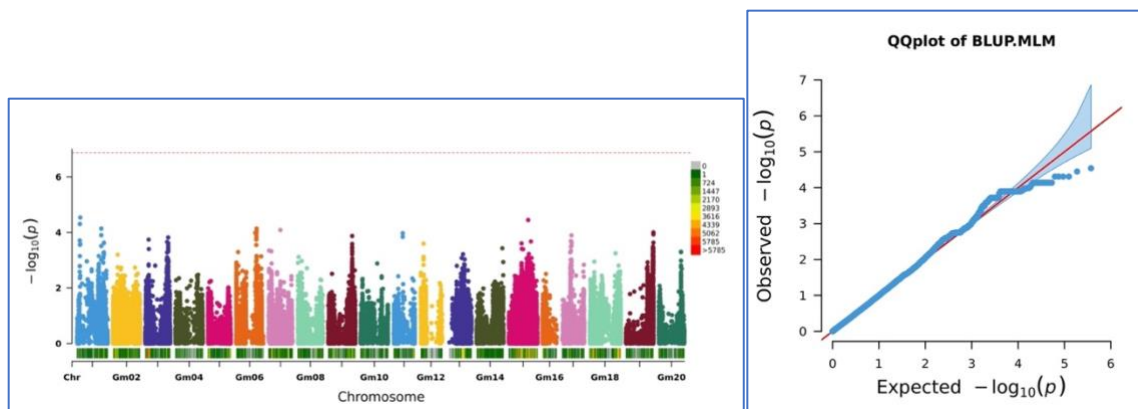
Figure S2.19. Manhattan plot for VP GWAS of yield BLUPs across all environments (left), and QQ-plot for TP GWAS of yield BLUPs (right). The most significant marker from each 10,000 base pair window was used for genomic prediction of yield in TP cross-validation (rMVP).
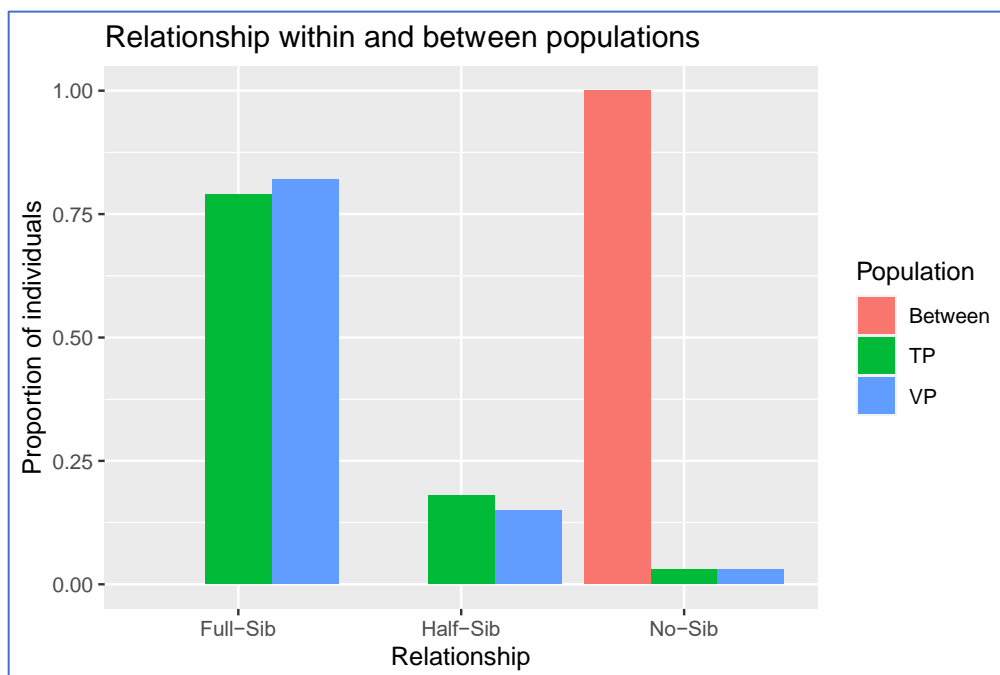
Figure S2.20. Proportion of individuals with at least one full sibling (shares both parents), at least one half-sibling (shares one parent), or no sibling within TP individuals, within VP individuals, and between TP and VP individuals. There are no shared parents between TP and VP individuals.
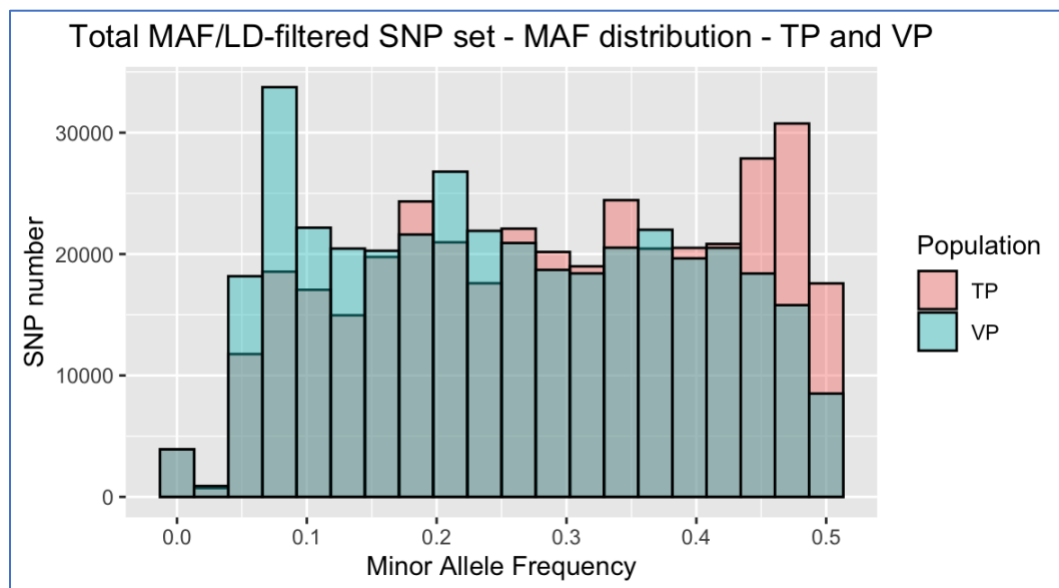
Figure S2.21. Minor allele frequency distribution across set of 373,348 SNP positions from MAF and LD-filtered WGS/imputation dataset.
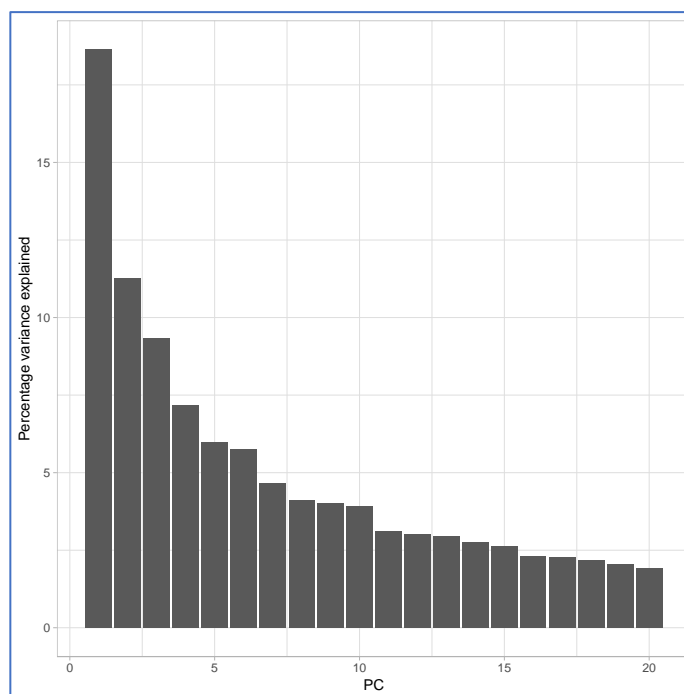
Figure S2.22. Principal components and variance explained for 190 TP and 165 VP individuals combined (goes main text Figure 3).

Figures S3.1 – S3.25
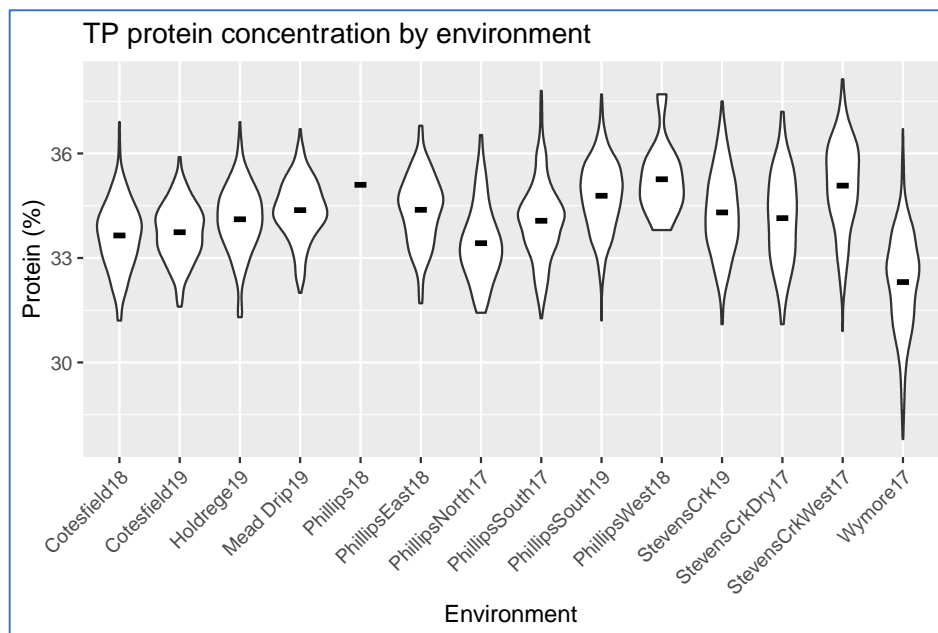
Figure S3.1. Training population protein distributions for all environments from 2017 to 2019. All environments were in Nebraska.
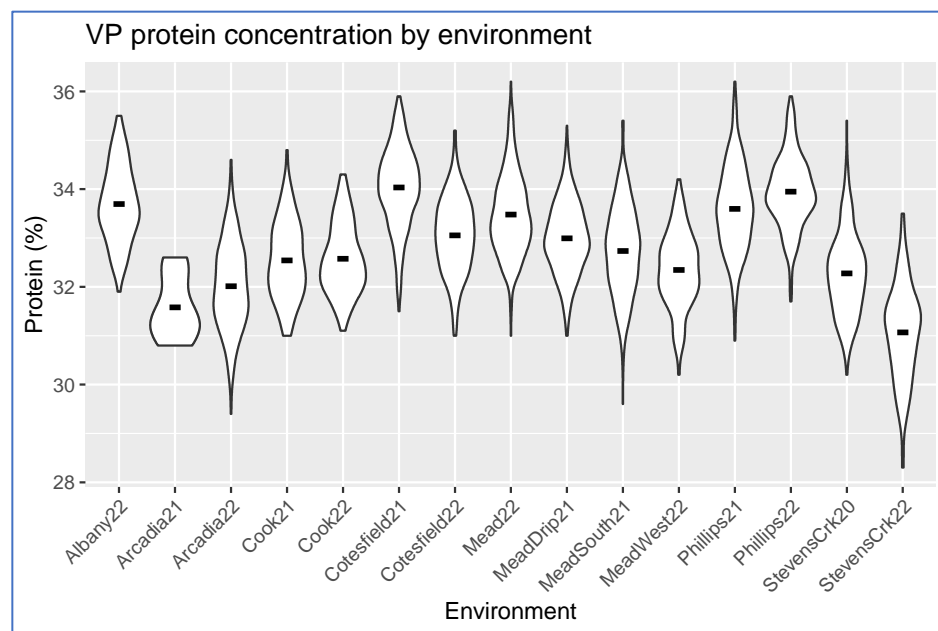
Figure S3.2. Validation population protein distributions for all environments from 2020 to 2022. Environments were in Nebraska, Missouri, and Iowa.
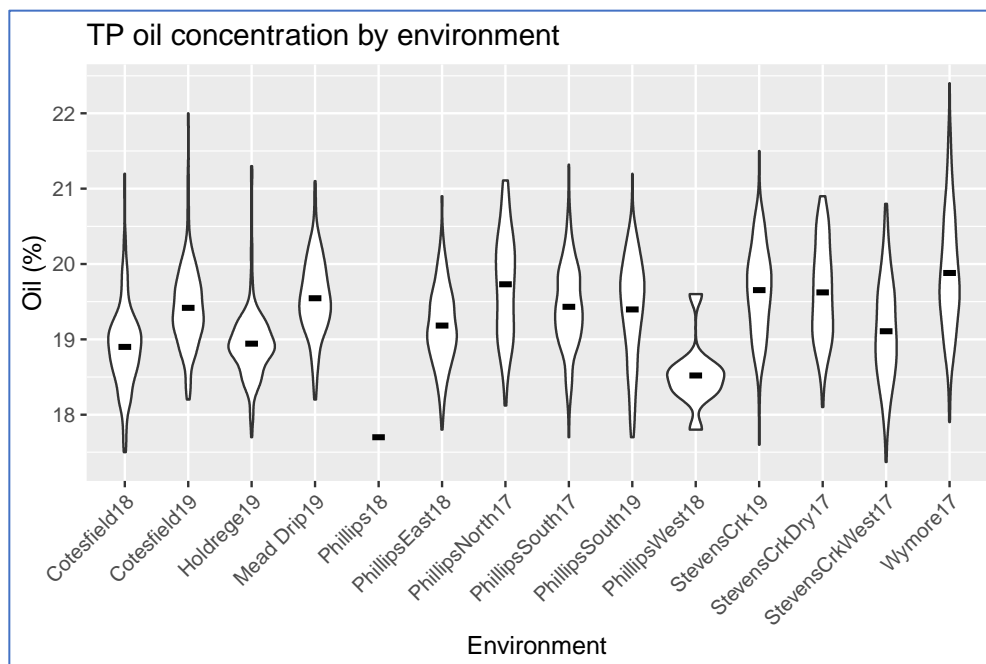
Figure S3.3. Training population oil distributions for all environments from 2017 to 2019. All environments were in Nebraska.
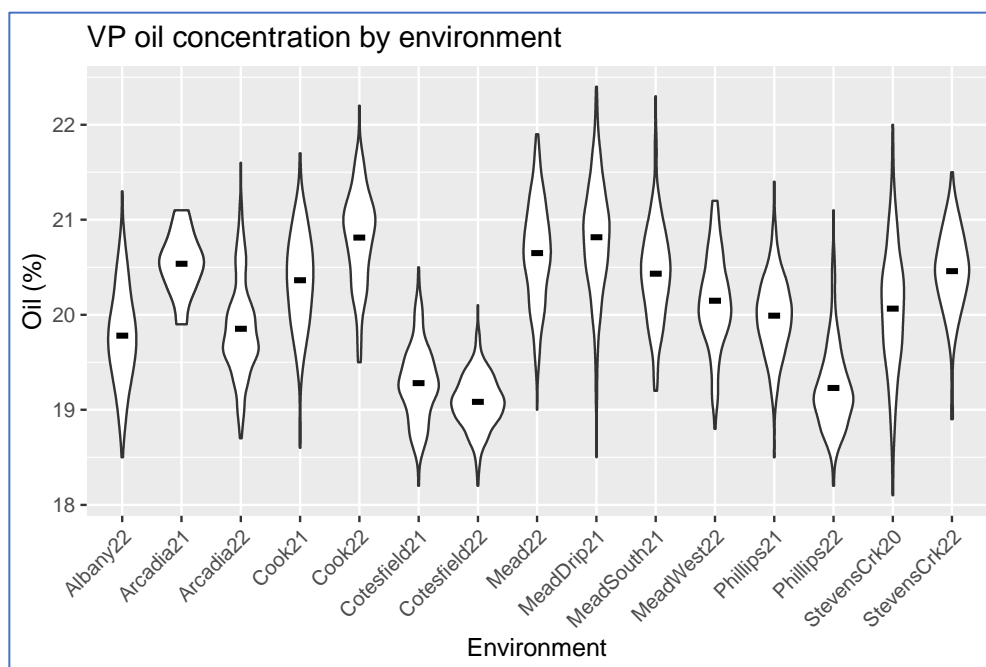
Figure S3.4. Validation population oil distributions for all environments from 2020 to 2022. Environments were in Nebraska, Missouri, and Iowa.
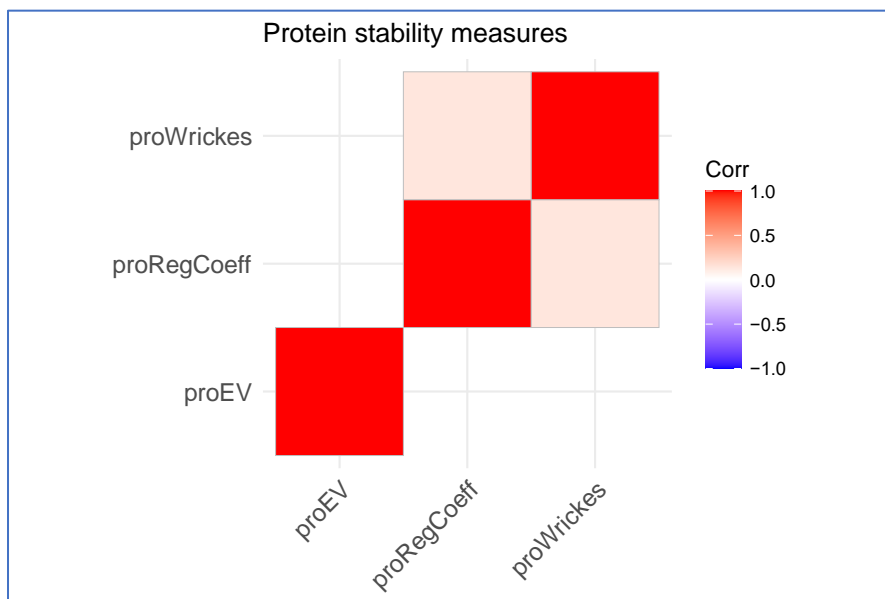
Figure S3.5. Correlation between protein stability measures: Environmental Variance (EV), Regression Coefficient (b$_i$; RegCoeff), Wricke's ecovalence (Wrickes).
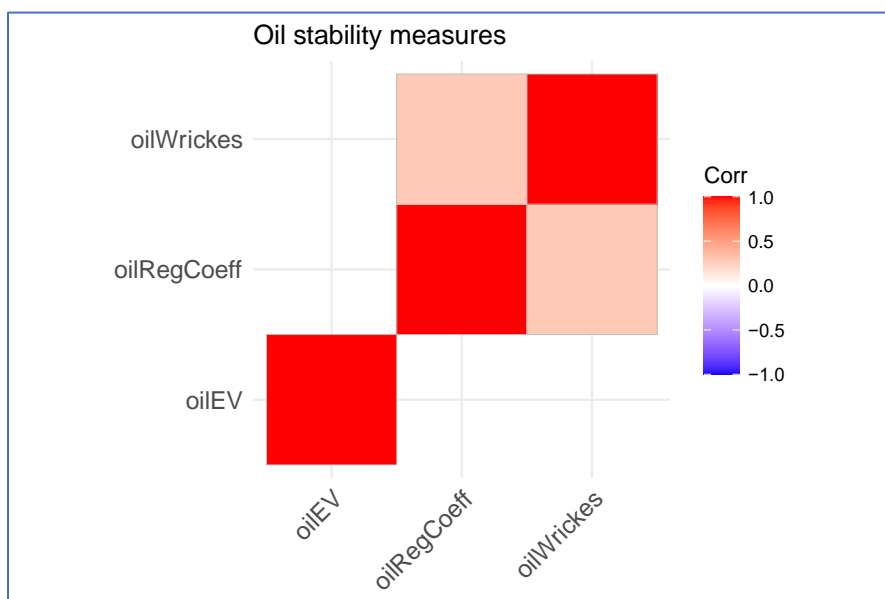
Figure S3.6. Correlation between oil stability measures: Environmental Variance (EV), Regression Coefficient (b$_i$; RegCoeff), Wricke's ecovalence (Wrickes).
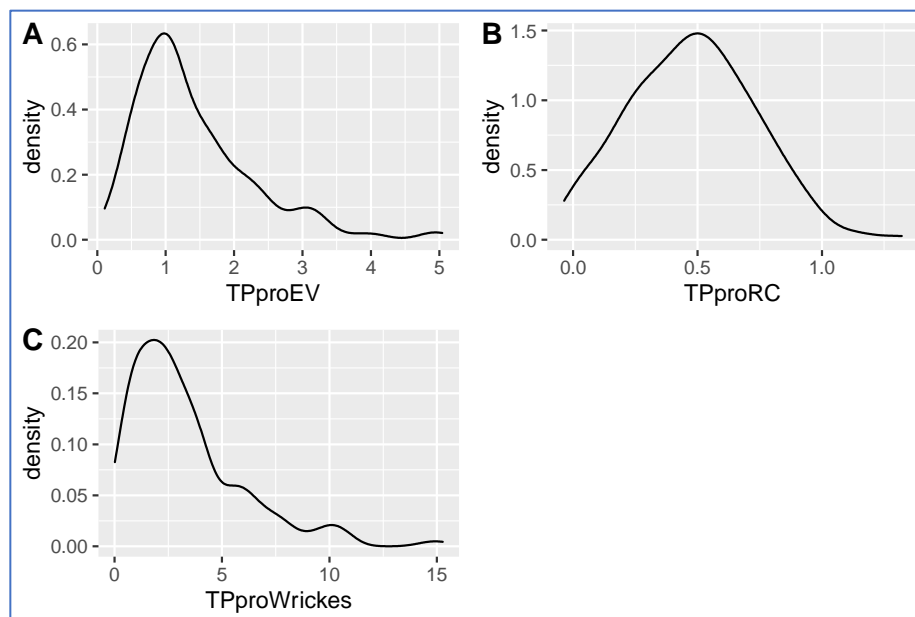
Figure S3.7 Distribution of training population protein content stability values A) environmental variance, B) regression coefficient, and C) Wricke's ecovalence.
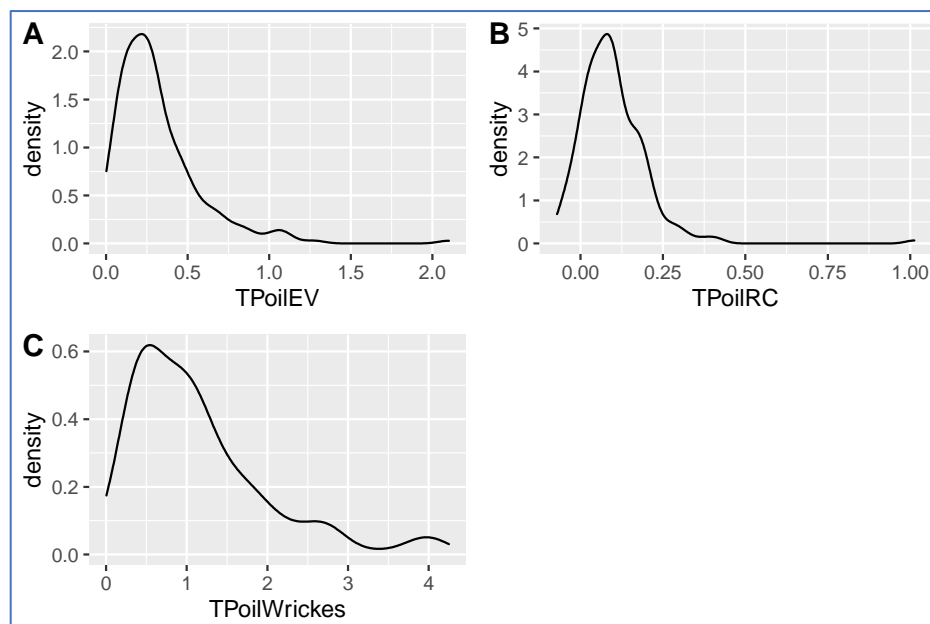
Figure S3.8. Distribution of training population oil content stability values A) environmental variance, B) regression coefficient, and C) Wricke's ecovalence.
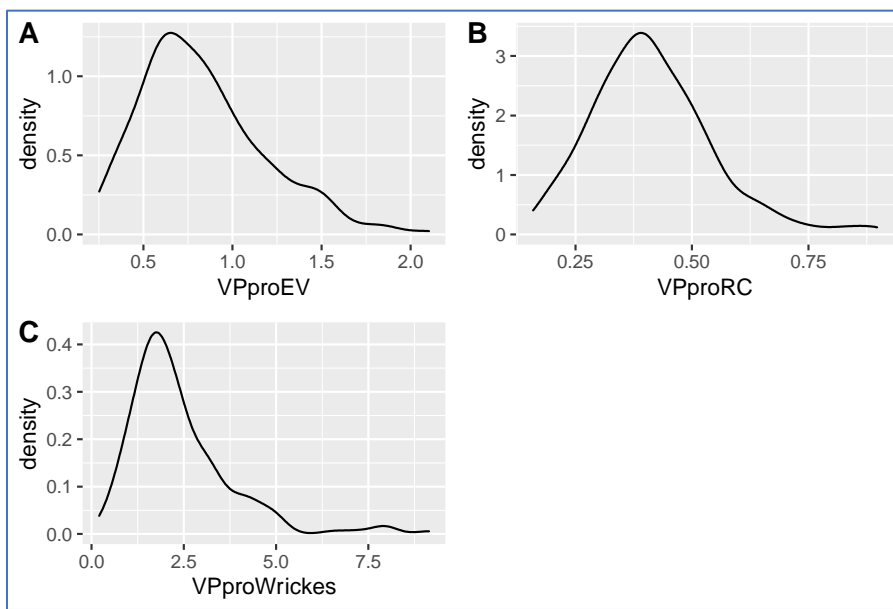
Figure S3.9. Distribution of validation population protein content stability values A) environmental variance, B) regression coefficient, and C) Wricke's ecovalence.
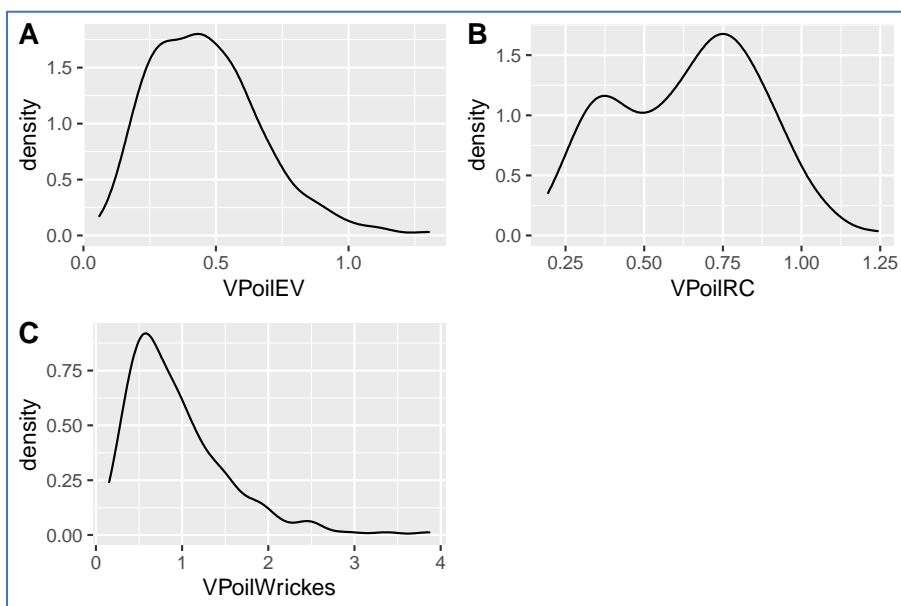
Figure S3.10. Distribution of validation population oil content stability values A) environmental variance, B) regression coefficient, and C) Wricke's ecovalence.
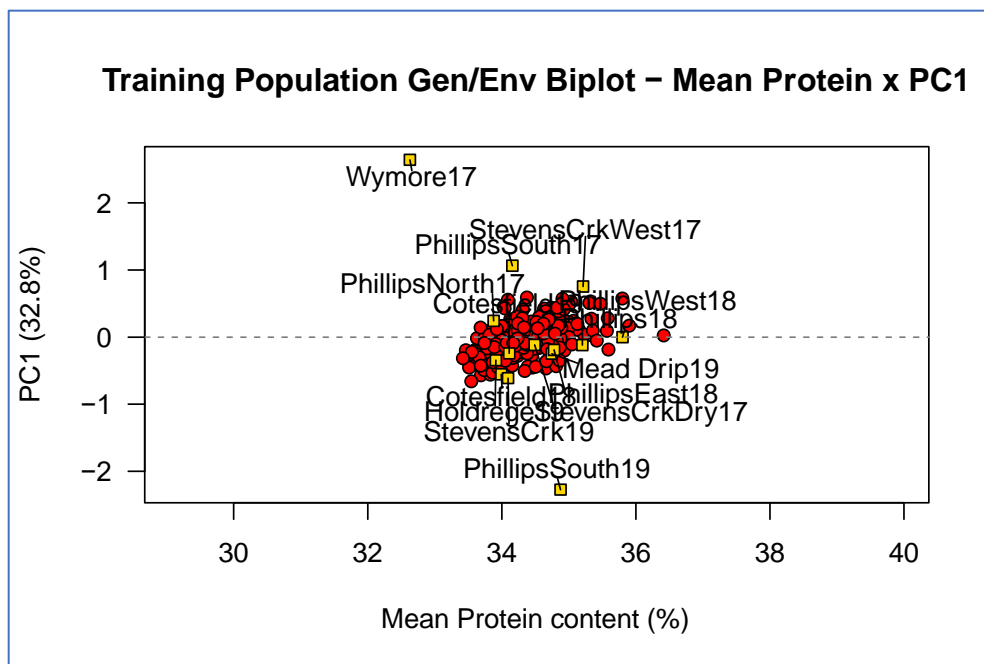
Figure S3.11. Biplot of mean protein content vs. AMMI PC1 (32.8%) for training population genotypes and environments. Genotypes are red circles and environments are yellow squares labelled by name.
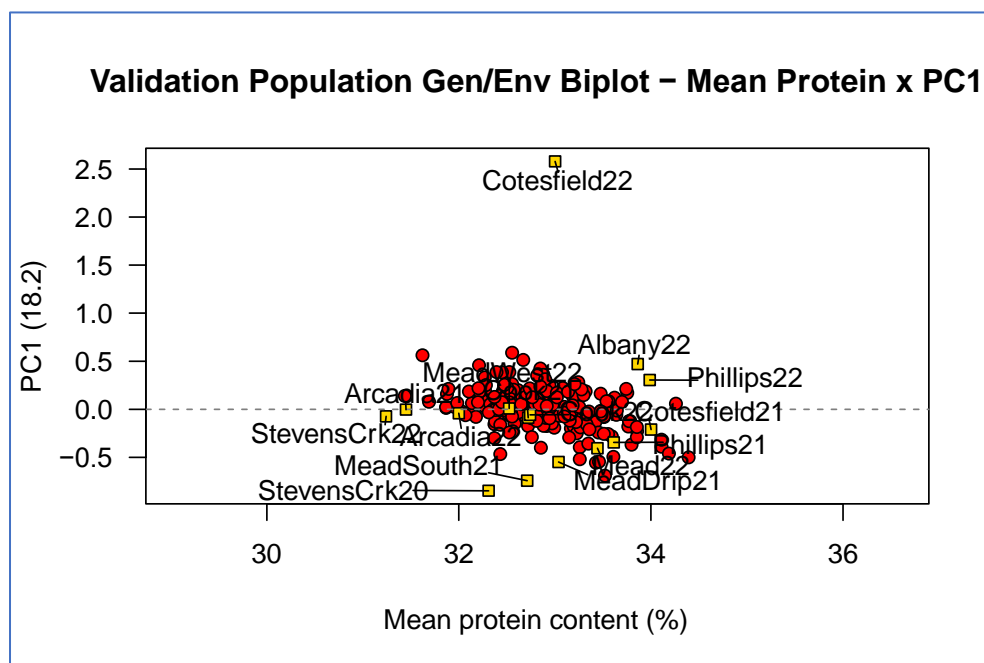
Figure S3.12. Biplot of mean protein content vs. AMMI PC1 (18.2%) for validation population genotypes and environments. Genotypes are red circles and environments are yellow squares labelled by name.
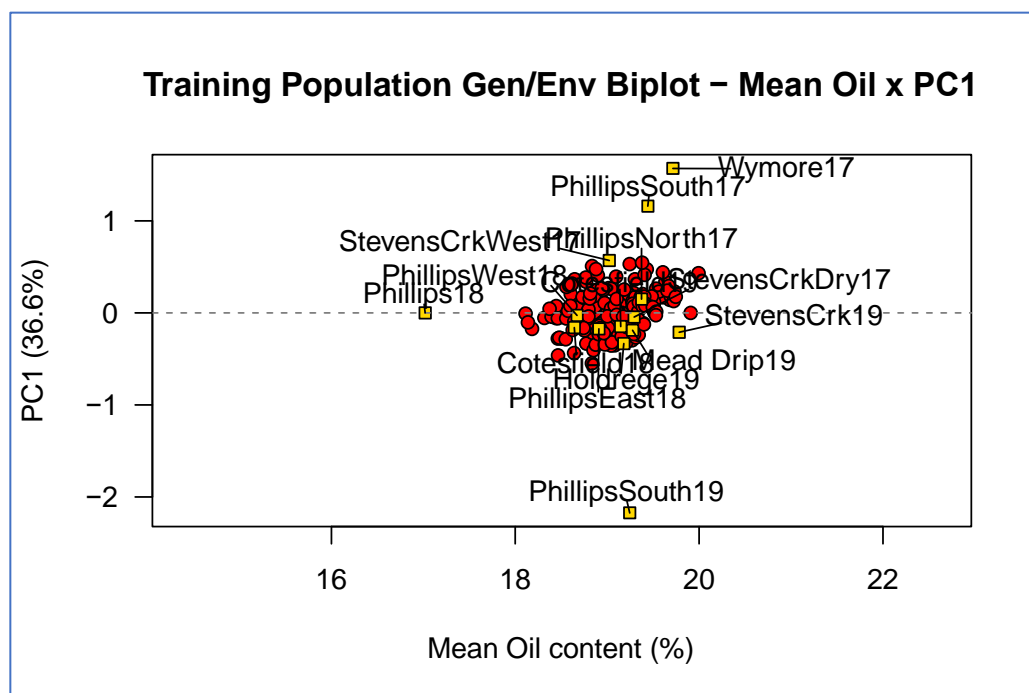
Figure S3.13. Biplot of mean oil content vs. AMMI PC1 (36.6%) for training population genotypes and environments. Genotypes are red circles and environments are yellow squares labelled by name.
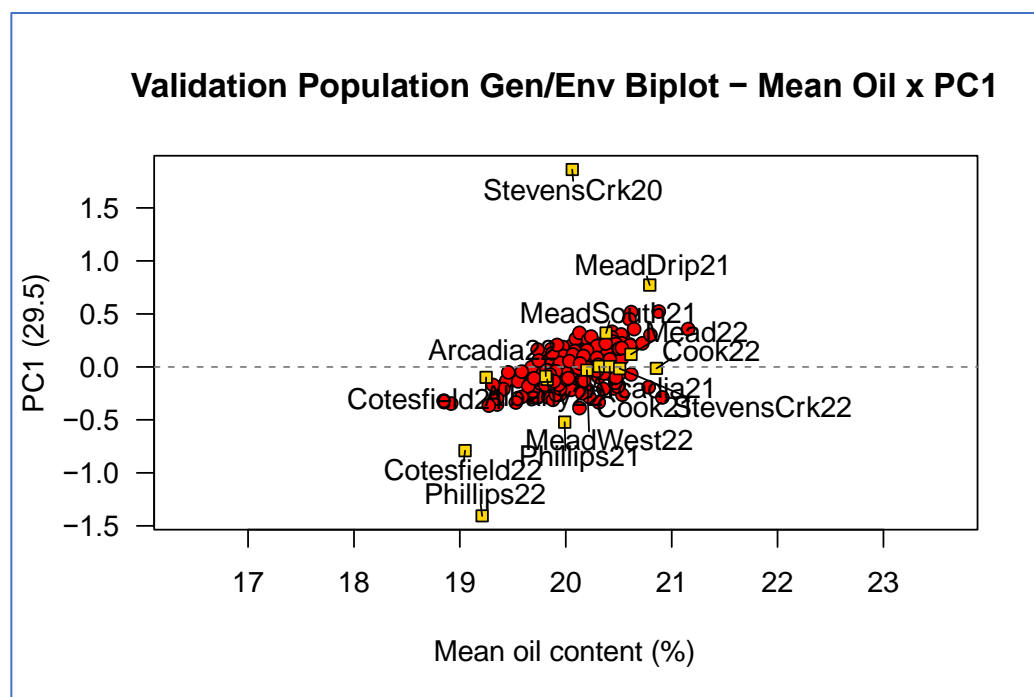

Figure S3.14. Biplot of mean oil content vs. AMMI PC1 (29.5%) for validation population genotypes and environments. Genotypes are red circles and environments are yellow squares labelled by name.
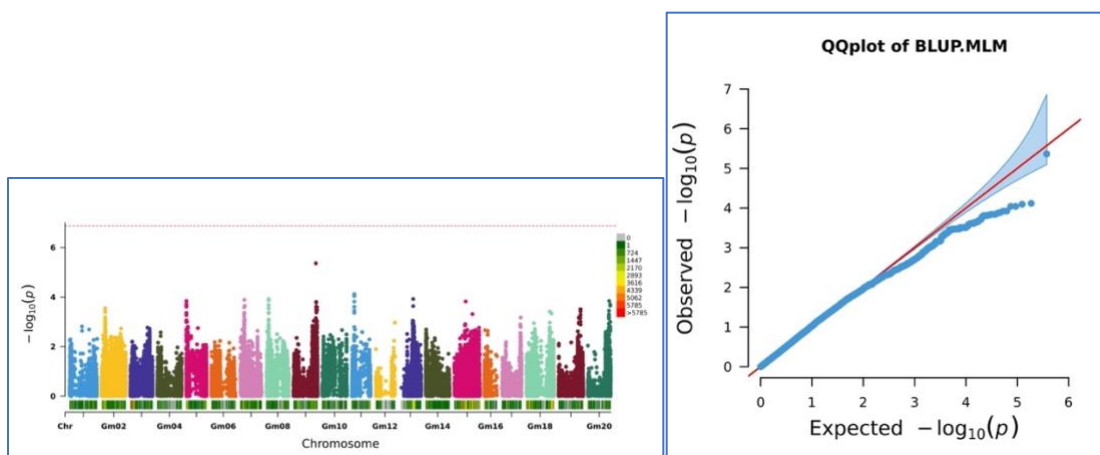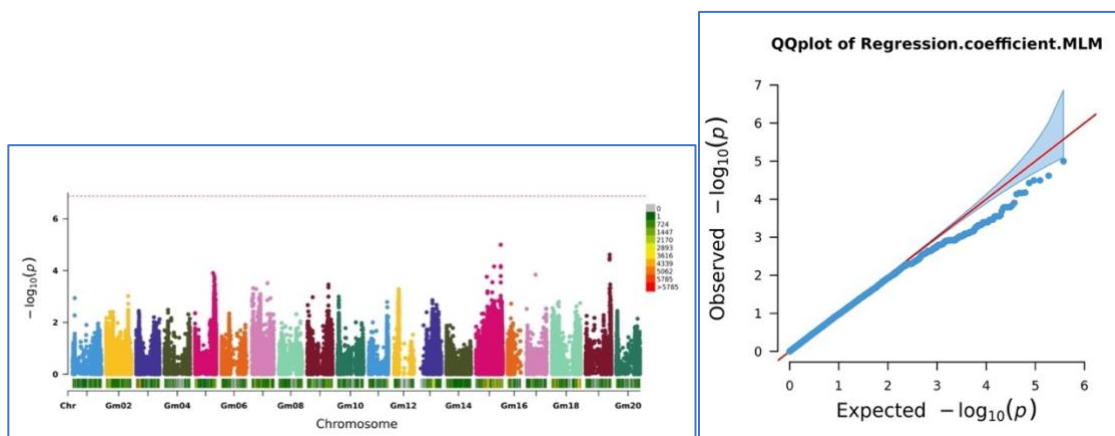
Figure S3.15. Manhattan plot (left) for TP GWAS of protein content BLUPs across all environments (left). QQ plot (right) for TP GWAS of protein content BLUPs. The most significant marker from each 10,000 base pair window was used for genomic prediction of yield for across-population prediction and VP cross-validation (rMVP).
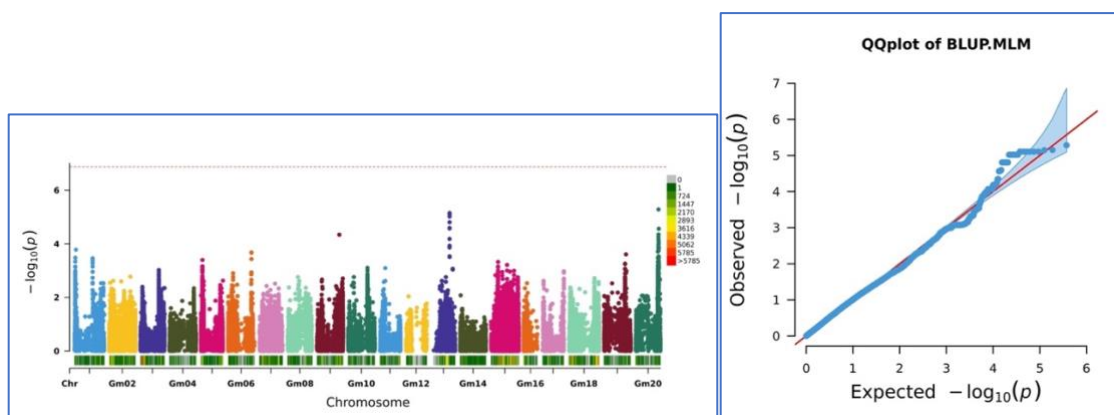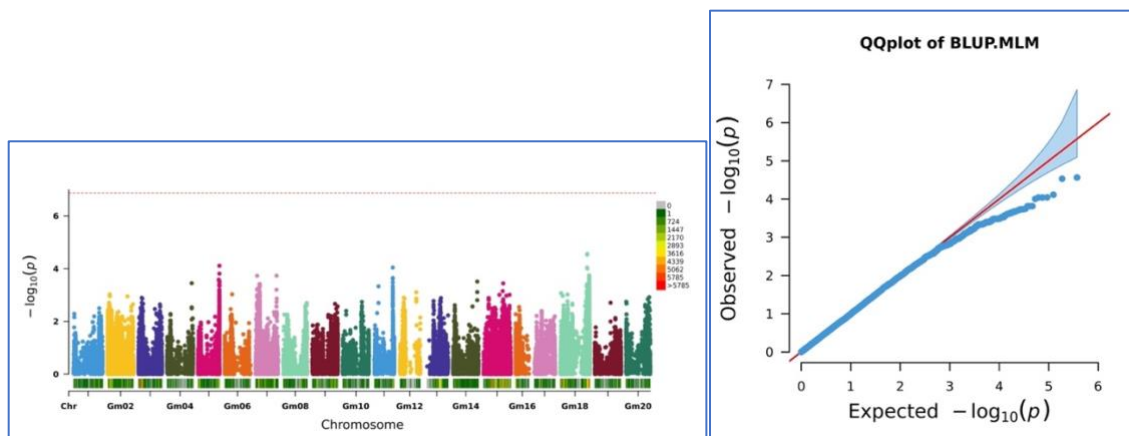

Figure S3.16. Manhattan plot (left) for TP GWAS of oil content BLUPs across all environments (left). QQ plot (right) for TP GWAS of oil content BLUPs. The most significant marker from each 10,000 base pair window was used for genomic prediction of yield for across-population prediction and VP cross-validation (rMVP).
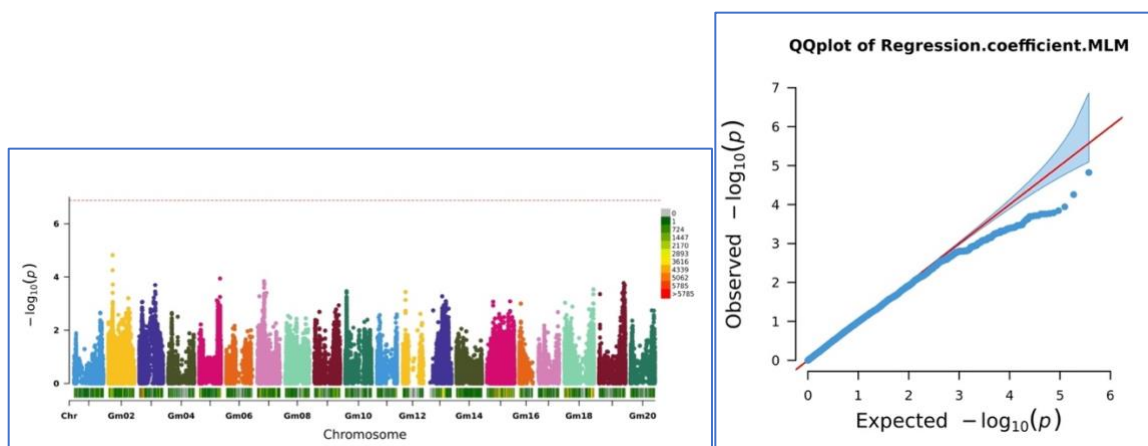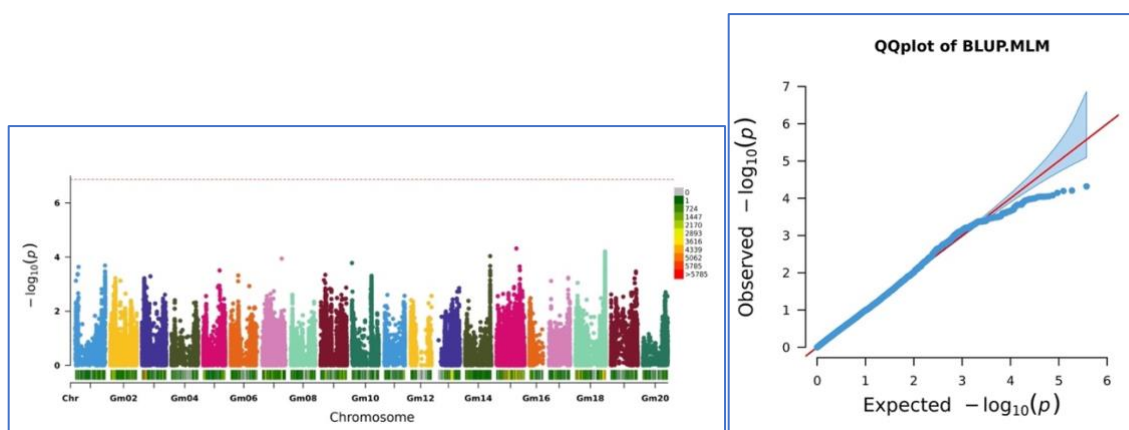
Figure S3.17. Manhattan plot (left) for TP GWAS for protein content stability (regression coefficient). QQ plot (right) for TP GWAS of protein content stability. The most significant marker from each 10,000 base pair window was used for genomic prediction across populations and VP cross-validation of static stability measures (rMVP).
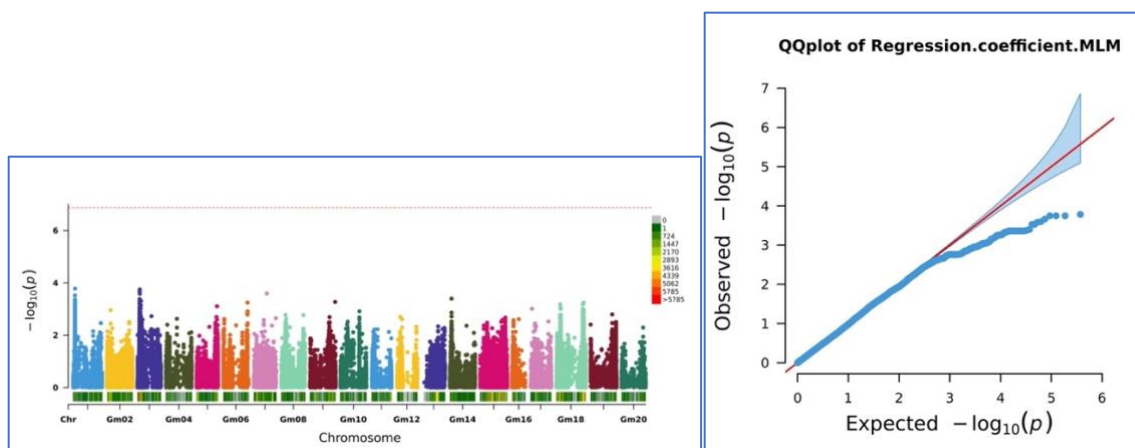

Figure S3.18. Manhattan plot (left) for TP GWAS for oil content stability (regression coefficient). QQ plot (right) for TP GWAS of oil content stability. The most significant marker from each 10,000 base pair window was used for genomic prediction across populations and VP cross-validation of static stability measures (rMVP).
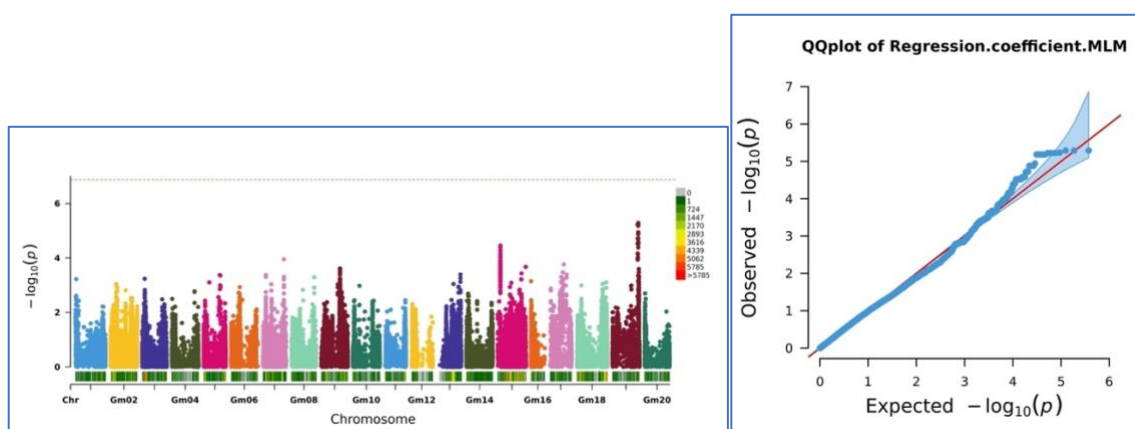
Figure S3.19. Manhattan plot (left) for VP GWAS of protein content BLUPs across all environments (left). QQ plot (right) for VP GWAS of protein content BLUPs. The most significant marker from each 10,000 base pair window was used for TP cross-validation (rMVP).


Figure S3.20. Manhattan plot (left) for VP GWAS of oil content BLUPs across all environments (left). QQ plot (right) for VP GWAS of oil content BLUPs. The most significant marker from each 10,000 base pair window was used for TP cross-validation (rMVP).

Figure S3.21. Manhattan plot (left) for VP GWAS for protein content stability (regression coefficient). QQ plot (right) for VP GWAS of protein content stability. The most significant marker from each 10,000 base pair window was used for TP cross-validation of static stability measures (rMVP).


Figure S3.22. Manhattan plot (left) for VP GWAS for oil content stability (regression coefficient). QQ plot (right) for VP GWAS of oil content stability. The most significant marker from each 10,000 base pair window was used for TP cross-validation of static stability measures (rMVP).
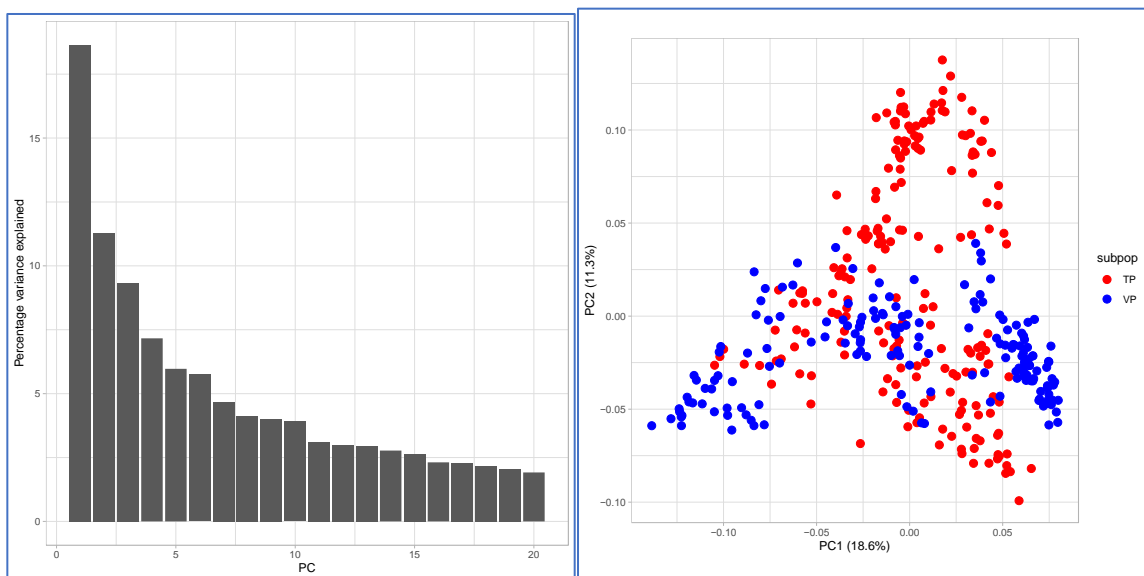
Figure S3.23. Principal components and variance explained for 190 TP and 165 VP individuals combined (left), and visualization of population structure (right) using the first two principal components between TP individuals (red) and VP individuals (blue).
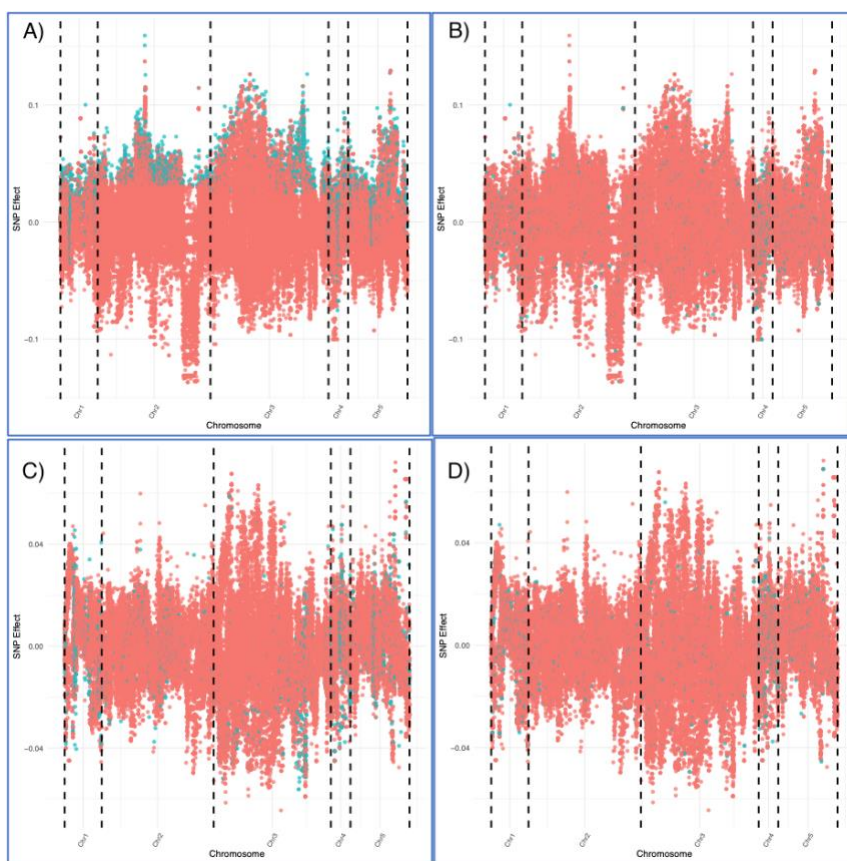
Figure S3.24. Effect size of training population GWAS-selected protein stability SNPs (A), effect size of random SNPs for TP protein stability (B), the effect of TP protein

stability GWAS SNPs in validation population (C), and the effect of random SNPs for VP protein stability (D). Chromosomes 1-5 are shown, with roughly 3,750 selected SNPs in blue in each panel. These SNPs were used for across-population prediction of VP protein stability (Table 4) and VP cross-validation of protein stability (Table 6). Selected SNPs for each scenario are blue.
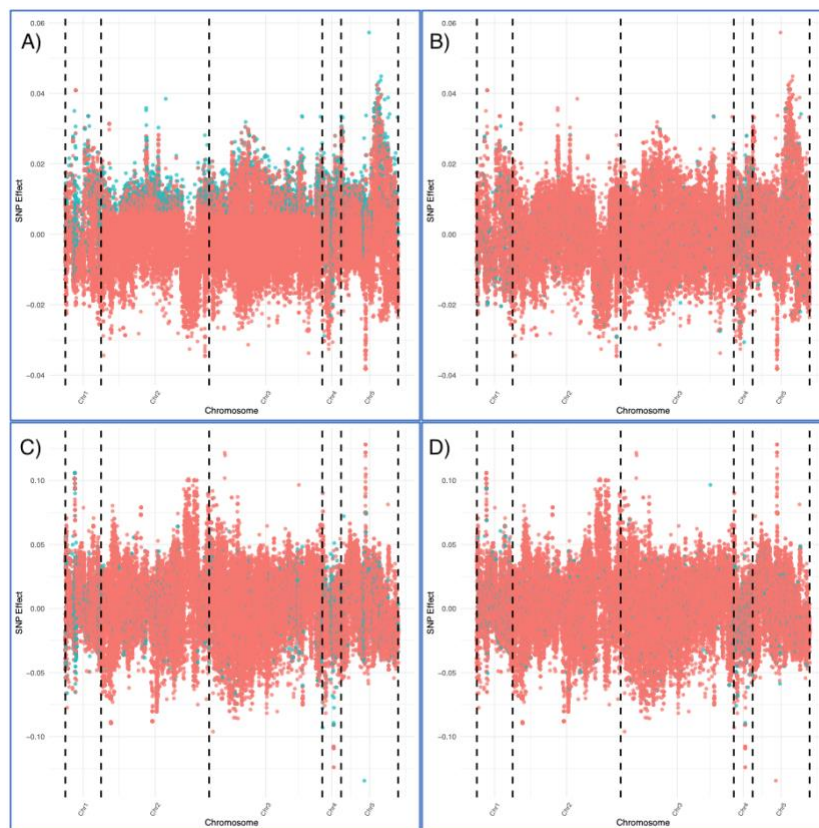
Figure S3.25. Effect size of training population GWAS-selected oil stability SNPs (A), effect size of random SNPs for TP oil stability (B), the effect of TP oil stability GWAS SNPs in validation population (C), and effect size of random SNPs for VP oil stability (D). Chromosomes 1-5 are shown, with roughly 3,750 selected SNPs in blue in each panel. These SNPs were used for across-population prediction of VP oil stability (Table 4) and VP cross-validation of oil stability (Table 6). Selected SNPs for each scenario are blue.

Supplementary Tables

Supplementary Tables S2.1 – S2.4

Table S2.1. Sequencing coverage and parent information for TP and VP.
        Attached table: TableS2.1.xlsx

Table S2.2. Training population locations, coordinates, years, and number of lines from each maturity group grown at each location.

Location	Coordinates	Years	MG I lines	MG II lines	MG III lines	MG IV lines
Cotesfield, NE	41.3581° N, 98.6338° W	2018, 2019	10	56	-	-
Lincoln, NE	40.8136° N, 96.7026° W	2017, 2018, 2019	-	-	75	48
Mead, NE	41.2286° N, 96.4892° W	2018, 2019	10	56	-	-
Phillips, NE	40.8980° N, 98.2132° W	2017, 2018, 2019	10	56	75	48
Wymore, NE	40.1222° N, 96.6623° W	2018	-	-	75	48

Table S2.3. Validation population locations, coordinates, years, and number of lines from each maturity group grown at each location.

Location	Coordinates	Years	MG II lines	MG III lines
Albany, MO	40.2492° N, 94.3317° W	2022	-	51
Arcadia, IA	42.0863° N, 95.0441° W	2020, 2021, 2022	134	-
Cook, NE	40.5098° N, 96.1619° W	2021, 2022	-	51
Cotesfield, NE	41.3581° N, 98.6338° W	2021, 2022	134	-
Lincoln, NE	40.8136° N, 96.7026° W	2020, 2022(MGIII only)	134	51
Mead, NE	41.2286° N, 96.4892° W	2021(2 MGII), 2022	134	51
Novelty, MO	40.0126° N, 92.2069° W	2021	-	51
Phillips, NE	40.8980° N, 98.2132° W	2021, 2022	134	51
Winterset, IA	41.3384° N, 94.0134° W	2020	-	51

Table S2.4. Comparison of GBLUP models for prediction of thirteen yield stability measures across populations.
        Attached Table: TableS2.4.xlsx

Supplementary Tables S3.1 and S3.2

Environmental Variance protein	r	rank correlation	rank coincidence	LogLik
ASReml standard GBLUP	0.164	0.145	0.59	-59.520
ASReml G + E	0.154	0.132	0.59	-57.794
ASReml G + E + GxE	0.154	0.132	0.59	-57.796
Regression Coefficient protein	r	rank correlation	rank coincidence	LogLik
ASReml standard GBLUP	-0.012	-0.034	0.51	210.442
ASReml G + E	-0.012	-0.037	0.50	216.540
ASReml G + E + GxE	-0.012	-0.037	0.50	216.540
Wricke's ecovalence protein	r	rank correlation	rank coincidence	LogLik
ASReml standard GBLUP	0.041	-0.001	0.49	-276.750
ASReml G + E	0.032	-0.001	0.49	-272.102
ASReml G + E + GxE	0.025	-0.001	0.50	-265.732
Environmental Variance oil	r	rank correlation	rank coincidence	LogLik
ASReml standard GBLUP	0.154	0.144	0.58	166.498
ASReml G + E	0.153	0.133	0.56	167.010
ASReml G + E + GxE	0.153	0.133	0.56	167.010
Regression Coefficient oil	r	rank correlation	rank coincidence	LogLik
ASReml standard GBLUP	-0.014	-0.033	0.48	400.673
ASReml G + E	-0.026	-0.039	0.49	407.079
ASReml G + E + GxE	-0.014	-0.024	0.49	408.430
Wricke's ecovalence oil	r	rank correlation	rank coincidence	LogLik
ASReml standard GBLUP	0.097	0.021	0.53	-52.368
ASReml G + E	0.086	0.027	0.53	-49.129
ASReml G + E + GxE	0.077	0.022	0.50	-47.450

Table S3.1. Comparison of three GBLUP models for TP training, VP prediction and validation of protein and oil stability for three stability measures (environmental variance, regression coefficient, Wricke's ecovalence). Predictive ability (r_{MP}), rank correlation, rank coincidence, and ASReml Log Likelihood for each model are provided. Log likelihood closer to zero indicates better model fit. SNPs used were 5,000 TP GWAS-selected SNP for Regression Coefficient stability.

Stability measure	r	rank correlation	rank coincidince	Heritability
EV protein	0.084	0.066	0.53	0.825
Regression Coefficient protein	-0.012	-0.034	0.51	0.699
Wricke's ecovalence protein	0.154	0.135	0.55	0.594
EV oil	-0.117	-0.152	0.43	0.718
Regression Coefficient oil	-0.014	-0.033	0.48	0.666
Wricke's ecovalence oil	0.108	0.061	0.50	0.525

Table S3.2. Across-population prediction using TP SNPs, where each stability measure was predicted using its own set of GWAS-selected SNPs. For environmental variance, prediction accuracy is lower using SNPs from its own GWAS that Regression coefficient SNPs in Table 4. Wricke's ecovalence prediction accuracies are slightly higher than in Table 4.