Summer 7-28-2023

Method Developments to Identify Loci and Selection Patterns Associated with Genotype by Environment Interactions in Soybean

Mary M. Happ
University of Nebraska - Lincoln

METHOD DEVELOPMENTS TO IDENTIFY LOCI AND SELECTION PATTERNS ASSOCIATED WITH GENOTYPE BY ENVIRONMENT INTERACTIONS IN SOYBEAN

by

Mary M. Happ

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Doctor of Philosophy

Major: Agronomy & Horticulture

(Plant Breeding & Genetics)

Under the Supervision of Professor David L. Hyten

Lincoln, Nebraska

July 20, 2023

METHOD DEVELOPMENTS TO IDENTIFY LOCI AND SELECTION PATTERNS ASSOCIATED WITH GENOTYPE BY ENVIRONMENT INTERACTIONS IN SOYBEAN

Mary M. Happ, Ph.D

University of Nebraska, 2023

Advisor: David L. Hyten

For many complex traits such as grain yield, genotype by environment (GxE) interactions are a prevalent source of phenotypic variation. Exploring the capacity of different methodologies to help describe and quantify the GxE interaction landscape for grain yield is an important step in informing plant breeders what the most viable strategies for management and exploitation of GxE may be. In this endeavor, we compared the results from multiple genome wide association studies (GWAS) that used either stability estimators as a phenotype to capture GxE variance, or directly mapped GxE in a mixed model for yield. Leading into this study, a method was developed to enable the cost-effective ascertainment of genotypic information via skim sequencing supplemented with imputation, where it was discovered that imputation accuracy could be maintained at ~98% down to a 0.3X genome coverage. This imputed genotype information was then used in the GWAS analysis that leveraged data from 213 elite local breeding lines tested over the course of three years at multiple sites in eastern Nebraska. Results from the GWAS showed minimal overlap in quantitative trait loci (QTL) discovered between the modeling methods, and that the majority of QTL discovered

displayed a crossover effect. These results prompted our final investigation, where high depth sequencing data was obtained for our study population and used to investigate the effect of artificial selection on genomic windows contributing to GxE interactions. As part of this exploration, several improvements were introduced in the modeling procedure to avoid the inherent biases associated with comparing variance estimates to selection statistics. It was determined through this combination of novel methods that GxE experiences less directional selection pressure than main genetic effects. Interestingly, in contrast to the GWAS results, this also revealed a rich landscape of small, conditionally neutral loci drove the majority of GxE interactions and appeared to be under more directional selection that other GxE effect types.

AUTHOR'S ACKNOWLEDGEMENTS

I would like to first and foremost, extend my gratitude to Dr. David L. Hyten for providing me with the opportunity pursue my Ph.D in his soybean genomics lab over the past seven years. Under his guidance, I feel as though I have been able to develop an exceptionally wide range of range of skills not only related to molecular plant breeding, but as a professional scientist. This is in large part due to the consistent emphasis he places on involving his students at all levels of the research process – from the grant and publication writing stages, to his encouragement to explore new methodologies in data collection and analysis, and more. It is certain that many of the competencies I leave his lab with are a direct result of the effort he has displayed in forming creative, independent researchers. Additionally, I would also like to thank the past and present members of the Hyten lab with whom I worked for both their friendship and support in facilitating many facets of my research experiments.

As I conclude the last chapter of my formal education here at UNL, I also feel an overwhelming sense of appreciation for my committee members - each of whom has inspired, encouraged, and guided me in their own unique and instrumental ways during my graduate education. Dr. Keenan Amundsen, I had no idea at the time, but your bioinformatics course my first semester at UNL set me down the path where I eventually found a niche I started a career with. Dr. George Graef, your lessons in the practical application of genomics to plant breeding were imperative to developing the lens through which I interpret research. Additionally, the friendliness and enthusiasm of the personnel within your breeding program in helping me prepare for and complete the field trials for

my research did not go unnoticed. Dr. Reka Howard, I am grateful not only for your technical expertise and interpretations in the realm of statistics, but your support as I transitioned from student into my first career position.

I would also like to take a moment to recognize and thank my best friend, Dr. Adam Striegel, whom I met while we both pursued our respective doctoral programs here at UNL. While our friendship is many things outside of an overlap on the education section of our resumés, the avid nature by which you pursued excellence in your graduate school endeavors and beyond has often served as inspiration to hold myself to a higher standard.

Finally, I would like to recognize and thank my parents, Dion and Cindy Happ. While it is certain that the experiences I had growing up as part of a farming family had a profound influence on my decision to pursue this doctoral degree, perhaps more important are the values that family instilled in me. I think often of the role they have played in the success I enjoy today, and directly attribute much of it to the positive examples of work ethic and personal discipline they set, as well as the scientific curiosity they nurtured in me from a young age.

PREFACE

Chapter 2 has been published in *G3: Genes, Genomes, Genetics*. (MM Happ, H Wang, GL Graef, and DL Hyten. “Generating high density, low cost genotype data in Soybean [Glycine max (L.) Merr.].”. 2019 Jul 1;9(7):2153–60.).

Chapter 3 has been published in *Frontiers in Plant Sciences*. (M.M. Happ, G.L. Graef, H. Wang, R. Howard, L. Posadas, and D.L. Hyten, “Comparing a Mixed Model Approach to Traditional Stability Estimators for Mapping Genotype by Environment Interactions and Yield Stability in Soybean [Glycine max (L.) Merr.]”. 2021 Mar 31; 12. 630175.).

Chapter 4 has been included in a manuscript that is currently in preparation for publication under the lead of Mary Happ (M.M. Happ, G.G. Graef, R. Howard, and D.L. Hyten. “Variable Selection Patterns Associated with Constitutive Genetic and GxE Effects for Grain Yield in a Locally Adapted Soybean Population” (currently being edited in preparation for submission to a peer reviewed journal).

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

LIST OF DATA FILES

Supplementary Data 2.1: Imputation And Study Genotype Relatedness Matrix (csv, 184.37 KB)

Supplementary Data 3.1: Significant Markers for Grain Yield Stability and GxE Interactions (csv, 35 KB)

Supplementary Data 4.1: Sample Metadata and Sequencing Depths (csv, 161 KB)

CHAPTER ONE: LITERATURE REVIEW

GxE Interactions: Challenges and Strategies in Crop Breeding

Improvements to crop productivity through plant breeding revolve around a general framework of identifying genotypes within a population that exhibit desirable phenotypic traits, and then recombining them to improve the genetic potential of the next generation of varieties. Differential performance among a group of genotypes tested in different environments complicates this endeavor, in a phenomenon that is commonly referred to as genotype by environment (GxE) interactions. For many important traits, such as grain yield, this component is reported to be highly influential. Consequently, it has long been recognized as a crucial factor that affects many levels of decision making within a crop breeding program (Bernardo, 2010).

Understanding the Genetic Architecture

As plant breeders are tasked with managing genetic variation in the pursuit of crop improvement, so they are accordingly challenged with decisions on implementing a variety of tools and analytical techniques that might aid in this endeavor. An important step in deciding what types of genomics assisted methods can help drive progress in a trait is determining the number, magnitude, and nature of the genetic loci affecting it – also broadly referred to as the genetic architecture. For example, traits under the control of a few loci with large effects would be well suited to improvement via a marker assisted selection approach. Conversely, traits with a more complex architecture involving many loci of small effects may be more suited to an application like genomic selection .

Furthermore, one can also combine the approaches by incorporating loci of larger effects into a selection model as fixed effects while still capturing smaller contributions through the random effects portion of the model (Kim et al., 2022; Liu et al., 2019; Spindel et al., 2015). Including an assessment of GxE interactions into the examination of a trait's genetic architecture can also help inform a breeder's choice of a genomics assisted selection methodology. Recent work suggests that in cases of high levels of GxE interactions, special adaptations to the prediction and modeling process that can be made to maximize effectiveness (Costa-Neto et al., 2021; Crossa et al., 2022; Gillberg et al., 2019; Jarquín et al., 2014).

Within the GxE interaction component of a trait's architecture lies another dimension of complexity in relation to a locus' effect in each individual environment. Thus, researchers often further categorize GxE loci according to the directions of the observed effects. If both significant positive and negative effects are observed depending on the environment, a locus is classified as having antagonistic pleiotropy. If the significant effects are all in one direction, a locus is said to display differential sensitivity. A special case of differential sensitivity is when the effect is observed in only one environment, termed conditional neutrality (Des Marais et al., 2013; El-Soda et al., 2014). It is important to keep in mind that classifications in this manner are both conditional on the environments the loci were measured in, and subject to bias relating to a lack of statistical power to detect small effects and stringency of significance thresholds. However, it stands to reason that studying GxE effects in this way can provide general insights into the predominant forces affecting variation for environmental interactions within a specific trait and breeding population.

In addition to quantifying the overall number and size of different loci contributing to trait variance, the degree to which main genetic effects and genotype by environment interactions for a trait may be controlled by the same loci is also of particular interest to plant breeders. If loci for both are genetically linked, then selection for one would affect the other. Interestingly, results from several studies currently indicate limited overlap between main genetic and GxE loci in a variety of traits and crops, including grain yield (Alvarez Prado et al., 2014; Diouf et al., 2020; Kusmec et al., 2017, 2018, Marguerit et al., 2012; Reymond et al., 2003) . This has important implications, as the relative independence of these components provides an opportunity for breeders to adapt cultivars in accordance with a range of objectives, including the development of more broadly stable material as well as those adapted to take advantage of highly specific environmental conditions.

Without variation for GxE responses, a population cannot readily adapt to changes in the surrounding environment. Thus, monitoring and preservation of sources of variability associated with GxE is an important endeavor. In order to do this, one must understand how artificial selection has historically shaped the present genetic architecture in a population. Selection leaves signals of its action across the genome in the form of allele frequency changes surrounding the loci under pressure. Genomic data collected to study the genetic architecture of GxE interactions can therefore also be used to investigate selection patterns at significant loci. A 2017 study in maize found that selection had significantly reduced variation for GxE between tropical and temperate maize varieties (Gage et al., 2017). However similar investigations have not been performed at a regional breeding program level for any crop, which is the method by

which most modern commercial varieties are produced, nor have they considered the potential for varying levels of selective consequences dependent on the type of GxE effect pattern.

Optimization of Multi Environment Trial Design

The overall goal of many crop breeding programs is to produce new and improved varieties and germplasm that performs well across multiple environments. However, when GxE interactions cause different genotypes to be superior in different environments, choosing which varieties to advance through the breeding pipeline presents a difficult dilemma. One approach to help alleviate this problem is to divide the breeding region into smaller groups of more homogeneous sites to be treated as separate breeding targets. This introduces more work for plant breeders (and seed producers), but can also lead to higher heritability, faster genetic gain, and ultimately better products for producers. Even in circumstances where the cost-benefit does not permit for this kind of breeding approach, identifying so called “mega environments” can be a useful for efficiently placing trial sites that capture the range of environmental conditions within a breeding region (Gauch Jr. & Zobel, 1997). In addition to more general information about potential sites (soil test results, historical weather patterns, management practices, etc.), identification of mega environments most often accomplished through clustering methods that take advantage of data from multi-year, multi-environment trials. Methodology in this realm is under consistent refinement, including recommendations on the indications and thresholds at which the breeder should consider separate selection at a new subset of sites (Gauch Jr. & Zobel, 1997; Yan et al., 2000, 2007; Yan, 2015; Yan et al., 2023).

Dividing a breeding region into subsets can help ease some of the challenge GxE interactions pose to breeding when they occur as the result of a wide range of large, but consistently different, environmental factors. However, those GxE interactions stemming from smaller and/or more variable sources, such as micro environmental variation and yearly fluctuating weather patterns, must still be accounted for and managed. Accurate and comprehensive capture of GxE variation through different considerations in the experimental design of multi environment trials is of critical importance to a plant breeder, as their selections are only as precise as the information upon which it is based. This may include the number of sites to test at, the degree of replication at which to test an experimental genotype both within and among environments, and the best strategy for incorporating highly replicated genotypes to be used as checks. When deciding these details, one must also consider the various resource limitations present within a breeding program. This is especially true in the early stages of the breeding process, where the number of experimental varieties is large but the seed availability for each is low. Current research suggests that deploying an unbalanced experimental design where more testing locations are used, but entries are not tested at every site, may increase the selection accuracy for yield. Utilizing common checks between sites in this scenario provides connectivity and boosts statistical power, allowing the breeder to further maximize the information obtained to make their selection decisions (Endelman et al., 2014; Lado et al., 2016; Ward et al., 2019). This alongside the somewhat recent introduction of efficient mixed modeling methodology and software has made it computationally feasibly to analyze large, unbalanced multi environment data quickly and more accurately than ever before (Isik et al., 2017).

Measurement

Many statistical methodologies have been proposed to help plant breeders quantify GxE interactions. One of the primary difficulties in evaluating data from multi environment trials is the summary and interpretation of all individual GxE interactions. Thus, reduction of the problem into a single, generalized 'stability' value is an appealing solution and provides an opportunity to easily use the result as a phenotype in a variety of common genomics-based analyses, in addition to direct use in selection decisions. Within the wide body of research dedicated to this endeavor, phenotypic stability is generally categorized as either "static" or "dynamic" in nature (Gage et al., 2017). A statically stable genotype maintains a constant phenotype across environments, while a dynamically stable genotype will have a constant difference from the mean phenotype from all tested genotypes in each environment. The latter is often considered more applicable to the demands of modern agriculture, where a positive response to agronomic inputs and other environmental conditions is desirable.

However, with the reduction in dimension of measuring GxE interactions, comes the potential for loss of important information. The presence of GxE interactions in multi environmental trials gives rise to a few statistical concerns which may be difficult to account for in more simplified analyses. This includes the likelihood of correlation between environments (Falconer, 1952), differences in error variance between environments (Hu et al., 2013, 2014), and the overall spatial variation within individual sites (Schabenberger & Gotway, 2017). Such aspects can be controlled within the framework of a mixed modeling approach when applied to directly capturing both a trait's main genetic and GxE components (Malosetti et al., 2013; Piepho, 2005;

Schabenberger & Gotway, 2017; van Eeuwijk et al., 2010). Context of the analysis being performed is likely to weigh heavily on the decision to compute more general stability measures versus application of a direct modeling approach to studying GxE. While the former may be better for creating rankings of genotypes to be selected upon, directly modeling GxE may be a more appropriate choice for precise mapping and prediction of the contributions GxE makes to the total architecture for a complex trait.

Exploitation

GxE interactions have often been considered a negative occurrence, due to the reductions in efficiency they generate in the plant breeding selection process. However, simply seeking to reduce GxE can come at the cost of throttling future breeding progress, a scenario that is especially concerning in the context of rapid population expansion and changing climatic conditions (Kusmec et al., 2018). While example of crossover (antagonistic) interactions are often used to illustrate the undesirable effects of GxE interactions when they result in rank changes, literature suggests that the predominant architecture underling GxE interactions is mostly differential sensitivity (Schabenberger & Gotway, 2017).

This has important implications in the context of exploitation for plant breeders, where alleles with no observed detriment in alternate environments could be recombined into superior genotypes that may be more readily adaptable to a range of environmental challenges. One noteworthy example of this in practice was the wide introgression of the *Sub1*A gene into rice varieties in Southeast Asia (Bailey-Serres et al., 2010). Interestingly, this gene confers both high flooding and drought tolerance, as well as

carrying no performance decrease in non-stress conditions (Fukao et al., 2011; Xu et al., 2006). While this strategy is feasible for loci of moderately large effect sizes, exploration of genomic prediction models that emphasizes stacks of GxE loci with small yet desirable contributions, is still lacking.

Current research also suggest that high levels of GxE variability can be related to higher phenotypic mean values and/or better stability. A recent paper investigating the phenotypic plasticity of seed yield among Argentinian soybean varieties discovered that increased GxE interactions were strongly correlated with both an increase in average yield, and a decrease in yield variability in highly productive environments. It was then concluded that this was a function of an extended seed filling period, which allowed the plant a longer time period over which to express plastic responses that captured positive growing conditions in highly productive environments (de Felipe & Alvarez Prado, 2021). Following a similar theme, other studies have also uncovered the contribution of GxE interactions in root and leaf architecture can make to yield (Pires et al., 2020; Xie et al., 2021). However the application of these findings in crop breeding for high input systems has been questioned, highlighting the critical nature of characterizing conditions in which GxE interactions will constitute a net benefit (Schneider & Lynch, 2020).

References

Alvarez Prado, S., Sadras, V. O., & Borrás, L. (2014). Independent genetic control of maize (Zea mays L.) kernel weight determination and its phenotypic plasticity. *Journal of Experimental Botany*, *65*(15), 4479-4487.

Bailey-Serres, J., Fukao, T., Ronald, P., Ismail, A., Heuer, S., & Mackill, D. (2010). Submergence Tolerant Rice: SUB1's Journey from Landrace to Modern Cultivar. Rice, 3(2), Article 2. https://doi.org/10.1007/s12284-010-9048-5

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

Costa-Neto, G., Crossa, J., & Fritsche-Neto, R. (2021). Enviromic Assembly Increases Accuracy and Reduces Costs of the Genomic Prediction for Yield Plasticity in Maize. Frontiers in Plant Science, 12. https://www.frontiersin.org/articles/10.3389/fpls.2021.717552

Crossa, J., Montesinos-López, O. A., Pérez-Rodríguez, P., Costa-Neto, G., Fritsche-Neto, R., Ortiz, R., Martini, J. W. R., Lillemo, M., Montesinos-López, A., Jarquin, D., Breseghello, F., Cuevas, J., & Rincent, R. (2022). Genome and Environment Based Prediction ModelsPrediction models and Methods of Complex TraitsComplex traits Incorporating Genotype × Environment Interaction. In N. Ahmadi & J. Bartholomé (Eds.), Genomic Prediction of Complex Traits: Methods and Protocols (pp. 245–283). Springer US. https://doi.org/10.1007/978-1-0716-2205-6_9

de Felipe, M., & Alvarez Prado, S. (2021). Has yield plasticity already been exploited by soybean breeding programmes in Argentina? Journal of Experimental Botany, 72(20), 7264–7273. https://doi.org/10.1093/jxb/erab347

Diouf, I., Derivot, L., Koussevitzky, S., Carretero, Y., Bitton, F., Moreau, L., et al. (2020). Genetic basis of phenotypic plasticity and genotype x environment interaction in a multi-parental population. bioRxiv [Preprint]. 2020.02.07.938456.

Des Marais, D. L., Hernandez, K. M., & Juenger, T. E. (2013). Genotype-by-Environment Interaction and Plasticity: Exploring Genomic Responses of Plants to the Abiotic Environment. Annual Review of Ecology, Evolution, and Systematics, 44(1), 5–29. https://doi.org/10.1146/annurev-ecolsys-110512-135806

El-Soda, M., Malosetti, M., Zwaan, B. J., Koornneef, M., & Aarts, M. G. M. (2014). Genotype × environment interaction QTL mapping in plants: Lessons from Arabidopsis. Trends in Plant Science, 19(6), 390–398. https://doi.org/10.1016/j.tplants.2014.01.001

Endelman, J. B., Atlin, G. N., Beyene, Y., Semagn, K., Zhang, X., Sorrells, M. E., & Jannink, J.-L. (2014). Optimal Design of Preliminary Yield Trials with Genome-Wide Markers. Crop Science, 54(1), 48–59. https://doi.org/10.2135/cropsci2013.03.0154

Falconer, D. S. (1952). The Problem of Environment and Selection. The American Naturalist, 86(830), 293–298. https://doi.org/10.1086/281736

Fukao, T., Yeung, E., & Bailey-Serres, J. (2011). The Submergence Tolerance Regulator SUB1A Mediates Crosstalk between Submergence and Drought Tolerance in Rice. The Plant Cell, 23(1), 412–427. https://doi.org/10.1105/tpc.110.080325

Gage, J. L., Jarquin, D., Romay, C., Lorenz, A., Buckler, E. S., Kaeppler, S., Alkhalifah, N., Bohn, M., Campbell, D. A., Edwards, J., Ertl, D., Flint-Garcia, S., Gardiner, J., Good, B., Hirsch, C. N., Holland, J., Hooker, D. C., Knoll, J., Kolkman, J., … de Leon, N. (2017). The effect of artificial selection on phenotypic plasticity in maize. Nature Communications, 8(1), 1348. https://doi.org/10.1038/s41467-017-01450-2

Gauch Jr., Hugh. G., & Zobel, R. W. (1997). Identifying Mega-Environments and Targeting Genotypes. Crop Science, 37(2), cropsci1997.0011183X003700020002x. https://doi.org/10.2135/cropsci1997.0011183X003700020002x

Gillberg, J., Marttinen, P., Mamitsuka, H., & Kaski, S. (2019). Modelling G×E with historical weather information improves genomic prediction in new environments. Bioinformatics, 35(20), 4045–4052. https://doi.org/10.1093/bioinformatics/btz197

Hu, X., Yan, S., & Li, S. (2014). The influence of error variance variation on analysis of genotype stability in multi-environment trials. Field Crops Research, 156, 84–90. https://doi.org/10.1016/j.fcr.2013.11.001

Hu, X., Yan, S., & Shen, K. (2013). Heterogeneity of error variance and its influence on genotype comparison in multi-location trials. Field Crops Research, 149, 322–328. https://doi.org/10.1016/j.fcr.2013.05.011

Isik, F., Holland, J., & Maltecca, C. (2017). Genetic Data Analysis for Plant and Animal Breeding. Springer International Publishing. https://doi.org/10.1007/978-3-319-55177-7

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. TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik, 127(3), 595–607. https://doi.org/10.1007/s00122-013-2243-1

Kim, G. W., Hong, J.-P., Lee, H.-Y., Kwon, J.-K., Kim, D.-A., & Kang, B.-C. (2022). Genomic selection with fixed-effect markers improves the prediction accuracy for Capsaicinoid contents in Capsicum annuum. Horticulture Research, 9, uhac204. https://doi.org/10.1093/hr/uhac204

Kusmec, A., de Leon, N., & Schnable, P. S. (2018). Harnessing Phenotypic Plasticity to Improve Maize Yields. Frontiers in Plant Science, 9, 1377. https://doi.org/10.3389/fpls.2018.01377

Kusmec, A., Srinivasan, S., Nettleton, D., & Schnable, P. S. (2017). Distinct genetic architectures for phenotype means and plasticities in Zea mays. Nature Plants, 3(9), 715–723. https://doi.org/10.1038/s41477-017-0007-7

Lado, B., Barrios, P. G., Quincke, M., Silva, P., & Gutiérrez, L. (2016). Modeling Genotype × Environment Interaction for Genomic Selection with Unbalanced Data from a Wheat Breeding Program. Crop Science, 56(5), 2165–2179. https://doi.org/10.2135/cropsci2015.04.0207

Liu, X., Wang, H., Hu, X., Li, K., Liu, Z., Wu, Y., & Huang, C. (2019). Improving Genomic Selection With Quantitative Trait Loci and Nonadditive Effects Revealed by Empirical Evidence in Maize. Frontiers in Plant Science, 10. https://www.frontiersin.org/articles/10.3389/fpls.2019.01129

Malosetti, M., Ribaut, J.-M., & van Eeuwijk, F. A. (2013). The statistical analysis of multi-environment data: Modeling genotype-by-environment interaction and its genetic basis. Frontiers in Physiology, 4, 44–44. https://doi.org/10.3389/fphys.2013.00044

Marguerit, E., Brendel, O., Lebon, E., Van Leeuwen, C., & Ollat, N. (2012). Rootstock control of scion transpiration and its acclimation to water deficit are controlled by different genes. *New Phytologist*, *194*(2), 416-429.

Piepho, H. P. (2005). Statistical tests for QTL and QTL-by-environment effects in segregating populations derived from line crosses. Theoretical and Applied Genetics, 110(3), 561–566. https://doi.org/10.1007/s00122-004-1872-9

Pires, M. V., de Castro, E. M., de Freitas, B. S. M., Souza Lira, J. M., Magalhães, P. C., & Pereira, M. P. (2020). Yield-related phenotypic traits of drought resistant maize genotypes. Environmental and Experimental Botany, 171, 103962. https://doi.org/10.1016/j.envexpbot.2019.103962

Reymond, M., Muller, B., Leonardi, A., Charcosset, A., & Tardieu, F. (2003). Combining quantitative trait loci analysis and an ecophysiological model to analyze the genetic variability of the responses of maize leaf growth to temperature and water deficit. Plant physiology, 131(2), 664-675.

Schabenberger, O., & Gotway, C. A. (2017). Statistical Methods for Spatial Data Analysis. CRC Press.

Schneider, H. M., & Lynch, J. P. (2020). Should Root Plasticity Be a Crop Breeding Target? Frontiers in Plant Science, 11, 546. https://doi.org/10.3389/fpls.2020.00546

Spindel, J., Begum, H., Akdemir, D., Virk, P., Collard, B., Redoña, E., Atlin, G., Jannink, J.-L., & McCouch, S. R. (2015). Genomic Selection and Association Mapping in Rice (Oryza sativa): Effect of Trait Genetic Architecture, Training Population Composition, Marker Number and Statistical Model on Accuracy of Rice Genomic Selection in Elite, Tropical Rice Breeding Lines. PLOS Genetics, 11(2), e1004982. https://doi.org/10.1371/journal.pgen.1004982

van Eeuwijk, F. A., Bink, M. C., Chenu, K., & Chapman, S. C. (2010). Detection and use of QTL for complex traits in multiple environments. Current Opinion in Plant Biology, 13(2), 193–205. https://doi.org/10.1016/j.pbi.2010.01.001

Ward, B. P., Brown-Guedira, G., Tyagi, P., Kolb, F. L., Van Sanford, D. A., Sneller, C. H., & Griffey, C. A. (2019). Multienvironment and Multitrait Genomic Selection Models in Unbalanced Early-Generation Wheat Yield Trials. Crop Science, 59(2), 491–507. https://doi.org/10.2135/cropsci2018.03.0189

Xie, X., Quintana, M. R., Sandhu, N., Subedi, S. R., Zou, Y., Rutkoski, J. E., & Henry, A. (2021). Establishment method affects rice root plasticity in response to drought and its relationship with grain yield stability. Journal of Experimental Botany, 72(14), 5208–5220. https://doi.org/10.1093/jxb/erab214

Xu, K., Xu, X., Fukao, T., Canlas, P., Maghirang-Rodriguez, R., Heuer, S., Ismail, A. M., Bailey-Serres, J., Ronald, P. C., & Mackill, D. J. (2006). Sub1A is an ethylene-response-factor-like gene that confers submergence tolerance to rice. Nature, 442(7103), 705–708. https://doi.org/10.1038/nature04920

Yan, W. (2015). Mega-environment Analysis and Test Location Evaluation Based on Unbalanced Multiyear Data. Crop Science, 55(1), 113–122. https://doi.org/10.2135/cropsci2014.03.0203

Yan, W., Hunt, L. a., Sheng, Q., & Szlavnics, Z. (2000). Cultivar Evaluation and Mega-Environment Investigation Based on the GGE Biplot. Crop Science, 40(3), 597–605. https://doi.org/10.2135/cropsci2000.403597x

Yan, W., Kang, M. S., Ma, B., Woods, S., & Cornelius, P. L. (2007). GGE Biplot vs. AMMI Analysis of Genotype-by-Environment Data. Crop Science, 47(2), 643–653. https://doi.org/10.2135/cropsci2006.06.0374

Yan, W., Nilsen, K. T., & Beattie, A. (2023). Mega-environment analysis and breeding for specific adaptation. Crop Science, 63(2), 480–494. https://doi.org/10.1002/csc2.20895

CHAPTER TWO: GENERATING HIGH DENSITY, LOW COST GENOTYPE DATA IN SOYBEAN [*GLYCINE MAX* (L.) MERR.]

Abstract

Obtaining genome-wide genotype information for millions of SNPs in soybean [*Glycine max* (L.) Merr.] often involves completely resequencing a line at 5X or greater coverage. Currently, hundreds of soybean lines have been resequenced at high depth levels with their data deposited in the NCBI Short Read Archive. This publicly available dataset may be leveraged as an imputation reference panel in combination with skim (low coverage) sequencing of new soybean genotypes to economically obtain high-density SNP information. Ninety-nine soybean lines resequenced at an average of 17.1X were used to generate a reference panel, with over 10 million SNPs called using GATK's Haplotype Caller tool. Whole genome resequencing at approximately 1X depth was performed on 114 previously ungenotyped experimental soybean lines. Coverages down to 0.1X were analyzed by randomly subsetting raw reads from the original 1X sequence data. SNPs discovered in the reference panel were genotyped in the experimental lines after aligning to the soybean reference genome, and missing markers imputed using Beagle 4.1. Sequencing depth of the experimental lines could be reduced to 0.3X while still retaining an accuracy of 97.8%. Accuracy was inversely related to minor allele frequency, and highly correlated with marker linkage disequilibrium. The high accuracy of skim sequencing combined with imputation provides a low cost method for obtaining

dense genotypic information that can be used for various genomics applications in soybean.

Introduction

Genomics research has yielded a variety of tools which allow for more efficient and precise translation of genetic variation into crop improvements. Panels of single nucleotide polymorphisms (SNPs) obtained through SNP arrays or genotyping-by-sequencing (GBS) are the most common tool used to explore and make associations between genetic and phenotypic variation. Genomics-assisted crop breeding continues to demand increasing densities of genotype information to successfully dissect and predict genetically complex traits (Hamblin et al. 2011; Lorenz et al. 2011). Current approaches of directly ascertaining a high density of SNP genotype data on large populations are cost prohibitive or fall short of being able capture the maximum amount of genetic space.

Fixed SNP arrays and GBS are popular options for SNP genotyping in crops. Panels ranging in densities of up to ~600,000 variants are now common in several crop species (Rasheed et al. 2017). However, recent genomics studies are utilizing datasets consisting of one million or more markers to answer complex, quantitative genetic questions. The need for this high density of markers is rendering current arrays and GBS approaches inadequate to generate the magnitude of data modern genomic studies require (Tian et al. 2011; Patil et al. 2016; Li et al. 2018). High-depth whole genome sequencing can achieve these marker densities. One study utilizing high-depth whole genome sequencing in soybean found 9,107,000 high quality SNPs (Valliyodan et al. 2016). Despite advances and the plummeting cost of next generation sequencing (NGS) data,

this approach still presents a heavy financial burden, as several reads are required at each variant site to ensure data quality and completeness.

Decreasing genome coverage in the interest of cost savings introduces missing data, which decreases power and can produce biased results. Imputation of missing data has the potential to allow the researcher to recover nearly all of the missing data points resulting from skim sequencing, drastically reducing genotyping expenses associated generating complete, high quality, high resolution SNP datasets. By predicting the unobserved genotypes based on the surrounding variants and their correlation to a complete reference panel, missing data can be amended to the correct allele genotype. This technique has been developed and extensively used in human genomic research, and is now commonly extended to other organisms (Pei et al. 2008; Howie et al. 2009; Howie et al. 2011). Seen frequently in plants is the use of imputation to fill missing data points in GBS data (Chan et al. 2016; Chung et al. 2017). Specially designed populations such as bi-parental, nested, and multi-parent where the founders are genotyped to a high depth and used for the reference haplotypes has been shown to boost accuracy (Tian et al. 2011; Swarts et al. 2014; Huang et al. 2014; Bayer et al. 2015; Cericola et al. 2018).

Crop breeding programs working with inbred species and/or inbred lines are uniquely positioned to leverage imputation algorithms in an extremely accurate manner. Near complete homozygosity through inbreeding or double haploids allows calling of genotypes despite having sampled one allele at the site. Large haplotype blocks in historically inbred crops theoretically permit imputation accuracy to extend across large physical regions, where genotyped markers are sparse but in high correlation with each other. Success with such a combinatorial approach has been reported in rice, using ~1X

coverage sequence data of 517 individuals. Imputation of the missing genotypes in these individuals without a reference panel to produce a SNP panel of ~3.6 million markers with >98% accuracy (Huang et al. 2010). This was confirmed in a later study that also included simulations performed down to 0.1X depth. Falling below a depth of 0.5X resulted in steep accuracy consequences, with concordance falling to 76% at the 0.1X level. (Wang et al. 2016).

Incorporation of a reference panel has been shown to result in large accuracy improvements at sequencing coverage less than <1X in humans, where imputation at the 0.1X level was improved from less than 5% accuracy to ~70% (Pasaniuc et al. 2012). With the growing amount of sequence data present in public databases for many common crops, it is possible to generate an extensive reference panel that might improve accuracy at ultra-low sequence coverage and further cut per sample genotyping cost. In this study, we report on a low coverage whole genome sequencing with imputation approach in a naturally inbred crop, soybean, for producing a low cost, high quality, high density SNP dataset. A reference panel was generated using publicly available high-depth sequencing data for 106 lines, and employed for imputing the missing genotypes of 114 lines sequenced at ultra-low depth. Coverages from 0.1X – 1X depth at intervals of 0.1X were evaluated. The factors influencing error rates and extensibility within/outside soybean were investigated, and the consequences of error rates and types of error on a typical genome-wide association study (GWAS) were explored.

Materials and Methods

Reference Panel

The reference panel for genotype imputation was generated using publicly available sequence data deposited in the NCBI Short Read Archive from study number SRP062245 (Valliyodan et al. 2016). This unfiltered, raw dataset consisted of 106 Glycine max lines sequenced at an average of 17.1X coverage (Supplementary Table 2.1). The raw reads were filtered for adapter sequence contamination, base quality, and truncated reads using Trimmomatic (Bolger et al. 2014). Bowtie2 was used to map reads to the Glycine max Wm82.a2.v1 reference genome with the "very sensitive" option (Langmead and Salzberg 2012). Reads with a mapping quality score of less than 20 were discarded. SNPs were called using the GATK3.7 HaplotypeCaller tool for an initial panel of 13,052,759 SNPs across all lines (Poplin et al. 2017). SNP calls with five or less reads supporting the call were filtered out, as well as calls with a confidence score of less than 20. To control for potential sample contamination/mixing, the inbreeding coefficient, also called the F statistic (Jain and Workman 1967), was calculated using the software Plink1.9 (Purcell et al. 2007). As soybean is historically an inbred crop, one can expect F statistics close to one in Glycine max. Seven samples fell below a cutoff of 0.9 and were discarded from the final reference panel. All heterozygous calls in the remaining 99 lines were filtered, leaving only biallelic SNPs for consideration. The final reference panel spanned 10,803,148 biallelic homozygous SNPs in 99 lines compared to 10,417,285 SNPs found by Valliyodan et al. using the same data set.

Imputation Panel

To generate a low sequence coverage panel for imputation, 114 experimental lines selected from the University of Nebraska soybean breeding program (Supplementary Table 2.1) were sequenced to a depth of 1X or greater on an Illumina NextSeq 500 (Illumina Hayward, Hayward, CA) using the manufacturer's protocol and 150 base pair paired end reads. DNA was isolated from lyophilized leaf tissue collected from twenty plants per genotype using a CTAB based extraction method (Keim 1988) scaled down for a 96 well plate by dividing all reagent volumes by 40. Extracted genomic DNA was fragmented using a Covaris S220 with the manufacturer's recommended settings for generating ~350 base pair length fragments (Covaris, Inc., Woburn, MA 01801). Double sided size selection was performed using KAPA Pure Beads to retain only fragments within the 250-450 base pair range using the manufacturer's protocol and eluted in 40 µl of TE buffer (Roche Sequencing Solutions, Santa Clara, CA 95050). After testing DNA concentration, samples were standardized to 62.5 ng /µl. Libraries were prepared using a custom protocol adapted from literature to perform A-tailing and end-repair in one reaction, and avoid PCR after adapter ligation by extending the incubation time (Kozarewa and Turner 2011; Knapp et al. 2012) . To perform end repair and A-tailing, 16 µl of fragmented genomic DNA for each sample was combined with 1 µl of T4 polynucleotide kinase (PNK) (10U/µl), 1 µl of T4 DNA polymerase (5U/µl), 1 µl of DreamTaq DNA Polymerase (5U/ µl), 2.7 µl of Cut Smart Buffer (10x), 2.2 µl of dATP (10mM), 0.8 µl of dNTP (10 mM), and 0.3 µl of ATP (10mM). Samples were incubated in a thermocycler for 30 min at 20°, and then immediately ramped to 65° and held at this temperature for 30 min. Samples then proceeded immediately to adapter ligation. To the

25 µl of end repaired and A-tailed product the following was added: 10 µl of T4 DNA Ligase Buffer, 3 µl of T4 DNA Ligase (2000U/µl), 3 µl of PEG 6000, 2 µl of PCR grade water, and 2 µl of uniquely barcoded adapters (30mM). Samples were incubated on a thermocycler for 45 min at 20°. After this time, samples were immediately cleaned using KAPA Pure Beads to retain fragments within the 350-550 base pair range and eluted in 20 µl of TE buffer. Multiplexing was performed by combining 5 µl of each individual library. Libraries were quantified using the KAPA Library Quantification Kit for Illumina platforms.

To create subsets simulating depths from 0.1X to 1X at intervals of 0.1X, reads were randomly selected from the raw datasets based upon the total number of reads obtained for each genotype. Each dataset was trimmed for adapter contamination, base quality and truncated reads using Trimmomatic, and then mapped to the Glycine max Wm82.a2.v1 reference genome with Bowtie2 using the "very sensitive" option. Mapped reads below a quality score of 20 were filtered. The genotypes at all 10,803,148 SNP positions in the reference panel were called in the low coverage imputation panel using GATK3.7 Haplotype Caller. Genotyping SNPs from a single read has been found accurate in rice whole genome sequencing and maize GBS applications (Swarts et al. 2014; Wang et al. 2016) . Any heterozygous calls were discarded, as well as calls not matching the two allele options at that position. For each subset, a random 5% of calls were masked and considered "true" genotypes for evaluating imputation accuracy.

To characterize how genetically distinct the experimental lines were from one another, a genomic relatedness matrix was constructed according to the van Raden metric using the R package "synbreed". Prior to calculation, the imputed dataset was filtered to

retain variants with a Beagle posterior genotype probability (GP) score above 0.9, pairwise r^2 LD metric below 0.4, and variant site missing rate below 5% using Plink1.9 (Purcell et al. 2007).

Imputation Concordance Evaluation

For the sake of computational efficiency, imputation was performed on a per chromosome basis using Beagle 4.1 (Browning and Browning 2016) with the low memory option. To assess accuracy, the imputed genotype calls were compared to the masked calls, and the percent of those in agreement constituted overall concordance using GATK 3.7's Genotype Concordance tool (McKenna et al. 2010). This accuracy assessment was performed across sequencing depths and minor allele frequencies. Three post imputation datasets were considered to quantify any accuracy improvement obtained by filtering poorly imputed sites. This included the raw imputed dataset, and two datasets filtered on GP. Values with GP scores under 0.45 and 0.9 were filtered for the latter two evaluation panels, respectively. VCFtools0.01.12a was used to bin by minor allele frequency, and Plink1.9 (Purcell et al. 2007; Danecek et al. 2011) was used to filter on GP score. GP score filtering thresholds were determined after examining their relationship to error rate (Supplementary Table 2.2)

Error and Linkage Disequilibrium

Error in relationship to linkage disequilibrium (LD) was examined as a potential metric of extensibility to other soybean population and crop species. D' and r^2 statistics were calculated for all pairwise reference panel SNPs using Plink1.9 (Gaunt et al. 2007; Purcell et al. 2007). Proportion of errors made at each SNP site across was calculated by

comparing the imputed values to the masked values across all subsets of depths. To reduce noise, data were smoothed through the application of a rolling average window with a width of 1500 SNPs after ordering by the respective pairwise LD metric. A second order polynomial was fit to describe the D' and error relationship, and a simple linear regression was fit to describe the relationship between r^2 and error.

Relationship Between Samples & Reference Panel

Close relatedness between the sample and reference genotypes has been previously reported to increase imputation precision. Relatedness matrices were generated based on five different coefficients and averaged the top five scores from each sample genotype as a metric for gauging degree of relatedness to the reference panel. These measures were plotted against concordance scores from the imputed data filtered for GP scores above 0.9 and averaged across all depth levels. A simple linear regression model was fit to assess potential correlation. Relatedness matrices were calculated using the R package “synbreed”, using options corresponding to measures described by vanRaden, Astle and Balding, Reif, Hayes and Goddard, and Euclidean distances (Wimmer et al. 2012).

Genome Representation

Genomic studies improve as the linkage between the genotyped polymorphism and underlying causative gene increases. The extent of LD between two markers therefore constitutes proxy for the correlation of the marker and underlying gene(s) of interest. To assess how well the panel represented variation across the genome, the distribution of LD in the imputed experimental dataset was compared to the SoySNP50k

Array positions extracted from the imputed experimental dataset (Song et al. 2015). SNPs with MAF below 0.05 were filtered out, a quality control step implemented in most genomic studies. Both D' and r^2 were calculated using Plink1.9, and distributions plotted in R3.4 (Team 2017).

Error and Beagle Posterior Genotype Probability

To explore the possibility of using GP values as a post imputation filtering metric, proportion of error across depth subsets was plotted against GP. A rolling average window with a width of 500 SNPs was applied to the proportion error after ordering by GP, and a second-degree polynomial was fit to describe the relationship in R3.4.

Error Type

Allele frequencies exhibit some degree of influence on the results of many genomics studies. Therefore, how imputation error skews this metric is of significant interest. Masked and imputed datasets were coded according to the major allele in the reference dataset. Errors were binned into four categories, homozygous major to minor, homozygous minor to homozygous major, homozygous major to heterozygous, and homozygous minor to heterozygous, based on which allele was incorrectly imputed and which allele was true. Because all heterozygous calls were filtered in the initial data generation, no heterozygous to major, or heterozygous to minor category exists.

Power Analysis

In the interest of determining the potential cost of imputation error, a basic power calculation for minor to major and major to minor errors in a GWAS was performed.

Using an R implementation of Purcell's "Genetic Power Calculator" (Purcell et al. 2003), power was calculated to detect a moderate effect QTL across minor allele frequency bins from <0.025 to 0.5. Simulations assumed an additive genetic model, 300 genotypes, LD between the QTL and marker of 0.8 D', a significance threshold that mirrored the Bonferroni correction for 1,716,234 SNPs (the final size of the SNP dataset after quality control filtering), and a QTL effect size of 1 standard deviation. Error rates from 1–10% were tested at intervals of 1%, with 100 iterations of the simulation performed at each error level. To investigate the possibility of including more genotypes to overcome power losses associated with imputation error, simulations were also performed for 150, 500, and 1000 genotypes for a 5% error rate at the same conditions as specified above.

Cost Analysis

Decreasing cost per sample allows a researcher to expand a study to overcome power loss introduced through the imputation error. To illustrate the impact of this, per sample sequencing costs were calculated using current Illumina NextSeq500 high throughput 300 cycle sequencing kit prices, cost analysis of a custom library prep protocol, and CTAB DNA extraction method (Supplementary Table 2.3). The retained cost per sample and average raw concordance were plotted as depth decreased.

Data Availability

Raw sequencing data directly generated by this project for use in creating the study panel has been deposited in the NCBI Short Read Archive under accession number PRJNA512147. The reference panel used for genotype imputation was generated using previously publicly available sequence data deposited in the NCBI Short Read Archive

from study number SRP062245 (Valliyodan et al. 2016). Supplementary figures and tables can be found in "Supplementary Figures and Tables" section of the appendix. The genomic relatedness matrix computed between the reference and imputation genotypes can be found in supplementary data file 2.1.

Results

SNP Genotyping & Imputation

The reference panel for imputation was constructed using 106 Glycine max lines sequenced at an average of 17.1X coverage using publicly available sequencing data deposited in the NCBI Short Read Archive (Valliyodan et al. 2016) (Supplementary Table 2.1). After quality control measures were applied to the raw and mapped sequence data (see Materials & Methods), a final reference panel of 10,803,148 biallelic homozygous SNPs across 99 lines was generated. SNPs discovered in the reference panel were used to genotype experimental lines in the study panel. This consisted of 114 lines that were sequenced to a depth of at least 1X. Coverages from 0.1X to 1X were analyzed by randomly subsetting reads from the raw sequence data. Of the 10,803,148 million markers discovered in the reference panel, the number of SNPs genotyped by this low coverage study panel subsets ranged from 133,747 to 1,288,463 markers. These subsets also ranged in missing data rates from 95.26 to 67.56% for those markers (Table 2.1).

Using the reference panel, genotype values for all missing positions were imputed. Close relatedness between the lines in the experimental panel may bias the overall accuracy of the imputation results. To evaluate this, vanRaden relatedness scores

Mean Depth	Genotyping Rate	Number of SNPs	Reads	Base Pairs
1	32.44%	1,288,463	6,327,889	949,183,385
0.9	30.41%	1,240,823	5,695,100	854,265,047
0.8	27.77%	1,174,619	5,062,311	759,346,708
0.7	24.91%	1,097,843	4,429,522	664,428,370
0.6	21.80%	1,005,880	3,796,734	569,510,031
0.5	18.47%	895,596	3,163,945	474,591,693
0.4	14.98%	760,167	2,531,156	379,673,354
0.3	11.40%	590,786	1,898,367	284,755,016
0.2	7.85%	375,343	1,265,578	189,836,677
0.1	4.74%	133,747	632,789	94,918,339

Table 2.1: The number of markers and genotyping rate in each low coverage subset from 0.1X to 1X sequencing depth. As coverage decreases, the total number of markers captured and completeness of the SNP panel decreases.

were calculated using real and imputed genotypes. The resulting values ranged from -0.44858 to 0.91112, with a median value of -0.02851, mean of -0.00286, and standard deviation of 0.17650. Strong relationships are generally indicated by values over 0.4. Our experimental panel exhibits few strongly related lines, with only 2.7% of all possible pairwise combinations showing a relationship above this threshold. Therefore, we would conclude the majority of our experimental genotypes to be distally/non-related.

An alternative to this whole genome sequencing approach are fixed SNP arrays. However, this method provides less total SNPs for genomic studies and may not capture as much of the genome. High LD between SNPs can be extended to assume a strong

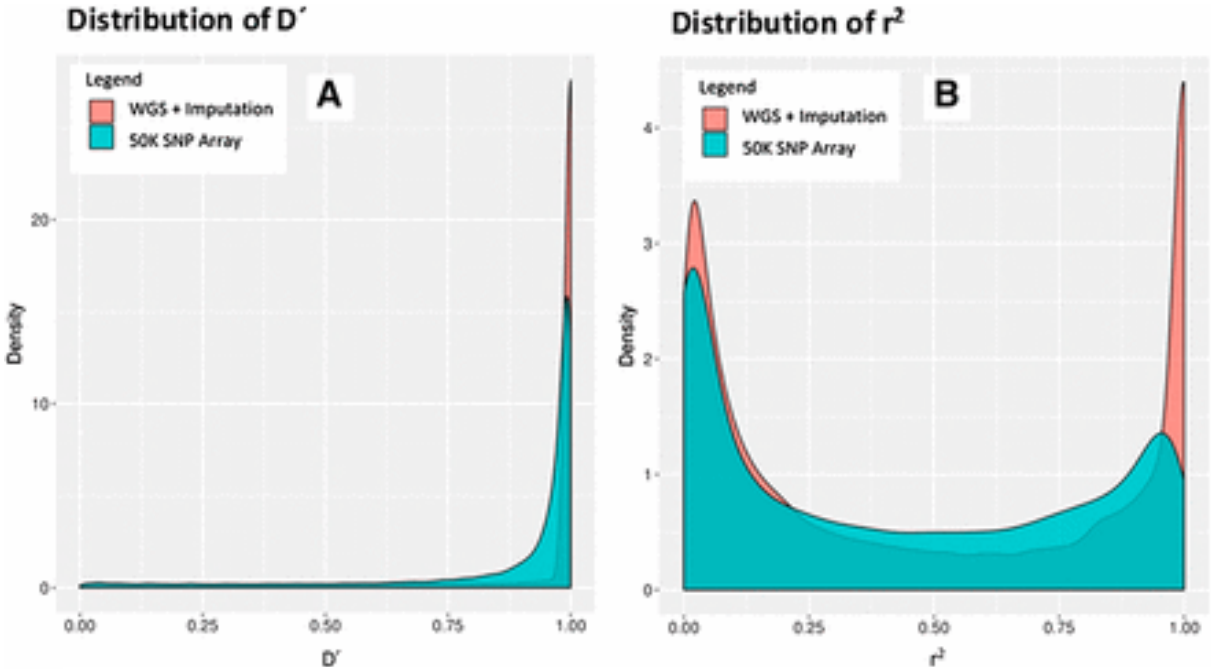


Figure 2.1: Comparing density plots for LD measures D' (A) and r^2 (B) demonstrates that using whole genome sequencing with imputation results in a dataset that has a higher proportion of SNPs is strong pairwise linkage with each other, represented in the heavier t ails in red near D' and r^2 values of 1.

correlation to other genomic variation between them. Plotting the density distributions of r^2 and D' LD measures for the Soy50KSNP Array and imputed dataset demonstrated that whole genome sequencing with imputation had a greater concentration of values toward higher linkage values. Generally, a D' or r^2 of over 0.8 between is considered "strong linkage". The imputed dataset provided 1,716,234 SNPs after common quality control filters, with 36.00% and 85.66% of r^2 and D' values above 0.8, respectively. This is in comparison to the 42,133 SNPs in the fixed array, where 24.20% and 80.00% of r^2 and D' values are above 0.8 (Figure 2.1). If high LD indicates a better tagging of underlying variation, the imputed dataset captures the genome's SNP variation better than the Soy50KSNP Array.

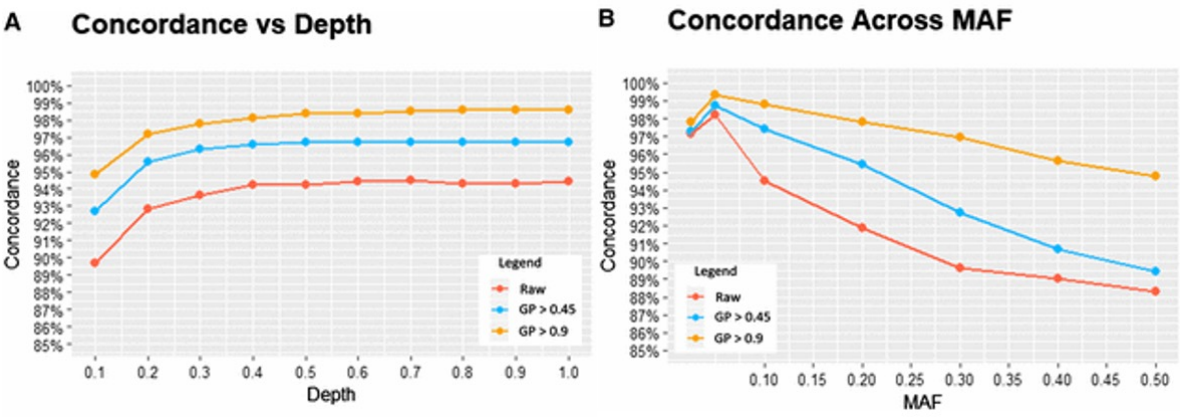


Figure 2.2: A) Overall accuracy of filtered and raw imputed datasets were plotted across the evaluated depths. For all study panels, concordance rapidly erodes below a sequencing depth of ~0.3X. B) Examining accuracy in the context of minor allele frequency (MAF) reveals that error occurs at higher rates as MAF approaches a maximum of 0.5.

Imputation Accuracy

Prior to imputation, 5% of genotype calls from the skim sequencing data were withheld to assess accuracy. Overall imputation accuracy was consistent for raw and filtered datasets as sequencing depth decreased from 1X until 0.3X, where accuracy drops off by an average of 3.5% from 0.3X to 0.1X (Figure 2.2A, Supplementary Table 2.2). Assessing the error type of this study showed that 53.13% of the errors made were incorrect imputation of the minor allele when the major allele was true. Of the remaining errors, 35.10% were incorrect imputation of the major allele when the minor allele was true, and 11.77% were incorrect imputation of heterozygous calls. No heterozygous to major/minor errors exist as at heterozygous calls were filtered in the initial panels (Figure 2.3).

Filtering on Beagle's posterior genotype probability (GP) to improve dataset quality was successful. When imputed positions with a GP score of less than 0.45 were discarded, accuracy improved by an average of 2.50% across sequencing depths. A more

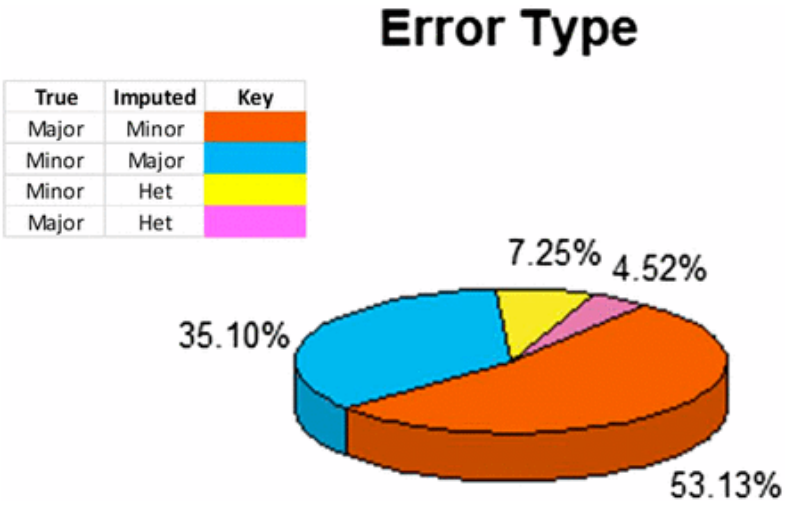


Figure 2.3: Proportion of errors made as categorized by whether the minor/major/heterozygous alleles was misimputed. In over half of all the errors made, Beagle overimputes the minor allele when the major allele is the true genotype. Incorrect heterozygous imputations make up a minor proportion of the total error and would likely be filtered out in inbred panels.

stringent filter that only kept positions with a GP score over 0.9 resulted in a 4.26% increase in accuracy (Supplementary Table 2.2). This practice did reintroduce some missing data, which varied across depth and filtering level. Data loss as a result of post imputation filtering was below 5% for all depths at a filtering level of GP > 0.45, but quickly inflates when filtering for imputation quality of GP > 0.9 to a missing data rate of 20.82% at 0.1X (Supplementary Figure 2.1). While filtering on Beagle's posterior genotype probability may reduce falsely imputed genotypes, it must be balance with the reintroduction of missing data it causes.

The error rate at individual marker loci may not be well captured by the overall concordance across all SNPs. Examining concordance in the context of minor allele frequency (MAF) reveals as MAF values approach a maximum of 0.5, concordance decreases. Application of post imputation filters of GP values increases overall accuracy through improved concordance at these increased MAFs (Figure 2.2B). This trend is

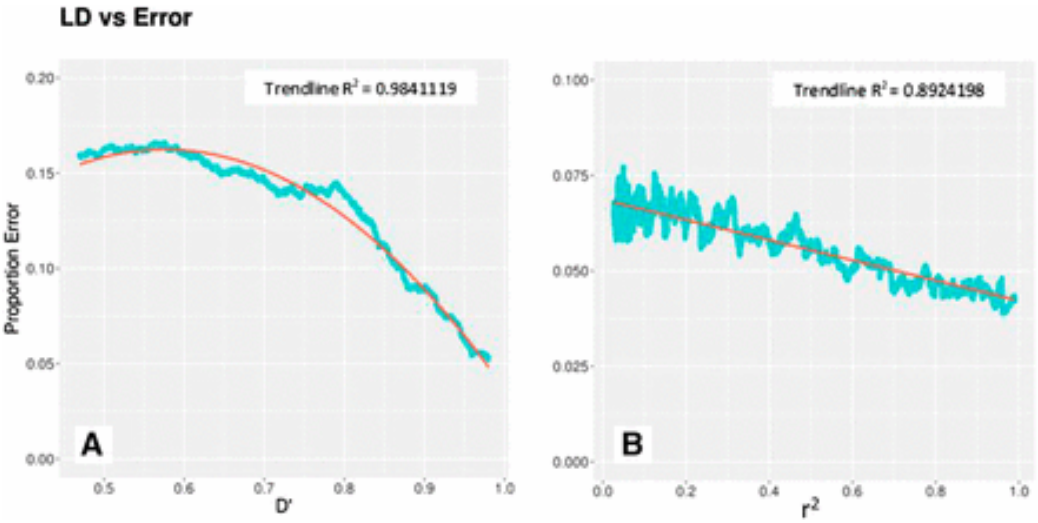


Figure 2.4: Comparing the smoothed frequency of errors made at individual SNP sites with LD measures D' (A) and r^2 (B) demonstrates the strong influence of linkage disequilibrium on imputation accuracy.

uniform across all sequencing depths (Supplementary Figure 2.2). Through examining imputation accuracy in this manner, it is apparent that higher error rates are occurring at SNP positions at MAFs nearest 0.5 than is described by the average concordance measure.

Error rates in imputation may be influenced by characteristics specific to the population and crop species to which it is applied. The correlation between variants is a cornerstone to the success of imputation. If the correlation between alleles is high then imputation accuracy should also be high and as the correlation between alleles decrease then the accuracy of imputation should also decrease. This correlation between alleles can be measured with LD. Soybean is a historically inbred crop with long ranging LD (Zhou et al. 2015). As D' and r^2 approach 1, where neighboring SNPs are in perfect linkage with each other, error rates are at their lowest. Both relationships demonstrate a very strong correlation with R^2 values of 0.98 and 0.89 for r^2 and D' respectively (Figure 2.4), indicating LD is an important factor to consider when applying this technique to other soybean populations or other crop species.

Relationship of the study genotypes to the reference panel genotypes has been suggested as a strong influencer of imputation accuracy. Plotting calculated values for five unique kinship metrics against concordance for each genotype did not demonstrate any strong linear relationships. The maximum correlation for any of the measures was for Reif's method, at an R^2 of 0.26. Examination of the standard error shows that the study population varies narrowly in terms of relatedness to the reference panel. Additionally, assessing the raw values suggests that the study population is weakly related to reference genotypes. This is best illustrated with the vanRaden and Astle & Balding measurements, where a "strong" relationship is usually indicated by values approximately >= 0.4. In both these cases, the largest measure does not exceed 0.18 and 0.16 (VanRaden 2008; Astle and Balding 2009). The combination of diminished values and narrow standard error indicates a weak relationship of the study panel to the reference panel (Supplementary Figure 2.3). The evidence of a weak relationship suggests that relationship was not a strong influencer of the high imputation accuracies obtained.

GWAS Power

Understanding the effect error rate has on genomic studies is important when selecting an appropriate genotyping technique. To determine the effect of the error rate of skim sequencing and imputation has on GWAS we performed power simulations of detecting a moderate effect QTL in a panel of 300 individuals. This power study showed significantly decreased power to detect QTL with increasing errors at MAF from 0.1-0.3. This was most pronounced when the minor allele was incorrectly imputed as the major

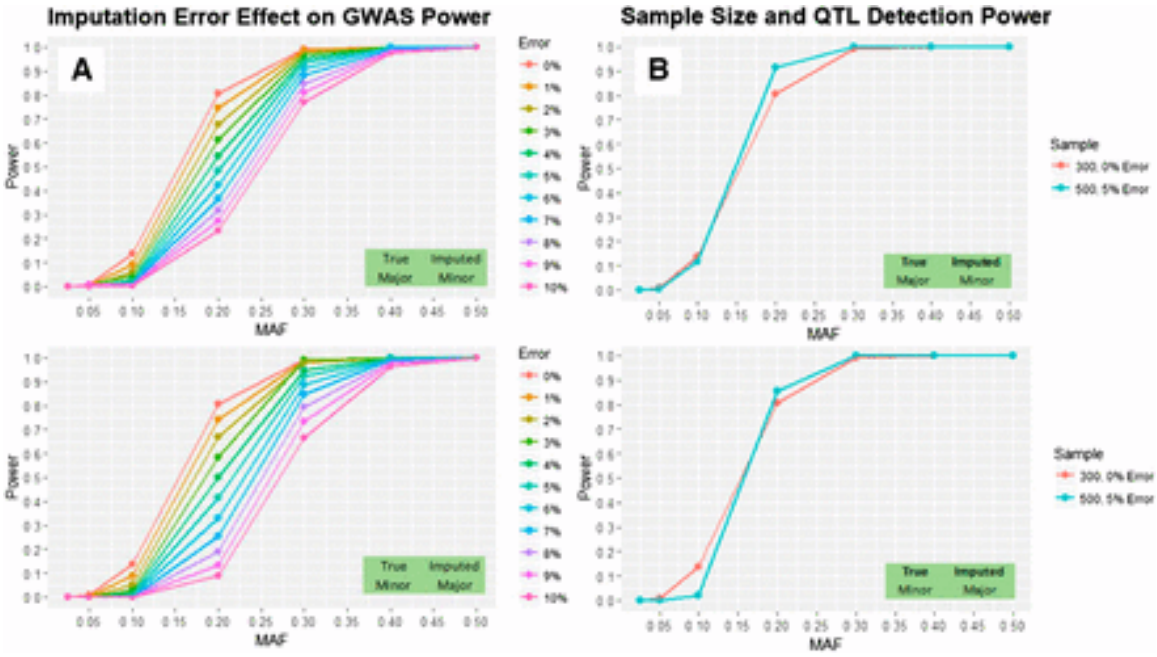


Figure 2.5: A) The power to detect a moderate effect QTL becomes increasingly sensitive to error for both major to minor and vice versa errors at intermediary MAFs. B) Comparing the power to detect the same QTL with 300 samples at a 0% genotyping error vs. 500 samples with a 5% error rate demonstrates that cost savings can be used to increase study sizes in order to recover power losses introduced by the imputation error of both major and minor alleles.

allele. Above 0.3 MAF, power for QTL detection was minimally affected by error (Figure 2.5A). Studying the effect of three additional sample sizes, while assuming a 5% error rate demonstrates the potential for experimenters to recover power losses through inclusion of more genotypes. Including 500 individuals at this fixed error rate recovers and even slightly improves power at mid-range MAFs over studying 300 genotypes with no genotyping error (Figure 2.5B).

Discussion

This study illustrates the potential of low coverage sequencing with imputation as an economical approach to obtaining high density SNP genotype information in soybean. Accelerating improvement of complex phenotypes through genomics necessitates high

quality, high resolution marker data. However, studies are often limited by the cost required to obtain this information through high coverage sequencing. The combination of low coverage sequencing with imputation presents an option that drastically cuts costs while retaining a high level of accuracy. Implementing a similar method in rice allowed researchers to generate a high quality, dense SNP dataset using 1X depth whole genome sequence (Huang et al. 2010; Wang et al. 2016). This analysis in soybean, which differs in the inclusion of a reference panel for imputation, determined sequencing depth could be reduced to 0.3X with no significant accuracy losses. Analogous results have been demonstrated in humans, where it was concluded that a reasonably accurate and dense dataset could be obtained from 0.2X coverage supplemented with imputation using a reference panel (Pasaniuc et al. 2012). To our knowledge, this is the first work to examine using imputation with real sequence data at less than 1X coverage in the construction of a high quality, highly affordable SNP dataset in plants. The effect of imputation method and structure of the reference panel have not been specifically examined in the context of application to skim sequencing, providing future avenues for research and improvement.

While SNP arrays and GBS are popular options for obtaining genotype information, high precision genomics demands markers to be in close linkage to the contributing genes. Regions of the genome with sparsely correlated markers may therefore contain overlooked causal variation (Hirschhorn and Daly 2005; Witte 2010). Skim sequencing with imputation, as investigated here, tags a significantly larger portion of the genome in tighter LD than the current soybean 50k array. This effect may be presumed to extend to GBS datasets of a similar size. Such a boost in resolution may

therefore reveal QTL in regions of the genome that would not have been captured through smaller datasets.

The accuracy and extensibility of this approach in other soybean populations, as well as other crops is based on several factors. To explore potential limitations in this method, population LD and the relatedness of reference panel to study lines were examined. Both of these factors have been implicated as strong influencers of imputation accuracy due to the innate reliance of the technique on the presence of sample haplotypes within the reference panel, as well as the extent of correlation between observed markers (Hickey et al. 2012; He et al. 2015). The strong inverse relationship observed between the proportion of SNPs incorrectly imputed at a given position and LD measurements suggests that for soybean populations and other crops with shorter range LD, imputation accuracy will likely decrease. There was no significant relationship detected between kinship measures and accuracy. However, the study genotypes exhibited little variation for any of the calculated metrics, which can be seen in the low standard deviations. Without a wide range of values to examine, identifying a clear trend is unlikely. The positive effect of relatedness on imputation accuracy is documented in other literature (Hickey et al. 2012; Ma et al. 2013; Boison et al. 2015), and should therefore be a consideration in expanding this method to other soybean populations and crop species. The overall weak kinship between study and reference panels in this data may also be viewed as a positive, since high levels of imputation accuracy were achieved despite this populations being interpreted as distally related.

The power to detect a QTL is partially dependent on the allele frequency at that loci (Ardlie et al. 2002; Tabangin et al. 2009). Therefore, the relationship between

imputation accuracy, minor allele frequency (MAF), and statistical power may be considered particularly important. In agreement with an analysis performed with maize, the data showed steadily decreasing imputation accuracy as MAF increased with the exception of very rare alleles (MAF < 0.05) (Hickey et al. 2012). An opposite tendency was observed with respect to statistical power losses across MAF, so it can be interpreted that at the loci a SNP dataset would display the highest imputation error rates, the GWAS is least affected by them. This trend has also been supported in human imputation analyses looking at sample size inflation factors under different imputation error types (Huang et al. 2009). In both cases, power consequences were greater for incorrect imputation of the minor allele. It is unclear how a combination of error types at a SNP locus would influence genomic studies. Decisions on the level of decreased coverage that can be tolerated should consequently be made not on the overall average concordance, but by examining the concordance across minor allele frequencies in relation to the maximum allowable error to retain power.

The cost savings associated with this method can be used to include more sample genotypes, not only recovering power losses at low minor allele frequencies, but potentially increasing total power. Similar results in humans have indicated sampling more genotypes with small error is more beneficial over fewer genotypes with perfect accuracy (Pasaniuc et al. 2012). Comparing the raw accuracy along sequencing depths along with per sample costs, shows that at the previously identified critical threshold of 0.3X coverage, there is only a 0.85% loss of accuracy relative to using a 1X sequence, while costs decreased 57% (Supplementary Figure 2.4). Moreover, the use of public sequence data to construct a broad reference panel eliminates the cost and limitations of

assembling special populations and sequencing the founders to a high coverage to serve as the reference haplotypes.

Conclusion

Here it is demonstrated that low coverage sequencing accompanied with imputation from a reference panel can be extended below 1X depth in soybean to capture high density, reasonably accurate SNP genotype information economically. The tremendous drop in per sample sequencing cost over high depth methods may allow researchers to expand the number of study genotypes in their investigations, while representing a larger portion of the genome than fixed SNP arrays and GBS data. The potential for success of this genotyping method within and outside of soybean is highly reliant on population LD. Furthermore, researchers should examine accuracy and power within the context of minor allele frequency to make informed decisions about sequencing depth tolerances. As genomics demands increasing SNP panel densities across a wide range of genotypes, skim sequencing with imputation constitutes a financially feasible and highly accurate way to meet these requirements.

Acknowledgments

Research reported in this publication was supported by the Nebraska Soybean Board project #1726. The authors also acknowledge Dr. Reka Howard and Dr. Keenan Amundsen for providing their technical perspective during the compilation of project results and manuscript drafting. This work was completed utilizing the Holland

Computing Center of the University of Nebraska, which receives support from the Nebraska Research Initiative.

References

Ardlie, Kristin G, Kathryn L Lunetta, and Mark Seielstad. 2002. "Testing for Population Subdivision and Association in Four Case-Control Studies." American Journal of Human Genetics 71 (2): 304–11. https://doi.org/10.1086/341719.

Astle, William, and David J. Balding. 2009. "Population Structure and Cryptic Relatedness in Genetic Association Studies." Statistical Science 24 (4): 451–71. https://doi.org/10.1214/09-STS307.

Bayer, Philipp E., Pradeep Ruperao, Annaliese S. Mason, Jiri Stiller, Chon-Kit Kenneth Chan, Satomi Hayashi, Yan Long, et al. 2015. "High-Resolution Skim Genotyping by Sequencing Reveals the Distribution of Crossovers and Gene Conversions in Cicer Arietinum and Brassica Napus." Theoretical and Applied Genetics 128 (6): 1039–47. https://doi.org/10.1007/s00122-015-2488-y.

Boison, S.A., D.J.A. Santos, A.H.T. Utsunomiya, R. Carvalheiro, H.H.R. Neves, A.M.Perez O'Brien, J.F. Garcia, J. Sölkner, and M.V.G.B. da Silva. 2015. "Strategies for Single Nucleotide Polymorphism (SNP) Genotyping to Enhance Genotype Imputation in Gyr (Bos Indicus) Dairy Cattle: Comparison of Commercially Available SNP Chips." Journal of Dairy Science 98 (7): 4969–89. https://doi.org/10.3168/JDS.2014-9213.

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. "Trimmomatic: A Flexible Trimmer for Illumina Sequence Data." Bioinformatics 30 (15): 2114–20. https://doi.org/10.1093/bioinformatics/btu170.

Browning, Brian L., and Sharon R. Browning. 2016. "Genotype Imputation with Millions of Reference Samples." The American Journal of Human Genetics 98 (1): 116–26. https://doi.org/10.1016/J.AJHG.2015.11.020.

Cericola, Fabio, Ingo Lenk, Dario Fè, Stephen Byrne, Christian S. Jensen, Morten G. Pedersen, Torben Asp, Just Jensen, and Luc Janss. 2018. "Optimized Use of Low-Depth Genotyping-by-Sequencing for Genomic Prediction Among Multi-Parental Family Pools and Single Plants in Perennial Ryegrass (Lolium Perenne L.)." Frontiers in Plant Science 9 (March): 369. https://doi.org/10.3389/fpls.2018.00369.

Chan, Ariel W, Martha T Hamblin, and Jean-Luc Jannink. 2016. "Evaluating Imputation Algorithms for Low-Depth Genotyping-By-Sequencing (GBS) Data." PloS One 11 (8): e0160733. https://doi.org/10.1371/journal.pone.0160733.

Chung, Yong Suk, Sang Chul Choi, Tae-Hwan Jun, and Changsoo Kim. 2017. "Genotyping-by-Sequencing: A Promising Tool for Plant Genetics Research and Breeding." Hortic. Environ. Biotechnol 58 (5): 425–31. https://doi.org/10.1007/s13580-017-0297-8.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, et al. 2011. "The Variant Call Format and VCFtools." Bioinformatics 27 (15): 2156–58. https://doi.org/10.1093/bioinformatics/btr330.

Emma Huang, B, Chitra Raghavan, Ramil Mauleon, Karl W Broman, and Hei Leung. 2014. “Efficient Imputation of Missing Markers in Low-Coverage Genotyping-by-Sequencing Data from Multi-Parental Crosses.” https://doi.org/10.1534/genetics.113.158014.

Gaunt, Tom R, Santiago Rodríguez, and Ian NM Day. 2007. “Cubic Exact Solutions for the Estimation of Pairwise Haplotype Frequencies: Implications for Linkage Disequilibrium Analyses and a Web Tool ‘CubeX.’” BMC Bioinformatics 8 (1): 428. https://doi.org/10.1186/1471-2105-8-428.

Hamblin, Martha T., Edward S. Buckler, and Jean-Luc Jannink. 2011. “Population Genetics of Genomics-Based Crop Improvement Methods.” Trends in Genetics 27 (3): 98–106. https://doi.org/10.1016/J.TIG.2010.12.003.

He, Sang, Yusheng Zhao, M Florian Mette, Reiner Bothe, Erhard Ebmeyer, Timothy F Sharbel, Jochen C Reif, and Yong Jiang. 2015. “Prospects and Limits of Marker Imputation in Quantitative Genetic Studies in European Elite Wheat (Triticum Aestivum L.).” BMC Genomics 16 (1): 168. https://doi.org/10.1186/s12864-015-1366-y.

Hickey, John M., Jose Crossa, Raman Babu, and Gustavo de los Campos. 2012. “Factors Affecting the Accuracy of Genotype Imputation in Populations from Several Maize Breeding Programs.” Crop Science 52 (2): 654. https://doi.org/10.2135/cropsci2011.07.0358.

Hirschhorn, Joel N., and Mark J. Daly. 2005. “Genome-Wide Association Studies for Common Diseases and Complex Traits.” Nature Reviews Genetics 6 (2): 95–108. https://doi.org/10.1038/nrg1521.

Howie, Bryan, Jonathan Marchini, and Matthew Stephens. 2011. “Genotype Imputation with Thousands of Genomes.” G3 (Bethesda, Md.) 1 (6): 457–70. https://doi.org/10.1534/g3.111.001198.

Howie, Bryan N., Peter Donnelly, and Jonathan Marchini. 2009. “A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies.” Edited by Nicholas J. Schork. PLoS Genetics 5 (6): e1000529. https://doi.org/10.1371/journal.pgen.1000529.

Huang, Lucy, Chaolong Wang, and Noah A. Rosenberg. 2009. “The Relationship between Imputation Error and Statistical Power in Genetic Association Studies in Diverse Populations.” The American Journal of Human Genetics 85 (5): 692–98. https://doi.org/10.1016/J.AJHG.2009.09.017.

Huang, Xuehui, Xinghua Wei, Tao Sang, Qiang Zhao, Qi Feng, Yan Zhao, Canyang Li, et al. 2010. “Genome-Wide Association Studies of 14 Agronomic Traits in Rice Landraces.” Nature Genetics 42 (11): 961–67. https://doi.org/10.1038/ng.695.

JAIN, S. K., and P. L. WORKMAN. 1967. “Generalized F-Statistics and the Theory of Inbreeding and Selection.” Nature 214 (5089): 674–78. https://doi.org/10.1038/214674a0.

KEIM, and P. 1988. “A Rapid Protocol for Isolating Soybean DNA.” Soybean Genet. Newsl. 15: 150–52. https://ci.nii.ac.jp/naid/10015372412/.

Knapp, Michael, Mathias Stiller, and Matthias Meyer. 2012. “Generating Barcoded Libraries for Multiplex High-Throughput Sequencing.” In Methods in Molecular Biology (Clifton, N.J.), 840:155–70. https://doi.org/10.1007/978-1-61779-516-9_19.

Kozarewa, Iwanka, and Daniel J. Turner. 2011. “Amplification-Free Library Preparation for Paired-End Illumina Sequencing.” In Methods in Molecular Biology (Clifton, N.J.), 733:257–66. https://doi.org/10.1007/978-1-61779-089-8_18.

Langmead, Ben, and Steven L Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” Nature Methods 9 (4): 357–59. https://doi.org/10.1038/nmeth.1923.

Li, Fengmei, Jianyin Xie, Xiaoyang Zhu, Xueqiang Wang, Yan Zhao, Xiaoqian Ma, Zhanying Zhang, et al. 2018. “Genetic Basis Underlying Correlations Among Growth Duration and Yield Traits Revealed by GWAS in Rice (Oryza Sativa L.).” Frontiers in Plant Science 9: 650. https://doi.org/10.3389/fpls.2018.00650.

Lorenz, Aaron J., Shiaoman Chao, Franco G. Asoro, Elliot L. Heffner, Takeshi Hayashi, Hiroyoshi Iwata, Kevin P. Smith, Mark E. Sorrells, and Jean-Luc Jannink. 2011. “Genomic Selection in Plant Breeding: Knowledge and Prospects.” Advances in Agronomy 110 (January): 77–123. https://doi.org/10.1016/B978-0-12-385531-2.00002-5.

Ma, P., R.F. Brøndum, Q. Zhang, M.S. Lund, and G. Su. 2013. “Comparison of Different Methods for Imputing Genome-Wide Marker Genotypes in Swedish and Finnish Red Cattle.” Journal of Dairy Science 96 (7): 4666–77. https://doi.org/10.3168/JDS.2012-6316.

McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, et al. 2010. “The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data.” Genome Research 20 (9): 1297–1303. https://doi.org/10.1101/gr.107524.110.

Pasaniuc, Bogdan, Nadin Rohland, Paul J McLaren, Kiran Garimella, Noah Zaitlen, Heng Li, Namrata Gupta, et al. 2012. “Extremely Low-Coverage Sequencing and Imputation Increases Power for Genome-Wide Association Studies.” Nature Genetics 44 (6): 631–35. https://doi.org/10.1038/ng.2283.

Patil, Gunvant, Tuyen Do, Tri D. Vuong, Babu Valliyodan, Jeong-Dong Lee, Juhi Chaudhary, J. Grover Shannon, and Henry T. Nguyen. 2016. “Genomic-Assisted Haplotype Analysis and the Development of High-Throughput SNP Markers for Salinity Tolerance in Soybean.” Scientific Reports 6 (1): 19199. https://doi.org/10.1038/srep19199.

Pei, Yu-Fang, Jian Li, Lei Zhang, Christopher J. Papasian, and Hong-Wen Deng. 2008. "Analyses and Comparison of Accuracy of Different Genotype Imputation Methods." Edited by Peter Heutink. PLoS ONE 3 (10): e3551. https://doi.org/10.1371/journal.pone.0003551.

Poplin, Ryan, Valentin Ruano-Rubio, Mark A. DePristo, Tim J. Fennell, Mauricio O. Carneiro, Geraldine A. Van der Auwera, David E. Kling, et al. 2017. "Scaling Accurate Genetic Variant Discovery to Tens of Thousands of Samples." BioRxiv, November, 201178. https://doi.org/10.1101/201178.

Purcell, S, S S Cherny, and P C Sham. 2003. "Genetic Power Calculator: Design of Linkage and Association Genetic Mapping Studies of Complex Traits." BIOINFORMATICS APPLICATIONS NOTE. Vol. 19. http://svn.donarmstrong.com/don/trunk/projects/research/linkage/papers/genetic_power_calculator_purcell_sham_bioinfor_19_1_149_2003_pmid_12499305.pdf.

Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A.R. Ferreira, David Bender, Julian Maller, et al. 2007. "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses." The American Journal of Human Genetics 81 (3): 559–75. https://doi.org/10.1086/519795.

Rasheed, Awais, Yuanfeng Hao, Xianchun Xia, Awais Khan, Yunbi Xu, Rajeev K. Varshney, and Zhonghu He. 2017. "Crop Breeding Chips and Genotyping Platforms: Progress, Challenges, and Perspectives." Molecular Plant 10 (8): 1047–64. https://doi.org/10.1016/J.MOLP.2017.06.008.

Song, Qijian, David L Hyten, Gaofeng Jia, Charles V Quigley, Edward W Fickus, Randall L Nelson, and Perry B Cregan. 2015. "Fingerprinting Soybean Germplasm and Its Utility in Genomic Research." G3 (Bethesda, Md.) 5 (10): 1999–2006. https://doi.org/10.1534/g3.115.019000.

Swarts, Kelly, Huihui Li, J. Alberto Romero Navarro, Dong An, Maria Cinta Romay, Sarah Hearne, Charlotte Acharya, et al. 2014. "Novel Methods to Optimize Genotypic Imputation for Low-Coverage, Next-Generation Sequence Data in Crop Plants." The Plant Genome 7 (3): 0. https://doi.org/10.3835/plantgenome2014.05.0023.

Tabangin, Meredith E, Jessica G Woo, and Lisa J Martin. 2009. "The Effect of Minor Allele Frequency on the Likelihood of Obtaining False Positives." BMC Proceedings 3 Suppl 7 (Suppl 7): S41. https://doi.org/10.1186/1753-6561-3-S7-S41.

Team, RC. 2017. "R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2017." https://scholar.google.com/scholar?hl=en&as_sdt=0,28&cluster=8918609904990403039.

Tian, Feng, Peter J Bradbury, Patrick J Brown, Hsiaoyi Hung, Qi Sun, Sherry Flint-Garcia, Torbert R Rocheford, Michael D McMullen, James B Holland, and Edward S Buckler. 2011. "Genome-Wide Association Study of Leaf Architecture in the Maize Nested Association Mapping Population." Nature Genetics 43 (2): 159–62. https://doi.org/10.1038/ng.746.

Valliyodan, Babu, Dan Qiu, Gunvant Patil, Peng Zeng, Jiaying Huang, Lu Dai, Chengxuan Chen, et al. 2016. "Landscape of Genomic Diversity and Trait Discovery in Soybean." Scientific Reports 6 (1): 23598. https://doi.org/10.1038/srep23598.

VanRaden, P.M. 2008. "Efficient Methods to Compute Genomic Predictions." Journal of Dairy Science 91 (11): 4414–23. https://doi.org/10.3168/jds.2007-0980.

Wang, Hongru, Xun Xu, Filipe Garrett Vieira, Yunhua Xiao, Zhikang Li, Jun Wang, Rasmus Nielsen, Chengcai Chu, and Jun Wang wangj. 2016. "The Power of Inbreeding: NGS-Based GWAS of Rice Reveals Convergent Evolution during Rice Domestication." https://doi.org/10.1016/j.molp.2016.04.018.

Wimmer, Valentin, Theresa Albrecht, Hans-Jürgen Auinger, and Chris-Carolin Schön. 2012. "Synbreed: A Framework for the Analysis of Genomic Prediction Data Using R." Bioinformatics 28 (15): 2086–87. https://doi.org/10.1093/bioinformatics/bts335.

Witte, John S. 2010. "Genome-Wide Association Studies and Beyond." Annual Review of Public Health 31: 9–20 4 p following 20. https://doi.org/10.1146/annurev.publhealth.012809.103723.

Zhou, Zhengkui, Yu Jiang, Zheng Wang, Zhiheng Gou, Jun Lyu, Weiyu Li, Yanjun Yu, et al. 2015. "Resequencing 302 Wild and Cultivated Accessions Identifies Genes Related to Domestication and Improvement in Soybean." Nature Biotechnology 33 (4): 408–14. https://doi.org/10.1038/nbt.3096.

CHAPTER THREE: COMPARING A MIXED MODEL APPROACH TO TRADITIONAL STABILITY ESTIMATORS FOR MAPPING GENOTYPE BY ENVIRONMENT INTERACTIONS AND YIELD STABILITY IN SOYBEAN [*GLYCINE MAX* (L.) MERR.]

Abstract

Identifying genetic loci associated with yield stability has helped plant breeders and geneticists begin to understand the role and influence of genotype by environment (GxE) interactions in soybean [Glycine max (L.) Merr.] productivity, as well as other crops. Quantifying a genotype's range of performance across testing locations has been developed over decades with dozens of methodologies available. This includes directly modeling GxE interactions as part of an overall model for yield, as well as methods which generate overall yield "stability" values from multi-environment trial data. Correspondence between these methods as it pertains to the outcomes of genome wide association studies (GWAS) has not been well defined. In this study, the GWAS results for yield and yield stability were compared in 213 soybean lines across 11 environments to determine their utility and potential intersection. Both univariate and multivariate conventional stability estimates were considered alongside a mixed model for yield that fit marker by environment interactions as a random effect. One-hundred and six total QTL were discovered across all mapping results, however, genetic loci that were significant in the mixed model for grain yield that fit marker by environment interactions were completely distinct from those that were significant when mapping using traditional

stability measures as a phenotype. Furthermore, 73.21% of QTL discovered in the mixed model were determined to cause a crossover interaction effect which cause genotype rank changes between environments. Overall, the QTL discovered via explicitly mapping GxE interactions also explained more yield variance that those QTL associated with differences in traditional stability estimates making their theoretical impact on selection greater. A lack of intersecting results between mapping approaches highlights the importance of examining stability in multiple contexts when attempting to manipulate GxE interactions in soybean.

Introduction

Establishing a better understanding of the genetic mechanisms which underlie a trait's variability can lead to greater progress for that phenotype. Grain yield is an example of a trait that displays a complex pattern of quantitative inheritance, dependent on the cumulative action of multiple genes (Falconer, 1996). It has long been recognized that the size and direction of these effects can be influenced differentially by the environmental conditions present over the growing season. These interactions between an individual's genetics with a wide range of environmental factors are commonly referred to as genotype by environment (GxE) interactions (Comstock and Moll, 1963; Crossa, 1990). This is a crucial consideration as a cultivar will be exposed to a variety of conditions in production settings that cannot be predicted in advance.

With regards to quantitative trait loci (QTL) modeling, plant breeders are often most interested in QTL that result in a consistent effect across environments. GxE interactions present a deviation from this simple additive model, but their contribution to overall phenotype makes them important none the less (Kang, 1997). The evaluation of

genotypes across several environments is therefore critical to understanding the contribution of GxE interactions to complex traits such as yield. Due to the contextual nature of GxE interactions, they are often considered a nuisance which obscures the ability to evaluate additive main genetic effects. However, categorizing QTL associated with GxE interactions based on their per environment effects can allow us to highlight those which may be useful for exploitation. Some QTL may have a positive effect on phenotype, but that effect is significantly stronger in some environments. Others are considered "conditionally neutral," only affecting trait values in some environments but having no effect in others. Both of these sources of GxE variation can have a positive impact on phenotype. Also critical, but less directly useful, are QTL contributing to GxE interactions that have opposing effects in different environments (El-Soda et al., 2014).

The structure of multi-environment trial (MET) data presents several statistical challenges which necessitate a more complex approach to analysis, including those for QTL detection. In a 1952 study, Falconer observed when measuring a trait in different environments that the correlation between those environments was a function of GxE (Falconer, 1952). That is, a high positive correlation is indicative of little to no GxE contribution, while values lower than one revealed GxE as a contributor to the measured trait. Another important consideration of MET data is the influence of GxE interactions on the error variance assumptions. Inherently, GxE interactions often cause the magnitude of genetic variance to differ between individual environments. Explicitly, this means the residual error variance in these analyses often break homogeneity assumptions and failure to account for this has the potential to inflate Type I error rates, especially when those include random GxE interaction terms (Hu et al., 2013, 2014). Assuming

genotype [i.e. marker] effects as random in a mixed model approach provides the flexibility to accommodate both differing correlation structures (Piepho, 2005; van Eeuwijk et al., 2010), as well as model a variety of residual error variance structures (Malosetti et al., 2013) and even spatial variation within the error term (Schabenberger and Gotway, 2017). A direct advantage of this analysis structure is the ability to test environment specific QTL effects alongside constitutive main genetic effects, allowing the categorization of QTL into those described above. Additionally, a direct mixed model approach has the advantage of accommodating incomplete and unbalanced datasets that are often common in agronomic field trials (Isik et al., 2017). However, with an increasing number of environments and incorporation of complicated model structures, the number of parameters to be estimated can inflate model size to such an extent that the time and resources to solve it may become impractical (Chen et al., 2010).

Plant breeders aim to select varieties which maintain their high performance across a target region. This trait is commonly referred to as phenotypic "stability," or sometimes "plasticity." Differences in stability among genotypes are the natural result of differing GxE interactions (Becker and Léon, 1988). Selecting varieties with superior stability can become difficult when a breeder has to consider all individual GxE interactions and multiple traits for many testing environments. In this case, transforming a multivariate problem such as GxE interactions into a univariate setup is attractive in that it lends itself to more classical analyses styles. Consequently, there has been a long term emphasis on developing methods that can quantify stability into single values that can then be used to rank and compare test genotypes (Lin et al., 1986; Becker and Léon, 1988; Crossa, 1990). Becker and Léon (1988) categorize stability as either static or

dynamic. Static phenotypic stability refers to the ability of genotype to produce a consistent phenotype independent from changes in environmental conditions. Dynamic stability describes the genotype's response to improved agronomic conditions. This is often considered more relevant in production settings where a variety's ability to respond positively to agronomic inputs such as irrigation and fertilizer is beneficial. However, static stability is often more repeatable and useful for traits such as seed composition which may be expected to meet a certain window of specifications. From the perspective of increasing grain yield, static stability is more relatively advantageous in unfavorable environmental conditions, which is particularly valuable in subsistence agriculture applications (Becker and Léon, 1988).

An increased knowledge of the genetic basis of GxE interactions opens avenues for breeders to manipulate stability through exploiting or minimizing the response to environmental aspects. Several stability measures have recently been used as phenotypes in genome wide association studies (GWAS) to identify novel genomic loci associated with GxE interactions (Bouchet et al., 2016; Xavier et al., 2018; Lozada and Carter, 2020). Explicit mapping of GxE as a marker by environment effect has also been explored, but less considered in stability analyses due to the logistical and computational demands needed to apply the methodology appropriately (Piepho and Pillen, 2004; van Eeuwijk et al., 2010; Malosetti et al., 2013). As yield stability estimates are used to quantify and explain the differences in GxE interactions between genotypes (Becker and Léon, 1988), conducting QTL mapping studies against these values as a phenotype would theoretically reveal some of the same significant loci as directly mapping GxE interactions. A study in barley using both real and simulated data found both static and

dynamic stability QTL for several phenotypes that co-located with loci significant in GxE interactions (Lacaze et al., 2009). Similar analyses in tomato reported a lesser degree of intersection, identifying that 24% of the plasticity QTL they discovered were also identified in a mixed model for GxE interactions (Diouf et al., 2020). To our knowledge, this hypothesis has not been tested in soybean population utilizing an unbalanced design for yield trials. Furthermore, past studies have been limited in the number of stability parameters tested in their comparisons. For this study, we report the results of fitting GxE into a mixed model for yield and compare them to using 29 traditional yield stability estimates to map genetic regions responsible for yield stability in a locally adapted soybean population. Yield estimates were obtained for 213 lines grown at five eastern Nebraska sites over three growing seasons. Mapping of yield stability genes was performed both through explicit modeling of marker by environment interactions, and a traditional GWAS approach for conventional stability measures. The potential overlap between identified QTL was investigated with an emphasis on exploring the ability of traditional stability measures to capture the GxE variation present in multi environment yield trials.

Materials and Methods

Field Sites and Experimental Design

The University of Nebraska-Lincoln soybean breeding program includes several testing sites across Nebraska but is mostly concentrated in the eastern half of the state where most soybean production occurs. Five testing sites from the breeding program

were selected for yield testing that took place over 3 years. Lines belonging to maturity groups I and II were evaluated at the Nebraska locations of Phillips, Cotesfield, and Mead. Lines belonging to group III were evaluated at the Nebraska locations of Phillips, Lincoln, and Wymore (Supplementary Table 3.1). Yield trials were grown in an augmented incomplete randomized block design at each site, with three replicates per site. Each block consisted of 21-24 entries, with checks assigned according to maturity group. Plots consisted of two rows in 2017, and four rows in 2018 and 2019 to minimize border effects. Rows were 6 meters in length with 0.76 meter spacing between rows. Seeds were sourced from a single location grown in the year prior to that growing season. Prior to planting, seeds were treated with CruiserMaxx at a rate of 1 ml per 200 g, to protect from early season insect and fungal diseases (Syngenta Crop Protection AG, CH-4002, Basel, Switzerland). Grain weight and moisture content were recorded at harvest, and adjusted to 13% moisture to calculate grain yield.

GWAS Panel Selection and Genotyping

The University of Nebraska-Lincoln soybean breeding program focuses on the improvement of soybean cultivars for producers in eastern Nebraska. Decades of intensive artificial selection through this program has resulted in a collection of genotypes that are highly refined for local conditions. Two-hundred and thirteen experimental lines from the University of Nebraska-Lincoln soybean breeding program were selected to explore and compare mapping methodologies related to GxE interactions across the lines' target growing region in eastern Nebraska. All lines are F4 derived lines created through bi-parental crosses and single seed descent. Lines selected represented a

range of both average yield and yield stability from a pool of genotypes that had existing yield data from 2013, 2014, and 2015 multi-environment yield trials. Yield stability was calculated using Wricke's ecovalence measure, which defines stability as the interaction of the genotype with its environment summed and squared across environments. Therefore, smaller values are considered more stable as they deviate less from the environmental means (Wricke, 1962).

DNA was isolated from lyophilized leaf tissue collected from twenty plants per genotype using a CTAB based extraction method scaled down for a 96 well plate by dividing all reagent volumes by 40 (Keim, 1988). To generate a high density marker panel that enabled a fine mapping resolution while remaining cost effective, whole genome skim sequencing with genotype imputation was used (Happ et al., 2019). The reference panel for imputation was generated from 99 soybean genotypes with publicly available whole genome sequence data, and consisted of 10,803,148 biallelic homozygous single nucleotide polymorphisms (SNPs). Study genotypes were sequenced at a target of < 1X coverage and imputation performed using Beagle 4.1 (Browning and Browning, 2016). All sequence data was deposited in the NCBI Short Read Archive database accession no: PRJNA699266. Pre imputation processing and quality control was performed according to the previously published protocol (Happ et al., 2019). Plink1.9 (Purcell et al., 2007) was used to eliminate individual low quality imputations with a genotype probability (GP) score of less than 0.9. To eliminate redundancy within the SNP panel, 1,129,769 SNPs in close linkage with a pairwise r^2 value of greater than 0.8 were removed using Plink1.9. Finally, 9,052,059 positions that were non-polymorphic or had a minor allele frequency (MAF) of less 0.05 were filtered out using Plink 1.9. The

final genotyping data for the study panel after these steps consisted of 621,320 high quality, homozygous, biallelic SNP markers.

Accounting for Kinship Between Study Genotypes

Controlling for population structure is an important procedure in association mapping to prevent false positives (Hayes, 2013; Korte and Farlow, 2013). In both scenarios, population structure was controlled through using the first eight principal components in a principal component analysis (PCA) performed in Plink1.9 with a reduced marker dataset. Plink 1.9 first constructs the variance-standardized genetic relationship matrix from marker data before extracting the top 20 principal components (Yang et al., 2011). Markers were first filtered to exclude those with pairwise r^2 linkage values over 0.4, to prevent the results from capturing linkage disequilibrium patterns. The generated eigenvalues were then visualized as a scree plot to determine the number of principal components to be included in the association mapping analysis (Supplementary Figure 3.1). As the plot levels off at approximately the eighth component, it was selected for the cutoff. Use of a genomic relatedness matrix to control for confounding relationships was also tested by computing the Balding-Nichols matrix in EMMAX, which estimates the pairwise relationship between individuals using genome wide SNP data (Kang et al., 2010). This was incorporated as a random effect into the described models. Inclusion of this matrix results in a 1.37 and 2.59 point increase in Akaike information criterion (AIC) and Bayesian information criterion (BIC) values, and therefore was dropped from the association analysis as it decreased modeling efficiency with no improvement.

Association Mapping

Association mapping of both GxE and stability measures required a flexible software that could allow us to fit both linear and mixed models. To this end, we used ASREML-R 4 (Butler et al., 2017) since it provides a wide range of options for modeling both fixed and random effects, as well as the option to include user defined residual error variances structures. Equation 3.1 describes the association analyses performed for explicitly mapping GxE by modeling raw yield averaged across replicates with genotype by environmental levels as a per marker random effects:

$$(3.1) \quad y = X\beta + Z\alpha + e$$

where y is the vector of raw yield estimates assumed to be normally distributed, X is the design matrix of fixed effects including the intercept, the top eight principal components to control for population structure, environment, and maturity grouping, β is the vector of fixed effect coefficients, Z is the incidence matrix of random effects including either marker, marker by year, marker by location, or marker by year by location effects, α the vector of random effect coefficients, and e is the vector of residuals. Allowing for an overall heterogeneous error variance structure resulted in model singularities. Residuals were instead specified as a direct sum of separate variance matrices for each environmental level. Each environmental "level" for the residual is defined as the unique year and location combination. Statistical significance of single markers fit in the linear mixed model was determined using the likelihood ratio test (LRT). This compares the log-likelihood of the model including the marker effect with the log-likelihood of the

model without the marker effect. A multiple testing correction was applied via a Bonferroni threshold ($\alpha = 0.05$) to define significant associations. Results were plotted in a Manhattan plot of –log10 p-values using R3.6 (R Core Team, 2019) with package "ggplot2" (Villanueva et al., 2016).

A wide variety of approaches for calculating yield stability pervades across scientific literature. Recently, Pour-Aboughadareh et al. (2019) reported the development of an R script to calculate a range of phenotypic stability estimates, providing a manageable way to calculate sixteen popular stability estimates using a single R function. This included Plaisted and Peterson's mean variance component, Plaisted's GE variance component, Wricke's ecovalence stability index, regression coefficient, deviation from regression, Shukla's stability variance, environmental coefficient of variance, Nassar and Huhn's statistics (S1 and S2), Huhn's equation (S3 and S6), Thennarasu's non-parametric statistics (NP1-4), and Kang's rank-sum (Happ et al., 2019). While this covered many of the prevalent univariate stability analysis methods, it did not include the multivariate additive main effect and multiplicative interaction (AMMI) analyses methods and subsequent stability values (Sabaghnia et al., 2008). AMMI modeling has been widely used in plant breeding programs to investigate GxE interactions and provide stability estimates through first isolating GxE interactions using a linear model that accounts for some of the main experimental design effects (Abera et al., 2004; Ezatollah et al., 2011; de Oliveira et al., 2014). An AMMI analysis was subsequently performed with the raw yield data and thirteen stability estimates calculated in R3.6 using package "ammistability" (Ajay et al., 2018), including the sum across environments of genotype by environment interactions (GEI) modeled by AMMI (AMGE), AMMI stability index

(ASI), AMMI stability value (ASV), AMMI based stability parameter (ASTAB), sum across environments of absolute value of GEI modeled by AMMI (AVAMGE), Annicchiarico's D parameter (DA), Zhang's D parameter (DZ), averages of the squared eigenvector values (EV), stability measure based on fitted AMMI model (FA), modified AMMI stability index (MASI), modified AMMI stability value (MASV), sums of the absolute value of the IPC scores (SIPC), absolute value of the relative contribution of IPCs to the interaction (Za). Equation 3.2 describes the typical linear model used for association mapping with each of stability measurement, which was also performed in ASREML-R 4 per SNP:

$$(3.2) \quad y = X\beta + e$$

where y is the vector of one of the yield stability estimates assumed to be normally distributed, X is the design matrix of fixed effects including the intercept, the top eight principal components to control for population structure, and the individual marker being tested, β is the vector of fixed effect coefficients, and e is the vector of residuals. The model assumes that e~N(0,Iσ2e). Fitting this model using Nassar and Huhn's S2 statistic, statistical significance of single markers fit in the linear mixed model was determined using the Wald test procedure that is part of the ASREML-R 4 package. A multiple testing correction was applied via a Bonferroni threshold ($\alpha = 0.05$) to define significant associations. Results were plotted in a Manhattan plot of –log10 p-values using R3.6 and package "ggplot2."

Overlap and GxE Variance Explained by QTL

If QTL, via association mapping with yield stability as a phenotype, captures genomic regions involved in GxE interactions, we would expect to see some degree of overlap with QTL identified in the explicit GxE association mapping. To visualize this, the bounds of significant QTL from each association model broadly classified as either GxE, multivariate conventional (AMMI), or univariate conventional were plotted using R3.6 and package "karyoploteR" from Bioconductor (Bernat and Serra, 2017). These were color coded according to model classification. Overlaps between QTL from each model classification were also plotted as a Venn Diagram using R3.6 with package "VennDiagram" (Chen and Boutros, 2011).

Contribution and impact of QTL can be characterized by computing their contribution to overall trait variance. If QTL discovered via association mapping with yield stability as a phenotype captures genomic regions involved in GxE interactions, it could be assumed that these regions would explain significant portions of GxE variance for yield. For each of the methods described, we computed the proportion of yield variance explained by GxE for the most significant SNP, that is, the SNP with the lowest p-value, in each individual QTL region. Equation 3.3 was used to calculate the proportion of yield variance explained by GxE after fitting equation 3.1 in ASREML-R 4 and extracting variance component estimates from the random effects' solutions for each marker:

$$\frac{marker \times environment genotype}{genotype + environment\ + (marker * environment) + residual} \tag{3.3}$$

For each model, the average and standard deviation of these values from all QTL was calculated. The results were plotted in R3.6 using ggplot2 and color coded according to model classification.

GxE Interaction Type

GxE interactions create noise in multi environment trials that make it difficult to identify which genotypes are superior. The two potential outcomes are changes in genotype ranking or a change in distance between rankings. To categorize the effect of each of the QTL discovered via direct GxE modeling, we extracted the effect size of each allelic state at individual environmental combinations for comparison. Effects larger than 6.8 kg/ha (0.25 bu/a) were significant at an alpha value of 0.05 and thus were the only effects considered for this analysis. If all effects were in one sign (all positive or all negative), the QTL was classified as a magnitude interaction. If one of more effects were of an opposite sign than the others, the QTL was considered a crossover interaction. This was performed for all 56 GxE QTL. At each QTL we also examined the overall and per environment adjusted yield distributions. Results of each were plotted in R3.6 using ggplot2.

Principal Component Analysis of Rankings

Plant breeders are often interested in ranking genotypes to make advancement selections. We compared the rankings from yield stability measurements to those ascertained from the Best Linear Unbiased Predictor (BLUP) of the various GxE interactions levels. BLUPs were calculated in ASREML-R 4 according to equation (1), where the incidence matrix Z instead included the random effects of genotype, genotype

by year, genotype by location, and genotype by year by location effects. To rank genotypes via these BLUPs, the absolute value of the BLUP values were taken and then ordered from smallest to largest. Therefore, the smallest GxE BLUP value denoted the most "stable" genotype. Rankings for the conventional yield stability measures were assigned according to their definition. In all cases, a ranking of "1" denoted the most stable genotype. A principal component analysis of these rankings was conducted in R3.6 using the "prcomp" function which is a part of basic R functionality. Results from the principal component analysis were plotted using the "ggplot2" package and color coded according to model classification.

Data Availability

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: NCBI BioProject, PRJNA699266; European Variation Archive, Project: PRJEB43548 and Analyses: ERZ1756748. Supplementary figures and tables can be found in "Supplementary Figures and Tables" section of the appendix. All significantly associated markers for the association models can be found in supplementary data file 3.1.

Results

Phenotype and Genotype Data

The 213 soybean experimental lines were yield tested in an augmented incomplete randomized block design at five eastern Nebraska locations over 3 years to assess grain yield stability. Grain yield over the course of these trials ranged from

2162.74 to 7080.70 kg/ha, with an average of 4976.41 kg/ha and standard deviation of 810.34 kg/ha. The highest yielding year was 2017 with an average grain yield of 5204.80 kg/ha and highest yielding location was Phillips, which averaged 5570.40 kg/ha (Supplementary Table 3.2). Distribution of yield values were approximately normal when examined visually per environment (Supplementary Figure 3.2). Likewise, the association panel captured a wide range of stability values both in the univariate and multivariate measures (Supplementary Tables 3.3, 3,4). Additionally, correlations between univariate stability parameters were much lower than the multivariate stability parameters computed for this study. This suggests capture of different aspects of stability and GxE interactions with the exception of perfect correlations between Wricke's ecovalence, Shukla's stability variance, the GE variance component, and the mean variance component (Supplementary Figure 3.3).

Construction of genotype information was performed using low coverage whole genome sequence data with imputation using a reference panel of deep sequenced soybean genotypes. DNA extracted from leaf tissue collected in 2016 from the study genotypes was used to perform whole genome sequencing at a minimum coverage of 0.3X. After post imputation quality control, the final genotyping panel consisted of 621,320 high quality, homozygous, biallelic SNPs with 1.79% of marker genotypes missing. Per marker missing data rates ranging from 0.34 to 7.74% with a standard deviation of 0.86%.

Association Mapping

Using multiple approaches to map genetic loci associated with grain yield stability in the 213 genotypes revealed 106 significant QTL via the Bonferroni threshold. 86 of these were determined to be independent between all mapping approaches when considering overlaps between QTL bounds (Supplementary Figures 3.4A–3.32A and Figure 3.1). The majority of QTL associated with GxE interactions were found in the marker∗location and marker∗year∗location terms, with some degree of overlap between all interactive terms (Figures 3.1B–D). The number of QTL for overall yield was affected by inclusion of environmental interaction terms, resulting in one additional QTL

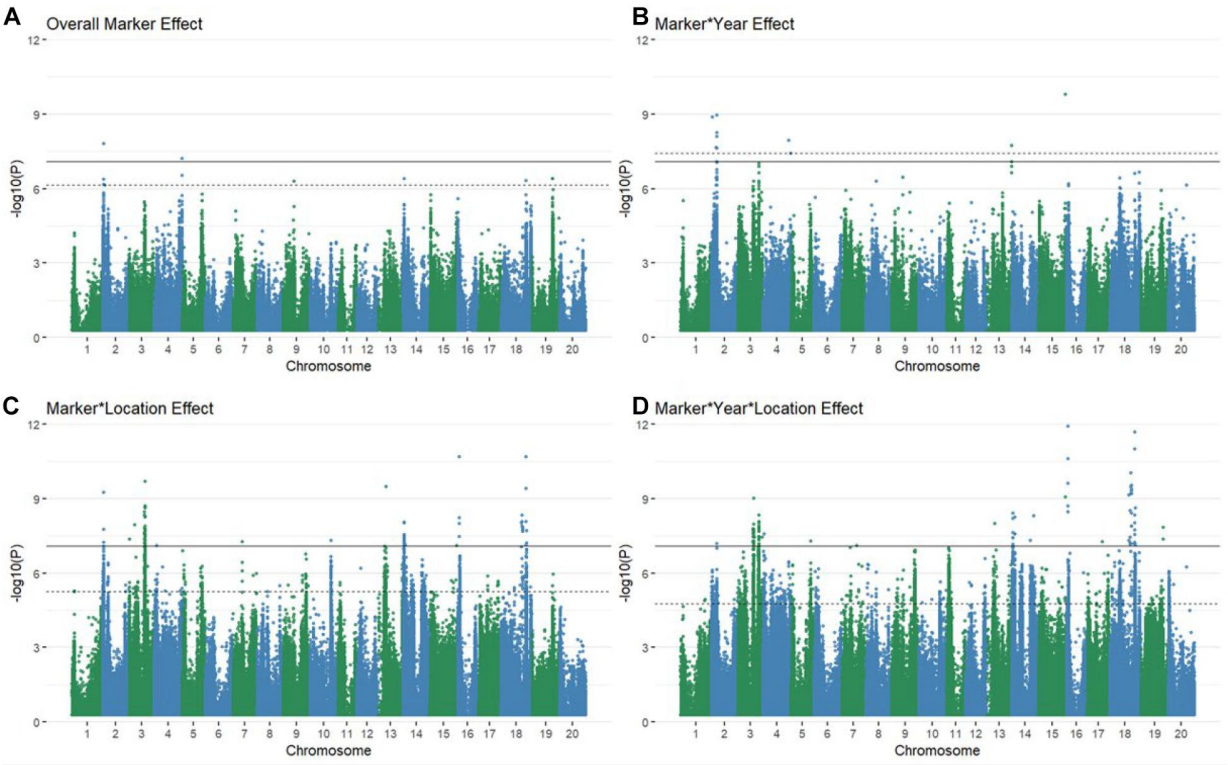


Figure 3.1: Manhattan plots of marker (A) and marker by environment (B–D) levels modeled explicitly as random explanatory variables of raw grain yield. Several associations are significant at every level via both the Bonferroni correction (solid black line) and a 5% FDR (dashed line) with some overlap between QTL discovered for varying levels of GxE interactions (B–D).

significant via the Bonferroni threshold and shifting which QTL were significant at a FDR of 5% (Figure 3.1A and Supplementary Figure 3.4A). Models fitting the coefficient of variation, Finlay Wilkinson, Sum Across Environments of Absolute Value of GEI Modeled by AMMI, and Zhang's D Parameter as phenotypes did not return any associations that were significant by either the Bonferroni correction or a FDR of 5%. The AMMI stability value and AMMI stability index only returned associations that were significant using a 5% FDR threshold. Model inflation was assessed by examining the quantile-quantile plots of p-values produced by each model fit (Supplementary Figures 3.4B–3.32B, 3.33). Deviation from the diagonal suggested considerable inflation in models fitting the GE variance component, mean variance component, Shukla's stability variance, Thennarasu NP2 statistic, and Wricke's ecovalence, and were therefore dropped from consideration in further analyses. Conventional yield stability measures are assumed to explain genotype differences in GxE interactions of multi-environment trials and distill them into a singular value. Overlap between QTL discovered using conventional measures as a phenotype and explicitly modeling GxE interactions may indicate the extent of their interchangeability. Considering the boundaries of the 86 independent QTL discovered in the association mapping, we found only one QTL shared among all three modeling approaches. Univariate and multivariate approaches shared eight intersecting QTL with each other, but only two and one QTL with explicit GxE modeling, respectively (Figure 3.2).

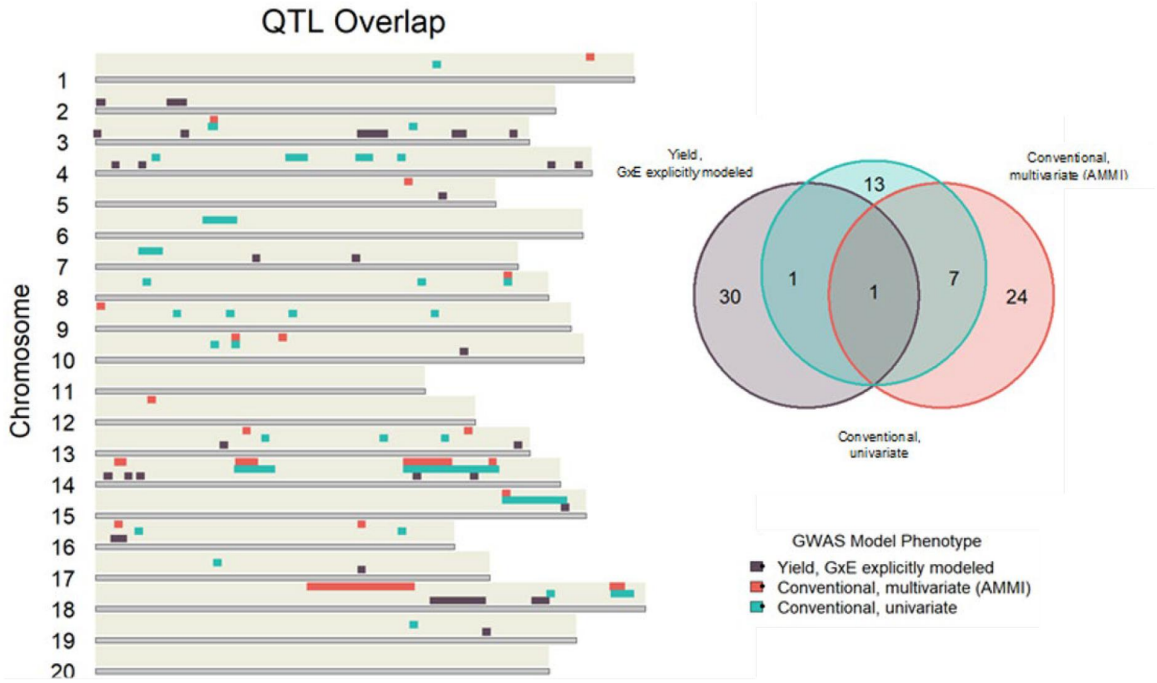


Figure 3.2: Independent QTL discovered using conventional measures as a GWAS phenotype share very little overlap with loci significant in the explicit GxE model.

Comparing the average yield variance explained by GxE at each of the QTL among approaches revealed that significant loci as reported by the explicit GxE model accounted for more GxE yield variance than either conventional approach. The largest number of QTL were discovered for the marker by year by location, and marker by location interaction effects, however the average effect size was lower than those QTL associated with additive main genetic effects and genotype by year interactions (Figure 3.3). These results suggest that using conventional yield stability estimates as a phenotype for GWAS is not a substitute for directly modeling GxE interactions.

Classification of GxE Interactions

If a locus is involved in creating changes in yield stability, it can often be seen as a difference in dispersion between the phenotypic distributions between alleles

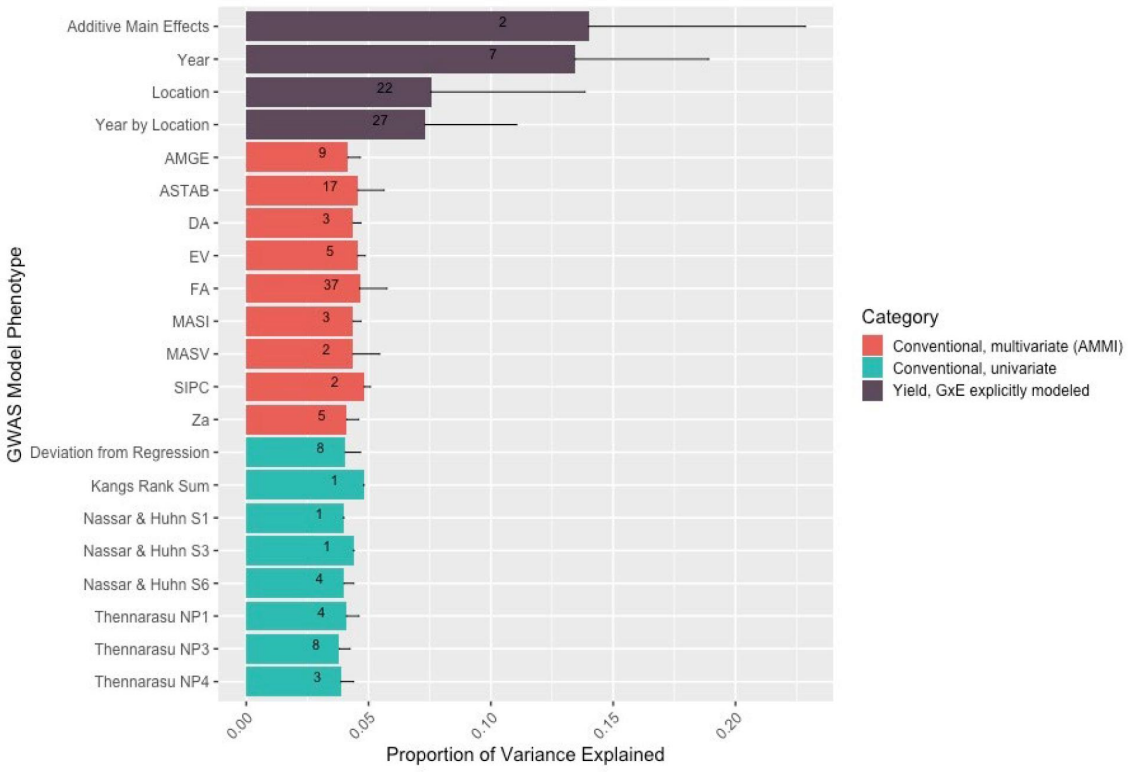


Figure 3.3: The number and variance explained by the QTL discovered in the explicit GxE model is greater than that discovered by GWAS models using either type of conventional measurement as a phenotype. Numbers within the bars represent the number of QTL discovered for that model/model level. The thin dark line from the top of the bar represents the standard deviation for yield variation explained among the QTL for that level.

(Rönnegård and Valdar, 2012). The distributions of adjusted phenotypes according to allele states at the QTL discovered by explicitly modeling GxE interactions for yield did not appear to follow this trend, with many of the distributions appearing to be nearly identical (Supplementary Figure 3.34). For example, the QTL at chromosome 3 physical position 9732856 shows a more characteristic difference in adjusted phenotype distributions that the QTL at chromosome 14 physical position 5032332 (Figure 3.4). Genotypes with allele "A" at the former QTL are less stable, as indicated by the flatter and wider distribution of adjusted phenotypes. When looking at all the adjusted phenotypes pooled across environments for QTL 14:5032332, we do not see an initial

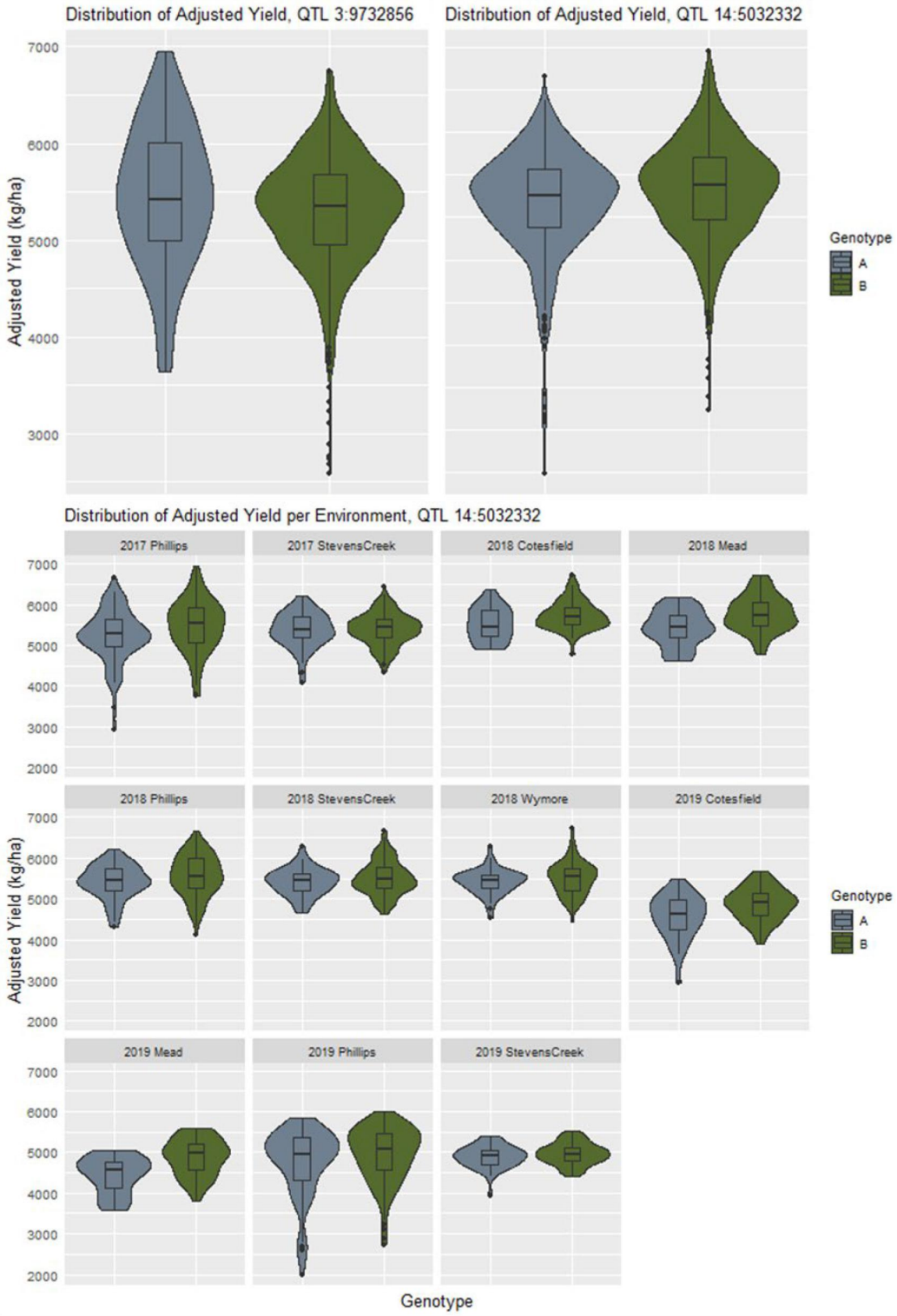


Figure 3.4: The contrast in distribution of adjusted yield between allelic states at the QTL on chromosome 3 indicates a difference yield stability as compared to the QTL on chromosome 14 which initially appears to be falsely associated. However, when examining the adjusted yield from a per environment basis, differences in mean and spread according to specific site combinations become more apparent.

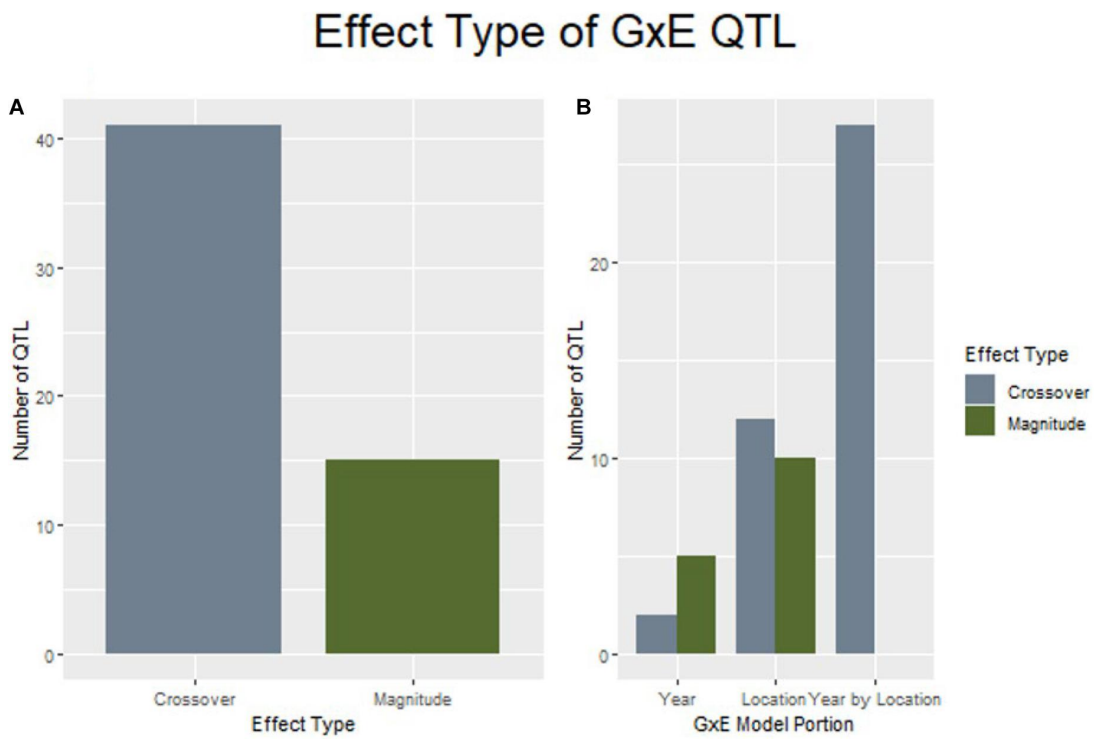


Figure 3.5: QTL of the crossover effect type are more prevalent in this study than magnitude changes (A), and are especially common in the marker by year by location interaction (B)

difference in distributions despite it being reported as a GxE QTL by the model. Upon examining the same data on a per environment basis, we see contrasts in mean and dispersion that are not consistent across year and location combinations (Figure 3.4). This is an indication of a crossover interaction occurring at this locus which has a canceling effect when assessing data combined across environments.

GxE interactions complicate the breeding process the most when they result in a crossover interaction that changes genotype rankings between growing environments. When examining the effect size and direction of QTL discovered by explicitly modeling GxE interactions for yield, 41 of the 56 loci were considered to produce crossover interactions. Further, it was noted that all QTL discovered through modeling marker by year by location were determined to be crossover interactions (Figure 3.5).

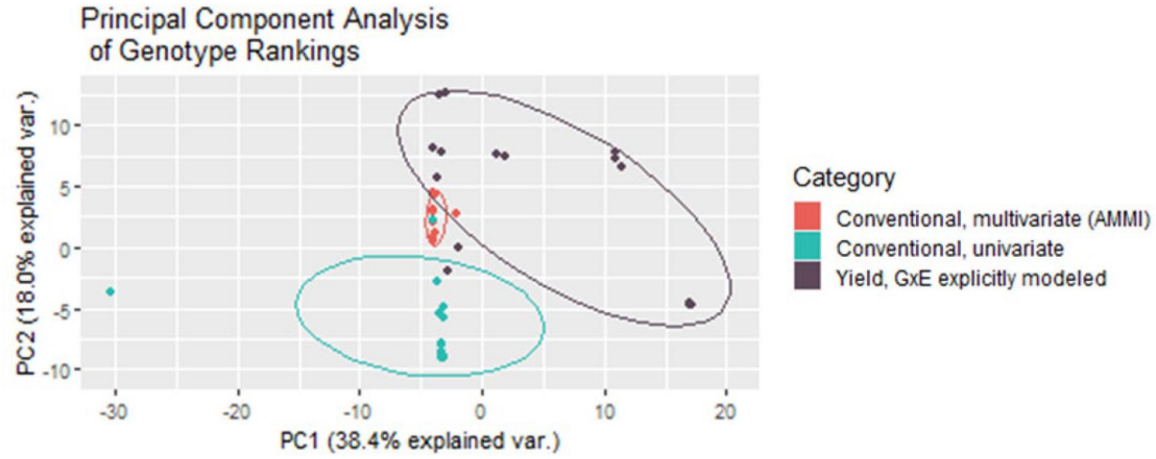


Figure 3.6: Multivariate conventional yield stability rankings group much tighter and closer to rankings generated from the BLUPs from fitting GxE interaction effects as random in the mixed model for yield.

Selection Rankings

Both multivariate and univariate conventional yield stability measures are used to create rankings that help breeders make selection decisions. Conducting a principal component analysis of these rankings in comparison to the rankings given by the BLUPs of the direct GxE model showed that multivariate conventional measures generated from AMMI modeling grouped very tightly together, and the closest with GxE BLUP rankings. Univariate yield stability measures also generally grouped together, intersecting far less with the GxE groupings than multivariate statistics (Figure 3.6). Together the first two principal components explained 56.4% of variability in the rankings indicating that these groupings still do not capture nearly half of the data variance.

Discussion

This analysis revealed that using conventional stability estimates to capture variation in GxE interactions for a genetic mapping study gave considerably different results from directly modeling GxE interactions when evaluating grain yield in a local soybean population. GxE interactions often heavily influence the per environment rankings of quantitative phenotypes such as grain yield, complicating the breeders' task of developing a stable variety. As a result, the modeling of phenotypic stability and identification of the involved genes has been the focus of many recent scientific studies (Bouchet et al., 2016; Xavier et al., 2018; Lozada and Carter, 2020). QTL affecting stability have been discovered using either a direct approach to modeling GxE interactions, or first calculating a yield stability "value" from the phenotypic data to then be used a phenotype in the GWAS. This study evaluates both of these approaches using the same association panel and demonstrated that the results were not interchangeable. To our knowledge this is the first study to directly compare QTL discovered using multiple methods of evaluating GxE interactions and yield stability in soybean.

Conventional yield stability estimates are a popular way to assess the influence GxE interactions without the burden of interpreting values for every testing environment. Such approaches are both appealing from a computational standpoint due to their simplicity, as well as pragmatic when discussing plasticity and stability within the scientific community. The historical implementations of traditional stability methods presented here are built upon variations on standard linear regressions – that is, a statistical model that only has fixed effects. When considering datasets from multi-

environment trials with missing observations, the results may be to varying extents, erroneous (Piepho, 1997). This is one potential rationale for the large difference in QTL results presented by this study compared to present research. However, unbalanced experimental designs are common in plant breeding programs both due to random loss within trials (pest/disease/weather damage/etc.) or by explicit design, and their accommodation should be prioritized. Mixed model analyses have been an effective tool in this regard. Recent studies have shown they can also be used to adapt traditional stability analyses, suggesting that conventional yield stability estimates can still provide useful insights from more complex field designs (Piepho, 1997, 1998; Meyer, 2009). Explicitly testing the consequences of using conventional yield stability statistics on unbalanced datasets may help refine the understanding of these results and future applications.

QTL confirmation testing would be an important step to validating the modeling approach used in their discovery. Due to unpredictable fluctuations in environment, the results of GxE interaction research are often difficult to replicate. With regards to QTL studies, this adds difficulty to the confirmation process. If effective, the results might not only be used to support the existence of a QTL affecting GxE interactions, but also serve to refine the understanding of what general effect it causes. This may be especially important when considering antagonistic crossover effects, which have the largest confounding effect on making selections in the plant breeding process. The results presented here indicate that the greater majority of GxE interactions in grain yield for soybean generate a crossover effect, with neither observed allele advantageous in all testing environments. In fact, in some cases opposite effects were nearly equal and

appeared to cancel each other out when observing the pooled data. Extensive testing of GxE QTL in new genetic backgrounds and new environments may reveal a shift in these QTL classifications and ultimately their utility to breeders. Detecting and accounting for crossover effects may be important to breeding decisions for local adaptations as well as separating out these noisy interactions from those that are more straightforward to incorporate.

Plant breeders often use a ranking approach to eliminate or progress genotypes in their breeding program (Sjoberg et al., 2020). Similarly, to the GWAS results, comparing the genotype rankings from each of the methods demonstrated the dissimilarity between conventional and explicit GxE modeling approaches. A multivariate approach (AMMI) which first starts with a model that retains some of the experimental design components best match the rankings from the GxE BLUPs themselves, further illustrating the importance of accounting for this additional variation in this experiment. The rigidity of software created for the purpose of computing traditional stability estimates limits the inclusion of non-genetic design components, such as replicate, location, blocking, and/or other spatial factors that may have been valuable in partitioning genetic variance from the phenotypic variance. Without prior adjustment, these artifacts have the potential to bias results and decrease selection accuracy for phenotypic stability.

Conclusion

This analysis determined that performing association mapping for grain yield GxE interactions in soybean using conventional yield stability measurement as a phenotype

provided nearly independent results from explicitly modeling marker by environment interactions in a mixed model for grain yield. While several QTL were discovered using both approaches, only one region overlapped between models and QTL discovered via conventional stability estimates explained far less GxE variance for grain yield. The results presented may have been influenced by the incomplete and unbalanced data structure utilized in the multi-environment trials, however this is a common occurrence in field trials and is often intentional to sample more genotypes and environments. Researchers and breeders interested in manipulating adaptation via GxE interactions need to consider the potential influences their modeling approach will have on their desired outcome.

Acknowledgments

Research reported in this publication was supported by the Nebraska Soybean Board project #1726. This work was completed utilizing the Holland Computing Center of the University of Nebraska, which receives support from the Nebraska Research Initiative. We would also like to acknowledge the members of the Graef lab for their assistance in preparation, planting, maintenance, and harvesting of field trials.

References

Falconer DS. Introduction to quantitative genetics. Pearson Education India; 1996.

Comstock RE, Moll RH. Genotype environment interactions. Statistical genetics and plant breeding. 1963.

Crossa J. Statistical Analyses of Multilocation Trials. In: Brady NC, editor. Advances in Agronomy [Internet]. Academic Press; 1990 [cited 2020 Oct 13]. p. 55–85. Available from: http://www.sciencedirect.com/science/article/pii/S0065211308608184

de Felipe, M., & Alvarez Prado, S. (2021). Has yield plasticity already been exploited by soybean breeding programmes in Argentina? Journal of Experimental Botany, 72(20), 7264–7273. https://doi.org/10.1093/jxb/erab347

Kang MS. Using Genotype-by-Environment Interaction for Crop Cultivar Development. In: Sparks DL, editor. Advances in Agronomy [Internet]. Academic Press; 1997 [cited 2021 Feb 12]. p. 199–252. Available from: https://www.sciencedirect.com/science/article/pii/S0065211308605696

El-Soda M, Malosetti M, Zwaan BJ, Koornneef M, Aarts MGM. Genotype × environment interaction QTL mapping in plants: lessons from Arabidopsis. Trends in Plant Science. 2014 Jun;19(6):390–8.

Falconer DS. The Problem of Environment and Selection. The American Naturalist. 1952 Sep 1;86(830):293–8.

van Eeuwijk FA, Bink MC, Chenu K, Chapman SC. Detection and use of QTL for complex traits in multiple environments. Current Opinion in Plant Biology. 2010 Apr;13(2):193–205.

Piepho HP. Statistical tests for QTL and QTL-by-environment effects in segregating populations derived from line crosses. Theor Appl Genet. 2005 Feb 1;110(3):561–6.

Chen X, Zhao F, Xu S. Mapping Environment-Specific Quantitative Trait Loci. Genetics. 2010 Nov 1;186(3):1053–66.

Becker HC, Léon J. Stability Analysis in Plant Breeding. Plant Breeding. 1988;101(1):1–23.

Lin CS, Binns MR, Lefkovitch LP. Stability Analysis: Where Do We Stand?1. Crop Science. 1986;26(5):cropsci1986.0011183X002600050012x.

Bouchet A-S, Laperche A, Bissuel-Belaygue C, Baron C, Morice J, Rousseau-Gueutin M, et al. Genetic basis of nitrogen use efficiency and yield stability across environments in winter rapeseed. 2016;

Xavier A, Jarquin D, Howard R, Ramasubramanian V, Specht JE, Graef GL, et al. Genome-wide analysis of grain yield stability and environmental interactions in a multiparental soybean population. G3: Genes, Genomes, Genetics. 2018 Feb;8(2):519–29.

Lozada DN, Carter AH. Insights into the Genetic Architecture of Phenotypic Stability Traits in Winter Wheat. Agronomy. 2020 Mar;10(3):368–368.

Piepho HP, Pillen K. Mixed modelling for QTL x environment interaction analysis. Euphytica. 2004;137(1):147–53.

Malosetti M, Ribaut J-M, van Eeuwijk FA. The statistical analysis of multi-environment data: modeling genotype-by-environment interaction and its genetic basis. Frontiers in Physiology. 2013 Mar 12;4:44.

Isik F, Holland J, Maltecca C. Genetic Data Analysis for Plant and Animal Breeding. Genetic Data Analysis for Plant and Animal Breeding. Springer International Publishing; 2017.

Wricke G. Evaluation Method for Recording Ecological Differences in Field Trials. Z Pflanzenzücht. 1962;47:92–6.

KEIM, P. A rapid protocol for isolating soybean DNA. Soybean Genet Newsl. 1988;15:150–2.

Happ MM, Wang H, Graef GL, Hyten DL. Generating high density, low cost genotype data in Soybean [Glycine max (L.) Merr.]. G3: Genes, Genomes, Genetics. 2019 Jul 1;9(7):2153–60.

Browning BL, Browning SR. Genotype Imputation with Millions of Reference Samples. The American Journal of Human Genetics. 2016 Jan 7;98(1):116–26.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics. 2007 Sep 1;81(3):559–75.

Korte A, Farlow A. The advantages and limitations of trait analysis with GWAS: A review. Vol. 9, Plant Methods. BioMed Central; 2013. p. 29.

Hayes B. Overview of Statistical Methods for Genome-Wide Association Studies (GWAS). In Humana Press, Totowa, NJ; 2013. p. 149–69.

Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: A Tool for Genome-wide Complex Trait Analysis. Am J Hum Genet. 2011 Jan 7;88(1):76–82.

Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, et al. Variance component model to account for sample structure in genome-wide association studies. Nature Genetics. 2010 Apr;42(4):348–54.

Butler DG, Cullis BR, Gilmour AR, Gogel BJ, Thompson R. ASReml-R Reference Manual Version 4 . 2017.

Team RC. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2019.

Villanueva RAM, Chen ZJ, Wickham H. ggplot2: Elegant Graphics for Data Analysis Using the Grammar of Graphics [Internet]. Springer-Verlag New York; 2016. 160 p. Available from: https://doi.org/10.1080/15366367.2019.1565254

Villanueva RAM, Chen ZJ, Wickham H. ggplot2: Elegant Graphics for Data Analysis Using the Grammar of Graphics. Springer-Verlag New York; 2016. 160–167 p.

Pour-Aboughadareh A, Yousefian M, Moradkhani H, Poczai P, Siddique KHM. STABILITYSOFT: A new online program to calculate parametric and non-parametric stability statistics for crop traits. Applications in Plant Sciences. 2019 Jan;7(1).

Pour-Aboughadareh A, Yousefian M, Moradkhani H, Poczai P, Siddique KHM. STABILITYSOFT: A new online program to calculate parametric and non-parametric stability statistics for crop traits. Applications in Plant Sciences. 2019 Jan 1;7(1).

Sabaghnia N, Sabaghpour SH, Dehghani H. The use of an AMMI model and its parameters to analyse yield stability in multi-environment trials. Journal of Agricultural Science. 2008 Oct;146(5):571–81.

Ezatollah Farshadfar, Nasrin Mahmodi, Anita Yaghotipoor. AMMI stability value and simultaneous estimation of yield and yield stability in bread wheat (Triticum aestivum L.). Australian Journal of Crop Science. 2011 Nov;

de Oliveira EJ, de Freitas JPX, de Jesus ON. AMMI analysis of the adaptability and yield stability of yellow passion fruit varieties. Scientia Agricola. 2014;71(2):139–45.

Abera W, van Rensburg JBJ, Labuschagne MT, Maartens H. Genotype-environment interactions and yield stability analyses of maize in Ethiopia. South African Journal of Plant and Soil. 2004;21(4):251–4.

Ajay BC, Aravind J, Abdul Fiyaz R. ammistability: Additive Main Effects and Multiplicative Interaction Model Stability Parameters. R package version 011 [Internet]. 2018; Available from: https://ajaygpb.github.io/ammistability/https://CRAN.R-project.org/package=ammistability.

Bernat Gel, Eduard Serra. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data | Bioinformatics | Oxford Academic. Bioinformatics. 2017 May 29;33(19).

Chen H, Boutros PC. VennDiagram: A package for the generation of highly-customizable Venn and Euler diagrams in R. BMC Bioinformatics. 2011 Jan 26;12(1):35.

Rönnegård L, Valdar W. Recent developments in statistical methods for detecting genetic loci affecting phenotypic variability. Vol. 13, BMC Genetics. BioMed Central; 2012. p. 63.

Piepho H-P. Analyzing Genotype-Environment Data by Mixed Models with Multiplicative Terms. Biometrics. 1997 Jun;53(2):761.

Piepho HP. Empirical best linear unbiased prediction in cultivar trials using factor-analytic variance-covariance structures. Theoretical and Applied Genetics. 1998 Jul;97(1–2):195–201.

Meyer K. Factor-analytic models for genotype × environment type problems and structured covariance matrices. Genetics Selection Evolution. 2009 Dec 30;41(1):21.

Endelman JB, Atlin GN, Beyene Y, Semagn K, Zhang X, Sorrells ME, et al. Optimal Design of Preliminary Yield Trials with Genome-Wide Markers. Crop Science. 2014 Jan 1;54(1):48–59.

Bernardo RN. Breeding for quantitative traits in plants. 2nd ed. Stemma Press; 2010.

CHAPTER FOUR: VARIABLE SELECTION PATTERNS ASSOCIATED WITH CONSTITUITIVE GENETIC AND GXE EFFECTS FOR GRAIN YIELD IN A LOCALLY ADAPTED SOYBEAN POPULATION

Abstract

It is widely accepted that artificial selection has reduced the overall amount of genetic variation present for most modern crops, including soybean [Glycine max (L.) Merr.] How traits with a complex genetic architecture, such as yield, may experience selection pressure differently in accordance with the diverse combination of main genetic and environmentally interactive components, is still unclear. For this study, we first identified the contribution of genomic windows to either main or genotype by environment (GxE) effects for soybean grain yield using 203 elite soybean genotypes tested in eleven environments. Tajima's D and pairwise weighted F_{st} values from wild, landrace, and three other elite populations were used in conjunction with these windowed estimates of phenotypic variance to compare the effect of selection among varying combinations of effect direction and prevalence among environments. Genomic regions with higher genetic diversity and lower divergence were significantly associated with higher GxE variance but not constitutive variance, indicating selection is relaxed at these interactive loci. Among GxE effect types we found prominent evidence for both negative and positive selection, and a markedly higher level of selection signatures at conditionally neutral loci. By obtaining whole genome resequencing data for our lines, modeling all marker effects simultaneously, and leveraging results from permutation datasets, we were

able to avoid some of the common pitfalls in analysis of this type and more accurately report on the complicated nature of response to selection for soybean yield.

Introduction

By recombining favorable genetic variation to produce new phenotypes, intensive artificial selection has led to remarkable gains in productivity within several crop species. A common concern is that this may lead to a reduction in total genetic diversity, since selective breeding only allows for a subset of the population to reproduce (Gepts, 2006). Preserving genetic variation is extremely important for future crop resilience and adaptation, especially considering rapidly changing climate conditions (Khoury et al., 2022). Loss of genetic diversity stemming from the domestication process and initial improvement of landraces into modern cultivars has been well documented in many crop species. Examinations at the individual breeding program level, however, have contested the assumption that modern selection always further erodes genetic variation (Bruce et al., 2019; Fu, 2015; van de Wouw et al., 2010; Wouw et al., 2010). Since producers mainly grow cultivars developed in regional breeding programs, studying how artificial selection influences a population's corresponding genetic variation for valuable traits on this scale is critical.

As selection changes an allele's prevalence within a population, the allele frequencies of the surrounding variation are similarly influenced due to linkage with the selected mutation (Barton, 1998; Smith & Haigh, 1974). Many of the standard statistics used to quantify genetic variation are based on summaries of the allele frequency spectrum, therefore providing a means of studying selection pressures among loci when

applied to genomic windows (Nielsen et al., 2005). Outliers in this context signal dramatic allele frequency shifts and/or genetic diversity loss, indicative of a directional selection causing a ”sweep” (Hermisson & Pennings, 2005). This footprint occurs when a beneficial mutation is fixed rapidly under strong selection, likely due to a large and sustained advantage it confers.

The overall genetic architecture of a trait has a direct influence on shaping genetic variation in response to selection pressures. Grain yield is a classic example of valuable quantitative trait controlled by many small effect loci. Under this scenario, there are many potential allelic combinations that can give rise to equally optimal phenotypes (Kremer & Le Corre, 2012). As a result, selection pressure at an individual locus may be relatively weak and the overall response occurs via subtle shift in allele frequencies at many loci, also referred to as polygenic adaptation (Barton & Keightley, 2022). In contrast to the selective sweep scenario, alleles contributing to polygenic adaptation are likely to never reach fixation. Or if they do, it occurs over such a long time period that recombination erodes the signature within linked variation (Chevin & Hospital, 2008; Lande, 1983).

Phenotypic variation can also be the result of contributions from loci whose effect varies between environments (Falconer, 1996). GxE interactions in the context of agriculture and plant breeding have often been thought of as a nuisance factor due to their transient nature. While low GxE may help improve the predictability of a genotype’s performance across a range of environments, it can also have negative consequences when it constrains cultivars from taking advantage of their environmental conditions. The genetic potential for a population to adapt to novel environments is becoming an

increasingly important attribute of plant breeding programs in the context of rapid climate change and a growing human population. For these reasons, it is important that we also consider the behavior of loci reflecting GxE interactions in response to selection pressure. In the presence of GxE interactions, the optimal genetic combinations fluctuate with space and time. Accordingly, selection pressure at individual GxE loci will also be variable, and therefore we hypothesize less intense.

The direction of an allele's effect, in combination with the proportion of selection environments it has an effect in, can be used to further categorize GxE interactions. It is logical to suggest then, that selection pressure at the GxE loci is dependent on the type of effect it has. When the environment determines whether the trait value is increased or decreased for an allele, this is termed antagonistic pleiotropy. Alternatively, the outcome may simply be a change in magnitude of the effect between environments, referred to as differential sensitivity. Conditional neutrality is the most extreme case of differential sensitivity and occurs when an allele only has an effect in one environment but not in the others (Des Marais et al., 2013; El-Soda et al., 2014). If two alleles with opposite conditional effects in different environments are in tight linkage with each other, they may appear as one quantitative trail loci (QTL) with an antagonistic pleiotropic effect. But the effect on surrounding variation would theoretically be the same as it elicits the same selection pressure (Anderson et al., 2013). While all types of GxE interactions are understood to play an important role in local polygenic adaptation, antagonistic pleiotropy is associated with the direct maintenance of genetic variation (Anderson et al., 2011, 2013). This is owed to the fact that no allele at the locus is 'optimal', and thus not selected for.

While weakened selection pressures allow for the continued survival of valuable genetic variation within a population, it also makes it harder to differentiate evidence of it from the background variation utilizing selection scan data alone (Kemper et al., 2014; Pritchard & Di Rienzo, 2010). By examining measures of genetic variation and divergence around genetic markers that explain high levels of phenotypic variance, subtle selection signatures may be realized (Berg & Coop, 2014). This approach is not without its challenges. Allele frequency is a confounding variable in this context, as loci with more intermediate allele frequencies will be calculated to contribute more to phenotypic variance while also being categorized as having higher genetic diversity and less evidence of selection (Josephs et al., 2017). Researchers can avoid false conclusions due to this feature by leveraging permutations of their association models to establish a null expectation of results (Josephs et al., 2015; Stanton-Geddes et al., 2013). Another important consideration is ascertainment bias in the marker dataset. QTL mapping studies often employ genotyping with panels of common variation. This misses a large proportion of rare and unique variation which may be critical to calculations of any number of selection scan metrics. If an investigator plans to examine selection metrics alongside results from a mapping population, it may be appropriate to perform moderate to deep whole genome sequencing of their lines in order to fully capture all the variation contained within it.

The primary goal of this study was to assess how selection has influenced constitutive and GxE variance for soybean grain yield in a locally adapted, elite breeding population, and how this may vary among effect complexities. We first identified genomic regions that contribute to grain yield using a previously published phenotypic

dataset after making several important improvements, including obtaining ~9.27x whole genome sequencing coverage for 203 genotypes and deploying a variance modeling approach that allowed us to simultaneously solve for all SNP effects. Using this resequencing data, we then calculated Tajima's D (Tajima 1989) and weighted pairwise F_{st} (Weir and Cockerham 1984) from a wild, landrace, and three other elite populations for evaluation in combination with the variance estimates, and the effect classifications we designated based upon them. Deploying a permutation testing approach, we were able to determine several important conclusions about variable selection pressures among GxE loci free from the major bias of allele frequency. Despite recent progress in statistical models allowing us to map marker by environment interactions, to our knowledge no investigations have included an examination of selection patterns at such loci in a modern breeding population (Malosetti et al., 2013; Piepho & Pillen, 2004; van Eeuwijk et al., 2010).

Materials and Methods

Panel Selection & Phenotype Collection

This study utilized 203 experimental lines from the University of Nebraska-Lincoln soybean breeding program that were previously selected to explore and compare mapping methodologies related to GxE interactions across the lines' target growing region in eastern Nebraska. Lines selected represented a range of both average yield and yield stability according to Wricke's ecovalence, selected from a pool of genotypes that had existing yield data from 2013, 2014, and 2015 multi-environment yield trials. Five

highly productive testing sites in eastern Nebraska were selected for multi environment yield trials which took place over three years (2017, 2018, 2019). Yield trials were grown in an augmented incomplete randomized block design at each site and included three replicates per site. More details about the study population and yield trials can be found in Happ et. al (2021).

DNA Extraction and Whole Genome Sequencing

DNA was isolated from lyophilized leaf tissue collected from twenty plants per genotype using a CTAB base extraction method scaled down for a ninety-six well plate by dividing all reagent volumes by forty (KEIM & P., 1988). Extracted DNA was normalized to a concentration of twenty ng/µL and sequencing libraries were constructed using the iGenomx RIPTIDE High Throughput Rapid Library Prep Kit (Twist Bioscience, South San Francisco, CA 94080). Libraries were quantified using the KAPA Library Quantification Kit for Illumina platforms (Roche Sequencing Solutions, Santa Clara, CA 95050) and then sequenced on an Illumina NovaSeq 6000 instrument (Illumina Hayward, Hayward, CA 94545) by the genome sequencing facility at the University of Kansas Medical Center.

Sequence Mapping and Variant Discovery

To assess historical divergence patterns, we obtained raw sequence data from 1,318 lines used in previous studies that included wild, landrace, and elite soybean genotypes across many geographies for comparison to our local breeding population (Fang et al., 2017; Torkamaneh et al. 2018.; Valliyodan et al., 2016; Zhou et al., 2015). Raw reads were filtered for adapter sequencing contamination, base quality, and

truncated reads using Trimmomatic v0.39 (Bolger et al., 2014). Bowtie v2.4 (Langmead & Salzberg, 2012) was used to map reads to the Glycine max Wm.82.a4.v1 reference genome (Valliyodan et al., 2019) using the "very sensitive" setting. Picard v2.9 was used to add sample names to the read groups in the resulting .bam file. To avoid the computational cost of holding over 1,000 .bam files open in memory at once, individual .bam files were combined into one large .bam file using Sambamba v0.8.1 (Tarasov et al., 2015). Then, the combined .bam file was split by chromosome using Sambamba so that variant discovery could proceed in parallel per chromosome without loading all the mapped reads associated with the other nineteen chromosomes into memory. Finally, to increase accuracy, duplicate reads were marked in the mapped .bam files with Sambamba. In summary, the input for our variant caller was one of twenty .bam files that contained all the reads for a single chromosome across all sample genotypes. Variants discovery was then performed per chromosome, and within each chromosome in parallel chunks, using Freebayes version 1.3 (Garrison & Marth, 2012). To minimize the variation in computational time, chunk sizes were determined with respect to sequencing coverage as recommended in the Freebayes manual. To achieve this, coverage estimates were generated from the combined .bam file using Samtools v1.6, and then the "coverage_to_regions.py" script provided with the Freebayes software was used to determine the physical bounds of chunks with an even amount of sequencing coverage. To avoid missing variation on the edges of each chunk, chunks were then adjusted to have 200 bp of overlap with the both the preceding and/or following chunk. Freebayes was ran with the standard filters, as well as a requiring a minimum site coverage of three. As suggested for increased computational performance, only the four best alleles were

evaluated at each SNP variant site. The initial results of the variant calling contained 14,769,136 sites (both InDels and SNPs) with a phred-scaled quality score greater than forty. Further quality filtering was performed with the vcffilter utility of vcflib, by employing the following recommended hard filters: DP > 5, QUAL/AO > 10, QUAL/DP > 2, SAF > 0 & SAR > 0, RPR > 1 & RPL > 1, MQM / MQMR > 0.9 & MQM / MQMR < 1.05 PAIRED > 0.05 & PAIREDR > 0.05 & PAIREDR / PAIRED < 1.75 & PAIREDR / PAIRED > 0.25 | PAIRED < 0.05 & PAIREDR < 0.05. This left a remaining 10,829,817 variants. Samples and variants with over fifty percent missing data were filtered next, which removed eleven genotypes and left 10,645,412 variants over 1,276 samples. To assess any possibly cross contamination of samples, we computed the inbreeding coefficient (F) using plink 1.9. Thirty-seven samples with an F statistic of under 0.9 were removed, as soybean is an inbred crop and near complete homozygosity from these lines was expected. At this stage, indels were also removed resulting in a final SNP dataset of 7,137,085 variants across 1,237 genotypes.

Calculation of Genetic Diversity and Divergence

Signatures of artificial selection can be recognized in genetic variation by using measures of genetic diversity and divergence from other populations. To assess the former, ANGSD version 0.934/0.935 was used to calculate Tajima's D per chromosome using the mapped .bam files from bowtie2 (Korneliussen et al., 2014). Filtering was performed by enabling BAQ computation (Li, 2011), setting the coefficient for downgrading reads with excessive mismatches to fifty, setting a minimum mapping quality of thirty, and a minimum base quality of twenty. Respectively, these filters were

implemented from within the ANGSD command with the flags "-baq 1 -C 50 -minMapQ 30 -minQ 20". Sliding window estimates were performed with a window size of 20,000 bp and step size of 5,000 bp using the thetaStat utility of the ANGSD package. To assess divergence, we calculated pairwise weighted F_{st} using vcftools version 0.1 from one wild, one landrace, and three elite soybean populations using biallelic SNP data discovered in whole genome sequence data (Danecek et al., 2011).

Variance Contribution of Main Genetic and GxE Effects

Models which fit one marker at a time are subject to inflation biases, but fitting all markers directly in a single model was not computationally feasible. To handle these challenges, we implemented a two-stage approach, where BLUPs for the genotype and GxE effects were extracted first, and then fit as the phenotype in a GBLUP model. Prediction values obtained via GBLUP modeling have been proven equivalent to those from SNP-BLUP, indicating one can then back-solve for individual marker values from these solutions. (Goddard, 2009; Strandén & Garrick, 2009). To achieve this, we utilized ASREML-R 4 since it provides a wide range of options for modeling both fixed and random effects, as well as the option to include user defined variance structures. Equation 4.1 describes the model used to extract the best linear unbiased predictors (BLUPs) for main and GxE effects for grain yield in our Nebraska population:

$$(4.1) \quad y = X\beta + Z\alpha + e$$

where y is the vector of raw yield estimates assumed to be normally distributed, X is the design matrix of fixed effects including the intercept and maturity grouping, β is the vector of fixed effect coefficients, Z is the incidence matrix of random effects including genotype and GxE effects, α is the vector of random effect coefficients, and e is the vector of residuals. Residuals were specified as a direct sum of separate variance matrices for each environmental level. Each environmental "level" for the residual is defined as the unique year and location combination. . Each environmental "level" for the residual is defined as the unique year and location combination. BLUPs were extracted from these model results, and were then used to fit a GLBUP model described in equation 4.2:

$$(4.2) \quad y = \mathrm{X}\beta + \mathrm{Z}\alpha + \mathrm{e}, \mathrm{var}(\alpha) = G\sigma^2_{\alpha'}$$

where y is a vector of BLUPs for either the main genetic effects or GxE effects for a single environment, X is the design matrix of fixed effects including the intercept, β is the vector of fixed effect coefficients, Z is the incidence matrix of random effects including the genotype, α is the vector of random effect coefficients, e is the vector of residual, and G is the scaled genomic relationship. . To compute G, we used genome-wide SNP information and constructed it according to equation 4.3:

$$(4.3) \quad G = \frac{MM'}{N_{SNP}}$$

where M is the n x N_{SNP} centered and scaled marker matrix (n being 203 genotyped individuals and N_{SNP} being the 7,137,085 SNP markers). After fitting the GBLUP model, we backsolved for individual SNP effects and variances, a method first outlined by

VanRaden (2008), which has since been further augmented and widely used in animal applications. Equations 4.4 and 4.5 below describe this approach: (Gualdrón Duarte et al., 2014; Legarra et al., 2018; VanRaden, 2008).

$$(4.4) \quad \hat{u} = MM^{-1}\hat{\alpha}$$

$$(4.5) \quad \mathrm{var}(\hat{u}) = \mathrm{M}'\mathrm{G}^{-1}\,(\mathrm{G}\sigma_{\alpha}^{2} - C^{aa})\mathrm{G}^{-1}\mathrm{M}$$

where $\hat{u}$ is the vector of SNP effects to be solved, $\hat{\alpha}$ is the vector of solutions from the GBLUP evaluation in equation 2, and C^{aa} is a matrix calculated according to equation 4.6:

$$(4.6) \quad C^{aa} = \sigma_e^2(I + G^{-1}\lambda), \lambda = \frac{\sigma_e^2}{\sigma_{\alpha}^2}$$

Using these results, the variance explained per marker was solved using equation 4.7:

$$(4.7) \quad VE_i = \frac{2\hat{u}_i^{\,2}(MAF_i)(1 - MAF_i)}{2\hat{u}_i^{\,2}(MAF_i)(1 - MAF_i) + (\mathrm{var}(\hat{a}_i)^2)2n(MAF_i)(1 - MAF_i)}$$

where for the ith marker, VE represents the total variance explained, û represents the predicted marker effect, MAF represents the marker's minor allele frequency, and n represents the total number of genotyped individuals. After performing these individual SNP calculations, measures were computed along sliding windows of 20,000 bp with a step size of 5,000 bp summing the results for all SNPs within each window. Permutation testing was performed by shuffling the grain yield phenotypes with respect to the

genotypes and then fitting the modeling approach outlined in equations 4.1-4.7. One hundred permutations were used for establishing a null distribution of the grain yield variance explained by GxE and main genetic effects.

Significance and Classification of Effect Types

Since the purpose of this study was not to identify traditional QTLs but perform a comparison of selection statistics to phenotypic variance explained by the underlying genetics, we designated windows as statistically significant for further comparison if the variance explained was more than three standard deviations higher than the mean. From this point, we classified significant GxE windows as either conditionally neutral, antagonistic, or differentially sensitive using the number of environments it reached significance in, as well as sign of the significantly estimated effects in each environment. Calculations of the overall contribution of effect types to the genetic variance were performed in such a way that for significant overlapping windows, the same SNPs were not included in the calculation multiple times.

Visualizations

All visualizations were performed in R/4.1 using the “ggplot2” package in conjunction with “ggpubr”, “RColorBrewer”, and “ggtext” (Alboukadel Kassambara, 2022; Claus O. Wilke & Brenton M. Wiernik, 2022; Erich Neuwirth, 2022; R Core Team, 2021). The package “data.table” was used for reading in datasets. For basic data manipulations, “matrixStats”, ““tidyr”, and “dplyr” were leveraged (Hadley Wickam et al., 2022; Hadley Wickam & Maximilian Girlich, 2022; Henrik Bengtsson, 2021).

Data Availability

Raw sequencing data directly generated by this project for use in creating the study panel has been submitted to the NCBI Short Read Archive under submission numbers SUB13622898 & SUB13716130. Supplementary figures and tables can be found in "Supplementary Figures and Tables" section of the appendix. Sample metadata and sequencing depth of the lines used in the study can be found in the supplementary data file 4.1.

Results

DNA Sequencing and SNP Discovery

Variant discovery was performed in order to carry out analyses that would allow us to determine how genomic regions within our local Nebraska population contributed to grain yield, as well as the extent of divergence from other populations. In addition to sequencing the improved Nebraska cultivars, we obtained raw sequence data from several wild, landrace, and other improved cultivar populations via the NCBI short read archive with the intent to calculate F_{st} between the populations using biallelic SNP data. Improved cultivars were further divided based upon their origin into East Asian, Canadian, and United States (excluding our local Nebraska genotypes) populations. The average sequencing coverage for all genotypes was 13.33, with a standard deviation of 8.27. Per population, the improved United States cultivars had the highest average sequencing coverage of 22.71, and the improved Nebraska cultivars had the lowest average sequencing depth of 9.27. After all quality control steps, the final marker panel

for F_{st} calculations consisted of 7,137,085 high quality, homozygous, biallelic SNPs across 1,237 individual genotypes with 6.32% of marker genotype calls missing. Per marker missing data rates ranged from 0 to 50%, with a standard deviation of 7.81%. Per individual missing data rates ranged from 0.26 to 46.68% with a standard deviation of 6.38%. 2,640,097 of these SNPs were polymorphic within our Nebraska improved cultivar population and were subset for use in calculating the contribution of windowed genomic regions to grain yield variance (Supplementary Table 4.1). This variation is unique to our population and therefore free from ascertainment biases associated with using a common genotyping panel.

Partitioning Main and GxE Variance and Effects on Grain Yield

Spanning five sites and eleven unique environments, 213 soybean experimental lines that were initially tested from the University of Nebraska – Lincoln breeding program were evaluated for multi environment yield trials in an augmented incomplete randomized block design. Grain yield over the course of these trials ranged from 2,162.74 to 7,080.70 kg/ha, with an average of 4,976.41 kg/ha and standard deviation of 810.34 kg/ha, and were distributed approximately normally per location (Happ et al., 2021). We assessed contributors to grain yield variance using a multi-section mixed effects model and found that environmental effects explained the most variance of any component, contributing 51.61 to 63.30% to the total variance of the raw yield esimates. Genotype effect explained another 8.32 to 10.21%. GxE interactions contributed between 10.99 to 13.48%, a result which highlights the overall importance of their contribution to grain yield variance for this population (Supplementary Table 4.2).

To gain perspective on the contributions of specific DNA regions to soybean grain yield, the proportion of variance explained and per environment effects were calculated on sliding genomic windows of 20,000 bp with 203 lines that passed genotyping filters. The average proportion of variance explained per window was approximately the same for main and GxE effects at 1.00e-5 and 1.01e-5, respectively. However, the median value for GxE effects and the overall distribution was skewed distinctly higher than the main genetic effects, suggesting greater contribution to overall yield variance. Interactions at the Phillips sites constituted the three highest values for proportion of variance explained of any environments, as well as broadest distributions for proportion of variance explained (Supplementary Table 4.3, Supplementary Figure 4.1A). This did not match the top three widest ranges of effects per environment, which were observed at the 2019 Cotesfield, 2018 Phillips, and 2018 Wymore sites (Supplementary Table 4.4, Supplementary Figure 4.1B).

GxE interactions can be further classified by the frequency it is observed among the selection environments, and the direction of those effects. We selected regions that explained more than three standard deviations above the average for yield variance as significant. Between main and GxE effects, this amounted to 15,528 windows or 15.86% of the total number of windows analyzed. Of these, we found that 11,299 were categorized as conditionally neutral. (Figure 4.1A). Conditionally neutral effects constituted approximately a third of the total measurable genetic variance (Figure 4.2A),

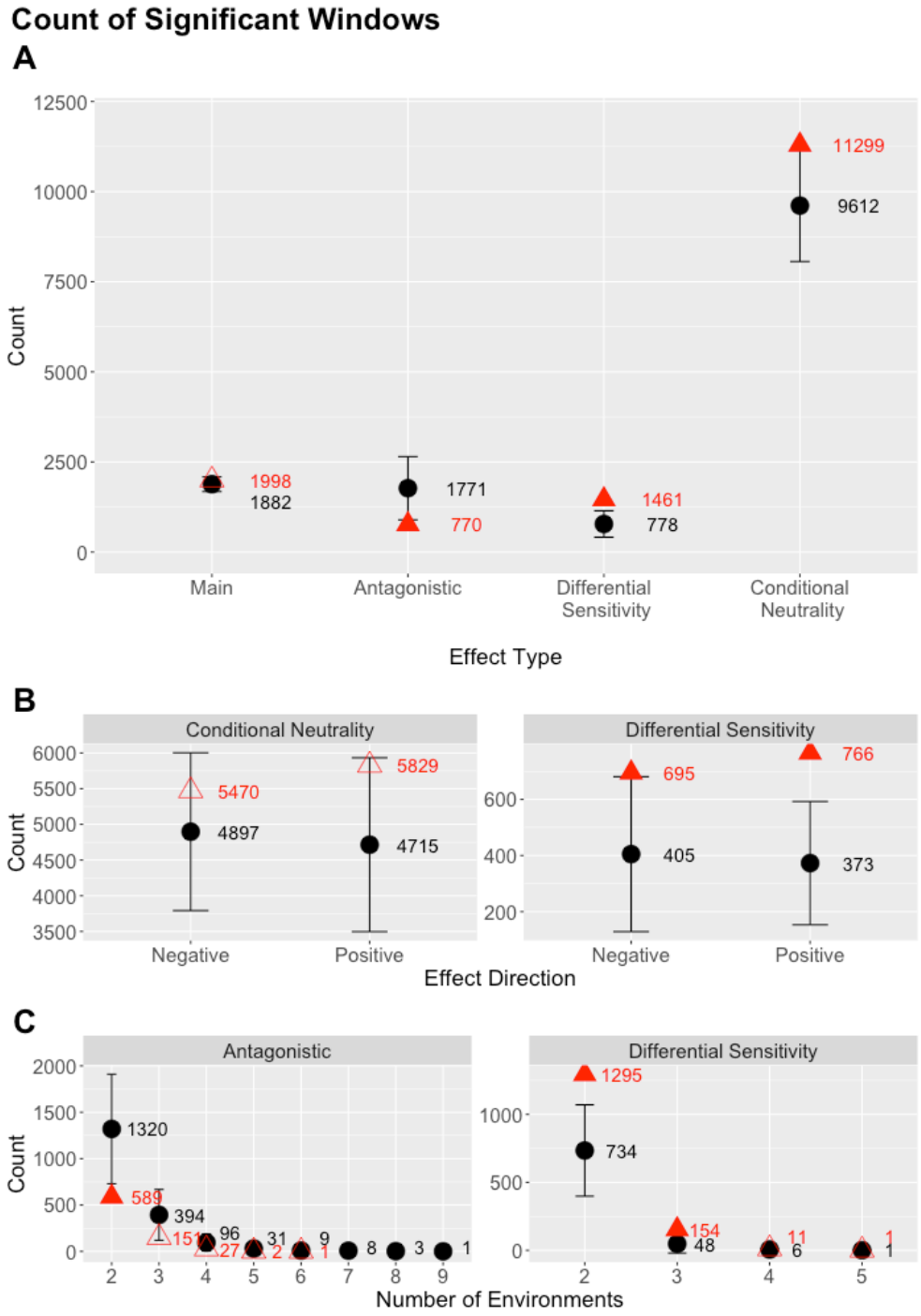


Figure 4.1: Comparisons of the number of significant windows among various levels of effect type (A), direction (B), and prevalence among environments (C). Black points represent the average number of windows sigificant for that effect classification based on the permutations, and the lines extending from those points the standard deviation. Red triangles represent the number of windows significant for that effect classification in the observed data, and a solid fill of the triangle reprsents that value fell outside the error bar for the permutation.

however had the lowest average proportion of variance explained per genomic window (Figure 4.2D). Interestingly, the majority of these loci were detected as having an effect in the Phillips location, specifically an overwhelming amount in 2019 (Supplementary Figure 4.2). This was also our highest average yielding location, but also our most variable, suggesting GxE effects play a role in increasing overall yield by allowing a genotype to take advantage of positive conditions. Main genetic effects were the next most prevalent effect type (Figure 4.1A) and explained the second highest amount of total genetic variance. (Figure 4.2A). When also considering they displayed the highest average proportion of variance explained per individual window, constitutive effects still clearly play a very important role in the determination of soybean grain yield and cannot be discounted (Figure 4.2D). Differentially sensitive effect patterns were observed almost as frequently as main genetic effects and antagonistic effects were the rarest (Figure 4.1A), though effect sizes were similar between the two (Figure 4.2D). When categorizing these windows by the number of environments they had an effect in, we observed fewer windows and a sharp decrease in the total genetic variance explained as the number of environments increased (Figure 4.1C & Figure 4.2C). However, the average proportion of genetic variance explained per window increased as the number of significant environments increased (Figure 4.2F). These data suggest that soybean grain yield is mostly determined by the accumulation of many small effect loci, that are often only detectable in a few or single environments.

Comparison of these counts to those from the permutations can shed light on the mechanisms of selection present for soybean grain yield. The number of significant

windows for all GxE effect types diverged significantly from the null expectation set by the permutations, albeit in different directions, indicating a mix of positive and negative selection dependent on effect type. Further examination also finds that loci with positive effects are more common than negative (Figure 4.1B) and explain more variance (Figure 4.2B). This supports the conclusion that artificial selection has significantly enriched the GxE landscape for grain yield with beneficial alleles. Additionally, average variance explained between positive and negative windows did not appreciably differ (Figure 4.2E). Reinforcing the narrative of negative selection at antagonistic loci, we observed effects in up to six environments in a single window in our real data, while the upper

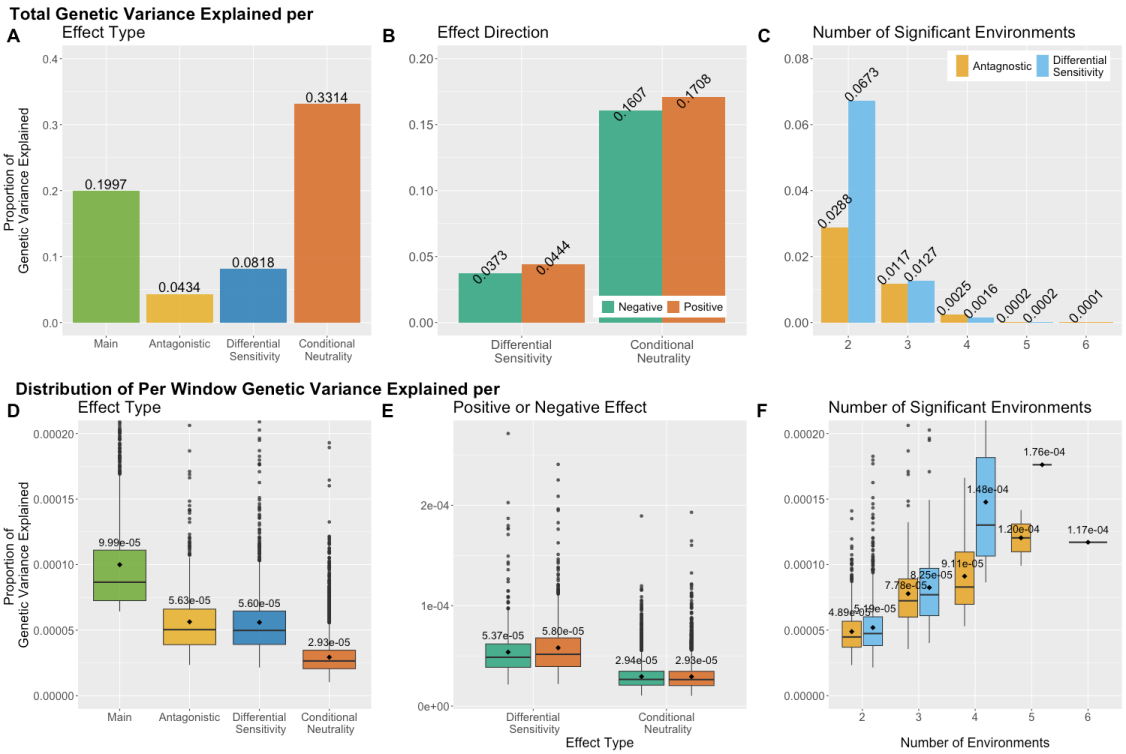


Figure 4.2: When summed together conditionally neutral loci constitute the greatest proportion of genetic variance explained (A), but per window explain the least (D). Positive effects explained a greater proportion of the total genetic variance in both differential sensitivity, and conditional neutrality (B), although the average per window variance explained was only slighlty higher for positive differential sensitvity effects (E). Windows affecting increasing numbers of environments make up very small proportion of the total genetic variance for grain yield, but explain more variance per window on average (C,F). Main genetic effects explain the second most amount of genetic variance for grain yield as a whole, and have the largest effect size per window (A,D).For the boxplots, the diamond shape and text represent the mean.

limit of the permutations suggests we should expect up to nine (Figure 4.1C). As the overall number and contribution of antagonistic effects is small in comparison to others, we interpret these data to mean that GxE for soybean grain yield in our population is generally under positive selection.

Grain Yield Variance Explained at Regions of Low/High Tajima's D and F_{st}

Quantifying changes in the genetic variation both within a population, and between populations, can help determine how selection has shaped a population's overall genomic landscape. To examine this, we calculated Tajima's D, as well as pairwise weighted F_{st} values with five other populations, on sliding 20,000 bp windows across the entire genome. Tajima's D values ranged from -2.90 to 4.04, with an average of -1.52 and were skewed towards more negative values indicating a strong general presence of directional selection. (Supplementary Figure 4.3A). Weighted pairwise F_{st} estimates for the entire genome followed the expected trend, with the highest value found in comparing our improved Nebraska cultivars to the wild population, and the lowest in comparison to the other United States improved cultivars. This trend was consistent with examinations of weighted F_{st} values on sliding windows (Supplementary Table 4.5). Distributions of windowed weighted F_{st} values were approximately normal for comparisons to the wild cultivars and skewed towards 0 for all other comparisons (Supplementary Figure 4.3B).

Measurements of genetic diversity and divergence on windows along the genomic space also provides a means by which to compare the level of artificial selection experienced between different parts of the genome. To explore the effect of artificial selection on loci with either main or GxE effects, we compared the amount of grain yield

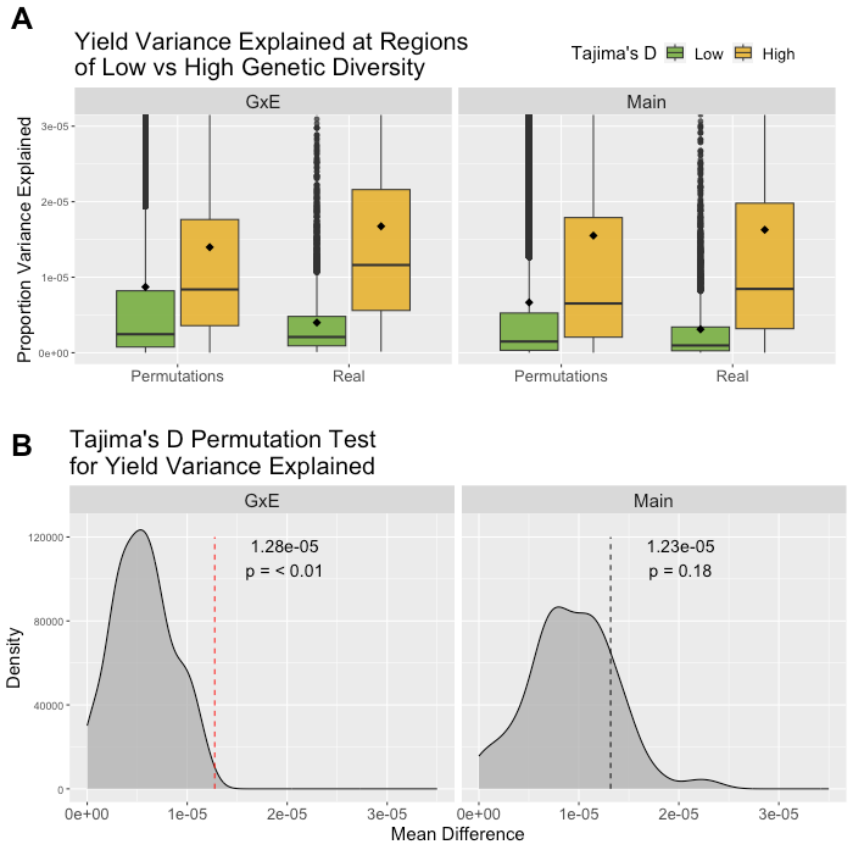


Figure 4.3: GxE effects, but not main genetic effects, explain significantly more grain yield variance on average in windows that make up the top 5% of Tajima's D values. Diamonds in the boxplots in (A) represent the subset mean. Grey distributions in (B) represent the mean differences between Tajima's D subsets in the permutations, and the dashed line represents the true mean difference between Tajima's D subsets in the real data.

variance attributable to each between the lowest and highest 5% of the sliding windows for Tajima's D and weighted pairwise F_{st} values. However, comparisons of variance explained between these subsets alone were likely to be biased by differences in allele frequencies between subsets. That is, the same allele found at a higher frequency would be calculated to contribute more to variance than if it was found at a lower frequency. Indeed, for low and high subsets of both Tajima's D and pairwise weighted F_{st} in this study, differences in minor allele frequency distributions were apparent (Supplementary Figure 4.3B).

To account for this bias, we utilized a permutation testing approach that would allow us to separate true differences in contribution to grain yield from those that appear solely to unequal allele frequency distributions between subsets (see Materials and Methods). Results of the permutation testing showed that the grain yield variance explained by GxE effects was significantly larger than expected ($p < 0.01$) at regions of high Tajima's D values, but this was not true for main genetic effects (Figure 4.3A-B). The same analysis performed for the weighted F_{st} calculations revealed that for GxE effects there was a significantly increased amount of grain yield variance explained by regions of high divergence from the wild population. However, there was a significant decrease in amount of grain yield variance explained by regions of high divergence from Landrace and East Asia populations (Figure 4.4A-B). When examining the main genetic effects, only a significant decrease in grain yield variance was found in regions of divergence from the East Asia population. (Figure 4.4C-D). The pervasive increased level of grain yield variance explained by GxE, but not main genetic effects, at regions of high genetic diversity and low divergence was consistent with a pattern of relaxed selection in regions associated with environment interactions.

Tajima's D and F_{st} in Relation to Effect Type, Direction, and Number of Significant Environments

The effect pattern displayed by a locus can influence the selection it experiences. To assess this, we contrasted the Tajima's D and weighted pairwise F_{st} values in regions significantly contributing to grain yield variance by effect type classification, as well as by effect direction and number of significant environments for the relevant classifications. Compared to the permutation data, all GxE effect types displayed a higher

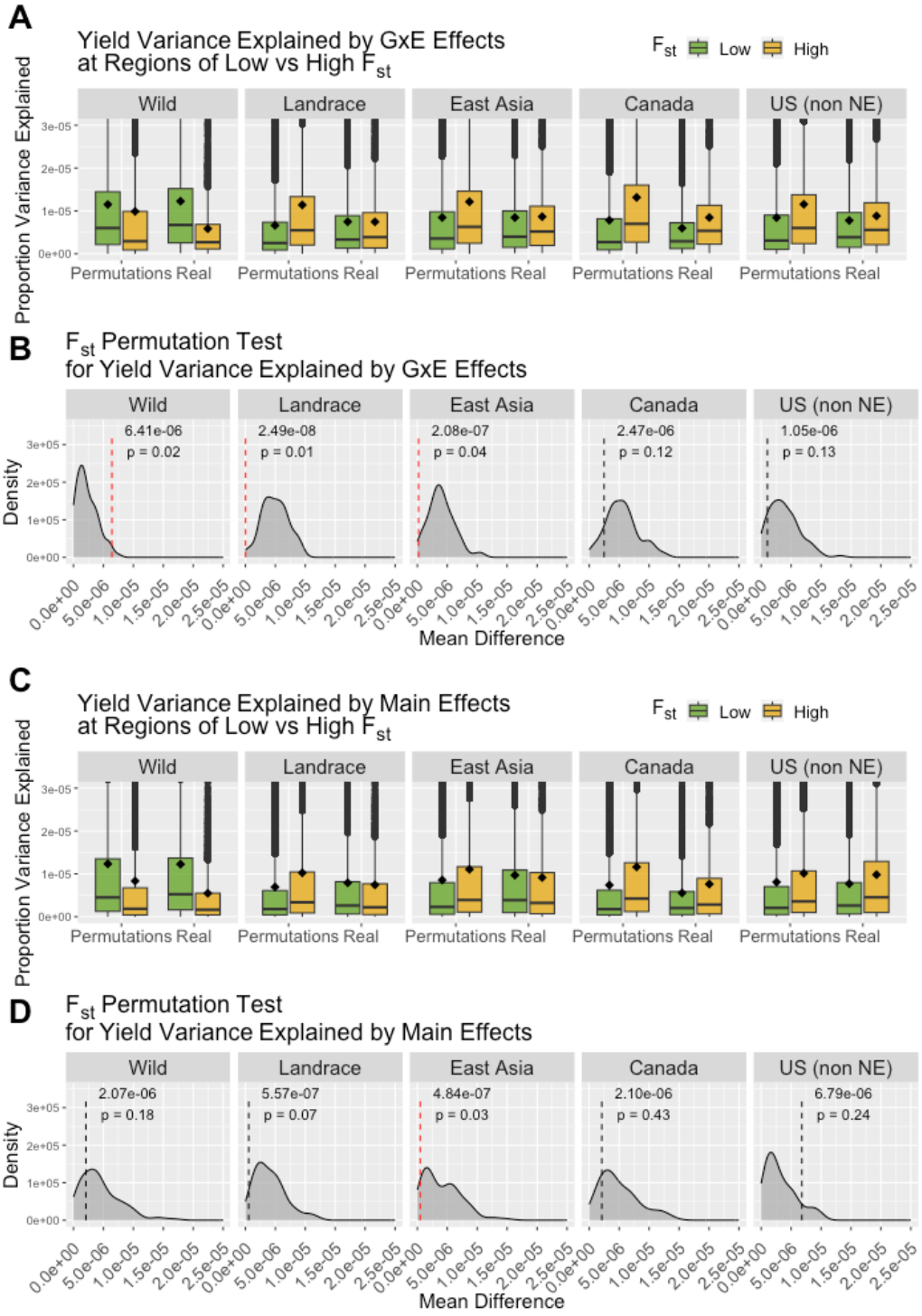


Figure 4.4:(A-B) Significantly less GxE variance than expected is observed for regions of high divergence from Landrace and East Asian populations, as opposed to comparison to the wild population. (C-D) For main genetic effects, less variance was observed at high divergence from the East Asian population. Diamonds in the boxplots in (A & C) represent the subset mean. Grey distributions in (B & D) represent the mean differences between F_{st} subsets in the permutations, and the dashed line represents the true mean difference between F_{st} subsets in the real data.

average Tajima's D and lower F_{st} value than expected, while main genetic effects only displayed a lower F_{st} value. This was generally consistent with the comparison of variance explained by each effect type at regions of contrasting diversity and divergence in the previous section. Among GxE effects, conditionally neutral effects consistently had the lowest Tajima's D and lowest F_{st}, a pattern that was not expected based on the permutation data. (Figure 4.5A and 4.6A). In fact, weighted F_{st} at conditionally neutral loci was higher than the genome average for four out of five of the comparisons. These results suggest that conditionally neutral loci have been under greater selection pressure than other types of GxE loci. Further parsing the data by effect direction revealed that

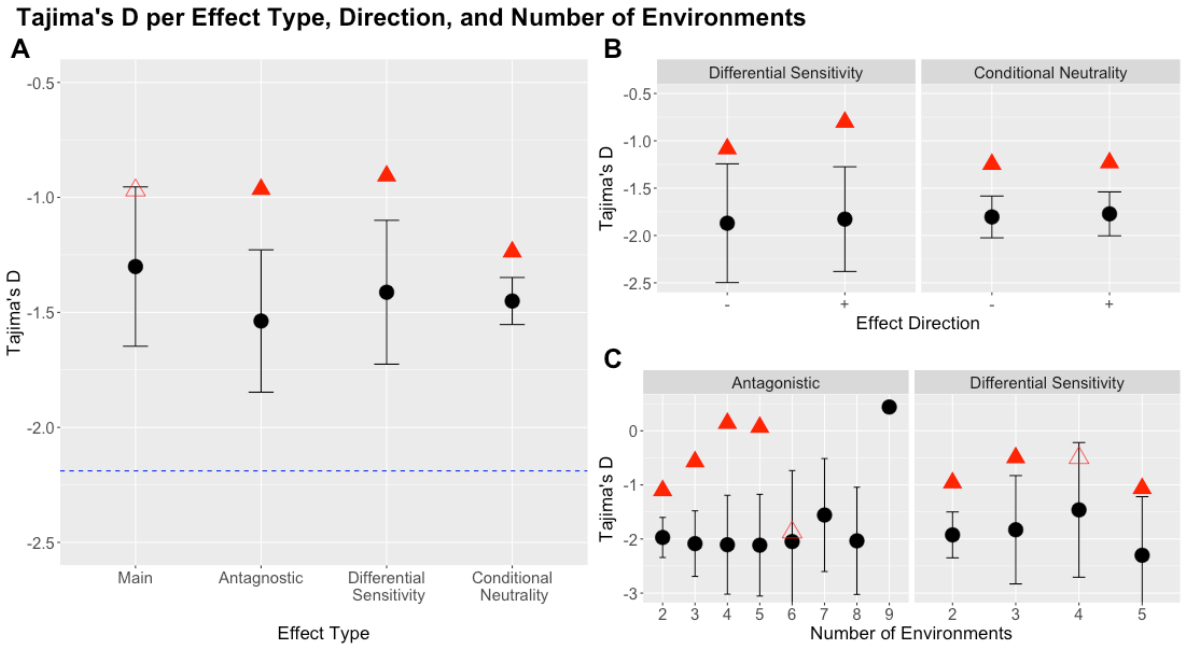


Figure 4.5: (A) While the average Tajima's D per effect type is higher than the genome wide median for all effect type classifications, conditionally neutral loci has a markedly lower average value than other effects, a difference not predicted by the permutations. (B) Tajima's D was also noticeably lower for negative differentially sensitive effects. (C) and roughly increases with an increasing number of environments a window effects. For (A-C), black points represent the average number of windows sigificant for that effect classification based on the permutations, and the lines extending from those points the standard deviation. Red triangles represent the number of windows significant for that effect classification in the observed data, and a solid fill of the triangle reprsents that value fell outside the error bar of the permutations. The dashed line in (A) represents the genome wide average Tajima's D.

:

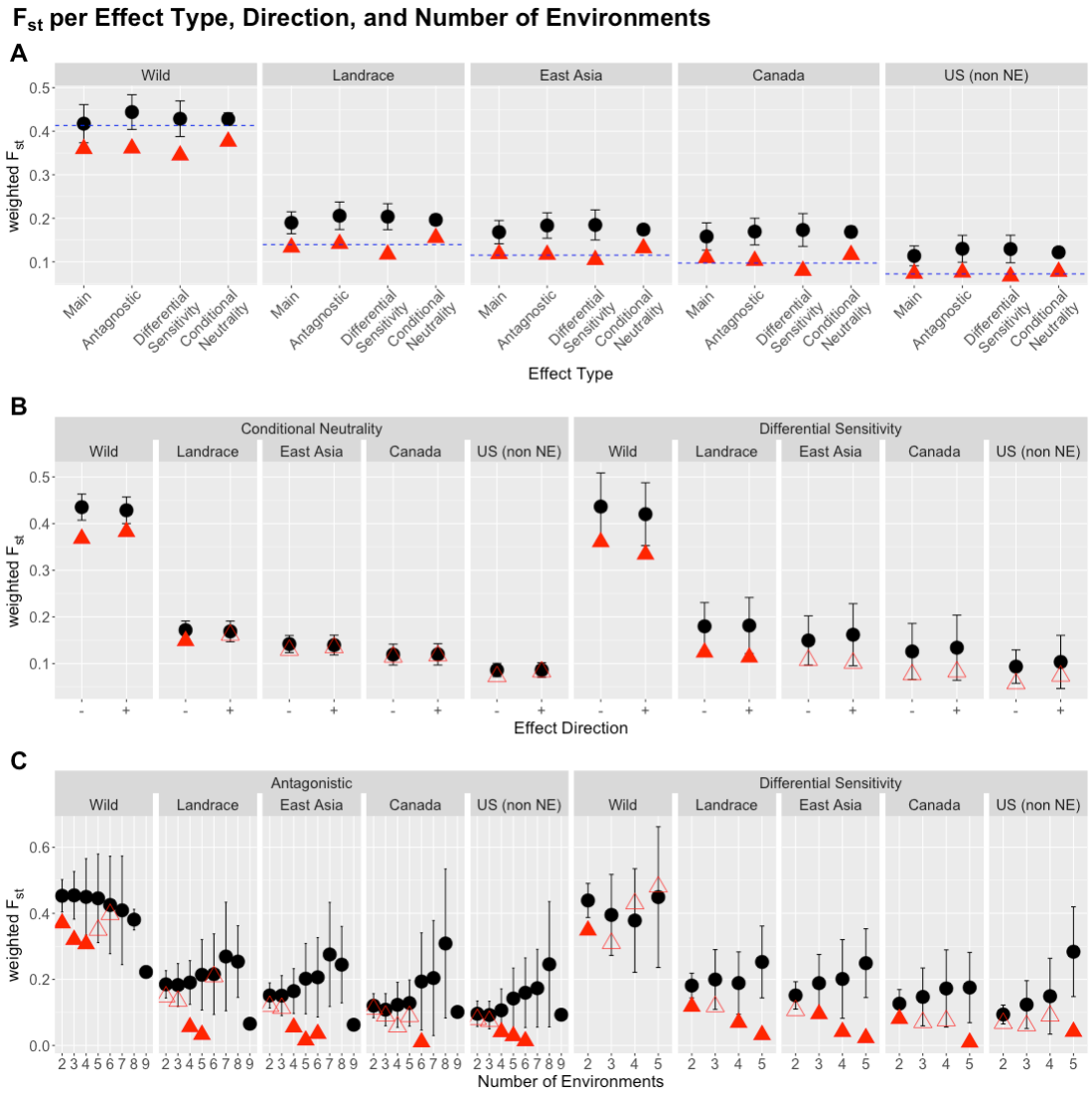


Figure 4.6: (A) Per effect type, the average weighted F_{st} value was lower than the permutation data and varied about the genome wide mean in all comparisons but consistently higher in conditionally neutral loci compared to other effect types. (B) No clear difference in divergence was detected between conditionally neutral and differentially sensitive effects for positive vs negative effects (C) Weighted F_{st} roughly decreased with an increasing number of environments with effects. For (A-C), black points represent the average number of windows sigificant for that effect classification based on the permutations, and the lines extending from those points the standard deviation. Red triangles represent the number of windows significant for that effect classification in the observed data, and a solid fill of the triangle reprsents that value fell outside the error bar of the permutations. The dashed line in (A) represents the genome wide average pairwise weighted F_{st} for that population.

diversity at negative differentially sensitive effects was lower than positive, suggesting stronger selection to remove these undesirable alleles (Figure 4.5B). A significant trend in weighted F_{st} between positive and negative effects was not appreciably clear (Figure 4.6B). When examining significant windows categorized by the number of environments they affect, Tajima's D generally increased and weighted F_{st} generally decreased with the number of environments. This pattern is loosely the opposite of what was predicted by the permutations and is most apparent for the antagonistic effects which reached a greater number of affected environments than the differentially sensitive loci, indicating negative selection on loci with effects in more environments (Figure 4.5C & 4.6C). Divergence from this trend for windows with a greater number of affected environments may be a result of the decreased number of windows observed to have such wide-spread effects (Figure 4.1C).

Discussion

In this paper we highlight the contribution of GxE to the complexity of the genomic landscape for soybean grain yield in an elite breeding population, and present evidence for varying levels of directional selection among these regions with respect to effect type, direction, and number of environments affected. Understanding the basis of GxE interactions for any trait and how the current landscape may have been shaped by selection within a crop breeding program is vital when seeking appropriate avenues to further drive genetic progress in crop productivity. Here, we first demonstrate that GxE is a major contributor to grain yield in an elite soybean population, via a disproportionately large number of small, conditionally neutral effects. This is consistent with hypotheses in the literature that conditionally neutral effects will be more predominant in selfing

species and when gene flow is restricted, as in a high intensity breeding program (Wadgymar et al., 2017). We then followed up with examinations of genetic diversity and divergence and show how selection appears to be acting less intensely in regions of the genome involved in environmental interactions.

Because the patterns of GxE interaction are variable and diverse, we hypothesized that selection pressure accordingly varies and does affect all GxE loci the same. Interestingly, in our study it was apparent that conditionally neutral loci displayed markedly increased signatures of selection relative to other effect types. This conflicts with the logic that loci with smaller effects in a smaller number of environments would experience weaker selection than those with more effects in a greater number of environments (Josephs, 2018). One possible explanation for this is that many of these loci classified as conditionally neutral truly have an effect in more than one environment, but it may be nearly impossible to distinguish this if selection has pushed allele frequencies in these regions towards fixation and their contribution to the overall phenotypic variance becomes diminished. This would also explain why we observed increased genetic diversity and decreased divergence in tandem with the increasing number of environments a genomic region affected. It is important to keep in mind the data presented in this study essentially provides a snapshot in time of an elite soybean breeding program, with loci in multiple stages of being selected upon. An interesting future direction for this research might involve studying the allele frequency changes at these antagonistic and differentially sensitive loci within the breeding population over multiple generations as it is subjected to artificial selection.

Additionally, further characterization of the environmental and plant growth factors contributing to GxE interactions could allow breeders to leverage positive GxE sources. A study of Argentinian soybean varieties recently discovered that increased GxE interactions were seen in genotypes with extended seed filling periods, and that this seems to translate into a net benefit for average seed yield as well as yield stability (de Felipe & Alvarez Prado, 2021). Our study also observed an increased level of GxE interactions in our highest yielding environment, possibly warranting the further study of conditions at this site that are possible leading to an advantageous GxE response.

Previously, we reported widespread antagonistic effects in this population using heavily imputed genotype data and single-locus genome wide association models (Happ et al., 2021). This contrasts with the results from this study, which concludes that the GxE landscape for grain yield is dominated by small, conditionally neutral effects. We speculate this difference in results is due to improvements in methods deployed in this study, namely assaying genetic variation from high depth sequencing data and deploying a modeling approach that simultaneously evaluated the contributions made by SNPs to grain yield. Performing SNP discovery from whole genome sequence data as opposed to imputing from a distantly related reference panel allowed us to capture more rare variation unique to this specific breeding population. Having the ability to analyze this low frequency variation may have been a key contributor in the discovery of small, conditionally neutral effect loci, as evidenced by the decreased Tajima's D value associated with them, indicative of more rare variation relative to other effect types. In this study we also considered genomic windows of multiple SNPs to be significant if the variance they explained was more than three standard deviations greater than the average

of all windows. Previous results were based on by identifying QTLs using a conservative, Bonferroni corrected p-value threshold to control for the multiple testing problem while modeling one variant at a time. The power to detect small effects at single SNPs, especially at low allele frequencies, may have been diminished in the previous approach. It is additionally unclear what the effect of simultaneously calculating all the SNP effects in a panel has on the power to discover GxE loci of varying effect patterns, which may be worth exploring given the gap in results seen between these studies.

Performing selection scans alone may overlook subtle changes in genetic variation that result from polygenic selection on a quantitative trait, and attempting to draw conclusions based on examining these values at QTL are likely to be biased by allele frequency (Josephs et al., 2017). To combat this, Josephs et al. (2015) compared QTL discovered in permutations to real data in order to avoid drawing false conclusions, a strategy we similarly implemented. This approach allowed us to elucidate that yield variance was higher at regions of diminished signatures of selection for GxE, but not main genetic effects - a pattern that would not have been clear without the null distributions established by the permutation data. A previous study in corn presented similar results where they compared the GxE variance for grain yield explained by groups of low and high F_{st} between temperate and tropical lines, and found it to be significantly lower in regions of high F_{st} (Gage et al., 2017). They concluded this was an indicator of selection against GxE. But we would suggest this could also be interpreted as reduced efficacy of selection to drive strong changes in allele frequency at these loci, due to the transient nature of environmental interactions. Yet another valuable insight made

possible by the permutation data, is that genetic variation for soybean grain yield is under both positive and negative selection, dependent on effect type and direction.

Conclusions

The data analyzed in our study suggest that loci involved with GxE effects are under varying levels of relaxed or weakened directional selection compared to the rest of the genome, while constitutive effects were not. GxE effects were found to be a massive driver of genetic variance for soybean grain yield principally via small effects detectable in single environments, which displayed relatively stronger selection signatures than other GxE loci. The evidence of even diminished directional selection at such an important component of genetic variance highlights the importance of incorporating novel material into breeding programs to preserve the future adaptability of a population to new environmental changes. Additionally, with this new understanding of the complexity surrounding the genetic architecture of soybean grain yield, breeders and researchers may find value in developing association and prediction models that explicitly take advantage of positive GxE interactions to aid in boosting genetic gain. Future studies which identify examine the change in allele frequencies at GxE loci over multiple generations could help resolve a more dynamic picture of how selection shapes the genetic landscape for soybean grain yield.

Acknowledgements

The research reported in this publication was supported by the Nebraska Soybean Board project #1726. This work was completed utilizing the Holland Computing Center of the University of Nebraska, which receives support from the Nebraska Research Initiative.

References

Alboukadel Kassambara. (2022). ggpubr: "ggplot2" Based Publication Ready Plots. https://CRAN.R-project.org/package=ggpubr

Anderson, J. T., Lee, C.-R., Rushworth, C., Colautti, R., & Mitchell-Olds, T. (2013). Genetic tradeoffs and conditional neutrality contribute to local adaptation. Molecular Ecology, 22(3), 699–708. https://doi.org/10.1111/j.1365-294X.2012.05522.x

Anderson, J. T., Willis, J. H., & Mitchell-Olds, T. (2011). Evolutionary genetics of plant adaptation. Trends in Genetics, 27(7), 258–266. https://doi.org/10.1016/j.tig.2011.04.001

Barton, N. H. (1998). The effect of hitch-hiking on neutral genealogies. Genetics Research, 72(2), 123–133. https://doi.org/10.1017/S0016672398003462

Barton, N. H., & Keightley, P. D. (2022). Understanding quantitative genetic variation | Nature Reviews Genetics. Nature Reviews Genetics, 3, 11–21. https://doi.org/10.1038/nrg700

Berg, J. J., & Coop, G. (2014). A Population Genetic Signal of Polygenic Adaptation. PLOS Genetics, 10(8), e1004412. https://doi.org/10.1371/journal.pgen.1004412

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics, 30(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bruce, R. W., Torkamaneh, D., Grainger, C., Belzile, F., Eskandari, M., & Rajcan, I. (2019). Genome-wide genetic diversity is maintained through decades of soybean breeding in Canada. Theoretical and Applied Genetics, 132(11), 3089–3100. https://doi.org/10.1007/s00122-019-03408-y

Chevin, L.-M., & Hospital, F. (2008). Selective Sweep at a Quantitative Trait Locus in the Presence of Background Genetic Variation. Genetics, 180(3), 1645–1660. https://doi.org/10.1534/genetics.108.093351

Claus O. Wilke & Brenton M. Wiernik. (2022). ggtext: Improved Text Rendering Support for "ggplot2." https://CRAN.R-project.org/package=ggtext

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., & Durbin, R. (2011). The variant call format and VCFtools. Bioinformatics, 27(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Des Marais, D. L., Hernandez, K. M., & Juenger, T. E. (2013). Genotype-by-Environment Interaction and Plasticity: Exploring Genomic Responses of Plants to the Abiotic Environment. Annual Review of Ecology, Evolution, and

Systematics, 44(1), 5–29. https://doi.org/10.1146/annurev-ecolsys-110512-135806

El-Soda, M., Malosetti, M., Zwaan, B. J., Koornneef, M., & Aarts, M. G. M. (2014). Genotype × environment interaction QTL mapping in plants: Lessons from Arabidopsis. Trends in Plant Science, 19(6), 390–398. https://doi.org/10.1016/j.tplants.2014.01.001

Erich Neuwirth. (2022). RColorBrewer: ColorBrewer Palettes. https://CRAN.R-project.org/package=RColorBrewer

Falconer, D. S. (1996). Introduction to quantitative genetics. Pearson Education India.

Fang, C., Ma, Y., Wu, S., Liu, Z., Wang, Z., Yang, R., Hu, G., Zhou, Z., Yu, H., Zhang, M., Pan, Y., Zhou, G., Ren, H., Du, W., Yan, H., Wang, Y., Han, D., Shen, Y., Liu, S., … Tian, Z. (2017). Genome-wide association studies dissect the genetic networks underlying agronomical traits in soybean. Genome Biology, 18(1), 161. https://doi.org/10.1186/s13059-017-1289-9

Fu, Y.-B. (2015). Understanding crop genetic diversity under modern plant breeding. Theoretical and Applied Genetics, 128(11), 2131–2142. https://doi.org/10.1007/s00122-015-2585-y

Gage, J. L., Jarquin, D., Romay, C., Lorenz, A., Buckler, E. S., Kaeppler, S., Alkhalifah, N., Bohn, M., Campbell, D. A., Edwards, J., Ertl, D., Flint-Garcia, S., Gardiner, J., Good, B., Hirsch, C. N., Holland, J., Hooker, D. C., Knoll, J., Kolkman, J., … de Leon, N. (2017). The effect of artificial selection on phenotypic plasticity in maize. Nature Communications, 8(1), 1348. https://doi.org/10.1038/s41467-017-01450-2

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. ArXiv:1207.3907 [q-Bio]. http://arxiv.org/abs/1207.3907

Gepts, P. (2006). Plant Genetic Resources Conservation and Utilization: The Accomplishments and Future of a Societal Insurance Policy. Crop Science, 46(5), 2278–2292. https://doi.org/10.2135/cropsci2006.03.0169gas

Goddard, M. (2009). Genomic selection: Prediction of accuracy and maximisation of long term response. Genetica, 136(2), 245–257. https://doi.org/10.1007/s10709-008-9308-0

Gualdrón Duarte, J. L., Cantet, R. J., Bates, R. O., Ernst, C. W., Raney, N. E., & Steibel, J. P. (2014). Rapid screening for phenotype-genotype associations by linear transformations of genomic evaluations. BMC Bioinformatics, 15(1), 246. https://doi.org/10.1186/1471-2105-15-246

Hadley Wickam & Maximilian Girlich. (2022). tidyr: Tidy Messy Data. https://CRAN.R-project.org/package=tidyr

Hadley Wickam, Romain François, Lionel Henry, & Kirill Müller. (2022). dplyr: A Grammar of Data Manipulation. https://CRAN.R-project.org/package=dplyr

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. https://www.frontiersin.org/articles/10.3389/fpls.2021.630175

Henrik Bengtsson. (2021). matrixStats: Functions that Apply to Rows and Columns of Matrices (and to Vectors). https://CRAN.R-project.org/package=matrixStats

Hermisson, J., & Pennings, P. S. (2005). Soft Sweeps: Molecular Population Genetics of Adaptation From Standing Genetic Variation. Genetics, 169(4), 2335–2352. https://doi.org/10.1534/genetics.104.036947

Identification of candidate domestication-related genes with a systematic survey of loss-of-function mutations—Torkamaneh—2018—The Plant Journal—Wiley Online Library. (n.d.). Retrieved November 16, 2021, from https://onlinelibrary.wiley.com/doi/full/10.1111/tpj.14104

Josephs, E. B. (2018). Determining the evolutionary forces shaping G × E. New Phytologist, 219(1), 31–36. https://doi.org/10.1111/nph.15103

Josephs, E. B., Lee, Y. W., Stinchcombe, J. R., & Wright, S. I. (2015). Association mapping reveals the role of purifying selection in the maintenance of genomic variation in gene expression. Proceedings of the National Academy of Sciences of the United States of America, 112(50), 15390–15395. https://doi.org/10.1073/pnas.1503027112

Josephs, E. B., Stinchcombe, J. R., & Wright, S. I. (2017). What can genome-wide association studies tell us about the evolutionary forces maintaining genetic variation for quantitative traits? New Phytologist, 214(1), 21–33. https://doi.org/10.1111/nph.14410

KEIM & P. (1988). A rapid protocol for isolating soybean DNA. Soybean Genet. Newsl., 15, 150–152.

Kemper, K. E., Saxton, S. J., Bolormaa, S., Hayes, B. J., & Goddard, M. E. (2014). Selection for complex traits leaves little or no classic signatures of selection. BMC Genomics, 15(1), 246. https://doi.org/10.1186/1471-2164-15-246

Khoury, C. K., Brush, S., Costich, D. E., Curry, H. A., de Haan, S., Engels, J. M. M., Guarino, L., Hoban, S., Mercer, K. L., Miller, A. J., Nabhan, G. P., Perales, H. R., Richards, C., Riggins, C., & Thormann, I. (2022). Crop genetic erosion: Understanding and responding to loss of crop diversity. New Phytologist, 233(1), 84–118. https://doi.org/10.1111/nph.17733

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinformatics, 15(1), 356. https://doi.org/10.1186/s12859-014-0356-4

Kremer, A., & Le Corre, V. (2012). Decoupling of differentiation between traits and their underlying genes in response to divergent selection. Heredity, 108(4), Article 4. https://doi.org/10.1038/hdy.2011.81

Lande, R. (1983). The response to selection on major and minor mutations affecting a metrical trait. Heredity, 50, 47–65.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods, 9(4), 357–359. https://doi.org/10.1038/nmeth.1923

Legarra, A., Ricard, A., & Varona, L. (2018). GWAS by GBLUP: Single and Multimarker EMMAX and Bayes Factors, with an Example in Detection of a Major Gene for Horse Gait. G3: Genes|Genomes|Genetics, 8(7), 2301–2308. https://doi.org/10.1534/g3.118.200336

Li, H. (2011). Improving SNP discovery by base alignment quality. Bioinformatics, 27(8), 1157–1158. https://doi.org/10.1093/bioinformatics/btr076

Malosetti, M., Ribaut, J.-M., & van Eeuwijk, F. A. (2013). The statistical analysis of multi-environment data: Modeling genotype-by-environment interaction and its genetic basis. Frontiers in Physiology, 4. https://www.frontiersin.org/articles/10.3389/fphys.2013.00044

Nielsen, R., Williamson, S., Kim, Y., Hubisz, M. J., Clark, A. G., & Bustamante, C. (2005). Genomic scans for selective sweeps using SNP data. Genome Research, 15, 1566–1575.

Piepho, H.-P., & Pillen, K. (2004). Mixed modelling for QTL × environment interaction analysis. Euphytica, 137(1), 147–153. https://doi.org/10.1023/B:EUPH.0000040512.84025.16

Pritchard, J. K., & Di Rienzo, A. (2010). Adaptation – not by sweeps alone. Nature Reviews Genetics, 11(10), Article 10. https://doi.org/10.1038/nrg2880

R Core Team. (2021). R: A Language and Environment for Statistical Computing (4.1). R Foundation for Statistical Computing.

Smith, J. M., & Haigh, J. (1974). The hitch-hiking effect of a favourable gene. Genetics Research, 23(1), 23–35. https://doi.org/10.1017/S0016672300014634

Stanton-Geddes, J., Paape, T., Epstein, B., Briskine, R., Yoder, J., Mudge, J., Bharti, A. K., Farmer, A. D., Zhou, P., Denny, R., May, G. D., Erlandson, S., Yakub, M., Sugawara, M., Sadowsky, M. J., Young, N. D., & Tiffin, P. (2013). Candidate Genes and Genetic Architecture of Symbiotic and Agronomic Traits Revealed by Whole-Genome, Sequence-Based Association Genetics in Medicago truncatula. PLOS ONE, 8(5), e65688. https://doi.org/10.1371/journal.pone.0065688

Strandén, I., & Garrick, D. J. (2009). Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. Journal of Dairy Science, 9. 2(6), 2971–2975. https://doi.org/10.3168/jds.2008-1929

Tajima F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 123(3):585–95.

Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J., & Prins, P. (2015). Sambamba: Fast processing of NGS alignment formats. Bioinformatics, 31(12), 2032–2034. https://doi.org/10.1093/bioinformatics/btv098

Valliyodan, B., Cannon, S. B., Bayer, P. E., Shu, S., Brown, A. V., Ren, L., Jenkins, J., Chung, C. Y.-L., Chan, T.-F., Daum, C. G., Plott, C., Hastie, A., Baruch, K., Barry, K. W., Huang, W., Patil, G., Varshney, R. K., Hu, H., Batley, J., … Nguyen, H. T. (2019). Construction and comparison of three reference-quality genome assemblies for soybean. The Plant Journal, 100(5), 1066–1082. https://doi.org/10.1111/tpj.14500

Valliyodan, B., Dan Qiu, Patil, G., Zeng, P., Huang, J., Dai, L., Chen, C., Li, Y., Joshi, T., Song, L., Vuong, T. D., Musket, T. A., Xu, D., Shannon, J. G., Shifeng, C., Liu, X., & Nguyen, H. T. (2016). Landscape of genomic diversity and trait discovery in soybean. Scientific Reports, 6(1), 23598–23598. https://doi.org/10.1038/srep23598

van de Wouw, M., van Hintum, T., Kik, C., van Treuren, R., & Visser, B. (2010). Genetic diversity trends in twentieth century crop cultivars: A meta analysis. Theoretical and Applied Genetics, 120(6), 1241–1252. https://doi.org/10.1007/s00122-009-1252-6

van Eeuwijk, F. A., Bink, M. C., Chenu, K., & Chapman, S. C. (2010). Detection and use of QTL for complex traits in multiple environments. Current Opinion in Plant Biology, 13(2), 193–205. https://doi.org/10.1016/j.pbi.2010.01.001

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

Wadgymar, S. M., Lowry, D. B., Gould, B. A., Byron, C. N., Mactavish, R. M., & Anderson, J. T. (2017). Identifying targets and agents of selection: Innovative methods to evaluate the processes that contribute to local adaptation. Methods in Ecology and Evolution, 8(6), 738–749. https://doi.org/10.1111/2041-210X.12777

Weir B.S. & Cockerham C.C. (1984) Estimating F-Statistics for the Analysis of Population Structure. Evolution, 38(6):1358–70.

Wouw, M. van de, Kik, C., Hintum, T. van, Treuren, R. van, & Visser, B. (2010). Genetic erosion in crops: Concept, research results and challenges. Plant Genetic Resources, 8(1), 1–15. https://doi.org/10.1017/S1479262109990062

Zhou, Z., Jiang, Y., Wang, Z., Gou, Z., Lyu, J., Li, W., Yu, Y., Shu, L., Zhao, Y., Ma, Y., Fang, C., Shen, Y., Liu, T., Li, C., Li, Q., Wu, M., Wang, M., Wu, Y., Dong, Y., … Tian, Z. (2015). Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nature Biotechnology, 33(4), 408–414. https://doi.org/10.1038/nbt.3096

APPENDIX

Supplemental Tables

Supplemental Table 2.1: Sequencing depth of individual genotypes used in the generation of the reference and study panels. Samples with text struck through were dropped from final reference panel due to suspected contamination.

Genotype	Panel	Mean Depth
G100540	Study	1
G100534	Study	1.01
G100585	Study	1.02
G100613	Study	1.02
G100538	Study	1.04
G100579	Study	1.04
G100555	Study	1.05
G100565	Study	1.05
G100526	Study	1.06
G100600	Study	1.07
G100573	Study	1.08
G100553	Study	1.09
G100602	Study	1.09
G100545	Study	1.11
G100561	Study	1.12
G100582	Study	1.13
G100562	Study	1.14
G100605	Study	1.15
G100506	Study	1.17
G100529	Study	1.17
G100535	Study	1.17
G100541	Study	1.17
G100583	Study	1.19
G100591	Study	1.19
G100536	Study	1.2
G100559	Study	1.21

G100574	Study	1.22
G100576	Study	1.25
G100504	Study	1.26
G100554	Study	1.26
G100601	Study	1.26
G100508	Study	1.29
G100515	Study	1.3
G100532	Study	1.3
G100569	Study	1.3
G100596	Study	1.3
G100516	Study	1.31
G100595	Study	1.31
G100537	Study	1.32
G100548	Study	1.32
G100531	Study	1.34
G100566	Study	1.36
G100599	Study	1.36
G100597	Study	1.37
G100606	Study	1.37
G100575	Study	1.38
G100503	Study	1.4
G100517	Study	1.4
G100547	Study	1.4
G100523	Study	1.44
G100519	Study	1.51
G100552	Study	1.52
G100507	Study	1.55
G100581	Study	1.55
G100513	Study	1.57
G100578	Study	1.59
G100592	Study	1.61
G100510	Study	1.64
G100607	Study	1.65
G100590	Study	1.69
G100593	Study	1.69
G100550	Study	1.71
G100563	Study	1.75
G100521	Study	1.82
G100560	Study	1.84
G100522	Study	1.85

G100557	Study	1.86
G100527	Study	1.87
G100568	Study	1.9
G100570	Study	1.91
G100501	Study	1.99
G100514	Study	1.99
G100588	Study	2
G100524	Study	2.04
G100587	Study	2.04
G100544	Study	2.05
G100511	Study	2.1
G100528	Study	2.15
G100610	Study	2.16
G100546	Study	2.17
G100584	Study	2.17
G100509	Study	2.21
G100577	Study	2.27
G100551	Study	2.28
G100580	Study	2.3
G100608	Study	2.32
G100539	Study	2.36
G100571	Study	2.36
G100549	Study	2.38
G100589	Study	2.44
G100609	Study	2.44
G100525	Study	2.52
G100604	Study	2.59
G100594	Study	2.64
G100530	Study	2.74
G100586	Study	2.87
G100564	Study	2.91
G100603	Study	2.91
G100518	Study	2.93
G100543	Study	3.06
G100611	Study	3.11
G100502	Study	3.3
G100572	Study	3.66
G100598	Study	4.2
G100505	Study	4.25
G100558	Study	4.47

G100556	Study	4.58
G100512	Study	6.78
G100614	Study	7.43
G100542	Study	9.15
G100612	Study	14.65
G100533	Study	14.78
G100567	Study	16.65
G100520	Study	36.47
PI_518664	Reference	14.8
Peking	Reference	15.68
PI_089772	Reference	16.3
PI_090763	Reference	14.28
PI_404166	Reference	16.58
PI_407788A	Reference	15.5
PI_424298	Reference	16.18
PI_437655	Reference	13.37
PI_495017C	Reference	15.6
PI_468915	Reference	12.51
PI_507354	Reference	13.29
PI_567305	Reference	13.39
S05-11482	Reference	15.76
PI_548667	Reference	13.64
PI_437654	Reference	15.77
PI_567387	Reference	20.48
PI_437725	Reference	19.8
PI_437690	Reference	14.15
PI_548402	Reference	18.37
PI_088788	Reference	12
PI_209332	Reference	15.82
PI_404198B	Reference	14.84
PI_424608A	Reference	12.58
PI_548316	Reference	17.68
PI_567516C	Reference	13.22
PI_612611	Reference	14.9
S10-11227	Reference	17.17
Holladay	Reference	17.6
IA3023	Reference	16.27
Maverick	Reference	16.34
PI_079691-4	Reference	15.08

PI_086006	Reference	19.54
PI_087617	Reference	19.64
PI_087631-1	Reference	16.36
PI_196175	Reference	18.59
PI_200508	Reference	14.84
~~PI_248515~~	~~Reference~~	~~12.67~~
~~PI_366121~~	~~Reference~~	~~16.97~~
PI_378702	Reference	14.29
PI_398593	Reference	18.63
PI_398595	Reference	19.64
PI_398610	Reference	19.95
PI_398614	Reference	16.93
PI_407162	Reference	19.51
PI_407184	Reference	18.81
PI_407729	Reference	17.84
PI_407965	Reference	20.48
PI_408105A	Reference	17.38
PI_416937	Reference	19.48
PI_424078	Reference	18.29
PI_424079	Reference	18.4
PI_424088	Reference	17.93
PI_437169B	Reference	19.19
PI_437679	Reference	17.68
PI_437863A	Reference	19.3
~~PI_438258~~	~~Reference~~	~~17.98~~
~~PI_458515~~	~~Reference~~	~~14.3~~
PI_464920B	Reference	20.7
PI_467312	Reference	17.78
PI_471938	Reference	16.82
PI_475783B	Reference	15.31
PI_483463	Reference	14.96
PI_518751	Reference	14.69
PI_542044	Reference	22.72
PI_547862	Reference	20.43
PI_548317	Reference	18.84
PI_548349	Reference	20.39
PI_548415	Reference	16.23
PI_548511	Reference	16.39

PI_548657	Reference	16.94
PI_549031	Reference	15.95
PI_552538	Reference	17.52
PI_561271	Reference	17.65
PI_567230	Reference	17.4
PI_567336B	Reference	17.94
PI_567343	Reference	14.6
PI_567354	Reference	17.61
PI_567357	Reference	15.85
PI_567383	Reference	18.44
~~PI_567519~~	~~Reference~~	~~17.33~~
PI_567611	Reference	18.14
PI_567651	Reference	15.03
PI_567690	Reference	17.16
PI_567719	Reference	15.11
PI_567731	Reference	15.6
PI_591539	Reference	19.09
PI_593258	Reference	17.8
PI_594012	Reference	18.38
PI_594512A	Reference	20.49
~~PI_594599~~	~~Reference~~	~~14.87~~
~~PI_597387~~	~~Reference~~	~~17.35~~
PI_603154	Reference	17.62
PI_603170	Reference	17.44
PI_603175	Reference	24.6
PI_603176A	Reference	18.92
PI_603497	Reference	15.56
PI_605869A	Reference	15.96
PI_639740	Reference	13.69
PI_647086	Reference	18.2
PI_658519	Reference	18.71
S07-5049	Reference	20.88
V71-370	Reference	20.88
FC_31721	Reference	20.2
PI_438471	Reference	24.01
PI_417091	Reference	22.45
PI_417015	Reference	16.22

Depth	Unfiltered Accuracy	GP > 0.45 Accuracy	Improvement	GP > 0.9 Accuracy	Improvement
0.1	89.70%	92.70%	3.00%	94.80%	5.10%
0.2	92.80%	95.60%	2.80%	97.20%	4.40%
0.3	93.60%	96.30%	2.70%	97.80%	4.20%
0.4	94.20%	96.60%	2.40%	98.10%	3.90%
0.5	94.20%	96.70%	2.50%	98.40%	4.20%
0.6	94.40%	96.70%	2.30%	98.40%	4.00%
0.7	94.50%	96.70%	2.20%	98.50%	4.00%
0.8	94.30%	96.70%	2.40%	98.60%	4.30%
0.9	94.30%	96.70%	2.40%	98.60%	4.30%
1	94.40%	96.70%	2.30%	98.60%	4.20%
Averages	**93.64%**	**96.14%**	**2.50%**	**97.90%**	**4.26%**

Supplementary Table 2.2: Accuracy improvement as a result of filtering on Beagle's genotype posterior probability after imputation.

Item	1X	0.9X	0.8X	0.7X	0.6X	0.5X	0.4X	0.3X	0.2X	0.1X
DNA Extraction Reagents	$0.10	$0.10	$0.10	$0.10	$0.10	$0.10	$0.10	$0.10	$0.10	$0.10
DNA Extraction Disposables/Overhead	$0.30	$0.30	$0.30	$0.30	$0.30	$0.30	$0.30	$0.30	$0.30	$0.30
Library Prep Reagents	$8.67	$8.67	$8.67	$8.67	$8.67	$8.67	$8.67	$8.67	$8.67	$8.67
Library Prep Disposables/Overhead	$2.00	$2.00	$2.00	$2.00	$2.00	$2.00	$2.00	$2.00	$2.00	$2.00
NextSeq 500/550 High Output Kit v2.5 (300 cycles)	$48.75	$43.88	$39.00	$34.13	$29.25	$24.39	$19.50	$14.63	$9.75	$4.88
Total per Sample Cost	**$59.82**	**$54.95**	**$50.07**	**$45.20**	**$40.32**	**$35.46**	**$30.57**	**$25.70**	**$20.82**	**$15.95**

Supplementary Table 2.3: A breakdown of the per item costs involved in the DNA extraction, sample library preparation, and sequencing of genotypes from 0.1X – 1X sequencing depths in USD.

Location	Coordinates	Years	Maturity Groups	Lines Tested				
				MG I	*early MG II*	*late MG II*	*MG III*	*Total*
Cotesfield	41.3581° N, 98.6338° W	2018, 2019	I, early II	23	85	x	x	108
Lincoln	40.8136° N, 96.7026° W	2017, 2018, 2019	late II, III	x	x	64	41	105
Mead	41.2286° N, 96.4892° W	2018, 2019	I, early II	23	85	x	x	108
Phillips	40.8980° N, 98.2132° W	2017, 2018, 2019	I, II, III	23	85	64	41	213
Wymore	40.1222° N, 96.6623° W	2018	late II, III	x	x	64	41	105

Supplementary Table 3.1: Five testing sites across eastern Nebraska were used to evaluate the yield performance of 213 soybean lines over three years for a combination of eleven unique environments. Lines were grouped according to maturity and assigned to field sites accordingly.

	Overall	Year			Location				
		2017	*2018*	*2019*	*Cotesfield*	*Mead*	*Phillips*	*Lincoln*	*Wymore*
Average	4976.41	5204.80	4919.81	4859.99	5199.87	3995.49	5570.40	4491.35	4671.49
Minimum	2162.74	3227.30	2176.19	2162.74	2949.56	2176.19	2162.74	3327.51	3619.37
Maximum	7080.70	7080.70	6832.55	6353.06	6568.93	5419.64	7080.70	5672.50	5811.03
Standard Deviation	810.34	733.96	920.07	663.02	636.17	660.77	673.28	423.59	362.30

Supplementary Table 3.2: The overall grain yield ranged from 2162.74 to 7080.70 kg/ha with an average of 4976.41 kg/ha and standard deviation of 810.34 kg/ha. 2017 was the highest average yielding year and Phillips the highest yielding average location.

	Thennarasu's non-parametric statistics				Huhn's and Nassar and Huhn's non-parametric statistics				Wricke's Ecovalence	Shukla's Stability Variance	Deviation from regression	Regression coefficient	Coefficient of variance	GE variance component	Mean variance component	Kangs Rank Sum
	NP1	***NP2***	***NP3***	***NP4***	***S1***	***S2***	***S3***	***S6***	W_i^2	σ^2_i	S^2_{di}	$^*b_i$	***CVi***	$\vartheta_{(i)}$	ϑi	***KR***
Average	49.89	0.59	0.57	0.62	62.08	3436.49	347.33	5.10	287.32	28.86	72.59	0.54	12.19	28.86	28.93	na
Minimum	14.73	0.09	0.19	0.13	18.29	282.47	20.55	0.88	15.37	1.42	0.00	0.00	0.00	27.95	15.27	na
Maximum	88.64	6.56	1.52	1.46	117.16	11996.16	1189.88	13.28	2245.88	226.53	302.64	1.23	33.87	28.99	127.31	na
Standard Deviation	16.82	0.81	0.24	0.29	21.11	2150.63	251.54	2.60	315.48	31.84	43.95	0.21	4.04	0.15	15.85	na

Supplementary Table 3.3: Univariate stability measures reflect a wide range of stability levels using multiple approaches.

	Sum Across Environments of GEI	AMMI Stability Index	AMMI Stability Value	AMMI Based Stability Parameter	Sum Across Environments of Absolute Value of GEI	Annicchiarico's D Parameter	Zhang's D Parameter	Averaged of the Squared Eigenvector Values	Stability Measure Based on Fitted AMMI Model	Modified AMMI Stability Index	Modified AMMI Stability Value	Sums of the Absolute Value of the IPC Scores	Absolute Value of the Relative Contribution of IPCs to the Interaction
	AMGE	***ASI***	***ASV***	***ASTAB***	***AVAMGE***	***DA***	***DZ***	***EV***	***FA***	***MASI***	***MASV***	***SIPC***	***Za***
Average	4.14E-18	3.13E-01	1.94E+00	1.74E+00	3.01E+01	1.36E+01	1.06E-01	4.61E-03	2.32E+02	3.30E-01	2.11E+00	1.81E+00	3.85E-02
Minimum	-1.69E-13	2.24E-02	1.39E-01	8.25E-02	8.76E+00	3.48E+00	2.44E-02	1.99E-04	1.21E+01	5.99E-02	4.11E-01	4.65E-01	7.61E-03
Maximum	1.63E-13	1.18E+00	7.35E+00	9.81E+00	8.48E+01	3.96E+01	2.51E-01	2.11E-02	1.57E+03	1.19E+00	7.44E+00	4.56E+00	1.11E-01
Standard Deviation	5.63E-14	2.09E-01	1.30E+00	1.75E+00	1.56E+01	6.90E+00	5.06E-02	4.47E-03	2.53E+02	2.05E-01	1.27E+00	8.84E-01	2.05E-02

Supplementary Table 3.4: Multivariate stability measures calculated from an AMMI model fit reflect a wide range of stability levels using multiple approaches.

Population	# of Lines	# of SNPs	Sequencing Depth				Missing Data						
			Mean	*S.D.*	*Min*	*Max*	Total	Per Marker			Per Individual		
								S.D.	*Min*	*Max*	*S.D.*	*Min*	*Max*
ALL	1237	7,137,085	13.33	8.27	0.05	65.54	6.32%	7.81%	0.00%	50%	6.38%	0.26%	46.68%
Wild	51	5,979,677	12.96	4.72	3.46	23.39	4.09%	8.10%	0.00%	49.02%	3.38%	1.65%	23.25%
Landrace	455	4,497,817	13.76	9.25	5.85	64.54	5.57%	8.47%	0.00%	49.89%	4.50%	0.65%	34.95%
East Asia	352	3,790,485	12.96	6.81	5.77	37.63	5.14%	8.37%	0.00%	50.00%	3.43%	0.40%	27.54%
Canada	112	2,533,440	14.26	6.79	3.44	39.56	7.86%	7.85%	0.00%	50.00%	8.20%	0.46%	39.49%
United States (non-Nebraska)	64	1,723,191	22.71	9.85	6.92	55.75	3.22%	6.48%	0.00%	50.00%	2.65%	0.36%	14.80%
Nebraska	204	2,640,097	9.27	5.83	0.05	33.00	10.53%	9.72%	0.00%	50.00%	9.78%	0.18%	45.74%

Supplementary Table 4.1: In total, whole genome sequence data from 1237 lines across six populations of soybean were analyzed in this study. Sequencing depth ranged from 9.27 to 22.71, and missing data rates ranged from 3.22% to 10.53% based upon the SNP calls.

Site	Variance Component			
	Genotype	Environment	Genotype x Environment	Residual
2017 Phillips	8.39%	52.02%	11.08%	28.52%
2017 Stevens Creek	8.86%	54.95%	11.70%	24.49%
2018 Cotesfield	8.75%	54.27%	11.56%	25.42%
2018 Mead	10.01%	62.09%	13.22%	14.67%
2018 Phillips	8.32%	51.61%	10.99%	29.07%
2018 Stevens Creek	9.03%	55.98%	11.92%	23.07%
2018 Wymore	9.87%	61.18%	13.03%	15.93%
2019 Cotesfield	9.53%	59.09%	12.58%	18.80%
2019 Mead	8.56%	53.08%	11.31%	27.05%
2019 Phillips	8.36%	51.85%	11.04%	28.75%
2019 Stevens Creek	10.21%	63.30%	13.48%	13.01%

Supplementary Table 4.2: Environmental effects explained the most variance at any location using a multi-section mixed effects model, followed by GxE interactions, and finally the constitutive genotype effects.

		Summary Statistics for Proportion of Yield Variance Explained per Genomic Window				
		Mean	Median	S.D.	Min	Max
Main		1.00E-05	3.90E-06	1.81E-05	4.94E-14	4.19E-04
GxE		1.01E-05	5.63E-06	1.29E-05	4.74E-08	2.71E-04
GxE per Environment	2017 Phillips	1.11E-06	4.17E-07	2.03E-06	1.10E-15	5.87E-05
	2017 Stevens Creek	3.77E-07	1.50E-07	7.01E-07	2.5-e-16	2.68E-05
	2018 Cotesfield	1.15E-07	4.27E-08	2.11E-07	4.15E-17	4.96E-06
	2018 Mead	2.45E-07	9.84E-08	4.26E-07	3.63E-16	1.62E-05
	2018 Phillips	1.02E-06	3.67E-07	1.99E-06	7.54E-20	5.52E-05
	2018 Stevens Creek	2.62E-07	1.03E-07	4.66E-07	2.04E-16	1.13E-05
	2018 Wymore	5.19E-07	2.08E-07	9.81E-07	9.02E-17	5.39E-05
	2019 Cotesfield	3.93E-07	1.54E-07	6.98E-07	1.293-16	3.18E-05
	2019 Mead	1.32E-07	5.16E-08	2.35E-07	1.07E-16	5.49E-06
	2019 Phillips	2.90E-06	1.13E-06	5.21E-06	2.47E-15	1.24E-04
	2019 Stevens Creek	2.25E-07	8.72E-08	4.15E-07	8.82E-17	1.25E-05

Supplementary Table 4.3: The average proportion of variance explained by either main or GxE effects was approximately the same among all windows, however the median GxE effects were markedly higher than median main genetic effects. Among the individual locations, the largest contributions to yield variance were seen at the Phillips locations.

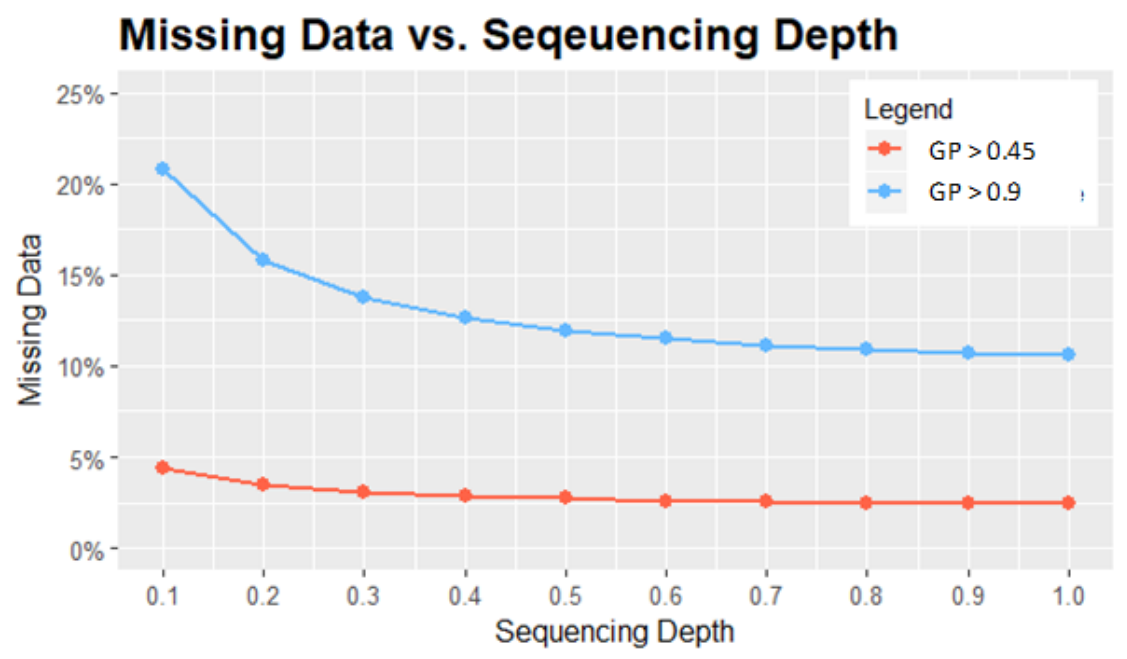
		Summary Statistics for Yield Effect (kg/ha) per Genomic Window					
		Mean	Median	S.D.	Min	Max	Range
Main		8.57E-05	-5.93E-05	1.09E-02	-0.18	0.19	0.37
Environment	2017 Phillips	4.32E-04	1.66E-04	1.26E-02	-0.23	0.17	0.4
	2017 Stevens Creek	-2.93E-05	8.21E-06	7.90E-03	-0.15	0.17	0.32
	2018 Cotesfield	1.82-E4	4.58E-06	9.19E-03	-0.16	0.15	0.31
	2018 Mead	7.67E-05	3.24-e05	1.25E-02	-0.14	0.27	0.41
	2018 Phillips	1.19E-04	5.04E-05	1.46E-02	-0.28	0.29	0.57
	2018 Stevens Creek	4.38E-04	1.65E-04	8.55E-03	-0.12	0.12	0.24
	2018 Wymore	8.31E-05	-4.22E-05	1.18E-02	-0.32	0.16	0.48
	2019 Cotesfield	1.68E-04	-1.42E-04	1.71E-02	-0.29	0.5	0.79
	2019 Mead	1.01E-04	4.29E-05	7.24E-03	-0.12	0.11	0.23
	2019 Phillips	8.08E-05	-5.47E-05	1.09E-02	-0.18	0.19	0.37
	2019 Stevens Creek	2.32e-04	-7.83E-05	6.43E-03	-0.13	9.32e-02	0.22

Supplementary Table 4.4: The average effect size for GxE effects in any environment was approximately the same or larger than the main genetic effects. The widest range of effect sizes was not consistent by sight, with the top three seen at the 2019 Cotesfield location, 2018 Phillips, and 2018 Wymore sites.

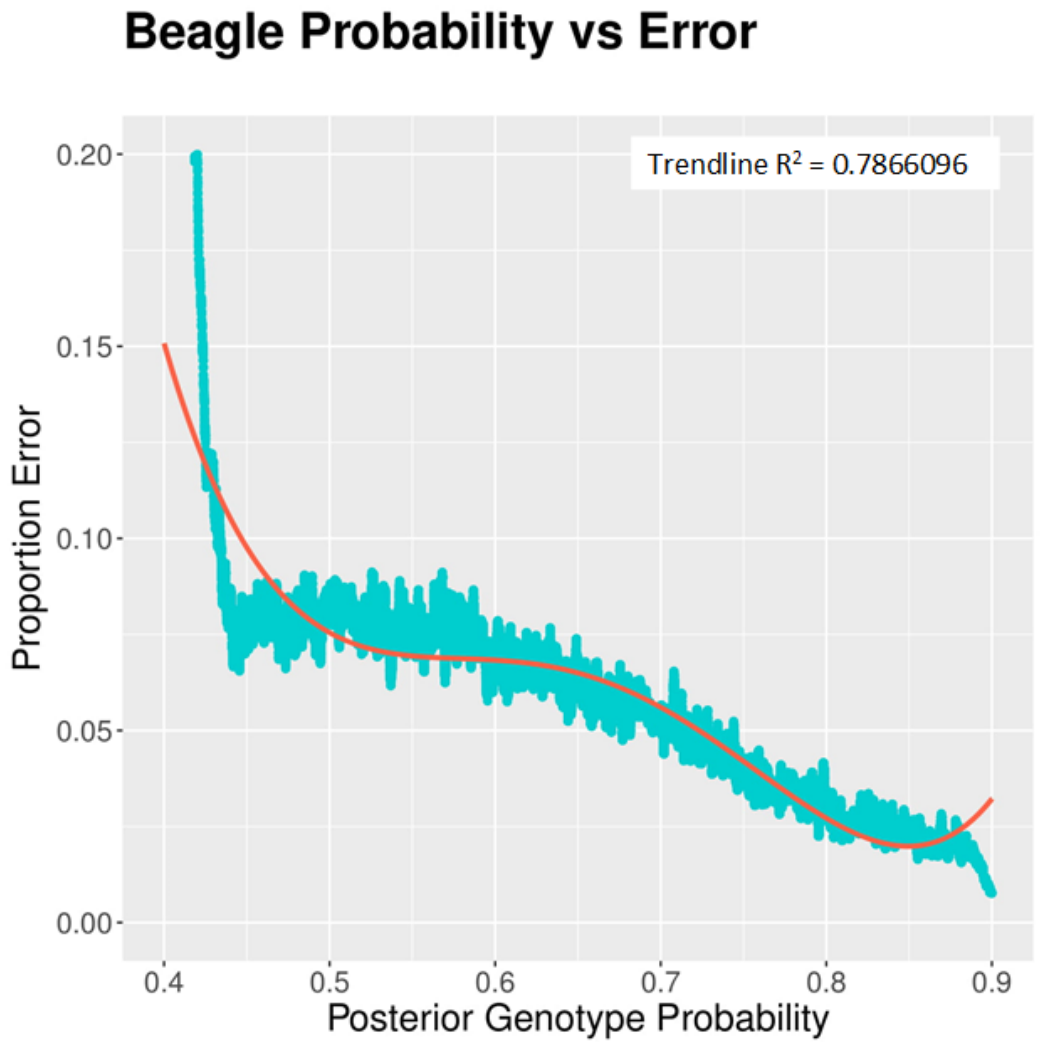
Population	Overall Weighted F_{st} Estimate	Windowed Weighted F_{st}				
		Mean	*Median*	*S.D.*	*Min*	*Max*
Wild	0.4867	0.4193	0.4132	0.1749	-0.1255	0.9665
Landrace	0.2079	0.1745	0.1396	0.1373	-0.0066	0.8695
East Asia	0.1873	0.1521	0.1152	0.1313	-0.0072	0.7589
Canada	0.1931	0.1484	0.0972	0.1511	-0.1280	0.8547
United States (non-Nebraska)	0.1261	0.1080	0.0722	0.1124	-0.0152	0.6790

Supplementary Table 4.5: Trends in overall and windowed F_{st} were as expected, with the highest value found in comparison to the wild lines and the lowest in comparison to a broad population of other elite US lines.

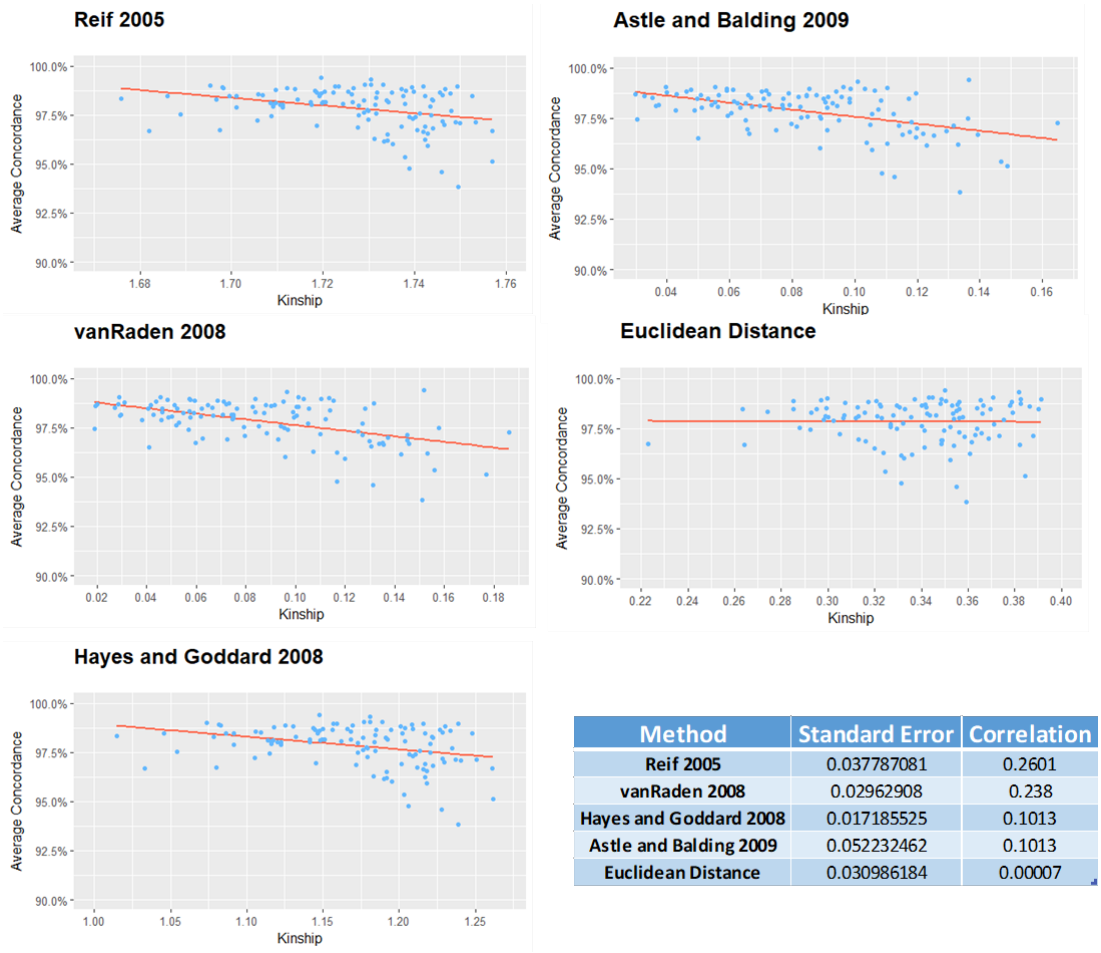
Supplemental Figures



Supplementary Figure 2.1: The extent of missing data points reintroduced by filtering for Beagle posterior genotype probabilities greater than 0.45 and 0.9. A mild filter of GP > 0.45 keeps missing data below 5%, while missing data greatly inflates to over 20% at 0.1X with a filter of GP > 0.9.

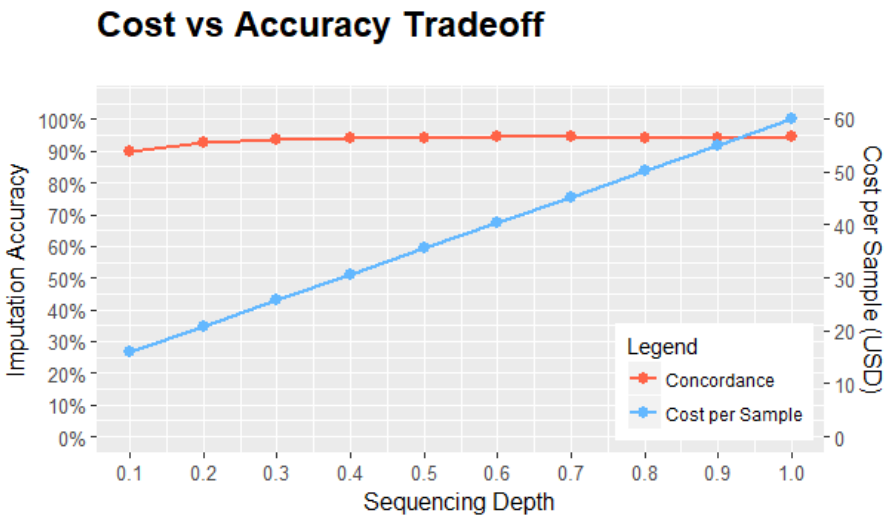


Supplementary Figure 2.2: Comparing the frequency of error at individual sites across sequencing depths with the assigned genotype probability for imputed values by Beagle demonstrates a strong correlation, making the posterior genotype probability a useful metric for post imputation filtering for data improvement.

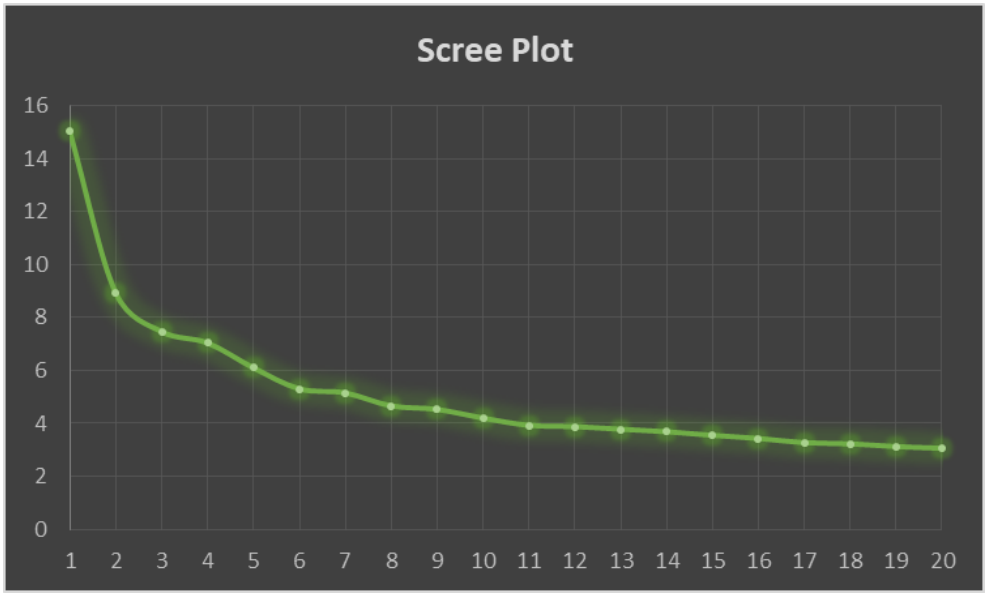


Method	Standard Error	Correlation
Reif 2005	0.037787081	0.2601
vanRaden 2008	0.02962908	0.238
Hayes and Goddard 2008	0.017185525	0.1013
Astle and Balding 2009	0.052232462	0.1013
Euclidean Distance	0.030986184	0.00007

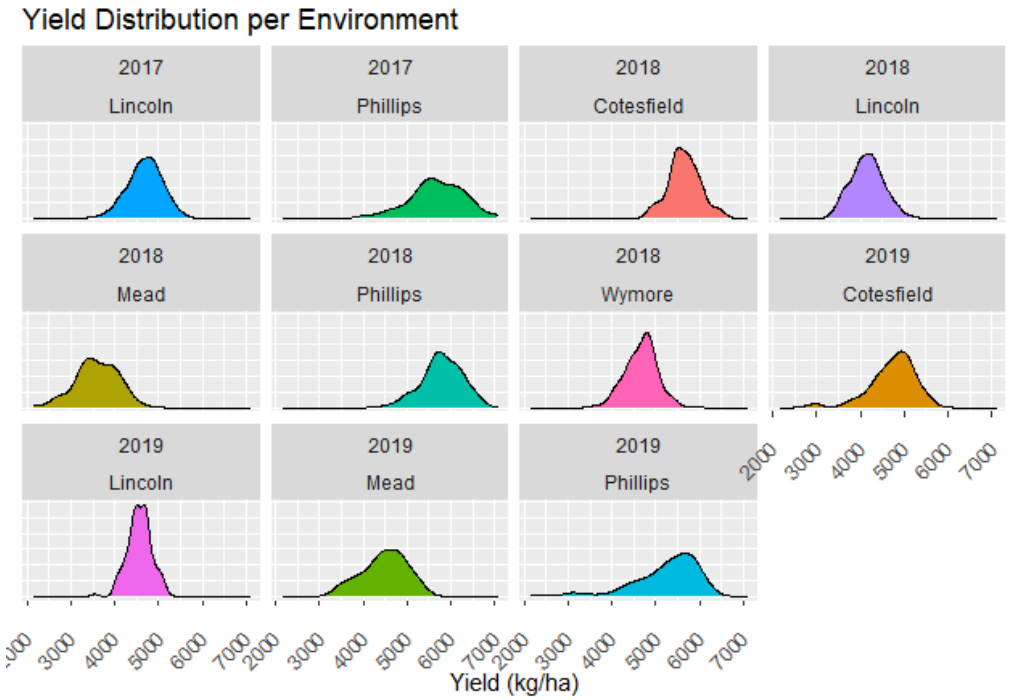
Supplementary Figure 2.3: The average of the top five scores for five relatedness metrics are plotted against the average concordance across depths for that genotype. Correlations and standard errors reported in the bottom right hand corner reveal no strong relationship for any metric, which may be explained by the low level variation within our study panel in terms of degree of kinship to the reference panel and overall weak kinship.



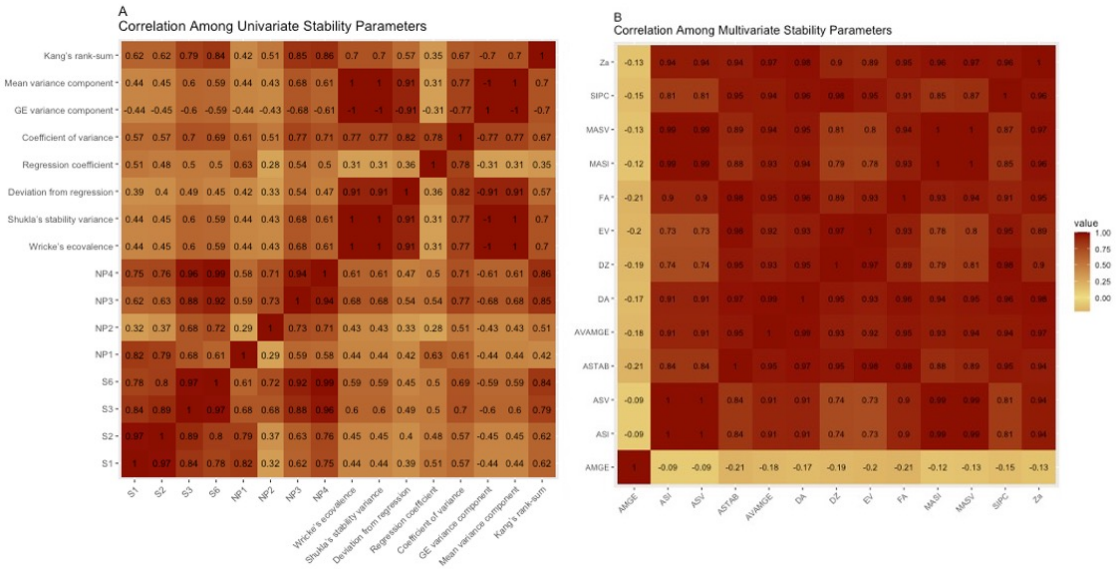
Supplementary Figure 2.4: The proportion of accuracy retained relative to 1X from raw imputed values vs the retained cost per sample. Decreasing coverage from 1X to 0.3X results in a nearly negligible loss in accuracy of 0.85%, while decreasing per sample costs by 57.04%



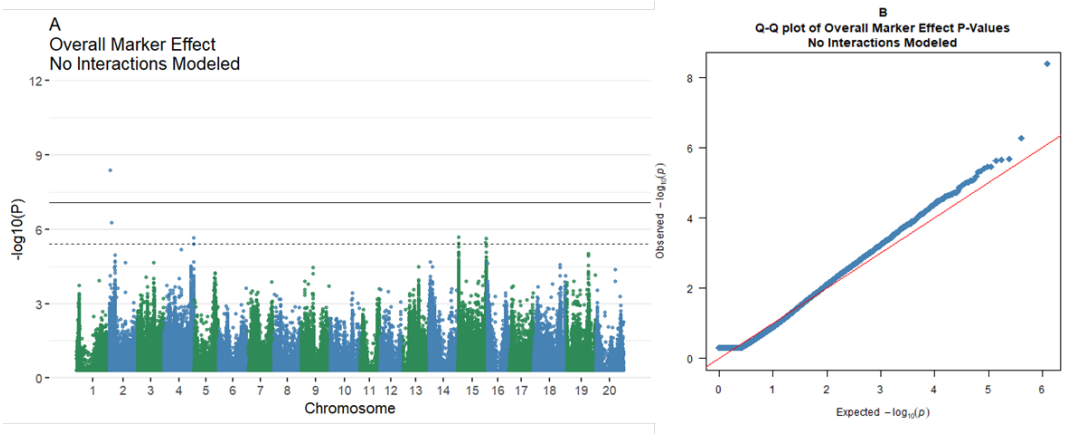
Supplementary Figure 3.1: The trend line drawn between eigenvalues from a principal component analysis of the genotypic dataset shows a reduction in slope for subsequent points around the eighth principal component. Thus, the first eight principal components were used in our GWAS analysis.



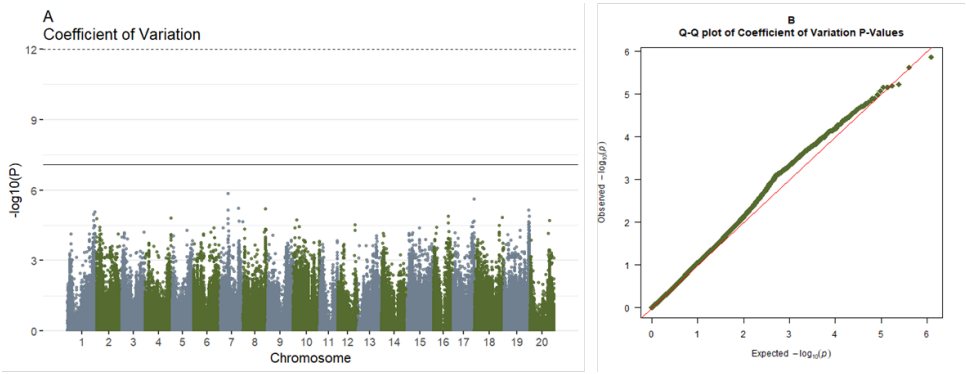
Supplementary Figure 3.2: Yield distribution is approximately normal per year and location combination.



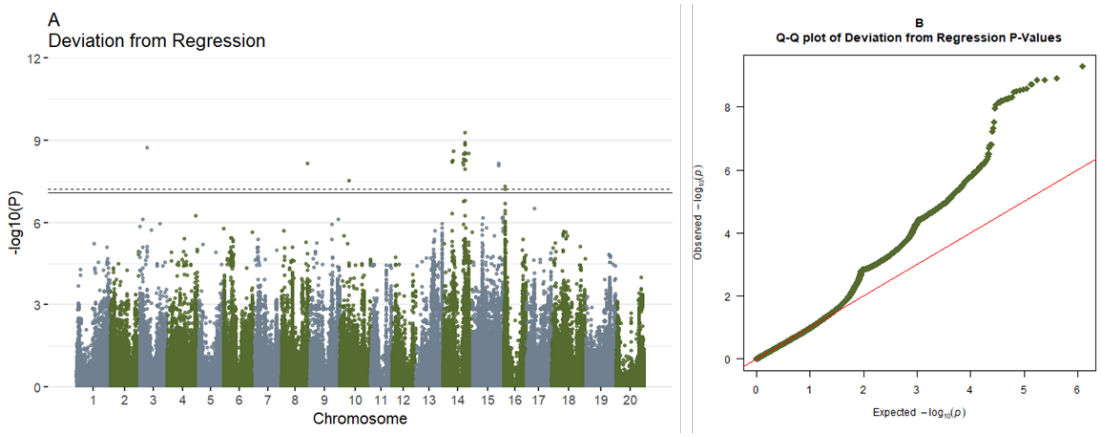
Supplementary Figure 3.3: Correlation matrices for traditional stability statistics reveals that while many of the multivariate stability parameters are highly correlated with each other, many of the univariate measures are fairly distinct. Notably, Wricke's Ecovalence, Shukla's Stability Variance, the GE variance component and mean variance components were all perfectly correlated with values of 1 or -1.



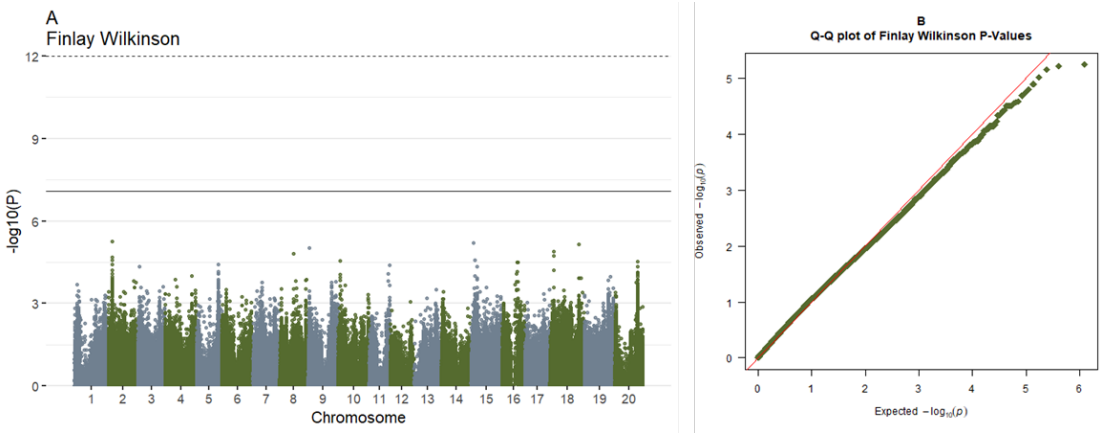
Supplementary Figure 3.4: Manhattan plot (A) and q-q plot (B) of the GWAS results for yield without fitting genotype by environment interactions 1 QTL is significant via Bonferroni correction, and an additional 3 QTL are significant when considering a FDR of 5%.



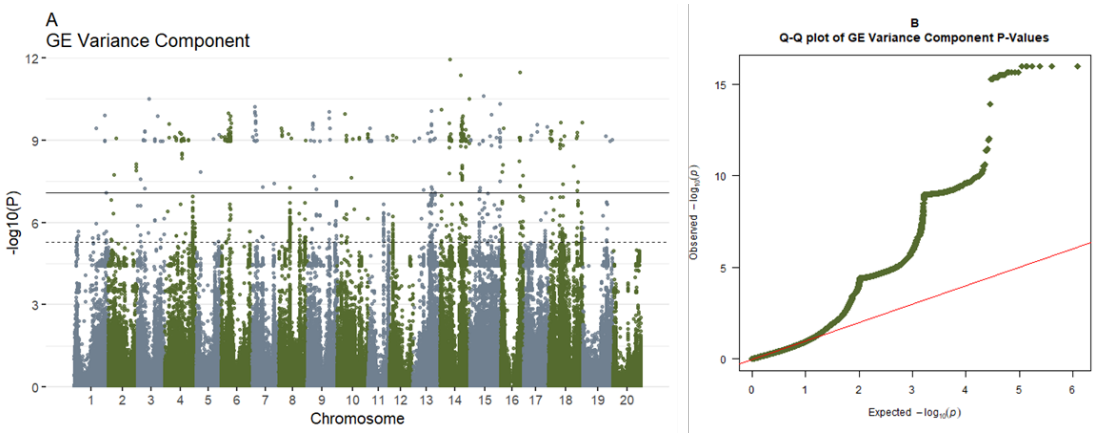
Supplementary Figure 3.5: Manhattan plot (A) and q-q plot (B) of GWAS results using the coefficient of variation as the model phenotype.



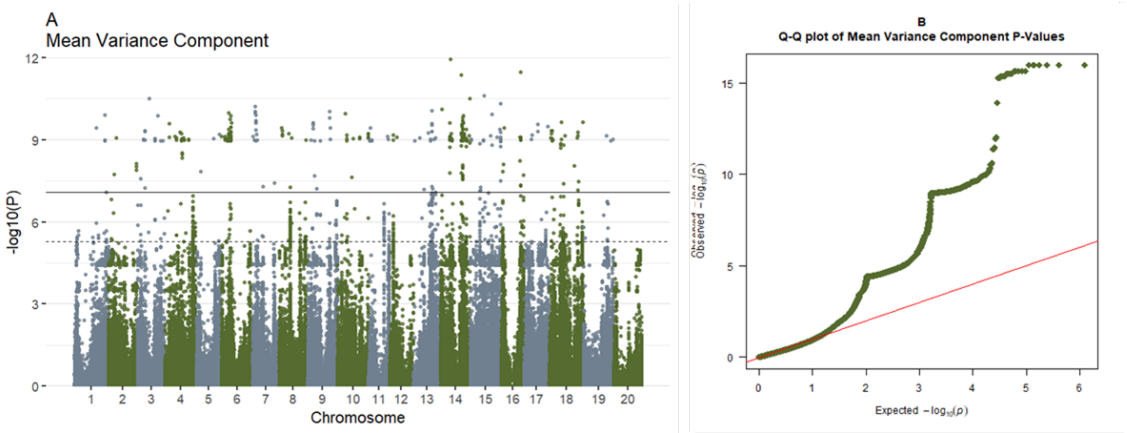
Supplementary Figure 3.6: Manhattan plot (A) and q-q plot (B) of GWAS results using the deviation from regression as the model phenotype.



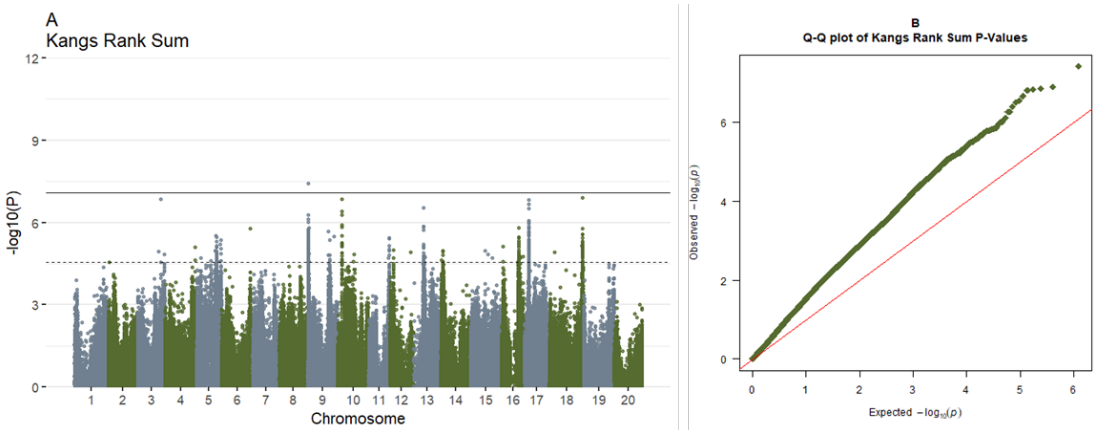
Supplementary Figure 3.7: Manhattan plot (A) and q-q plot (B) of GWAS results using the Finlay Wilkinson value as the model phenotype.



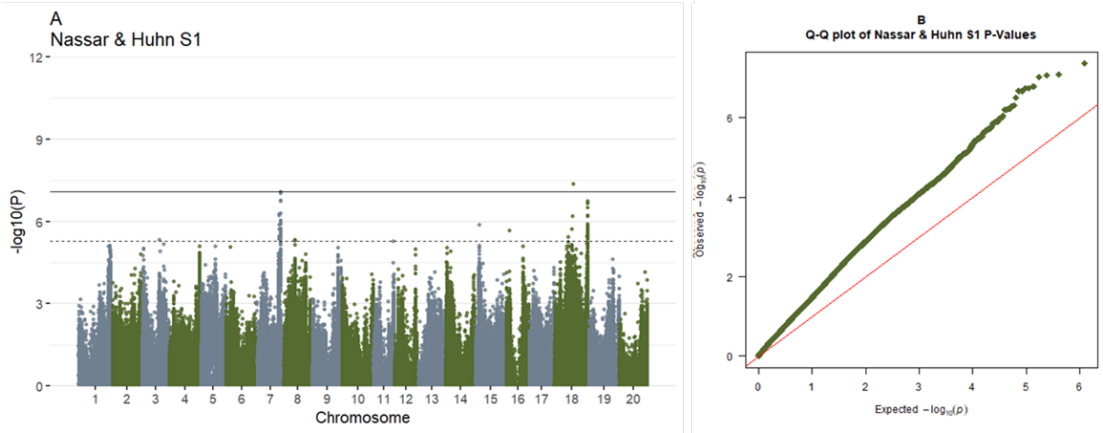
Supplementary Figure 3.8: Manhattan plot (A) and q-q plot (B) of GWAS results using the GE variance component as the model phenotype.



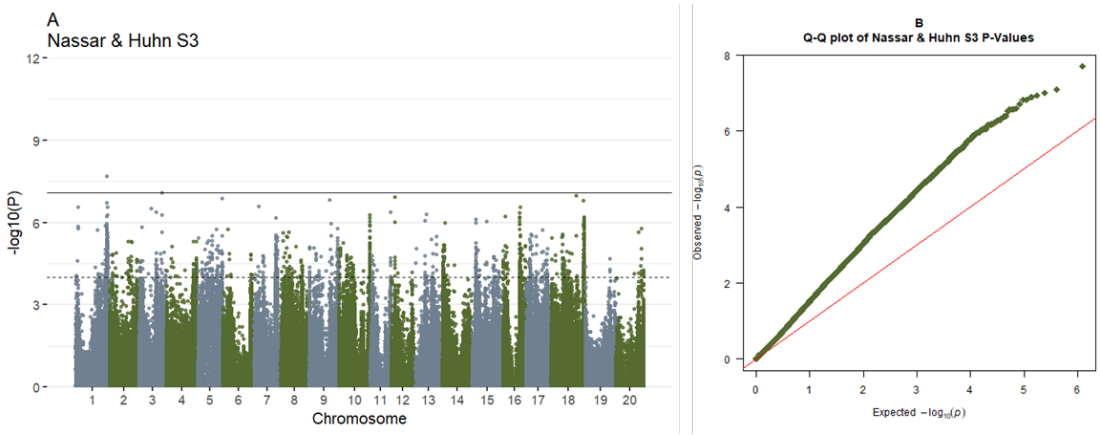
Supplementary Figure 3.9: Manhattan plot (A) and q-q plot (B) of GWAS results using the mean variance component as the model phenotype.



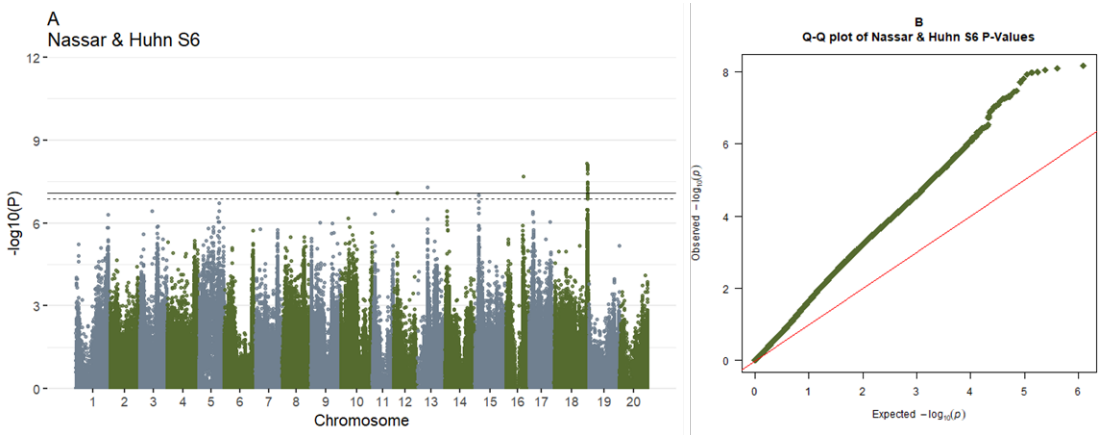
Supplementary Figure 3.10: Manhattan plot (A) and q-q plot (B) of GWAS results using Kangs Rank Sum as the model phenotype.



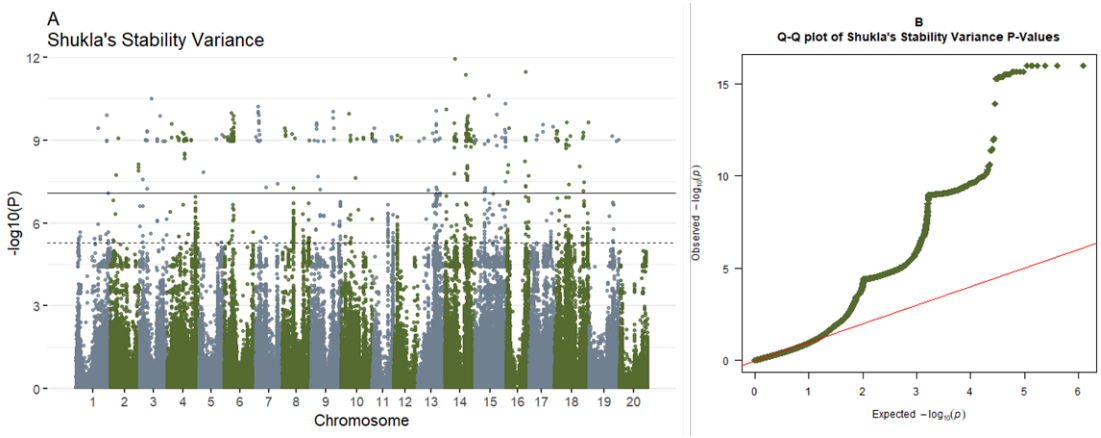
Supplementary Figure 3.11: Manhattan plot (A) and q-q plot (B) of GWAS results using the Nassar and Huhn S1 statistic as the model phenotype.



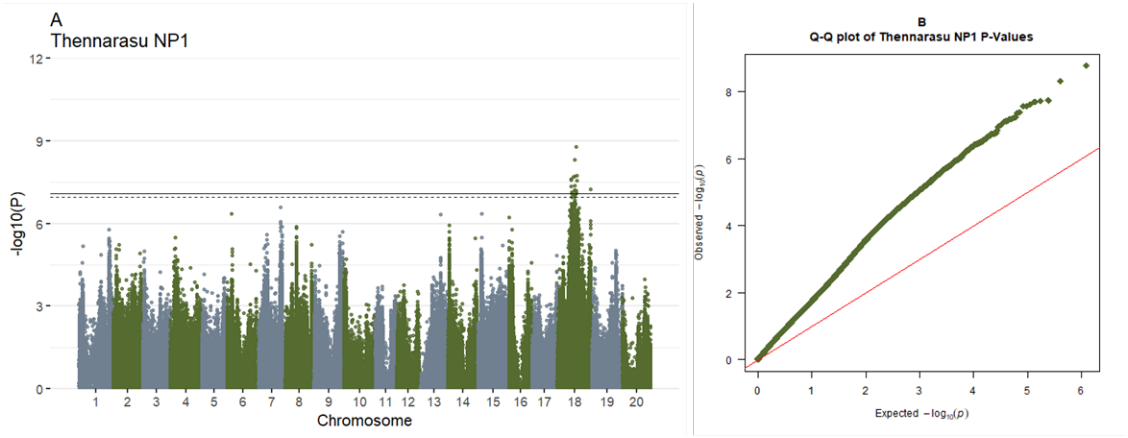
Supplementary Figure 3.12: Manhattan plot (A) and q-q plot (B) of GWAS results using the Nassar and Huhn S3 statistic as the model phenotype.



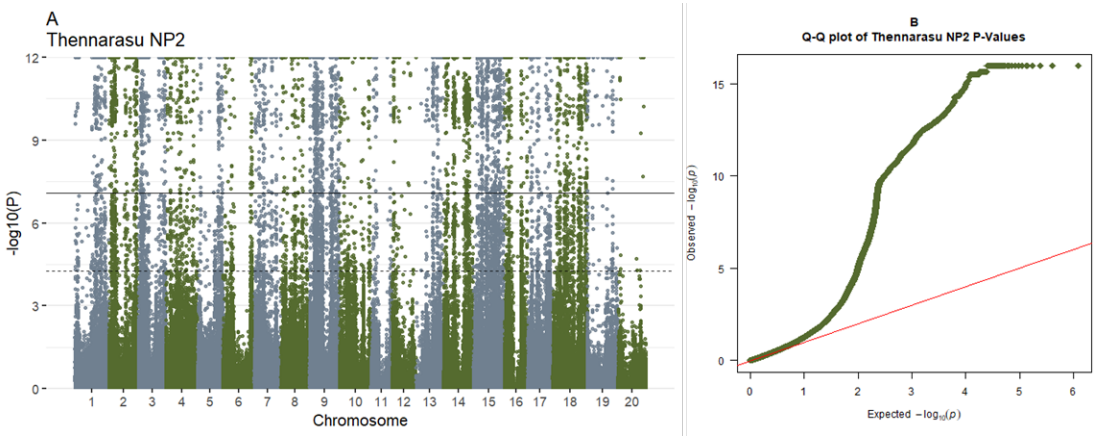
Supplementary Figure 3.13: Manhattan plot (A) and q-q plot (B) of GWAS results using the Nassar and Huhn S6 statistic as the model phenotype.



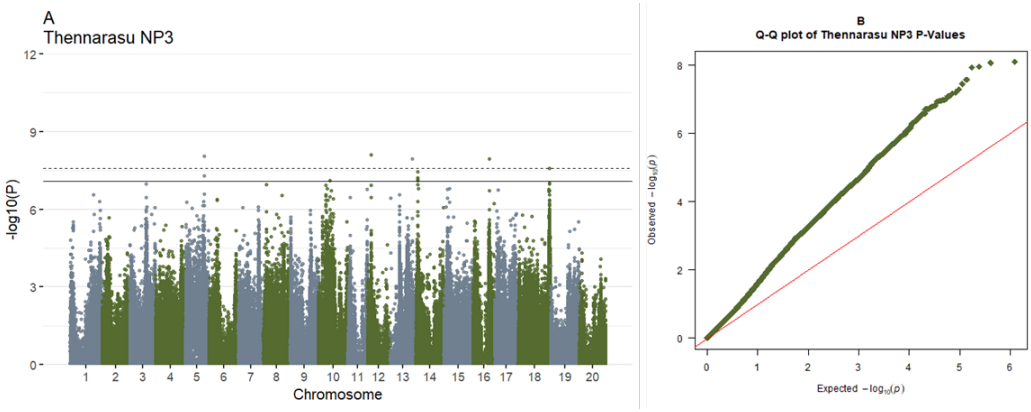
Supplementary Figure 3.14: Manhattan plot (A) and q-q plot (B) of GWAS results using Shukla's Stability Variance as the model phenotype.



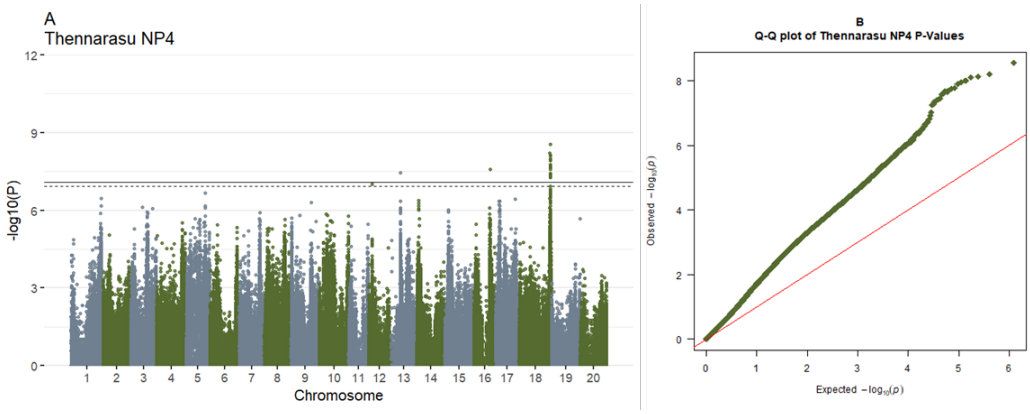
Supplementary Figure 3.15: Manhattan plot (A) and q-q plot (B) of GWAS results using the Thennarasu NP1 statistic as the model phenotype.



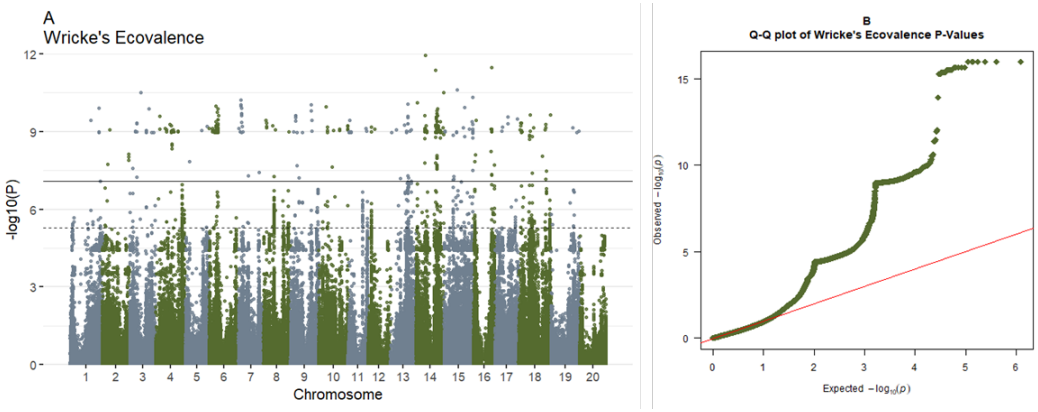
Supplementary Figure 3.16: Manhattan plot (A) and q-q plot (B) of GWAS results using the Thennarasu NP2 statistic as the model phenotype.



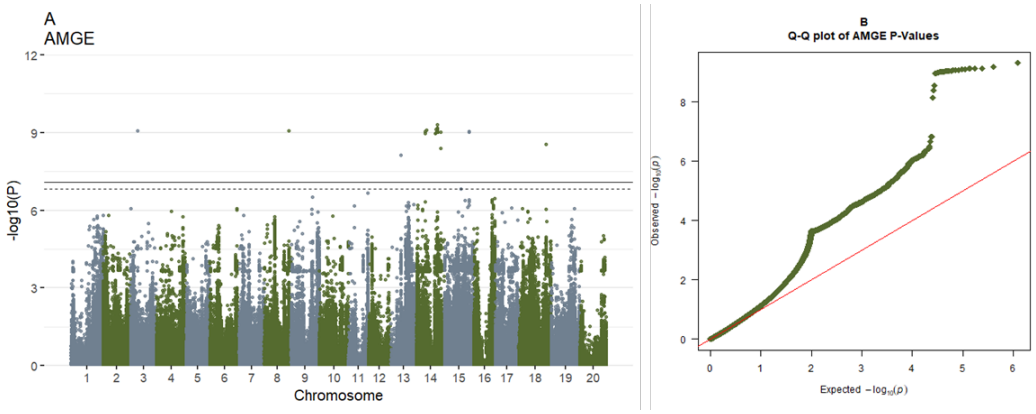
Supplementary Figure 3.17: Manhattan plot (A) and q-q plot (B) of GWAS results using the Thennarasu NP3 statistic as the model phenotype.



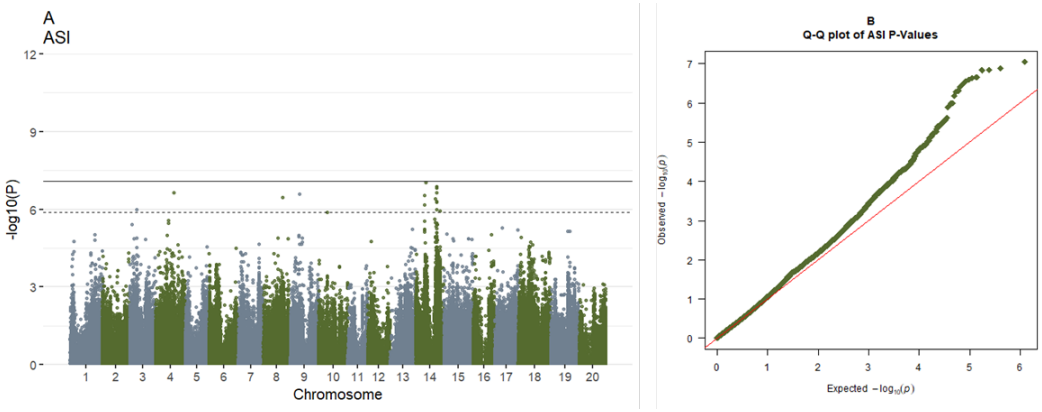
Supplementary Figure 3.18: Manhattan plot (A) and q-q plot (B) of GWAS results using the Thennarasu NP4 statistic as the model phenotype.



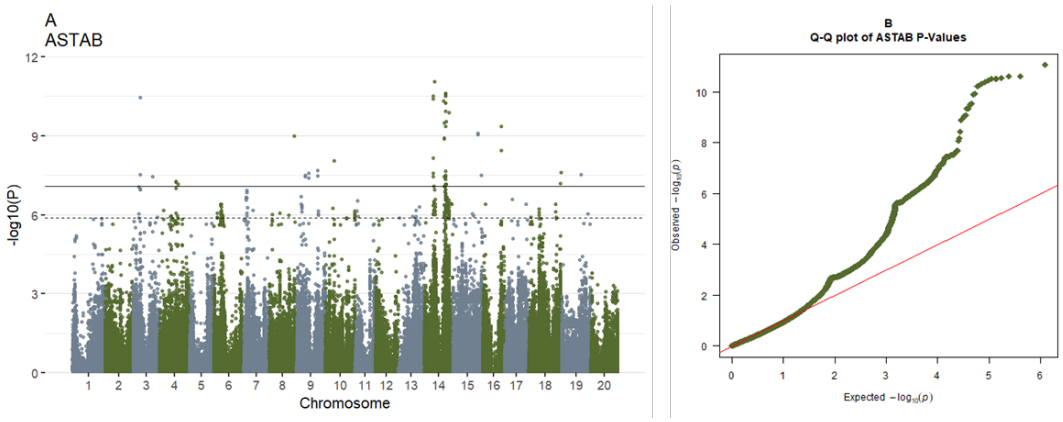
Supplementary Figure 3.19: Manhattan plot (A) and q-q plot (B) of GWAS results using Wricke's Ecovalence as the model phenotype.



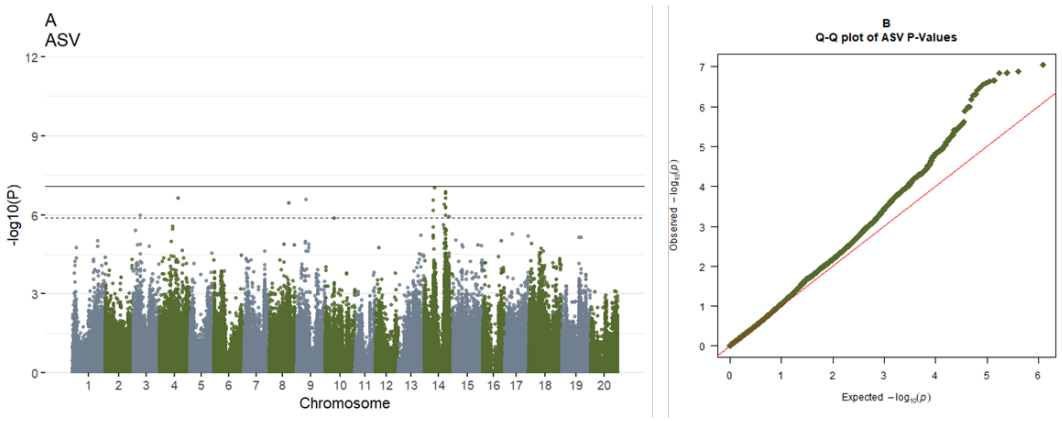
Supplementary Figure 3.20: Manhattan plot (A) and q-q plot (B) of GWAS results using the sum across environments of GEI modeled by AMMI as the model phenotype.



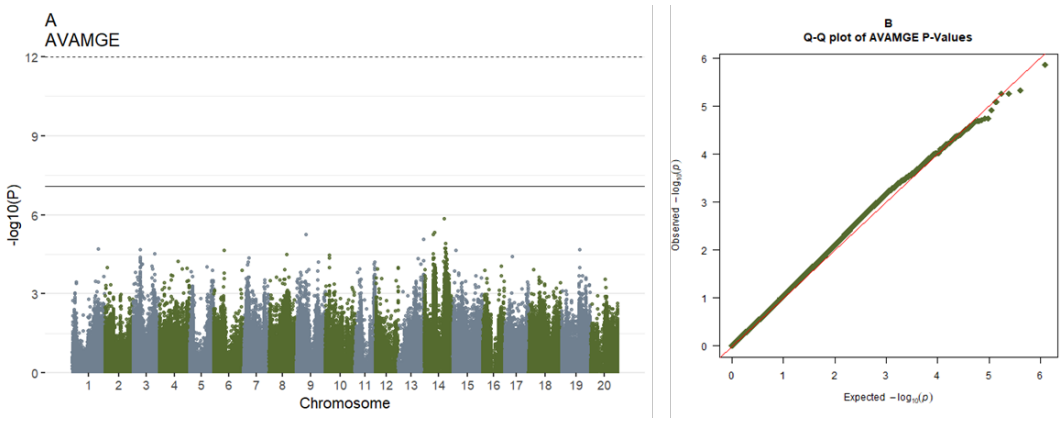
Supplementary Figure 3.21: Manhattan plot (A) and q-q plot (B) of GWAS results using the AMMI stability value as the model phenotype.



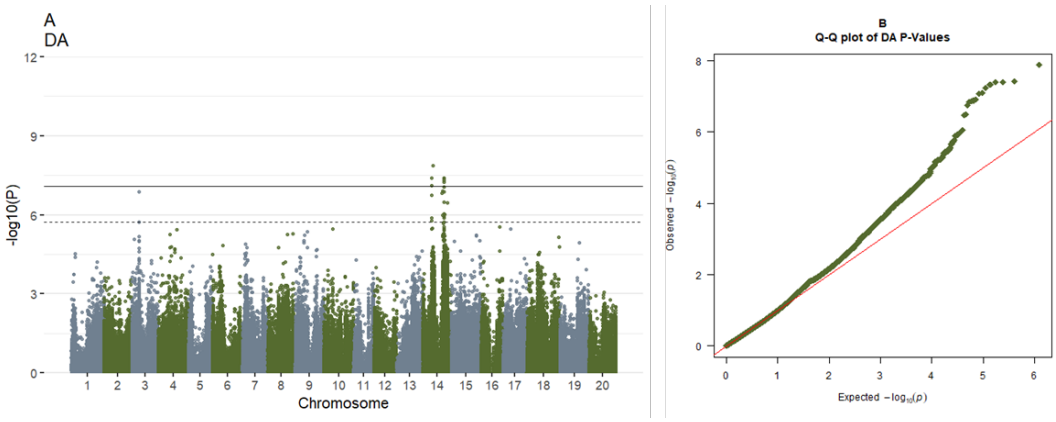
Supplementary Figure 3.22: Manhattan plot (A) and q-q plot (B) of GWAS results using the AMMI based stability parameter as the model phenotype.



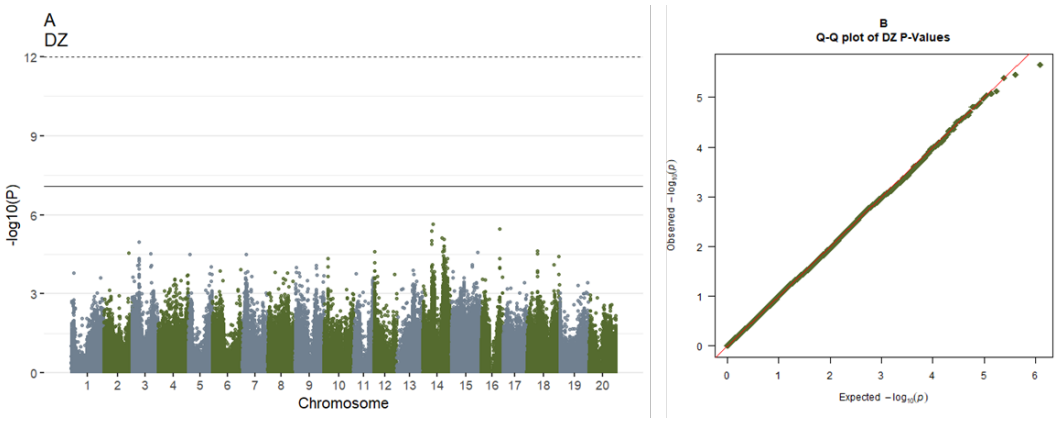
Supplementary Figure 3.23: Manhattan plot (A) and q-q plot (B) of GWAS results using the AMMI stability value as the model phenotype.



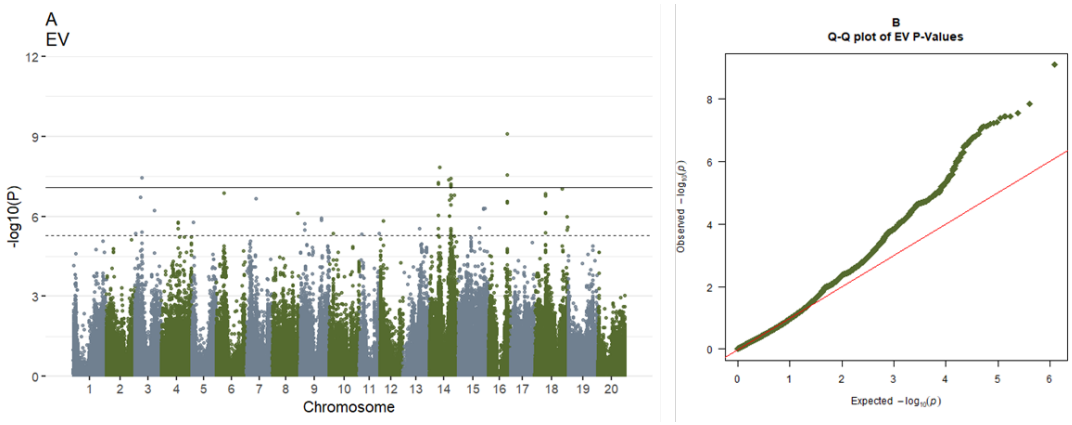
Supplementary Figure 3.24: Manhattan plot (A) and q-q plot (B) of GWAS results using the sum across environments of absolute value of GEI modelled by AMMI as the model phenotype.



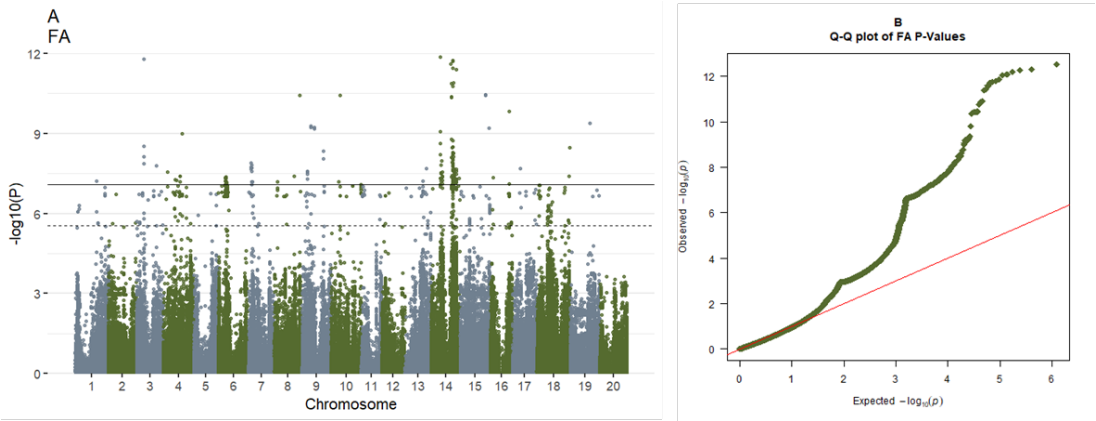
Supplementary Figure 3.25: Manhattan plot (A) and q-q plot (B) of GWAS results using Annicchiarico's D parameter as the model phenotype.



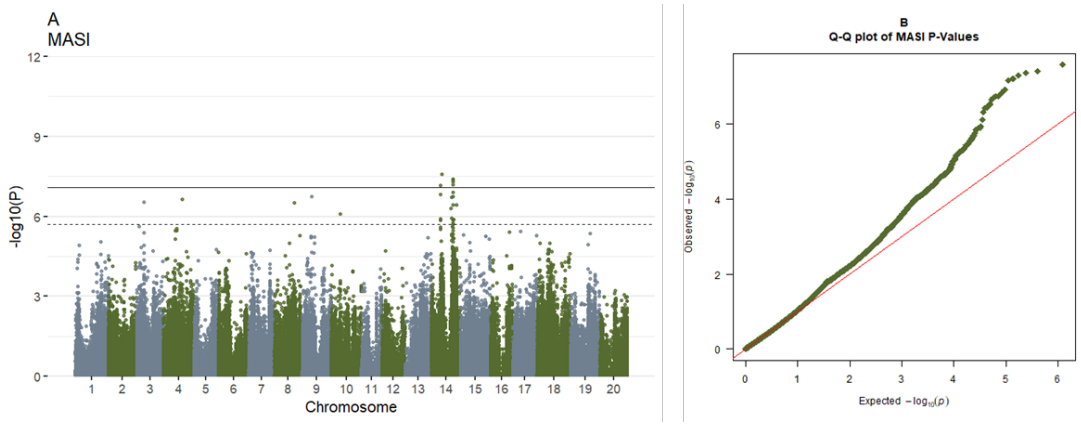
Supplementary Figure 3.26: Manhattan plot (A) and q-q plot (B) of GWAS results using Zhang's D parameter as the model phenotype.



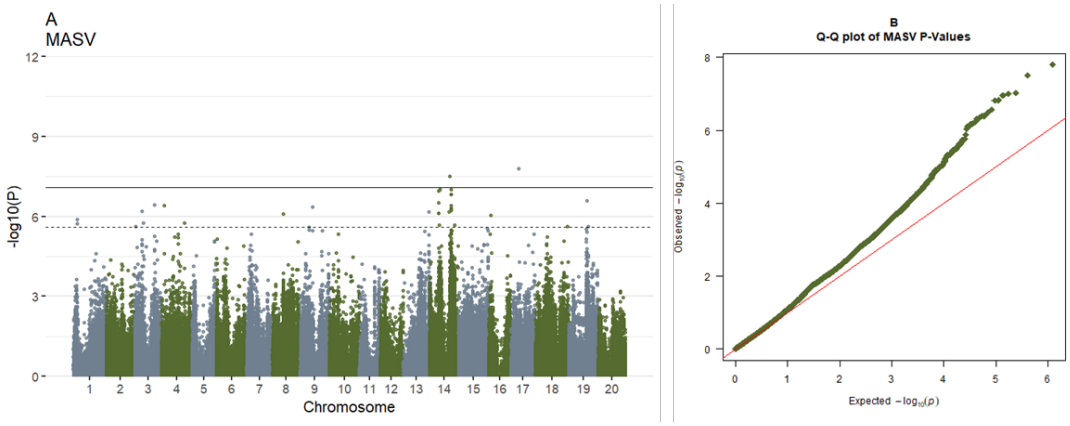
Supplementary Figure 3.27: Manhattan plot (A) and q-q plot (B) of GWAS results using the averages of the squared eigenvector values as the model phenotype.



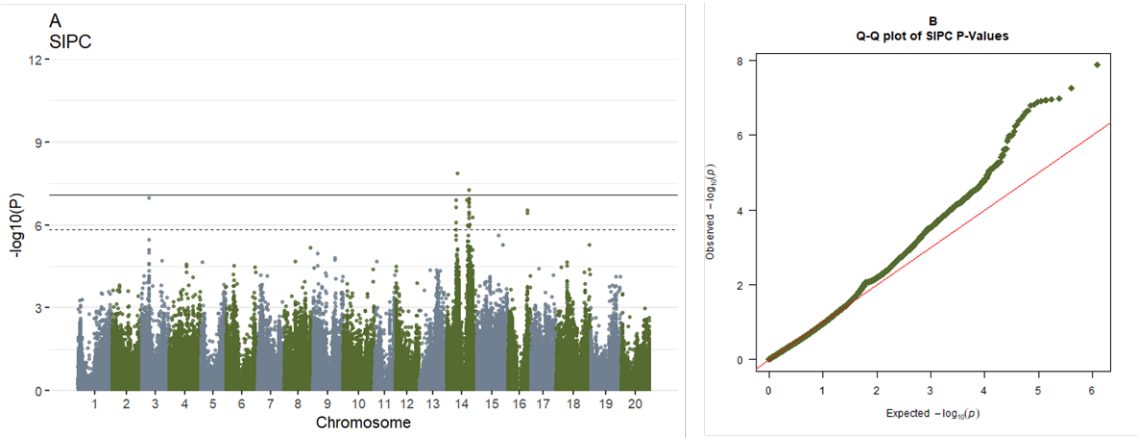
Supplementary Figure 3.28: Manhattan plot (A) and q-q plot (B) of GWAS results using the stability measure based on fitted AMMI model value as the model phenotype.



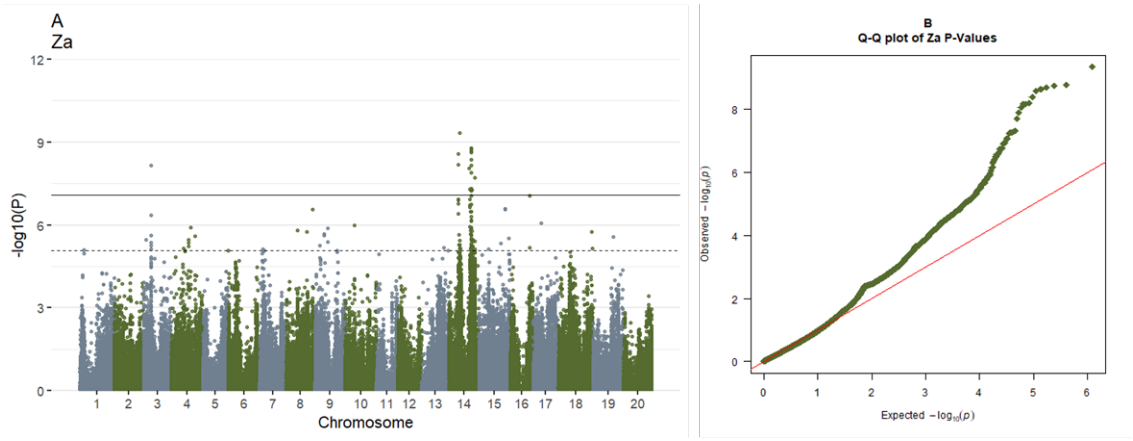
Supplementary Figure 3.29: Manhattan plot (A) and q-q plot (B) of GWAS results using the modified AMMI stability index as the model phenotype.



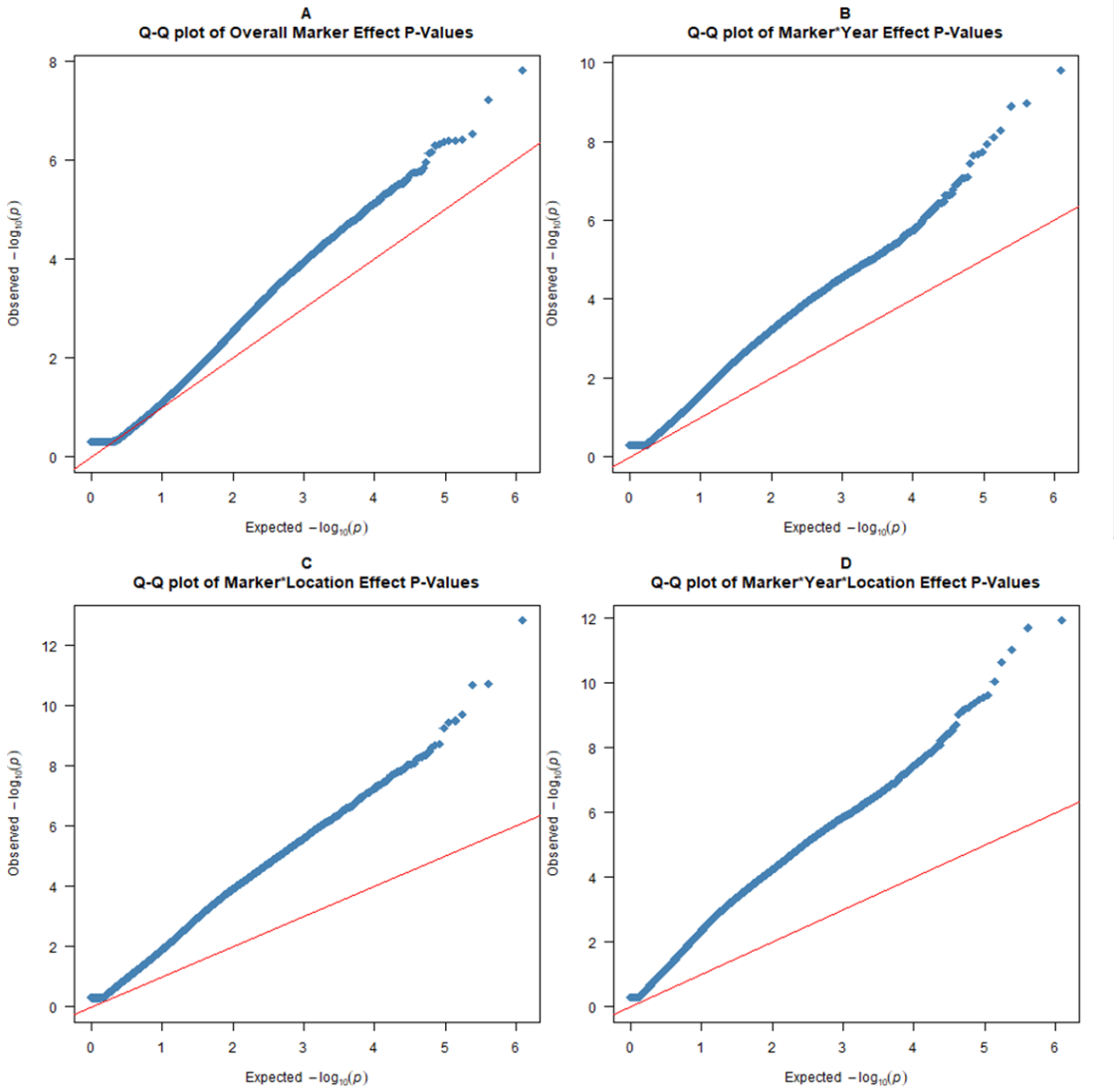
Supplementary Figure 3.30: Manhattan plot (A) and q-q plot (B) of GWAS results using the modified AMMI stability value as the model phenotype.



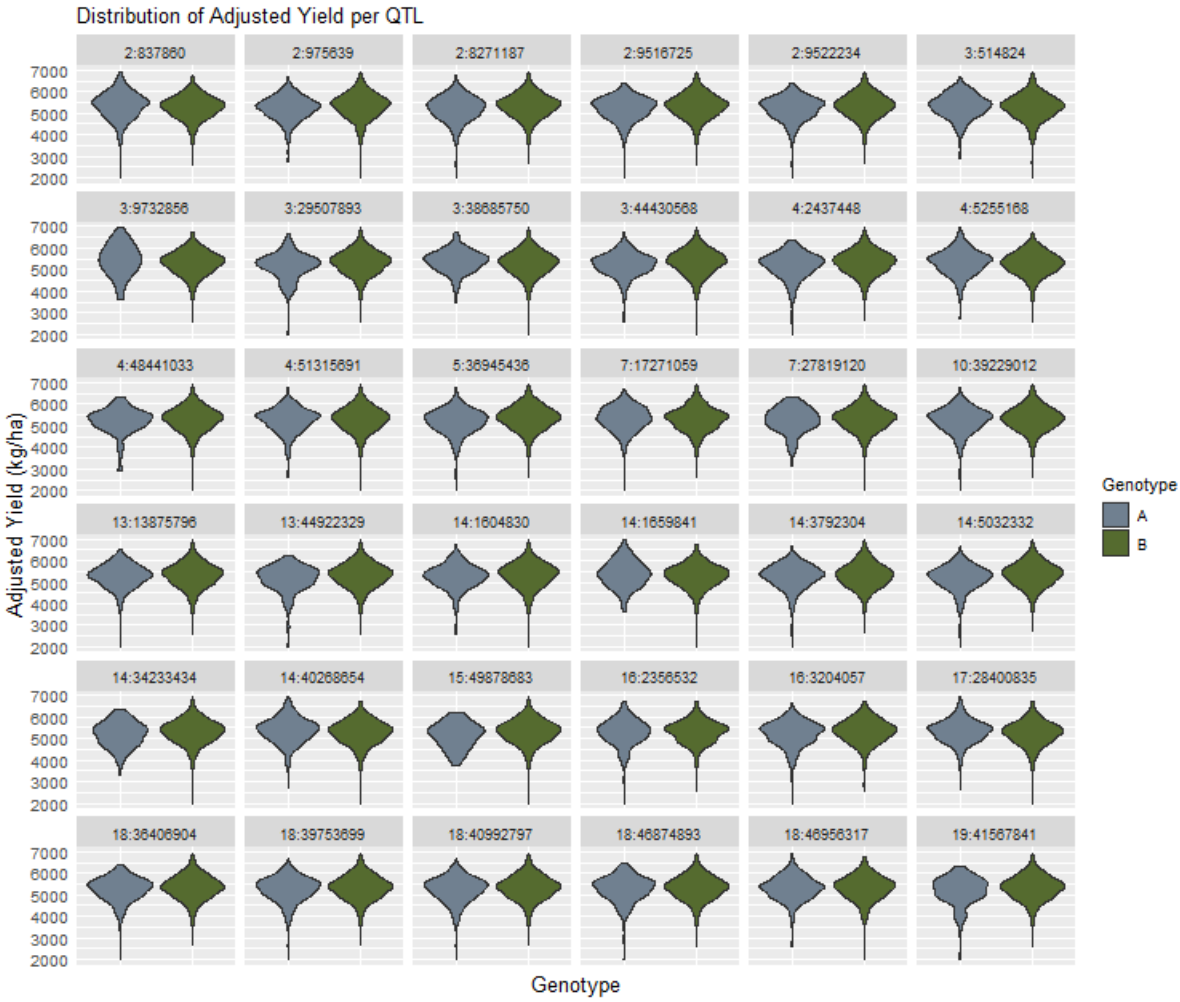
Supplementary Figure 3.31: Manhattan plot (A) and q-q plot (B) of GWAS results using the sums of the absolute value of the IPC scores as the model phenotype.



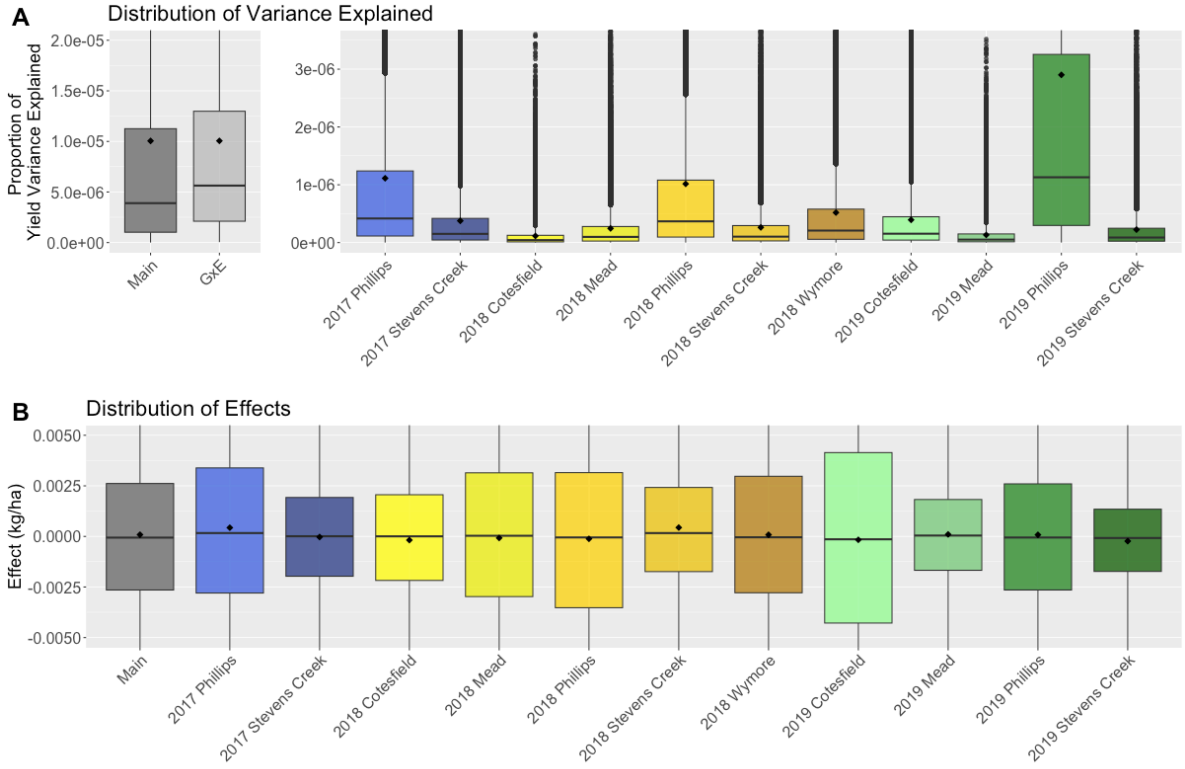
Supplementary Figure 3.32: Manhattan plot (A) and q-q plot (B) of GWAS results using the absolute value of the relative contribution of IPCs to the interaction as the model phenotype.



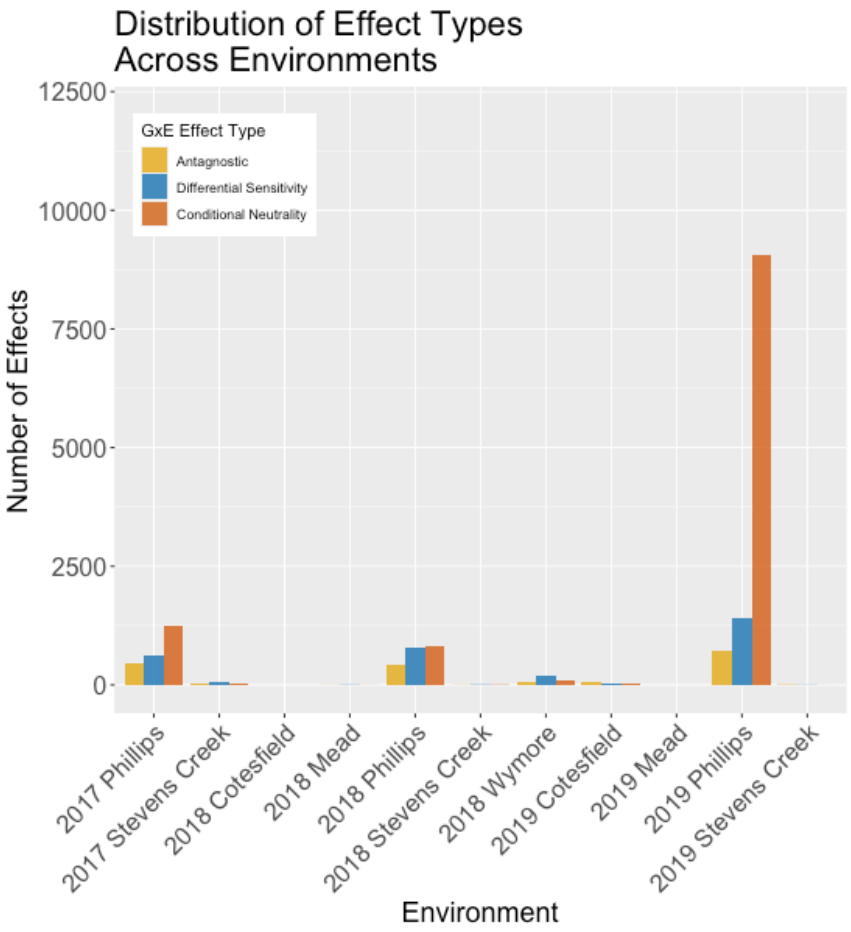
Supplementary Figure 3.33: Q-Q plots of each of the various marker interaction levels of explicitly modeling GxE interactions show that the results from more complex interactions are more inflated.



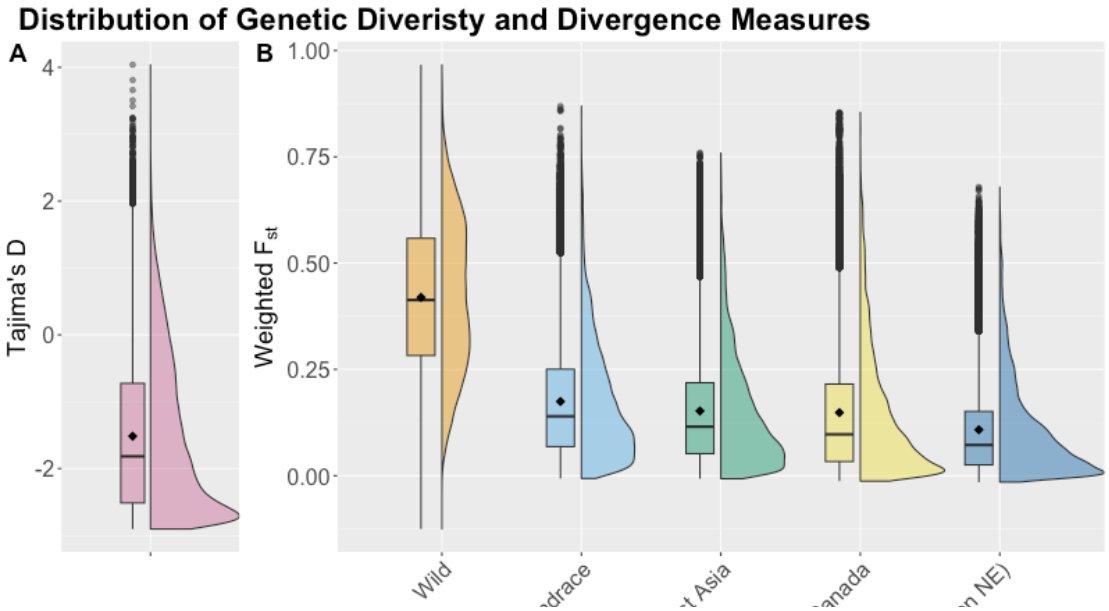
Supplementary Figure 3.34: Comparing the adjusted yield distributions of GxE QTL across environments shows that many QTL have no obviously advantageous allele when assessing the pooled data.



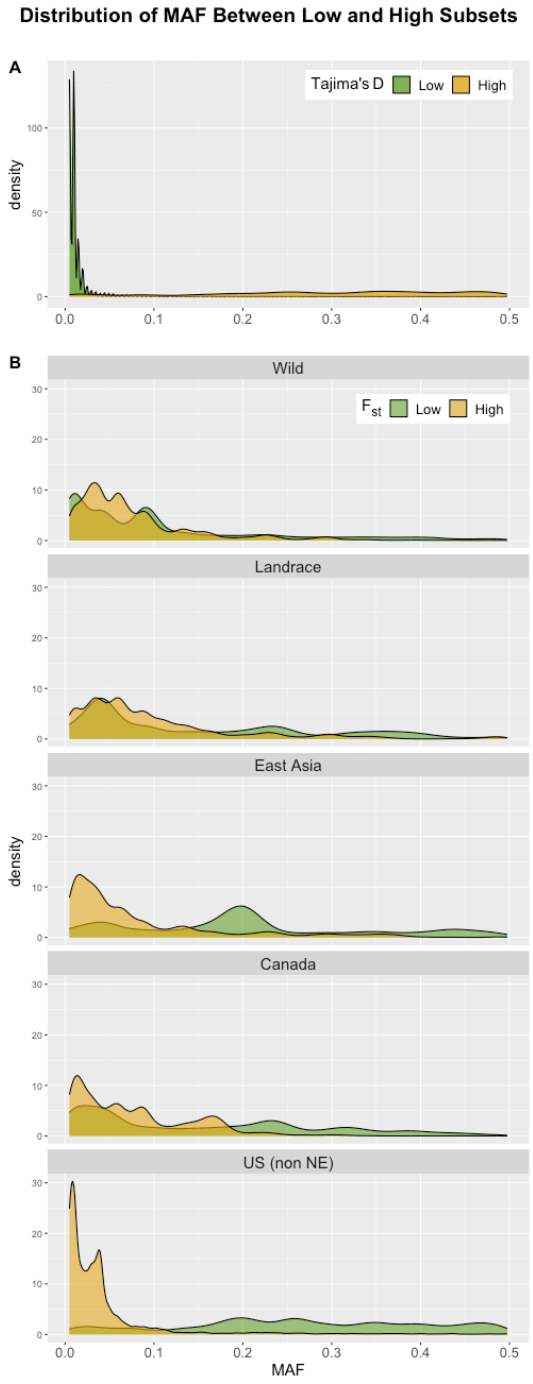
Supplementary Figure 4.1: A) The distribution of proportion of variance explained by GxE effects per window is skewed slightly higher than main genetic effects. This is made up of a variety of per environment variance distributions B) The distribution of per window effect sizes is varied among environments.



Supplementary Figure 4.2: The overwhelming majority of GxE interactions occur in the 2019 Phillips location.



Supplementary Figure 4.3: A) The distribution of per window Tajima's D values is skewed towards the negative. B) The distribution of per window weighted pairwise F_{st} values is skewed continuously lower as the comparison is made to theoretically more related material.



Supplementary Figure 4.4: The minor allele frequency distribution of markers in windows for the top and bottom 5% of Tajima's D (A) and pairwise weighted F_{st} values (B) shows clear differences, which is an important confounding factor to control for in this type of analysis.