GENOMIC SELECTION FOR KANSAS WHEAT

by

ROBERT C GAYNOR

M.S., Oregon State University, 2010

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Agronomy
College of Agriculture

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2015

Abstract

Wheat breeders are constantly working to develop new wheat varieties with improved performance for agronomically important traits such as yield and disease resistance. Identifying better ways of phenotyping germplasm, developing methods for predicting performance based on genetic information, and identifying novel sources of genetic disease resistance can all improve the efficiency of breeding efforts. Three studies relating to these research interests were conducted. Synthetic hexaploid wheat lines were screened for resistance to root-lesion nematodes, an economically important pest of wheat. This resulted in the identification of three lines resistant to the root-lesion nematode species *Pratylenchus thornei*. Grain yield data from multi-location yield trials and average yields for counties in Kansas were used to identify wheat production areas in Kansas. Knowledge obtained from this study is useful for both interpreting data from yield trials and deciding where to place them in order to identify new higher yielding varieties. These data also aided the final research study, developing a genomic selection (GS) model for yield in the Kansas State University wheat breeding program. This model was used to assess the accuracy of GS in conditions experienced in a breeding project. Available measurements of GS have been constructed using simulations or using conditions not typical of those experienced in a wheat breeding program. The estimate of accuracy determined in this study was less than many of the reported measurements. This measure of accuracy will aid in determining if GS is a cost efficient tool for use in wheat breeding.

GENOMIC SELECTION FOR KANSAS WHEAT

by

ROBERT C GAYNOR

M.S., Oregon State University, 2010

A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Agronomy
College of Agriculture

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2015

Approved by:

Major Professor
Dr. Allan Fritz

Copyright

Abstract

Wheat breeders are constantly working to develop new wheat varieties with improved performance for agronomically important traits such as yield and disease resistance. Identifying better ways of phenotyping germplasm, developing methods for predicting performance based on genetic information, and identifying novel sources of genetic disease resistance can all improve the efficiency of breeding efforts. Three studies relating to these research interests were conducted. Synthetic hexaploid wheat lines were screened for resistance to root-lesion nematodes, an economically important pest of wheat. This resulted in the identification of three lines resistant to the root-lesion nematode species *Pratylenchus thornei*. Grain yield data from multi-location yield trials and average yields for counties in Kansas were used to identify wheat production areas in Kansas. Knowledge obtained from this study is useful for both interpreting data from yield trials and deciding where to place them in order to identify new higher yielding varieties. These data also aided the final research study, developing a genomic selection (GS) model for yield in the Kansas State University wheat breeding program. This model was used to assess the accuracy of GS in conditions experienced in a breeding project. Available measurements of GS have been constructed using simulations or using conditions not typical of those experienced in a wheat breeding program. The estimate of accuracy determined in this study was less than many of the reported measurements. This measure of accuracy will aid in determining if GS is a cost efficient tool for use in wheat breeding.

Table of Contents

List of Figures

List of Tables

Chapter 1 - Root-Lesion Nematode (*Pratylenchus spp.*) Resistance in Synthetic Hexaploid Wheat

Abstract

Genetic resistance in bread wheat [*Triticum aestivum* L. (2n=6x=42, AABBDD)] to root-lesion nematodes (*Pratylenchus* spp.) in Kansas has not been identified. Synthetic hexaploid wheats (2n=6x=42, AABBDD) are a potential source of resistance to root-lesion nematodes based on screenings conducted in Australia. A set of 200 synthetic wheats and 10 bread wheats was screened for resistance to *P. thornei* in a greenhouse study using two replications. The 40 most resistant and three least resistant synthetic wheats were reexamined with five bread wheats in a four replication screening for resistance to *P. thornei* and a four replication screening for resistance to *P. neglectus*. Both screenings showed significant ($P < 0.01$ for *P. neglectus* and $P < 0.001$ for *P. thornei*) differences between lines, but no correlation was observed between resistance to *P. thornei* and *P. neglectus* ($r = -0.1$, $P = 0.5$). Multiple comparison testing revealed synthetic wheats with resistance to *P. thornei* better than the screened bread wheats (Tukey adjusted $P < 0.05$), but not for *P. neglectus*. A six replication screening of *P. thornei* resistance confirmed two synthetic wheats with resistance to *P. thornei* better than the best tested Kansas variety (Tukey adjusted $P < 0.05$). These two synthetic wheats are a potential source of genetic resistance to *P. thornei* that can be utilized for breeding more resistant bread wheats.

Introduction

Root-lesion nematodes (*Pratylenchus* spp.) are an economically important pest to bread wheat [*Triticum aestivum* L. (2n=6x=42, AABBDD)] production. They have been estimated to cost the Australian wheat industry $36 million per year (Brennan and Murray, 1998). A study conducted in the Pacific Northwest (Oregon, Washington, and Idaho) estimated root-lesion nematodes to be responsible for up to a five percent reduction in yield resulting in a loss of up to $51 million per year (Smiley, 2009). In Kansas, they are estimated to reduce average annual grain yield by two percent (Appel et al., 2013). Field surveys have identified *P. neglectus* as the most common species of root-lesion nematode in Kansas and *P. thornei*, the second most common, occurred in just one percent of soil samples (Todd et al., 2014).

The best way to control root-lesion nematodes in wheat is to plant resistant varieties. However, there are no varieties adapted to Kansas that have been identified as resistant. To develop resistant varieties adapted to Kansas, a source of resistance to nematodes found in Kansas must first be identified. Synthetic hexaploid wheat (2n=6x=42, AABBDD), produced by the hybridization of *Triticum turgidum* L. subsp. *durum* (2n=4x=28, AABB) and *Aegilops tauschii* (2n=2x=14, DD), has been identified as a potential source of resistance to root-lesion nematodes in studies conducted in Australia (Thompson, 2008). These synthetic wheats can readily be crossed with bread wheat to transfer resistance genes (Gill and Raupp, 1987). One synthetic wheat, CPI133872, has been crossed to bread wheat to produce a genetic mapping population (Zwart et al., 2005). Quantitative Trait Loci (QTLs) identified in this mapping population showed alleles conferring resistance from the synthetic wheat line were present on chromosomes 6D and 4B, consistent with previous work (Zwart et al., 2005). Since screening of synthetic wheats was carried out on nematodes found in Australia, it is unclear if their resistance would be effective on nematode populations found in Kansas. To address this issue, this study seeks to screen synthetic wheats for resistance to *P. thornei* and *P. neglectus* populations found in Kansas. The goal is to identify a synthetic wheat line that can be used as a source of resistance for breeding new resistant wheat varieties.

Materials and Methods

Plant Material

- 200 synthetic wheats produced by CIMMYT (International Maize and Wheat Improvement Center); seed obtained from J. Fellers was used in the first screening; seed from J. Raupp and B. Gill was used in subsequent screenings
- Five wheat varieties adapted to Kansas were used as checks: Jagger, Karl 92, Armour, Overley, and Everest
- Five wheat varieties from Oregon and Washington that have been screened for tolerance to root-lesion nematodes or are derived from lines which have been screened (Smiley, 2009): Alpowa, Goetze, ORCF-102, Tubbs 06, and OR4081056
- An association mapping panel containing 234 common wheat varieties from the Great Plains region; seed provided by Eduard Akhunov

Screening Trials

Screenings for synthetics were conducted in four separate greenhouse trials. The first trial screened all synthetic wheats, Kansas checks, and Oregon and Washington lines for *P. thornei* resistance. This trial was ran in two blocks, spaced two weeks apart. Each block included one entry for each synthetic wheat and up to three replications of each check. The Oregon and Washington varieties were only present in the second block due to unavailability of seed at time of planting. The second and third screening trials examined the Kansas checks and the 40 most resistant and three most susceptible synthetic wheats as identified in the first trial. Each trial was laid out in a randomized complete block design with four blocks. The final trial screened for *P. thornei* resistance using the Kansas checks and six synthetic wheats representing the four most and two least resistant lines to *P. thornei* as determined in the second *P. thornei* screening. This trial used six replications in a completely randomized design. The association mapping panel was used to screen for resistance to *P. thornei*. This trial used a randomized complete block design containing four blocks blocked by time. Each block was spaced one to two months apart.

Screening Protocol

Two seeds per replicate were planted in 500 cm^3 tubes. Soil and nematodes for screening were collected at sites previously identified to have high incidence of a single root-lesion nematode species. Soil for *P. thornei* screenings was collected from a field near Conway Springs, Kansas. A field near Caldwell, Kansas was used for *P. neglectus* screening. Prior to planting, soil was homogenized and the number of nematodes was determined. If nematode counts were low, additional nematodes from previous screenings were added. The first two screenings for *P. thornei* used nematodes already present in the soil and each trial contained between 2000 and 2400 nematodes per tube. The final *P. thornei* screening used additional nematodes from an earlier screening applied 2 cm below the seeds to bring the number of nematodes per tube to 1500. The *P. neglectus* screening contained about 370 nematodes per tube already present in the soil and an additional 720 nematodes per tube applied 2 cm below the seed. For the association mapping trial, the first replication used just the nematodes already present in the soil (520 nematodes per tube). All subsequent replications used additional nematodes bring the total number per tube to 1,600 to 1,800.

Plants were watered from the top as needed and grown for 8 to 9 weeks at 25°C. At which time, the roots were extracted, cleaned, and transferred to 250 ml beakers containing about 125 ml of water. The beakers were placed on racks and aerated to collect nematodes inside the roots. After one week, the water was passed through a sieve and the roots were rinsed above the sieve to collect nematodes. The volume of water containing collected nematodes was brought to 20 ml for each sample. A 4 ml aliquot was taken, placed on a counting plate, and the number of nematodes present was determined by counting half of the plate with a dissecting microscope. Roots were dried and weighed. The number of nematodes per gram of root was calculated. This number was log transformed and analyzed in an ANOVA using R (R Core Team, 2013). One-way, simultaneous multiple comparison tests were conducted in the four replication trials to determine if any synthetic wheats were more resistant than any of the checks. A Tukey's honest significant difference (HSD) test was used on the six replication screening trial to determine which wheats were significantly different for nematodes per gram of root.

Results

An ANOVA for the initial two replication screening trial for *P. thornei* resistance found statistically significant (P < 0.001) differences between tested lines. The average number of nematodes per gram of root for each line ranged from 325 to 8704 (Figure 1.1). The four replication screening for *P. thornei* resistance also showed statistically significant (P < 0.001) differences between lines (Figure 1.2). A one-way, simultaneous multiple comparison test identified fourteen synthetics significantly (P < 0.05) more resistant than one to four of the checks (results not shown). No synthetic was significantly more resistant than all five checks. The four replication screening for *P. neglectus* resistance had an ANOVA with statistically significant (P < 0.01) differences between lines (Figure 1.3). However, no synthetic had resistance significantly less than any of the checks in the one-way, simultaneous multiple comparison test. No correlation was observed between resistance to *P. thornei* and resistance to *P. neglectus* (r = -0.1, P = 0.5). Statistically significant (P < 0.001) differences were observed in the six replication screening for *P. thornei* resistance (Figure 1.4). A Tukey's HSD test showed three of the four resistant synthetic lines tested were significantly (P < 0.05) more resistant than the check varieties (Table 1.1). The two susceptible synthetic lines were significantly (P < 0.05) more susceptible than all of the checks except for Karl 92. Among the checks, only Armour and

Karl 92 were significantly different (P = 0.008). However, they weren't significantly different in the four replication trial for *P. thornei*.

Very few nematodes were recovered in the fourth replication of the association mapping panel screening, so this replication was discarded. No significant difference was found between association mapping panel lines using the first three replications. A calculation of statistical power was performed to determined how many replications would be needed to find significant differences between the observed differences in the first three replications. This test determined that 11 replications would be needed. This was taken as an indication that there were not any meaningful differences in genetic resistance to *P. thornei* between association panel lines.

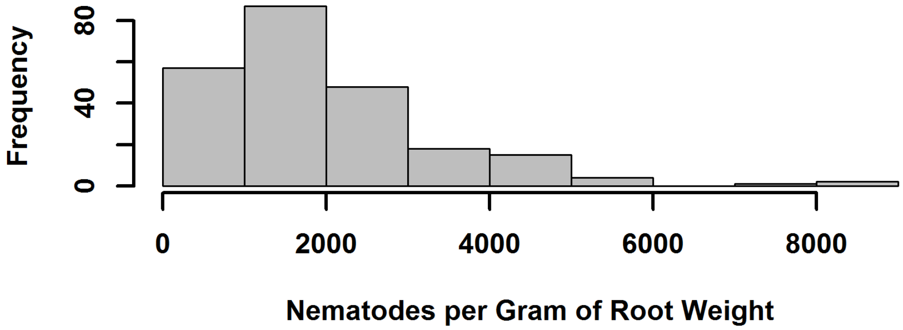

Figure 1.1: Histogram of average *P. thornei* nematodes per gram of root for synthetic wheats examined in the two replication trial.

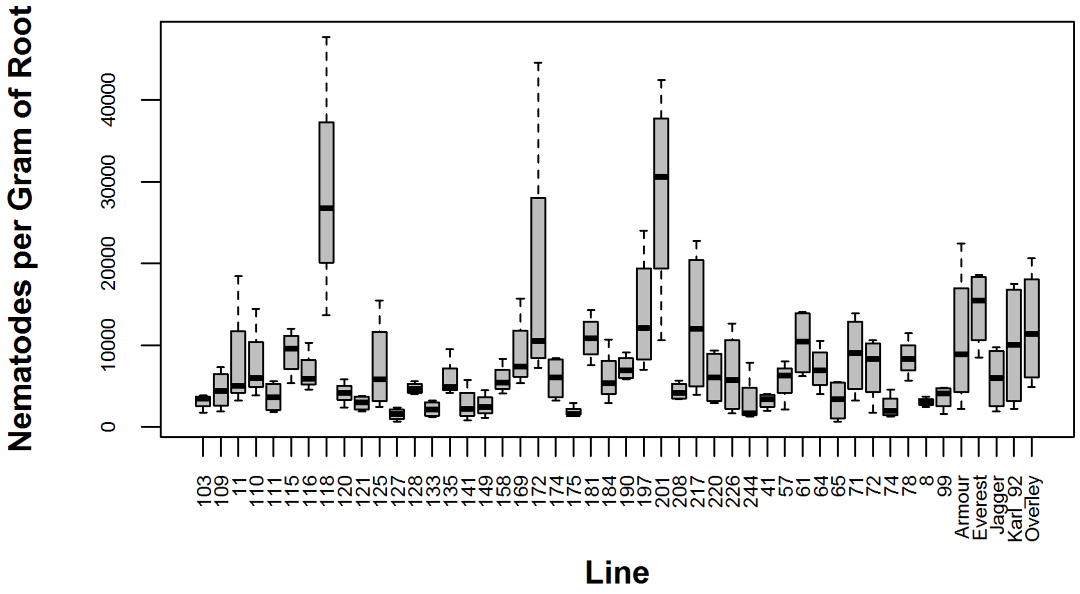

Figure 1.2: Box plot of nematodes per gram of root for lines[*] evaluated in the four replication *P. thornei* screening trial.

[*] Numbers correspond to CIMMYT's synthetic ID numbers (Mujeeb-Kazi and Hettel, 1995)

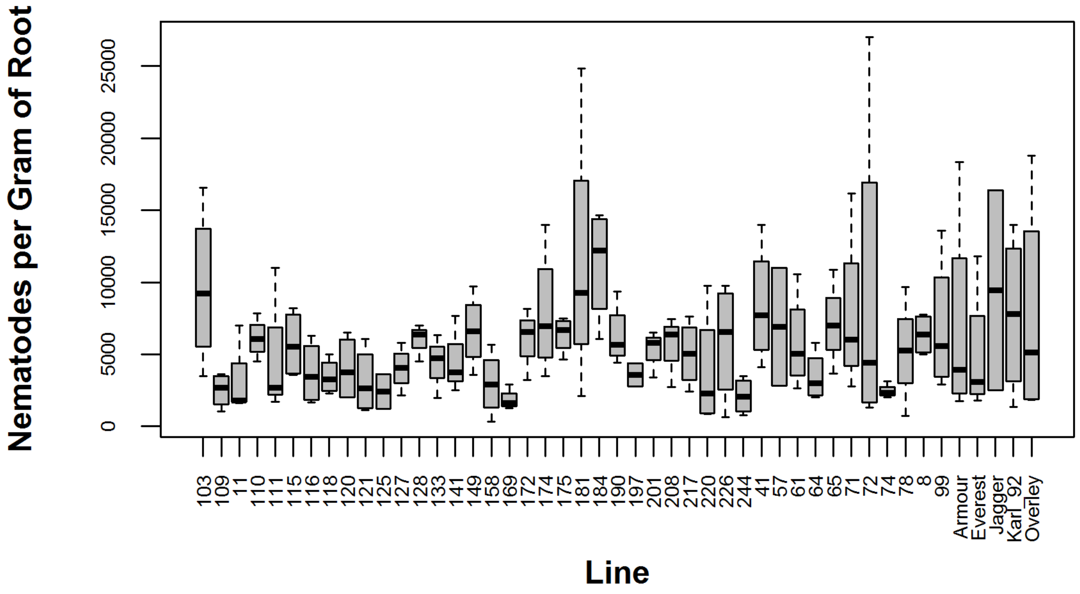

Figure 1.3: Box plot of nematodes per gram of root for lines[*] evaluated in the four replication *P. neglectus* screening trial.

[*]Numbers correspond to CIMMYT's synthetic ID numbers (Mujeeb-Kazi and Hettel, 1995)

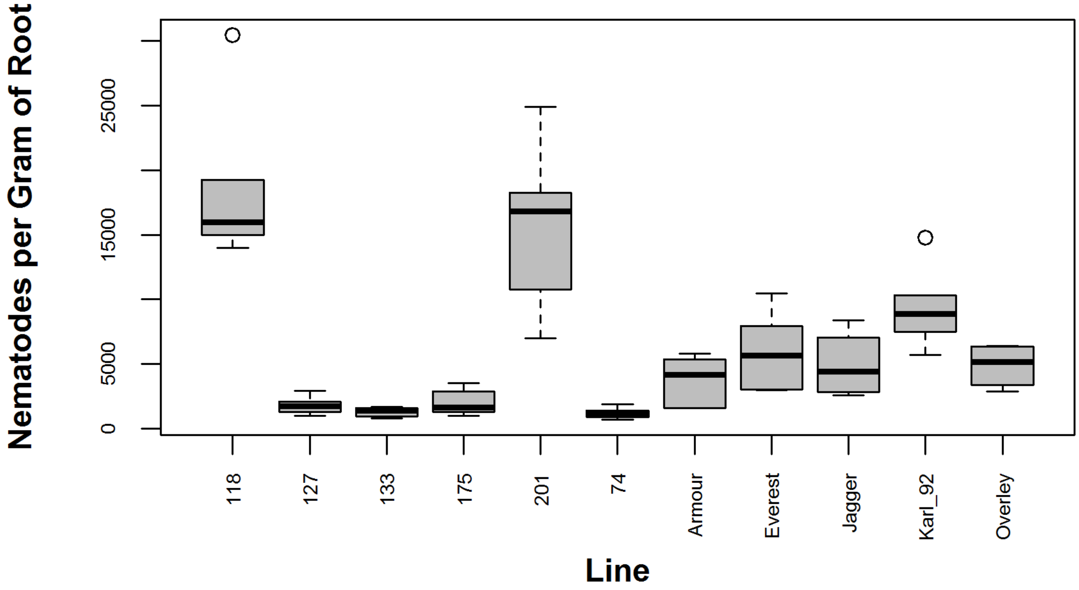

Figure 1.4: Box plot of nematodes per gram of root for lines[*] evaluated in the six replication *P. thornei* screening trial.

[*] Numbers correspond to CIMMYT's synthetic ID numbers (Mujeeb-Kazi and Hettel, 1995)

Table 1.1: P-values for Tukey's honest significant difference test of nematodes per gram of root in the six replication *P. thornei* screening trial.

Lines[1]:	118	127	133	175	201	74	Armour	Everest	Jagger	Karl 92
127	***									
133	***	NS								
175	***	NS	NS							
201	NS	***	***	***						
74	***	NS	NS	NS	***					
Armour	***	NS	**	NS	***	**				
Everest	***	**	***	**	**	***	NS			
Jagger	***	*	***	*	***	***	NS	NS		
Karl 92	NS	NS	***	***	NS	***	**	NS	NS	
Overley	***	NS	***	*	**	***	NS	NS	NS	NS

[1] Numbers correspond to CIMMYT's synthetic ID numbers (Mujeeb-Kazi and Hettel, 1995)

NS, *, **, and *** represent P-values > 0.05, < 0.05, < 0.01, and < 0.001, respectively

Discussion

The three synthetic lines with resistance greater than all check varieties in the six replication trial are prime candidates for sources of genetic resistance to *P. thornei* found in Kansas. Having been found resistant in the two previous trials strengthens confidence that these lines are more resistant than the examined checks. Unfortunately, these lines were not found to be more resistant to *P. neglectus* than check varieties. This species is the predominate root-lesion nematode species in Kansas, so resistance to *P. thornei* alone is not sufficient for producing a root-lesion nematode resistant variety for Kansas. Lack of correlation between *P. thornei* and *P. neglectus* resistance suggests that to develop a variety resistant to both, re-screening of a large set of synthetics for *P. neglectus* would have to occur. Resistance to both species from different synthetics would then have to be crossed into one line to develop a wheat variety that is resistant to both species.

Chapter 2 - Identifying Production Areas for Wheat in Kansas

Abstract

Accounting for genotype-by-environment interaction (GxE) in a breeding program can be made easier with an understanding of production areas within target environments. Wheat (*Triticum aestivum* L.) production areas in Kansas were examined using data from a multi-location yield trial which tests advanced breeding lines, and estimates for average grain yield per Kansas county. Hierarchical clustering and a factor analysis was applied to both data sets to identify putative production areas. A clear division into an eastern and a western production area was observed in all analyses. Indication for some greater subdivision into smaller production areas was observed in some analyses. Limitations in the data sets examined hindered the ability to draw firm conclusions about borders for these smaller production areas and the ability to adequately examine the impact of yearly fluctuations. Collection of new data is proposed to better model production areas in Kansas by modeling dynamic production area whose borders change according to yearly fluctuations.

Introduction

Genotype by environment interaction (GxE) can make developing plant varieties with improved performance challenging. This is due to the potential for relative variety rankings for agronomically important traits to differ according to environmental conditions. This makes it unlikely, and probably impossible, for there to be a single best variety for all conditions. Identifying a best variety is thus dependent on where and under what conditions it will be grown. A convenient way of handling this challenge is to partition locations into regions where GxE is minimal. These regions will be referred to as production areas. Relative rankings within production areas are less volatile due to decreased GxE. Selecting improved varieties for a production area can then be accomplished by averaging performance over sites contained within that area.

The mega-environment strategy used by CIMMYT (International Maize and Wheat Improvement Center) is an example of this approach. Mega-environments are defined as potentially non-contiguous areas spanning multiple countries that share similar biotic and abiotic stresses, cropping system requirements, and consumer preferences (Gauch and Zobel, 1997). As implemented by CIMMYT, this strategy is used to develop germplasm with high and stable yield across a wide range of environments (Braun et al., 1996). That is, they produce broadly adapted germplasm that can be utilized in several countries. Wheat (*Triticum aestivum* L.) production in Kansas is divided into two mega-environments (Braun and Payne, 2012). This split between mega-environments occurs at a point slightly west of the center of the state and divides Kansas into eastern and western halves.

Wheat breeding at Kansas State University (KSU) focuses on developing more narrowly adapted varieties, relative to CIMMYT. This is done to maximize yield potential within Kansas. Focusing on more narrowly adapted varieties means CIMMYT's mega-environments may not minimize GxE enough to best serve KSU breeding efforts. Identifying regions similar too, but on a smaller scale than CIMMYT's mega-environments may prove useful. Identification of these areas has previously been accomplished using replicated, multi-location yield trials. For example, Peterson used data from the Northern and Southern Regional Performance Nurseries to classify testing locations in the Great Plains (1992). His study utilized only four sites in Kansas, so its resolution was rather poor for the state. A testing nursery with more locations would have to be used to achieve adequate resolution.

The Kansas Intrastate Nursery (KIN) is a replicated, multi-location test for advanced wheat lines in Kansas. This nursery typically evaluates about a dozen locations each year. The locations are split between two breeding projects at KSU, the Hays project which focuses on western Kansas and the Manhattan project which focuses on central and eastern Kansas. The greater number of locations evaluated in this nursery should allow for adequate resolution. Unfortunately, the data set has its own limitations: test locations change over years; few long-term checks are used, resulting in high yearly turnover of test varieties; and the Hays and Manhattan projects evaluate slightly different sets of varieties. Each project evaluates between 30 and 40 varieties in the KIN each year. About 10 of those varieties are not common between the two projects.

Average county yield data may be useful as an alternative to multi-location yield trials. The United States Department of Agriculture's National Agriculture Statistics Service (NASS) makes yearly estimates of average wheat yield per acre for each county in Kansas. Since there are 105 counties in Kansas, this data set has a higher degree of resolution than the KIN data.

Explicit measurement of GxE with this data isn't possible, because yields aren't listed according to cultivar. Indirect measurement of GxE may be possible. An estimate for average yield in a county is a measurement for all wheat grown in that county. This is impacted by which cultivars were grown, production practices used, and environmental factors experienced. Counties with similar yearly fluctuations for average county yield are expected to be similarly impacted by these three factors. Since these factors are tied to GxE, grouping by fluctuations in county yield is expected to be roughly equivalent to grouping by GxE.

The goal of this study is to divide Kansas into production areas. Two data sets were used to accomplish this. Data from the KIN was used to identify similar testing locations using hierarchical clustering and factor analysis. Average county yield data from the NASS was used to find groups of similar Kansas counties using hierarchical clustering and factor analysis. Results from these analyses were compared to draw conclusions about the general structure of presumptive production areas in Kansas.

Materials and Methods

Eighteen years (1996-2013) of grain yield data from the KIN were examined. Only trials with a coefficient of variation less than thirteen and locations evaluated for more than six years were considered. Average cultivar yields were calculated on a by location by year basis according to the most appropriate statistical analysis for each trial. Correlations between locations were calculated using the method described by Peterson (1992). This involved first calculating all within year correlations between locations, applying a Fisher z-transformation to these correlations, averaging the transformed values across years, and finally transforming back to correlation coefficients. A factor analysis using the principal component extraction method was performed on the resulting correlation matrix. The “psych” package in R was used to determine the recommended number of factors based on complexity one very simple structure (VSS) and Wayne Velicer's minimum average partial (MAP) (Revelle, 2013). If the two criteria disagreed, a number of factors within the recommended numbers was chosen based on interpretability of the solution. An oblimin rotation was applied to the factors to aid in interpretability. This rotation was chosen because it allows for correlated factors. Since the underlying factors represent regions, they are expected to be correlated based on their proximity to each other. Hierarchical clustering of locations was performed using the average distance method and a distance matrix calculated using one minus the correlation.

The NASS website (http://www.nass.usda.gov/Quick_Stats/index.php) was used to find average wheat yield per acre for each county in Kansas over a twenty-five year period (1983-2007). Groups of similar counties were determined using hierarchical clustering with multiscale bootstrap resampling implemented in the R package 'pvclust' (Suzuki and Shimodaira, 2006). The number of bootstraps was set to 10,000. Tree selection was performed using a 0.9 cutoff of a biased adjusted measure for the bootstrap probability known as the approximately unbiased (AU) p-value (Shimodaira, 2002). The number of clusters to retain was also examined using partitioning around medoids (PAM). A PAM clustering was performed using the 'fpc' package in R (Hennig, 2014). The optimum number of clusters was determined using the average silhouette width (ASW) criterion. A factor analysis using the principal component extraction method was also conducted to identify regions of similar counties. The number of factors to retain was determined with the same method used for the KIN data. An oblimin rotation was applied to aid in interpretability. Factor loadings for each county were plotted using Quantum GIS (Quantum GIS Development Team, 2013). All statistical analyses were performed using the R language (R Core Team, 2013).

Results

The VSS and MAP criteria both suggested a two component solution for the factor analysis of KIN data. Two factors explained 58% of the total variance and had a correlation of 0.15 between them. Factor loadings with an absolute value greater than 0.4 are presented in Table 2.1. The St. John location was the only one to show split loadings. This is a site managed by the Hays program and is farther southeast of than any of their other sites. The remaining sites split into factors cleanly according to if they were managed by the Hays or Manhattan program. For this reason, factors were labeled as Hays and Manhattan. The cluster analysis dendrogram is presented in Figure 2.1. Its first split divides all sites according the program that managed them, except for the St. John site which clusters with Manhattan program sites. The locations of these site are presented in Figure 2.2 with sites colored according to the first split in the dendrogram.

Table 2.1: Factor loadings greater than 0.4 for testing locations in KIN data.

Manhattan Factor		Hays Factor	
Manhattan	0.78	Hays	0.78
Hutchinson	0.71	Colby	0.67
Belleville	0.77	Garden City	0.75
Everest	0.78	Ness	0.79
Gypsum	0.71	Osborne	0.68
Hesston	0.70	Graham	0.77
Caldwell	0.61	Ford	0.62
Barber	0.75	St. John	0.57
St. John	0.46		

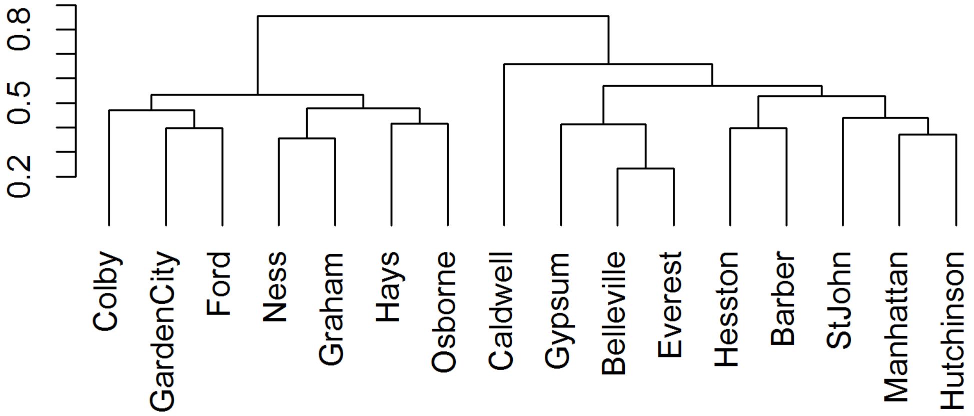

Figure 2.1: Dendrogram for hierarchical clustering of locations in the KIN.

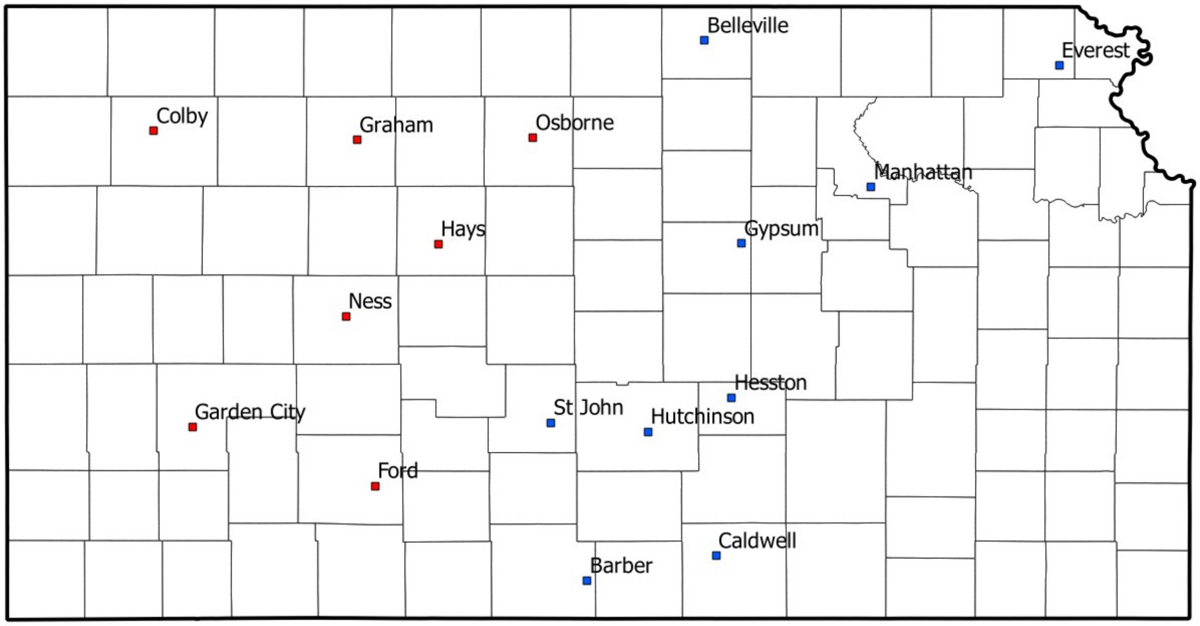

Figure 2.2: Analyzed KIN locations. Locations with blue symbols belong to the "Manhattan" cluster and red symbols belong to the "Hays" cluster.

Five and two factor solutions for average county yield data were suggested by MAP and VSS, respectively. The elbow method suggests a two factor solution which explains 76% of the total variance and has a 0.16 correlation between factors. Factor loadings for this solution indicate a vertical split that divides Kansas roughly in half (results not shown). A three factor solution explaining 81% of the total variance was chosen as the 'best' representation of the data and is presented herein. This solution was chosen since it explained a higher proportion of total variance, better accomplished the goal of dividing the state into smaller regions, and was readily interpretable when examining factor loadings. Solutions with more factors were deemed less interpretable. The three factors were named central, southeastern, and western based on geographic distribution of their loadings. The central factor correlated to the western and southeastern factors at 0.28 and 0.53, respectively. Correlation between the western and southeastern factors was -0.13. Factor loadings are presented in Figures 2.3, 2.4, and 2.5. Five clusters of counties were identified with the 0.9 AU p-value cutoff (Figure 2.6). Using a 0.88 AU p-value cutoff would reduce the number of clusters to two. The two western clusters would merge into one and the three eastern clusters would form the other cluster. The PAM clustering analysis suggested two clusters using the ASW criterion.

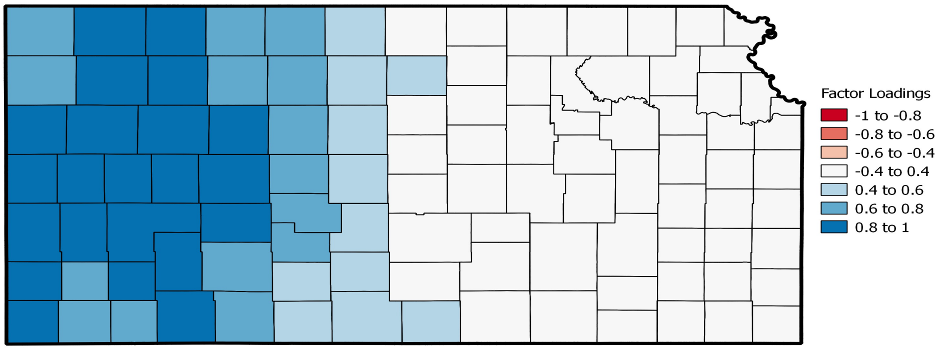

Figure 2.3: Factor loadings for western factor.

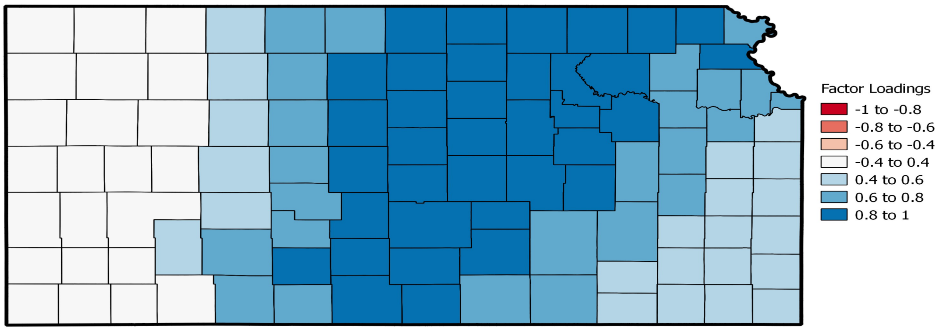

Figure 2.4: Factor loadings for central factor.

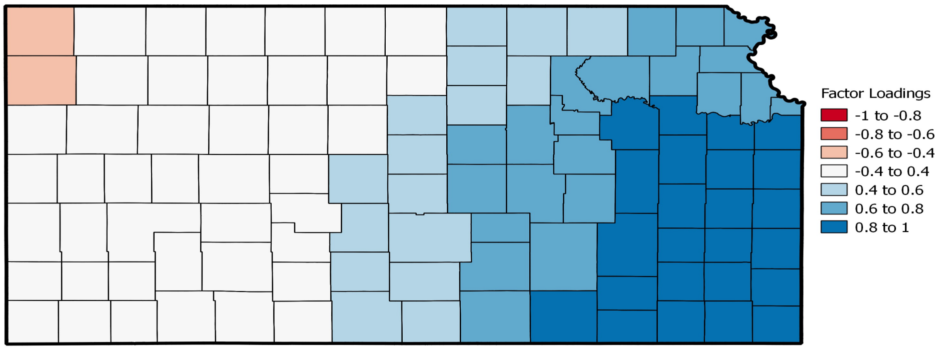

Figure 2.5: Factor loadings for southeastern factor.

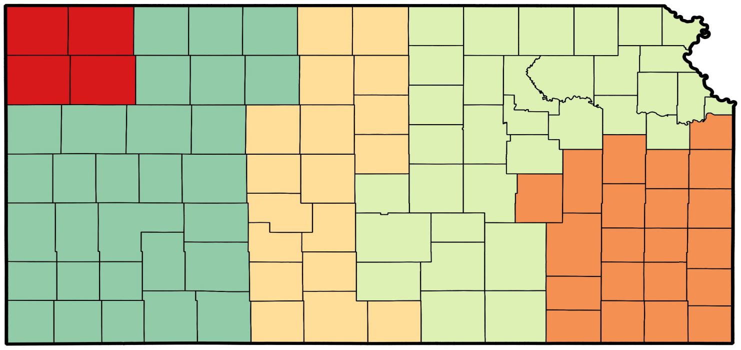

Figure 2.6: Clusters identified using bootstrapping and an AU p-value cutoff of 0.9.

Discussion

The KIN sites were well explained by splitting them into an eastern and western group based on factor loadings and the prominent split in the cluster dendrogram. Except for the St. John site, these sites split strictly according to whether they were managed by the Hays or Manhattan projects. The Manhattan and Hays projects target their breeding efforts for the eastern and western halves of states, respectively. Thus, the east/west division was expected due to this and due to CIMMYT's mega-environments. This split may be confirmation that these projects have picked their sites well. However, differences in data between the projects could also have influenced this division. Most concerning is that the difference in which lines are evaluated by each project. This difference could be making sites from the same program appear more similar than they actually are. The actual division between sites may not be as clear cut as suggested by data if this is the case. There was no strong evidence for a division into smaller groups than these two. Thus, no additional insight was discovered about how to handle GxE in Kansas, beyond the current Hays/Manhattan split. Limited power due to short comings in this data set may be at fault, or smaller regions may not be as meaningful as expected.

Average county yield data gave a more nuanced view of production areas in Kansas, but still supported a two production area split. For example, the factor analysis identified three factors with a fair amount of correlation between them. When considering these correlations and split loadings for many counties, the factor analysis shows strong evidence of two main production areas. Some degree of finer subdivision may be inferred. Two correlated factors in the eastern portion of the state indicates there is a subdivision between the southeastern corner and the rest of the eastern part of the state. This may be explained by average annual precipitation. The southeastern corner is the wettest part of Kansas (U.S. Geological Survey, 2005). The rest of the state follows an east-west gradient for precipitation that gets dryer in the west. Counties in the center of the state had split loadings on eastern and western factors. This may suggest a gradual transition from one region to another mirroring the precipitation gradient. Clusters identified using the bootstrapping method appear to orient according to the precipitation gradient, giving further evidence of a gradual transition.

It should be noted that the AU p-value used to identify clusters was originally developed for phylogenetic analyses where the number of observations exceeds the number of individuals

(Shimodaira, 2002). Since this condition isn't met (i.e. individuals = counties = 105, observations = years = 25), the AU p-values should be viewed as inexact. A relatively lenient p-value cut-off of 0.9 was also used, so results from this method of clustering are lacking a high degree rigor. The ASW criterion, presumed to be more rigorous, identified just two clusters. These two methods tackle the problem of identifying clusters in very different manners, so a difference in their results wasn't unexpected. The bootstrapping method searches for clusters by identify counties that can reliably be grouped together. The ASW criterion seeks to identify an optimal number of clusters which maximizes differences between clusters while minimizing differences within a cluster. Taken together, results from these two methods can be interpreted in the same manner as results from the factor analysis. That is, there are prominent western and eastern production areas with evidence of a gradual transition between them.

Dividing Kansas into two production areas is supported by results from all analyses. Where to make that division is less clear, as is whether any further subdivisions should be made. Weather and disease patterns are different each year. How representative a location is relative to a production area probably changes according to these yearly fluctuations. Dynamic borders for production areas that change according to yearly conditions are probably more representative of true patterns. Under these conditions, drawing firm borders between production areas and partitioning into production areas with minor differences would be difficult. This may be the source of ambiguity in the results from this experiment. Unfortunately, available data isn't sufficient to adequately examining these yearly fluctuations. Data from a multi-location yield trial with several years of data consisting of consistent locations and varieties tested across years is needed to thoroughly examine these yearly fluctuations and create a dynamic model. Until such a data set becomes available, dividing Kansas into two production areas, east and west, is best. Care should be taken when interpreting data in the boundary between production areas to ensure that they are included in the most representative production area. Care needs to also be taken when considering the southeastern corner of Kansas, because the data might not be representative of the rest of the eastern production area.

Chapter 3 - GSwGBS: an R package for Genomic Selection with Genotyping-by-Sequencing Markers

'GSwGBS' is an add-on package for the R programming language designed for performing genomic selection (GS) using markers obtaining via genotyping-by-sequencing (GBS). Specifically, it is for GBS markers obtained using the two restriction enzyme, *de novo* technique described by Poland *et al.* (2012a). There are three functions in the package to facilitate this task: GBSPipeline, hap2marker, and GS.model. 'GSwGBS' is freely available on GitHub (https://github.com/gaynorr/GSwGBS) and listed under the GNU General Public License.

GBSPipeline

The GBSPipeline function implements bioinformatic steps needed to convert raw sequence data in fastq or qseq format to single nucleotide polymorphism (SNP) markers. Most code needed to perform these steps is written in Java, so the 'rJava' package is used (Urbanek, 2013). The 'rJava' package is used to initialize the Java Virtual Machine (JVM) and create an instance of a Java class named GBSPipeline which is included in the 'GSwGBS' package. The GBSPipeline Java class depends on the TagsToSNPsNoAnchor Java class also included in the 'GSwGBS' package and several classes and libraries found within the software program TASSEL 4.0 (Bradbury et al., 2007). To access classes and libraries within TASSEL, the user must have a copy of TASSEL version 4.0 standalone installed on their computer. This software is freely available online (http://tassel.bitbucket.org/TasselArchived.html). A path to the user's copy of TASSEL is supplied as an argument to the GBSPipeline function so it can be added to the JVM's classpath. This makes classes and libraries within TASSEL available to the GBSPipeline Java class.

In addition to supplying a path to a copy of TASSEL, the user also supplies GBSPipeline with paths to their sequence data and a key file. These data are used to perform the GBS bioinformatic pipeline described by Poland *et al.* (2012a). Resulting output is written to multiple files in the user's work directory. Descriptions of these files can be found in TASSEL's help documentation for its GBS pipeline. The SNP markers are contained within three files with .hap extensions. These files are assigned numbers based on the number allowable mismatches used during sequence alignment (i.e., 1, 2, and 3).

hap2marker

The hap2marker function converts the three .hap files generated by GBSPipeline to a matrix of numerically coded SNPs for use in subsequent GS model construction. This is accomplished by sequentially reading in each .hap file and filtering markers for the amount of missing data based on a user supplied parameter. The SNPs from each file are merged together and any duplicated SNP markers are removed. An optional step can remove SNP data from any entries containing the phrase “BLANK” in their names. Such entries are used to denote a blank well in DNA isolation plates and is used to check that plates were properly processed.

The SNP allele scores, which are coded in letter format, are then converted to a numeric format. Major alleles are given a value of 1, heterozygotes become 0, and minor alleles become -1. When allele frequencies are equal, the first listed allele becomes 1 and the second allele becomes -1. Any SNPs with equivalent numeric scores are reduced to a single representative SNP marker. SNPs with equal allele frequencies are tested both as scored and with a sign change applied to their scoring to determine SNPs with equivalent scores. Finally, missing data can be imputed before returning the numerically scored SNPs. Methods for imputing missing data include using the mean, median, EM algorithm, or a random forest.

GS.model

The GS.model function is a wrapper for obtaining GS predictions using statistical models implemented in other R packages. The function was designed to minimize the amount of user generated coded needed to run these models, create a consistent method for calling each model, and to allow for fast computation. Genetic data is intended to come from numerically coded markers produced by hap2marker, but any markers similarly coded can be used. Up to five different GS models are used for predictions. Where possible, the 'foreach' package is used for parallel computing to reduce run time (Revolution Analytics and Watson, 2014).The 'rrBLUP' package is used to generate predictions based on estimated marker effects using a ridge regression best linear unbiased prediction approach or estimated line effects using a Gaussian kernel (Endelman, 2011). Random forest regression is implemented using the 'randomForest' package (Liaw and Wiener, 2002). A partial least squares regression model is implemented using the 'pls' package (Mevik et al., 2013). The number of components retained in this model is determined using 10-fold cross-validation (CV) on the training population to minimize the bias-corrected CV estimate. The fifth model is an elastic net model produced using the 'glmnet' package (Friedman et al., 2010). The elastic net mixing parameter and lambda are both set using a grid selection, 10-fold CV approach. A sequence of mixing parameters ranging from 0 (equivalent to the ridge regression penalty) to 1 (lasso penalty) is examined with a sequence of lambdas generated by the glmnet function to identify a pair of values which produce the lowest CV error. Predictions from all chosen methods are returned in a data frame with the average of all selected methods if more than one method was chosen.

Chapter 4 - Genomic Selection for Kansas Wheat

Abstract

Advances in inexpensive, high-throughput genotyping have lead to a re-envisioning of marker-assisted selection (MAS) into what is known as genomic selection (GS). This technique uses all available genome-wide markers as predictors in a statistical model intended to improved predictions, relative to classical MAS, for complex traits such as yield in wheat. Current values for the accuracy of GS have been determined using simulation and/or conditions not representative of those in a breeding project. To assess GS accuracy under more representative conditions, GS models were developed using germplasm and historic yield data from the Kansas State University (KSU) wheat breeding project. These models were used to predict grain yield in preliminary yield trial nurseries in 2011-2014. A correlation of 0.22 between predicted yield and observed multi-location means was obtained by the best performing model. This is lower than many previously reported values. Potential factors resulting in this lower value were examined. Continuation of this study for additional years is recommended to improve the scope of inference.

Introduction

Grain yield in wheat is a complex trait controlled by many genes and highly influenced by genotype-by-environment interactions. Obtaining an accurate phenotype for this trait requires using multi-location yield trials. These trials are expensive and use a large amount of seed for planting. This limits their use to screenings of later generation breeding material. To account for these limitations, it is typical for a breeding program to screen earlier generation lines at a small number of locations using trials with little or no replication. The best performing lines in these screenings are advanced into subsequent trials which increase the number of locations and/or replications and the process repeated. Lines advancing to the end of this process should represent those with the highest yield and be well characterized for this trait. Lower yielding lines are ideally eliminated early in the process, so resources aren't needlessly spent on them. A well functioning marker-assisted selection (MAS) protocol for yield could accelerate this process of phenotyping and/or reduce the investment in screening poor performing lines that are ultimately discarded. Genomic selection (GS), a relatively new technique for MAS, may make this possible.

Traditional MAS is a multiple step procedure that requires initial research to identify and confirm quantitative trait loci (QTLs). Molecular markers closely linked to important QTLs are used to screen lines for favorable alleles (reviewed by Collard et al., 2005). The QTL identification step is prone to biased estimates of QTL effects and is unable to detect QTLs with small effects (Jannink et al., 2010). This results in poor line performance predictions and partially explains why large numbers of QTLs for agronomically important traits have been reported, but few reports exist of these QTLs being successfully utilized in MAS breeding programs (Bernardo, 2008).

The GS procedure was proposed as an alternative to traditional MAS which can improve prediction of complex quantitative traits (Meuwissen et al., 2001). The GS procedure is accomplished without the need to identify QTLs. A training population with known genotypic and phenotypic values is used to develop a statistical model for predicting phenotypic performance of lines with known genotypic values and unknown phenotypic values. The genotypic data comes from a large set of genome-wide molecular markers. Phenotypic values are calculated using average line performance over a target set of environments. GS uses the full set of markers to generate line performance predictions instead of just selected markers, as used in traditional MAS. Simulation studies (Meuwissen et al., 2001; Bernardo and Yu, 2007) and empirical results (Heffner et al., 2010; Poland et al., 2012b) indicate GS outperforms traditional MAS for complex traits like yield.

Determining if GS for wheat yield would be effective in a breeding program requires a cost-benefit analysis. An important component of this cost-benefit analysis is the accuracy at which GS predicts line performance. The actual performance of a line isn't known and can only be estimated using observed performance in yield trials. An estimate for GS accuracy can be made using the correlation between predicted performance and observed performance. Using this method, correlations between 0.17 and 0.65 have been observed in studies examining GS for wheat yield (Burgueño et al., 2012; Campos et al., 2009; Crossa et al., 2010; Heffner et al., 2011; Poland et al., 2012b). One or more factors could be contributing to this wide range of correlations: differences in genotypic and phenotypic variances between studies, differences in statistical methods used for GS, and differences in phenotyping and genotyping methods. These factors make it difficult to determine what level of accuracy can be obtained under conditions present in a breeding program.

Many previous studies examining GS accuracy for wheat yield have relied on cross-validation (CV) approaches that randomly assign lines to training and validation populations. Random sampling generates populations that are representative of the larger population from which they are drawn. This technique fails to replicate conditions expected in GS breeding programs where older germplasm with phenotypic data would be used to build a model for new germplasm without phenotypic data. Breeders are constantly working to improve their germplasm by making selections. They also employ germplasm exchange to introduce new material in their breeding program. Both of these activities change the genetic structure of a breeding program's germplasm at a population level. The new germplasm, although derived from older germplasm, represents a different population from the older germplasm. These changes may introduction of new alleles or change linkage-disequilibrium (LD) patterns between the populations. It is LD between markers and genes controlling yield that makes GS for yield possible (Heffner et al., 2009). Changing these patterns of LD or introducing new alleles is expected to reduce GS accuracy, so previous studies may report overly optimistic correlations for GS accuracy. There may be other meaningful difference between these studies and how a GS breeding program would be implemented that could influence accuracy. Thus, reproducing conditions in a GS breeding program as closely as possible is needed to obtain a meaningful estimate for accuracy.

The goal of this study is to measure accuracy of GS for wheat yield. To accomplish this, breeding material and phenotypic data generated by the Kansas State University (KSU) wheat breeding program, in the course of performing its typical breeding operations, will be examined. An existing conventional breeding program was chosen to simulate conditions expected in a GS breeding program without making a prohibitively large investment in time and money to actually create one. A GS protocol will be optimized for this program to obtain a high level of prediction accuracy in early generation yield trials. Predicted performance in these trials will be compared to observed performance to assess accuracy of GS for wheat yield. Factors potentially limiting accuracy of GS, such as training population size and genetic differences between training and test populations, will be examined to identify which may have the greatest impact.

Materials and Methods

KSU Wheat Breeding

Wheat breeding at KSU is performed by two separate programs. One program is based at Manhattan, Kansas and focuses on breeding hard red winter wheat for central and eastern Kansas. The other program is based in Hays, Kansas and focuses on breeding hard red and hard white winter wheat for western Kansas. The Manhattan program is the focus of this study. It produces F_5 derived lines using a selected bulk breeding strategy. New lines are derived from a single plant and seeded at a single location in small plots (0.75 x 2.25 m) using an unreplicated, augmented design. This nursery is referred to as the individual plant short rows (IPSRs). Grain yield is measured on these plots, but is not a major basis for selection. This is due to low accuracy attributed to lack of replication and differences in seeding rates due to variability in seed harvested from the F_5 plants used to seed these plots. Lines selected in the IPSRs are advanced to the preliminary yield nursery (PYN). PYNs are seeded using an augmented design in unreplicated plots (1.5 x 4.5 m) at multiple locations. This is the first nursery where yield is an important selection criteria. Selected lines are advanced to the advanced yield nursery (AYN). AYNs include lines from both the Hays and Manhattan projects and are evaluated at multiple locations by both projects. They are seeded in two replicate plots (1.5 x 4.5 m) using using an alpha lattice design for locations managed by the Manhattan program and seeded in three replicate plots (1.5 x 4.5 m) using a randomized complete block design (RCBD) at locations managed by the Hays project. Lines selected from AYNs are advanced to the Kansas intrastate nursery (KIN). The KIN is evaluated at the largest amount of locations and is seeded in four replicate plots (1.5 x 4.5 m) in a RCBD at sites managed by the Hays program and seeded in three replicate plots (1.5 x 4.5 m) in an alpha lattice design for Manhattan program sites. Lines are evaluated for multiple years in the KIN and regional performance nurseries before a decision on variety release is made. All of these trials are managed according to typical practices where they are planted. These trials are not sprayed for disease, so disease resistance has a large influence on observed yields.

Genotypic Data

A total of 4,966 lines predominately from KSU's Manhattan wheat breeding program were selected for genotyping. The majority of these lines were new additions to the IPSR nurseries in 2012 (harvest year; 492), 2013 (1283), and 2014 (1,920). The remaining genotyped lines were taken from the 2012 PYNs (331), 2011 PYNs (106), and an assortment of other lines available in storage. The bulk of these other lines consisted of older germplasm and varieties from other programs that have been important parents in the Manhattan program. Most of the 2012 IPSR lines genotyped entered the 2013 PYNs (382). The 2014 PYNs contained contained 330 of the 2013 IPSRs.

DNA extraction was performed in four sets at separate times. The first set included the 2011 PYNs and older lines. The second, third, and forth sets were the 2012 IPSRs and PYNs, the 2013 IPSRs, and 2014 IPSRs, respectively. Four seeds from each line were grown for two weeks prior to DNA extraction. DNA extraction was performed on bulked leaf tissue from each line using the BioSprint 96 DNA Plant Kit (Qiagen) with the BioSprint 96 Workstation (Qiagen). Genotyping-by-sequencing was used to identify single nucleotide polymorphisms (SNPs) in the extracted DNA following the methods of Poland *et al.* (2012a). The 'GSwGBS' package presented in chapter 3 was used to run the genotyping-by-sequencing pipeline for calling SNPs and to convert them to numeric coding (i.e. 1 for lines homozygous for the most common allele, 0 for heterozygotes, and -1 for lines homozygous for the less frequent allele). Missing marker data was imputed using the “EM” imputation method.

Phenotypic Data

Grain yield data from 21 locations over seven years (2008-2014) were evaluated. These data consisted of 81 sites (year by location combinations), because not all locations were present each year. Three yield trial nurseries (PYNs, AYNs, and KINs) were examined. The number of entries and locations in each nursery varied depending on the year. The KINs were evaluated at between eight and thirteen locations and contained 35 to 54 entries. The AYNs contained between 56 and 130 entries evaluated at two to nine locations. The PYNs contained between 211 and 481 entries evaluated at three to seven locations. Depending on the year, the AYNs and PYNs may have been split into as many as three smaller nurseries and evaluated separately to account for field variation.

Each year by location by nursery combination is referred to as a trial. Each trial was analyzed separately using the most appropriate statistical method for its experimental design to determine line means for that trial. Data from any trial with a coefficient of variation greater than 13% were discarded. A total of 232 trials were retained resulting in 15,836 trial line means.

A core set of locations (Manhattan, Hutchinson, McPherson, Barber County, Sumner County, Gypsum, Belleville, and Ellsworth) was chosen by the researcher, based on historical knowledge and results from chapter 2, as the locations most representative of the target environment for the Manhattan breeding program. The core set of locations reduced the number of locations to 8, sites to 41, trials to 173, and trial line means to 12,866.

Overall line means were calculated for use in GS training populations. This was accomplished using only data which would have been available when a training population was used. For example, the training population for the 2012 PYNs only uses data which would have been available before the 2012 PYNs were harvested. These trial line means were then used to calculate site line means. This was accomplished by fitting a mixed linear model containing a random effect for trial and a random lines effect, each site was analyzed separately. All site means were combined into a single mixed model containing a random site effect and a random line effect. This final mixed model was used to obtain the overall line means. Repeatability for training populations was measured using trial line means and a mixed model with random effects for site and line. Repeatability was calculated by taking variance due to lines divided by the sum of variance due to lines and variance due to error. The 'lme4' package in R was used to fit these mixed linear models (Bates et al., 2013).

GS Models

Five statistical methods for GS were considered. The 'GSwGBS' package was used to build models for: ridge regression of marker effects (RR), Gaussian kernel (GAUSS), partial least squares regression (PLSR), elastic net (ELNET), random forest with 1000 trees (RF), and the average of all methods (AVE).

Allowable missing marker data was optimized by testing maximum allowable missingness from the original SNP calling procedure (80%) down to 30% in increments of 10. Random sampling without replacement was used to choose a testing population of 100 lines out of all lines with genotypic and phenotypic data, from the core locations, available in the year 2013. The remaining lines constituted the training population. A GS model was built using RR and the correlation of predicted values to observed line means was determined. This procedure was repeated 100 times and average correlation was used to choose the optimal value for maximum allowable marker missingness. The RR method was chosen, because previous studies suggest it performs reasonably well and it is considerably computationally faster than other methods. The observed optimum was then used in all subsequent analyses.

Evaluation of GS Accuracy

Grain yield for the 2012, 2013, and 2014 PYNs were predicted by constructing GS models using all methods. The training populations for each PYN consisted of all older lines and older phenotypic data (i.e., the training population for the 2013 PYNs contained only data from 2008-2012). Line means used in these training populations were constructed separately from all available locations and from only core locations. The precision of each GS model was determined by calculating correlations between predicted yields and observed multi-location means. Correlations between predictions from the best performing GS model and observed yields from individual locations were used to compare predictiveness of GS to that of a single location yield trial.

Single year GS models using just data from PYNs were constructed using leave-one-out CV. Only the RR and GAUSS methods were examined in these models. Prediction accuracies measured using these models were used to confirm methods used in this study could achieve accuracies measured in previous empirical GS studies. They also served as a baseline for comparing GS using historical data in a breeding program to performing GS using a single year's data.

To measure the impact of training population size on 2013 PYN predictions, the same random sampling and CV procedures used for optimization of allowable marker missingness was used to evaluate different sized training populations. These training populations ranged in size from 500 to 850 in increments of 50. One hundred samples were made for each population size.

The 2013 IPSR nursery was used to compare predictions made by GS to those made by a breeder. The KSU breeder chose 50 lines out of the nursery he believed to represent the highest yielding lines. The GAUSS method was used to select 50 lines predicted to have the highest yields based on GS. A permutation test using 10,000 permutations was used to determine if there was a statistically significant overlap between lines selected by the breeder and GS. These two sets of selected lines were then combined and grown as one section in the 2014 PYN. A linear model was constructed using trial means from this PYN section to determine if there was a significant difference in yield between selection methods. The linear model contained a fixed effect for selection method and a random effect for location.

Results

A total of 56,000 genetic markers were identified at the 80% level of allowable marker missingness. The CV tests for optimal allowable missingness determined a value of 70% as optimal. This reduced the number of markers to 38,432. A correlation of predicted yield to observed yield of 0.33 was obtained during optimization using this cutoff. The first three principal component (PC) scores for the imputed marker scores were taken to access the genetic structure of all genotyped lines. These PCs explained 8 percent of the total marker variation. Plots of these PC scores for each PYN and their training populations were used assess how similar a PYN was to its training population. The distribution of 2012 PYN lines was very different from its training population (Figure 4.1). The other PYNs showed distributions more similar to their training populations (Figures 4.2 and 4.3).

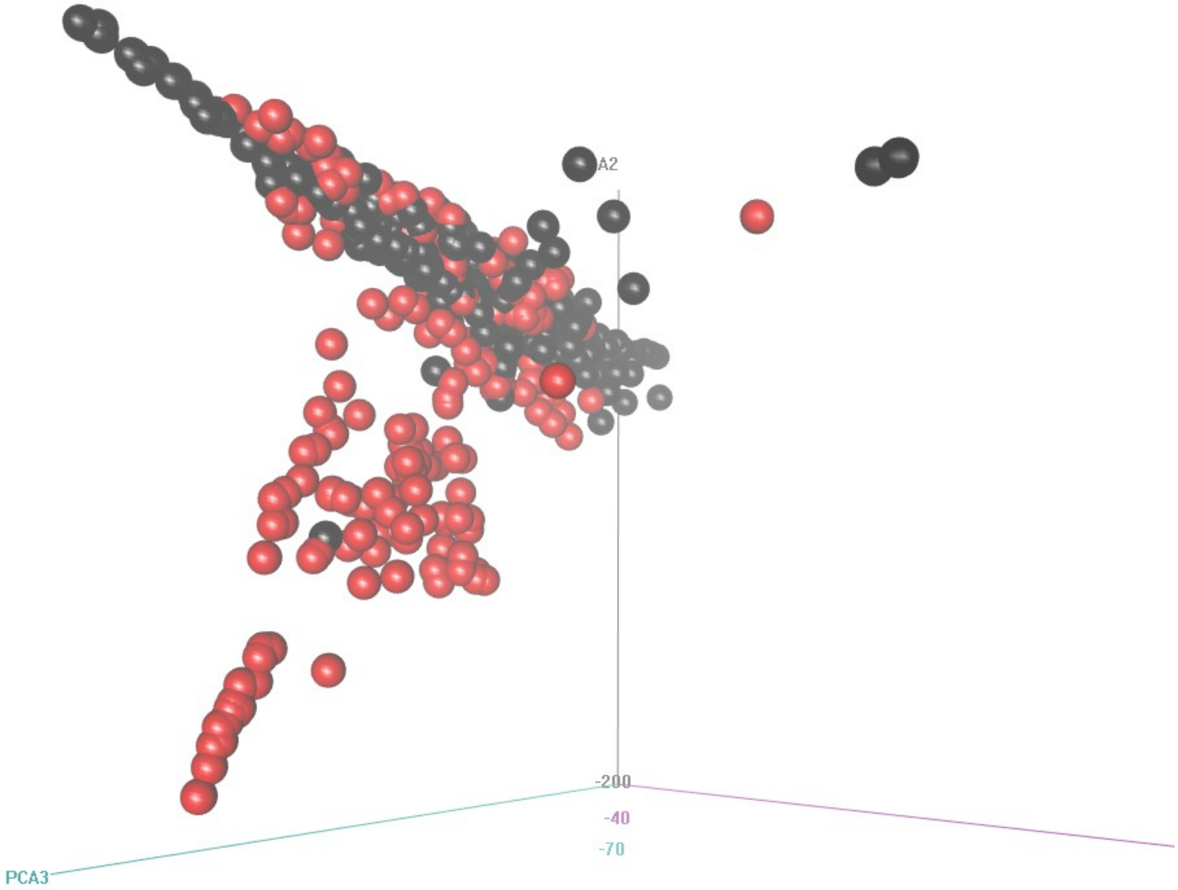

Figure 4.1: Principal component scores for imputed marker scores of 2012 PYN lines (red) and lines in its training population (black).

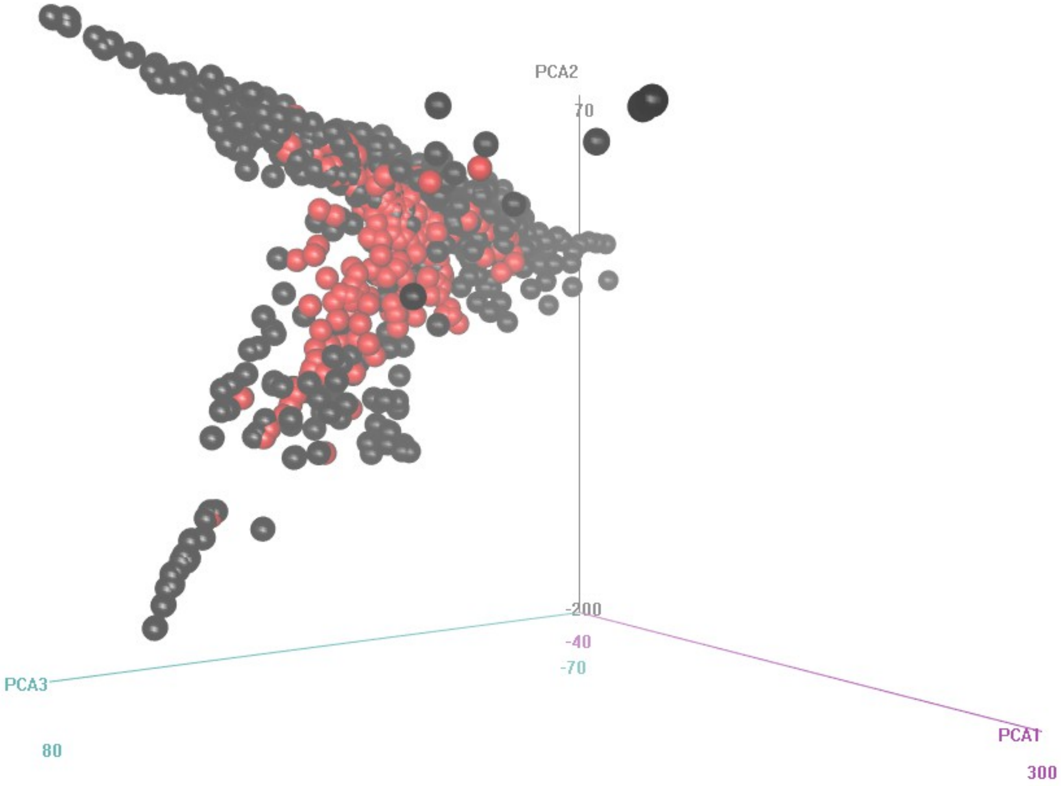

Figure 4.2: Principal component scores for imputed marker scores of 2013 PYN lines (red) and lines in its training population (black).

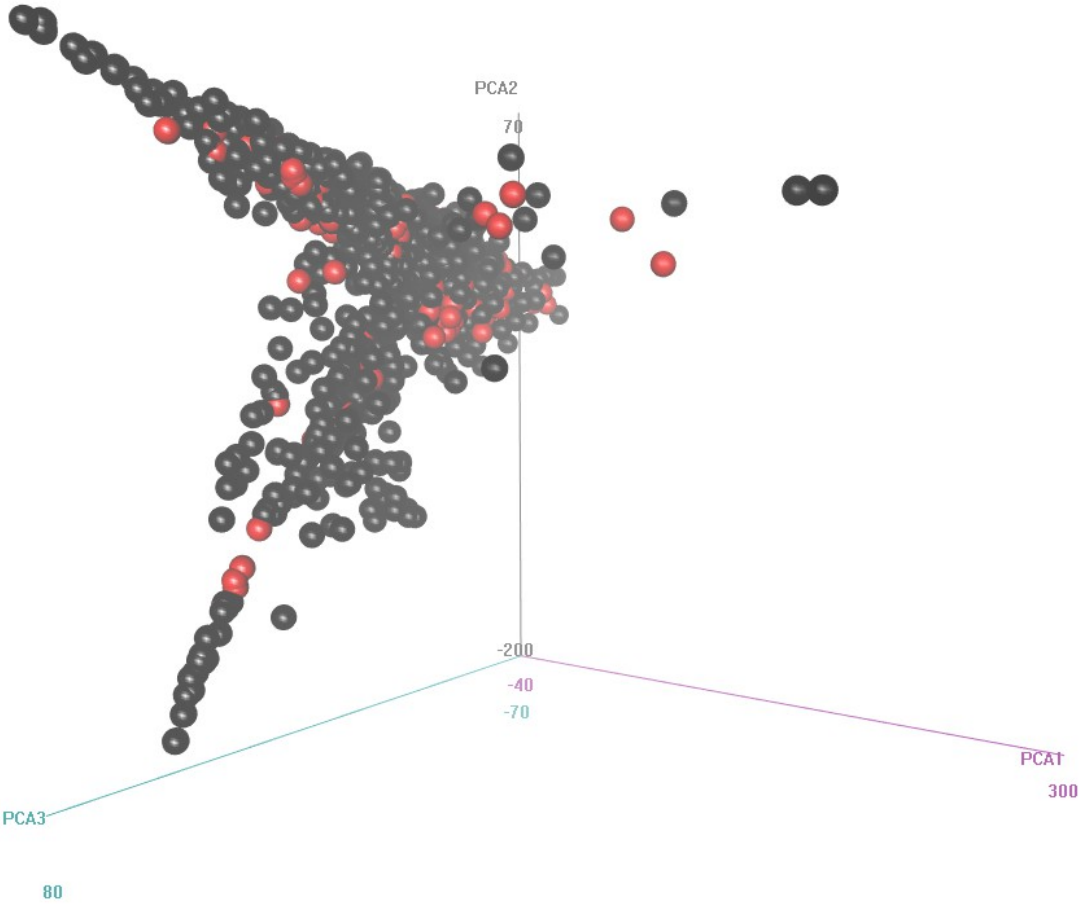

Figure 4.3: Principal component scores for imputed marker scores of 2014 PYN lines (red) and lines in its training population (black).

Histograms of line means for grain yield across all years generated using all locations (Figure 4.4) and just core locations (not shown) both appear to follow a Gaussian distribution. Most lines had means within 0.2 metric tons per hectare (MT/ha) of the mean for all lines. Year to year variation for line means was examined by calculating correlations between years with more than 50 common lines (Table 4.1). Only adjacent years met this criteria. Except for 2012-2013, all between year correlations were non-significant (P>0.05). Lines common between years are limited to check varieties and lines selected, in part, for high yield in the earlier year. Thus, these correlations are only considering a portion of the variability present in each year. Figure 4.5 shows an example of this relationship for the 2012-2013 comparison.

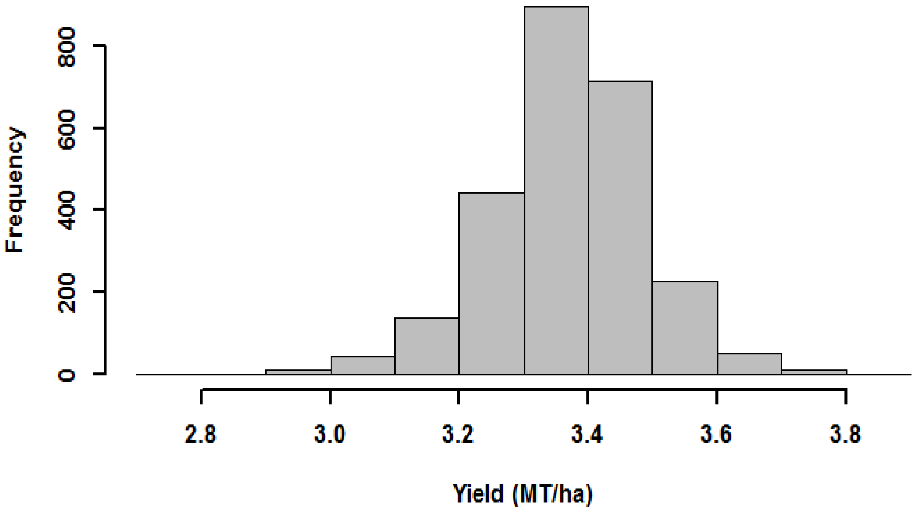

Figure 4.4: Histogram of all overall line means using all locations and years 2008-2014.

Table 4.1: Correlations between single year line means for years with more than 50 common lines.

Years Compared	Correlation	P-value	Shared Lines
2009-2010	0.06	0.63	61
2010-2011	0.21	0.10	62
2011-2012	0.14	0.28	58
2012-2013	0.30	0.01	76
2013-2014	0.15	0.23	64

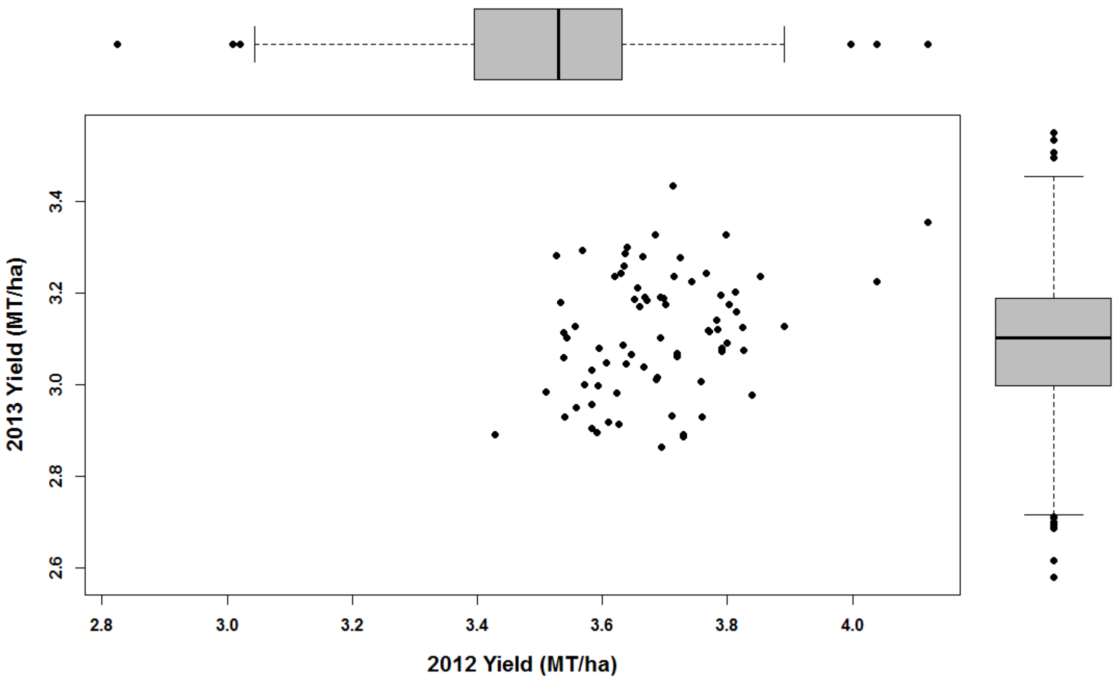

Figure 4.5: Line means for 2012 and 2013. Means for shared lines are in the scatterplot and means for all lines are displayed in boxplots.

The size of training populations used for each predicted PYN (2012-2014) increased as the year advanced due to previous PYNs being included in the new training population (Table 4.2). Repeatabilities for these training populations showed little variability (0.256-0.294). Variability in repeatabilities for the PYNs was greater (0.263-0.471). Line means for each predicted PYN decreased in successive years and the spread of line means was greatest in 2014 (Figure 4.6). The majority of data in each training population came from core locations and means calculated using all locations and core locations were highly correlated (Table 4.2).

Table 4.2: Summary data for predicted PYNs and their training populations.

Predicted PYN			Training Pop (all locs)			Training Pop (core locs)			Corr Training
Year	Lines	H	Lines	Trial Means	H	Lines	Trial Means	H	Pops
2012	331	0.317	566	6847	0.294	563	5138	0.299	0.94
2013	382	0.263	899	9423	0.278	895	7544	0.292	0.96
2014	330	0.471	1281	12993	0.256	1277	10745	0.263	0.96

pop = population, loc = location, H= repeatability, corr = correlation

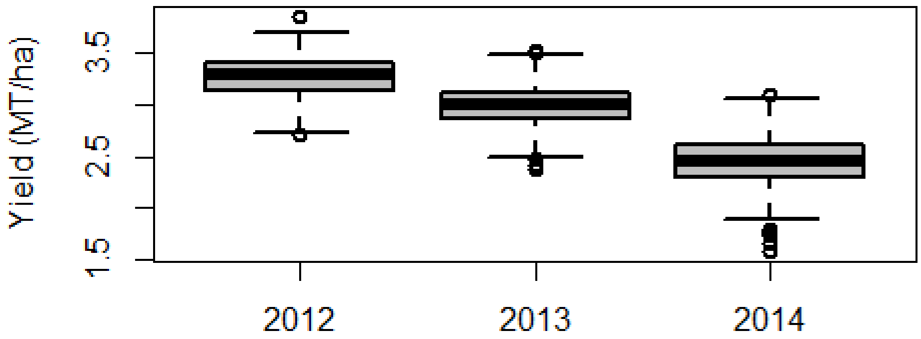

Figure 4.6: Multi-location line means for predicted PYNs.

Correlations for GS predictions of each PYN to their observed multi-location means are presented in Table 4.3. No correlation for the 2012 PYN was significant (P<0.05), and almost all correlations for the 2013 and 2014 PYNs were significant. The GAUSS method produced the highest correlation in 2013 using the core locations (r=0.21) and 2014 using all locations (r=0.22). Training populations using just core locations generated higher correlations in 2012 and 2013, but lower correlations in 2014. Predictions using the GAUSS method in 2013 with core locations (Table 4.4) and 2014 with all locations (Table 4.5) were not as predictive of multi-location means as individual locations in each year's PYN. Predictions produced using leave-one-out CV in the PYNs (Table 4.6) produced more accurate predictions of line performance than those made using historical data (Table 4.3).

Table 4.3: Correlations of predicted PYN yields to observed multi-location yields.

	2012 PYN		2013 PYN		2014 PYN	
Locations[1]	All	Core	All	Core	All	Core
RRBLUP	0.07	0.07	0.15^{**}	0.18^{***}	0.22^{***}	0.19^{***}
GAUSS	0.08	0.09	0.15^{**}	0.21^{***}	0.22^{***}	0.20^{***}
PLSR	0.05	0.10	0.13	0.17^{***}	0.17^{**}	0.15^{**}
ELNET	0.07	0.07	0.15^{**}	0.18^{***}	0.22^{***}	0.19^{***}
RF	-0.08	-0.08	0.08	0.15^{**}	0.10	0.11^{*}
AVE	0.04	0.07	0.15^{**}	0.19^{***}	0.20^{***}	0.18^{**}

[1] Locations included in the training population

*, **, and *** represent P-values <0.05, <0.01, and <0.001, respectively

Table 4.4: Correlations for 2013 PYN site means, overall mean, and Gaussian kernel predicted means (using core locations). P<0.001 for all correlations unless noted.

	Belleville	Gypsum	Hutchinson	Manhattan	McPherson	Sumner	PYN Mean
Gypsum	0.31						
Hutchinson	0.18	0.35					
Manhattan	0.38	0.27	0.17				
McPherson	0.41	0.40	0.23	0.34			
Sumner	0.23	0.20	0.20	0.18	0.46		
PYN Mean[1]	0.45	0.48	0.35	0.37	0.55	0.38	
GS Prediction	0.12^{NS}	0.14^{**}	0.1^{NS}	0.11^{**}	0.22	0.11^{*}	0.21

[1] Site being compared is excluded from mean.

^{NS}, *, and ** indicate P>0.05, P<0.05, and P<0.01, respectively.

Table 4.5: Correlations for 2014 PYN site means, overall mean, and Gaussian kernel predicted means (using all locations). P<0.001 for all correlations unless noted.

	Belleville	Gypsum	Lane	Manhattan	McPherson	Sumner	PYN Mean
Gypsum	0.61						
Lane	0.63	0.64					
Manhattan	0.41	0.43	0.30				
McPherson	0.56	0.53	0.54	0.41			
Sumner	0.60	0.56	0.55	0.30	0.46		
PYN Mean[1]	0.73	0.69	0.67	0.42	0.62	0.62	
GS Prediction	0.15^{**}	0.25	0.30	0.04^{NS}	0.14^{*}	0.16^{**}	0.22

[1] Site being compared is excluded from mean.

^{NS}, *, and ** indicate P>0.05, P<0.05, and P<0.01, respectively.

Table 4.6: Prediction accuracies in PYNs using GS models created with a leave-one-out CV approach. P<0.001 for all correlations.

GS Model	2012	2012	2014
RR	0.42	0.38	0.57
GAUSS	0.44	0.38	0.55

Testing for the impact of training population size on predicted 2013 PYN yields showed a slight increasing in performance as training population size increased and decreasing variability in predictions (Figure 4.7). Performance improved from a correlation of 0.17 using 500 lines to 0.19 with 850 lines. These findings suggest that increasing training population size alone isn't accounting for increased GS prediction accuracies.

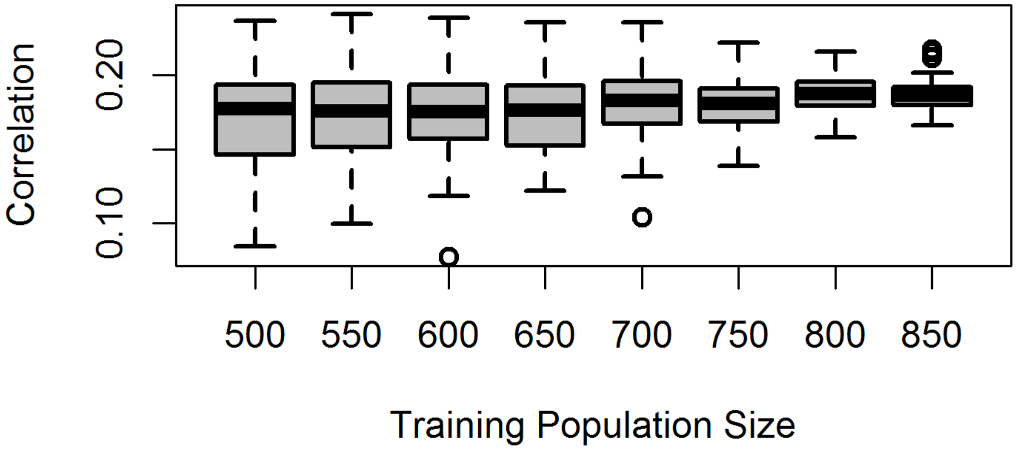

Figure 4.7: Box plots for effect of simulated increase in training population size on GS prediction accuracy.

A comparison between predictions from GS and those of KSU's wheat breeder used 1,282 IPSR lines from 2013. Both GS and the breeder chose 50 lines as their best guess for the highest yielding lines in the next year's PYN. Five lines were chosen by both GS and the breeder. The permutation based p-value for the probability of selecting five or more common lines was determined to be 0.038 using 10,000 permutations. No significant difference was observed for average yield for lines chosen using GS, the breeder, or both.

Discussion

The GS models were unable to predict 2012 PYN yields and obtained modest accuracy predicting 2013 and 2014 PYN yields. The inability to predict 2012 PYN is likely due to the lack representativeness of the training population at that time (Figure 4.1). Training populations using data from just core locations outperformed training populations using all available data in 2012 and 2013, but underperformed in 2014. The core locations were chosen with the intention of improving GS performance by minimizing adverse effects caused by genotype-by-environment interaction. Poorer performance of models using just these locations in 2014 could be due to exceptional weather experienced that year. The 2014 Kansas wheat crop experienced an unusually cold winter and widespread drought which resulted in the lowest statewide average yield since 1995 (“Kansas Wheat History,” 2014). Testing GS performance in subsequent years would be very useful for determining if 2014 was an anomaly due to this atypical weather.

Weather and disease pressures were atypical for all three PYN years examined. The 2012 Kansas wheat crop experienced more stripe rust (*Puccinia striiformis* f.sp. *tritici*) and barley yellow dwarf than typical, but other economically important foliar diseases had very little impact on the crop (Appel et al., 2014). Several record high temperatures were experienced during the growing season resulting in the 2012 Kansas wheat harvest being the earliest on record (“Kansas Wheat History,” 2014). The 2013 and 2014 wheat crops in Kansas were virtually unaffected by foliar diseases (Appel et al., 2014). They also experienced cool temperatures delaying both years' harvest and moderate to severe drought (“Kansas Wheat History,” 2014). Continuing this experiment for additional years would be useful for determining if the GS accuracy values reported in this study are relevant to more “typical” years in Kansas.

It should be noted that the true values of mean line yields are not known. This means the observed measurements of yield which were used for comparison have error associated with them. Thus, the best way to assess the observed GS performance is to consider it in relation to phenotypic selection. One way of examining this was assessed using data from individual PYN sites to predict average performance over all other sites (Tables 4.4 and 4.5). Each individual site was more predictive of multi-location yield than the best performing GS models. However, this isn't a fair comparison. The GS models account for average performance over several earlier years while the PYN site means are only considering performance in a single year. Thus, it is possible GS is actually providing a better prediction of performance in subsequent years. To examine this possibility, phenotypic correlation across years was assessed by measuring correlation between the same lines in different years (Figure 4.5). Unfortunately, there is a very high turnover of lines in breeding programs, so only adjacent years had enough shared lines to be useful for comparison. Five year to year comparisons ended up being made and only two of them showed correlations as high or higher than the best performing GS models. This would appear to indicate that GS is performing very well in relation to phenotypic correlation. However, this comparison also has its flaws. Lines tend to only be present multiple years in a breeding program if they are either being used as agronomic checks or they have been selected due to desirable traits. In both cases, these lines are expected to be relatively high yielding. Whereas, the correlations reported for GS use earlier generation breeding material which has experienced relatively less selection for yield. This results in less variability for yield in the year to year comparisons and thus correlations are expected to be lower than those obtained if the GS lines were grown in multiple years. Thus, it is difficult to determine precisely how well GS is performing in relationship to phenotypic selection.

The comparison made between the breeder and GS probably offers the best way of measuring GS performance. This comparison was only made in one year and that year didn't show any statistical difference in yield between lines chosen by the breeder or GS. This bodes well for GS. Unfortunately, the year in which this comparison was made was the year with unusually harsh weather. These harsh conditions severely stressed the crop, as seen in the much lower yields for the 2014 PYNs relative to other years (Figure 4.6). This environmental stress is probably also responsible for the higher repeatability observed in the 2014 PYNs (Table 4.2). Lack of statistical difference between the two methods of selection may just be a case of neither method being able to predict performance in this unusual year. A comparison of GS to the breeder in years more representative of 'typical' years in Kansas would provide a better comparison of their relative performances.

Several factors can limit performance of GS. These include: small training population size, poor genotyping, poor phenotyping, genetic differences between training and testing populations, and limitations of the GS procedures.

Training population size is the most straightforward issue to assess. Examination of training population size for the 2013 PYNs showed only a very small increase in precision going from 500 lines to 850 lines (Figure 4.7). Indicating that training population size is not likely to be a limiting factor. Extrapolating this trend suggests a drastic increase in training population size would be needed to achieve even a small improvement in prediction accuracy. Variance in prediction accuracy decreased as the training population increased. This may be due to both more consistent sampling of important genetic diversity decreasing occurrence of low correlations and less frequent sampling of “ideal” training populations, those which produce higher correlations than a training population using all lines. The existence of these “ideal” training populations suggests it may be possible to limit lines used in a training population to improve GS accuracy. How to choose these lines is an important avenue to explore in future research. The CV method used to examine population size is also partially reponisble for the decrease in variance. As sampled population size approaches the amount of lines available for sampling, there will be increasing correlation between samples due to a higher proportion of shared lines between those samples.

The quality of genotyping was not assessed in this study. Ideal genotyping would have several markers in LD with all important genes controlling yield. It is not possible to determine if this is occurring. If genes controlling yield are assumed to be spread over the whole genome, assessing how well the markers cover the genome could be done instead. Genetic markers for wheat obtained using genotyping-by-sequencing have shown a high degree of coverage in reference mapping populations, so the quality of genotyping is not expected to be a limitation (Saintenac et al., 2013).

The quality of phenotyping in this study is limited, because the data comes directly from an applied breeding program. The primary goal of these programs is to turn out new varieties. This is done by choosing methods on the basis cost efficiency. This often means only making rough measurements for traits when that is all that is needed to be effective. Many selection decisions in these programs are made using visual assessments based on personal experience and thus don't generate data that can be analyzed. The sheer volume of data these programs generate is bound to contain mistakes and these mistakes often go unnoticed. These limitations in data quality generated by breeding programs can be problematic for building GS models. However, GS needs to be able to deal with these limitations, because it is unlikely breeders will find it cost effective to make large changes in their programs to accommodate building GS models.

Genetic differences between training populations and testing populations can be a major concern and is probably the reason the 2012 PYNs were accurately predicted. The graphs of marker PCs show a very complex set of genetic relationships between lines in the breeding program (Figures 4.1, 4.2, and 4.3). This complex relationship develops as a result of inter-mating, selection, and introduction of new material. Introduction of new material is probably of the greatest concern. This can introduce new alleles not present in the training populations. When this is the case, predictions for lines containing these alleles are not expected to be very accurate. The difference observed between the 2012 PYNs and its training population are probably the result of such a new introduction. To account for these occurrences, it may be necessary to exclude some crosses from evaluation using GS to ensure that favorable new alleles aren't lost. Further research in this area is recommended. This could be accomplished by collecting additional years of data and tracking performance of GS in relation to the pedigree of lines being predicted.

Another key piece of evidence showing that genetic differences between training populations and testing populations may be a limiting factor comes from the PYN within year GS predictions using leave-one-out CV (Table 4.6). The values observed using this approach were much higher than the values obtaining using GS with historical data (Table 4.3). The CV uses just lines in the PYN for a training population, so they should be representative of the line being predicted. Year-to-year variation is also ignored using this approach, because training data only comes from a single year. The low observed correlations between years in Table 4.1 suggest that year-to-year variation has a very significant impact on yield in Kansas. The higher correlations for GS using leave-one-out CV are closer to those obtained in earlier GS experiments which used CV. The observed lower values for GS using historical data illustrates the need for measuring GS accuracies under conditions representative of those experienced in an actual breeding project.

There is also the possibility that the statistical methods just aren't up to the task. The low observed between year correlations (Table 4.1) and the relatively high values for leave-one-out CV (Table 4.6) suggest GS statistical methods can perform relatively well at predicting a phenotype using genotypic data. The bottle neck appears to be generating a representative phenotypic value. This will require careful consideration of genotype-by-environment interactions. This study used a very simple approach for dealing with genotype-by-environment interactions by looking at using a training population consisting of only data from a “core” set of locations. Predictions using this approach were improved in two out of the three years. However, more sophisticated approaches need to be considered. Such an approach could involve modeling environments with covariates. Improved modeling of genotype-by-environment interactions could greatly improve GS predictions.

This study used data from an actual breeding program. Whereas, previous studies have relied on simulation data or specially designed experiments to measure GS accuracy. Thus the lower observed accuracy for GS using historic data suggests many of the previously reported accuracies are overly optimistic, because they failed to accounted for all the intricacies of a breeding program. At present, these lower values probably represent a better estimate of GS accuracy for performing a cost-benefit analysis to determine the usefulness of GS in a breeding program. However, further research could increase the accuracy of GS. Improving the handling of genotype-by-environment interaction and the selection of lines to use for a training population appear to be the most promising areas for further research.

Bibliography

Appel, J., E. DeWolf, W. Bockus, and T. Todd. 2013. Preliminary 2013 Kansas Wheat Disease Loss Estimates. Available at http://agriculture.ks.gov/docs/default-source/pp-disease-reports-2012/2013-ks-wheat-disease-loss-estimates44D2D289EE71.pdf?sfvrsn=6 (verified 17 November 2013).

Appel, J., E. DeWolf, T. Todd, and W. Bockus. 2014. Preliminary 2014 Kansas Wheat Disease Loss Estimates. Available at http://agriculture.ks.gov/docs/default-source/PP-Disease-Reports-2014/2014-ks-wheat-disease-loss-estimates.pdf (verified 5 February 2015).

Bates, D., M. Maechler, B. Bolker, and S. Walker. 2013. lme4: Linear mixed-effects models using Eigen and S4.

Bernardo, R. 2008. Molecular Markers and Selection for Complex Traits in Plants: Learning from the Last 20 Years. Crop Sci. 48(5): 1649.

Bradbury, P.J., Z. Zhang, D.E. Kroon, T.M. Casstevens, Y. Ramdoss, and E.S. Buckler. 2007. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23(19): 2633–2635.

Braun, H.-J., and T. Payne. 2012. Mega-Environment Breeding. *In* Reynolds, M., Pask, A., Mullan, D. (eds.), Physiological Breeding I: Interdisciplinary Approaches to Improve Crop Adaptation. CIMMYT, Mexico, D.F.

Braun, H.-J., S. Rajaram, and M. van Ginkel. 1996. CIMMYT's approach to breeding for wide adaptation. Euphytica 92(1-2): 175–183.

Brennan, J.P., and G.M. Murray. 1998. Economic Importance of Wheat Diseases in Australia. NSW Agriculture.

Burgueño, J., G. de los Campos, K. Weigel, and J. Crossa. 2012. Genomic Prediction of Breeding Values when Modeling Genotype × Environment Interaction using Pedigree and Dense Molecular Markers. Crop Sci. 52(2): 707.

Campos, G. de los, H. Naya, D. Gianola, J. Crossa, A. Legarra, E. Manfredi, K. Weigel, and J.M. Cotes. 2009. Predicting Quantitative Traits With Regression Models for Dense Molecular Markers and Pedigree. Genetics 182(1): 375–385.

Collard, B.C.Y., M.Z.Z. Jahufer, J.B. Brouwer, and E.C.K. Pang. 2005. An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. Euphytica 142(1-2): 169–196.

Crossa, J., G. de los Campos, P. Pérez, D. Gianola, J. Burgueño, J.L. Araus, D. Makumbi, R.P. Singh, S. Dreisigacker, J. Yan, V. Arief, M. Banziger, and H.-J. Braun. 2010. Prediction of Genetic Values of Quantitative Traits in Plant Breeding Using Pedigree and Molecular Markers. Genetics 186(2): 713–724.

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

Friedman, J., T. Hastie, and R. Tibshirani. 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. J. Stat. Softw. 33(1): 1–22.

Gauch, H.G., and R.W. Zobel. 1997. Identifying Mega-Environments and Targeting Genotypes. Crop Sci. 37(2): 311.

Gill, B.S., and W.J. Raupp. 1987. Direct Genetic Transfers from Aegilops squarrosa L. to Hexaploid Wheat1. Crop Sci. 27(3): 445.

Heffner, E.L., J.-L. Jannink, and M.E. Sorrells. 2011. Genomic Selection Accuracy using Multifamily Prediction Models in a Wheat Breeding Program. Plant Genome J. 4(1): 65.

Heffner, E.L., M.E. Sorrells, and J.-L. Jannink. 2009. Genomic Selection for Crop Improvement. Crop Sci. 49(1): 1.

Hennig, C. 2014. fpc: Flexible procedures for clustering.

Jannink, J.-L., A.J. Lorenz, and H. Iwata. 2010. Genomic selection in plant breeding: from theory to practice. Brief. Funct. Genomics 9(2): 166–177.

Kansas Wheat History. 2014. Available at http://www.nass.usda.gov/Statistics_by_State/Kansas/Publications/Crops/whthist.pdf (verified 16 February 2015).

Liaw, A., and M. Wiener. 2002. Classification and Regression by randomForest. R News 2(3): 18–22.

Meuwissen, T.H.E., B.J. Hayes, and M.E. Goddard. 2001. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics 157(4): 1819–1829.

Mevik, B.-H., R. Wehrens, and K.H. Liland. 2013. pls: Partial Least Squares and Principal Component regression.

Mujeeb-Kazi, A., and G.P. Hettel. 1995. Utilizing wild grass biodiversity in wheat improvement: 15 years of wide cross research at CIMMYT. Available at http://repository.cimmyt.org/xmlui/handle/10883/1204 (verified 4 December 2013).

Peterson, J.C. 1992. Similarities among Test Sites Based on Cultivar Performance in the Hard Red Winter Wheat Region. Crop Sci. 32(4): 907.

Poland, J.A., P.J. Brown, M.E. Sorrells, and J.-L. Jannink. 2012a. Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach. PLoS ONE 7(2): e32253.

Poland, J., J. Endelman, J. Dawson, J. Rutkoski, S. Wu, Y. Manes, S. Dreisigacker, J. Crossa, H. Sánchez-Villeda, M. Sorrells, and J.-L. Jannink. 2012b. Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing. Plant Genome J. 5(3): 103.

Quantum GIS Development Team. 2013. Quantum GIS Geographic Information System. Open Source Geospatial Foundation Project.

R Core Team. 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Revelle, W. 2013. psych: Procedures for Psychological, Psychometric, and Personality Research. Northwestern University, Evanston, Illinois.

Revolution Analytics, and S. Watson. 2014. foreach: Foreach looping construct for R.

Saintenac, C., D. Jiang, S. Wang, and E. Akhunov. 2013. Sequence-Based Mapping of the Polyploid Wheat Genome. G3 GenesGenomesGenetics 3(7): 1105–1114.

Shimodaira, H. 2002. An Approximately Unbiased Test of Phylogenetic Tree Selection. Syst. Biol. 51(3): 492–508.

Smiley, R.W. 2009. Root-Lesion Nematodes Reduce Yield of Intolerant Wheat and Barley. Agron. J. 101(6): 1322.

Suzuki, R., and H. Shimodaira. 2006. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics 22(12): 1540–1542.

Thompson, J.P. 2008. Resistance to Root-Lesion Nematodes (pratylenchus Thornei and P. Neglectus) in Synthetic Hexaploid Wheats and Their Durum and Aegilops Tauschii Parents. Aust. J. Agric. Res. 59(5): 432–446.

Todd, T.C., J. Appel, J. Vogel, and N.A. Tisserat. 2014. Survey of Plant-Parasitic Nematodes in Kansas and Eastern Colorado Wheat Fields. Plant Health Prog.Available at http://www.plantmanagementnetwork.org/php/elements/sum2.aspx?id=10778.

Urbanek, S. 2013. rJava: Low-level R to Java interface.

U.S. Geological Survey. 2005. Precipitation. Available at http://nationalatlas.gov/printable/images/pdf/precip/pageprecip_us3.pdf.

Zwart, R.S., J.P. Thompson, and I.D. Godwin. 2005. Identification of quantitative trait loci for resistance to two species of root-lesion nematode (Pratylenchus thornei and P. neglectus) in wheat. Aust. J. Agric. Res. 56(4): 345–352.