ABSTRACT

SARINELLI, JOSÉ MARTÍN. Genomic Selection and Association Mapping with a Historical Data Set of Southeastern USA Soft Red Winter Wheat. (Under the direction of Dr. Gina Brown-Guedira).

While world population is growing at a steady pace, wheat (*Triticum aestivum* L.) yields and the area planted with the crop are stagnated. The small annual increase in production needs to be reverted to satisfy the increasing demand for food. Resources have been invested to develop tools that can help plant breeders to increase the rate of genetic gain per breeding cycle by combining phenotypic records and molecular marker information. One such tool is Genomic Selection (GS) which utilizes prediction models to obtain genomic estimated breeding values (GEBVs) and make selections of un-phenotyped individuals using molecular marker data. With association mapping, phenotypic records and molecular marker information are utilized to detect marker-trait associations based on historical linkage disequilibrium in the population under study. The primarily aim of this research study was to validate the use of historical unbalanced soft red winter wheat cooperative nurseries from Southern and Mid-Atlantic USA from 2008 to 2016 to predict GEBVs. Additionally, genome wide scans were performed to detect marker-trait associations for resistance to powdery mildew of wheat (caused by *Blumeria graminis* sp. *tritici*).

In the first chapter, we used cross validation to evaluate predictive ability of genomic predictions from the historical Gulf-Atlantic Wheat Nursery (GAWN). We evaluated the impact of training population size and selection methods (Random, Clustering, PEVmean, PEVmean1) and the effect of inclusion major QTL as fixed covariates. Maximum predictive abilities were obtained for training population sizes of 200 individuals or more and using the selection method that minimized the prediction error variance between the training

population and the validation set (PEVmean, PEVmean1). Major gene covariates always increase predictive ability with the greatest increase when multiple covariates were combined. Maximum predictive abilities were 0.64, 0.56, 0.71, 0.73, 0.60 for grain yield, test weight, heading date, plant height and powdery mildew resistance respectively.

In the second chapter, we validated the results from the previous analysis with a scheme of forward prediction as occurs in ongoing breeding pipelines using two nurseries in different testing stages. The GAWN nursery was utilized as the training population while the Southern Uniform Wheat (SUNWHEAT) nursery, which is at an earlier testing stage, was used as a validation set. We measured prediction accuracy for different statistical frameworks using G-BLUP mixed models (two-step versus one-step and univariate versus multivariate) for grain yield and test weight. The one-step univariate model outperformed the two-step model by 9 percent with maximum prediction accuracy of 0.84 for grain yield. Results from the first two chapters provide empirical evidence for use unbalanced uniform nurseries as training population to incorporate genomic selection in the pipeline of ongoing breeding programs.

In the last chapter, we performed GWAS to detect marker trait associations for powdery mildew resistance in a panel of 862 winter wheat genotypes. A set of more than 14,000 SNPs obtained using genotyping by sequencing was utilized to study population structure, family relatedness and marker trait-association. We identified six regions associated with powdery mildew resistance located in chromosome 1A (2), 2B (1), 6B (1) and 7A (2). Loci identified in this study can be utilized for marker assisted selection for breeding new cultivars with multiple sources of resistance. It was demonstrated that a decrease in the average powdery mildew severity over time was associated with an increase

of the frequency of SNP alleles associated with resistance in modern cultivars. However, resistance was primarily related with single major genes.

Together these studies support the use of marker based breeding tools to improve complex traits and increase the rate of genetic gain by breeding cycle. Results from this dissertation were the basis for the adoption of GS within breeding programs of the SUNGRAINS cooperative breeding effort.

Genomic Selection and Association Mapping with a Historical Data Set of Southeastern USA Soft Red Winter Wheat

by
José Martín Sarinelli

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Crop Science

Raleigh, North Carolina

2017

APPROVED BY:

_____	_____
Dr. Gina Brown-Guedira Committee Chair	Dr. J. Paul Murphy
_____	_____
Dr. Consuelo Arellano	Dr. James B. Holland

DEDICATION

I dedicate this dissertation to my wife Cecilia Paoppi for her constant support, unconditional love, and for trust in me during this journey. To my parents, Blanca and José for their support and encouragement to make my dreams reality.

BIOGRAPHY

José Martín Sarinelli was born in 1980 of Blanca Meyssan and José Sarinelli. He grew up in Chivilcoy, an agricultural town in the province of Buenos Aires, Argentina. He has two sisters Julia and Celeste. He graduated in 1999 from high school. At the age of 18, he left Chivilcoy to start college at the Universidad de La Plata, Argentina. In 2004, he graduated with honors as an Agronomic Engineer. During summers, in the time-off high school and college, he worked with his father as a beekeeper. After graduation in 2004, he started his professional career in the seed industry. First, he worked as a Corn Assistant Breeder in Asociados Don Mario. Then, he decided to move and started a new job as a Corn Breeder in Advanta Seeds from 2008 to the 2010. At the end of 2010, he was recruited by Monsanto to work as a Commercial Breeder Associate in Pergamino, Buenos Aires. During his ten years working as an applied Corn Breeder in Argentina, he gained a lot of field experience and jointly with other breeders developed hybrids that remain available for farmers in the temperate and subtropical corn production regions of Argentina. At the beginning of 2014, he moved to Raleigh, North Carolina to start his Ph.D. in Plant Breeding and Genetics at NC State University. After graduation, Martin will continue working at NC State University in a post-doctoral research position under the direction of Dr. Gina Brown-Guedira.

ACKNOWLEDGMENTS

I would like to express my gratitude to my major advisor Dr. Gina Brown-Guedira, for her support, suggestions, advice and for pushing me one-step forward to make me a better graduate student. I would also like to recognize the constant support and guidance of my other committee members: Dr. Paul Murphy, Dr. James Holland and Dr. Consuelo Arellano, for their helpful advices related with courses, data analysis and manuscript preparation. I also owe my gratitude to Dr. David Marshall for being my graduate representative.

I would like to make a special recognition to Dr. Mohammed Guedira, my co-equipper for field and greenhouse activities, for his support and friendship. And to my first advisor in Argentina, Dr. Pedro Balatti for his advices, time and friendship during all my career.

To my family and friends for their encouragement, friendship and support. Specially, to my friend José (Pepe) Curioni, who helped me with his advice to move forward in the most difficult time of my graduate career at NC State University.

I would like to thank to all the people from the USDA Eastern Region Small Grains Genotyping Lab, Kim Howell, Sharon Williamson, Priyanka Tyagi, Jared Smith, Mai Xiong and Eddie Lauer for their support and for the time we shared at work. I would also like to make a special recognition to all SUNGRAINS breeders who developed the germplasm and provide me the phenotypic records to build my projects as well as their assistants who worked in the field collecting data.

Thank you to Dr. Charles Stuber for his trust and for supporting me during this time. Finally, a special acknowledgment to the Monsanto Fellowship in Plant Breeding for providing the funding for my graduate studies at NC State University.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1. LITERATURE REVIEW

Wheat

World production and importance

Common wheat (*Triticum aestivum* L.) is the most widely cultivated cereal in the world with more than 220 million ha planted across a variable range of climatic and edaphic conditions with an estimated mean production during the last decade of 619 million metric tons annually (FAO, http://faostat.fao.org/). Wheat placed third in the rank of cereals produced worldwide after corn (*Zea mays* L.) and rice (*Oryza sativa* L.). The top five world wheat producers are: China, India, United States of America, Russia Federation and France with 107, 73, 60, 44 and 36 million metric tons per year, respectively (http://faostat.fao.org/). Over the past five decades, world wheat yields have grown at an average rate of 2.2 percent per annum (Shiferaw et al., 2013).

World wheat yields during the three decades after the beginning of the Green Revolution had annual increases of about 4 percent per year that coincided with the development of dwarf, high yielding cultivars, combined with an increase in the use of fertilizers and irrigation. From the beginning of the 2000, annual yield increase stagnated to 1.3 percent per annum. As a result, global production expanded from 222 million metric tons in 1961 to around 620 million tons in 2008-10 (Shiferaw et al., 2013). The main driver for the increase in total production was genetic and management improvements and not due to a substantial increase in the area cultivated with the crop (Curtis et al., 2002, Marshall et al., 2001).

Wheat is classified into different categories depending on grain hardiness (soft, medium-hard and hard) and grain color (red and white). These categories can be also subdivided into subclasses based on spring and winter growth habit (Peña, 2002).

Wheat grains are rich in carbohydrates, minerals and proteins used primarily for human consumption in a wide range of products such as bread, pasta or as whole grain. Wheat contributes about 20 percent of daily calories and protein to the human diet with an average per capita consumption of 92 kg/year (Shiferaw et al., 2013). Wheat is a major dietary component in many countries because wheat exhibits agronomic adaptability, ease of storage and ease of converting grain into flour for making foods. Wheat can also be used for animal feed as grain, silage or consumption of the whole plant. (Curtis et al., 2002). Worldwide, 71 percent of wheat is used for food, 18 percent is used for animal feed and the remaining 11 percent is destined to industrial processing and other uses (Shiferaw et al., 2013).

With the global population expected to surpass nine billion in 2050, the FAO in 2009, using different predictions models, suggested it is necessary to increase the agricultural production 60 or 70 percent from 2005 to 2050 (Wise, 2013). The improvement of crop yields is critical for food security. There are several ways to increase crop yield per unit area, for example, by increasing the use of inputs (fertilizer, water, pesticides) or by improving the intrinsic yield of hybrids and cultivars through genetic improvement.

United States and North Carolina production

In the United States of America (USA), most of the wheat is grown in the Great Plains from Texas to North Dakota. During the period 2003-2013 the total production was 59 million metric tons per year with a mean yield of 2.9 metric ton/ha and an average area of 21 million hectares across the country (http://faostat.fao.org/).

Around 70-80 percent of the planted area in the USA is winter wheat, sown in the fall and harvested in the spring or early summer while the other 20-30 percent is spring wheat planted in the spring and harvested in late summer or early fall (Mergoum et al., 2009). The five principal market classes grown in the USA listed in order of acreage are: hard red winter wheat (HRWW) in the central and southern Great Plains, hard red spring wheat (HRSW) in the Northern Great Plains, soft red winter wheat (SRWW) in the Southeastern area, white wheat primarily in Pacific Northwest and durum wheat is concentrated in North Dakota (Curtis et al., 2002).

The eastern USA wheat growing region comprises an area from Louisiana to New York and from eastern Kansas to the Atlantic coastal plain which also includes North Carolina. In this region the most important wheat class is SRWW, characterized by relatively low protein content, used in flat breads, cakes, pastries, crackers, waffles, pretzels, soup thickeners, biscuits and as animal feed (Baenziger et al., 2009). In North Carolina the mean area planted with wheat during the last three years was approximately 340,000 hectares with an average yield of 1.5 metric ton/ha (http://www.usda.gov).

Origin, Evolution and Genetics

During the Neolithic revolution around 10,000-9,000 BC a transition from nomadic to agrarian lifestyles coincided with the beginning of agriculture. The Fertile Crescent, a geographic region located in the Middle East extending from Jordan, Israel, Lebanon, and Syria through southeast Turkey and along the Tigris and Euphrates rivers through Iraq and western Iran, was one of the domestication centers for different crops and animals. Two of the ancestors of common wheat, the diploid wild einkorn (*Triticum monococcum* ssp *aegilopoides* L.) and tetraploid wild emmer (*T. turgidum* ssp. *dicoccoides* L.), were grown in that area and used as staple crops for early civilization during that period (Faris, 2014, Feldman, 2001).

Common wheat is an annual, self-pollinated allohexaploid species (2n=6x=42, AABBDD) with 21 pairs of chromosomes grouped into three different genomes A, B, D. The three different genomes had divergent but related origins. The basic number of chromosomes in each genome is seven and they share homoeologous relationships with corresponding chromosomes in the other genomes (Sear, 1954). The chromosomes exhibit diploid segregation during meiosis controlled by the Ph1 (pairing homoeologous) gene located on chromosome 5B, which suppress the pairing between homoeologous chromosomes (Riley and Chapman, 1958). The wheat genome is complex with a total size of around 17 Gigabase-pairs (Gbp, Brenchle et al., 2012). Sequence and de novo assembly of each chromosome arm in hexaploid wheat cultivar Chinese Spring yield an average genome size for each genome of 5.727, 6.274 and 4.937 Gbp for the A, B and D genome respectively. In the same experiment,

also were identified 133,090 genes of which 40,253 where located in A genome, 44,523 in the B genome and 39,425 in the D genome (Mayer et al., 2014).

There are no wild hexaploid progenitors of common wheat. Cultivated wheat originated from a spontaneous hybridization event and posterior chromosome doubling between cultivated allotetraploid *T. turgidum* ssp. *dicoccum* (2n=4x=28, AABB), the donor of the A and B genome, and a wild diploid species *Aegilops tauschii* ssp. *strangulata* (2n=2x=14, DD), the donor of the D genome (Feldman, 2001, McFadden and Sears, 1946, Petersen et al., 2006). The allotetraploid progenitor, *T. turgidum* ssp. *dicoccum*, originated through hybridization and further chromosome duplication between two wild diploids *Triticum urartu* which contributed the A genome and possibly *Aegilops speltoides* contributing the B genome. However, despite decades of intensive research the origin of the B genome remains controversial (Gustafson et al., 2009, Petersen et al., 2006).

Genetic basis of Wheat Improvement

Wheat domestication involved three cultivation phases each characterized by phenotypic changes to adjust the crop to the farming environment (Feldman, 2001). During the first phase, coincident with the transition from wild to cultivated habitat, the most important changes were non-brittle spikes, free-threshing, non-dormant seeds, uniform and rapid germination. During the second phase and associated with the production of landraces in diverse environments, the most important changes were adaptation to new regional environments, increased tillering, increased plant height, development of a canopy with wide horizontal leaves, increased competitiveness with weeds and other wheat genotypes and

modifications in processes that control the timing of various growth stages. The last cultivation phase occurred during the last century in monomorphic fields associated with modern breeding procedures. This phase involved increased yield in densely planted fields, a canopy with erect leaves, reduced height, improved harvest index, enhanced response to fertilizers and agrochemicals, increased resistance to lodging, pests and diseases (Faris, 2014, Feldman, 2001).

All the phenotypic character changes that occurred in the different cultivation phases have complex genetic system with a few or several hundreds of loci involved in their expression which were consciously or unconsciously manipulated by breeders in order to obtain cultivars better adapted to each local environment.

Genetics of plant height

Wheat breeders have expended considerable effort to incorporate elite agronomic characters in their genetic stocks giving rise to uniform high yielding cultivars. During the 1960s, for example, the deployment of reduced plant height or "dwarfing" genes which reduced plant height from 150 to around 85 centimeters were key in the implementation of the "Green Revolution" leading to increases in grain yield (Reynolds and Borlaug, 2006, Worland and Snape, 2001). The dwarfing genes, classified according to their sensitivity to gibberellic acid, resulted in plants having shorter stems thereby being better able to resist lodging in high input environments. To date, 21 *Rht* genes with major effects on decreasing plant height have been identified in wheat (McIntosh, 1998).

The most widely used dwarfing genes, *Rht-B1b* and *Rht-D1b*, are gibberellic acid insensitive and located in the short arm of the chromosomes 4B and 4D respectively. They were introduced from the Japanese cultivar "Norin 10" (Gale et al., 1975, Gale and Marshall, 1976). In 1948, Orville Vogel at Washington State University, in an attempt to reduce plant height and avoid lodging in highly fertilized environments, grown wheat lines with the dwarfing genes from "Norin 10" into the cultivar "Brevor 14". Genotypes derived from that cross become the main source of dwarfing genes *Rht-B1b* and *Rht-D1b* in U.S. and Mexican germplasm (Gale and Youssefian, 1985). Under the leadership of N.E. Borlaug at the Centro Internacional de Mejoramiento de Maiz y Trigo (CIMMYT) in Mexico, the dwarfing genes *Rht-B1b* and *Rht-D1b* were transferred to Mexican spring wheat, and in the course of the Green Revolution, the resulting semi-dwarf Mexican cultivars were taken to India, Pakistan, Iran and other Mediterranean countries (Worland and Snape, 2001). Today, more than 70 percent of the wheat cultivars worldwide contain at least one of the dwarfing genes used during the Green Revolution (Evans, 1998, Zhang et al., 2006). In the USA, over 90 percent of modern hard and soft winter wheat germplasms carry either *Rht-B1b* or *Rht-D1b* (Guedira et al., 2010).

Another important group of dwarfing genes sensitive to gibberellic acid, such as *Rht8c*, were introduced into Italian germplasm and then spread into Europe. *Rht8c* was derived from the Japanese cultivar Akakomugi and it is closely linked to the photoperiod insensitivity allele *Ppd-D1a* located in the short arm of chromosome 2D (Gale and Youssefian, 1985, Korzun et at., 1998). In USA wheat germplasm, the *Rht8c* allele is present in less than 10 percent of the commercial cultivars (Guedira et al., 2010).

Genetics of flowering time

The cultivation of common wheat in a wide range of environments became possible due to selection of optimum flowering time, determined by three groups of genes that collaborate and interplay permitting wide adaptation of wheat to different areas and climates from the Equator to the Artic (Kamran et al., 2014). Three groups of genes, Vernalization (*Vrn*), Photoperiod (*Ppd*) and earliness per se, are responsible for the transition from vegetative to reproductive stages in response to environmental stimuli, to ensure that floral initiation occurs at optimum temperatures (Kamran et al., 2014, Kiss et al., 2014, Stelmakh, 1997).

Vernalization is the acquisition or acceleration of the ability to flower by a chilling treatment (Chouard, 1960). The spring/winter cultivar growth habit is determined by the sensitivity of the alleles present at the vernalization response loci to different cold treatments. Winter wheats are sensitive to vernalization and tend to be sown during the autumn. During winter, they experience a cold treatment without floral initiation until temperatures becomes favorable. Spring wheats are sown in the spring and generally do not need cold temperatures to induce flowering (Worland and Snape, 2001).

Natural variation in vernalization requirement is regulated by three groups of genes: *Vrn1*, *Vrn2* and *Vrn3*. *Vrn1* is the most important group and regulates the transition of the apical meristem from vegetative to reproductive phase (Distelfeld and Dubcovsky, 2009). The group *Vrn1*, comprised of a series of homoeologous loci called *Vrn-A1*, *Vrn-B1* and *Vrn-D1*, are located on chromosomes 5A, 5B and 5D respectively (Kamran et al., 2014, Law et al., 1976, Law and Worland, 1997). The group of *Vrn2* loci act as flowering repressors and

are located on chromosome group 5 (Yan et al., 2004). The *Vrn3* series comprises *Vrn-A3*, *Vrn-B3* and *Vrn-D3* located on chromosome 7A, 7B, 7D, and accelerate the reproductive development of the apical meristem with the consequent acceleration of flowering time. There is a well document effect of the *Vrn3* genes as integrators of the vernalization and photoperiod response (Distelfeld and Dubcovsky, 2009, Kamran et al., 2014, Yan et al., 2006).

The presence of at least one dominant allele in any genome of the groups *Vrn1* and *Vrn3* genes confers spring growth habit, whereas the presence of homozygous recessive genotypes confers winter growth habit. In addition, the vernalization requirement for winter wheat varies as a function of the number of recessive homozygous genes present in each group (Guedira et al., 2016).

Photoperiod response is the second important genetic system determining flowering time. Photoperidism refers to the phenomenon occurring in plants which senses and responds to different day or night length by receiving signals in the form of phytochrome to induce flowering (Fosket, 1994). A photoperiod sensitive wheat cultivar will remain vegetative until increasing day lengths during the spring satisfy the photoperiod requirements and enable the plant to initiate flowering. On the other hand, a photoperiod insensitive variety can immediately switch to reproductive growth once temperatures increase during the spring.

The increased day length to satisfy photoperiod requirements in photoperiod sensitive cultivars is primarily genetically regulated by a series of homeologous loci, *Ppd1*, in the short arm of group 2 chromosomes. To date, three loci have been identified: *PpdA1*, *PpdB1*, and *PpdD1* located on chromosomes 2A, 2B, and 2D, respectively (Kamran et al., 2014, Law et

al., 1978, Scarth and Law, 1983, Welsh et al., 1973). Dominant alleles at *Ppd1* loci confer day length insensitivity, whereas recessive alleles at *Ppd1* loci make wheat day length sensitive (Pugsley, 1966). *PpdD1* is the most photoperiod insensitive locus followed by *PpdB1* and *PpdA1* (Worland and Snape, 2001).

Finally, earliness per se influences flowering time independently of environmental signals. It is a quantitatively inherited trait controlled by minor genes. It is not well documented compared with *Vrn* and *Ppd* genes because of the lesser relative importance, and the difficulties in discriminating the effects usually cofounded with photoperiod and vernalization (Kamran et al., 2014, Worland and Snape, 2001).

Genetics of diseases resistance

One of the most economically and environmentally friendly mechanism to control diseases in crops is through the use of genetic resistance. A key selection objective for wheat breeders consists of breeding for disease resistance. Disease resistance can be classified as qualitative, associated with a single resistance gene recognizing avirulence factors in the pathogen, and quantitative resistance usually conditioned by many genes with small effect (Poland and Rutkoski, 2016, Worland and Snape, 2001).

Globally important diseases of wheat include rusts (*Puccinia* spp.), powdery mildew (*Blumeria graminis* (DC) Speer f. sp. *tritici* emend. E. J. Marchal), fusarium head blight or scab (*Fusarium* spp.), *Stagonospora nodorum blotch* (*Parastagonospora nodorum*), bacterial diseases (*Xanthomonas* spp., *Pseudomonas* spp.) and viral diseases (Soil-borne mosaic virus, Wheat streak mosaic virus, Barley Yellow Dwarf Virus), producing a wide range of

symptoms in different parts of the plant and with variable degree of yield loss. There are numerous genes identified that confer resistance to different diseases in wheat (Cowger et al, 2012, McIntosh 1995, McIntosh et al., 2013).

In this study, we will make focus on resistance to powdery mildew. This disease is a highly prevalent disease in North Carolina and the wheat breeding program at North Carolina State University have a long tradition related with the incorporation of resistant genes in elite cultivars.

Powdery mildew is one of the most important foliar disease of common wheat worldwide. In the USA, powdery mildew is often observed during the early spring in wheat grown in the southeastern region including North Carolina (Parks et al., 2009). Plants affected by powdery mildew often have reduced tillering and tiller survival, produced fewer kernels per head and decreased kernel weight, test weight and protein content (Everts and Leath, 1992). Reported crop losses by powdery mildew in susceptible cultivars are up to 34 percent in severe epidemics (Johnson et al., 1979, Leath and Bowen, 1989). These deficiencies may also negatively impact milling and baking quality traits, including flour yield, kernel softness, and grain protein content (Everts et al., 2001).

The fungus is an obligate parasite, growing only on living tissue. Powdery mildew can affect all above ground parts of wheat including stem, leaves and spikes through the production of white to gray pustules primarily on the upper leaf surface. The optimal growing conditions are temperatures in the range of 10-22 °C and relative humidity above 95 percent (Parry, 1990).

One efficient way to control this fungal disease is using resistant germplasm. Genes associated with resistance to powdery mildew are designated *Pm*, and, in general, are major genes conferring hypersensitive responses. Cowger et al. (2012) in a detailed review of the disease described over 40 different loci located on different chromosomes, for example *Pm3a*, *Pm3b* and *Pm3c* located in the short arm of chromosome 1A (Sear and Briggle, 1966), *Pm1* and *Pm37* in the long arm of chromosome 7A (Perugini et al., 2008, Sear and Briggle, 1969). These genes confer resistance to different races of the pathogen and come from different germplasm sources including spring and winter wheat cultivars, landraces and related species.

Alien introgression

Oftentimes, wheat breeders introgress novels genes into their germplasm pools from wild relatives and other related species to incorporate alleles conferring disease and pest resistance, stress and salt tolerance and winter hardiness (Friebe et al., 1996). The level of difficulty in transferring genes from related species to wheat depends greatly on the evolutionary distance between the species involved.

Harlan and de Wet (1971) defined three different genes pools. In wheat for example, species belonging to the primary gene pool share homologous genomes and gene transfer between these species can be achieved by direct hybridization, homologous recombination, backcrossing and selection. The secondary gene pool of common wheat includes other polyploid *Triticum* spp. and *Aegilops* spp. having at least one homologous genome in common with *T. aestivum*. Gene transfer from these species is possible by homologous

recombination if the target gene is located on a homologous chromosome. Species belonging to the tertiary gene pool are more distantly related. Their chromosomes are not homologous to those of common wheat. Consequently, other strategies need to be employed, because gene transfer from these species cannot be achieved by homologous recombination (Friebe et al., 1996).

The term alien translocation refers to a chromosome abnormality caused by rearrangement of DNA segments between non-homologous chromosomes. Some of the most important and widely used alien introgressions in common wheat are the translocations created by artificial hybridization with *Secale cereale* (rye) *t1BL:1RS* and *t1AL:1RS*. These are carriers of several disease resistance genes such as *Pm8* resistance to powdery mildew, *Sr31* conferring resistance to stem rust, *Lr26* resistance to leaf rust and *Yr9* resistance to yellow rust (Worland and Snape, 2001). Another key translocation is the *t2BS-2GS:2GL-2BL* from the hybridization of common wheat with the allotetraploid *Triticum timopheevii*, which is the source of stem rust resistance gene *Sr36* and closely linked to powdery mildew resistance gene *Pm6* (Friebe et al., 1996).

Molecular Breeding

Prior to development of molecular markers in 1980 most of the genetic improvement in plant breeding was based on phenotypic selection of superior genotypes. The gradual incorporation of markers in plant breeding schemes introduced what is known today as "molecular breeding" or "marker assisted breeding". The primary application of molecular breeding is to indirectly identify and select superior individuals for the trait of interest, thus

removing the need of extensive and expensive phenotyping. Marker based breeding methods include different variants of marker assisted selection (MAS) including population enrichment, marker assisted recurrent selection, marker assisted backcrossing, marker assisted gene pyramiding and marker assisted background selection. These different variants of MAS tools rely on a few markers in high linkage disequilibrium (LD) with the QTL or QTLs associated with the trait of interest and explain a substantial proportion of the variance associated with the trait. With the advent of high throughput genotyping platforms and increased computing power, methods that simultaneously predict the effect of thousands of small QTLs in LD with markers gave rise to what is known as Genomic Selection (Ben-Ari and Lavi, 2012, Meuwissen et al., 2001, Yang et al., 2015).

Other uses of markers, indirectly related with the application of markers in plant breeding, include: identification and mapping of QLTs, studies of genetic diversity, as an aid to efficiently identification of parental germplasm (Ben-Ari and Lavi, 2012), conservation of genetic resources and test of genetic purity (Singh et al., 2015).

Molecular markers

A genetic marker is defined as a trait that is polymorphic and indicates the genotype of individuals that exhibit the trait. Genetic markers are tools utilized for plant identification, plant improvement as well as genetics studies. They allow assaying genetic variation and provide an efficient means to link phenotypic and genotypic variation (Henry, 2012, Varshney et al., 2005). Genetic marker can be grouped in three different categories:

morphological, proteins and DNA markers. Oftentimes the last two categories are known collectively as molecular markers.

The earliest genetic markers were morphological markers which are based on phenotyping easily identifiable polymorphism, for example leaf shape or leaf color. These markers were replaced by isozymes pertaining to the group of protein markers. Isozymes are alternative forms or structural variants of an enzyme that have different molecular weights and electrophoretic mobility but have the same catalytic activity or function. Isozymes reflect the products of different alleles rather than different genes because the difference in electrophoretic mobility is caused by point mutation as a result of amino acid substitution (Singh et al., 2015, Xu, 2010).

Molecular markers based on DNA polymorphism were initiated in the early 1980's. In the interim, DNA markers have become the most important group of genetic markers and technological advances allow them to be utilized to satisfy different needs such as: increased throughput, reduced cost, higher reproducibility, greater abundance and user friendliness (Singh et al., 2015). A variety of DNA markers, including restriction fragment length polymorphism (RFLP), random amplification of polymorphic DNA (RAPD), amplified fragment length polymorphisms (AFLP), microsatellites or simple sequence repeats (SSR), and single nucleotide polymorphism (SNP) have been developed in different crop plants (Agarwal et al., 2008, Phillips and Vasil, 2001, Singh et al., 2015). Among others classifications, DNA markers can be divided into two groups: non-polymerase chain reaction (PCR) based methods like RFLP and PCR methods including RAPDs, AFLPs, SSRs, SNPs.

A brief description of the DNA markers follows with special emphasis in SNP and sequencing technologies.

Restriction fragment length polymorphisms (RFLP)

RFLP were the first DNA markers utilized. The procedure consists in the digestion of purified DNA using restriction enzymes leading to the formation of different DNA fragments which are separated by gel electrophoresis. DNA fragments are denatured and transferred to a solid support (Southern blotting) to posterior exposure and hybridization to a labeled DNA probe. Finally, the bands involved in the hybridization are identified by autoradiography or by color development (Botstein et al., 1980).

Most of the RFLP markers are co-dominant and locus specific. RFLP genotyping is highly reproducible and the methodology is simple. However, RFLP analysis require high amounts of high quality DNA, has low genotypic throughput and is really difficult to automate. Most genotyping requires radioactive methods so it use is limited to specific laboratories (Xu, 2010).

Random amplification of polymorphic DNA (RAPD)

This technique refers to the utilization of a single, short, random-sequence oligonucleotide primer in a low stringency polymerase chain reaction (PCR) for the simultaneous amplification of several DNA fragments. The primer, which binds to many different loci, is used to amplify random sequences from a complex DNA template that it complementary to it. The amplified products are usually separated in an agarose gel and visualized using ethidium bromide staining (Williams et al., 1990).

The advantages of RAPD technology include: neither DNA probe or sequence information is required for the design of specific primers, the procedure does not involve blotting or hybridization steps making the technique quick, simple and efficient, requires less DNA than RFLP and the procedure can be automated. Development of species-specific markers is not required and the technique can be applied to any organism with minimal initial development. Low reproducibility is a major limitation of RAPD makers, particularly in ongoing genetic and plant breeding programs. Another issue with RAPD markers is their predominantly dominant nature (Xu, 2010).

Amplified fragment length polymorphisms (AFLP)

AFLP is a group of markers based on restriction digestion of DNA combined with PCR. The technique involves three steps: 1) restriction of the DNA and ligation of oligonucleotide adapters, 2) selective amplification of sets of restriction fragments, and 3) gel electrophoresis analysis of the amplified fragments. The selective amplification is achieved by the use of primers that extend into the restriction fragments, amplifying only those fragments in which the primer extensions match the nucleotides flanking the restriction sites (Vos et al., 1995).

In general, AFLP assays can be carried out using small amount of DNA, they have a very high multiplex ratio and genotyping throughput, and they are relatively reproducible across laboratories. They can be applied to any organism with no formal marker development required and a set of primers can be used with different species. However, there are some limitations because high DNA quality is needed, marker development is complicated and not

cost effective and they have low reproducibility compared with RFLP. AFLP are mostly dominant markers (Xu, 2010).

Microsatellites or simple sequence repeats (SSR)

Simple sequence repeats are tandem repeated motifs of 1-6 bp (Litt and Luty, 1989) which have a frequent occurrence in all plant and animal genomes (Tautz and Renz, 1984). One of the most important attributes of SSR is the high level of allele variation even within the same species (different number of repeats), making them valuable as genetic markers. Microsatellite loci are individually amplified by PCR using pairs of oligonucleotide primers specific to unique DNA sequence flanking the SSR sequence. Polymorphism are identified by analysis of variations in the length of the amplified fragment by gel electrophoresis or capillary electrophoresis systems (Zane et al., 2002). Microsatellite markers are used for genetic diversity studies, population genetics, evolutionary studies, genome analysis, gene mapping, marker-assisted selection, and association mapping (Kalia et al., 2011).

SSR markers have gained considerable importance in plant genetics and breeding owing to many desirable attributes including high polymorphism, codominant inheritance, high reproducibility, relative abundance, extensive genome coverage (including organellar genomes), amenability to automation and multiplexing. This technology requires very small amounts of DNA per assay. Limitations of SSRs markers include the technically complicated development, labor intensiveness and cost (Parida et al., 2009, Xu, 2010).

Single nucleotide polymorphisms (SNP)

SNP have become the most popular choice of molecular marker for genetic analyses in plant and animals. A SNP is an individual base pair difference between two DNA

sequences. SNP are the most abundant markers present in a genome (Henry, 2012). SNP are produced by either transition (C/T or A/G) or transversion (A/T, A/C, G/T, G/C). SNP may fall within coding or non-coding regions of genes, within the intergenic regions between genes, with different frequencies and in different chromosome regions. This marker system yields reliable and reproducible results, which are amenable to automation and high-throughput genotyping. The chief limitations of SNP are the high cost of equipment particularly for high-throughput genotyping (Xu, 2010).

The use of SNP as a molecular marker system utilizing already identified SNP involves a preliminary phase of polymorphism discovery at the nucleotide base pair level. The discovery of SNP has been achieved in a variety of ways but is now probably best achieved by DNA sequencing. A second phase related to the use of SNP genotyping platforms to score from a single SNP marker to a very large number of markers assayed using high-density SNP chips. The different genotyping platforms can be classified based on the different detection methods utilized: 1) DNA hybridization, 2) PCR amplification, 3) oligonucleotide ligation and 4) DNA replication (Singh et al., 2015, Sobrino et al., 2005).

To date, there are several high-throughput genotyping platforms commercially available that use SNP markers. KASP™ (Kompetitive Allele-Specific PCR, LGC Genomics, UK) is an allele-specific PCR-based assay with low level multiplexing. It uses two forward primers, one reverse primer, and two fluorescent oligonucleotides (FRET). The 3’ terminal base of one of the forward primers is complementary to one of the two SNP alleles, while that of the other primer is complementary to the other allele of the SNP locus (Singh et al., 2015).

Microarray platforms like Illumina Infinium BeadChips (Illumina Inc., San Diego, CA) and Affymetrix GeneChip (Affymetrix Inc., Santa Clara, CA) allow high levels of multiplexing. Both microarrays platforms mentioned are available for use in wheat with different numbers of SNP (Cavanagh et al., 2013, Kumar et al., 2014, Want et al., 2014). By using these platforms, cost per data point is less than SSR-based genotyping, but the SNP information of the target organism is a prerequisite for designing a genotyping project, resulting in increased cost and longer experimental time. Nonetheless, these platforms are still useful when combined with economical SNP discovery pipelines using Next Generation Sequencing (NGS) technologies (Kim et al., 2015).

Advances in NGS technologies that rely on massively parallel sequencing and imaging capture techniques yield several hundreds of millions of bases per run in tens to hundreds of individuals allowing simultaneous identification and genotyping of SNP loci (Davey et al., 2011, Shendure and Hi, 2008). Such techniques can be performed directly on genomic DNA in a single sequencing step and with mostly parallelized library preparation. NGS methods can be utilized to sequence the whole or just a fraction of the genome. For species with a big genome size like common wheat the use of whole genome sequencing is time and cost prohibitive for implementation on a routine basis. Therefore, methods to reduce genome complexity have been developed in order to avoid sequencing repetitive DNA and reduce the size of the genome to be sequenced. The reduced representation library (RRL) approach, for example, cuts the entire genome with specific restriction enzyme to reduce genome complexity for the organism of interest. One limitation of RRL is that important genomic regions may not be captured by these libraries when restriction sites are not

available surrounding those regions. In addition, in order to increase the throughput of the sequencing run, barcoded adapters need to be incorporated into the system to keep track of different individuals (Kim et al., 2015).

Genotyping by sequencing (GBS)

The first high throughput method that combines NGS with RRL and DNA barcoded adapters to sequence multiplex libraries in parallel was restriction site-associated genomic DNA (RAD-seq) allowing high-density SNP discovery and genotyping (Baird et al., 2008). In this method each individual in the population is digested with a single rare cutter restriction enzyme that produces sticky ends. The fragments produced are ligated to a barcoded adapter (P1) which allows identification of the different individuals. DNA fragments from different individuals with the P1 barcode adapter are pooled, randomly sheared and size-selected. DNA is then ligated to a second adapter (P2), a Y adapter that has divergent ends. The structure of the P2 adapter ensures that only P1 adapter ligated DNA fragments will be amplified during the PCR amplification step. Finally, these amplified pools are sequenced in the sequencer instrument. The raw sequence reads can be aligned to a reference genome sequence of the species under study and SNP can be identified using an adequate software. In case a reference genome is not available, sequence reads representing the same genomic region are identified and classified into groups based on their sequence similarity. Polymorphisms are then identified between these reads.

Genotyping by sequencing (GBS) is an even more cost-effective genotyping procedure based on NGS technology and was developed as a modification of RAD-seq

protocol to simplify the DNA library construction. This protocol was initially developed for maize and then used for many other different species (Elshire et al., 2011, Poland et al., 2012). When compared with RAD-seq, the GBS protocol utilizes methylation sensitive enzymes which allow reduce genome complexity and target low copy number regions. In terms of barcoding strategy, it is similar to RAD-seq but modulation of barcode nucleotide composition and length results in fewer sequence phasing errors. Furthermore, in GBS, the generation of restriction fragments with appropriate adapters is more straightforward (single-well digestion of genomic DNA and adapter ligation results in reduced sample handling), and there are fewer DNA purification steps and fragments may or may not be size selected (Mir and Varshney, 2013). The flexibility of GBS in regards to species, populations and research objectives makes this an ideal tool for plant genetics studies like genome wide association studies (GWAS) and genomic selection (GS) (Poland and Rife, 2012, Yang et al., 2015). At this time, a large volume of plant research performed by GBS is available in scientific journals (ie. Arruda et al, 2016, Deschamps et al., 2012, He et al., 2014, Kim et al., 2015, Philomin et al., 2017, Poland and Rife, 2012, Sun et al. 2017).

Poland et al. (2012) introduced a modification to the original GBS protocol by replacing the original single restriction enzyme cutter with a two-enzyme system that includes one rare-cutter and one common-cutter to generate uniform libraries consisting of a forward (barcoded) adaptor and a reverse (Y) adaptor on alternate ends of each fragment. The use of two enzymes in this GBS approach enables the capture of most fragments associated with the rare-cutting enzyme.

The common steps after generating NGS reads are de-multiplexing into groups of individuals with the same barcode and aligning them to a reference genome using a variety of multiple alignment tools, or among themselves in cases where a reference genome is not available. Bowtie (Langmead et al., 2009) and Burrows-Wheeler-Aligner (BWA) (Li and Durbin, 2009) are frequently used tools for GBS reads alignment. The alignment information is processed by actual genotyping pipelines such as TASSEL-GBS which was designated for species with reference genomes or incomplete genome assemblies (Glaubitz et al., 2014). For species without reference genomes, an alternative pipeline "UNEAK" (Lu et al., 2013) embedded in TASSEL-GBS can be used but they generally require higher read coverage and stringent filtering steps to remove false positive SNPs (Kim et al., 2015).

One of the chief limitations of GBS is the large number of missing data produced. Polymorphic loci scored by GBS can contain a large proportion of missing data across samples because low sequence read depth, library complexity or multiplexing level, leading to some loci with no coverage in some individuals (Davey et al., 2011, Poland and Rife, 2012). However, most statistical analyses involving molecular marker data require a complete marker dataset to proceed, thus marker imputation is required.

Imputation can be performed in species with or without a reference genome. In the first case, the existence of a reference genome allows one to physically anchor the SNP loci in the correct order and to construct haplotypes. Examples of imputation software based on haplotype construction to impute the missing genotypes include Beagle (Browning and Browning, 2007) or fastPHASE (Scheet and Stephens, 2006) which rely on Hidden Markov

model algorithm (HMM) to assign missing genotypes by flanking markers. These methods are highly reliable and have been widely used in humans and cattle.

In the absence of a reference genome and due to the lack of ordered markers the previous methods cannot be utilized. The solutions in these cases are imputation methods that do not require markers in physical order. Several different algorithms to impute unordered missing data are available; however, they were not originally developed for this purpose. Rutkoski et al. (2013) compared five different algorithms to impute unordered missing data including: mean imputation, k-nearest neighbors imputation (Troyanskaya et al., 2001), singular value decomposition imputation (Troyanskaya et al., 2001), expectation maximization imputation (Dempster et al., 1977, Poland et al., 2012) and random forest imputation (Stekhoven and Bühlmann, 2011) concluding that random forest imputation is the more accurate but, in terms computational time required to perform imputation is highly demanding in comparison with the other methods utilized. LD-kNNi is an imputation method for unordered markers based on the k-nearest neighbor algorithm which consider the linkage disequilibrium (LD) between SNP when choosing the nearest neighbors. This approach gives similar accuracies to methods that have access to a reference genome (Money et al., 2015).

Linkage disequilibrium

Linkage disequilibrium (LD), also known as gametic phase disequilibrium is the nonrandom association of alleles at different loci within a population (Hedrick, 1987). LD is a population genetic term utilized to describe a population where the occurrence of certain

haplotypes is more frequent than we can expect based on independent segregation of alleles alone.

Oftentimes the terms linkage and LD are confused, but the phenomena are different. In linkage analysis, alleles of two different loci are inherited together because they are physically linked in the same chromosome. While, LD is the occurrence of nonrandom associations between alleles of two loci in a population irrespective of their physical location in the genome. It should be noticed that physical linkage of loci will create LD; however, LD can still be significant even for unlinked loci (Flint-Garcia et al., 2003).

The first measure utilized to estimate LD, "D", was proposed by Lewontin (1964). D quantifies disequilibrium as the difference between the observed frequency of an allelic combination in a sample and the product of the expected frequencies of the alleles under linkage equilibrium. So, for the case of two loci A and B with two alleles each one A1, A2 and B1, B2:

$$D_{AB} = P_{A1B1} - (P_{A1} * P_{B1})$$

where, P_{A1B1} are the observed frequency of the allelic combination A1B1 in the sample and P_{A1} and P_{B1} are the observed frequencies of the alleles in the same sample (Singh et al., 2015). Values of D varies from -0.25 to +0.25 where zero represents linkage equilibrium or independent segregation of alleles and measures that are significantly different from zero indicate that LD exist.

LD is usually found in natural populations between loci for which recombination has not had enough time to dissipate the initial disequilibrium. Under random mating the amount of disequilibrium is reduced every generation. The LD decay in random mating population is

a function of recombination fraction and time measured in generations from the new mutation was created (time 0). Falconer and Mackay (1996) describe a formula for which LD decays with time and recombination frequency:

$$D_t = (1-\theta)^t * D_0$$

where, D_0 and D_t represent the LD at generation 0 and t respectively, while θ is the recombination fraction and t is time in generations.

Because D is allele frequency dependent, it is difficult to compare LD between loci within the same population. Several different methods to make standardized measures of LD have been proposed: r^2, D', D2, d, Q and δ. The preferred LD measure in plants are r^2 and D'. With D' accounting for recombination history and less sensitivity to allele frequency changes and r^2 accounting for recombination and mutation history. Neither methods perform really well under low allele frequency and small population sizes (Gupta et al., 2005).

The most common measure of LD utilized in association analysis studies is r^2. It represents the square correlation coefficient between the alleles of the two loci considered.

$$r^2 = \frac{(D_{AB})^2}{(P_{A1} * P_{B1} * P_{A2} * P_{B2})}$$

where, P_{A1}, P_{B1}, P_{A2} and P_{B2} are the observed frequencies of alleles A1, B1, A2, and B2 respectively in the population and (D_{AB}) is the estimate of D between the two loci. The magnitude of r^2 ranges from 0 for loci with independent segregation, to 1, for completely linked loci with identical allele frequency (Ardlie et al., 2002).

Measures of LD in populations are affected by recombination and allele frequency. As a result, there are several factors creating LD patterns in populations including selection,

migration, mutation, drift, admixture, population structure, population size and mating system (Flint-Garcia et al., 2003). In general, LD decays faster in outcrossing than in self-pollinated species because recombination is less effective in self-pollinated species where individuals are more homozygous than in outcrossing species. In terms of populations, LD decays faster in landraces or wild germplasm which represent a more diverse source of germplasm compared to elite germplasm which have been selected for only a few traits of agronomic interest. Lastly and not less important, LD decay for the most important crop species varies across chromosomes, within a chromosome and between different genomes in polypoid species (Gupta et al., 2005).

Analysis of LD in several populations of common wheat using different marker systems revealed that LD patterns vary substantially from one population to the other and also with the marker type used. A set of 251 winter wheat lines from US uniform scab nurseries evaluated between 2008 and 2010, screened using 346 DArT molecular markers had a mean distance at which LD decayed below r^2= 0.20, of 9.9 cM (Benson et al., 2012). Another study of LD decay conducted by Cabrera et al. (2014), in soft winter wheat with a set of historical and elite germplasm genotyped using 392 DArT molecular markers, reported LD decayed to r^2= 0.20 within 5 to 10 cM. Both studies concluded that these variations are associated with the source of germplasm, the wheat genome evaluated (A, B, D) and the position within each chromosome. Furthermore, a denser genome coverage with molecular markers would provide a greater insight into the extent of LD.

Population structure and family relatedness

The term population structure refers to the phenomenon that individuals within a population do not always represent a single homogeneous group, but they accommodate in subgroups based on the degree of relatedness between them (Sneller et al., 2009, Wright, 1951). Some of the causes of structure in plant populations are genetic bottlenecks, genetic drift, selection, local adaptation, breeding history and relatedness or shared pedigrees between individuals (Crossa et al., 2007, Yu et al., 2006).

Investigations of population structure and family relatedness in plant populations have frequently utilized genomic relationship matrices and principal components analyses. These methods use molecular marker data from all the individuals in the population and model the degree of association between them. For example, use of relationship matrix or kinship matrix (K) allows estimates of all possible pairwise covariance between the individuals in the population based on the proportion of molecular markers shared (VanRaden, 2008). The dimension reduction technique called Principal Component Analysis (PCA) explains the variance-covariance structure of all markers available in the population through a few linear combinations of these variables. The objective is to reduce the dimensionality of the data and allow an easy interpretation of the result in two or three-dimensional space (Price et al., 2010, Sneller et al., 2009).

In genome wide association or association mapping studies, structure within populations and degree of genetic relatedness between individuals can lead to spurious associations between markers and QTL associated with the trait of interest (Cabrera et al., 2015, Myles et al., 2009, Sorrells et al., 2009). Population structure and family relatedness

also plays a key role in the accuracy of genomic estimated breeding values predictions from genomic selection, with models yielding higher accuracies when the individuals in the training population and the validation set are more closely related (Akdemir et al., 2015, Habier et al., 2013 Isidro et al., 2015, Rincent et al., 2012).

Application of molecular marker in plant breeding

QTL and association mapping

The main goal of association mapping in plants is to identify potential allelic variants closely linked with a gene/QTL responsible of the phenotypic variation in a trait of interest and for application of molecular markers to germplasm characterization. Association mapping studies can be classified into family and population based mapping categories.

Family mapping, also referred to as linkage mapping or QTL mapping, involves the development of bi-parental or more complex populations using homozygous parents for linkage mapping of markers and QTLs. These populations comprise closely related families derived from common parents using a specific mating scheme. Linkage mapping considers phenotypic and genotypic variation among offspring of relatively few genotypes and relies solely on the recombination events that occur between the QTL and the marker after the two selected parents are crossed (Myles et al., 2009).

Thousands of QTL mapping studies are available in the literature in different crops, traits, mapping populations, marker types and statistical analyses. A search in Google Scholar for the title “QTL mapping in wheat” yields 44,600 results on the subject (July 19, 2017). Most of the QTL mapping experiments include four basic requirements. First, a

mapping population where two or more contrasting parents for the trait of interest are crossed to develop an F2, doubled haploid, recombinant inbred line population or complex populations, developed by crossing multiple parents. Second, a marker linkage map where different type of polymorphic markers covering the entire genome are used to construct a linkage map for the population, which depicts the order of the markers in each linkage group and the genetic distances between them in centimorgan (cM). Third, a collection of phenotypic data for the mapping population for the trait of interest, often in replicated trails over different years and locations. Fourth, QTL detection involving different approaches to detect the association between the markers and trait.

The different methods to detect marker trait association are grouped into single QTL mapping (single marker analysis, single interval mapping) or multiple QTL mapping (composite interval mapping, multiple interval mapping). All methods rely on regression, analysis of variance (ANOVA), maximum likelihood parameter estimation or Bayesian statistical methods for QTL detection. A detailed explanation of how each method works can be found in Rifkin (2012) and Singh et al. (2015).

The power and resolution of QTL mapping to detect markers trait association depend of the amount of recombination events that occurred, the number of markers utilized, the sample size and the heritability of the QTL under consideration. Often QTL mapping cannot be performed with large population sizes over many generations due to the cost and time required to perform the experiment, in that cases the resolution can be poor. Although, QTL studies provide a good way to localize genes to individual chromosomes, chromosomes arms

or genomic regions but, usually do not have enough resolution to locate the gene or the functional polymorphism (Singh et al., 2015).

Nonetheless, thousands of QTL mapping studies have been published in the literature. In wheat, a number of QTL have been successful applied in plant breeding (Bernardo, 2008). The QTL *Fhb1* from the Chinese cultivar Sumai 3, located on chromosome 3B and conferring resistance to Fusarium head blight (FHB, caused by *Fusarium graminearum*), constitutes one example of effective discovery and introduction of a QTL in wheat breeding schemes. Introgression of the QTL into different genetic backgrounds has shown a mean reduction of disease severity of 23 percent (Pumphrey et al., 2007).

In population mapping, also referred to as genome wide association mapping (GWAS) or LD based mapping, the mapping population consists of a diverse set of lines from natural populations, breeding populations or multiparent populations like nested association mapping populations (NAM) or multiparent advanced generation intercross populations (MAGIC). GWAS exploits the phenotypic and genotypic variation present in natural populations and makes inferences based on past recombination events and the historical LD generated between the QTL and the markers (Ersoz et al., 2007). Additionally, NAM and MAGIC populations allow the exploitation of the LD resulting from crosses between the parents of these populations and the historical LD present between them (Cavanagh et al., 2008, Yu et al., 2008). Because LD mapping takes advantage of historical recombination in the population, if large number of molecular markers (to cover the complete genome) and adequate sample size are utilized, it is considered potentially more efficient for

detecting marker trait associations and the contributions of rare alleles compared with bi-parental QTL mapping (De Silva and Ball, 2007).

For LD mapping studies to effectively identify marker trait associations, it is of primary importance to understand the LD structure of the population under consideration. The power of associations are determined by the extent of LD and sample size utilized in the study. If LD decays too fast within a region, large number of markers are required to scan that target region of the genome. On the other hand, if LD decays too slowly, the size of the haplotype blocks would be too large to reveal the underlying causative locus. So, the decay of LD over physical distance in the population determines the marker density required and the level of resolution that may be obtained in an association study (Flint-Garcia et al., 2003).

GWAS was first utilized in human genetics to map genes associated with complex diseases such as Alzheimer (Saunders et al., 1993). However, many of the initial associations discovered have not been consistently replicated and many of them have been spurious because the statistical model utilized to explain the association cannot account for the presence of problems related with population structure and family relatedness. In plants, factors that limit the application of LD mapping or lead to spurious associations include: population structure, pleiotropic and epistatic interactions, genotype by environment interactions, experimental design, weak statistical tests, small sample size, type of markers utilized, marker genome coverage, complexity of the trait under study and quality of phenotypic data utilized (Zhu et al., 2008).

In comparison with QTL mapping, GWAS benefits include: 1) does not require the development of populations because they are usually samples from existing materials, 2)

markers identified with GWAS can be directly used in MAS because they were identified in a collection of diverse germplasm, 3) higher resolution, and 4) possibility to utilize data available from previous studies. On the other hand, limitations of GWAS vs QTL mapping include: 1) results impacted by several populations parameters including selection history and population structure, 2) require higher genome marker coverage, and 3) power of association affected by low allele frequency.

The general procedure of LD mapping includes five different stages: 1) selection of the population for the study, 2) phenotyping, 3) genotyping, 4) determination of the level of relatedness between individuals in the population and the influence of population structure, and 5) testing the genotypes and phenotypes for their associations (Ersoz et al., 2007).

From the point of view of plant breeding programs Breseghello and Sorrells (2006) suggested three different types of populations for the implementation of GWAS in plants: 1) germplasm bank collections containing all the genetic diversity of the species, 2) elite breeding materials integrated with lines and checks evaluated in nurseries and 3) synthetic populations which can be a close approximation to random mating assumption because are designated to minimized inbreeding. Later, other types of populations were developed, including NAM populations, first developed for maize, which use recombinant inbred lines developed from a diverse set of lines crosses with a common parent (Yu et al., 2008) and MAGIC populations obtained by sequential crossing of multiple parents (Cavanagh et al., 2008).

The success of GWAS like QTL mapping depends on the accurate phenotyping and genotyping in the population of interest. It includes field or greenhouse evaluations with

replications extended through locations and years to increase the power of the analysis. In GWAS experiments using germplasm collections or a set of highly diverse materials, special care is needed regarding phenotyping due to the wide range of variation present that can affect accurate evaluation of all experimental entries.

Genotyping can be performed using different platforms. However, in modern GWAS studies, bi-allelic SNP markers are the genetic marker of choice because of their high frequency, low mutation rate and amenability to automation.

Because perfect random mating probably does not exist, complex patterns of population structure and relatedness between individuals have been created by nonrandom mating of individuals. These patterns need to be addressed before GWAS analysis because markers that appear significantly associated with the trait of interest, in reality, may only capture the relatedness between individuals in the population. This gives rise to spurious associations between the marker and the trait under study. Currently the statistical method of choice for GWAS studies in plants is a mixed linear model that accounts for population structure and pairwise genetic relatedness simultaneously called Q + K model (Yu et al., 2006, Myles et al., 2009). The mixed model equation for the Q + K mixed model method is:

$$Y = S\alpha + Q\delta + Zu + e$$

where, Y is the vector of phenotypic observations, α and δ are vectors of fixed effects due to marker and population structure respectively, u and e are the vectors of random effect due to polygene background and residuals respectively. Q is a matrix containing principal components, while S and Z are design matrices that relate Y with α and u. The variance structure associated with the random polygenic effect is defined as Var(u): $K\sigma_g^2$ with K being

the genomic relationship matrix and σ_g^2 the genetic variance. While the residual variance structure is Var(e): R σ_e^2 with R equal to an identity matrix and the residual variance σ_e^2.

A modification of the Q + K mixed model is the multi-locus mixed model approach (MLMM) that uses a stepwise mixed model regression with forward inclusion of significant loci and backward elimination. The criteria to stop forward addition is based on reduction of the variance of the polygenic effect as a consequence of the addition of fixed covariates to the model. The backward elimination is performed from the last forward addition in order to evaluate the best combination of covariates in the model. This MLMM is an improvement in comparison with the mixed model in terms of power to detect marker trait associations and reduce false discovery rate (Segura et al., 2012).

The original Q + K mixed model approach is robust in plants, animals and humans, however it can be computationally intense to deal with large data sets due to the extensive number of individuals tests performed during the analysis. To avoid the computational issue, two optimization algorithms were developed: 1) CMLM, to compress the number of individuals in the overall population by clustering them into groups, which is useful for large numbers of individuals in the population (Zhang et al., 2010), and 2) EMMAx, an efficient mixed model association algorithm which performs the GWAS analysis in two steps. First, it efficiently calculates variance parameters without individual markers as fixed effects in the model, and second, fits a new mixed model for each marker using the parameters fitted in the first step (Kang et al., 2010).

The GWAS analysis can be performed in different statistical software packages including the R-packages: GAPIT (Lipka et al., 2012), rrBLUP (Endelman, 2011), MLMM

(Segura et al., 2012), or a software package developed in Java called TASSEL (Bradbury et al., 2007).

Another important consideration in whole genome GWAS scans refers to the control of false positives (Type I error) due to the large amount of individual statistical tests performed, which can lead to spurious marker trait association if the cutoffs utilized are too liberal. In the GWAS context, Type I error refers to incorrect rejection of the null hypothesis of no association between the marker and trait under consideration. One alternative to control the Type I error rate for multiple comparison is the Bonferroni correction (familywise error rate) in which an α level for each individual test is previously chosen and then divided by the number of tests or markers in the experiment ($\alpha' = \alpha/m$). Then the modified α' level is utilized as the new threshold to declare a marker trait association significant or not. The Bonferroni correction is too conservative leading to a reduction in the power to detect marker trait associations as the number of markers evaluated increases. Another less stringent statistical method to address the multiple testing problem is the False Discovery Rate (FDR), which measures the rate of significant test that are truly null. FDR is estimated as proportion of false positives from the total number of associations declared significant (Benjamini and Hochberg 1995). An extension of FDR is the q value which provides a measure of significance for each individual test taking into account that simultaneously several tests are being performed (Storey and Tibshirani, 2003).

Several GWAS studies in wheat have been published identifying marker trait association for different diseases and key agronomic traits (ei. Breseghello and Sorrells, 2006, Crossa et al., 2007, Macaferri et al., 2015, Reif et al., 2011, Zhang et al., 2014). Other

examples include, Yu et al. (2011) who used 276 spring wheat breeding lines from CIMMYT genotyped with Diversity Array Technology and phenotyped in Njoro, Kenya for stem rust (*Puccinia graminis* sp. *tritici*) resistance. They identified 15 different loci associated with adult plant resistance. Arruda et al. (2016) characterized resistance to FHB resistance in a panel of 273 breeding lines from the midwestern and eastern United States phenotyped in the field in Urbana, IL. They were genotyped using GBS (19,992 SNP identified), and 10 different marker trait associations were found related to FHB resistance. Zanke et al. (2015) used GWAS analysis in a panel of 372 European winter and spring wheat lines to identify marker trait associations for grain weight. Wheat lines were genotyped with a 90k iSelect array in addition to 372 SSR markers. In this experiment the authors found more than 86 significant marker trait associations in 17 of the 21 wheat chromosomes.

Genomic Selection

Since the 1980's, molecular marker technologies have used several strategies to promote marker information as an aid to selection. Many approaches have been tested in order to increase genetic gains per cycle compared with phenotypic selection alone. Initially the detection of marker trait association and incorporation into breeding programs involved indirect selection for the trait of interest using markers. This approach was limited to major effect QTL in high LD with the marker and only a limited number of successful examples have been described, primarily related with monogenic traits. One main limitation of the initial MAS methods was the lack of accurate prediction of polygenic traits, such as yield, which typically involve several small effect QTLs (Bernardo, 2008, Holland, 2004).

To deal with complex polygenic traits, Meuwissen et al. (2001) introduced the concept of Genomic Selection (GS) in animal breeding, which allow the prediction of Genomic Estimated Breeding Value (GEBV) of individuals using all identified molecular markers simultaneously. Instead of considering only significant QTL in the model, this approach incorporates all loci regardless of the effect of each locus. In plant breeding, the concept is referred as Genomic Selection or Genome Wide Selection (Bernardo and Yu, 2007). To date, several GS simulation and empirical studies in major crops like maize (*Zea Mays* L.), wheat (*Triticum aestivum* L.), rice (*Oryza sativa* L.), oat (*Avena sativa* L.) and barley (*Hordeum vulgare* L.) have been published showing moderate to high accuracies for different traits (Asoro et al., 2011, Cabrera et al., 2015, Lian et al., 2014, Lorenz et al., 2012, Poland et al., 2012, Rutkoski et al., 2015, Spindel et al., 2015).

The main advantages of GS include greater genetic gains per unit time than phenotypic selection, in particular for plant and animal species with long generation time intervals because it permits the estimation of GEBV for each individual without the need for phenotyping. This reduces the time for each breeding cycle. In annual crops the genetic gain per cycle can be increased by modifying the selection intensity per cycle instead of reducing the generation time through evaluation of a greater number of genotypes without the need to phenotype. When compared with MAS, GS predicts breeding values more accurately for highly polygenic traits as well as identifies new parental germplasm for population development early in the breeding cycle, accelerating cultivar developments programs (Heslot et al., 2015, Lorenz et al., 2011).

Genomic Selection involves two different but related populations, 1) a training population and, 2) a population with selection candidates to which GS will be applied. The training population contains individuals with both phenotypic and genotypic data and is used to train the prediction model and obtain marker effect estimates. These marker effect estimates are subsequently used to calculate the GEBV of selection candidates with genotypic information available alone. The GEBVs of selection candidates are then used to select individuals for advance in the breeding cycle without phenotypic information (Hayes et al., 2009, Heffner et al., 2009, Jannink et al., 2010).

One main challenge for the implementation of GS in cultivar development programs is how to combine phenotypic and genotypic data to design the training population for prediction of GEBVs of selection candidates accurately. In wheat, data from historical uniform cooperative nurseries evaluated across different environments represent an interesting source for use as a training population (Storlie and Charmet, 2013, Rutkoski et al., 2015). These historical nurseries, even when they were not originally thought as training population in the context of GS, provide phenotypic records for different traits of interest evaluated across different environments. However, one drawback of these nurseries is the high degree of unbalance in terms of the number of identical genotypes evaluated in the historical series across years with only a few commercial cultivars or performance checks replicated across the complete series which help to maintain the connectivity across years. Additionally, there is another unbalance because not all genotypes evaluated in the historical series are available for genotyping creating an unbalance between the total number of lines with phenotypic records and the total number of lines that can be genotyped. To overcome

this, issue a two-step GS protocol can be implemented where first a best linear unbiased estimate (BLUE) of inbred lines using all phenotypic records can be obtained and then integrated with the genotypic information available to train the GS model.

The continued advance of high-throughput marker technologies which allow dense genome coverage and the reduction of genotyping cost per individual make the implementation of GS in different plant breeding programs an attractive alternative for selection even in complex or polyploid genomes such as common wheat (Crossa et al., 2010, Poland et al., 2012). The idea behind the use of dense genome wide coverage for GS is to capture with markers most of the genetic variance associated with the trait of interest through maximizing the number of QTL effects that can be estimated because they are in high LD with at least one marker (Hefner et al., 2009).

In the majority of the GS models using dense genome wide coverage the number of marker or predictor effects (p) that need to be estimated largely exceed the number of phenotypic records (n) available and in some cases, there is a high degree of collinearity between predictors. Under this condition the number of degrees of freedom available in the model to estimate each parameter under standard multiple linear regression is not enough, giving rise to what is called “large p, small n problem”. To confront the “large p, small n problem” several statistical estimation procedures have been developed including variable selection methods, shrinkage estimation methods, kernel methods, dimension reduction methods and machine learning methods (de los Campos et al., 2013, Jannink et al 2010, Lorenz et al., 2011).

Random regression best linear unbiased prediction (RR-BLUP), a shrinkage estimation method, also known as ridge regression (Whittaker et al., 2000) was utilized in the landmark GS paper of Meuwissen et al., (2001) for calculating the best linear unbiased predictor estimates simultaneously for all the markers in the model, by treating the markers as random effects. In RR-BLUP, all the marker effects are assumed to be normally distributed with mean zero with a common variance σ_g^2, where σ_g^2 is obtained by dividing the total genetic variance by the number of marker present in the study. Thus, in ridge regression models all markers are equally shrunk toward zero by the same scaling factor (all markers explain the same proportion of genetic variance) although the marker effects estimates are likely to differ from each other in the experiment (de los Campos et al., 2013, Lorenz et al., 2011, Piepho, 2009). Despite, the unrealistic assumption that individual markers have the same variance the model is still useful and widely utilized in plant breeding yielding higher accuracies when compared with stepwise regression and phenotypic BLUP methods (Habier et al., 2007, Meuwissen et al., 2001).

An alternative method, genomic best linear unbiased prediction (G-BLUP), equivalent to RR-BLUP, also has been used widely to estimate GEBVs. In G-BLUP each genotype is considered a random effect in the model and a genomic relationship matrix is utilized to estimate variance-covariance relationships between individuals based on all markers (VanRaden 2008).

The genomic relationship matrix is calculated using markers data, irrespective of the trait genetic architecture or marker density and distribution (Habier et al., 2007, Heslot et al., 2015, VanRaden, 2008). The genomic relationship matrix used in G-BLUP models captures

the degree of genetic relationship between individuals by estimation of the proportion of identical by state alleles shared by individuals in the population. Estimates of genomic relationship matrix between different genotypes in the population is more precise using genetic relationship than based on pedigree information alone. A detailed description of a widely-used method utilized to derived the genomic relationship matrix based on marker information can be found in VanRaden, 2008. The G-BLUP mixed model offers the flexibility to modeling different genomic prediction scenarios including multiple traits and multi-environment trials simultaneously without big computational challenges in terms of memory requirements when compared with other predictions approaches included RR-BLUP (de los Campos et al., 2013, Jia and Jannink, 2012).

Another group of widely used methods for GS based on shrinkage or penalized regression are the Bayesian models which differ from RR-BLUP by allowing predictor variables to have different variances following a specified prior distribution, meaning that each marker can be shrunken to zero at a different degree. The different variants of Bayesian regression methods including BayesA, BayesB, BayesC, BayesCπ differ in the way in which marker effects are considered in the model. For example, with BayesA, the marker variance distribution is sampled from an inverted chi-square distribution (χ^2) and allows each marker in the model to have an effect with his own variance N(0, σ_g^2) , BayesB considers that only some markers have an effect while others do not have effect at all with a probability π. For markers with an effect different from zero, an independent variance is estimated sampled from the same inverted χ^2 distribution as in BayesA. BayesCπ makes the same assumption for marker effect as BayesB but the variance for all markers with an effect different from

zero is assumed to be the same (de los Campos et al., 2013, Gianola et al., 2009, Habier et al., 2011, Meuwissen et al., 2001).

In the simulation experiment conducted by Meuwissen et al. (2001) different prediction methods were evaluated and they concluded that Bayesian methods outperformed Stepwise regression and RR-BLUP methods. BayesB was the method that reached higher accuracies. However, studies with empirical data showed that the performance of Bayesian methods compared with RR-BLUP or G-BLUP are similar and in many scenarios the latter methods produce higher accuracies (Heffner et al., 2011, Rutkoski et al., 2014, Zhong et al., 2009).

The methods previously described for GS do not distinguish between the effect of known major genes or random markers. Bernardo (2014), in a simulation study, evaluated the change in the accuracy of the GS model due to the incorporation of major known QTL under different trait heritabilities, number of major genes associated with the trait of interest, as well as the proportion of variance explained by each major gene. He concluded that the incorporation of known QTL as fixed effect in the model never was disadvantageous if the marker effect explained 10 percent or more of the total variance associated with the trait. Further, a cross validation study with a wheat data set containing more than 300 lines evaluated for stem rust resistance showed that GS models incorporating the effect of known major effect QTL outperformed scenarios where all markers were considered random effects (Rutkoski et al., 2014).

GS model predictive ability is evaluated as the Pearson correlation (r) between the GEBV and the estimated breeding values (EBV) obtained from empirical phenotypic data

through cross validation procedures or using empirical validation (Lorenz et al., 2011). The correlation r(GEBV/ EBV) provides an estimate of selection accuracy and can be further related to selection response or the genetic gain equation, $R=ir\sigma_A$, where i is the selection intensity and σ_A is the standard deviation of the additive variance (Falconer and Mackay, 1996). The accuracy of GS models are influenced by several factors including TP size (Arruda et al., 2015), population structure (Akdemir et al., 2015, Isidro et al., 2015, Rincent et al., 2012), marker type, marker density and imputation method utilized to impute genotypic missing data (Jacobson et al., 2015, Poland et al., 2012, Rutkoski et al., 2013), LD between markers and QTLs (Hayes et al., 2009) , presence of major effect QTLs (Arruda et al., 2016, Bernardo, 2014, Bian and Holland, 2017), heritability and genetic architecture of trait under evaluation (Rutkoski et al., 2015), statistical model utilized (Arruda et al., 2015, Storlie and Charmet 2013), and degree of relationship between TP and validation set (Akdemir et al., 2015, Habier et al., 2013, Isidro et al., 2015, Lorenz et al., 2012).

References

Agarwal, M., Shrivastava, N., and Padh, H. 2008. Advances in molecular marker techniques and their applications in plant sciences. Plant cell reports, 27: 617-631.

Akdemir, D. Sanchez, J.I. and Jannink, J.L. 2015. Optimization of genomic selection training populations with a genetic algorithm. Genetics Selection Evolution 47: 1-10.

Andolfatto, P., Davison, D., Erezyilmaz, D., Hu, T.T., Mast, J., Sunayama-Morita, T., and Stern, D.L. 2011. Multiplexed shotgun genotyping for rapid and efficient genetic mapping. Genome research, 21: 610-617.

Ardlie, K.G., Kruglyak, L., and Seielstad, M. 2002. Patterns of linkage disequilibrium in the human genome. Nature Reviews Genetics 3: 299-309.

Arruda, M.P., Brown, P.J., Lipka, A.E., Krill, A.M., Thurber, C., and Kolb, F.L. 2015. Genomic Selection for Predicting Head Blight Resistance in a Wheat Breeding Program. The Plant Genome, 8: 1-12.

Arruda, M.P., Brown, P., Brown-Guedira, G., Krill, A.M., Thurber, C., Merrill, K.R., Foresman, B.J., and Kolb, F.L. 2016. Genome-Wide Association Mapping of Fusarium Head Blight Resistance in Wheat using Genotyping-by-Sequencing. The Plant Genome, 9: 1-14.

Arruda, M.P., Lipka, A.E., Brown, P.J., Krill, A.M., Thurber, C., Brown-Guedira, G., Dong, Y., Foresman, B.J., and Kolb, F.L. 2016. Comparing genomic selection and marker-assisted selection for Fusarium head blight resistance in wheat (*Triticum aestivum* L.). Molecular Breeding, 36: 1-11.

Asoro, F.G., Newell, M.A., Beavis, W.D., Scott, M.P., and Jannink, J.L. 2011. Accuracy and training population design for genomic selection on quantitative traits in elite North American oats. The Plant Genome, 4: 132-144.

Baenziger, P., Graybosch, R., Van Sanford, D., and Berzonsky, W. 2009. Winter and specialty wheat. In Cereals, pp. 251-265. Springer US.

Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A., Selker, E.U., Cresko, W.A., and Johnson, E.A. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One, 3: 1-7.

Ben-Ari, G., and Lavi, U. 2012. Marker-assisted selection in plant breeding. Plant Biotechnology and Agriculture: Prospects for the 21st Century: 163-184.

Benjamini, Y., and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological) 289–300.

Benson, J., Brown-Guedira, G., Paul Murphy, J., and Sneller, C. 2012. Population structure, linkage disequilibrium, and genetic diversity in soft winter wheat enriched for fusarium head blight resistance. The Plant Genome, 5: 71-80.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., and Boutell, J.M. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature, 456: 53-59.

Bernardo, R. 2008. Molecular markers and selection for complex traits in plants: learning from the last 20 years. Crop Science, 48: 1649-1664.

Bernardo, R. 2014. Genomewide selection when major genes are known. Crop Science, 54: 68-75.

Bernardo, R., and Yu, J. 2007. Prospects for genome-wide selection for quantitative traits in maize. Crop Science, 47: 1082-1090.

Bian, Y., and Holland, J.B. 2017. Enhancing genomic prediction with genome-wide association studies in multiparental maize populations. Heredity, 118: 585-593.

Botstein, D., White, R.L., Skolnick, M., and Davis, R.W. 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. American journal of human genetics, 32: 314-331.

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

Brenchley, R., Spannagl, M., Pfeifer, M., Barker, G.L., D'Amore, R., Allen, A.M., McKenzie, N., Kramer, M., Kerhornou, A., Bolser, D., and Kay, S. 2012. Analysis of the bread wheat genome using whole-genome shotgun sequencing. Nature, 491: 705-710.

Breseghello, F., and Sorrells, M. E. 2006. Association analysis as a strategy for improvement of quantitative traits in plants. Crop Science, 46: 1323-1330.

Breseghello, F., and Sorrells, M.E. 2006. Association mapping of kernel size and milling quality in wheat (*Triticum aestivum* L.) cultivars. Genetics, 172: 1165-1177.

Briggle, L.W., and Sears, E.R. 1966. Linkage of resistance to *Erysiphe graminis* f sp. *tritici* (*Pm3*) and Hairy Glume (*Hg*) on Chromosome 1A of Wheat. Crop Science, 6: 559-561.

Browning, S.R., and Browning, B.L. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. The American Journal of Human Genetics, 81: 1084-1097.

Cabrera, A., Guttieri, M., Smith, N., Souza, E., Sturbaum, A., Hua, D., Griffey, C., Barnett, M., Murphy, P., Ohm, H., and Uphaus, J. 2015. Identification of milling and baking quality QTL in multiple soft wheat mapping populations. Theoretical and Applied Genetics, 128: 2227-2242.

Cabrera, A., Souza, E., Guttieri, M., Sturbaum, A., Hoffstetter, A., and Sneller, C. 2014. Genetic diversity, linkage disequilibrium, and genome evolution in soft winter wheat. Crop Science, 54: 2433-2448.

Cavanagh, C.R., Chao, S., Wang, S., Huang, B.E., Stephen, S., Kiani, S., Forrest, K., Saintenac, C., Brown-Guedira, G., Akhunova, A., and See, D. 2013. Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. Proceedings of the national academy of sciences, 110: 8057-8062.

Cavanagh, C., Morell, M., Mackay, I., and Powell, W. 2008. From mutations to MAGIC: resources for gene discovery, validation and delivery in crop plants. Current opinion in plant biology, 11: 215-221.

Chagné, D., Batley, J., Edwards, D., and Forster, J.W. 2007. Single nucleotide polymorphism genotyping in plants, In Association Mapping in Plants (pp. 77-94). Springer New York.

Chouard, P. 1960 Vernalization and its relation to dormancy. Annual Rev Plant Physiol, 11:191-237.

Cowger, C., Miranda, L., Griffey, C., Hall, M., Murphy, J.P., and Maxwell, J. 2012. Wheat powdery mildew. Disease resistance in wheat. CABI, Oxfordshire, 84-119.

Crossa, J., Burgueno, J., Dreisigacker, S., Vargas, M., Herrera-Foessel, S.A., Lillemo, M., Singh, R.P., Trethowan, R., Warburton, M., Franco, J., and Reynolds, M. 2007. Association analysis of historical bread wheat germplasm using additive genetic covariance of relatives and population structure. Genetics, 177: 1889-1913.

Crossa, J., de Los Campos, G., Pérez, P., Gianola, D., Burgueño, J., Araus, J.L., Makumbi, D., Singh, R.P., Dreisigacker, S., Yan, J., and Arief, V. 2010. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics, 186: 713-724.

Curtis, B.C., Rajaram, S., and Gómez Macpherson, H. 2002. Bread wheat: improvement and production. Food and Agriculture Organization of the United Nations (FAO).

Davey, J.W., Hohenlohe, P.A., Etter, P.D., Boone, J.Q., Catchen, J.M., and Blaxter, M.L. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Reviews Genetics, 12: 499-510.

de los Campos, G., Hickey, J.M., Pong-Wong, R., Daetwyler, H.D., and Calus, M.P. 2013. Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics, 193: 327-345.

De Silva, H.N., and Ball, R.D. 2007. Linkage disequilibrium mapping concepts. In Association mapping in plants (pp. 103-132). Springer New York.

Dempster, A.P., Laird, N.M., and Rubin, D.B. 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society. Series B (methodological), 1-38.

Deschamps, S., Llaca, V., and May, G.D. 2012. Genotyping-by-sequencing in plants. Biology, 1: 460-483.

Distelfeld, A., Li, C., and Dubcovsky, J. 2009. Regulation of flowering in temperate cereals. Current opinion in plant biology, 12: 178-184.

Dvorak, J., Akhunov, E.D., Akhunov, A.R., Deal, K.R., and Luo, M.C. 2006. Molecular characterization of a diagnostic DNA marker for domesticated tetraploid wheat provides evidence for gene flow from wild tetraploid wheat to hexaploid wheat. Molecular biology and evolution, 23: 1386-1396.

Dvorak, J., Luo, M.C., and Yang, Z.L. 1998. Genetic evidence on the origin of *Triticum aestivum* L. In The origins of agriculture and crop domestication. Proceedings of the Harlan symposium. ICARDA, Aleppo. 235-251.

Edwards, D., Forster, J.W., Chagné, D., and Batley, J., 2007. What Are SNPs?. In Association mapping in plants (pp. 41-52). Springer New York.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S., and Mitchell, S.E. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One, 6: 1-10.

Endelman, J.B. 2011. Ridge regression and other kernels for genomic selection with R package rrBLUP. The Plant Genome, 4: 250-255.

Ersoz, E.S., Yu, J., and Buckler, E.S. 2007. Applications of linkage disequilibrium and association mapping in crop plants. In Genomics-assisted crop improvement (pp. 97-119). Springer Netherlands.

Evans, L.T. 1998. Feeding the ten billion. Plant and population growth. Cambridge Univ. Press, Cambridge, UK.

Everts K.L., and Leath S. 1992. Effect of early season powdery mildew on development, survival, and yield contribution of tillers of winter wheat. Phytopathology 82:1273-1278.

Everts, K.L., Leath, S., and Finney, P.L. 2001. Impact of powdery mildew and leaf rust on milling and baking quality of soft red winter wheat. Plant Dis 85:423-429.

Falconer, D.S., and Mackay, T.F. 1996. Introduction to quantitative genetics. Fourth ed. Longman Science and Technology, Harlow, UK.

Faris, J.D. 2014. Wheat domestication: Key to agricultural revolutions past and future. In Genomics of plant genetic resources (pp. 439-464). Springer Netherlands.

Feldman, M., 2001 Origin of Cultivated Wheat. The World Wheat Book: A History of Wheat Breeding, edited by Bonjean, A. P., and Angus, W. J. 3- 56. Lavoisier Publishing Inc. in Paris, France.

Food and Agriculture Organization of the United Nations. FAOSTAT. FAOSTAT (Database). (Latest update: 07 Mar 2014) Accessed (19 May 2015). URI: http://data.fao.org/ref/262b79ca-279c-4517-93de-ee3b7c7cb553.html?version=1.0.

Fosket, D.E. 1994. Plant growth & development, a molecular approach. Academic Press, San Diego.

Friebe, B., Jiang, J., Raupp, W.J., McIntosh, R.A., and Gill, B.S. 1996. Characterization of wheat-alien translocations conferring resistance to diseases and pests: current status. Euphytica, 91: 59-87.

Gale, M.D., and Marshall, G.A. 1976. The chromosomal location of *Gai1* and *Rht1*, genes for gibberellin insensitivity and semi-dwarfism, in a derivative of Norin 10 wheat. Heredity, 37: 283-289.

Gale, M.D., and Youssefian, S. 1985. Dwarfing genes in wheat. p. 1–35. In G.E. Russell (ed.) Progress in plant breeding. Butterworths and Co., London.

Gale, M.D., Law, C.N., and Worland, A.J. 1975. The chromosomal location of a major dwarfing gene from Norin 10 in new British semi-dwarf wheats. Heredity, 35: 417-421.

Gianola, D., de los Campos, G., Hill, W.G., Manfredi, E., and Fernando, R. 2009. Additive genetic variability and the Bayesian alphabet. Genetics, 183: 347-363.

Glaubitz, J.C., Casstevens, T.M., Lu, F., Harriman, J., Elshire, R.J., Sun, Q., and Buckler, E.S. 2014. TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. PLoS One: 9: 1-11.

Gupta, P.K., Rustgi, S., and Kulwal, P.L. 2005. Linkage disequilibrium and association studies in higher plants: present status and future prospects. Plant molecular biology, 57: 461-485.

Gustafson, P., Raskina, O., Ma, X., and Nevo, E., 2009. Wheat evolution, domestication, and improvement. Wheat: science and trade. Wiley, Danvers, pp.5-30.

Habier, D., Fernando, R.L., and Dekkers, J.C.M. 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics, 177: 2389-2397.

Habier, D., Fernando, R.L., Kizilkaya, K., and Garrick, D.J. 2011. Extension of the Bayesian alphabet for genomic selection. BMC bioinformatics, 12: 1-12.

Habier, D., Fernando, R.L., and Garrick, D.J. 2013. Genomic BLUP decoded: a look into the black box of genomic prediction. Genetics, 194: 597-607.

Haider, N. 2013. The origin of the B-genome of bread wheat (*Triticum aestivum* L.). Russian Journal of Genetics 49: 263-274.

Harlan, J.R., and de Wet, J.M. 1971. Toward a rational classification of cultivated plants. Taxon, 509-517.

Hayes, B.J., and Goddard, M.E. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics, 157: 1819-1829.

Hayes, B.J., Bowman, P.J., Chamberlain, A.J., and Goddard, M.E. 2009. Invited review: Genomic selection in dairy cattle: Progress and challenges. Journal of dairy science, 92: 433-443.

He, J., Zhao, X., Laroche, A., Lu, Z.X., Liu, H., and Li, Z. 2014. Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. Frontiers in plant science, 5: 1-8.

Hedden, P. 2003. The genes of the green revolution. Trends Genet 19:5-9.

Hedrick, P.W. 1987. Gametic disequilibrium measures: proceed with caution. Genetics, 117: 331-341.

Heffner, E.L., Sorrells, M.E., and Jannink, J.L. 2009. Genomic selection for crop improvement. Crop Science, 49: 1-12.

Henry, R.J. 2012. Evolution of DNA Marker Technology in Plants, in Molecular Markers in Plants (ed R. J. Henry), Blackwell Publishing Ltd., Oxford, UK.

Heslot, N., Jannink, J.L., and Sorrells, M.E. 2015. Perspectives for genomic selection applications and research in plants. Crop Science, 55: 1-12.

Holland, J.B. 2004. Implementation of molecular markers for quantitative traits in breeding programs—challenges and opportunities. In New Directions for a Diverse Planet: Proceedings for the 4th International Crop Science Congress. Regional Institute, Gosford, Australia, www. cropscience. org. au/icsc2004.

International Wheat Genome Sequencing Consortium. 2014. A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. Science, 345(6194), 1251788.

Isidro, J., Jannink, J.L., Akdemir, D., Poland, J., Heslot, N. and Sorrells, M.E. 2015. Training set optimization under population structure in genomic selection. Theoretical and applied genetics, 128:145-158.

Jacobson, A., Lian, L., Zhong, S. and Bernardo, R. 2015. Marker imputation before genomewide selection in biparental maize populations. The Plant Genome, 8: 1-9.

Jannink, J.L., Lorenz, A.J., and Iwata, H. 2010. Genomic selection in plant breeding: from theory to practice. Briefings in functional genomics, 9: 166-177.

Jia, Y., and Jannink, J.L. 2012. Multiple-trait genomic selection methods increase genetic value prediction accuracy. Genetics, 192: 1513-1522.

Johnson, J.W., Baenziger, P.S., Yamazaki, W.T., and Smith, R.T.1979. Effects of powdery mildew on yield and quality of isogenic lines of Chancellor wheat. Crop Science 19: 349-352.

Kalia, R.K., Rai, M.K., Kalia, S., Singh, R., and Dhawan, A.K. 2011. Microsatellite markers: an overview of the recent progress in plants. Euphytica,177: 309-334.

Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.Y., Freimer, N.B., Sabatti, C., and Eskin, E. 2010. Variance component model to account for sample structure in genome-wide association studies. Nature genetics, 42: 348-354.

Kim, C., Guo, H., Kong, W., Chandnani, R., Shuang, L.S. and Paterson, A.H. 2016. Application of genotyping by sequencing technology to a variety of crop breeding programs. Plant Science, 242: 14-22.

Kiss, T., Balla, K., Veisz, O., Láng, L., Bedő, Z., Griffiths, S., Isaac, P., and Karsai, I. 2014. Allele frequencies in the VRN-A1, VRN-B1 and VRN-D1 vernalization response and PPD-B1 and PPD-D1 photoperiod sensitivity genes, and their effects on heading in a diverse set of wheat cultivars (*Triticum aestivum* L.). Molecular Breeding, 34: 297-310.

Korzun, V., Röder, M.S., Ganal, M.W., Worland, A.J., and Law, C.N. 1998. Genetic analysis of the dwarfing gene (*Rht8*) in wheat. Part I. Molecular mapping of *Rht8* on the short arm of chromosome 2D of bread wheat (*Triticum aestivum* L.). Theoretical and Applied Genetics, 96: 1104-1109.

Kumar, S., Wang, Z., Banks, T.W., Jordan, M.C., McCallum, B.D., and Cloutier, S. 2014. *Lr1*-mediated leaf rust resistance pathways of transgenic wheat lines revealed by a gene expression study using the Affymetrix GeneChip® Wheat Genome Array. Molecular Breeding, 34: 127-141.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome biology, 10: 1-10.

Law, C.N., and Worland, A.J. 1997. Genetic analysis of some flowering time and adaptive traits in wheat. New Phytologist, 137: 19-28.

Law, C.N., Sutka, J., and Worland, A.J. 1978. A genetic study of day-length response in wheat. Heredity, 41:185-191.

Law, C.N., Worland, A.J., and Giorgi, B. 1976. The genetic control of ear-emergence time by chromosomes 5A and 5D of wheat. Heredity, 36: 49-58.

Leath, S., and Bowen, K.L. 1989. Effects of powdery mildew, triadimenol seed treatment, and triadimefon foliar sprays on yield of winter-wheat in North Carolina. Phytopathology 79:152-155.

Lewontin, RC. 1964. The interaction of selection and linkage. I. General considerations; heterotic models. Genetics, 49: 49-67.

Li, H., and Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25:1754-1760.

Lian, L., Jacobson, A., Zhong, S., and Bernardo, R. 2014. Genomewide prediction accuracy within 969 maize biparental populations. Crop Science, 54: 1514-1522.

Lipka, A.E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P.J., Gore, MA., Buckler, E. S., and Zhang, Z. 2012. GAPIT: genome association and prediction integrated tool. Bioinformatics, 28: 2397-2399.

Litt, M., and Luty, J.A. 1989. A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. American journal of human genetics, 44: 397-401.

Lorenz, A.J., Smith, K.P., and Jannink, J.L. 2012. Potential and optimization of genomic selection for Fusarium head blight resistance in six-row barley. Crop Science, 52: 1609-1621.

Lorenz, A.J., Chao, S., Asoro, F.G., Heffner, E.L., Hayashi, T., Iwata, H., Smith, K.P., Sorrells, M.E., and Jannink, J.L. 2011. 2 Genomic Selection in Plant Breeding: Knowledge and Prospects. Advances in agronomy, 110: 77-123.

Lu, F., Lipka, A.E., Glaubitz, J., Elshire, R., Cherney, J.H., Casler, M.D., Buckler, E.S., and Costich, D.E. 2013. Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. PLoS Genet, 9: 1-14.

Maccaferri, M., Zhang, J., Bulli, P., Abate, Z., Chao, S., Cantu, D., Bossolini, E., Chen, X., Pumphrey, M., and Dubcovsky, J. 2015. A genome-wide association study of resistance to stripe rust (*Puccinia striiformis* f. sp. *tritici*) in a worldwide collection of hexaploid spring wheat (*Triticum aestivum* L.). G3: Genes| Genomes| Genetics, 5: 449-465.

Mascher, M., Wu, S., Amand, P.S., Stein, N., and Poland, J. 2013. Application of genotyping-by-sequencing on semiconductor sequencing platforms: a comparison of genetic and reference-based marker ordering in barley. PLoS One, 8: 1-11.

McFadden, E.S., and Sears, E.R., 1946. The origin of *Triticum spelta* and its free-threshing hexaploid relatives. Journal of Heredity, 37: 107-116.

McIntosh, R.A., Wellings, C.R., and Park, R.F. 1995. Wheat rusts: an atlas of resistance genes. Csiro Publishing.

McIntosh, R.A., Yamazaki, Y., Dubcovsky, J., Rogers, J., Morris, C., and Appels, R. 2013. Catalogue of Gene Symbols for Wheat. 12th International. Wheat Genetic. Symposium (pp. 1-31). Yokohama, Japan.

McIntosh, R.A. 1997. Breeding wheat for resistance to biotic stresses. In Wheat: prospects for global improvement (pp. 71-86). Springer Netherlands.

McVittie, J.A., Gale, M.D., Marshall, G.A., and Westcott, B. 1978. The intrachromosomal mapping of the Norin 10 and Tom Thumb dwarfing genes. Heredity, 40: 67-70.

Mergoum, M., Singh, P.K., Anderson, J.A., Peña, R.J., Singh, R.P., Xu, S.S., and Ransom, J.K. 2009. Spring wheat breeding. In Cereals (pp. 127-156). Springer US.

Mir, R.R., and Varshney, R. K. 2013. Future prospects of molecular markers in plants. Molecular markers in plants. Wiley, New York. Pp. 169-190.

Monson-Miller, J., Sanchez-Mendez, D.C., Fass, J., Henry, I.M., Tai, T.H., and Comai, L. 2012. Reference genome-independent assessment of mutation density using restriction enzyme-phased sequencing. BMC genomics, 13: 1-15.

Myles, S., Peiffer, J., Brown, P.J., Ersoz, E.S., Zhang, Z., Costich, D.E. and Buckler, E.S. 2009. Association mapping: critical considerations shift from genotyping to experimental design. The Plant Cell, 21: 2194-2202.

Nelson, G.C., Rosegrant, M.W., Palazzo, A., Gray, I., Ingersoll, C., Robertson, R., Tokgoz, S., Zhu, T., Sulser, T.B., Ringler, C., and Msangi, S. 2010. Food security, farming, and climate change to 2050: Scenarios, results, policy options (Vol. 172). Intl Food Policy Res Inst, Washington, D.C.

Parida, S.K., Kalia, S.K., Kaul, S., Dalal, V., Hemaprabha, G., Selvi, A., Pandit, A., Singh, A., Gaikwad, K., Sharma, T.R., and Srivastava, P.S. 2009. Informative genomic microsatellite markers for efficient genotyping applications in sugarcane. Theoretical and Applied Genetics, 118: 327-338.

Parks, R., Carbone, I., Murphy, J.P., and Cowger, C. 2009. Population genetic analysis of an eastern US wheat powdery mildew population reveals geographic subdivision and recent common ancestry with UK and Israeli populations. Phytopathology, 99: 840-849.

Parry, D.W. 1990. Diseases of small grain cereals. In: Plant Pathology in Agriculture Cambridge University Press, Cambridge, pp. 159-248.

Peña, R.J., 2002. Wheat for bread and other foods. Bread wheat improvement and production. Food and Agriculture Organization of the United Nations (FAO). Plant Production and Protection Series, Rome, Italy. pp. 483-542.

Perugini, L.D., Murphy, J.P., Marshall, D., and Brown-Guedira, G. 2008. *Pm37*, a new broadly effective powdery mildew resistance gene from Triticum timopheevii. Theoretical and Applied Genetics, 116: 417-425.

Petersen, G., Seberg, O., Yde, M., and Berthelsen, K. 2006. Phylogenetic relationships of Triticum and Aegilops and evidence for the origin of the A, B, and D genomes of common wheat (*Triticum aestivum* L.). Molecular phylogenetics and evolution, 39: 70-82.

Peterson, B.K., Weber, J.N., Kay, E.H., Fisher, H.S., and Hoekstra, H.E. 2012. Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS One, 7: 1-11.

Phillips, R.L., and Vasil, I.K. 2001. DNA-based markers in plants (Vol. 6). Springer Science & Business Media.

Philomin, J., Singh, R.P., Singh, P.K., Crossa, J., Rutkoski, J.E., Poland, J.A., Bergstrom, G.C., and Sorrells, M.E. 2017. Comparison of Models and Whole-Genome Profiling Approaches for Genomic-Enabled Prediction of Septoria Tritici Blotch, Stagonospora Nodorum Blotch, and Tan Spot Resistance in Wheat. The Plant Genome, 10: 1-16.

Poland, J.A., and Rife, T.W. 2012. Genotyping-by-sequencing for plant breeding and genetics. The Plant Genome, 5: 92-102.

Poland, J.A., Brown, P.J., Sorrells, M.E., and Jannink, J.L. 2012. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS One, 7: 1-8.

Poland, J., and Rutkoski, J. 2016. Advances and Challenges in Genomic Selection for Disease Resistance. Annual Review of Phytopathology, 54: 79-98.

Poland, J., Endelman, J., Dawson, J., Rutkoski, J., Wu, S., Manes, Y., Dreisigacker, S., Crossa, J., Sánchez-Villeda, H., Sorrells, M., and Jannink, J.L. 2012. Genomic selection in wheat breeding using genotyping-by-sequencing. The Plant Genome, 5: 103-113.

Price, A.L., Zaitlen, N.A., Reich, D., and Patterson, N. 2010. New approaches to population stratification in genome-wide association studies. Nature Reviews Genetics, 11: 459-463.

Pugsely, A.T. 1966. The photoperiodic sensitivity of some spring wheats with special reference to the variety Thatcher. Crop and Pasture Science, 17: 591-599.

Pumphrey, M.O., Bernardo, R., and Anderson, J.A. 2007. Validating the QTL for Fusarium head blight resistance in near-isogenic wheat lines developed from breeding populations. Crop Science, 47: 200-206.

Reif, J.C., Gowda, M., Maurer, H.P., Longin, C.F.H., Korzun, V., Ebmeyer, E., Bothe, R., Pietsch, C., and Würschum, T. 2011. Association mapping for quality traits in soft winter wheat. Theoretical and Applied Genetics, 122: 961-970.

Reynolds, M.P., and Borlaug, N.E. 2006. Impacts of breeding on international collaborative wheat improvement. The Journal of Agricultural Science, 144: 3-17.

Rifkin, S.A. 2012. Quantitative trait loci (QTL): methods and protocols. Humana Press.

Riley, R., and Chapman V. 1958. Genetic control of the cytological diploid behavior of hexaploid wheat. Nature, 182: 713-715.

Rincent, R., Laloë, D., Nicolas, S., Altmann, T., Brunel, D., Revilla, P., Rodriguez, V.M., Moreno-Gonzalez, J., Melchinger, A., Bauer, E. and Schön, C.C. 2012. Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (Zea mays L.). Genetics, 192: 715-728.

Rothberg, J.M., Hinz, W., Rearick, T.M., Schultz, J., Mileski, W., Davey, M., Leamon, J.H., Johnson, K., Milgrew, M.J., Edwards, M., and Hoon, J. 2011. An integrated semiconductor device enabling non-optical genome sequencing. Nature, 475: 348-352.

Rutkoski, J.E., Poland, J.A., Singh, R.P., Huerta-Espino, J., Bhavani, S., Barbier, H., Rouse, M.N., Jannink, J.L., and Sorrells, M.E. 2014. Genomic selection for quantitative adult plant stem rust resistance in wheat. The plant genome, 7: 1-10.

Rutkoski, J.E., Poland, J., Jannink, J.L., and Sorrells, M.E. 2013. Imputation of unordered markers and the impact on genomic selection accuracy. G3: Genes Genomes Genetics, 3: 427-439.

Rutkoski, J., Singh, R.P., Huerta-Espino, J., Bhavani, S., Poland, J., Jannink, J.L., and Sorrells, M.E. 2015. Efficient use of historical data for genomic selection: a case study of stem rust resistance in wheat. The Plant Genome, 8: 1-10.

Salamini, F., Özkan, H., Brandolini, A., Schäfer-Pregl, R., and Martin, W. 2002. Genetics and geography of wild cereal domestication in the Near East. Nature Reviews Genetics, 3: 429-441.

Saunders, A.M., Strittmatter, W.J., Schmechel, D., George-Hyslop, P.S., Pericak-Vance, M.A., Joo, S.H., Rosi, B.L., Gusella, J.F., Crapper-MacLachlan, D.R., Alberts, M.J., and Hulette, C. 1993. Association of apolipoprotein E allele ϵ4 with late-onset familial and sporadic Alzheimer's disease. Neurology, 43: 1467-1467.

Scarth, R., and Law, C.N. 1983. The location of the photoperiod gene, *Ppd2* and an additional genetic factor for ear-emergence time on chromosome 2B of wheat. Heredity, 51: 607-619.

Scheet, P., and Stephens, M. 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. The American Journal of Human Genetics, 78: 629-644.

Sears, E.R. 1954. The aneuploids of common wheat. Mol. Agr. Exp. Sta. Res. Bull. 572: 1-59.

Sears, E.R., and Briggle, L.W. 1969. Mapping the gene *Pml* for resistance to *Erysiphe graminis* f. sp. *tritici* on chromosome 7A of wheat. Crop science, 9: 96-97.

Segura, V., Vilhjálmsson, B.J., Platt, A., Korte, A., Seren, Ü., Long, Q., and Nordborg, M. 2012. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nature genetics, 44: 825-830.

Shendure, J., Ji, H. 2008. Next-generation DNA sequencing. Nat. Biotechnol, 26: 1135-1145.

Shiferaw, B., Smale, M., Braun, H. J., Duveiller, E., Reynolds, M., and Muricho, G. 2013. Crops that feed the world 10. Past successes and future challenges to the role played by wheat in global food security. Food Security, 5: 291-317.

Singh, B.D., and Singh, A.K. 2015. Marker-assisted plant breeding: principles and practices. New Delhi: Springer.

Sneller, C.H., Mather, D.E., and Crepieux, S. 2009. Analytical approaches and population types for finding and utilizing QTL in complex plant populations. Crop science, 49: 363-380.

Sobrino, B., Brión, M., and Carracedo, A. 2005. SNPs in forensic genetics: a review on SNP typing methodologies. Forensic science international, 154: 181-194.

Sorrells, M.E., and Yu, J. 2009. Linkage disequilibrium and association mapping in the Triticeae. In Genetics and genomics of the Triticeae (pp. 655-683). Springer US.

Spindel, J., Begum, H., Akdemir, D., Virk, P., Collard, B., Redoña, E., Atlin, G., Jannink, J.L., and 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 Genet, 11: 1-25.

Stekhoven, D.J., and Bühlmann, P. 2012. MissForest non-parametric missing value imputation for mixed-type data. Bioinformatics, 28: 112-118.

Stelmakh, A.F. 1997. Genetic systems regulating flowering response in wheat. In Wheat: Prospects for Global Improvement (pp. 491-501). Springer Netherlands.

Stolle, E., and Moritz, R.F. 2013. RESTseq–efficient benchtop population genomics with RESTriction Fragment SEQuencing. PLoS One, 8: 1-5.

Storey, J.D., and Tibshirani, R. 2003. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences, 100: 9440-9445.

Storlie, E., and Charmet, G., 2013. Genomic selection accuracy using historical data generated in a wheat breeding program. The Plant Genome, 6: 1-9.

Sun, J., Rutkoski, J.E., Poland, J.A., Crossa, J., Jannink, J.L., and Sorrells, M.E. 2017. Multitrait, Random Regression, or Simple Repeatability Model in High-Throughput Phenotyping Data Improve Genomic Prediction for Wheat Grain Yield. The Plant Genome, 10: 1-12.

Tautz, D., and Renz, M. 1984. Simple sequences are ubiquitous repetitive components of eukaryotic genomes. Nucleic acids research, 12: 4127-4138.

Thudi, M., Li, Y., Jackson, S.A., May, G.D., and Varshney, R.K. 2012. Current state-of-art of sequencing technologies for plant genomics research. Brief Funct Genomics 11: 3-11.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R.B. 2001. Missing value estimation methods for DNA microarrays. Bioinformatics, 17: 520-525.

Truong, H.T., Ramos, A.M., Yalcin, F., de Ruiter, M., van der Poel, H.J., Huvenaars, K.H., Hogers, R.C., van Enckevort, L.J., Janssen, A., van Orsouw, N.J., and van Eijk, M.J. 2012. Sequence-based genotyping for marker discovery and co-dominant scoring in germplasm and populations. PLoS One, 7: 1-9.

Tsunewaki, K. 1968. Origin and phylogenetic differentiation of common wheat revealed by conparative gene analysis, pp. 71–85 in Proceedings of the 3rd International Wheat Genetics Symposium, edited by K. W. Finley and K. W. Sheperd. Canberra, Australia.

Van Tassell, C.P., Smith, T.P., Matukumalli, L.K., Taylor, J.F., Schnabel, R.D., Lawley, C.T., Haudenschild, C.D., Moore, S.S., Warren, W.C., and Sonstegard, T.S. 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nature methods, 5: 247-252.

VanRaden, P.M. 2008. Efficient Methods to Compute Genomic Predictions. Journal of Dairy Science 91: 4414-4423.

Varshney, R.K., Graner, A., and Sorrells, M.E. 2005. Genic microsatellite markers in plants: features and applications. TRENDS in Biotechnology, 23: 48-55.

Vos, P., Hogers, R., Bleeker, M., Reijans, M., Van de Lee, T., Hornes, M., Friters, A., Pot, J., Paleman, J., Kuiper, M., and Zabeau, M. 1995. AFLP: a new technique for DNA fingerprinting. Nucleic acids research, 23: 4407-4414.

Wang, S., Wong, D., Forrest, K., Allen, A., Chao, S., Huang, B.E., Maccaferri, M., Salvi, S., Milner, S.G., Cattivelli, L., and Mastrangelo, A.M. 2014. Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array. Plant biotechnology journal, 12: 787-796.

Welsh, J.R., Keim, D.L., Pirasteh, B., and Richards, R.D. 1973. Genetic control of photoperiod response in wheat. In Proceedings of the 4th International Wheat Genetics Symposium (pp. 879-884). Columbia, Missouri.

Whittaker, J.C., Thompson, R., and Denham, M.C. 2000. Marker-assisted selection using ridge regression. Genetical research, 75: 249-252.

Williams, J.G., Kubelik, A.R., Livak, K.J., Rafalski, J.A., and Tingey, S.V. 1990. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic acids research, 18: 6531-6535.

Wise, T.A. 2013. Can we feed the world in 2050? A scoping paper to assess the evidence (Working Paper No. 13-04). Tufts University, Global Development and Environment Institute.

Worland, A.J., Sayers, E.J., and Börner, A. 1994. The genetics and breeding potential of *Rht12*, a dominant dwarfing gene in wheat. Plant Breeding, 113: 187-196.

Worland, T., and J. Snape, 2001 Genetic basis of worldwide wheat varietal improvement, in The World Wheat Book: A History of Wheat Breeding, edited by A. P. Bonjean and W. J. Angus. The World Wheat Book, Lavoisier Publishing Inc., Paris, France.

Wright, S. 1951. The genetical structure of populations. Ann. Eugen. 15: 323-354.

Xu, Y. 2010. Molecular plant breeding. CAB International, Wallingford.

Yan, L., Fu, D., Li, C., Blechl, A., Tranquilli, G., Bonafede, M., Sanchez, A., Valarik, M., and Dubcovsky, J. 2006. The wheat and barley vernalization gene VRN3 is an orthologue of FT. Proc Natl Acad Sci USA 103: 19581-19586.

Yan, L., Loukoianov, A., Blechl, A., Tranquilli, G., Ramakrishna, W., SanMiguel, P., Bennetzen, J.L., Echenique, V., and Dubcovsky, J. 2004. The wheat VRN2 gene is a flowering repressor down-regulated by vernalization. Science, 303: 1640-1644.

Yang, H., Li, C., Lam, H.M., Clements, J., Yan, G., and Zhao, S. 2015. Sequencing consolidates molecular markers with plant breeding practice. Theoretical and Applied Genetics, 128: 779-795.

Yu, J., Holland, J.B., McMullen, M.D., and Buckler, E. S. 2008. Genetic design and statistical power of nested association mapping in maize. Genetics, 178: 539-551.

Yu, J., Pressoir, G., Briggs, W.H., Bi, I. V., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Holland, J.B., and Kresovich, S. 2006. A unified mixed-

model method for association mapping that accounts for multiple levels of relatedness. Nature genetics, 38: 203-208.

Yu, L.X., Lorenz, A., Rutkoski, J., Singh, R.P., Bhavani, S., Huerta-Espino, J., and Sorrells, M.E. 2011. Association mapping and gene–gene interaction for stem rust resistance in CIMMYT spring wheat germplasm. Theoretical and Applied Genetics, 23: 1257-1268.

Zane, L., Bargelloni, L., and Patarnello, T. 2002. Strategies for microsatellite isolation: a review. Molecular ecology, 11: 1-16.

Zanke, C.D., Ling, J., Plieske, J., Kollers, S., Ebmeyer, E., Korzun, V., Argillier, O., Stiewe, G., Hinze, M., Neumann, F., and Eichhorn, A. 2015. Analysis of main effect QTL for thousand grain weight in European winter wheat (*Triticum aestivum* L.) by genome-wide association mapping. Frontiers in plant science, 6: 1-14.

Zhang, D., Bowden, R.L., Yu, J., Carver, B.F., and Bai, G. 2014. Association analysis of stem rust resistance in US winter wheat. PLoS One, 9: 1-10.

Zhang, X., Yang, S., Zhou, Y., He, Z., and Xia, X. 2006. Distribution of the *Rht-B1b*, *Rht-D1b* and *Rht8* reduced height genes in autumn sown Chinese wheats detected by molecular markers. Euphytica, 152:109-116.

Zhang, Z., Ersoz, E., Lai, C.Q., Todhunter, R.J., Tiwari, H.K., Gore, M.A., Bradbury, P.J., Yu, J., Arnett, D.K., Ordovas, J.M. and Buckler, E.S. 2010. Mixed linear model approach adapted for genome-wide association studies. Nature genetics, 42: 355-360.

Zhong, S., Dekkers, J.C., Fernando, R.L., and Jannink, J.L. 2009. Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study. Genetics, 182: 355-364.

Zhu, C., Gore, M., Buckler, E.S., and Yu, J. 2008. Status and prospects of association mapping in plants. The plant genome, 1: 5-20.

CHAPTER 2. Training population selection and use of fixed covariates to optimize genomic predictions in a historical Southeastern USA winter wheat panel

As prepared for submission to the journal Theoretical and Applied Genetics.

Embargoed until publication of wheat reference genome

J. Martin Sarinelli, J. Paul Murphy, Priyanka Tyagi, James B. Holland, Jerry W. Johnson, Mohamed Mergoum, Richard E. Mason, Ali Babar, Stephen Harrison, Russell Sutton, Carl A. Griffey and Gina Brown-Guedira

J.M. Sarinelli, J.P. Murphy and P. Tyagi, Dep. Of Crop Science and Soil Sciences, North Carolina State University, Raleigh, NC 27695; J.B. Holland and G. Brown-Guedira, Department of Crop and Soil Sciences and USDA-ARS, North Carolina State University, Raleigh, NC 27695; J.W. Johnson and M. Mergoum, Department of Crop and Soil Sciences, University of Georgia, Athens, GA 30602; R.E. Mason, Department of Crop Soil and Environmental Sciences, University of Arkansas, Fayetteville, AR 72701; S. Harrison, Department of Agronomy, Louisiana State University, Baton Rouge, LA 70803; R. Sutton, AgriLife Research, Texas A&M University, College Station, TX 77843; C.A. Griffey, Department of Crop and Soil Environmental Sciences, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061; A. Babar, Agronomy Department, University of Florida, Gainesville, FL 32611.

Keywords

Wheat, Genomic Selection, Optimization, Covariates, Historic Data

Conflict of Interest

The authors declare that they have no conflict of interest.

Author Contribution Statement

JMS: Collected phenotypic and genotypic data, data analysis and wrote the manuscript.

GBG, JPM, JBH: provided guidance during the study, manuscript edits and review.

PT: Processed genotypic data.

JH, MM, CG, EM, SH, RS, JPM: developed the germplasm and phenotypic data collection.

Acknowledgments

Personnel from USDA small grains lab for genotypic data collection.

Funding

Funding for this research was provided by the United States Department of Agriculture National Institute of Food and Agriculture (nifa.usda.gov) Triticeae Coordinated Agricultural Project (grant No. 2011-68002-30029) to Gina Brown-Guedira and a Monsanto Graduate Fellowship supporting the graduate student.

Key Message

The optimization of training populations and the use of diagnostic markers as fixed covariates increases the predictive ability of genomic prediction models in a Southern USA cooperative wheat breeding panel.

Abstract

The use of historical data from unbalanced breeding nurseries as training populations offers the possibility to integrate genomic predictions into the existing pipeline of breeding programs. We used cross-validation to evaluate the predictive ability of genomic predictions in a set of 467 genotypes from the Gulf Atlantic Wheat Nursery (GAWN) evaluated from 2008 to 2016. We evaluated the impact on predictive ability of different training population sizes and selection methods (Random, Clustering, PEVmean and PEVmean1) and the effect of inclusion of major QTL for heading date, plant height and powdery mildew resistance as fixed covariates. Increases in predictive ability as the size of the training population increased were more evident for the Random and Clustering selection methods and reached a maximum at the largest population size of 350 individuals. Selection methods based on minimization of the prediction error variance outperformed the other methods evaluated across all the population sizes with an optimum of 200 to 250 individuals. Major gene covariates always improved prediction accuracy, with the greatest increases in performance from the use of multiple covariates in combination. Maximum prediction abilities after optimization of the training population were 0.64, 0.56, 0.71, 0.73, 0.60 for grain yield, test weight, heading date, plant height and powdery mildew resistance, respectively. Our results demonstrate the potential value of combining these unbalanced sets of phenotypic records with genome wide marker data for prediction of untested genotypes in collaborative breeding efforts in soft red winter wheat.

Introduction

Phenotypic data generated from historical plant breeding nurseries offer the possibility to integrate these sets for use as genomic selection training populations to increase genetic gain by reducing breeding cycle times (Hayes et al. 2009, Dawson et al. 2013, Crossa et al. 2013). Public winter wheat breeding programs in the United States have a long tradition of germplasm exchange and collaborative testing of advanced experimental lines prior to cultivar release. Breeders enter lines into various collaborative testing nurseries grown in subsets of environments throughout different wheat regions within the USA. For example, in the eastern and southern winter wheat growing region, the annual Gulf Atlantic Wheat Nursery (GAWN) is a cooperative evaluation of advanced generation breeding lines from public breeding programs in Arkansas, Florida, Georgia, Louisiana, North Carolina, South Carolina, Texas and Virginia. Phenotypic data are collected on a variety of traits evaluated in diverse germplasm and environments. GAWN entries are predominantly F8 or later generations in the pipeline of cultivar development and few lines are entered into the nursery in more than one year. A set of common checks provides a connection between location-year combinations.

The prediction of Genomic Estimated Breeding Values (GEBVs) for genotypes with no available phenotypic data is central to genomic selection approaches. GEBVs are estimated from prediction models that relate trait values to genome-wide marker profiles using training populations containing both phenotypic and genotypic information on a set of individuals (Meuwissen et al. 2001). Model accuracy is measured as the Pearson correlation between GEBVs and true breeding value of each individual (TBV). However, because the

TBV is usually unknown, predictive ability is measured as the correlation between GEBVs and the phenotypic estimated breeding value (EBV) instead (Heffner et al. 2009, Jannink et al. 2010, Lorenz et al. 2011).

Recent studies have examined the predictive ability of genomic selection models using historical datasets from multi-environment nurseries. Storlie and Charmet (2013) reported genomic selection predictive abilities between 0.2 and 0.5 for grain yield using an unbalanced historical set of elite winter wheat lines evaluated over 11 years at different locations in France. Rutkoski et al. (2012, 2015) showed the potential value of historical unbalanced data to predict GEBVs for resistance to wheat stem rust (caused by *Puccinia graminis* f. sp. *tritici*) with predictive abilities between 0.1 and 0.4. Additionally, several studies have examined factors concerning genomic selection approaches aimed at increasing predictive ability and rate of genetic gain. Factors examined include: size of the training population (Arruda et al. 2015, Cericola et al. 2017), population structure (Rincent et al. 2012, Akdemir et al. 2015, Isidro et al. 2015), degree of relationship between training population and validation set (Lorenz et al. 2012, Akdemir et al. 2015, Isidro et al. 2015) and addition of fixed covariates in the genomic selection model (Bernardo 2014, Hoffstetter et al. 2016). However, the optimal approach to the use of historical unbalanced data for genomic selection in commercial breeding programs is uncertain.

Of the statistical methods available to estimate GEBVs of individuals, one of the most utilized in genomic selection is ridge regression best linear unbiased prediction (RR-BLUP), a penalized regression method where all molecular marker effects are estimated from the training population and then used to predict GEBVs for individuals for which only genotypic

data are available. RR-BLUP assumes that markers are random effects with common variance and all markers are equally shrunk toward zero by the same scaling factor (Meuwissen et al. 2001, Piepho 2009, Lorenz et al. 2011, de los Campos et al. 2013). An alternative method, genomic best linear unbiased prediction (G-BLUP), equivalent to RR-BLUP, also has been used widely to estimate GEBVs. In G-BLUP each genotype is considered a random effect in the model and a genomic relationship matrix is utilized to estimate variance-covariance relationships between individuals based on all markers (VanRaden 2008).

Although genomic selection procedures estimate the effect of many loci with small effects simultaneously, variability for major effect QTL influencing agronomic traits is present in the southeastern USA wheat germplasm pool. The dwarfing alleles *RhtB1b* and *RhtD1b* that are major determinants of plant height are each present at relatively high frequencies in eastern winter wheat germplasm (Guedira et al. 2010). Variation has also been reported for the homeologous VERNALIZATION1 genes *Vrn-A1* and *Vrn-B1* as well as the homeologous PHOTOPERIOD1 genes *Ppd-A1*, *Ppd-B1* and *Ppd-D1* that are important loci controlling heading date in winter wheat (Guedira et al. 2014, 2016). Each year the USDA-ARS Eastern Regional Small Grains Genotyping evaluates entries in collaborative wheat breeding nurseries with marker assays that are predictive for alleles of these major plant height and heading data genes, as well as assays for genes conferring resistance to disease and affecting end-use quality (https://triticeaetoolbox.org/wheat/). In addition, the regional lab and individual breeding programs assay these markers on thousands of experimental lines

at different stages in the breeding pipeline. As a result, genetic information about the presence of major genes affecting agronomic traits is often available to the breeder.

Markers linked to major QTL associated with important agronomic traits may also be detected using the complete panel of genotypes and phenotypes available through association analysis. One approach to account for the effect of major genes in genomic selection is to consider the effect of the markers associated with major QTL as fixed while maintaining all other marker effects as random. Bernardo (2014) used simulation to examine the effect of modelling major QTLs modeled as fixed covariates on genomic selection response. This study suggested that QTLs explaining more than 10 percent of the variation associated with the trait should be included in prediction models as fixed covariates.

Chromosomal translocations involving related genomes segregate in SRWW (Olson et al. 2010) and are associated with subpopulation differentiation (Benson et al. 2012). Properly accounting for population structure and genetic relatedness within a population can increase prediction accuracy of genomic selection models (Crossa et al. 2014, Gou et al. 2014). In an empirical study using different populations of rice (*Oryza sativa*) and wheat (*Triticum aestivum*), Isidro et al. (2015) demonstrated that that the best method for training population selection depended on the genetic architecture of the trait and the level of population structure. Akdemir et al. (2015) developed an algorithm for efficient training population selection based on the minimization of the prediction error variance of individuals included in the validation set, demonstrating that this method outperforms random training population selection.

The objectives of this research were to evaluate the effects of different training population sizes and different optimization strategies for training population design considering population structure. In addition, we evaluated the effect of modelling known major genes for plant height and heading date and newly identified markers affecting resistance to powdery mildew as fixed effects in the models. Overall, our goal was to determine how these factors affect the predictive ability of the genomic selection model in an unbalanced historical set of winter wheat germplasm.

Materials and Methods

Plant material

A set of 483 soft red winter wheat elite lines in the F8 or later generations plus nine cultivars serving as checks were evaluated in field environments from 2008 to 2016. The experiments were part of the annual GAWN uniform cooperative testing program of elite germplasm from the breeding programs at Virginia Polytechnic Institute and State University (VA), North Carolina State University (NC), Clemson University (SC), The University of Georgia (GA), The University of Florida (FL), Louisiana State University (LA), The University of Arkansas (AR), and Texas A&M AgriLife Research (TX). The data set was unbalanced; the number of entries per year varied between 44 and 82 and the level of participation of each program varied across the historical series (Table 2.1). Besides the check cultivars, the number of entries evaluated in more than one year was 36.

Fifty five percent of elite lines came from bi-parental populations, and 40 percent from three-parent or backcross populations. The remaining five percent came from more

complex-cross populations. The breeding method utilized prior to entry in the GAWN cooperative test was bulk-pedigree. Depending on the program, the F2 to F5 generations were advanced by families in bulk. Likewise, depending on the program, spikes were selected in the F2 to F5 generations for entry into a pedigree selection protocol in head rows. Lines were evaluated for at least three seasons by the originating breeding program in their home state prior to entry into the GAWN.

Phenotypic data collection and analyses

The nursery was evaluated at one location in up to seven states per year from 2008 to 2016: Arkansas (Stuttgart or Marianna), Florida (Citra or Quincy), Georgia (Plains), Louisiana (Winnsboro), North Carolina (Kinston), Texas (Farmersville) and Virginia (Warsaw). Experimental designs at each environment were randomized complete block designs with two to three replications, although data for some traits were recorded on only one replication. Plot size varied but were typical of yield trial plots for wheat in the region with a minimum of 1.3 meters wide and 3.1 meters long. Data were obtained from four to seven locations annually, with an average of 5.9 locations per year over the nine seasons of the historical series. Data were available for grain yield measured in Mg ha^{-1}, and test weight, expressed as kg m^{-3}. Heading date was recorded as day of year when 50 percent of the plants in a plot were at Zadoks growth stage 59 (Zadoks et al. 1974). Plant height was recorded as the distance in cm from soil level to the tip of the average head, excluding awns. Powdery mildew resistance (caused by *Blumeria graminis* f. sp. *tritici* syn. *Erysiphe graminis*) was recorded using a 0 to 9 scale, incorporating both height and intensity of conidia in the canopy. A value of 0 indicated complete absence of conidia in the canopy and a value of 9

indicated conidia throughout the canopy and on the flag leaf. Out of a total of 63 potential location-year combinations, data for grain yield, test weight, heading date, plant height and powdery mildew resistance were collected at 49, 49, 53, 43 and 18 environments, respectively.

During the 2015-2016 growing season, an experiment was grown to collect data for high heritability traits in a common environment. A total of 391 lines for which adequate seed was available were grown in Raleigh, North Carolina in a randomized complete block design experiment with two replicates. Experimental units were 1-meter rows spaced 30 cm apart. The traits recorded were heading date, plant height and reaction to a natural epidemic of powdery mildew. These data were included in the analysis to determine if addition of measurements from a common environment would improve prediction accuracy.

The total number of data points available for analysis were 7028, 5301, 4861, 4780 and 2246 for grain yield, test weight, heading date, plant height and powdery mildew resistance, respectively. The following linear mixed model was utilized for the analysis of grain yield, test weight, heading date and plant height:

$$y_{ijkl} = \mu + Y_i + L_j + B(YL)_{jk} + YL_{ij} + G_l + YG_{il} + LG_{jl} + YGL_{ijl} + \varepsilon_{ijkl}$$

where, y_{ijkl} was the phenotypic observation of genotype *l* in the i^{th} year in the j^{th} location in the k^{th} block, μ was the overall mean, Y_i was the year effect, L_j was the location effect, $B(YL)_{ijk}$ was the block effect nested within year and location, G_l was the genotypic effect, YL_{ij}, YG_{il}, LG_{jl}, YLG_{ijl} were the interaction terms representing year by location, genotype by year, genotype by location and genotype by year by location, respectively and $\varepsilon_{ijkl,}$

represented the residual term. For this model, the overall mean and the genotypic effect were considered fixed and all the remaining terms random. Random effects (u) and residuals (e) were assumed to be normally and independently distributed u~IDD (0, Iσ_u^2) and e~IDD (0, Iσ_e^2) with the error variance component allowed to be heterogeneous across environments (year-location combination) for grain yield and uniform for the other four traits considered. A simplified version of the statistical model was used for powdery mildew, in which the effects of year and location were combined into a single term called environment due to the low amount of data to estimate all parameters in the full model. Best linear unbiased estimates (BLUE) of each genotype were calculated as the estimated genotypic effect plus overall mean.

The models were implemented using ASReml-R (Butler et al. 2009). Estimates of broad sense heritability on a plot basis for each trait were calculated using a statistical model identical to the previously described with the overall mean as fixed effect and all other terms random. The residual variance was uniform for all traits considered. The broad sense heritability on a per plot basis of each trait was computed according to Holland et al. (2003) as:

$$H^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_{gy}^2 + \sigma_{gl}^2 + \sigma_{gly}^2 + \sigma_e^2}$$

where σ_g^2 represent the genotypic variance, σ_{gy}^2, σ_{gl}^2, σ_{gyl}^2, are the variances component due to the genotype by year, genotype by location and genotypes by location by year, respectively, and σ_e^2 is the variance associated with the residual term.

Genotypic Data

Genotyping by sequencing (GBS, Elshire et al. 2011) using the protocol described by Poland et al. (2012) was conducted for 467 of the 492 lines that were phenotyped, including elite genotypes and commercial cultivars. Seeds of 25 of the older lines were no longer available. Briefly, DNA was extracted from tissue collected from 10 days old plants using DNEasy 96 Plant Kits (Qiagen, Venlo, Netherlands). Genome complexity was reduced using a combination of two enzymes, *MspI* (CCGG) a common cutter and *PstI* (CTGCAG) a rare cutter, and barcoded adaptors were ligated to each sample. Ninety-six individual samples were pooled into a single library and polymerase chain reaction amplified. Each pooled library was sequenced on an Illumina HiSeq 2500. SNP calling on raw sequence data was done with Tassel5GBSv2 pipeline (Glaubitz et al., 2014) using aligner method of bwa version 0.7.12 (Li and Durbin 2009) for aligning SNPs to reference sequence. The International Wheat Genome Sequencing Consortium (IWGSC) genome assembly v0.4 was used as a reference genome to align the SNP with a physical position (https://wheat-urgi.versailles.inra.fr/Seq-Repository/Assemblies). A total of 630,543 polymorphic SNP were identified and saved in a file in hapmap format.

A series of filters for minor allele frequency, marker heterozygosity, and proportion of missing data for markers were applied to increase the quality of the marker set. Only SNP with ≤ 50 percent missing data, ≥ 5 percent minor allele frequency and ≤ 10 percent of heterozygous calls per marker locus were retained, yielding 63,693 SNP that were imputed using the Beagle4 function in R package Synbreed (Wimmer et al. 2012, Browning and Browning 2016). After imputation, redundant SNPs were binned based on SNP pair distance

less than 64 bp and with ≥ 10 percent of heterozygosity for downstream analysis. The final number of SNP utilized for analyses was 34,095.

All lines were evaluated with KASP assays diagnostic for genes having major effects on plant height and heading date that are routinely used for characterization of lines in southeastern breeding programs. These included assays associated with plant height (*Rht-B1* and *Rht-D1*), vernalization requirement duration (*Vrn-A1* and *Vrn-B1*), and photoperiod (*Ppd-A1*, *Ppd-B1*, *Ppd-D1*). In addition, assays that detect alien translocations including the *t1AL:1RS* and *t1BL:1RS* translocations from rye (*Secale cereale* L.), the *t2BS:2GS·2GL:2BL* translocation from *Triticum timopheevii* and the translocation *t2AS:2NS* from *Triticum ventricosum* were screened (S. Table 2.1). These data are available for download at the T3 database (https://triticeaetoolbox.org/wheat/).

Genomic information was organized in a matrix where data for lines was organized in rows and marker scores for the 34,095 SNP from GBS plus the KASPs assays were arranged in columns. Principal Components Analysis of this matrix was implemented with the prcomp function in R version 3.3.1 (R Core Team 2016).

Training Population and Validation Set

The predictive ability of genomic selection models for the five traits was measured as mean Pearson correlations between BLUE and GEBV across 50 replicates of cross validation. For cross validation, we selected 50 different validation sets of size 60 (the average number of unique entries evaluated each year) as a random sample of genotypes from the 467 wheat lines with phenotypic and genotypic data. The same 50 validation sets

were consistently utilized throughout the experiment to ensure accurate comparisons between different genomic selection methods.

The effect of training population size on the predictive ability of the genomic selection models was assessed using seven different population sizes (50, 100, 150, 200, 250, 300 and 350). For each population size, we implemented four approaches for optimization of the training population design using historical data:

Random: Random training population selection was utilized as the reference method to compare with the other three methods of training population optimization. For this method, a random sample of individuals was selected as training population for each validation set according to the size of the training population.

Clustering: We identified population structure associated with the translocation *t2BS:2GS·2GL:2BL*. Thus, we attempted to optimize the training population design by assigning the same proportion of individuals with and without the translocation in the training population and in each validation set. For each validation set randomly selected from the 467 genotypes, the proportion of individuals with the translocation was estimated. The remaining individuals (potential training population candidates) were split based on presence or absence of the translocation, and sample sizes for each group were chosen so that the translocation group frequencies were the same between training and validation sets. A random sample of the individuals required in each translocation set was selected and merged to form the training population. The same procedure was utilized across the 50 different validation set and for each training population size utilized.

PEVmean: This approach utilized a training population optimization algorithm for each validation set that minimized the mean prediction error variance (Rincent et al. 2012; Akdemir et al. 2015). The PEVmean algorithm used genomic information from all genotypes to measure the reliability of the GEBVs for individuals in the validation set. An optimal training population from all genotypes available was selected to minimize the mean prediction error variance in the validation set. The estimation of the prediction error variance requires the inversion of the genomic relationship matrix estimated from genomic information of the individuals in the data set, which is computationally intensive, especially since the optimization process has to be repeated several times with larger sample sizes. Thus, we used the approach suggested by Akdemir et al. (2015) for an efficient approximation to the prediction error variance using the first few principal components of the genotypes to estimate the genomic relationship matrix. The PEVmean strategy was implemented using the function GenAlgForSubsetSelection from the R package STPGA (Akdemir 2016). Principal components were estimated as previously described for population structure and the first 100 principal components were chosen for error variance estimation. The best training population for each of the 50 different validation sets for each of the seven different population sizes was selected after 300 iterations, while other parameters in the function were set with default values.

PEVmean1: This was a modification of PEVmean whereby instead of estimating an optimal training population for each of the 50 validation sets, we estimated the optimum training population for each individual within a validation set replication that minimize the prediction error variance. This procedure was replicated for each individual genotype in the

validation set replication with the goal of identifying the best training population to predict each genotype. The procedure was repeated for each validation set and for the different population sizes utilized in this study.

Genomic selection and association analysis

The RR-BLUP model described by (Meuwissen et al. 2001; Piepho 2009) was used to estimate GEBV. The model was specified as follows:

$$y = X\beta + Zu + e$$

where *y* was a *n x 1* vector of BLUEs for each wheat genotype obtained for one trait, β was a *p x 1* vector of fixed effects which included the overall mean and fixed covariates (major gene and association mapping markers), $\boldsymbol{X}$ and $\boldsymbol{Z}$ were the design matrices for fixed and random effects, with dimensions *n x p* and *n x m*, respectively, *u* was a *m x 1* vector of random marker effects and *e* was an *n x 1* vector representing the residual terms. The variance-covariance structure associated with the random term was u ~ N (0, Iσ_u^2) and for the residual term was e ~ N (0, Iσ_e^2). The estimates of *u* were obtained from the mixed.solve function using the package rrBLUP in R (Endelman 2011).

The impact of incorporating markers previously associated with one or more major effect QTL as fixed effects in the genomic selection model was measured as the change in predictive ability of models with and without covariates. Markers predictive of the dwarfing alleles *Rht-D1b* and *Rht-B1b* were used as covariates in models to predict plant height. Polymorphisms in exon 4 of *Vrn-A1* (Diaz et al. 2012, Yan et al. 2015, Guedira et al. 2016) and the first intron of *Vrn-B1* (Guedira et al. 2014, 2016) associated with differences in

duration of vernalization duration requirement and heading date in winter wheat were included as covariates in models to predict heading date. In addition, an assay predictive of the insensitive allele *Ppd-D1a* of the major photoperiod locus on chromosome 2D (Beales et al. 2007) was included in the models to predict heading date.

Association analysis using a mixed linear model implemented in the R package GAPIT (Lipka et al. 2012; Tang et al. 2016) was performed to identify significant markers for grain yield, test weight and powdery mildew resistance. To avoid bias in the calculation of predictive ability for the genomic selection model, the identification of markers to use as covariates was based on genome-wide association study (GWAS) specific to each training set of lines only. Each GWAS accounted for population structure and degree of relatedness between individuals in the population to avoid spurious associations (Yu et al. 2006). Markers were declared significant based on the Bonferroni corrected p-value at $\alpha = 0.1$. The most significant GBS SNP marker associated with resistance to powdery mildew was selected as fixed covariate to be incorporated in the genomic selection model for powdery mildew reaction. No markers were identified as significantly associated with grain yield or test weight. The impact of incorporating markers as fixed effects in the genomic selection models for plant height, heading data and powdery mildew reaction was measured using the Random and PEVmean training population selection methods in combination with each of the seven training set sizes previously described.

Pairwise comparisons of predictive ability for different training population optimization methods and for the addition of fixed covariates in genomic selection models were performed across the different training population sizes and traits considered in the

analysis. The effect of different prediction methods was tested using a one-way ANOVA using the model:

$$r = \mu + Treatment + error$$

where *r* is the vector of predictive abilities, μ is the overall mean predictive ability, *Treatment* is the fixed effect of different training population optimization methods or the different models including fixed covariates and *error* represents the residual term in the model.

Results

Phenotypic summary

Nine years of phenotypic data for grain yield, test weight, heading date, plant height and powdery mildew resistance from the historical GAWN data set were curated and analyzed (Table 2.2). The effect of genotype was significant for all traits analyzed. Genotype by environment interactions were significant for all traits except plant height, for which genotype by year and genotype by location interactions were not significant. Broad sense heritability on a per plot basis was low (0.14) for grain yield and intermediate for test weight (0.28). Heritabilities for heading date, plant height and powdery mildew resistance were higher (0.49-0.57). Line means for grain yield ranged from 0.42 to 8.27, with a mean of 4.24 Mg ha^{-1} and a large range of variation was observed for the other four traits considered. These data indicated that the GAWN data set represented a diverse set of germplasm evaluated over a diverse set of locations and years.

Genotypic data and population structure

In the panel of 467 genotypes analyzed, 34,095 SNP from GBS and 12 KASP markers were retained after filtering. The level of polymorphism detected by GBS varied for each genome with 13,190, 16,522 and 4,575 polymorphic SNP detected in the A, B and D genomes, respectively (S. Table 2.2).

Two distinct clusters were observed on the first principal component axis that was related to the presence of the *t2BS:2GS-2GL:2BL* translocation derived from *T. timopheevii* (Figure 2.1). The correlation between the first principal component and the diagnostic KASP based on SNP marker IWA8085 that resides in the translocation was 0.89. Of the 467 individuals in the data set, 145 possessed the translocation and 322 did not. The proportion of genotypes containing the translocation varied by breeding program: AR (26 percent), FL (0 percent), GA (9 percent), LA (13 percent), NC (81 percent), SC (8 percent), TX (8 percent) and VA (32 percent). Although a separation of lines into two clusters based the presence/absence of the alien translocation was observed, the first principal component explained less than 8 percent of the total variation in the dataset based on markers. Within the cluster of lines lacking the translocation, there was some separation of lines developed by programs in the more southern states of GA and SC, from lines developed in the Mid-Atlantic programs in VA and NC. Lines from GA were noticeable for their diversity on both the first and the second principal components.

Genomic selection prediction ability and training population optimization

Predictive ability of genomic selection models varied with training population sizes and optimization methods (Figure 2.2). Mean predictive abilities from cross validation for

grain yield (0.40 to 0.64) and test weight (0.31 to 0.56) were particularly encouraging given the importance of these traits as primary selection goals in wheat breeding. Slightly higher predictive abilities were observed for heading date (0.44 to 0.70) and plant height (0.47 to 0.64). Predictive ability for powdery mildew resistance ranged from 0.36 to 0.57. The percent change in predictive abilities due to the addition of fixed covariates evaluated for heading date, plant height and powdery mildew resistance ranged between -9 to +22, -4 to +22 and +8 to +14, respectively when compared with a model without fixed covariates at different training population sizes. The effect of incorporating Raleigh 2016 as a common environment for measurement of plant height, heading date and powdery mildew resistance for 391 lines increased the mean predictive ability of heading date by 12 percent when using the PEVmean training population selection criteria (Table 2.3). For plant height and powdery mildew resistance, the effect of data from a common environment on the overall mean predictive ability of the models was marginal or null. Based on observed improvements for heading date predictions, the Raleigh 2016 data were retained in subsequent analyses.

Optimization of the training population by methods that minimize the prediction error variance of the individuals in the validation set (PEVmean, PEVmean1) outperformed models that used a purely random approach (Random) or a combination of clustering and random sampling (Clustering, Figure 2.2). This trend was observed for all traits and at all training population sample sizes. However, the magnitude of the difference in predictive ability for different training population selection methods was reduced as the training population size increased. Significant differences between optimization methods were typically found for training population sizes between 50 and 300, while for training

population size 350, the different methods tended to converge. For example, for grain yield, training population selection methods PEVmean and PEVmean1 had accuracies that were better ($p < 0.05$) than Clustering and Random for sample sizes from 50 to 200 genotypes. PEVmean and PEVmean1 continued to perform better with training population sample sizes from 250 to 350, but the difference in accuracy with the other methods was not significant, with the exception of the significant difference between PEVmean versus Random for training population size 250 ($p < 0.05$).

Considering the effect of training population size alone, we found that the model predictive ability generally increased as the training population size increased. In the Random and Clustering methods, the magnitude of increase in accuracy for all traits was greatest between training population size 50 to 250, while increases in the population sizes over 250 individuals did not have a significant impact. On the other hand, an increase in the training population size beyond 150, when using methods based on minimizing the PEV, did not increase predictive ability for grain yield, test weight, plant height and powdery mildew resistance. For heading date, this point was reached at a training population size of 200. It was noteworthy that for grain yield, when using training population selection PEVmean and PEVmean1, there was a high degree of accuracy even when the training population size was 50, and there was no substantial increase in predictive ability for any other training population size. An interesting anomaly was observed for PEVmean whereby the training population size of 100 had a lower predictive ability than training population size 50 for all traits but still outperformed training population selection methods Random and Clustering.

Effect of fixed covariates on accuracies

The investigation into the impact on genomic selection model predictive ability through inclusion of marker covariates as fixed versus random effects utilized two training population selection methods, PEVmean and Random, with all training population sizes. For heading date and plant height, markers in previously identified causal genes were used as covariates. No significant markers were identified for grain yield and test weight after association analyses, so models having covariates were not evaluated for these traits. The most significant marker trait association for powdery mildew resistance varied depending of the independent training set utilized. However, for all training sets, the most significant markers were located in a 12 Mb region on the long arm of chromosome 7A (S. Table 2.3). For these analyses, the most significant marker identified in each independent run was utilized as the fixed covariate for prediction of the respective validation set. Overall, it was observed that including markers associated with major effect genes or QTL as fixed effects increased the prediction accuracy similarly for both TP selection methods when compared with models without covariates (S. Fig. 2.1, S. Fig. 2.2, S. Fig. 2.3). Results based on the random model with and without covariates are shown in Table 2.4 for different TP sizes.

The markers for heading date utilized as covariates in genomic selection models were based on polymorphisms in the *Vrn-A1*, *Vrn-B1* and *Ppd-D1* loci that are known to be important determinants of flowering time in wheat. The frequency of the early flowering allele in the panel for each marker was 0.21, 0.28 and 0.63 for *Vrn-A1*, *Vrn-B1* and *Ppd-D1*, respectively (Table 2.5). The addition of the marker for *Vrn-A1* had the largest impact on accuracy (Table 2.4). Using a combination of markers in the *Vrn-A1*, *Vrn-B1* and *Ppd-D1*

genes, the accuracy of predictions increased with all population sizes and was significantly different from models that considered one marker only. However, the impact of adding fixed covariates to the model decreased as training population size increased. When training populations contained 300 or more individuals, the only models that differ from the no covariate control were those that included multiple covariates. Given a training population size of 50, an average 10 percent increase in model predictive ability was observed when markers in the *Vrn-A1*, *Vrn-B1* and *Ppd-D1* genes were included as covariates in the model, compared with the model having no covariates. The improvement in predictive ability when adding covariates was three percent when the training population size was 350.

Genomic selection models for plant height included genotypes at SNP in the reduced height genes *Rht-B1* and *Rht-D1* indicative of the dwarfing alleles. Most of the lines in the population were semi-dwarf having either *Rht-D1b* (74 percent) or *Rht-B1b* (20 percent). The effect of the *Rht-D1b* allele in the genomic selection models was greater than that observed for *Rht-B1b* (Table 2.5). The genomic selection model that incorporated *Rht-B1* as a fixed covariate did not significantly increase accuracy across different training population sizes, whereas the model with *Rht-D1* alone significantly increased accuracy over all training population sizes (Table 2.4). Furthermore, including both diagnostic markers as fixed covariates in the model simultaneously produced increases in the predictive ability ranging from 9 to 17 percent compared with the models without covariates across the different training population sizes evaluated.

For powdery mildew resistance, we used the most significant SNP marker, detected by association analysis in each training set after masking the phenotypes of the validation set

replication associated to it avoiding biases in the calculation of the fixed covariate in the prediction model. The SNP markers selected as fixed covariates are detailed in S. Table 2.3 and the number of times that the SNP appear as the most significant in the 50 independent set evaluated is 8, 1, 3, 11, 2, 24 and 3 times respectively. The effect on model predictive ability of using the most significant SNP detected by GWAS as a fixed covariate was significant for all training population sizes compared with the effect of no covariates in the model (Table 2.4). Given a training population size of 50, an average 14 percent increase in predictive ability was observed when the most significant SNP was included as fixed covariate in the model, compared with the model having no covariates. The improvement in predictive ability when adding the covariate was 8 percent when the training population size was 350.

Discussion

Questions arise when breeders integrate genomic selection into an ongoing cultivar development program concerning material to be utilized in the training population and how to optimize prediction ability of the model based on the germplasm available. This study is the first report on the utility of historical elite wheat lines from eight southeastern USA public breeding programs evaluated in a cooperative nursery across the southeastern USA for genomic selection. These public programs are currently developing commercial cultivars of soft red winter wheat for the region, thus the study provides valuable empirical results on the use of genomic selection.

Overall, cross validation results from this study were encouraging regarding the use of unbalanced historical data for genomic selection predictions, even for highly polygenic

and complex traits like grain yield. The mean predictive ability for grain yield was 0.64 for a training population size of 350 individuals while using the training population optimization method PEVmean, congruent with Crossa et al. (2010). Moreover, our cross-validation results for grain yield in wheat outperformed those of Storlie and Charmet (2013), who used historical unbalanced data of 318 lines grown over an 11-year period in France and the results presented by Poland et al. (2012) in a panel of 254 wheat lines evaluated in Mexico during 2010.

Grain yield is the main goal for improvement in wheat cultivar development programs. Plant breeders incorporating genomic selection in their pipelines to predict the GEBV of individuals generally assign the cost of genotyping to improvement of grain yield (Poland and Rutkoski 2016). However, the marker information is available to predict other traits including resistance to diseases, grain quality and agronomic traits if phenotypic data are available for the training population. Multi-location data for grain yield, test weight, heading date, plant height and powdery mildew resistance were collected for the GAWN nursery. Average cross validation results for test weight, heading date, plant height and powdery mildew resistance showed moderate to high prediction ability for these traits, which reinforces the potential of the unbalanced GAWN nursery as a training population. Furthermore, we demonstrated that for heading date, the inclusion of one year of inexpensive phenotyping for most lines in a common environment increased the prediction ability by up to eight percent. Although heritability of heading date was moderately high in this study (0.54 on a plot basis), it was nonetheless influenced by the year to year variation in winter and early spring temperatures experienced in southern USA locations. Significant genotype

by environment interactions were observed for heading date (Table 2.2). The addition of data from a common nursery experiment did not result in a similar improvement in predictive ability for plant height, as this trait was less influenced by genotype by environment interactions. Similarly, addition of data from a common environment for reaction to powdery mildew did not result in significant increases in predictive ability. It is possible that the multi-year data provided a broader sampling of the powdery mildew isolates encountered by the training population over time.

The adjustment of training population optimization criteria in the genomic selection model led to an increase in model accuracy. Subpopulations can affect the predictive ability of genomic predictions if not accounted for during model building. Training population selection containing individuals more closely related with the validation set should lead to an increase in the precision of the GEBV estimates. Evaluation of population structure in our data set indicated there were two subpopulations associated with the presence or absence of the *t2BS:2GS 2GL:2BL* translocation derived from *T. timopheevii*. However, when the translocation was utilized as a criteria to optimize the design of the training population (Clustering method), the accuracies were not different from the Random training population selection method for all traits and training population sizes considered. The lack of improvement in the accuracy of the Clustering method compared with Random suggested that population structure associated with the alien translocation did not affect the traits under consideration. Isidro et al. (2015) also did not observe significant increases in model prediction ability when comparing random training population selection with clustering methods based on the origin of the 1127 wheat genotypes to optimize training populations.

In contrast, methods of training population design based on reduction of PEV mean of the validation set (PEVmean and PEVmean1) were more accurate compared with methods that selected individuals at random. This was especially true with small training population sizes, because they better accounted for the relationship between the individuals in the training population and the validation set (Habier et al. 2013). The optimum size of the training population for these methods was about 200 for all five traits considered. Further significant increases in model accuracy with increased population size were not observed, suggesting that not all individuals in the training population need to be utilized to get adequate levels of predictive ability in the model. Similar findings were reported with different populations of wheat (*Triticum aestivum*, Isidro et al. 2015, Rutkoski et al. 2015), rice (*Oryza sativa*, Akdemir et al. 2015, Isidro et al. 2015), arabidopsis (*Arabidopsis thaliana*, Akdemir et al. 2015) and maize (*Zea mays*, Rincent et al. 2012; Akdemir et al. 2015). The PEVmean1 and PEVmean methods were different at training population size 100 where PEVmean had a drop-in model predictive ability. Akdemir et al. (2015) also found a decay or stagnation in model predictive ability for training population of 100 individuals in a highly structured maize population when using PEVmean and Random training population selection when compared with training population of sizes ranging from 50 to 200.

In terms of computational time required to perform the analysis, the methods PEVmean and PEVmean1 were time consuming in comparison with Random and Clustering methods. However, PEVmean1 had the most intensive requirement with an average of 40 hours of computer time to get a full set of training populations (3000) for each training population size analyzed using a desktop computer DELL with 16 GB of RAM and 3.4 GHz.

Given the small improvement in predictive ability between an optimized training population for each individual compared with the PEVmean method that optimized the training population for a validation set having 60 members, utilizing the PEVmean1 method is not recommended due to the time and effort involved in implementing this protocol.

Other genomic selection studies using fixed covariates have been published (Arruda et al. 2016, Hoffstetter et al. 2016). Bian and Holland (2017) conclude that adding SNP associated with the trait as fixed covariates in genomic predictions models yield higher predictive abilities when compared with models that only consider the polygenic background for traits that are not highly polygenic. We used two approaches to avoid the potential bias in selecting markers for inclusion as fixed co-variates. For plant height and heading date, we assayed polymorphisms in major known genes to affect these traits. For powdery mildew, we identified the most significant SNP marker in each training set associated with powdery mildew resistance after masking the phenotypic data of individuals in each of the 50 different validation sets.

The addition of markers as fixed covariates in the genomic selection models was demonstrated to be useful for plant height and powdery mildew resistance across the complete range of training population samples sizes evaluated. These results were in agreement with Bernardo (2014) who pointed out that the accuracy of genomic selection models can be increased by adding major genes as fixed effects when they represent a large proportion of the total variance associated with the trait ($\geq$ 10 percent) under consideration. However, that was not the case for heading date, for which an increase in training population sample size led to a reduction of the influence of the covariate in the model. In this study,

addition of fixed effects for markers associated to known major genes or for significant SNP markers detected through association analysis both led to an increase of the model predictive ability, reinforcing the utility of the addition of fixed covariates in the models when available. The trend of increasing the predictive ability of the model was also observed when a combination of markers was included in the model as fixed covariates. This was observed for different sample sizes and in some cases, the combination of markers outperformed the predictive ability for models with the addition of only one maker. Heading date, plant height and powdery mildew resistance considered in this study are highly heritable traits in comparison with grain yield, where the identification of a major effect QTL is not straightforward due the complex polygenic genetic architecture. For these traits, the effect of adding fixed covariates on model predictive ability was never worse than the model considering all markers as random effects. The maximum response observed was the effect of combining *Rht-D1* and *Rht-B1* for plant height with an average of 11 percent predictive ability increase in comparison with a model without covariates across all training population size considered in the study.

Conclusion

Use of historical unbalanced phenotypic data from cooperatives nurseries from different southeastern USA breeding programs was a reliable and accurate way to incorporate genomic selection predictions into the breeding pipeline, even for a highly polygenic trait like grain yield. The training population optimization algorithm that reduced PEVmean increased the accuracy of predictions for each trait analyzed, particularly for small population

sizes. This was the case even under the presence of population structure in the breeding population driven by the presence of a chromosomal translocation. We demonstrated that the effect of adding diagnostic markers associated with large effect genes or QTL as fixed effects in the model increased the overall model predictive ability for most samples sizes evaluated. Our results have implications for the use of training populations from 50 to 350 individuals. For all traits, minimization of PEV in the validation set and/or the addition of data for markers closely linked to or representing causal polymorphisms in genes affecting the traits had the greatest impact in improvement of accuracy when training population size was between 50 to 150 individuals. In some cases, breeders may have genotyped representatives of the germplasm to be predicted, and would like to target a small population of lines as a training set. In this scenario, training population optimization and targeted phenotyping of a small number of lines for expensive and /or difficult phenotyping could be done. Nonetheless, the observed predictive abilities for all training population selection methods tended to converge as training population sizes increased to 350 individuals. Thus, if the phenotypic records are available, utilizing all lines and incorporating selected markers as fixed covariates can produce models with high accuracies. In an applied breeding pipeline, it is likely that lines from the previous year's nurseries will be added to the training population, increasing the size of the training population over time. However, this may also increase the degree of divergence between genotypes in the training population and the selection set. In this case, optimization of training population could be used to increase the accuracy of the predictions.

References

Akdemir D (2016) STPGA: Selection of Training Populations by Genetic Algorithm. R package version 3.0. https://CRAN.R-project.org/package=STPGA.

Akdemir D, Sanchez JI, Jannink JL (2015) Optimization of genomic selection training populations with a genetic algorithm. Genetics Selection Evolution, 47: 1-10.

Arruda MP, Brown PJ, Lipka AE, Krill AM, Thurber C, Kolb FL (2015) Genomic selection for predicting head blight resistance in a wheat breeding program. The Plant Genome, 8: 1-12.

Arruda MP, Lipka AE, Brown PJ, Krill AM, Thurber C, Brown-Guedira G, Dong Y, Foresman BJ, Kolb FL (2016) Comparing genomic selection and marker-assisted selection for Fusarium head blight resistance in wheat (Triticum aestivum). Mol Breed 36: 1-11.

Beales J, Turner A, Griffiths S, Snape JW, Laurie, D.A. (2007) A pseudo-response regulator is misexpressed in the photoperiod insensitive Ppd-D1a mutant of wheat (Triticum aestivum L.). Theoretical and Applied Genetics, 115: 721-733.

Benson J, Brown-Guedira G, Murphy PJ, Sneller C (2012) Population structure, linkage disequilibrium, and genetic diversity in soft winter wheat enriched for fusarium head blight resistance. The Plant Genome, 5: 71-80.

Bernardo R (2014) Genomewide Selection when Major Genes Are Known. Crop Science 54: 68-75.

Bian Y and Holland JB (2017) Enhancing genomic prediction with genome-wide association studies in multiparental maize populations. Heredity, 118: 585-593.

Browning BL, Browning SR (2016) Genotype imputation with millions of reference samples. The American Journal of Human Genetics, 98: 116-126.

Butler DG, Cullis BR, Gilmour AR, Gogel BJ (2009) ASReml-R reference manual, release 3. Technical report. NSW Department of Primary Industries.

Cericola F, Jahoor A, Orabi J, Andersen JR, Janss LL, Jensen J (2017) Optimizing Training Population Size and Genotyping Strategy for Genomic Prediction Using Association Study Results and Pedigree Information. A Case of Study in Advanced Wheat Breeding Lines. PloS one, 12: 1-20.

Crossa J, Beyene Y, Kassa S, Pérez P, Hickey JM, Chen C, de los Campos G, Burgueño J, Windhausen VS, Buckler E, Jannink JL (2013) Genomic prediction in maize breeding populations with genotyping-by-sequencing.Genes Genomes Genetics, 3: 1903-1926.

Crossa J, Pérez P, Hickey J, Burgueño J, Ornella L, Cerón-Rojas J, Zhang X, Dreisigacker S, Babu R, Li Y, Bonnett D (2014) Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity, 112: 48-60.

Dawson JC, Endelman JB, Heslot N, Crossa J, Poland J, Dreisigacker S, Manès Y, Sorrells ME, Jannink JL (2013) The use of unbalanced historical data for genomic selection in an international wheat breeding program. Field Crops Research, 154: 12-22.

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

Díaz A, Zikhali M, Turner AS, Isaac P, Laurie DA, (2012) Copy number variation affecting the Photoperiod-B1 and Vernalization-A1 genes is associated with altered flowering time in wheat (Triticum aestivum). PLoS One, 7: 1-11.

Ellis M, Spielmeyer W, Gale K, Rebetzke G, Richards R (2002) " Perfect" markers for the Rht-B1b and Rht-D1b dwarfing genes in wheat. Theoretical and Applied Genetics, 105: 1038-1042.

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PloS one, 6: 1-10.

Endelman, JB (2011) Ridge regression and other kernels for genomic selection with R package rrBLUP. The Plant Genome, 4: 250-255.

Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q, Buckler ES (2014) TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. PloS one, 9: 1-11.

Guedira M, Brown-Guedira G, Sanford DV, Sneller C, Souza E, Marshall D (2010) Distribution of genes in modern and historic winter wheat cultivars from the Eastern and Central USA. Crop Science, 50: 1811-1822.

Guedira M, Maloney P, Xiong M, Petersen S, Murphy JP, Marshall D, Johnson J, Harrison S, Brown-Guedira G (2014) Vernalization duration requirement in soft winter wheat is associated with variation at the VRN-B1 locus. Crop Science, 54: 1960-1971.

Guedira M, Xiong M, Hao YF, Johnson J, Harrison S, Marshall D, Brown-Guedira G (2016) Heading Date QTL in Winter Wheat (Triticum aestivum L.) Coincide with Major Developmental Genes VERNALIZATION1 and PHOTOPERIOD1. PloS one, 11: 1-21.

Guo Z, Tucker DM, Basten CJ, Gandhi H, Ersoz E, Guo B, Xu Z, Wang D, Gay G (2014) The impact of population structure on genomic prediction in stratified populations. Theoretical and applied genetics, 127: 749-762.

Habier D, Fernando RL, Garrick DJ (2013) Genomic BLUP decoded: a look into the black box of genomic prediction. Genetics, 194: 597-607.

Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009) Invited review: Genomic selection in dairy cattle: Progress and challenges. Journal of dairy science, 92: 433-443.

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

Helguera M, Khan IA, Kolmer J, Lijavetzky D, Zhong-Qi L, Dubcovsky J (2003). PCR assays for the cluster of rust resistance genes and their use to develop isogenic hard red spring wheat lines. Crop Science, 43: 1839-1847.

Hoffstetter A, Cabrera A, Huang M and Sneller C (2016) Optimizing training population data and validation of genomic selection for economic traits in soft winter wheat.G3: Genes| Genomes| Genetics, 6: 2919-2928.

Holland JB, Nyquist WE, Cervantes-Martinez CT (2003) Estimating and interpreting heritability for plant breeding: an update. Plant Breeding Rev 22:9–112.

Isidro J, Jannink JL, Akdemir D, Poland J, Heslot N, Sorrells ME (2015) Training set optimization under population structure in genomic selection. Theoretical and applied genetics, 128: 145-158.

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

Li H and Durbin R, (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25: 1754-1760.

Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ, Gore MA, Buckler ES, Zhang Z (2012) GAPIT: genome association and prediction integrated tool. Bioinformatics, 28: 2397-2399.

Lorenz AJ, Chao S, Asoro FG, Heffner EL, Hayashi T, Iwata H, Smith KP, Sorrells ME, Jannink JL (2011) Genomic selection in plant breeding: knowledge and prospects. Advances in agronomy, 110: 77-123.

Lorenz AJ, Smith KP, Jannink JL (2012) Potential and optimization of genomic selection for Fusarium head blight resistance in six-row barley. Crop Science, 52: 1609-1621.

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

Milus EA, Lee KD, Brown-Guedira G (2015) Characterization of stripe rust resistance in wheat lines with resistance gene Yr17 and implications for evaluating resistance and virulence. Phytopathology, 105: 1123-1130.

Nishida H, Yoshida T, Kawakami K, Fujita M, Long B, Akashi Y, Laurie DA, Kato K (2013) Structural variation in the 5′ upstream region of photoperiod-insensitive alleles Ppd-A1a and Ppd-B1a identified in hexaploid wheat (*Triticum aestivum* L.), and their effect on heading time. Molecular breeding, 31: 27-37.

Olson EL, Brown-Guedira G, Marshall DS, Jin Y, Mergoum M, Lowe I, Dubcovsky J (2010) Genotyping of US Wheat Germplasm for Presence of Stem Rust Resistance Genes, and. Crop science, 50: 668-675.

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

Poland J, Endelman J, Dawson J, Rutkoski J, Wu S, Manes Y, Dreisigacker S, Crossa J, Sánchez-Villeda H, Sorrells M, Jannink JL (2012) Genomic selection in wheat breeding using genotyping-by-sequencing. The Plant Genome, 5: 103-113.

Poland J, Rutkoski J (2016) Advances and Challenges in Genomic Selection for Disease Resistance. Annual Review of Phytopathology, 54: 79-98.

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

R Core Team (2016) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rincent R, Laloë D, Nicolas S, Altmann T, Brunel D, Revilla P, Rodriguez VM, Moreno-Gonzalez J, Melchinger A, Bauer E, Schön CC (2012) Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (*Zea mays* L.). Genetics, 192: 715-728.

Rutkoski J, Benson J, Jia Y, Brown-Guedira G, Jannink JL, Sorrells ME (2012) Evaluation of genomic prediction methods for Fusarium head blight resistance in wheat. The Plant Genome, 5: 51-61.

Rutkoski J, Singh RP, Huerta-Espino J, Bhavani S, Poland J, Jannink JL, Sorrells ME (2015) Efficient use of historical data for genomic selection: a case study of stem rust resistance in wheat. The Plant Genome, 8: 1-10.

Storlie E, Charmet G (2013) Genomic selection accuracy using historical data generated in a wheat breeding program. The Plant Genome, 6: 1-9.

Tang Y, Liu X, Wang J, Li M, Wang Q, Tian F, Su Z, Pan Y, Liu D, Lipka AE, Buckler ES (2016) GAPIT Version 2: an enhanced integrated tool for genomic association and prediction. The Plant Genome, 9: 1-9.

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

Wimmer V, Albrecht T, Auinger HJ, Schön CC (2012) synbreed: a framework for the analysis of genomic prediction data using R. Bioinformatics, 28: 2086-2087.

Yan L, Li G, Yu M, Fang T, Cao S, Carver BF, (2015) Genetic Mechanisms of Vernalization Requirement Duration in Winter Wheat Cultivars. In Advances in Wheat Genetics: From Genome to Field, 117-125. Springer Japan.

Yu J, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature genetics, 38: 203-208.

Zadoks JC, Chang TT, Konzak CF (1974) A decimal code for the growth stages of cereals. Weed research, 14: 415-421.

Table 2. 1: Test year (Year), entries per cooperating state (Breeding programs) and total numbers of checks and elite advanced line entries in the Gulf-Atlantic Wheat Nursery (GAWN) from 2008 to 2016.

Year	Breeding Programs								Commercial Checks	Total by year
	AR	FL	GA	LA	NC	SC	TX	VA		
2008	12	7	12	12	12	12	0	12	3	82
2009	13	0	12	10	12	12	0	12	3	74
2010	12	1	12	11	12	12	0	12	3	75
2011	0	0	12	12	12	4	0	12	4	56
2012	12	1	12	11	12	12	0	12	4	76
2013	12	1	12	11	12	0	0	12	4	64
2014	5	0	7	6	6	0	0	10	4	38
2015	6	0	6	6	6	0	6	10	4	44
2016	4	0	6	7	9	0	6	10	3	45
Total by program	76	10	91	86	93	52	12	102		

Table 2. 2: Summary phenotypic information for grain yield, test weight, heading date, plant height and powdery mildew resistance. Including number of environment where each trait was evaluated (No. Env), number of data points for the analysis of each trait (No. Data Points), descriptive statistics of each trait including minimum (Min), average (Mean), maximum (Max) and standard deviation (SD). Variance components estimates for random effects: Location (Loc), Year (Year), Year x Location interaction (YL), Replication (Rep(YL)), Genetic (Gen), Genetic x Year interaction (GY), Genetic x Location interaction (GL), Genetic x Year x Location interaction (GYL Var) and residual term in the model. Broad sense heritability on a per plot basis (Heritability) estimated for each trait. Significant model variance component at 0.05 α level labels as (*).

	Traits				
	Grain yield	Test weight	Heading date	Plant height	Powdery mildew
	Mg ha^{-1}	Kg m^{-3}	Days	Centimeters	0-9 scale
No. Env	49	49	54	44	19
No. Data Points	7028	5075	4861	4780	2446
Min	0.42	41.70	63.00	53.34	0.00
Mean	4.24	57.05	105.56	88.34	2.16
Max	8.27	65.80	131.00	137.16	9.00
SD	1.30	6.60	11.91	11.07	2.01
Variance components estimates					
Loc	52.60*	3.12*	37.95*	3.24*	
Year	53.82	0.97	43.26*	2.82*	0.60*
YL	142.09*	5.41*	21.67*	4.05*	
Rep(YL)	3.76*	0.07*	0.08*	0.15*	0.10*
Gen	19.84*	1.07*	14.42*	4.85*	1.66*
GY	6.54*	0.41*	1.26*	0.13	
GL	26.24*	0.32*	5.39*	0.06	0.62*
GYL	50.71*	1.26*	4.12*	1.10*	
Residual	39.65	0.80	1.45	2.42	1.03
Heritability	0.14	0.28	0.54	0.57	0.49

Table 2. 3: Mean prediction ability after 50 cycles of cross validation using two methods (Random and PEVmean) of TP optimization methods for heading date, plant height and powdery mildew resistance were calculated and averaged across the seven different TP sizes to estimate the effect in the genomic selection model of using all phenotypic data available from the historical series vs a model that incorporate phenotypic data from the historical series plus a common environment (Raleigh2016) were 391 genotypes were grown together. Test for statistical significance was performed for each trait. Significant different at $p<0.01$ was label as (***).

	All Locations	Raleigh2016_Removed
	Heading Date	
Random	0.59***	0.53
PEVmean	0.66***	0.58
	Plant Height	
Random	0.57	0.57
PEVmean	0.62	0.62
	Powdery Mildew Resistance	
Random	0.48	0.48
PEVmean	0.53	0.52

Table 2. 4: Mean prediction ability across 50 cycles of cross validation for heading date, plant height and powdery mildew resistance according to genomic selection models that consider the addition of diagnostic markers associated with major effect QTLs as fixed covariates incorporated in the model using different pre-defined training population sizes. For heading date and plant height known diagnostic markers associated with the trait were utilized, while for powdery mildew resistance the most significant SNP detected in each validation cycle was utilized as the fixed covariate. Analysis performed with training population selection method Random. Each trait had a different set of markers related with major QTLs. * Significantly different from the model that did not include covariates at level. 0.1. *** Significantly different from the model that did not include covariates at level 0.05.

	TRAINING POPULATION SIZE						
	TP050	TP100	TP150	TP200	TP250	TP300	TP350
				Heading Date			
No_Covariate	0.46	0.53	0.56	0.61	0.63	0.67	0.68
PpdD1	0.42*	0.54	0.58	0.62	0.64	0.68	0.69
VrnA1	0.51***	0.58***	0.60***	0.63*	0.65*	0.68	0.69
VrnB1	0.49	0.56	0.59	0.63	0.64	0.67	0.68
VrnA1_PpdD1	0.51***	0.59***	0.62***	0.65***	0.67***	0.70***	0.70***
VrnA1_VrnB1_PpdD1	0.56***	0.61***	0.64***	0.67***	0.68***	0.71***	0.71***
				Plant Height			
No_Covariate	0.49	0.53	0.55	0.58	0.60	0.61	0.63
RhtB1	0.47	0.52	0.55	0.58	0.60	0.61	0.62
RhtD1	0.56***	0.60***	0.62***	0.64***	0.66***	0.67***	0.68***
RhtB1_RhtD1	0.59***	0.64***	0.67***	0.69***	0.71***	0.72***	0.73***
				Powdery Mildew Resistance			
No_Covariate	0.36	0.44	0.48	0.51	0.52	0.53	0.55
Most Significant SNP	0.42***	0.50***	0.53***	0.56***	0.56***	0.57***	0.60***

Table 2. 5: Mean allelic effect (Mean effect), allele frequency and SNP position in Megabase pairs of fixed covariates utilized for heading date and plant height in a model that considered all covariates selected for each trait using the Random training population selection method with a training population size of 350 individuals across the 50 validation sets. Frequency are indicated for the dwarfing alleles at *Rht-B1* and *Rht-D1*, and early flowering alleles for *Vrn-A1*, *Vrn-B1* and *Ppd-D1*.

Trait	Marker Name	Favorable Allele	Chromosome	Position (Mbp)	Allele frequency	Mean effect [a]
Plant height	*Rht-B1*	T	4B	30.86	0.20	-3.78
	Rht-D1	T	4D	18.78	0.74	-5.38
Heading date	*Vrn-A1*	C	5A	587.42	0.21	-1.37
	Vrn-B1	C	5B	573.81	0.18	-1.28
	Ppd-D1	Deletion	2D	33.96	0.63	-0.91

[a] Calculated as the average effect of the marker estimated from 50 different training population of size 350 selected at Random utilized to predict GEBVs in each of the 50 validation set for a model that considered all markers simultaneously.

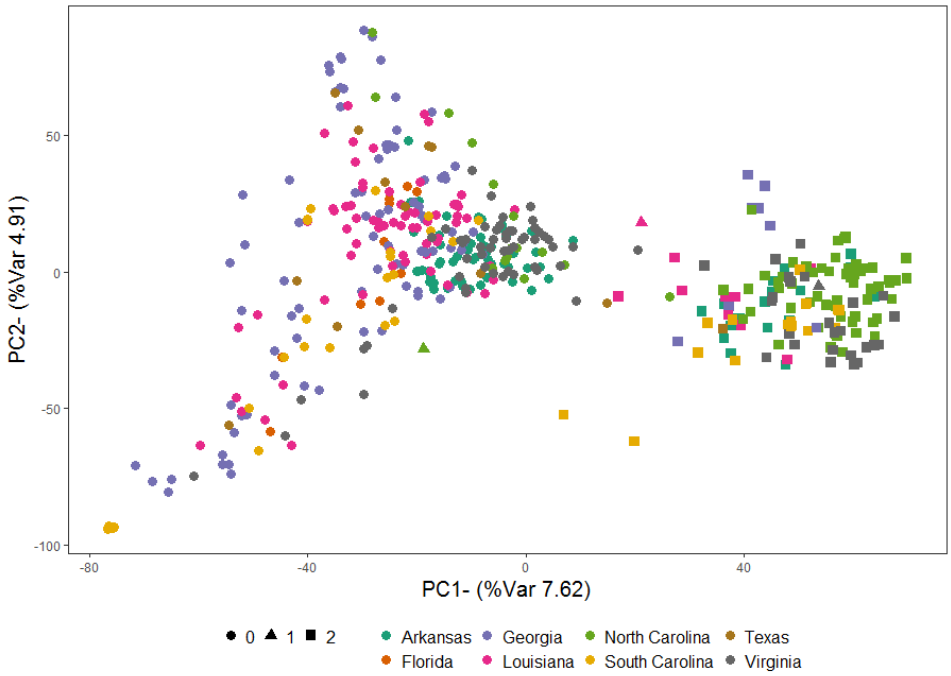


Figure 2. 1: Scatter plot of the first two principal components after analysis of 467 winter wheat lines using 34,107 SNPs. Points are color coded according to the origin of genotypes. Arkansas, University of Arkansas; Florida, University of Florida; Georgia, University of Georgia; Louisiana, Louisiana State University; North Carolina, North Carolina State University; South Carolina, Clemson University; Virginia, Virginia Tech; Texas, Texas A&M AgriLife Research. Different shapes represent the number of copies of the allele of the diagnostic marker *Sr36* linked to the *t2BS:2GS 2GL:2BL* translocation from *T. timopheevii*. Percentages in each axis represents the proportion of variance explained by each principal component.

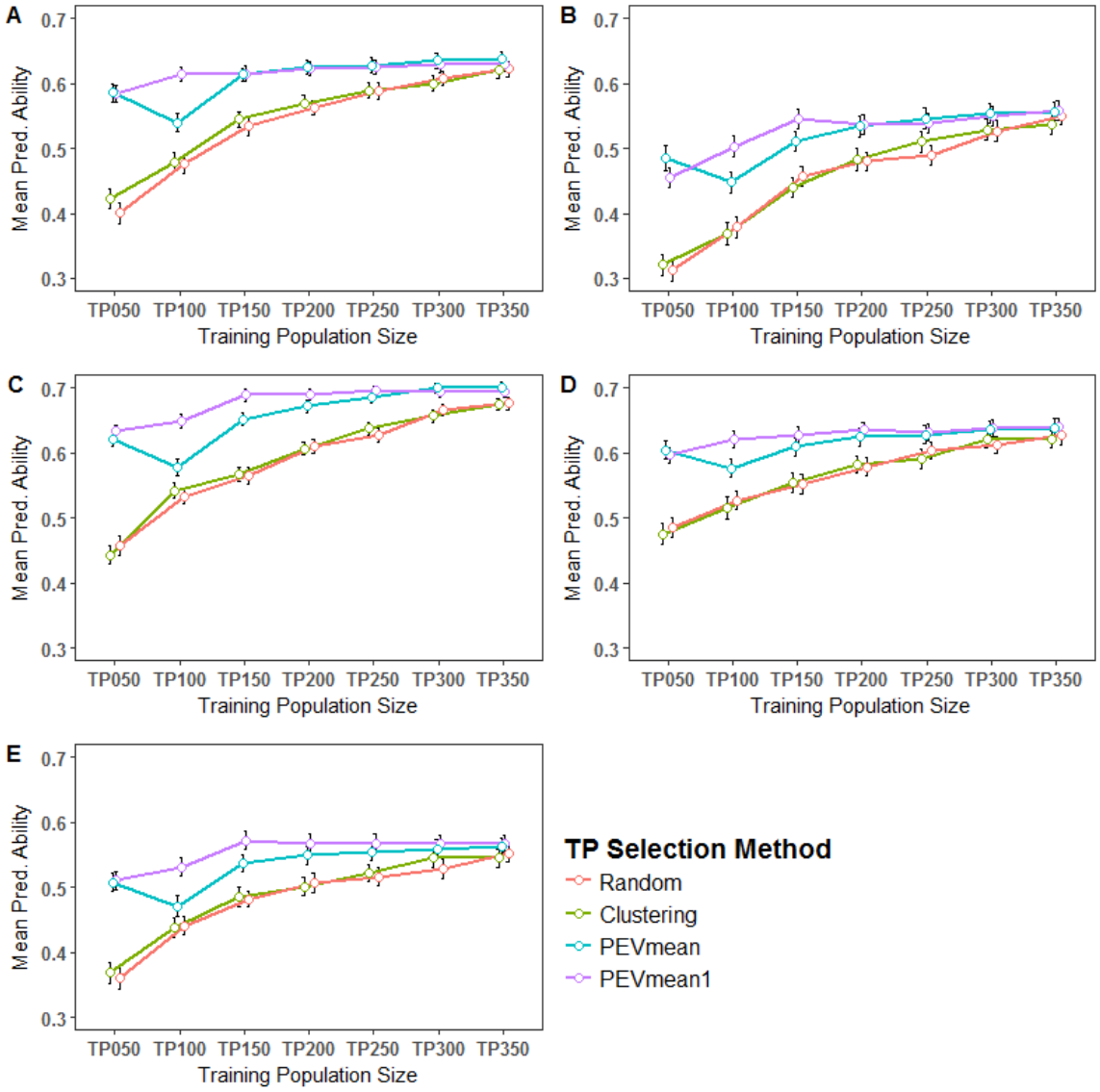


Figure 2. 2: Comparison of mean predictive ability (Mean Pred. Ability) for grain yield (A), test weigh (B), heading date (C), plant height (D) and powdery mildew resistance (E) for four training population optimization methods: Clustering (Weighted proportion of translocation *t2BS:2GS 2GL:2BL* in the training population and validation set), PEVmean (training population selected by minimization of the PEV mean in the validation set), PEVmean1 (training population selected by minimization of the PEV of each individual in the validation set) and Random (random training population selection). All methods were evaluated for seven different training population sizes (50, 100,150, 200, 250, 300 and 350). Error bars represent ± one standard error of the mean.

Supplemental Material for Chapter 2

S. Table 2. 1: Molecular markers utilized for characterization of major genes associated with heading date, plant height and translocations.

Gene	Chromosome	Marker	Allele	Variant	Reference
Ppd-A1	2A	*TaPpdA1_Prodel*	Deletion Insertion	Photoperiod insensitive Photoperiod sensitive	Nishida et al. 2013, Guedira et al. 2016
Ppd-B1	2B	*TaPpdBJ001*	Insertion Deletion	Photoperiod insensitive Photoperiod sensitive	Beales et al. 2007
Ppd-B1	2B	*TaPpdBJ003*	Insertion Delition	Photoperiod insensitive Photoperiod sensitive	Diaz et al. 2012
Ppd-D1a	2D	*TaPpdDD001*	Deletion Insertion	Photoperiod insensitive Photoperiod sensitive	Beales et al. 2007
Vrn-A1	5A	*Vrn-A1_exon4*	T/T C/C	Long vernalization Short vernalization	Diaz et al. 2012
Vrn-B1	5B	*TaVrnB1_1752*	C/C G/G	Short vernalization Long vernalization	Guedira et al. 2014
Rht-B1	4B	*Rht-B1*	T/T C/C	*Rht-B1b* Dwarfing Rht-B1a Wild type	Ellis et al. 2002
Rht-D1	4D	*Rht-D1*	T/T G/G	*Rht-D1b* Dwarfing Rht-D1a Wild type	Ellis et al. 2002
Translocation 2AS:2NS	2A	*Lr37*	G/G A/A	Lr37 Present Lr37 Absent	Helguera et al. 2003, Milus et al. 2015
Translocation 2BS:2GS·2GL:2BL	2B	*IWA8085*	T/T G/G	Sr36 Present Sr36 Absent	Brown-Guedira unpublished
Translocation 1RS:1AL	1A	*1RS:1AL_8035*	T/T C/C	1RS:1AL Present 1RS:1AL Absent	Brown-Guedira unpublished
Translocation 1RS:1BL	1B	*1RS:1BL_6110*	A/A G/G	1RS:1AL Present 1RS:1AL Absent	Brown-Guedira unpublished

S. Table 2. 2: Single nucleotide polymorphism (SNP) distribution across the A, B, and D genomes. SNP came from the 467 genotypes genotyped using genotyping by sequencing (GBS).

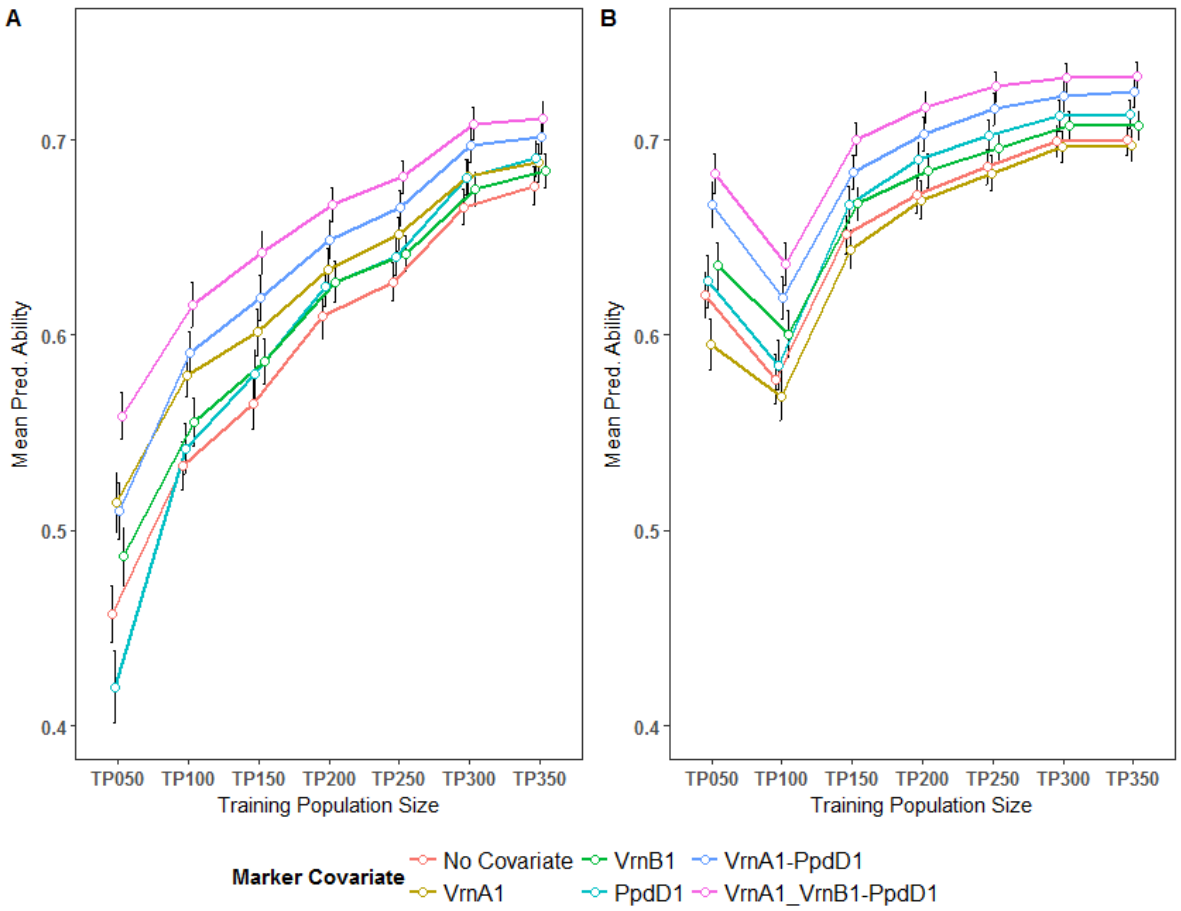
	Genome		
Chromosome	A	B	D
1	1658	2001	1005
2	2027	2490	940
3	1797	3458	503
4	2047	1373	350
5	1860	2374	400
6	1379	2449	653
7	2418	2191	722

S. Table 2. 3: Mean allelic effect, allele frequency and SNP position of fixed covariates utilized for powdery mildew resistance in a model that considered all covariates selected for each trait using the Random training population selection method with a training population size of 350 individuals across the 50 validation sets. Frequency are indicated for the resistant allele.

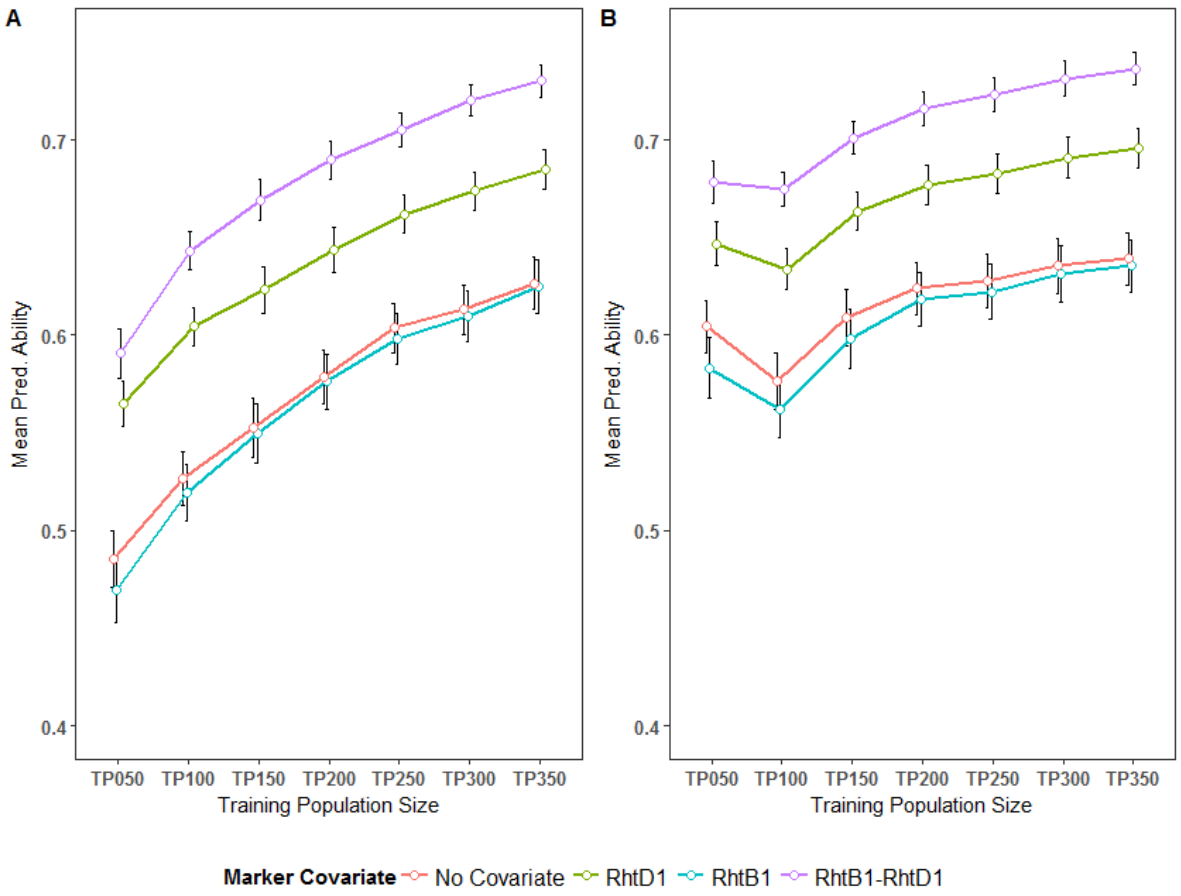
Marker Name	N[a]	Favorable Allele	Chr	Position (Mbp)	Allele frequency	Mean effect [b]
S7A_722257467	8	G	7A	722.26	0.10	-0.46
S7A_724846061	1	G	7A	724.85	0.11	-0.37
S7A_724875049	3	G	7A	724.88	0.11	-0.44
S7A_726607932	11	G	7A	726.61	0.41	-0.45
S7A_730187393	2	G	7A	730.19	0.11	-0.47
S7A_731168451	24	G	7A	731.17	0.12	-0.44

[a] Number of times the marker is the most significant after run association analysis masking the phenotypes that are part of each validation set.

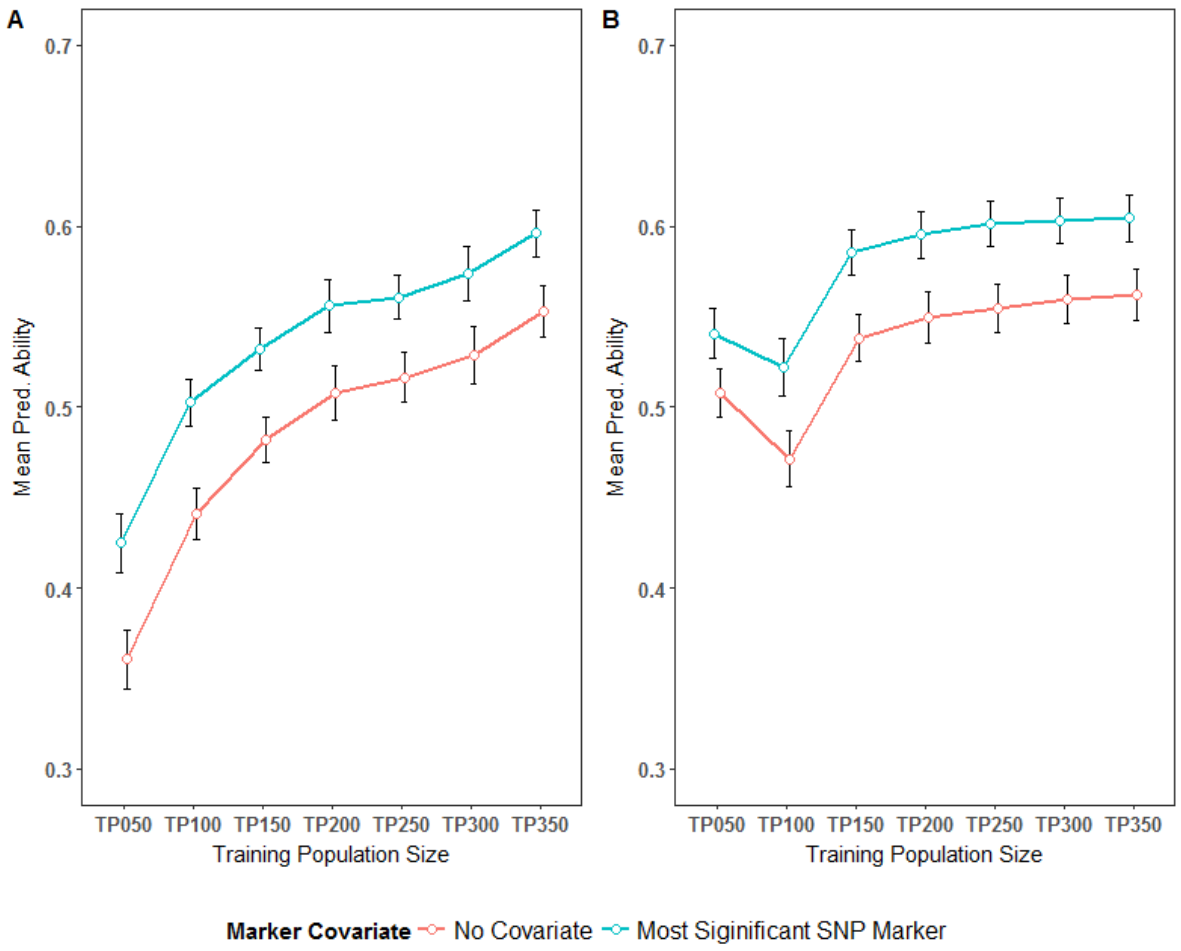
[b] Calculated as the average fixed effect of the marker for training populations were the marker was the most significant using training population size 350 individuals selected at Random.



S. Figure 2. 1: Comparison of mean predictive ability (Mean Pred. Ability) of heading date according to genomic selection models that consider or not the addition of diagnostic markers associated with major effect QTLs as fixed covariates incorporated in the model using different pre-defined training population sizes. Results are presented for training population optimization method Random (A) and PEVmean (B). Error bars represent ± one standard error of the mean.



S. Figure 2. 2: Comparison of mean predictive ability (Mean Pred. Ability) of plant height according to genomic selection models that consider or not the addition of diagnostic markers associated with major effect QTLs as fixed covariates incorporated in the model using different pre-defined training population sizes. Results are presented for training population optimization method Random (A) and PEVmean (B). Error bars represent ± one standard error of the mean.



S. Figure 2. 3: Comparison of mean predictive ability (Mean Pred. Ability) of powdery mildew resistance according to genomic selection models that consider or not the addition of the most significant SNP marker associated with major effect QTL as fixed covariates incorporated in the model using different pre-defined training population sizes. Results are presented for training population optimization method Random (A) and PEVmean (B). Error bars represent ± one standard error of the mean.

CHAPTER 3. Forward genomic predictions in ongoing Southeastern USA wheat breeding programs

As prepared for submission to the journal Crop Science.

J. Martin Sarinelli, J. Paul Murphy, Priyanka Tyagi, James B. Holland, Jerry W. Johnson, Mohamed Mergoum, Richard E. Mason, Ali Babar, Stephen Harrison, Russell Sutton, Carl A. Griffey and Gina Brown-Guedira

J.M. Sarinelli, J.P. Murphy and P. Tyagi, Dep. Of Crop Science and Soil Sciences, North Carolina State University, Raleigh, NC 27695; J.B. Holland and G. Brown-Guedira, Department of Crop and Soil Sciences and USDA-ARS, North Carolina State University, Raleigh, NC 27695; J.W. Johnson and M. Mergoum, Department of Crop and Soil Sciences, University of Georgia, Athens, GA 30602; R.E. Mason, Department of Crop Soil and Environmental Sciences, University of Arkansas, Fayetteville, AR 72701; S. Harrison, Department of Agronomy, Louisiana State University, Baton Rouge, LA 70803; R. Sutton, AgriLife Research, Texas A&M University, College Station, TX 77843; C.A. Griffey, Department of Crop and Soil Environmental Sciences, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061; A. Babar, Agronomy Department, University of Florida, Gainesville, FL 32611.

Abstract

Genomic selection is a molecular breeding tool, that allows prediction of genomic estimated breeding values (GEBVs) of lines prior to phenotyping thus increasing the efficiency of cultivar development programs. Unbalanced historical data sets can be utilized as training populations for the implementation of genomic selection without additional phenotyping. This study evaluated forward genomic predictions in soft red winter wheat (*Triticum aestivum* L.) in the Southern University Small Grains (SUNGRAINS) cooperative testing effort. The training population was based on field testing of an advance generation nursery, the Gulf Atlantic Wheat Nursery (GAWN), and the validation population was based on the earlier generation of field testing of the Southern Uniform Wheat Nursery (SUNWHEAT). Grain yield and test weight were investigated using the G-BLUP mixed model for GEBVs predictions including, two-step univariate and multivariate analysis as well as one-step univariate analysis. We also studied the degree of genetic relationship between the genotypes in the training and validation set. Prediction accuracies for grain yield ranged between 0.44 and 0.84, while for test weight ranged between 0.15 and 0.46. Prediction accuracy using one-step univariate analysis was superior to the two-step analysis by 12 percent for grain yield, but there was a negative difference between one-step and two-step models for test weight. The degree of genetic relationship between individuals in the training and validation set was conserved across different years with an average genetic distance of 0.20. We conclude that unbalanced historical uniform nurseries were a valuable resource for the incorporation of genomic selection in ongoing wheat cultivar development breeding programs in the southeastern United States.

Introduction

Wheat cultivar development programs have traditionally selected superior genotypes based on phenotypic information from annual multi-environment trials. More recently, inexpensive genotyping has been combined with molecular breeding tools to assist selection decisions and increase the rate of genetic gain per breeding cycle (Bernardo, 2010, Singh et al., 2015). One promising tool, genomic selection (Meuwissen et al., 2001), was first proposed for animal breeding but is becoming a common approach in plant breeding also (Heffner et al., 2009, Jannink et al., 2010). Genomic selection predicts genomic estimated breeding values of genotypes (GEBVs) for quantitative traits where many loci with small effects contribute to trait variation. A training population containing individuals with genotypic and phenotypic information is utilized to train a prediction model for the estimation of GEBVs of individuals based on genotypic information alone (Heffner et al., 2009, Jannink et al., 2010).

Several studies using genomic selection in wheat (*Triticum aestivum* L.) have been published for grain yield (Poland et al., 2012, Crossa et al., 2014, Michel et al., 2016), disease resistance traits (Daetwyler et al., 2014, Arruda el at., 2015, Rutkoski et al., 2015) and end-use quality traits (Battenfield et al., 2016). Results have demonstrated the potential of this molecular tool when incorporated into the pipeline of ongoing programs. In general, these studies measured model predictive ability as the Pearson correlation between GEBVs and genotypic values obtained from phenotypic records as best linear unbiased estimate (BLUE) by means of cross validation partitions of the data. Sarinelli et al. (2017) used nine years of unbalanced historical data from the Gulf Atlantic Wheat Nursery (GAWN) that is

part of a cooperative field testing effort among southeastern USA public wheat breeding programs. Maximum predictive abilities for grain yield, test weight, plant height, heading date and powdery mildew resistance of 0.64, 0.56, 0.73, 0.71, and 0.60, were reported These results demonstrated the potential of utilizing an unbalanced historical data set as a training population for genomic selection.

Genomic best linear unbiased predictor (G-BLUP) is one of the statistical methods utilized in genomic selection to predict GEBVs of individuals. The G-BLUP mixed model approach offers flexibility in modeling different genomic prediction scenarios including multiple traits and multi-environment trials simultaneously (de los Campos et al., 2013, Bernal-Vasquez et al., 2017). This method uses a genomic relationship matrix constructed from dense molecular marker information to describe the genomic relationship between individuals (VanRaden, 2008, Hayes et al., 2009). In genomic selection, the phenotypic information available to train the prediction model can be utilized in one-step or two-step analyses. In the one-step analysis the prediction model is constructed using raw phenotypic records from multi-environment evaluation. The two-step analysis first estimates the genotypic value of each genotype (BLUE) from phenotypic records collected in multi-environment trials, and the BLUEs are used to train the prediction model in the second stage (Oakey et al., 2016). Most reports of genomic predictions utilize genomic selection models for single trait (univariate) and a two-step approach to obtain GEBVs (Schulz-Streeck et al., 2013). However, statistical methods that simultaneously use multiple traits and account for the entire variance-covariance structure of the observed data, including the effect of genotype by environment interaction, in more complex model structures can improve genomic

prediction accuracies (Calus and Veerkamp, 2011, Burgueño, et al., 2012, Schulz-Streeck et al., 2013, Crossa et al. 2016, Bernal-Vasquez et al., 2017). Multivariate models take advantage of the genetic correlation between traits to increase prediction accuracy. Jia and Jannink (2012) showed an advantage of multivariate models when the genetic correlation between traits considered in the analysis was high and when phenotypic records were not available for all individuals. On the other hand, the incorporation of the genotype by environment interaction in the model better accounted for the relationship between individuals evaluated in different environments to obtain more precise GEBV estimates. Lado et al. (2016), using a highly unbalanced wheat data set, found increases in predictive ability of genomic selection models after including the genotype by environment interaction effect in the prediction equation.

Plant breeders conduct annual field evaluations for multiple genotypes in multiple environments to assess performance for different target traits (Cullis et al., 2006). The use of unbalanced historical data sets of breeding lines as training population provides phenotypic records for different traits evaluated in multi-environment trials (Dawson et al., 2013, Rutkoski et al., 2015).The amount and quality of the information available for each trait in each environment varies and this can impact model prediction accuracy. Initial data curation based on coefficient of variation, repeatability or minimum thresholds for certain traits can help to identify trials with low precision, which once removed, increases the quality of the final data set.

Studies using historical data sets for forward predictions in ongoing breeding programs with genetic materials at a different stage in the breeding pipeline are limited. Southeastern USA public wheat breeders evaluate two cooperative uniform nurseries designated the Southern Uniform Wheat Nursery (SUNWHEAT) and Gulf Atlantic Wheat Nursery (GAWN). These are sequential in the cultivar development pipeline and are evaluated annually in up to seven locations. The objectives of this study were to: (1) evaluate the prediction accuracy of genomic selection in an empirical forward prediction scheme using the historical GAWN as a training population to predict grain yield and test weight of entries in multiple years of SUNWHEAT; (2) determine how genetic relatedness between the training population and selection candidates varies across time and breeding programs participating in the nurseries; (3) compare genomic prediction accuracy of multivariate vs univariate approaches and one-step vs two-step genomic selection models using G-BLUP; and (4) evaluate the effect of phenotypic data curation on model accuracy.

Materials and Methods

Plant material

We utilized phenotypic and genotypic data from two separate uniform cooperative nurseries of soft red winter wheat that were part of the multi-location evaluation efforts of southeastern USA public wheat cultivar development programs. The GAWN nursery typically contains F_8 or later generation lines that have undergone three generations of yield evaluation, while the SUNWHEAT nursery typically contains lines in the immediate prior

generation. Typically, the superior lines in the SUNWHEAT nursery are entered in to the GAWN nursery. The GAWN dataset utilized in this study contained 430 elite lines developed by The University of Arkansas (AR), The University of Florida (FL), The University of Georgia (GA), Louisiana State University (LA), North Carolina State University (NC), Clemson University (SC), Texas A&M AgriLife Research (TX) and Virginia Polytechnic Institute and State University (VA). Data from 2008 to 2015 were utilized for the training population. GAWN datasets were unbalanced across years with respect to genotypes with a few commercial checks and entries repeated across years. A detailed description of this material can be found in Sarinelli et al. (2017). The SUNWHEAT dataset utilized in this study contained lines from the same institutions entering genotypes in the GAWN, with the exception of Clemson University and Virginia Polytechnic Institute and State University that do not participate at this stage. The SUNWHEAT nursery, grown from 2014 to 2016, was utilized as a validation set in this study. It contained 75 to 79 different experimental lines and four checks cultivars per year.

The experimental lines in SUWNWHEAT were derived primarily from bi-parental (59 percent) and three way (40 percent) crosses, with complex cross-populations representing only 1 percent of the complete set of genotypes. Populations were created using 181 different parents, of which only 13 were evaluated in the GAWN nursery in some year of the historical series considered in this study. The breeding method utilized prior to entry in the cooperative nurseries was bulk-pedigree. Depending on the program, the F_2 to F_5 generations were advanced by families in bulk and spikes were selected at different generations for entry into a pedigree selection protocol that utilized visual head row evaluations (Sarinelli et al., 2017).

Following head row generations, and before entry in the uniform cooperative nurseries, entries underwent one or more years of yield and test weight evaluations in their states of origin.

Phenotypic data collection and analyses

Training populations and validation sets were evaluated at one location in up to seven southern states per year: Arkansas (Stuttgart or Marianna), Florida (Citra or Quincy), Georgia (Plains), Louisiana (Winnsboro), North Carolina (Kinston), and Texas (Farmersville). The GAWN training population was also evaluated at Warsaw, Virginia. Grain yield, expressed as Mg ha^{-1} and test weight, expressed as kg m^{-3} were the traits analyzed in this study. Experimental designs in each environment (location by year combination) were randomized complete block designs with up to three replications, although in some environments, data for grain yield and test weight were recorded in only one replication. Plot sizes were at least 1.3 meters wide and 3.1 meters long, typical of wheat yield trial plots in the region.

Two scenarios were evaluated to determine the effect of the phenotypic data quality of the training population on the prediction accuracy of the model. One scenario used all data points available in the historical series (No-Data-Curation). In the second scenario, only data from environments with coefficient of variation lower that 15 percent and an average grain yield above 2.69 Mg ha^{-1} were utilized (Data-Curated). The following linear model was implemented to analyze grain yield and test weight in each environment:

$$y_{ij} = \mu + B_i + G_j + \varepsilon_{ij}$$

where, y_{ij} was the phenotypic observation of each individual genotype in the i^{th} block, μ was the overall mean, G_j was the genotypic effect and $\varepsilon_{ij,}$ represented the residual term. This

model considered the residual term as random effect with ε~IDD (0, σ_e^2) and the other terms as fixed effects. Then coefficient of variation for each environment was calculated as:

$$\text{CV}= \frac{\sqrt{\sigma_e^2}}{\mu} * 100.$$

Additionally, a linear mixed model was utilized for the combined analysis of grain yield and test weight over environments:

$$y_{ijk} = \mu + E_i + B(E)_{ij} + G_k + EG_{ik} + \varepsilon_{ijk}$$

with, y_{ijk} representing the phenotypic observation of genotype *k* in the i^{th} environment in the j^{th} block, μ was the overall mean, E_i was the environment effect (location by year combination), $B(E)_{ij}$ was the block effect nested within environment, G_k was the genotypic effect, EG_{ik} is the interaction term of genotype by environment and $\varepsilon_{ijk,}$ represented the residual term. In this model the overall mean and the genotypic effect were considered fixed and all the remaining terms random. Random effects (u) and residuals (ε) were assumed to be normally and independently distributed u~IDD (0, $I\sigma_u^2$) and e~IDD (0, $I\sigma_e^2$). Best linear unbiased estimates (BLUEs) of grain yield and test weight for each genotype in the training population were estimated considering the genotypic effect and the intercept as fixed effects while maintaining the rest of the model terms as random effects.

Estimates of broad sense heritability on a plot basis for each trait and for each training population were calculated using the statistical model previously described with the overall mean (μ) as a fixed effect and all the remaining terms as random. The broad sense heritability of each trait was computed according to Holland et al. (2003) as:

$$H^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_{gl}^2 + \sigma_e^2}$$

where, σ_g^2 represented genotypic variance, σ_{gl}^2 are the variances component due to the genotype by environment interaction and σ_e^2 is the variance associated with the residual term. Also we estimated heritability on line mean-basis by using the formula suggested by Cullis et al. (2006) as:

$$H_c^2 = 1 - \frac{\bar{V}_{BLUP}}{2\sigma_g^2}$$

where, $\bar{V}_{BLUP}$ is the average variance of a difference of two BLUPs (obtained from ASReml output in the .pvs file).

To estimate BLUEs and variance components in the validation sets, we utilized the same statistical mixed model as in the combined analysis for the training population, but in this case the environments represent different locations within the same year. Variance components from each validation set were utilized to compute broad sense heritability on per plot basis (Holland et al., 2003) and in entry mean basis (Cullis et al., 2006) as previously specified.

The models were implemented using ASReml (Gilmour et al., 2015) to obtain restricted maximum likelihood estimates (REML) of variance components and solve the mixed linear model equations.

Genotypic Data

A total of 634 wheat genotypes evaluated in GAWN from 2008 to 2015 and SUNWHEAT from 2014 to 2016 underwent genotyping by sequencing (GBS, Elshire et al., 2011), using the protocol described by Poland et al. (2012). Briefly, DNA was extracted from 10 day old leaves using DNEasy 96 Plant Kits (Qiagen, Venlo, Netherlands). Genome

complexity reduction was carried out using a combination of two enzymes *MspI* (CCGG) a common cutter and *PstI* (CTGCAG) a rare cutter and barcoded adaptors were ligated to each sample. Ninety six individual samples were pooled into a single library and polymerase chain reaction amplified. Each pooled library was sequenced in an Illumina HiSeq 2500 instrument at the Genomic Science Laboratories (North Carolina State University). After sequencing, single nucleotide polymorphisms (SNP) were scored using the Tassel5 pipeline (Glaubitz et al., 2014) using aligner method of bwa version 0.7.12 (Li and Durbin 2009) for aligning SNPs to reference sequence. Polymorphic SNP were identified and saved in a file with hapmap format.

A series of specific filters for minor allele frequency (MAF), marker heterozygosity, and proportion of missing data for markers were applied to increase the quality of the marker set. Only SNP with less than 50 percent missing data, more than five percent MAF and less than 10 percent of heterozygous calls per marker locus were retained, yielding 58,955 SNP that were imputed with Beagle4 using a function in the R package synbreed (Wimmer et al. 2012, Browning and Browning 2016). After imputation, SNP with more than 10 percent of heterozygosity and MAF lower than 5 percent were removed for downstream analysis. The final number of GBS SNPs utilized for the analysis was 44,844.

Additionally, 12 KASP markers routinely used for marker assisted selection in southeastern breeding programs were screened and incorporated into the GBS SNPs panel. These included markers predictive for different alleles for reduced plant height (*Rht-B1* and *Rht-D1*), vernalization duration (*Vrn-A1* and *Vrn-B1*), and photoperiod response (*Ppd-A1*, *Ppd-B1*, *Ppd-D1*). Markers were also associated with presence of alien translocated

chromosomes including the *t1AL:1RS* and *t1BL:1RS* translocations from rye (*Secale cereale* L.), the *t2BS:2GS·2GL:2BL* translocation from *Triticum timopheevii* (Zhuk) and the *t2AS:2NS* from *Triticum ventricosum* (Tausch).

Training Populations and Validation Sets

Plant breeders select superior genotypes at the end of the crop season based on phenotypic data collected from multi-environment trials to advance elite materials to the following stage of the breeding pipeline. Thus, our scheme of forward prediction to evaluate the accuracy of genomic selection used the GAWN as a training population to predict performance of the earlier generation materials entered across three field seasons of the SUNWHEAT as validation sets. In each year, new data were added to the training population such that three sets of genotypes from the historical GAWN nursery were utilized to construct the training populations: genotypes from 2008 to 2013 (Subset1), genotypes from 2008-2014 (Subset2) and genotypes from 2008 to 2015 (Subset3). The subsets were utilized to predict GEBVs for genotypes included in the SUNWHEAT nursery as follows, Subset1 was utilized to predict genotypes evaluated in the SUNWHEAT 2014, Subset2 to predict genotypes in the SUNWHEAT 2015 and Subset3 to predict genotypes in the SUNWHEAT 2016. Genotypes were not allowed to be present in the training population and validation set simultaneously.

Model accuracy is measured as the Pearson correlation between GEBVs and true breeding value of each individual (TBV). However, because the TBV is usually unknown, predictive ability is measured as the ratio of the Pearson correlation between GEBVs and the genotypic value estimated from phenotypic data (BLUE) and the square root of the

heritability in each validation set considered (Heffner et al. 2009, Jannink et al. 2010, Lorenz et al. 2011), calculated as:

$$r = \frac{\rho(GEBV/BLUE)}{\sqrt{H^2}}$$

To compare genomic selection with selection based on field evaluation, genotypes in each validation set were ranked from the most to least favorable for grain yield using adjusted phenotypic means (BLUE) and GEBVs, respectively. The number of individuals that are simultaneously selected by phenotypic selection using a selection intensity of 20 percent and genomic selection using a selection intensity of 40 percent was determined.

We studied the genetic similarity between the individuals in the validation and training population set using genetic distances. The average minimum genetic distance of genotypes from different breeding program participating in each validation set with genotypes evaluated in the training population were measured. The genetic distance was calculated using software TASSEL version 5 as: 1 - IBS (identical by state), with IBS defined as the probability that alleles drawn at random from two different genotypes at the same locus are the same (Bradbury et al., 2007). Values close to zero represented genotypes that more closely related. To perform the analysis, we first calculated the genetic distances in the complete data set. Then we partitioned the analysis for each validation set and by breeding program within validation set. We also partitioned the genotypes in the training population based on the genotypes that were evaluated each year of the historical series. For example, from the 80 genotypes evaluated in SUNWHEAT 2016, we created six subsets of genotypes representing the six breeding programs. The training population utilized for

predictions of SUNWHEAT 2016 included the historical GAWN nursery from 2008 to 2015 which was also partitioned by year as previously mentioned. Then the minimum genetic distance between the genotypes of each breeding program with the genotypes in each subset of the training population was extracted and the average minimum genetic distance by program calculated. We also recorded the absolute minimum genetic distance for each subset.

Genomic selection

Predictions of GEBVs for genotypes in different validation sets were obtained using single and two-step genomic selection approaches. For the single-step analyses, raw phenotypic records available for the training population were used to construct a genomic selection model that accounted for the random effects of block, environment and genotype by environment interaction simultaneously to obtain GEBVs of individuals in the validation set. In the stepwise approach (two-steps), BLUEs of genotypes in the training population were estimated using a linear mixed model as described above. Then BLUEs were used in combination with the genotypic information to train the model to predict GEBVs of genotypes in the validation set. Additionally, for the stepwise approach the analysis was made one trait at a time (univariate) and combining both traits in the same model (multivariate) in an effort to capture the genetic correlation between traits. Multivariate analysis with a one-step approach was not performed due to the high degree of unbalance in the data set that difficult convergence of the statistical model.

A linear mixed model was used to predict GEBVs known as genomic best linear unbiased predictors (G-BLUPS; Meuwissen et al., 2001, Piepho, 2009) with the genomic relationship matrix (**G**) calculated according to VanRaden (2008), method 1, as follows:

$$\mathrm{G} = \frac{\mathrm{ZZ}'}{2\sum p_i(1-p_i)}$$

Z= M - P, where M was a matrix of dimension *nxm* with genotypes in rows and markers in columns, in which each element represents the number of minor alleles carried by an individual at a locus; and P was another *nxm* matrix containing the minor allele frequency of each marker multiplied by 2 in each column. p_i, was the observed minor allele frequency of marker *i*.

- Stepwise univariate analysis:

$$y = \mu + Zu + e$$

y was an *nx1* vector of BLUEs of the *n* wheat genotypes, *μ* represented the overall mean, *u* was an *nx1* vector of random genotypic effects, Z was the design matrix that related *y* with random genotypic effects, with dimension *nxn* and *e* was an *nx1* vector representing the residual terms. The variance-covariance structure associated with the random term was u ~ N (0, $\mathbf{G}\sigma_u^2$) where σ_u^2 was the additive genetic variance. The variance associated with the residual term was e ~ N (0, $\mathrm{I}\sigma_e^2$) where σ_e^2 is the residual variance and I was an identity matrix of dimension *nxn*.

- Stepwise multivariate analysis:

$$y = X\beta + Zu + e$$

The basic form of the statistical model was similar as that used for univariate analysis where, *y* was an *nxt* matrix of BLUEs of the *n* wheat genotypes for the *t* traits evaluated, β was a vector containing the overall mean for grain yield and test weight respectively and X was a design matrix of dimension *nxt*, *u* was an *nxt* matrix of random genotypic effects on each trait, Z was the design matrix that related *y* with random genotypic effects, with dimension *nxn* and *e* was an *nxt* matrix representing the residual terms. The variance-covariance structure associated with the random term was u ~ N (0, $\boldsymbol{G} \otimes \sum_u$) where $\sum_u$ was variance-covariance matrix of dimension *txt* for the multiple traits considered. The variance-covariance structure associated with the residual term was e ~ N (0, $I \otimes \sum_e$) with $\sum_e$ as the variance-covariance matrix of dimension *txt* for residuals for each trait and I was an identity matrix of dimension *nxn*.

- One step univariate analysis:

$$y = \mu + Z_s u_s + Z_g u_g + Z_{ge} u_{ge} + e$$

y was an *nx1* vector of phenotypic observations of the *g* wheat genotypes within different environments, μ represented the overall mean, u_s was a vector of random environmental effects which considered the effect of environment and block, u_g was a *gx1* vector of random genotypic effect, and u_{ge} was a vector of genotype by environment interaction effects. Z_g, Z_s and Z_{ge} were the design matrices that relate *y* with random genotypic effects and the remaining random terms in the model. *e* was an *nx1* vector representing the residual term in the model. The variance-covariance structure associated with the genotype random term was u_g ~ N (0, $\mathbf{G}\sigma_g^2$) where σ_g^2 was the additive genetic variance and **G** represented the realized

relationship matrix of dimension *gxg*. u_s, was normally and independently distributed u_s~N (0, Iσ_s^2). The variance component associated with u_{ge} was ~ N (0, Jσ_{ge}^2) where J is a block diagonal matrix, where each block is the G matrix and corresponds to covariance within an environment, with no covariance between environments, and σ_{ge}^2 was the genotypic by environment variance. The variance component associated with the residual term was e ~ N (0, Iσ_e^2).

Results

Descriptive Statistics

The number of lines in the training populations utilized for forward prediction were 363, 395 and 430 for GAWN2008-13, GAWN2008-14 and GAWN2008-15, respectively. Genotypes were evaluated at a maximum of seven locations each year. Each training population was utilized to predict a set of genotypes evaluated in the next year of the SUNWHEAT nursery. The size of the validation sets varied between 74 and 80 (Table 3.1).

The total number of environments from which training population data were collected varied from 38 to 47, although most lines in the training set were evaluated in only a subset of the environments, because the training population was updated annually. On average, training populations contained 30 percent more data points for grain yield than for test weight. Average mean grain yield and test weight for each training population utilized in this study ranged from 4.32 to 4.38 Mg ha^{-1} and from 57.2 to 57.3 kg m^{-3}, respectively. Broad sense heritability for each trait on line mean basis for the training populations range from

0.59 to 0.62 and from 0.79 to 0.81 for grain yield and test weight, respectively. Trait heritabilities did not vary much with the updating of the training population for the three subsets utilized in this study (Table 3.1).

Individual phenotypic analysis was performed for validation sets SUNWHEAT2014, SUNWHEAT2015 and SUNWHEAT2016 with mean grain yields of 5.13, 3.56 and 3.66 Mg ha^{-1} respectively. Mean test weights were 58.0 kg m^{-3} for SUNWHEAT 2014 and SUNWHEAT 2015 and 53.5 kg m^{-3} for SUNWHEAT 2016. Broad sense heritability estimated on line mean basis ranged from 0.40 to 0.52 for grain yield and 0.63 to 0.83 for test weight (Table 3.1).

Coefficient of variation and summary statistics for individual environments where the historical training population was evaluated exhibited a wide range of variation for both traits. Mean grain yield ranged from 2.3 to 6.2 Mg ha^{-1}, while coefficient of variation ranged from 4.7 to 17.3 percent, while the average test weight varied from 50.5 to 63.4 kg m^{-3} and coefficient of variation varied from 0.4 to 3 (Table 3.2).

Genetic distance and genotypic information

The number of markers varied by chromosome, with limited polymorphism detected in the D genome (Figure 3.1A). The minor allele frequency distribution for the complete data set of 634 genotypes (Figure 3.1B) was similar to the MAF distribution observed in other wheat panels (Fu, 2014).

We studied the genetic similarity between the individuals in the validation and training population set using genetic distance. The mean genetic distance of genotypes in each validation set developed by different breeding programs and genotypes in the training

populations varied between 0.11 to 0.24 (Table 3.3). The overall mean genetic distance was 0.20. When genotypes entering the training population in different years were compared with the three validation sets, we did not see important changes in relatedness over time. The relatively small and uniform genetic distance values observed across breeding programs and years suggested a high degree of relationship between the training populations and validation sets. For the three validation sets considered in this study, wheat genotypes from University of Georgia were consistently more closely related to individuals in the training population, with average genetic distance of 0.17. In each year, lines from the University of Georgia had the lowest minimum genetic distance with individuals in the training population (Table 3.3).

Genomic selection accuracies

High predictions accuracies for grain yield were obtained for the SUNWHEAT 2014 and 2015 validation sets with maximum accuracies across models of 0.83 and 0.84, respectively, while for SUNWHEAT 2016 the maximum accuracy was 0.64. For grain yield the model that gave the highest accuracies was the one-step univariate analysis. Moderate accuracies were obtained for test weight, with maximum accuracies of 0.43, 0.37 and 0.46 for the validation sets evaluated in 2014, 2015 and 2016. Genomic selection methods that gave the highest accuracy for test weight varies by year (Table 3.4). The effect of phenotypic data curation (Data-Curated) before training the prediction model for grain yield evaluated as a change in prediction accuracy for the three methods considered in this study was favorable with an overall 16 percent increase in model accuracy while, that was not the case for test weight where a reduction in prediction accuracy in a similar proportion was observed (Table 3.4).

Model accuracy for the two-step univariate and multivariate analysis gave similar results with accuracies ranging from 0.55 to 0.78 and 0.30 to 0.46 for grain yield and test weight, respectively, after phenotypic data curation (Table 3.4). For the two-step multivariate analysis, the genetic correlation between both traits analyzed was 0.37 for the training population GAWN 2008-13, 0.34 for GAWN 2008-14 and 0.40 for the training population GAWN 2008-15. Model accuracy in the one-step univariate analysis was always higher compared with the two-step univariate and multivariate analysis for grain yield across the three validation sets, with an average increase of 12 percent. The maximum increase in prediction accuracy was 15 percent in the validation set SUNWHEAT 2015. Comparison of prediction accuracies between the one-step and the two-step analysis for test weight found an overall 5.8 percent reduction in accuracy when one-step was utilized.

Finally, we evaluated the required selection intensity to apply to GEBVs in order to capture a high proportion of the superior individuals in validation sets when applying a 20 percent selection intensity to phenotypic data that would be typical of the SUNWHEAT nursery. Across the three years evaluated we found that a selection intensity of 40 percent applied to the GEBVs for grain yield captured 65 percent or more of the genotypes selected based on phenotypic data. Genotypes from the University of Georgia had high overall performance, based on phenotypic observations and genomic predictions (Figure 3.2, Figure 3.3, Figure 3.4).

Discussion

Plant breeding programs generate thousands of new data points per year for genotypes at different stages in the cultivar development pipeline. Only lines with superior performance are evaluated in more than one season. The configuration of the testing structure generates highly unbalanced data sets connected by genotypes utilized as performance checks. These historical unbalanced phenotypic data sets are a valuable resource for training population construction if genomic information is also available to predict GEBVs of untested genotypes. In this study, we used data from two uniform nurseries that represented different testing stages of the SUNGRAINS cooperative breeding effort. The unbalanced historical data set GAWN was used as a training population to predict performance of genotypes in SUNWHEAT, a nursery at an earlier stage in the breeding pipeline. Validation results in our study shown high prediction accuracies for grain yield and moderate accuracies for test weight. Maximum prediction accuracies were 0.84 and 0.46 for grain yield and test weight, respectively. These results support the use of historical unbalanced uniform nurseries as training sets for genomic selection in wheat breeding in the southeastern United States.

By using phenotypic selection, plant breeders attempt to discriminate superior genotypes in earlier stages to advance to the next generation based on records from a limited number of environments. This is due in part to the limited seed availability, but also to the expense of testing large numbers of genotypes. Thus, genomic selection has the potential to perform similar or better than phenotypic selection from field evaluation in a limited number of environments. In a study comparing the efficiency of phenotypic selection versus genomic selection for grain yield and resistance to Fusarium head blight in barley (*Hordeum vulgare*

L.), Sallam and Smith (2016) concluded that genetic gains by genomic selection were similar to gains obtained with phenotypic selection. They found that the relative efficiency of genomic selection was greater than phenotypic selection when the accuracy of predictions were higher than the square root of heritability for the trait (accuracy of phenotypic selection). In our study, we found that the accuracy of genomic prediction for grain yield were higher than the square root of heritability in validation set SUNWHEAT2014 and SUNWHEAT2015 but not for validation set SUNWHEAT2016. Test weight prediction accuracies were never higher than the square root of heritability on an entry mean basis. From these results, we can infer similar efficiencies between selection based on phenotypic and genotypic information for grain yield but not for test weight.

Different criteria can be implemented for data curation of phenotypic records including repeatability, coefficient of determination and coefficient of variation (Dawson et al., 2013). We evaluated use of the coefficient of variation and environment average grain yield for environment selection because they were the most common criteria utilized for SUNGRAINS breeders to determine which data include in combined analyses. We found that data curation for grain yield increased the accuracy of genomic predictions by 16 percent on average for different validation sets and statistical models considered in the study. However, when the curated data for grain yield was utilized to predict test weight, a reduction of 20 percent on average was observed. This result suggested that data curation must be trait specific and not just based on the grain yield.

We found differences between G-BLUP models utilized for predictions with an advantage for the single trait one-step analysis in comparison with the two-step univariate

and multivariate two-step analyses. One advantage of univariate one-step models can be attributed to the use of a more complex model structure with the incorporation of the genotype by environment interaction effects into the model. Other studies that included the genotype by environment interaction in the statistical model observed increased model prediction accuracy when compared primarily with the two-step analysis (Jarquin et al. 2014, Crossa et al., 2015, Lado et al. 2016, Oakey et al. 2016).

A multivariate approach takes advantage of the genetic correlation between traits to improve the accuracy of genomic predictions. Comparisons between multivariate versus univariate analysis were performed only for the two-step analysis because the one-step model cannot converge to a restricted maximum likelihood due to the high degree of unbalanced data in the set. Comparisons of the two-step univariate and multivariate analysis did not shown differences in terms of increasing the prediction accuracy of the models. Jia and Jannink (2012) concluded that to obtain an advantage in prediction accuracy from multivariate models the genetic correlation between traits must be high. The genetic correlation between grain yield and test weight in our study varied between 0.34 and 0.40. Another possible explanation could arise from the fact that heritability of the traits in the multivariate analysis was low. Previous studies of genomic selection using multivariate models suggested that at least one of the traits in the model must have high heritability to increase prediction accuracy of the trait with low heritability in comparison with univariate analysis because low heritable traits can borrow information from high heritable traits if they are correlated (Jia and Jannink, 2012, Guo et al., 2014).

Overall the genetic distance between training population and validation set was low (0.20), with a minimum for the cultivar development program from Georgia (0.17). Lines derived from Georgia were in general more used as parents than any other germplasm from the remaining breeding programs involved in the cooperative nursery. Genetic distance between selection candidates and genotypes from the historical training population did not increase over the period considered. Previous reports demonstrated that the degree of genetic relationship between individuals in the training population and the validation set influenced the accuracy of genomic prediction (Habier et al., 2013). Clark et al. (2012) reported increases in prediction accuracy when the individuals in the training population and validation sets were more closely related. In this study, the high prediction accuracies for the traits evaluated in combination with the low genetic distance between training and validation sets reinforced the use this unbalanced historical data to train prediction models in ongoing breeding programs.

Finally, we evaluated the required selection intensity to be imposed on GEBVs in order to capture a high proportion of the individuals selected based on phenotypic evaluation. We found that a selection intensity of 40 percent on GEBV permitted a recovery of at least 65 percent of the phenotypic selections when a selection intensity of 20 percent was imposed on phenotypic results. These values were consistent across the different validation sets evaluated. These results were promising with respect to improving the efficiencies in cultivar development programs. For example, if the number of new lines advanced to the yield evaluation stage by a breeding program are kept constant, selecting the best 40 percent based on GEBVs will reduce the number of field plots by 60 percent. Another option could be to

keep the number of field plots constant, but increase the number of potential selection candidates to be selected based on GEBVs by 150%. The incorporation of genomic selection in the pipeline of cultivar development programs will allow breeders to evaluate and incorporate different breeding strategies with a direct impact in the genetic gain and the efficiency of the program.

Conclusion

Results from this study are encouraging regarding the use of uniform historical unbalanced nurseries as training populations for genomic predictions of untested genotypes. This strategy offers the possibility of discarding genotypes with inferior GEBVs before planting them in replicated trials and concentrating resources on genotypes with superior GEBVs. This could increase the rate of genetic gain per breeding cycle. Results reported in this study are the basis for the adoption of Genomic Selection in SUNGRAINS wheat breeding programs.

The G-BLUP model using multi-environment trials in one-step to simultaneously account the genotypic effect as well as other factors, such as genotype by environment interaction, outperformed both univariate and multivariate two-step models. We did not find any benefit to the multivariate approach for grain yield and test weight when a two-step approach was utilized to predict GEBVs.

Future Analysis

Overall prediction accuracies for each validation set evaluated were high for all years considered. However, prediction accuracy by individual breeding program varies across each

year. This can be a consequence of the genetic structure within each breeding program or because genetic distances between lines in the training and validation sets differ among breeding programs. The use of training population optimization methods could be a strategy to investigate if predictive accuracy by breeding program can be improved.

Acknowledgments

Special thanks to the technical staff of USDA small grains lab, Kim Howell and Jared Smith for their assistance in genotypic data collection. Funding resources utilized in this research were from United States Department of Agriculture National Institute of Food and Agriculture (nifa.usda.gov) Triticeae Coordinated Agricultural Project (grant No. 2011-68002-30029) and Monsanto Graduate Fellowship supporting graduate research assistantship of Martin Sarinelli.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

Arruda, M.P., Brown, P.J., Lipka, A.E., Krill, A.M., Thurber, C. and Kolb, F.L. 2015. Genomic selection for predicting head blight resistance in a wheat breeding program. The Plant Genome, 8: 1-12.

Battenfield, S.D., Guzmán, C., Gaynor, R.C., Singh, R.P., Peña, R.J., Dreisigacker, S., Fritz, A.K. and Poland, J.A., 2016. Genomic selection for processing and end-use quality traits in the CIMMYT spring bread wheat breeding program. The Plant Genome, 9: 1-12.

Bernal-Vasquez, A.M., Gordillo, A., Schmidt, M. and Piepho, H.P. 2017. Genomic prediction in early selection stages using multi-year data in a hybrid rye breeding program. BMC genetics, 18: 1-17.

Bernardo, R. 2010. Breeding for quantitative traits in plants. 2nd ed. Stemma Press, Woodbury, MN. ISBN 978-0-9720724-1-0.

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

Browning, B.L. and Browning, S.R. 2016. Genotype imputation with millions of reference samples. The American Journal of Human Genetics, 98: 116-126.

Burgueño, J., de los Campos, G., Weigel, K. and Crossa, J. 2012. Genomic prediction of breeding values when modeling genotype× environment interaction using pedigree and dense molecular markers. Crop Science, 52: 707-719.

Calus, M.P. and Veerkamp, R.F. 2011. Accuracy of multi-trait genomic selection using different methods. Genetics Selection Evolution, 43: 26.

Clark, S.A., Hickey, J.M., Daetwyler, H.D. and van der Werf, J.H. 2012. The importance of information on relatives for the prediction of genomic breeding values and the

implications for the makeup of reference data sets in livestock breeding schemes. Genetics Selection Evolution, 44: 1-9.

Crossa, J., de los Campos, G., Maccaferri, M., Tuberosa, R., Burgueño, J. and Pérez-Rodríguez, P. 2016. Extending the marker× environment interaction model for genomic-enabled prediction and genome-wide association analysis in durum wheat. Crop Science, 56: 2193-2209.

Crossa, J., Pérez, P., Hickey, J., Burgueño, J., Ornella, L., Cerón-Rojas, J., Zhang, X., Dreisigacker, S., Babu, R., Li, Y. and Bonnett, D. 2014. Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity, 112: 48-60.

Cullis, B.R., Smith, A.B. and Coombes, N.E. 2006. On the design of early generation variety trials with correlated data. Journal of Agricultural, Biological, and Environmental Statistics, 11: 381-393.

Daetwyler, H.D., Bansal, U.K., Bariana, H.S., Hayden, M.J. and Hayes, B.J. 2014. Genomic prediction for rust resistance in diverse wheat landraces. Theoretical and Applied Genetics, 127: 1795-1803.

Dawson, J.C., Endelman, J.B., Heslot, N., Crossa, J., Poland, J., Dreisigacker, S., Manès, Y., Sorrells, M.E. and Jannink, J.L. 2013. The use of unbalanced historical data for genomic selection in an international wheat breeding program. Field Crops Research, 154: 12-22.

de los Campos, G., Hickey, J.M., Pong-Wong, R., Daetwyler, H.D. and Calus, M.P. 2013. Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics, 193: 327-345.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S. and Mitchell, S.E. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PloS one, 6: 1-10.

Fu, Y.B. 2014. Genetic diversity analysis of highly incomplete SNP genotype data with imputations: an empirical assessment. G3: Genes| Genomes| Genetics, 4: 891-900.

Gilmour, A.R., Gogel, B.J., Cullis, B.R., Welham, S.J. and Thompson, R. 2015. ASReml user guide release 4.1 structural specification. Hemel Hempstead: VSN International Ltd.

Glaubitz, J.C., Casstevens, T.M., Lu, F., Harriman, J., Elshire, R.J., Sun, Q. and Buckler, E.S. 2014. TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. PloS one, 9(2), p.e90346.

Guo, G., Zhao, F., Wang, Y., Zhang, Y., Du, L. and Su, G. 2014. Comparison of single-trait and multiple-trait genomic prediction models. BMC genetics, 15: 1-7.

Habier, D., Fernando, R.L. and Dekkers, J.C.M. 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics, 177: 2389-2397.

Habier, D., Fernando, R.L. and Garrick, D.J. 2013. Genomic BLUP decoded: a look into the black box of genomic prediction. Genetics, 194: 597-607.

Hayes, B.J., Visscher, P.M. and Goddard, M.E. 2009. Increased accuracy of artificial selection by using the realized relationship matrix. Genetics research, 91: 47-60.

Heffner, E.L., Sorrells, M.E. and Jannink, J.L. 2009. Genomic selection for crop improvement. Crop Science, 49: 1-12.

Holland, J.B., Nyquist, W.E. and Cervantes-Martínez, C.T. 2003. Estimating and interpreting heritability for plant breeding: an update. Plant breeding reviews, 22: 9-112.

Jannink, J.L., Lorenz, A.J. and Iwata, H. 2010. Genomic selection in plant breeding: from theory to practice. Briefings in functional genomics, 9: 166-177.

Jia, Y. and Jannink, J.L. 2012. Multiple-trait genomic selection methods increase genetic value prediction accuracy. Genetics, 192: 1513-1522.

Lado, B., Barrios, P.G., Quincke, M., Silva, P. and Gutiérrez, L. 2016. Modeling genotype× environment interaction for genomic selection with unbalanced data from a wheat breeding program. Crop Science, 56: 2165-2179.

Li, H. and Durbin, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25: 1754-1760.

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

Michel, S., Ametz, C., Gungor, H., Akgöl, B., Epure, D., Grausgruber, H., Löschenberger, F. and Buerstmayr, H. 2016. Genomic assisted selection for enhancing line breeding: merging genomic and phenotypic selection in winter wheat breeding programs with preliminary yield trials. Theoretical and Applied Genetics, 130: 363-376.

Oakey, H., Cullis, B., Thompson, R., Comadran, J., Halpin, C. and Waugh, R. 2016. Genomic selection in multi-environment crop trials. G3: Genes| Genomes| Genetics, 6: 1313-1326.

Poland, J., Endelman, J., Dawson, J., Rutkoski, J., Wu, S., Manes, Y., Dreisigacker, S., Crossa, J., Sánchez-Villeda, H., Sorrells, M. and Jannink, J.L. 2012. Genomic selection in wheat breeding using genotyping-by-sequencing. The Plant Genome, 5: 103-113.

Poland, J.A., and Rife, T.W. 2012. Genotyping-by-sequencing for plant breeding and genetics. The Plant Genome, 5: 92-102.

Rutkoski, J., Singh, R.P., Huerta-Espino, J., Bhavani, S., Poland, J., Jannink, J.L. and Sorrells, M.E. 2015. Efficient use of historical data for genomic selection: a case study of stem rust resistance in wheat. The Plant Genome, 8: 1-10.

Sallam, A.H. and Smith, K.P. 2016. Genomic Selection Performs Similarly to Phenotypic Selection in Barley. Crop Science, 56: 2871-2881.

Schulz-Streeck, T., Ogutu, J.O. and Piepho, H.P. 2013. Comparisons of single-stage and two-stage approaches to genomic selection. Theoretical and applied genetics, 126: 69-82.

Singh, B.D. and Singh, A.K. 2015. Marker-assisted plant breeding: principles and practices. New Delhi: Springer.

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. Journal of dairy science, 91: 4414-4423.

Wimmer, V., Albrecht, T., Auinger, H.J., and Schön, C.C. 2012. Synbreed: a framework for the analysis of genomic prediction data using R. Bioinformatics, 28: 2086-2087.

Table 3. 1: Training populations (TP) and validation sets (VS) summary phenotypic information for grain yield and test weight measured as Mg ha^{-1} and Kg m^{-3} respectively. Including number of environments where each trait was evaluated (No. Env), number of data points for the analysis of each trait (No. Data Points), number of genotypes (No. Genotypes TP, No. Genotypes VS) and descriptive statistics of each trait based on best linear unbiased estimate including minimum (Min), average (Mean), maximum (Max). Heritability was estimated as broad sense heritability on a per plot basis (H^2) and on entry means (H^2_C) by using Cullis et al. (2006) approximation for unbalanced data for the different traits and sets evaluated.

	Training populations					
	GAWN 2008-2013		GAWN 2008-2014		GAWN 2008-2015	
No. Genotypes TP	363		395		430	
	Grain Yield	Test Weight	Grain Yield	Test Weight	Grain Yield	Test Weight
No. Env	38	38	43	43	47	47
No. Data Points	5627	3867	6189	4377	6496	4595
Min	2.45	51.49	2.49	51.33	2.41	51.24
Mean	4.31	57.26	4.38	57.18	4.33	57.21
Max	5.30	60.69	5.44	60.47	5.57	60.74
H^2	0.17	0.38	0.18	0.38	0.17	0.38
H^2_C	0.62	0.81	0.62	0.81	0.59	0.79
	Validation Sets					
	SUNWHEAT 2014		SUNWHEAT 2015		SUNWHEAT 2016	
No. genotypes VS	76		74		80	
No. Env	5	5	4	4	5	4
No. Data Points	615	615	385	375	662	484
Min	4.06	54.62	2.06	55.97	2.76	46.85
Mean	5.13	58.00	3.56	58.01	3.66	53.47
Max	5.92	60.51	4.33	62.00	4.55	57.54
H^2	0.16	0.47	0.14	0.31	0.11	0.50
H^2_C	0.52	0.83	0.40	0.63	0.44	0.83

Table 3. 2: Summary phenotypic information for in each environment where genotypes of the training population (GAWN) were evaluated for grain yield (Mg ha^{-1}) and test weight (kg m^{-3}).

Environment	Grain Yield					Test Weight				
	Mean	Min	Max	N	CV	Mean	Min	Max	N	CV
FARMERSVILLE2008	2.8	0.6	4.3	140	14.1	54.6	48.7	58.6	140	1.4
FARMERSVILLE2010	3.6	1.3	4.8	149	5.7	58.1	54.2	62.3	75	
FARMERSVILLE2011	3.8	2.5	5.0	112	9.1	59.9	57.9	62.2	56	
FARMERSVILLE2012	5.1	2.9	6.4	146	5.9	57.7	53.2	61.0	73	
FARMERSVILLE2013	3.4	1.6	5.0	124	12.8	57.3	51.7	61.4	116	0.9
FARMERSVILLE2014	5.3	3.8	6.6	111	4.7	61.0	58.3	63.0	111	0.6
KINSTON2008	4.4	1.6	7.1	140	10.3	55.0	41.7	60.7	140	1.6
KINSTON2009	3.7	2.2	4.9	136	11.8	59.5	53.8	62.8	135	0.9
KINSTON2011	4.5	2.9	5.9	111	8.5	61.8	56.5	64.3	111	0.8
KINSTON2012	5.1	1.8	7.3	141	10.0	56.4	50.3	60.1	140	1.5
KINSTON2013	4.9	2.7	7.4	124	11.5	54.8	50.1	58.2	123	1.7
MARIANNA2015	4.3	1.7	5.9	84	8.9	60.0	53.2	64.7	83	1.4
PLAINS2008	5.3	2.9	6.6	210	7.4	58.5	52.0	62.0	70	
PLAINS2009	2.4	0.7	4.0	200	12.9	53.4	48.5	57.4	68	
PLAINS2010	3.5	2.2	4.4	225	7.1	60.2	58.0	62.5	75	
PLAINS2011	6.0	4.0	7.5	168	6.2	58.9	50.7	62.7	56	
PLAINS2012	4.1	1.5	5.7	216	6.7	59.4	51.7	63.1	72	
PLAINS2013	5.6	2.0	7.9	122	10.7	58.9	54.0	62.5	57	
PLAINS2014	4.6	1.6	6.3	74	8.1	57.3	51.0	60.0	36	
PLAINS2015	2.8	1.4	5.0	78	17.3	.	.	.	0	
QUINCY2008	4.7	1.7	7.3	209	14.3	53.6	48.6	58.2	70	
QUINCY2009	3.9	1.4	6.4	203	11.2	50.5	44.5	55.4	68	
QUINCY2010	3.6	1.7	6.0	218	11.5	53.9	46.1	58.2	75	
QUINCY2011	5.3	2.5	6.6	166	10.7	58.9	53.6	62.3	56	
QUINCY2012	2.8	0.5	4.9	218	15.9	54.7	48.0	59.2	72	
QUINCY2013	3.2	0.9	5.7	185	15.1	51.8	45.2	58.5	62	
QUINCY2014	4.3	2.0	7.2	109	13.7	50.9	43.5	56.6	111	2.9
STUTTGART2008	5.3	2.5	7.6	136	11.9	53.8	46.3	58.2	115	3.0
STUTTGART2009	3.0	2.1	4.4	68		57.2	50.9	59.2	68	
STUTTGART2012	3.9	2.9	5.2	146	6.3	63.4	60.2	65.8	146	0.5
STUTTGART2013	4.1	2.2	5.9	119	13.6	56.6	51.2	62.2	123	2.3
WARSAW2008	6.2	3.5	8.3	140	6.4	58.5	52.6	61.0	140	1.0
WARSAW2009	4.9	3.5	6.3	136	5.1	57.4	54.4	60.4	135	0.7
WARSAW2010	4.7	3.7	6.0	150	8.6	61.6	58.7	63.6	150	0.5
WARSAW2011	6.2	3.7	7.7	112	6.0	60.6	56.0	63.6	112	1.0
WARSAW2012	5.3	3.6	7.3	146	8.0	62.2	59.5	64.2	146	0.4
WARSAW2013	5.7	4.6	7.0	124	6.6	59.2	55.7	62.6	124	0.6
WARSAW2014	5.5	4.4	6.6	74	7.5	58.9	56.7	60.7	74	0.5
WARSAW2015	5.8	4.6	6.7	86	5.0	58.6	55.8	60.4	86	0.5
WINNSBORO2008	4.2	0.5	7.0	137	10.9	59.2	53.7	62.3	139	1.5
WINNSBORO2009	3.6	1.2	5.4	136	10.9	56.1	49.0	59.8	136	2.2
WINNSBORO2010	3.7	1.3	4.8	150	8.6	54.0	50.1	57.7	149	1.2
WINNSBORO2011	4.2	2.6	5.4	112	11.2	57.3	52.9	61.1	112	1.1
WINNSBORO2012	2.8	1.3	4.1	146	11.2	55.7	49.4	59.9	146	1.3
WINNSBORO2013	4.8	2.8	6.2	62		56.7	42.8	60.2	60	
WINNSBORO2014	5.5	4.1	7.0	111	5.8	54.8	48.4	58.7	111	2.7
WINNSBORO2015	2.3	0.4	3.3	86	11.6	52.7	48.0	56.7	72	1.6

Table 3. 3: Minimum (Min), and average minimum (Mean) genetic distance between the genetic material evaluated in each year of the historical series used as training population (GAWN) and the genetic material from each breeding program evaluated in each validation set (SUNWHEAT 2014, 2015 and 2016). Genetic distance was measured as 1 – IBS (probability of identical by state), where values close to 0 mean individuals more closely related.

		Validation Set 2014												
		Arkansas		Florida		Georgia		Louisiana		North Carolina		Texas		
		Mean	Min	Mean	Min	Mean	Min	Mean	Min	Mean	Min	Mean	Min	Average
TP GAWN 2008-13	2008	0.23	0.17	-	-	0.11	0.06	0.20	0.15	0.19	0.04	0.20	0.15	0.19
	2009	0.24	0.17	-	-	0.19	0.06	0.21	0.15	0.22	0.17	0.20	0.15	0.21
	2010	0.24	0.18	-	-	0.16	0.06	0.15	0.22	0.21	0.14	0.21	0.16	0.19
	2011	0.24	0.21	-	-	0.16	0.14	0.21	0.19	0.21	0.14	0.21	0.19	0.21
	2012	0.22	0.13	-	-	0.14	0.05	0.18	0.14	0.21	0.11	0.18	0.14	0.19
	2013	0.23	0.13	-	-	0.15	0.05	0.18	0.12	0.21	0.14	0.19	0.14	0.19
	Average	**0.23**	**0.17**	**-**	**-**	**0.15**	**0.07**	**0.19**	**0.16**	**0.21**	**0.12**	**0.20**	**0.16**	
		Validation Set 2015												
		Mean	Min	Mean	Min	Mean	Min	Mean	Min	Mean	Min	Mean	Min	Average
TP GAWN 2008-14	2008	0.21	0.10	-	-	0.17	0.04	0.20	0.16	0.18	0.14	0.22	0.17	0.20
	2009	0.23	0.18	-	-	0.19	0.07	0.18	0.15	0.22	0.18	0.21	0.16	0.21
	2010	0.22	0.13	-	-	0.18	0.07	0.19	0.14	0.21	0.17	0.23	0.20	0.21
	2011	0.22	0.13	-	-	0.19	0.14	0.22	0.20	0.22	0.16	0.21	0.13	0.21
	2012	0.19	0.03	-	-	0.18	0.10	0.19	0.12	0.21	0.17	0.21	0.13	0.20
	2013	0.22	0.16	-	-	0.17	0.09	0.18	0.13	0.20	0.16	0.22	0.18	0.20
	2014	0.23	0.14	-	-	0.17	0.06	0.20	0.17	0.20	0.13	0.17	0.22	0.19
	Average	**0.22**	**0.12**	**-**	**-**	**0.18**	**0.08**	**0.19**	**0.15**	**0.21**	**0.16**	**0.21**	**0.17**	
		Validation Set 2016												
		Mean	Min	Mean	Min	Mean	Min	Mean	Min	Mean	Min	Mean	Min	Average
TP GAWN 2008-15	2008	0.21	0.12	0.20	0.15	0.17	0.11	0.22	0.20	0.21	0.16	0.14	0.21	0.19
	2009	0.20	0.09	0.20	0.16	0.18	0.11	0.19	0.14	0.22	0.14	0.19	0.14	0.20
	2010	0.20	0.11	0.22	0.14	0.18	0.11	0.23	0.18	0.23	0.20	0.21	0.11	0.21
	2011	0.20	0.13	0.22	0.20	0.19	0.14	0.23	0.19	0.23	0.20	0.21	0.17	0.21
	2012	0.20	0.12	0.20	0.13	0.17	0.04	0.22	0.15	0.23	0.20	0.19	0.08	0.20
	2013	0.20	0.08	0.20	0.13	0.17	0.12	0.20	0.16	0.21	0.16	0.19	0.09	0.20
	2014	0.21	0.13	0.22	0.16	0.18	0.09	0.23	0.19	0.21	0.16	0.21	0.13	0.21
	2015	0.21	0.14	0.23	0.21	0.18	0.13	0.23	0.19	0.21	0.17	0.23	0.17	0.22
	Average	**0.20**	**0.12**	**0.21**	**0.16**	**0.18**	**0.11**	**0.22**	**0.18**	**0.22**	**0.17**	**0.20**	**0.14**	

Table 3. 4: Prediction accuracy for SUNWHEAT 2014, 2015 and 2016 using a different training population with historical data for each year which simulate forward predictions in a real plant breeding program. Traits evaluated were grain yield (Yield) and test weight (TW). We evaluated three different genomic selection statistical methods for predictions including Two-step univariate analysis, Two-step multivariate analysis and one-step univariate analysis.

	Data-Curated					
	SUNWHEAT2014		SUNWHEAT2015		SUNWHEAT2016	
	Yield	TW	Yield	TW	Yield	TW
Two-step univariate	0.78	0.41	0.74	0.37	0.55	0.40
Two-step multivariate	0.77	0.41	0.73	0.37	0.57	0.42
One-step univariate	0.83	0.43	0.84	0.30	0.64	0.46
	No-Data-Curation					
	Yield	TW	Yield	TW	Yield	TW
Two-step univariate	0.78	0.41	0.45	0.24	0.55	0.36
Two-step multivariate	0.77	0.41	0.44	0.24	0.57	0.35
One-step univariate	0.83	0.43	0.52	0.15	0.64	0.28

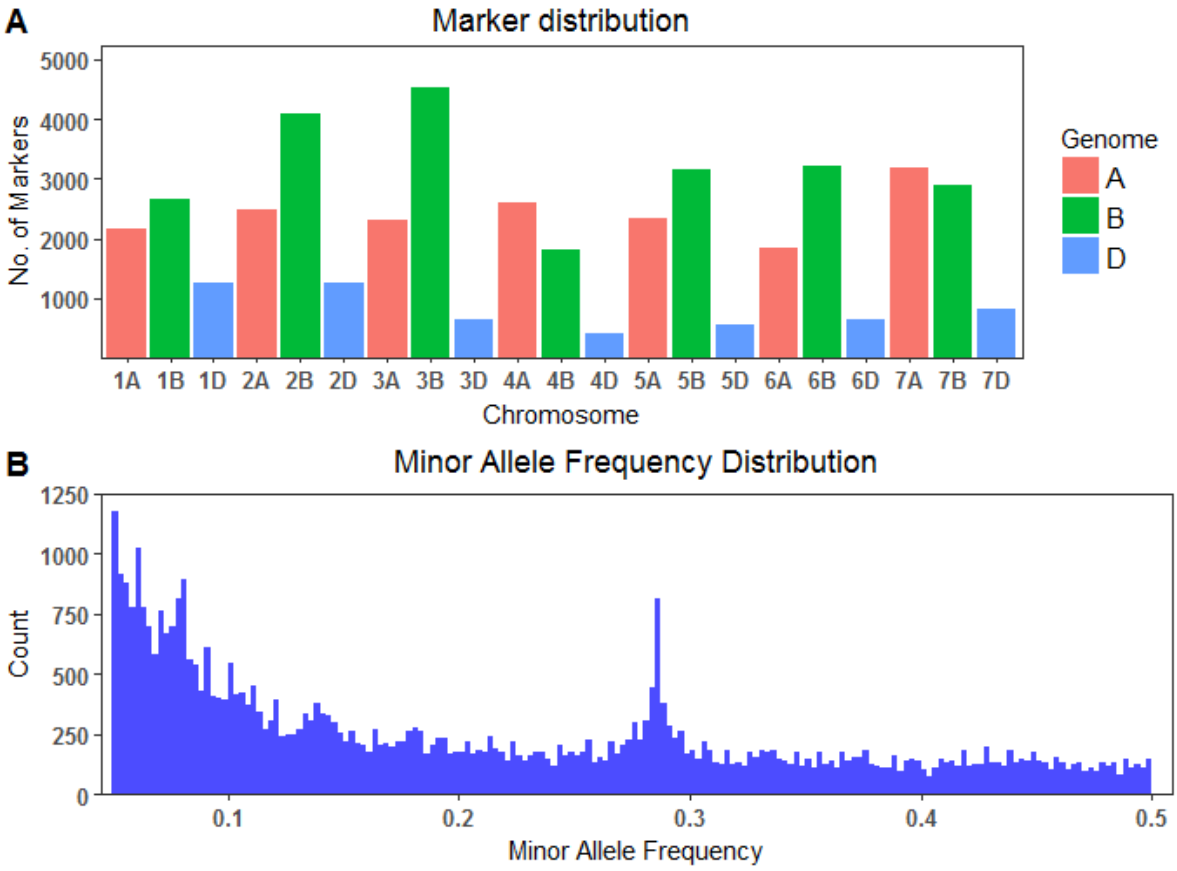


Figure 3. 1: Summary genotypic information including number of marker per chromosome (A) and minor allele frequency distribution (B) after marker filtering and imputation using the all genotypes the historical GAWN nursery from 2008 to 2015 and genotypes in SUNWHEAT from 2014 to 2016.

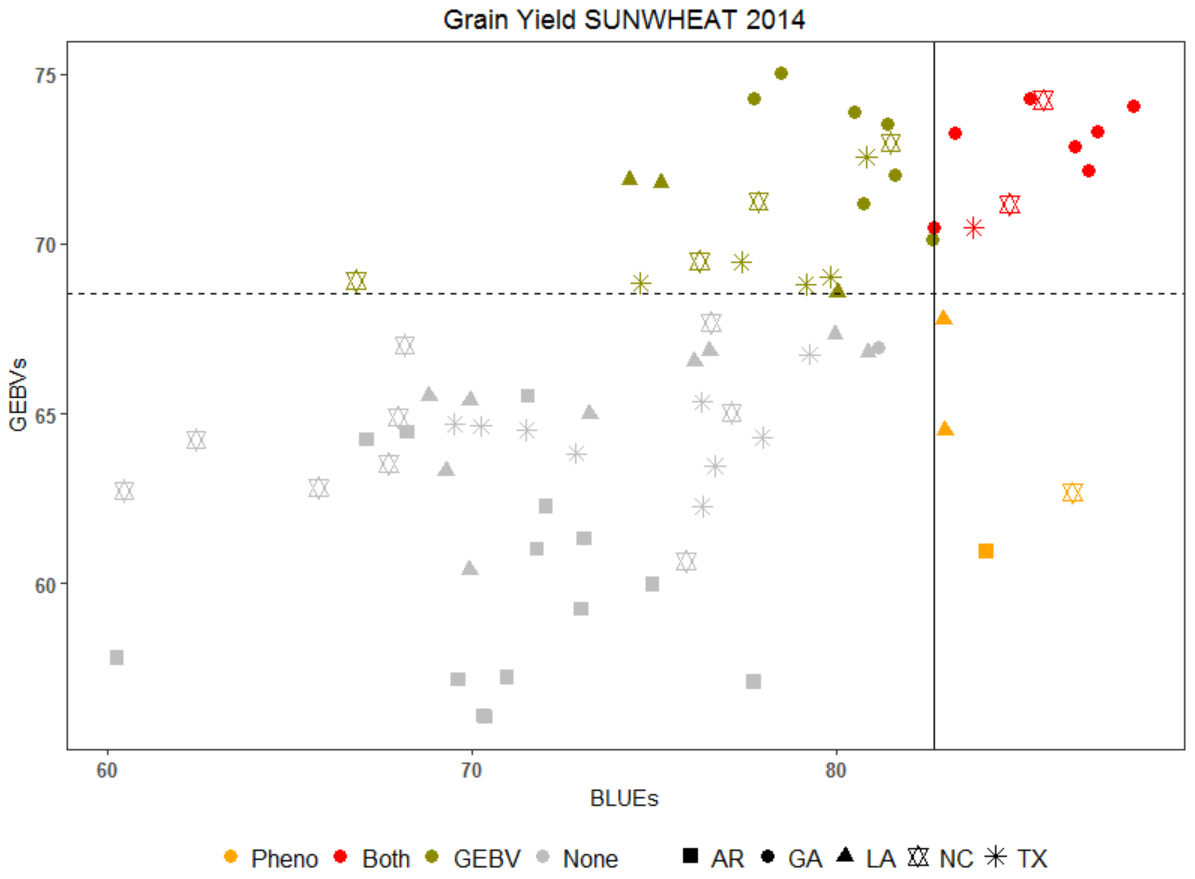


Figure 3. 2: Scatter plot of genotypes selected by applying different selection intensities when phenotypic or genomic selection is utilized for trait grain yield mesuared as Mg ha^{-1}. GEBVs are from the one step analysis using locations after data curation and BLUEs for SUNWHEAT 2014. Vertical solid line represent a cut-off selection intensity of the best 20 percent based on BLUEs for phenotypic selection. Horizontal dashed line represent a cut-off selection intensity of the best 40 percent based on GEBVs for genomic predictions. Points are color coded according to which genotypes were selected only from phenotypic selection (golden), only from genotypic selection (green), selected by phenotypic and genomic selection (red), not selected (gray) based on different selection intensities. Different shapes in the plot are the different breeding programs participating in the nursery as follow: square = University of Arkansas; circle = University of Georgia; triangle = Louisiana State University; star = North Carolina State University; asterisk = Texas A&M AgriLife Research.

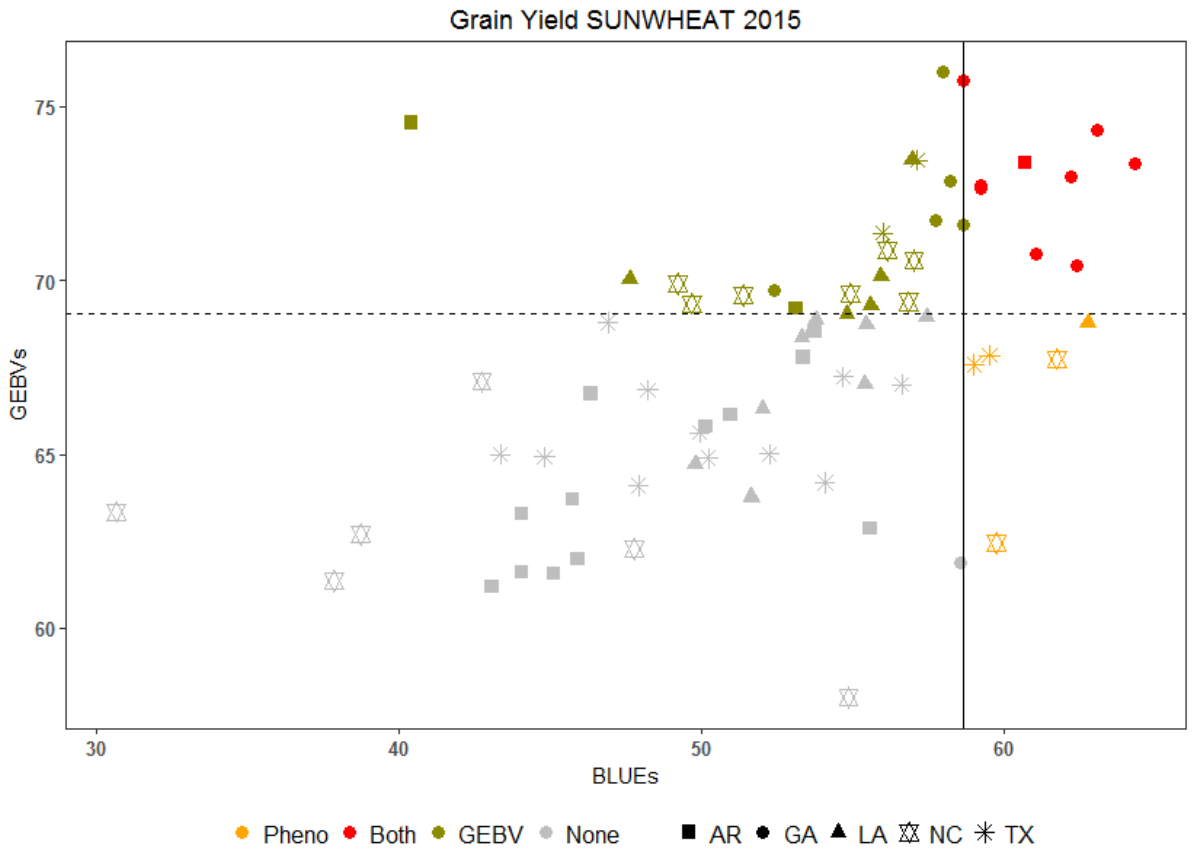


Figure 3. 3: Scatter plot of genotypes selected by applying different selection intensities when phenotypic or genomic selection is utilized for trait grain yield mesuared as Mg ha^{-1}. GEBVs are from the one step analysis using locations after data curation and BLUEs for SUNWHEAT 2015. Vertical solid line represent a cut-off selection intensity of the best 20 percent based on BLUEs for phenotypic selection. Horizontal dashed line represent a cut-off selection intensity of the best 40 percent based on GEBVs for genomic predictions. Points are color coded according to which genotypes were selected only from phenotypic selection (golden), only from genotypic selection (green), selected by phenotypic and genomic selection (red), not selected (gray) based on different selection intensities. Different shapes in the plot are the different breeding programs participating in the nursery as follow: square = University of Arkansas; circle = University of Georgia; triangle = Louisiana State University; star = North Carolina State University; asterisk = Texas A&M AgriLife Research.

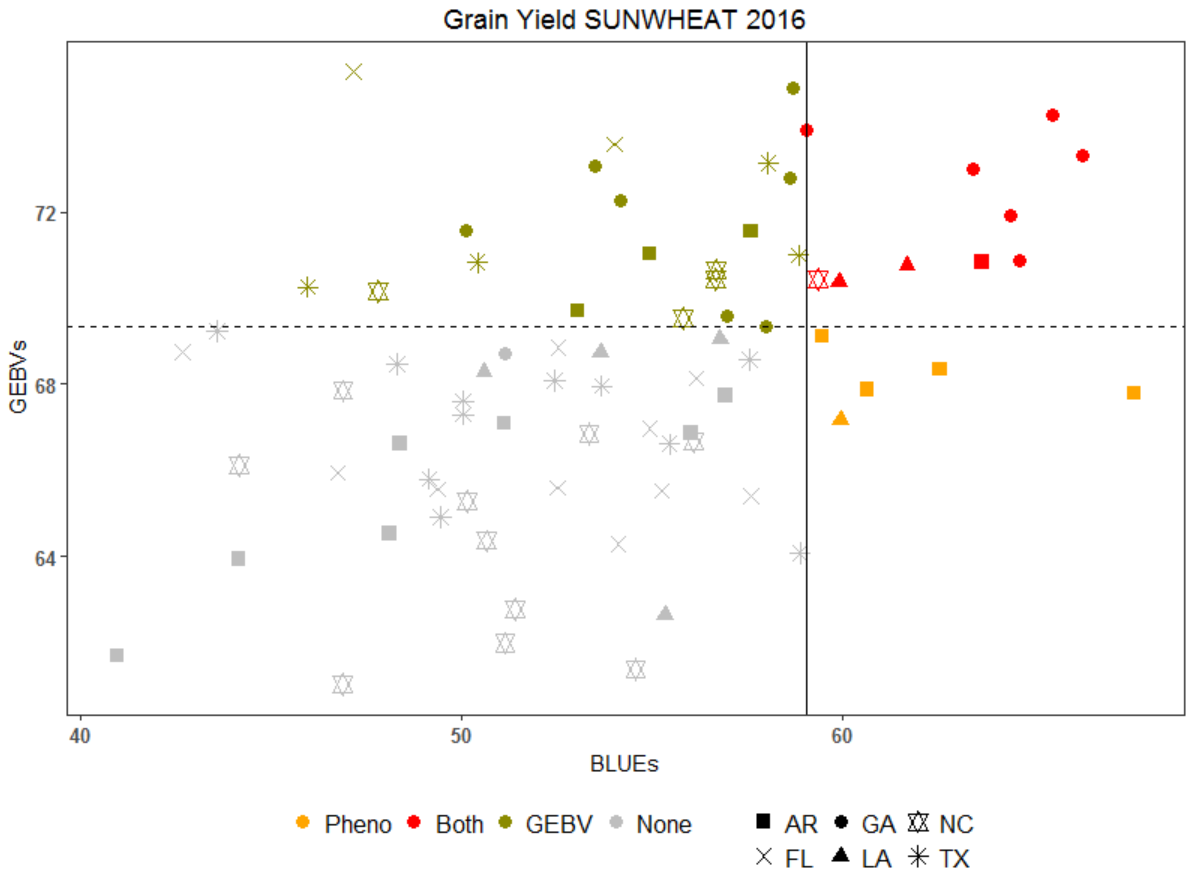


Figure 3. 4: Scatter plot of genotypes selected by applying different selection intensities when phenotypic or genomic selection is utilized for trait grain yield mesuared as Mg ha^{-1}. GEBVs are from the one step analysis using locations after data curation and BLUEs for SUNWHEAT 2016. Vertical solid line represent a cut-off selection intensity of the best 20 percent based on BLUEs for phenotypic selection. Horizontal dashed line represent a cut-off selection intensity of the best 40 percent based on GEBVs for genomic predictions. Points are color coded according to which genotypes were selected only from phenotypic selection (golden), only from genotypic selection (green), selected by phenotypic and genomic selection (red), not selected (gray) based on different selection intensities. Different shapes in the plot are the different breeding programs participating in the nursery as follow: square = University of Arkansas; x-sign = Florida, University of Florida; circle = University of Georgia; triangle = Louisiana State University; star = North Carolina State University; asterisk = Texas A&M AgriLife Research.

CHAPTER 4. Discovery and validation of powdery mildew resistance genes through association mapping in a soft red winter wheat (*Triticum aestivum* L.) panel

As prepared and submitted for consideration for special edition to the journal Theoretical and Applied Genetics on use of draft wheat genome sequence.

J. Martin Sarinelli, Keith R. Merrill, Mohammed Guedira, Priyanka Tyagi, J. Paul Murphy, James B. Holland, Jerry W. Johnson, Yuanfeng Hao, Christina Cowger and Gina Brown-Guedira

J.M. Sarinelli, M. Guedira, P. Tyagi and J.P. Murphy, Dep. Of Crop and Soil Sciences, North Carolina State University, Raleigh, NC 27695; K.R. Merrill, Monsanto Company, St. Louis, MO 63177; J.B. Holland, C. Cowger and G. Brown-Guedira, Department of Crop and Soil Sciences and USDA-ARS, North Carolina State University, Raleigh, NC 27695; J.W. Johnson and Y. Hao, Department of Crop and Soil Sciences, University of Georgia, Athens, GA 30602

Abstract

Powdery mildew, caused by the obligate biotroph *Blumeria graminis* (DC) Speer f. spp. *tritici* emend. E. J. Marchal (*Bgt*), is an important disease in USA wheat (*Triticum aestivum* L.) production regions. We performed association analysis to detect marker-trait association for powdery mildew resistance in a panel of 862 winter wheat genotypes. The germplasm originated in the soft red wheat growing regions of the eastern US, primarily from Southern and Mid-Atlantic breeding programs. Genotypes were screened in the field under natural infection in a wide range of environments and in the greenhouse using two *Bgt* isolates (Ken-2-5-B and Ken-2-5-D). A genome-wide set of more than 14,000 SNPs obtained using genotyping by sequencing was utilized to analyze population structure, family relatedness, linkage disequilibrium and marker-trait association in the population. We identified six regions associated with powdery mildew resistance located on chromosomes 1A (2), 2B, 6B and 7A (2). A stepwise method was used in the association analyses where the most significant makers were added to the models and sequential analysis performed until no additional marker was declared as significant at a false discovery rate (FDR) adjusted p-valued lower than 0.1. Loci declared as significant can be validated and incorporated into breeding populations via marker assisted selection or used to identify sources of resistant germplasm. These markers can also be utilized in genomic predictions models as covariates associated with major genes to improve the model accuracy. This study also provided an overview of average resistance and frequency of favorable alleles over time and by breeding program.

Introduction

The wheat (*Triticum aestivum* L.), foliar fungal disease powdery mildew, caused by the obligate biotrophic *Blumeria graminis* (DC) Speer f. sp. *tritici* emend. E. J. Marchal (*Bgt*), can reduce yield up to 40 percent under humid and high input conditions (Bennet, 1984, Johnson et al., 1979). Powdery mildew is an important wheat disease globally with high incidence in most wheat growing regions of Europe and Asia (Cowger et al., 2012, Morgounov et al. 2012). While powdery mildew occurs in most wheat growing regions of the USA, it is economically damaging in the states of Maryland, North Carolina, South Carolina and Virginia (Parks et al., 2009). Reported crop losses by powdery mildew on susceptible commercial cultivars Chancellor and Saluda were up to 34 percent in severe epidemics in the eastern USA (Johnson et al., 1979, Leath and Bowen, 1989).

Powdery mildew can affect all above-ground parts of a wheat, producing white to gray pustules primarily on the upper leaf surface (Murray et al 1998, Parry 1990). Plants affected by the fungus often have reduced tillering and tiller survival, kernel weight, test weight and protein content, and produce fewer kernels per head (Everts and Leath, 1992). These deficiencies may negatively impact milling and baking quality traits, including flour yield, kernel softness (Everts et al., 2001).

One effective and environmentally friendly practice implemented to control powdery mildew infections is the use of cultivars with genetic resistance (Petersen et al., 2015). To date, 58 different resistance genes have been designated from *Pm1* to *Pm58*. Loci *Pm1, Pm3, Pm4, Pm5* and *Pm24* have multiple resistance alleles. In addition, another 39 resistance genes have been temporarily designated (Bhullar et al., 2009, 2010, Hao et al., 2015, McIntosh et

al., 2013, 2014, 2016, 2017). The origin of powdery mildew resistance is diverse and includes spring and winter wheat, landraces, related species, and genera. Resistance to powdery mildew can be classified as race-specific or non-race-specific. Most of the genes deployed confer race-specific resistance expressed at all developmental stages and regulated by single major genes. These race-specific genes generally provide high levels of resistance to specific *Bgt* isolates but can be defeated in a short period, even before deployment in wheat cultivars, due to the high evolutionary potential of the pathogen (Cowger et al., 2009, Parks et al., 2008). Pyramiding multiple race-specific resistance genes has been more effective in conferring appropriate and durable levels of resistance (Singh et al., 2005). Several QTL have been identified conferring horizontal resistance, also called adult plant resistance or partial resistance, associated with non-race-specific genes. In general, partial resistance tends to be more durable than race-specific major gene resistance (Cowger et al., 2012).

Selection of resistant cultivars based on molecular makers offers the possibility of early generation selection, even in the absence of consistent infections, and allows for pyramiding multiple genes of varying effect. Several linkage mapping studies in biparental populations have successfully detected QTL associated with powdery mildew resistance in wheat (Hao et al., 2015, Petersen et al., 2015, Worthington et al., 2014). Association mapping, also known as linkage disequilibrium mapping, is an alternative approach that relies on historical recombination events. Association mapping is a useful tool for surveying germplasm collections to identify new allelic variants that confer resistance to diseases, making these studies complementary to biparental linkage mapping studies (Breseghello and

Sorrells, 2006, Myles et al., 2009, Zhu et al., 2008). Association mapping can be combined with high throughput genotyping technologies such as massively parallel SNP array-based genotyping and sequencing based technologies such as genotyping by sequencing (GBS) that combine SNP discovery and genotyping and produce dense genome coverage at a relatively low cost (Elshire et al., 2011). Alignment of sequence reads from GBS and source sequences of SNP on arrays to the draft reference sequence of the wheat genome RefSeq v1.0 developed by the International Wheat Genome Sequencing Consortium allows for translation of these polymorphisms to physical positions in the genome. Once marker-trait associations are detected, the reference genome can be used to facilitate the identification of genes of interest, examine local and genome-wide patterns of LD, and create PCR primers or sequence-specific probes for use in marker assisted selection programs. In wheat, several association mapping studies have been published for diseases of economic importance, including rusts (Macaferri et al., 2015, Turner et al., 2017, Zhang et al., 2014) and Fusarium head blight (Arruda et al., 2016, Kollers et al., 2013). However, the only report available for association mapping for powdery mildew resistance was made with a limited genome coverage and a small population of spring wheat lines derived from CIMMYT (Crossa et al., 2007).

The specific goals of the present study were (1) to carry out a genome wide scan to identify SNP related with powdery mildew resistance in winter wheat germplasm from the eastern USA, (2) to establish relationships between loci identified in this study and previously identified *Pm* genes, and (3) to study the distribution of most significant resistance

alleles across the historical time series and through Southern and Mid-Atlantic USA public wheat breeding programs.

Materials and Methods

Plant material

This study utilized two populations, referred to as Populations A and B, comprising 274 and 640 genotypes, respectively. Population A consisted of soft winter wheat landraces, breeding lines and cultivars collected from eastern USA breeding programs. These included several founding landraces as well key parental lines from the early and mid-1900s. The majority of lines in population A were modern, elite cultivars and breeding lines from public and private breeding programs in the region. Population B consisted of entries in two uniform cooperatives nurseries of soft red winter wheat, that were part of the multi-location evaluation scheme of collaborating public breeding programs, the Gulf Atlantic Wheat Nursery (GAWN) and Southern Uniform Wheat Nursery (SUNWHEAT). Genotypes included in this population included entries in the GAWN during harvest years 2008 to 2016 and entries in the SUNWHEAT nursery from 2014 to 2016. These nurseries included elite materials from public breeding programs in the Mid-Atlantic and Southern part of the eastern soft wheat production area, including Clemson University (SC), The University of Florida (FL), The University of Arkansas (AR), The University of Georgia (GA), Louisiana State University (LA), North Carolina State University (NC), Texas A&M AgriLife Research (TX) and Virginia Polytechnic Institute and State University (VA). A more detailed description of the research materials is provided in Appendix A and Appendix B.

Phenotypic data collection and analyses

The field based powdery mildew resistance screening was unbalanced because a different set of genotypes were evaluated in each environment (year, location, and nursery combination). The environments and the number of genotypes evaluated in each are summarized in S. Table 4.2. Powdery mildew resistance was assessed in the field when plants had reached Zadoks Growth Stage 39 to 50 (Zadoks et al., 1974) and symptoms had developed uniformly within the plots. A numeric 0-9 scale was adopted for scoring based on diseases severity (DS) where 0 indicated immunity, DS=0; 1, 0 < DS ≤ 10 percent; 2, 10 < DS ≤ 20 percent; 3, 20 < DS ≤ 30 percent; 4, 30 < DS ≤ 40 percent; 5, 40 < DS < 60 percent; 6, 60 ≤ DS < 70 percent; 7, 70 ≤ DS < 80 percent; 8, 80 ≤ DS < 90 percent; 9, full susceptibility, DS ≥ 90 percent (Bennett and Westcott, 1982). All phenotypic records available from field evaluations of population A and B across the complete set of environments were combined and best linear unbiased estimates (BLUE) for each genotype as well as variance components were estimated using mixed models in SAS 9.4 software (SAS Institute Inc., Cary, NC). The following linear mixed model was utilized for the analysis:

$$y_{ijk} = \mu + E_i + B(E)_{ij} + G_k + EG_{ik} + \varepsilon_{ijk}$$

where, y_{ijk} was the phenotypic observation of genotype *k* in the i^{th} environment in the j^{th} block, μ was the overall mean, E_i was the environment effect, $B(E)_{ij}$ was the block effect nested within environment, G_k was the genotypic effect, EG_{ik} was the genotype by environment interaction and ε_{ijk}, represented the residual term. For this model, the overall

mean and the genotypic effect were considered fixed and all the remaining terms random. Random effects (u) and residuals (e) were assumed to be normally and independently distributed u~IDD (0, Iσ_u^2) and e~IDD (0, Iσ_e^2). BLUE of each genotype was calculated as the estimated genotypic effect plus overall mean.

Broad sense heritability on a per plot basis was computed using a statistical model identical to the previously described with the overall mean as fixed effect and all other terms random according to Holland et al. (2003) as:

$$H^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_{ge}^2 + \sigma_e^2}$$

Where, σ_g^2 represent the genotypic variance, σ_{ge}^2 is the variance component due to genotype by environment interaction, and σ_e^2 is the variance associated with the residual term.

Greenhouse seedling resistance screening to powdery mildew was performed in December 2014 for population A and in December 2015 for population B. Experiments were planted in randomized complete block designs with three replicates for population A and two replicates for population B. Experimental units were small square pots planted with 6 seeds of a single line. Commercial check cultivars NC-Neuse, Saluda, USG3209 and Coker 68-15 were planted at 25-pot intervals along the greenhouse bench. The powdery mildew *Bgt* isolates ‘KEN 2-5-B’ and ‘KEN2-5-D’ obtained from the collection maintained by the USDA-ARS Plant Science Research Unit at North Carolina State University, were used as the inoculum sources for population A and B, respectively. The inoculum was increased on plants for the susceptible cultivar ‘Jagalene’ under greenhouse conditions prior to conducting the disease evaluation experiment. Seedlings were inoculated 15 days after planting at

Zadoks growth stages 13 to 20 (Zadoks et al., 1974) by gently shaking conidia from leaves of infected Jagalene plants onto the foliage. Disease reactions were scored 10 days after inoculation following a 0-4 rating scale, where 0 = resistant (no visible symptoms or flecks), 1 = moderately resistant (necrosis with low to medium sporulation), 2 = intermediate (medium sporulation), 3 = moderately susceptible (no necrosis with medium to high sporulation), and 4 = susceptible (no necrosis with very high sporulation). Phenotypic records of the seedling screening for each population were analyzed separately and BLUEs for each genotype as well as variance components were calculated using mixed models in SAS 9.4 software (SAS Institute Inc., Cary, NC). The following linear mixed model was used for the analysis:

$$y_{ij} = \mu + B_i + G_j + \varepsilon_{ij}$$

where, y_{ij} was the phenotypic observation of genotype *j* in the i^{th} replication, μ was the overall mean, B_i was the effect of replication, G_j was the genotypic effect and ε_{ij}, represented the residual term. For this model, the overall mean and the genotypic effect were considered fixed and the replication was considered random. The replication random effect (u) and residuals (e) were assumed to be normally and independently distributed u~IDD (0, $I\sigma_u^2$) and e~IDD (0, $I\sigma_e^2$). BLUE of each genotype was calculated as previously specified.

Broad sense heritability on a per plot basis was computed using a statistical model identical to the previously described with the overall mean as fixed effect and all other terms random according to Holland et al. (2003) as:

$$H^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}$$

where, σ_g^2 represent the genotypic variance, and σ_e^2 is the variance associated with the residual term.

Genotypic Data

Genotypes from population A and B underwent genotyping by sequencing (GBS, Elshire et al., 2011), using the protocol described by Poland et al. (2012). Briefly, DNA was extracted from 10 days old leaves using DNEasy 96 Plant Kits (Qiagen, Venlo, Netherlands). Genome complexity reduction was carried out using a combination of two enzymes *MspI* (CCGG) a common cutter and *PstI* (CTGCAG) a rare cutter and barcoded adaptors were ligated to each sample. Ninety six individual samples were pooled into a single library and polymerase chain reaction amplified. Each pooled library was sequenced in an Illumina HiSeq 2500 instrument at the Genomic Science Laboratories (North Carolina State University). After sequencing, single nucleotide polymorphisms (SNP) were scored using the Tassel5 pipeline (Glaubitz et al., 2014) using aligner method of bwa version 0.7.12 (Li and Durbin 2009) for aligning SNPs to reference sequence. The International Wheat Genome Sequencing Consortium (IWGSC) RefSeq v1.0 genome assembly (https://wheat-urgi.versailles.inra.fr/Seq-Repository/Assemblies) was used as a reference genome to align the SNP with a physical position. Polymorphic SNPs were identified and saved in a file with hapmap format. SNP names include the chromosome name and the physical distance in base pairs from the telomere of the short arm.

Only SNPs with less than 20 percent missing data, more than 5 percent minor allele frequency (MAF) and less than 10 percent of heterozygous calls per marker locus were

retained. Missing data were imputed using Beagle4 through a function in R package synbreed (Wimmer et al. 2012, Browning and Browning 2016). After imputation, SNPs with more than 10 percent heterozygosity and MAF below 5 percent were removed before further analysis. Redundant SNPs in the data set were binned based on two criteria: 1) pairwise intrachromosomal linkage disequilibrium (LD) between SNPs measured as r^2 = 1 and 2) distance between the SNPs pair less than 64 base pairs and r^2 greater or equal to 0.9. A single SNP in each bin was retained. Pairwise measures of intrachromosomal LD between pairs of SNP were estimated using Tassel 5 (Bradbury et al., 2007).

For the greenhouse seedling screening analysis of population B, 12 KASP markers associated with major QTL routinely used for marker assisted selection in southeastern breeding programs for plant height (*Rht-B1* and *Rht-D1*), vernalization (*Vrn-A1* and *Vrn-B1*), photoperiod (*Ppd-A1, Ppd-B1, Ppd-D1*), the *t1AL:1RS* and *t1BL:1RS* translocations from rye (*Secale cereale* L.), the *t2BS:2GS-2GL:2BL* translocation from *Triticum timopheevi* (Zhuk), as well as the translocation *t2AS:2NS* from *Triticum ventricosum* (Tausch) were screened and incorporated into the GBS SNPs panel.

Association mapping analysis

To avoid spurious marker-trait associations that can arise due to population structure and family relatedness, we utilized a unified linear mixed model that simultaneously accounts for population structure and genetic family relatedness called Q + K model (Yu et al., 2006). To account for population structure we included the eigenvectors of the first three principal components of the marker data set as fixed covariates, and to account for

relatedness we used the realized genomic relationship matrix (VanRaden 2008) in a mixed model framework. The linear mixed model for the analysis can be specified as follow:

$$y = X\beta + Qv + Zu + e$$

where $\boldsymbol{y}$ was a n x 1 vector of BLUE of each wheat genotype obtained for the trait, $\boldsymbol{\beta}$ was a p x 1 vector of fixed effects which included the model overall mean and the allelic effect of each marker, $\boldsymbol{v}$ was another vector of fixed effects with dimensions p' x 1 including the eigenvectors of the first three principal components and additional covariates (significant markers associated with the trait) added to the model when stepwise association analysis was performed, $\boldsymbol{u}$ was a n x 1 vector of random polygenic effects and $\boldsymbol{e}$ was a n x 1 vector for the residual terms. The design matrices $\boldsymbol{X}$ and $\boldsymbol{Q}$ were associated with the fixed effects with dimensions n x p and n x p'. $\boldsymbol{Z}$ was the design matrix associated with $\boldsymbol{u}$ with dimensions n x n. The variance-covariance structure associated with the random term was $u \sim N(0, \mathbf{K}\sigma_u^2)$ where **K** represented the realized relationship matrix. The distribution of residual effects was $\boldsymbol{e} \sim N(0, I\sigma_e^2)$.

The analysis was performed in R version 3.3.1 (R Core Team 2016) using the package genome association and prediction integrated tool (GAPIT) (Lipka et al., 2012) using a compressed mixed linear model (Zhang et al., 2010). To detect significantly associated markers of smaller effect that can be masked by QTLs with large effect, association analyses were performed in a step-wise approach. The stepwise association analysis utilizing markers associated with major effect QTL also excludes other markers in the region or on other chromosomes that are in high LD with the selected marker (Segura et al., 2012). The analysis was performed by selecting the most significant marker in each step

and sequentially adding this marker to the model until no additional marker can be declared as significant. Significant marker-trait associations were declared based on a false discovery rate (FDR) lower than 0.1 (Benjamini and Hochberg, 1995).

To determine the relationship of the marker-trait associations identified on chromosome 6B with the *Pm54* gene, linkage mapping and interval QTL mapping analysis was performed in the recombinant inbred line (RIL) population derived from the cross between AGS 2000 (PI 612956) and Pioneer brand 26R61 (26R61, PI612153) using JMP Genomics 8. Phenotypic data of powdery mildew screening of the 175 RILs included in the mapping population under natural infection in the field and in the growth chamber was extracted from Hao et al. (2015). The genetic map was constructed by Guedira et al. (2016) using a combined set of SSR markers and DArT previously used by Hao et al. (2015) and SNP markers from the Illumina iSelect 9K array for wheat described by Cavanagh et al. (2013). BLAST of the source sequences for SNP from the iSelect array was done to assigned markers to a physical position in the IWSGC RefSeq v1.0 assembly.

Results

Phenotypic summary

Disease reaction to powdery mildew in greenhouse experiments showed significant variation for each population evaluated, with mean disease scores of 3.1 and 1.8 in populations A and B, respectively (Table 4.1). A greater proportion of susceptible lines was observed in population A than population B (Fig. 4.1B, Fig. 4.1C). The isolate KEN 2-5-D

differed from KEN2-5-B in that lower reaction types were observed on susceptible check cultivars Coker68-15, Saluda and USG3209 (S. Table 4.1).

The number of genotypes evaluated in each environment varied from 39 to 577, and a wide range in mean powdery mildew severity score was observed among the 25 environments (S. Table 4.2). Powdery mildew disease score BLUEs averaged across all field environments ranged from 0 to 6.5 and were skewed toward resistance with an average of 2.5 on the 9-point scale (Table 4.1, Fig 4.1A). Broad sense heritability estimates on a per plot basis were 0.55 for the field experiments and from 0.59 to 0.61 for the greenhouse evaluations.

Linkage disequilibrium and population structure

The final number of SNPs after data curation and filtering for each population was 14,704, 14,566 and 16,519 for population A, population B and the combined population, respectively. The D genome had the lowest number of polymorphisms (Table 4.2). The average intra-chromosomal pairwise LD (r^2) was lower in population A ($r^2 = 0.37$) than population B ($r^2 = 0.43$) or the combined population ($r^2 = 0.46$). The lowest average LD value was observed on chromosome 4D, which had a low level of polymorphism. The average pairwise LD by genome was 0.42, 0.45, and 0.40 for the A, B and D genomes, respectively. The average physical distance between adjacent markers within each chromosome varied from 0.55 Mbp (chromosome 7A) to 7.51 Mbp (chromosome 4D) (Table 4.2).

The first two principal components of the marker data explained 7.3% and 4.3 % of the marker variation, respectively (Fig. 4.2). The sub-population differentiation observed

based on the first principal component was due to the presence or absence of the *t2BS:2GS-2GL:2BL* translocation chromosome derived from *Triticum timopheevii*. Twenty-nine percent of the genotypes had the translocation and the principal component was strongly correlated (r = 0.9) with a KASP assay based on SNP marker IWA8085 that is diagnostic for the translocation. No relationship was observed between the first two principal components and the geographic origin of the genotypes.

Population A. Greenhouse seedling screening *Bgt* isolate Ken-2-5-B

Population A comprised 264 genotypes and represented more variation in year of origin/release and breeding program origin. Association analysis for population A using BLUEs calculated based on the infection type reaction to *Bgt* isolate Ken-2-5-B revealed five significant loci at a FDR ≤ 0.1. One SNP was located on chromosome 2B and the other four were located on chromosome 7A. All significant associations had negative allelic effects, indicating that the minor alleles at all detected loci were associated with a reduction in powdery mildew infection (S. Fig. 4.1, S. Table 4.3). The most significant loci detected under the Q + K model, SNP 7A_724919354 (p-value= 2.8E-16) explained 28 percent of the observed variance among entry means. The frequency of 7A_724919354 was only 0.05 and lines containing the favorable allele had a mean infection type of 0.8 while the overall mean of the population was 3.1 on a scale of 0 to 4 (Table 4.3). No additional markers were significant when this SNP was utilized as a fixed covariate in a second scan of the genome.

The pairwise LD between SNP 7A_724919354 with the complete set of SNPs utilized for the marker-trait association scan indicated that it was in high LD (r^2 = 0.59) with the SNP

detected on chromosome 2B. The average LD for the other three significant markers co-located on chromosome 7A with SNP 7A_724919354 was $r^2 = 0.1$ (S. Fig. 4.2).

Population B. Greenhouse seedling screening with *Bgt* isolate Ken-2-5-D

The 559 elite lines and cultivars screened in population B from the combined cooperative wheat nurseries GAWN and SUNWHEAT from 2008 to 2016 mostly originated from Southern and Mid-Atlantic USA breeding programs. Association analysis for disease reaction when population B genotypes were exposed to *Bgt* isolate Ken-2-5-D revealed a total of 109 significant SNPs at a FDR ≤ 0.1. Significant SNPs were located across different chromosomes including 1A (36), 3A (1), 3D (1), 4D (1), 5A (2), 5B (2), 7A (62), 7B (2) and 7D (2) (S. Fig. 4.3, S. Table 4.4). The most significant SNPs detected after each step-wise analysis were 7A_724919354 (p-value= 6.3E-19), 1A_1236236 (p-value= 3.0E-10) and 1A_5149784 (p-value= 2.9E-06), explaining 10, 4 and 3 percent of the variance (Table 4.3, S. Fig. 4.3A, S. Fig. 4.3B, S. Fig. 4.3C, S. Fig. 4.3D). The average scores of the lines homozygous or heterozygous for the more resistant allele were 0.3, 0.7 and 1.4 for 7A_724919354, 1A_1236236 and 1A_5149784, respectively, compared to the overall population mean of 1.8 on the 0 to 4 scale. No line in the panel carried the favorable alleles for the three most significant markers simultaneously, while the frequency of lines with a combination of resistant alleles at two loci was below two percent.

Results of pairwise LD between the most significant markers (7A_724919354, 1A_1236236 and 1A_5149784) across the complete genome indicate that some of the many significant SNPs detected in the genome-wide scan were in high LD with the most significant markers (S. Fig. 4.4). High levels of local LD (average $r^2 > 0.72$) were observed between

marker 7A_724919354 and 29 markers located between 722 and 736 Mbp on chromosome 7A, declared significant in the initial analysis using the K+Q model (S. Fig. 4.4). In addition, high LD was observed for 7A_724919354 and SNP located on other chromosomes that were declared significant using the K+Q model. This included markers on other chromosomes including 3D_170070173, 5A_103749173, 7B_744996923, 7B_748713164 and 7D_634582986 that had an average pairwise LD of 0.87 with 7A_724919354. When 7A_724919354 was used a covariate in the model, these markers were no longer significantly associated with resistance (S. Fig. 4.3B). We also found that significant SNP 1A_1236236 was in high LD with other SNPs in chromosome 1A that were declared significant using the Q+K+7A_724919354 model (S. Fig. 4.4). The exception was SNP 1A_5149784 that was significant in the Q+K+7A_724919354+1A_1236236 model.

Combined Population. Field screening under natural infection

Association analysis using BLUEs of 862 different genotypes from field evaluation for powdery mildew under natural infections using Q + K mixed model and sequential addition of fixed covariates identified 61 significant SNPs at FDR $\leq$ 0.1 (Fig. 4.3A, Fig. 4.3B, Fig. 4.3C, Fig. 4.3D). Significant markers were on chromosomes 2A (3), 2B (9), 3A (5), 3D (2), 4D (1), 5A (1), 5B (1), 6B (8), 7A (28), 7B (2), 7D (1) (S. Table 4.5).

The final stepwise association model contained markers 7A_724919354 (p-value= 2.6E-15), 6B_695007016 (p-value= 2.8E-14), 7A_726648444 (p-value= 4.9E-07), and 2B_717237729 (p-value= 2.6E-06), associated with 5, 4, 2 and 1 percent of the total phenotypic variance, respectively, after accounting for population structure and family relatedness (Table 4.3). The largest allele effects were observed for 7A_724919354 and

6B_695007016. Average powdery mildew score for genotypes having at least one favorable allele of the four most significant makers were 0.93, 1.89, 2.21 and 2.19 for SNPs 7A_724919354, 6B_695007016, 7A_726648444 and 2B_717237729, respectively. No genotype in the panel carried favorable alleles for the four SNPs simultaneously and only eight genotypes out of 862 carried alleles of the two SNP of largest effect, 7A_724948361 and 6B_695007016, with a mean disease score of 0.92.

Varying patterns of LD were observed in the regions harboring the four most significant SNPs from the field evaluation of powdery mildew (Fig. 4.4). In the combined panel, the average pairwise LD between SNP 7A_724919354 and other 25 SNPs detected in a region between 711 and 736 Mbp in chromosome 7A was $r^2 = 0.89$. In addition, SNP 7A_724919354 was in high LD (average $r^2 = 0.84$) with markers 5A_103749173, 7B_744996923, 7B_748713164, and 7D_634582986 that were declared significant in the initial genome scan. When 7A_724919354 was added to the model as a covariate, significant effects were no longer observed for these markers (Fig. 4.3B). In contrast, SNP 7A_726648444 was perfectly correlated with SNP 7A_726681216 ($r^2 = 1$) but had was in low LD with the other significant markers in the 711 - 736 Mbp region of chromosome 7A than nearby marker 7A_724919354 (Fig. 4.4).

SNP 6B_695007016 and the other seven significant markers detected in the surrounding region between 695 and 697 Mbp were in high LD ($r^2 = 0.89$). These markers were no longer significant when 6B_695007016 was added as fixed covariate in the model (Fig. 4.3C). Mean LD between SNP 2B_717237729 and the other eight significant markers

on chromosome 2B was only r^2 = 0.37, but only SNP 2B_717237729 was included in the final model (Fig. 4.3D, Fig. 4.3E).

Frequency of resistant alleles

The mean powdery mildew score under field screening for lines developed in different time periods revealed an increase in the level of resistance in more current materials (Fig. 4.5A). In addition, a concurrent increase in the frequency of favorable alleles of the four most significant SNP from the field evaluation (7A_724919354, 6B_695007016, 7A_726648444 and 2B_717237729) was observed in modern elite genotypes evaluated or released after 1990 (Figure 4.5A).

Average powdery mildew field score across eight breeding programs from Southern and Mid-Atlantic States varied (Figure 4.5B). Mean diseases scores were lower in particular for the breeding programs from the Mid-Atlantic states (North Carolina State and Virginia) with an average of 1.7, while for the remaining Southern states, the mean disease score was 2.9 in a scale from 0-9. The cumulative frequency of the four favorable resistance alleles for each breeding program also varies, with highest frequency in lines from Georgia (1.17), followed by North Carolina (0.91) and Virginia (0.98), and lowest frequency in lines from Arkansas (0.58). The frequency of resistant alleles was negatively correlated (r = -0.59) with the mean level of resistance of each breeding program. The frequency of favorable alleles for SNP 2B_717237729 was the highest across breeding programs, however, the proportion of variance explained by this marker was the lowest. On the other hand, the frequency of favorable alleles for SNP 7A_724919354, associated with more phenotypic variance than any other SNP, was above twenty percent only in North Carolina and Georgia and less than 5

percent in all other programs. The frequency of the favorable allele for SNP 7A_726648444 was similar across all breeding programs with an average frequency of favorable alleles of 0.26 while for SNP 6B_695007016 the highest frequency of favorable alleles was observed in the South Carolina breeding program followed by Georgia.

Discussion

We performed association analysis to identify new sources of resistance to *Bgt* that can potentially facilitate the incorporation of new sources of resistance into elite germplasm. In this study, a total of 862 winter wheat genotypes from a diverse set of genetic material (population A) and two elite cooperative nurseries (population B) were analyzed. The germplasm primarily originated in breeding programs from the Southern and Mid-Atlantic USA regions. Genotypes were screened in the field under natural infection and in the greenhouse using two different *Bgt* isolates (Ken-2-5-B and Ken-2-5-D).

Marker data revealed the presence of two sub-populations correlated with the *t2BS:2GS-2GL:2BL* translocation derived from *Triticum timopheevii*, but no additional population stratification was found based on geographic origin of the genotypes.

The total number of significant SNPs associated with powdery mildew resistance at a FDR ≤ 0.1were 144. The six most significant markers located on chromosomes 1A, 2B, 6B and 7A were analyzed in detail. These QTL regions were potentially co-located with major powdery mildew resistance genes previously described. SNP 7A_724919354 was significant when genotypes were challenged with the isolates Ken-2-5-B and Ken-2-5-D and under natural infection in the field. SNPs 1A_1236236, and 1A_5149784 were significant when the

genotypes were challenged with the isolate Ken-2-5-D. SNPs 2B_717237729, 6B_695007016 and 7A_726648444 were significant when the analysis was performed using data from the field. These markers were selected to be the most significant markers detected in each step in the sequential analysis performed until no additional marker were declared as significant at a FDR lower than 0.1. The stepwise approach utilized in this study helped to detect significant markers with smaller effects that can be masked by large QTL effects. In addition, this approach excluded markers in high LD with the marker of major effect in the QTL region (Segura et al., 2012) or with unlinked QTL regions (e.g., markers 5A_103749173, 7B_744996923, 7B_748713164, 7D_634582986 that were in high LD with 7A_724919354).

The GBS SNP 1A_1236236 was significantly associated with resistance in greenhouse screening with the isolate Ken-2-5-D. This SNP was in high LD (r^2=0.86) with a KASP marker based on iSelect SNP marker IWA8035 that was also significant in our association analysis (p-value= 2.04E-07). The KASP assay is used to predict the presence of the translocation *t1RS:1AL* from rye (*Secale cereale* L.) in soft wheat germplasm (Brown-Guedira unpublished). Long range LD was observed between 1A_1236236 and other SNP distributed across the chromosome arm, as would be expected for markers associated with a whole arm alien translocation (S. Fig 4). The translocation *t1RS:1AL* is known to carry powdery mildew resistance gene *Pm17* (Heun et al., 1990). All lines having the favorable 1A_1236236 allele had the low infection types expected of lines having *Pm17* when challenged with an avirulent isolate. Based on powdery mildew reaction, 1A_1236236 was more predictive of the presence of *t1RS:1AL* in population B than KASP assay IWA8035.

Markers associated with *Pm17* were not detected in the field analysis as this gene no longer confers resistance to powdery mildew at most locations in the southeastern United States (Parks et al., 2008).

When markers 7A_724919354 and 1A_1236236 that were associated with major gene effects were added to the mixed model, an additional maker on chromosome 1A associated with resistance to Ken-2-5-D was detected. Marker 1A_5149784 was not in LD with 1A_1236236 (r^2=7.18E-04) suggesting the presence of another locus for resistance not associated with the translocation *t1RS:1AL*. BLAST analysis of the sequence of the cloned *Pm3* gene against the IWGSC RefSeq v1.0 places the *Pm3* locus on the short arm of chromosome 1A between 4,497,769 and 4,501,210 bp. The proximity of 1A_5149784, located at 5,149,784 bp on the chromosome, with the *Pm3* locus suggests that this marker may be detecting resistance to KEN-2-5-D conferred by one of the *Pm3* alleles. No markers in the *Pm3* region were detected in the field analysis.

SNP 6B_695007016, significantly associated with powdery mildew resistance in the field (p-value = 2.82E-14), was found to be in close proximity to the QTL region associated with the *Pm54* (Hao et al., 2015). Our analysis of the phenotypic data from Hao et al. (2015) using the map of Guedira et al. (2016) that integrates SNP markers from the 9K iSelect array for wheat allows a direct comparison of the location of the QTL from the previous study and the GBS SNP markers. Using the data from growth chamber screening from Hao et al. (2015), we located the *Pm54* locus on the linkage map of the AGS 2000 x Pioneer brand 26R61 recombinant inbred line population (Fig. 4.6). *Pm54* was located under the chromosome 6B QTL peak in all four field environments. The iSelect SNP IWA 6754

located on chromosome 6B at 696,149,546 bp in the IWGSC RefSeq v1.0 was 0.8 cM proximal to *Pm54*. Eight SNP in high LD and located on chromosome 6B between 695,007,016 and 696,144,281 bp were significantly associated with resistance in our association mapping analysis (S. Table 4.5).

SNP 2B_717237729 was significant in the analysis of field information once markers associated with major genes were added to the model as fixed covariates. This locus had a relatively small effect, explaining one percent of the total phenotypic variance. However, the frequency of the favorable allele at this locus was relatively high (0.29). Hao et al. (2015) identified a second QTL in the AGS 2000 x Pioneer brand 26R61 recombinant inbred line mapping population on the long arm of chromosome 2B for resistance to powdery mildew in the field during 2012 and 2013 in North Carolina. Flanking iSelect SNP markers IWA2678 and IWA3175 were located at 699,109,204 and 715,029,506 bp, respectively, in the IWGSC RefSeq v1.0 (data not shown). Although SNP 2B_717237729 was significant in our association analysis and is located nearby, this marker was not polymorphic between AGS 2000 and Pioneer brand 26R61. In chromosome 2B, near or in the region were the most significant markers from our genome scan were detected three formerly named genes (*Pm6, Pm33, Pm51*) and three temporarily named powdery mildew resistance genes (*MlZec1, MlAB10, PmJM22, MlLX99*) were identified (Maxwell et al., 2010, Zhan et al., 2014, Zhao et al., 2013, Zhu et al., 2005). The powdery mildew resistance gene *Pm6* is located in *t2BS:2GS-2GL:2BL* translocation derived from *Triticum timopheevii* (Jorgensen and Jensen, 1973). It is unlikely that SNP 2B_717237729 is associated with the *Pm6* gene because the correlation between 2B_717237729 and the KASP marker IWA8085 that resides in the

translocation was 0.37. It is likely that our association mapping analysis detected an important region conferring moderate resistance to powdery mildew in eastern soft winter wheat in the long arm of chromosome 2B. However, its relationship with the resistance genes other than *Pm6* is uncertain.

Our genome-wide scan revealed several significant marker-trait associations in the long arm of chromosome 7A in a region between 711 and 736.5 Mbp. There were markers located within that region significantly associated with resistance at different stages of our step-wise analysis. These markers exhibited different allele frequencies and allele effects, suggestive of the presence of multiple genes or alleles in this region conferring resistance. It is known that the long arm of chromosome 7A is a gene-cluster region conferring powdery mildew resistance. The *Pm1* gene with five designated alleles (a-e), *Pm9, Pm37, MlAG12* were located within that region (Hsam et al., 1998, Maxwell et al., 2009, Perugini et al., 2008, Schneider et al., 1991, Singrün et al., 2003). SNPs identified in the chromosome 7A in this study are potentially linked to genes in this cluster.

SNP 7A_724919354 was identified as the most significant SNP for powdery mildew resistance for the three screenings performed in this study. The proportion of variance explained by the marker was 28, 10 and 5.5 percent when the genotypes were challenged with the isolates Ken-2-5-B, Ken-2-5-D, and under natural field infections, respectively. The favorable resistant allele from SNP 7A_724919354 was present in 10 percent of lines overall, including the cultivars NC-Neuse and SS8641 thought to carry an allele of the *Pm1* resistance gene (Murphy et al., 2004). Three iSelect SNP markers co-segregated with the highly effective powdery mildew resistance gene from NC-Neuse in the AGS 2000 x NC-

Neuse recombinant inbred line mapping population: IWB1164 located in chromosome 7A at 725,433,740 bp, IWA501 located in chromosome 7A at 726,483,490 bp and IWB72104 located in 7A at 736,520,902 bp (Petersen et al., prepared for submission). The proximity of 7A_724919354 and other markers in high LD significantly associated with a powdery mildew resistance gene of major effect in this region of 7A suggests that these SNP are associated with the allele of *Pm1* present in NC-Neuse and SS8641. Pairwise LD analysis between SNP 7A_724919354 and other SNP in the region between 722 and 730 Mbp of 7A reveal the presence of a long range LD block that could be related with introgression of an alien segment containing resistance genes from wild relatives species (Fig. 4.4). The region surrounding the *Pm1* gene-cluster has suppressed recombination associated with an alien introgression (Liang et al., 2016, Neu et al., 2002).

SNP 7A_726648444 was significant only in the analysis of field data once markers associated with other major genes were added to the model (7A_724919354 and 6B_695007016). This marker is located in the same gene-cluster region harboring the *Pm1* gene as marker 7A_724919354. However, when both markers were compared we found several differences related with the frequency of the favorable allele, proportion of variance explained by the marker, allelic effect and LD between the markers (Table 4.3). The average powdery mildew score under field evaluations for genotypes carrying at least one allele of SNP 7A_726648444 was 2.21 while for marker 7A_724919354 was 0.93. Only eight genotypes carried at least one copy of the favorable alleles at each marker simultaneously with an average disease score of 0.69. This suggest the presence of two different sources of resistance or different allelic variants located in the same region.

Individual powdery mildew resistance genes have been overcome even before release in commercial cultivars due to the high evolutionary potential of *Bgt* population (Cowger et al. 2009). Wheat breeders have been successful in selecting for powdery mildew resistance cultivars over time based on field evaluation, in particular in the Southern and Mid-Atlantic regions of the USA, where the disease is frequent. This was demonstrated by the decrease in average field disease score over time and increase in the frequency of favorable alleles of SNP selected (Fig. 4.5A). However, more focus will be required for deploying combinations of resistance genes in the same cultivar as we identified very few lines having multiple SNP alleles associated with resistance. The strategy of discovery new sources of resistance and pyramiding different resistance genes can improve cultivar longevity because the pathogen must defeat multiple genes to cause disease. Selection of elite lines with more than one resistance gene can be facilitated with the assistance of molecular markers tightly linked to resistance genes (Liu et al., 2000).

The lowest mean disease score and the presence of several resistance alleles at intermediate frequency in Mid-Atlantic breeding programs was due to the effort made by plant breeders to incorporate new sources of resistance into the elite germplasm. The North Carolina State University wheat breeding program have long tradition of releasing resistant cultivars with resistance genes introgressed from different relatives (Murphy et al., 1998, Murphy et al., 1999, Murphy et al., 2002, Murphy et al., 2004, Murphy et al., 2007, Navarro et al., 2000, Petersen et al., 2015, Worthington et al., 2014).

Conclusion

Association analysis performed in a set of diverse and elite germplasm combining field screening with evaluations of specific isolates of *Bgt* allowed us to simultaneously explore the composition of the genetic resistance in the population, as well as validation of marker-trait associations in related populations. In this study, we identified six regions significantly associated with resistance to powdery mildew located on chromosome 1A (2), 2B, 6B and 7A (2). Markers found in this study could be associated with previously reported resistance genes/QTL and the surrounding sequences could be used for development of highly predictive assays. In addition, we found that most of the resistant germplasm evaluated contained only a single major resistance allele. The identification of markers associated with different QTL conferring resistance to powdery mildew can be utilized in marker assisted selection programs for pyramiding multiple resistance genes of varying effect in elite wheat cultivars. Significant markers associated with resistance to powdery mildew also have potential to be utilized as fixed covariates in genomic selection model to improve model prediction accuracy. Finally, as in other association analysis, additional validation will be required to confirm the true association between the marker detected in the analysis and the causal genes.

References

Arruda, M.P., Brown, P., Brown-Guedira, G., Krill, A.M., Thurber, C., Merrill, K.R., Foresman, B.J. and Kolb, F.L. 2016. Genome-Wide Association Mapping of Fusarium Head Blight Resistance in Wheat using Genotyping-by-Sequencing. The Plant Genome, 9: 1-14.

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

Breseghello, F. and Sorrells, M.E. 2006. Association analysis as a strategy for improvement of quantitative traits in plants. Crop Science, 46: 1323-1330.

Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological) 289–300.

Bennett, F.G. and Westcott B. 1982. Field assessment of resistance to powdery mildew in mature wheat plants. Plant Pathology 31: 261-268.

Bennett, F.G., 1984. Resistance to powdery mildew in wheat: a review of its use in agriculture and breeding programmes. Plant pathology, 33: 279-300.

Bhullar, N.K., Street, K., Mackay, M., Yahiaoui, N. and Keller, B. 2009. Unlocking wheat genetic resources for the molecular identification of previously undescribed functional alleles at the Pm3 resistance locus. Proceedings of the National Academy of Sciences, 106: 9519-9524.

Bhullar, N.K., Zhang, Z., Wicker, T. and Keller, B. 2010. Wheat gene bank accessions as a source of new alleles of the powdery mildew resistance gene Pm3: a large scale allele mining project. BMC plant biology, 10: 88.

Cavanagh, C.R., Chao, S., Wang, S., Huang, B.E., Stephen, S., Kiani, S., Forrest, K., Saintenac, C., Brown-Guedira, G.L., Akhunova, A. and See, D. 2013. Genome-wide

comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. Proceedings of the national academy of sciences, 110: 8057-8062.

Cowger, C., Parks, R. and Marshall, D. 2009. Appearance of powdery mildew of wheat caused by *Blumeria graminis* f. sp. *tritici* on *Pm17*-Bearing cultivars in North Carolina. Plant Disease, 93: 1219-1219.

Cowger, C., Miranda, L., Griffey, C., Hall, M., Murphy, J.P., and Maxwell, J. 2012. Wheat powdery mildew. Disease resistance in wheat. CABI, Oxfordshire, 84-119.

Crossa, J., Burgueno, J., Dreisigacker, S., Vargas, M., Herrera-Foessel, S.A., Lillemo, M., Singh, R.P., Trethowan, R., Warburton, M., Franco, J. and Reynolds, M. 2007. Association analysis of historical bread wheat germplasm using additive genetic covariance of relatives and population structure. Genetics, 177: 1889-1913.

Everts K.L., and Leath S. 1992. Effect of early season powdery mildew on development, survival, and yield contribution of tillers of winter wheat. Phytopathology 82:1273-1278.

Everts, K.L., Leath, S. and Finney, P.L. 2001. Impact of powdery mildew and leaf rust in milling and baking quality of soft red winter wheat. Plant Disease 85: 423-429.

Glaubitz, J.C., Casstevens, T.M., Lu, F., Harriman, J., Elshire, R.J., Sun, Q. and Buckler, E.S. 2014. TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. PloS one, 9: 1-11.

Guedira, M., Xiong, M., Hao, Y.F., Johnson, J., Harrison, S., Marshall, D. and Brown-Guedira, G. 2016. Heading Date QTL in Winter Wheat (Triticum aestivum L.) Coincide with Major Developmental Genes VERNALIZATION1 and PHOTOPERIOD1. PloS one, 11: 1- 21.

Hao, Y., Parks, R., Cowger, C., Chen, Z., Wang, Y., Bland, D., Murphy, J.P., Guedira, M., Brown-Guedira, G. and Johnson, J. 2015. Molecular characterization of a new

powdery mildew resistance gene *Pm54* in soft red winter wheat. Theoretical and Applied Genetics, 128: 465-476.

Hsam, S.L.K., Huang, X.Q., Ernst, F., Hartl, L. and Zeller, F.J. 1998. Chromosomal location of genes for resistance to powdery mildew in common wheat (*Triticum aestivum* L. em Thell.). 5 Alleles at the Pm1 locus. Theoretical and Applied Genetics, 96: 1129-1134.

Heun, M., Friebe, B. and Bushuk, W. 1990. Chromosomal location of the powdery mildew resistance gene of Amigo wheat. Phytopathology, 80: 1129-1133.

Holland, J.B., Nyquist, W.E. and Cervantes-Martínez, C.T. 2003. Estimating and interpreting heritability for plant breeding: an update. Plant breeding reviews, 22: 9-112.

Holland, J.B. 2007. Genetic architecture of complex traits in plants. Current opinion in plant biology, 10: 156-161.

Johnson, J.W., Baenziger, P.S., Yamazaki, W.T. and Smith, R.T. 1979. Effects of powdery mildew on yield and quality of isogenic lines of 'Chancellor' wheat. Crop Science 19: 349-352.

Jorgensen, J. and Jensen, C.J. 1973. Gene *Pm6* for resistance to powdery mildew in wheat. Euphytica, 22: 423-423.

Kollers, S., Rodemann, B., Ling, J., Korzun, V., Ebmeyer, E., Argillier, O., Hinze, M., Plieske, J., Kulosa, D., Ganal, M.W. and Röder, M.S. 2013. Whole genome association mapping of Fusarium head blight resistance in European winter wheat (*Triticum aestivum* L.). PLoS One, 8: 1-10.

Leath, S. and Bowen, K.L. 1989. Effects of powdery mildew, triadimenol seed treatment, and triadimefon foliar sprays on yield of winter wheat in North Carolina. Phytopathology, 79: 152-155.

Liang, J., Fu, B., Tang, W., Khan, N.U., Li, N. and Ma, Z. 2016. Fine mapping of two wheat powdery mildew resistance genes located at the cluster. The plant genome, 9: 1-9.

Lipka, A.E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P.J., Gore, M.A., Buckler, E.S. and Zhang, Z. 2012. GAPIT: genome association and prediction integrated tool. Bioinformatics, 28: 2397-2399.

Liu, J., Liu, D., Tao, W., Li, W., Wang, S., Chen, P., Cheng, S. and Gao, D. 2000. Molecular marker-facilitated pyramiding of different genes for powdery mildew resistance in wheat. Plant Breeding, 119: 21-24.

Maccaferri, M., Zhang, J., Bulli, P., Abate, Z., Chao, S., Cantu, D., Bossolini, E., Chen, X., Pumphrey, M. and Dubcovsky, J. 2015. A genome-wide association study of resistance to stripe rust (*Puccinia striiformis* f. sp. *tritici*) in a worldwide collection of hexaploid spring wheat (*Triticum aestivum* L.). G3: Genes| Genomes| Genetics, 5: 449-465.

Maxwell, J.J., Lyerly, J.H., Srnic, G., Parks, R., Cowger, C., Marshall, D., Brown-Guedira, G. and Murphy, J.P. 2010. : A Subsp. Derived Powdery Mildew Resistance Gene Identified in Common Wheat. Crop science, 50: 2261-2267.

McIntosh, R.A., Dubcovsky, J., Rogers W.J., Morris, C. and Xia, X.C. 2017. Catalogue of gene symbols for wheat: 2017 supplement.

McIntosh, R.A., Dubcovsky, J., Rogers W.J., Morris, C., Appels, R. and Xia, X.C. 2016. Catalogue of gene symbols for wheat: 2015-2016 supplement

McIntosh, R.A., Dubcovsky, J., Rogers, W.J., Morris, C., Appels, R., Xia, X.C. 2014. Catalogue of gene symbols for wheat: 2013-2014 supplement. Annual Wheat Newsletter 60:153-175.

McIntosh, R.A., Yamazaki, Y., Dubcovsky, J., Rogers, W.J., Morris, C., Appels, R., Xia, X.C. 2013. Catalogue of gene symbols for wheat. In: Ogihara Y (ed) Proceeding of the 12th international wheat genetics symposium, Yokohama, Japan, 8-13 Sept 2013.

Myles, S., Peiffer, J., Brown, P.J., Ersoz, E.S., Zhang, Z., Costich, D.E. and Buckler, E.S. 2009. Association mapping: critical considerations shift from genotyping to experimental design. The Plant Cell, 21: 2194-2202.

Morgounov, A., Tufan, H.A., Sharma, R., Akin, B., Bagci, A., Braun, H.J., Kaya, Y., Keser, M., Payne, T.S., Sonder, K. and McIntosh, R. 2012. Global incidence of wheat rusts and powdery mildew during 1969–2010 and durability of resistance of winter wheat variety Bezostaya 1. European journal of plant pathology, 132: 323-340.

Murphy, J.P., Leath, S., Huynh, D., Navarro, R.A. and Shi, A. 1998. Registration of NC96BGTD1, NC96BGTD2 and NC96BGTD3 wheat germplasm resistant to powdery mildew. Crop Sci, 38:570-571.

Murphy, J.P., Leath, S., Huynh, D., Navarro, R.A. and Shi, A. 1999. Registration of NC96BGTA4, NC96BGTA5 and NC96BGTA6 wheat germplasm. Crop Sci, 39:883-884.

Murphy, J.P., Leath, S., Huynh, D., Navarro, R.A. and Shi, A. 1999. Registration of NC97BGTD7 and NC97BGTD8 wheat germplasms resistant to powdery mildew. Crop Sci, 39:884-885.

Murphy, J.P., Navarro, R.A.and Leath, S. 2002. Registration of NC99BGTAG11 wheat germplasm resistant to powdery mildew. Crop Sci, 42:1382.

Murphy, J.P., Navarro, R.A., Leath, S., Bowman, D.T., Weisz, P.R., Ambrose, L.G., Pate, M.H. and Fountain, M.O. 2004. Registration of 'NC-Neuse' wheat. Crop Sci, 44:1479-1480.

Murphy, J.P., Navarro, R.A., Marshall, D., Cowger, C., Cox, T.S., Kolmer, J.A., Leath, S. and Gaines, C.S. 2007. Registration of NC06BGTAG12 and NC06BGTAG13 powdery mildew- resistant wheat germplasm. Journal of Plant Registrations 1:75.

Murray, T. D., Parry, D.W. and Cattlin, N.D. 1998. A color handbook of diseases of small grain cereal crops, Iowa State University Press, Ames.

Navarro, R.A., Murphy, J.P., Leath, S. and Shi, A. 2000. Registration of NC97BGTAB9 and NC97BGTAB10 wheat germplasm lines resistant to powdery mildew. Crop Sci, 40:1508- 1509.

Neu, C., Stein, N. and Keller, B. 2002. Genetic mapping of the *Lr20-Pm1* resistance locus reveals suppressed recombination on chromosome arm 7AL in hexaploid wheat. Genome, 45: 737-744.

Parks, R., Carbone, I., Murphy, J.P., Marshall, D. and Cowger, C. 2008. Virulence structure of the eastern US wheat powdery mildew population. Plant Disease, 92: 1074-1082.

Parks, R., Carbone, I., Murphy, J.P. and Cowger, C. 2009. Population genetic analysis of an eastern US wheat powdery mildew population reveals geographic subdivision and recent common ancestry with UK and Israeli populations. Phytopathology, 99: 840-849.

Parry, D.W. 1990. Diseases of small grain cereals. In: Plant Pathology in Agriculture. Cambridge University Press, Cambridge, pp. 160-224.

Perugini, L.D., Murphy, J.P., Marshall, D. and Brown-Guedira, G. 2008. *Pm37*, a new broadly effective powdery mildew resistance gene from *Triticum timopheevii*. Theoretical and Applied Genetics, 116: 417-425.

Petersen, S., Lyerly, J.H., Worthington, M.L., Parks, W.R., Cowger, C., Marshall, D.S., Brown-Guedira, G. and Murphy, J.P. 2015. Mapping of powdery mildew resistance gene *Pm53* introgressed from *Aegilops speltoides* into soft red winter wheat. Theoretical and applied genetics, 128: 303-312.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

SAS Institute Inc. 2012. SAS/STAT® 12.1 User's Guide. Cary, NC: SAS Institute Inc.

Segura, V., Vilhjálmsson, B.J., Platt, A., Korte, A., Seren, Ü., Long, Q. and Nordborg, M. 2012. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nature genetics, 44: 825-830.

Schneider, D.M., Heun, M. and Fischbeck, G., 1991. Inheritance of the powdery mildew resistance gene *Pm9* in relation to *Pm1* and *Pm2* of wheat. Plant Breeding, 107: 161-164.

Singh, R.P., Huerta-Espino, J. and William, H.M. 2005. Genetics and breeding for durable resistance to leaf and stripe rusts in wheat. Turkish Journal of Agriculture and Forestry, 29: 121-127.

Singrün, C.H., Hsam, S.L.K., Hartl, L., Zeller, F.J. and Mohler, V. 2003. Powdery mildew resistance gene Pm22 in cultivar Virest is a member of the complex *Pm1* locus in common wheat (*Triticum aestivum* L. em Thell.). Theoretical and Applied Genetics, 106: 1420-1424.

Turner, M.K., Kolmer, J.A., Pumphrey, M.O., Bulli, P., Chao, S. and Anderson, J.A. 2017. Association mapping of leaf rust resistance loci in a spring wheat core collection. Theoretical and Applied Genetics, 130: 345-361.

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. Journal of dairy science, 91: 4414-4423.

Worthington, M., Lyerly, J., Petersen, S., Brown-Guedira, G., Marshall, D., Cowger, C., Parks, R. and Murphy, J.P. 2014. *MlUM15*: an *Aegilops Neglecta* derived Powdery Mildew Resistance Gene in Common Wheat. Crop Science, 54: 1397-1406.

Yu, J., Pressoir, G., Briggs, W.H., Bi, I.V., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Holland, J.B., and Kresovich, S. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature genetics, 38: 203-208.

Zadoks, J.C, Chang, T.T. and Konzak, C.F. 1974. A decimal code for the growth stages of cereals. Weed research, 14: 415-421.

Zhan, H., Li, G., Zhang, X., Li, X., Guo, H., Gong, W., Jia, J., Qiao, L., Ren, Y., Yang, Z. and Chang, Z. 2014. Chromosomal location and comparative genomics analysis of powdery mildew resistance gene *Pm51* in a putative wheat-*Thinopyrum ponticum* introgression line. PLoS One, 9: 1-15.

Zhang, Z., Ersoz, E., Lai, C.Q., Todhunter, R.J., Tiwari, H.K., Gore, M.A., Bradbury, P.J., Yu, J., Arnett, D.K., Ordovas, J.M. and Buckler, E.S. 2010. Mixed linear model approach adapted for genome-wide association studies. Nature genetics, 42: 355-360.

Zhang, D., Bowden, R.L., Yu, J., Carver, B.F. and Bai, G. 2014. Association analysis of stem rust resistance in US winter wheat. PLoS One, 9: 1-10.

Zhao, Z., Sun, H., Song, W., Lu, M., Huang, J., Wu, L., Wang, X. and Li, H. 2013. Genetic analysis and detection of the gene MlLX99 on chromosome 2BL conferring resistance to powdery mildew in the wheat cultivar Liangxing 99. Theoretical and applied genetics, 126: 3081-3089.

Zhu, C., Gore, M., Buckler, E.S. and Yu, J. 2008. Status and prospects of association mapping in plants. The plant genome, 1: 5-20.

Zhu, Z., Zhou, R., Kong, X., Dong, Y. and Jia, J. 2005. Microsatellite markers linked to 2 powdery mildew resistance genes introgressed from *Triticum carthlicum* accession PS5 into common wheat. Genome, 48: 585-590.

Table 4. 1: Summary phenotypic analysis for the greenhouse seedling screeining for population A and B using two different isolates KEN-2-5-B and KEN-2-5-D respectively with a scale from 0 (resistance) to 4 (susceptible) and summary phenotypic information of the combined field evalaution under natural infection using a scale from 0 (resistance) to 9 (suceptible). Including number of environment evaluated (No. Env), number of data points for the analysis of each trait (No. Data Points), number of genotypes (No. Genotypes), descriptive statistics based on BLUEs including minimum (Min), average (Mean), maximum (Max). Broad sense heritability on a per plot basis (Heritability) was estimated.

	Population A Seedling Screening	Population B Seedling Screening	Combined (Pop A and B) Field Screening
Scale	Infection Type (0-4)	Infection Type (0-4)	Severity (0-9)
No. Genotypes	264	559	862
No. Env	1	1	25
No. Data Points	792	1122	3947
Min	0	0	0
Mean	3.1	1.8	2.5
Max	4	4	6.5
Heritability	0.59	0.61	0.55

Table 4. 2: Summary genotypic information by chromosome for population A, B and Combined population. Including number of marker by chromosome (No. Markers), Average intrachromosomal pairwise LD measured as r^2 and mean pairwise distance between makers in Megabase pairs (Mbp).

Chr.	Number of Markers			Average Pairwise LD (r2)			Mean distance (Mbp)		
	Pop A	Pop B	Combined	Pop A	Pop B	Combined	Pop A	Pop B	Combined
1A	802	771	859	0.39	0.43	0.47	0.74	0.77	0.69
2A	808	827	957	0.38	0.45	0.49	0.97	0.95	0.82
3A	858	882	966	0.36	0.41	0.45	0.87	0.85	0.78
4A	755	793	886	0.33	0.37	0.42	0.98	0.94	0.84
5A	870	855	970	0.43	0.47	0.51	0.82	0.83	0.73
6A	730	691	771	0.38	0.44	0.47	0.85	0.89	0.80
7A	1310	1251	1438	0.36	0.39	0.43	0.56	0.59	0.51
1B	878	864	964	0.37	0.41	0.44	0.78	0.80	0.71
2B	1235	1361	1559	0.38	0.47	0.51	0.65	0.59	0.51
3B	1366	1367	1524	0.37	0.41	0.45	0.61	0.61	0.54
4B	443	464	539	0.43	0.51	0.57	1.52	1.45	1.25
5B	1065	1102	1220	0.37	0.44	0.47	0.67	0.65	0.58
6B	1064	923	1079	0.39	0.45	0.48	0.68	0.78	0.67
7B	959	881	1027	0.46	0.49	0.54	0.78	0.85	0.73
1D	306	328	388	0.51	0.65	0.67	1.62	1.51	1.28
2D	399	378	474	0.45	0.50	0.53	1.63	1.72	1.37
3D	169	161	171	0.35	0.39	0.42	3.66	3.85	3.61
4D	64	68	75	0.11	0.15	0.18	8.07	7.59	6.87
5D	147	130	142	0.29	0.28	0.34	3.86	4.37	4.00
6D	250	230	254	0.41	0.43	0.47	1.90	2.07	1.87
7D	226	239	256	0.34	0.42	0.44	2.81	2.67	2.50

Table 4. 3: Most significant markers (SNP) utilized as fixed covariates detected through association analysis for powdery mildew resistance under field and greenhouse evaluation along with information regarding the chromosome number (Chr), physical position in Megabase pair (Pos Mbp), and allele for each SNP (Allele), unadjusted p-value, minor allele frequency (maf), number of genotypes utilized in the analysis (nobs), Allelic Effect, False Discovery Rate (FDR) and proportion of the variance explained by the marker when enter the model.

SNP Name	Chr	Pos (Mbp)	Allele (u/f)	p-value	maf	nobs	Allelic Effect	FDR (p-value)	Var Explained
	Field screening under natural infection								
S2B_717237729	2B	717.24	T/C	2.6E-06	0.29	862	-0.32	2.5E-02	1
S6B_695007016	6B	695.01	C/A	2.8E-14	0.12	862	-0.65	2.7E-10	4
S7A_724919354	7A	724.92	A/T	2.6E-15	0.10	862	-0.86	4.3E-11	5.5
S7A_726648444	7A	726.65	G/A	4.9E-07	0.23	862	-0.33	8.1E-03	2
	Greenhouse seedling screening population B. *Bgt* isolate 'Ken-2-5-D'								
S1A_1236236	1A	1.24	A/G	3.0E-10	0.04	559	-0.69	4.4E-06	4
S1A_5149784	1A	5.15	C/T	2.9E-06	0.10	559	-0.28	4.2E-02	3
S7A_724919354	7A	724.92	A/T	6.3E-19	0.13	559	-0.79	2.8E-15	10
	Greenhouse seedling screening population A. *Bgt* isolate 'Ken-2-5-B'								
S7A_724919354	7A	724.92	A/T	2.8E-16	0.05	264	-1.30	4.1E-12	28

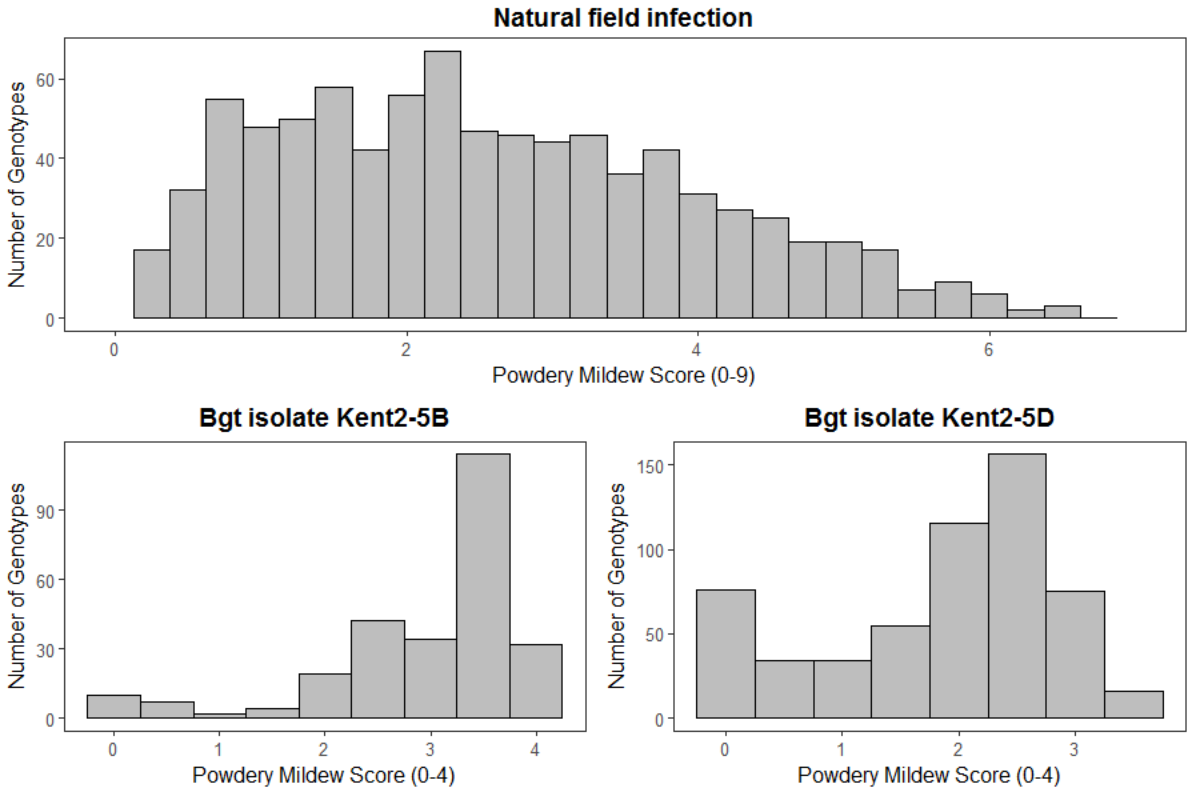


Figure 4. 1: Frequency distribution of disease severity to powdery mildew under natural infection in field conditions (A) and frequency distributions of infection type in response to the *Bgt* isolates 'KEN 2-5-B' in population A (B) and 'KEN2-5-D' in population B (C). Distributions are based on best linear unbiased estimates (BLUEs) for each data.

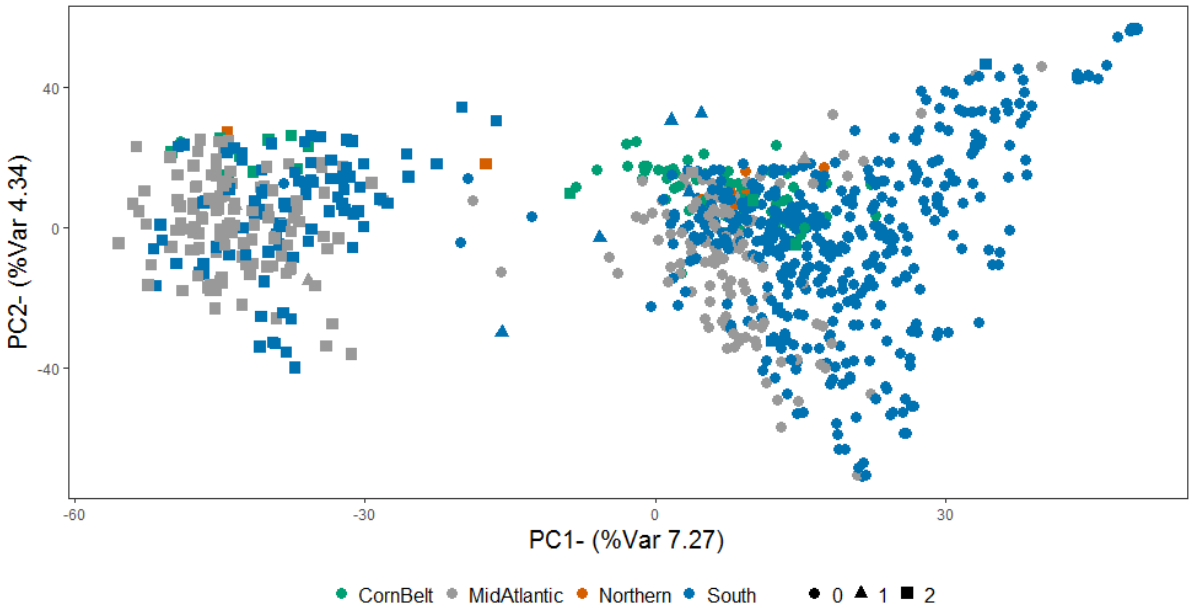


Figure 4. 2: Scatter plot of the first two principal components after analysis of 862 winter wheat lines using 15,559 SNPs. Points are color coded according to the origin of genotypes in Corn Belt, Mid Atlantic, Northern and South of USA. Different shapes represent the number of copies of the allele of the diagnostic marker *Sr36* linked to the 2BS:2GS 2GL:2BL translocation from *T. timopheevii*. Percentages in each axis represents the proportion of variance explained by each principal component.

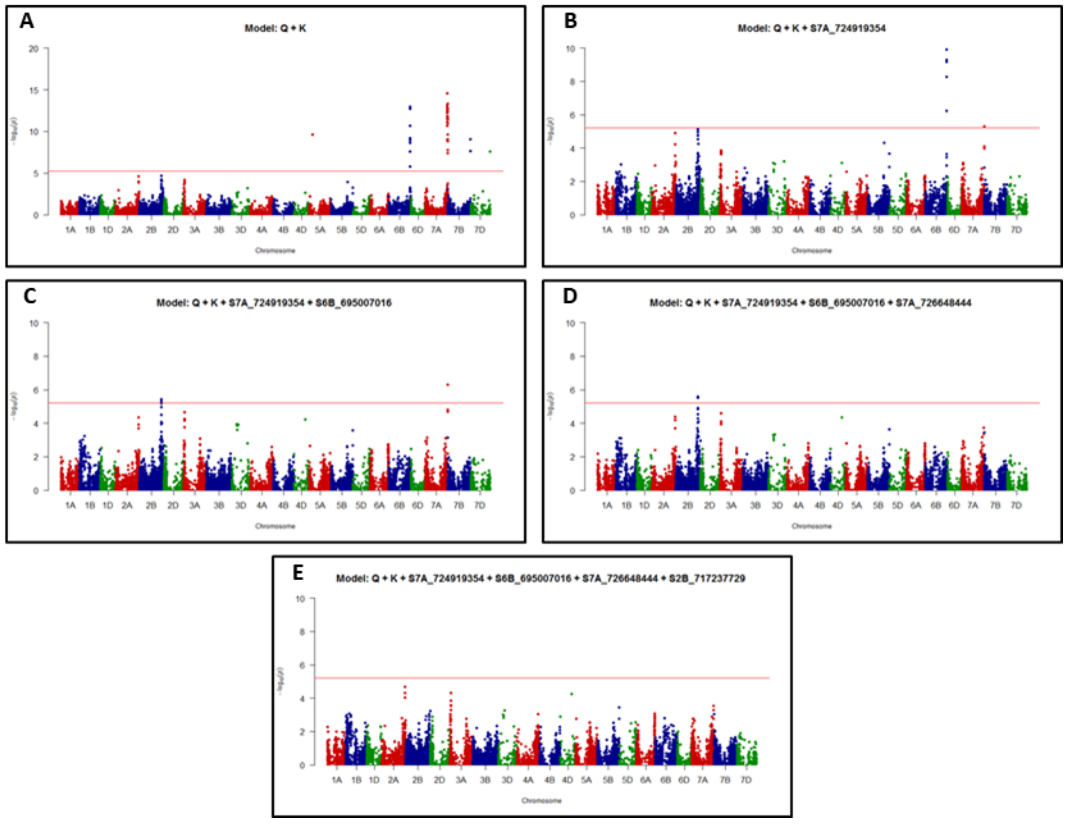


Figure 4. 3: Manhattan plot for each step of the association analysis for resistance to powdery mildew under natural infection in the field. The most significant markers were sequentially added as fixed covariates in different steps. Each analysis was represented in a different plot: (A) mixed model included the first three principal components and the polygenic effect of genotypes named Q + K model, (B): Q + K + 7A_724919354, (C) Q + K + 7A_724919354 + 6B_695007016, (D) Q + K + 7A_724919354 + 6B_695007016 + 7A_726648444 and (E) Q + K + 7A_724919354 + 6B_695007016 + 7A_726648444 + 2B_717237729. The y-axis represents the p-value of the marker-trait association on a -log10 scale. Markers are ordered by map physical position and grouped by chromosome (x-axis). The three different wheat genomes are represented with different colors: A genome (red), B genome (blue), and D genome (green). Solid horizontal line represents a significance threshold at a Bonferroni corrected alpha of 0.1.

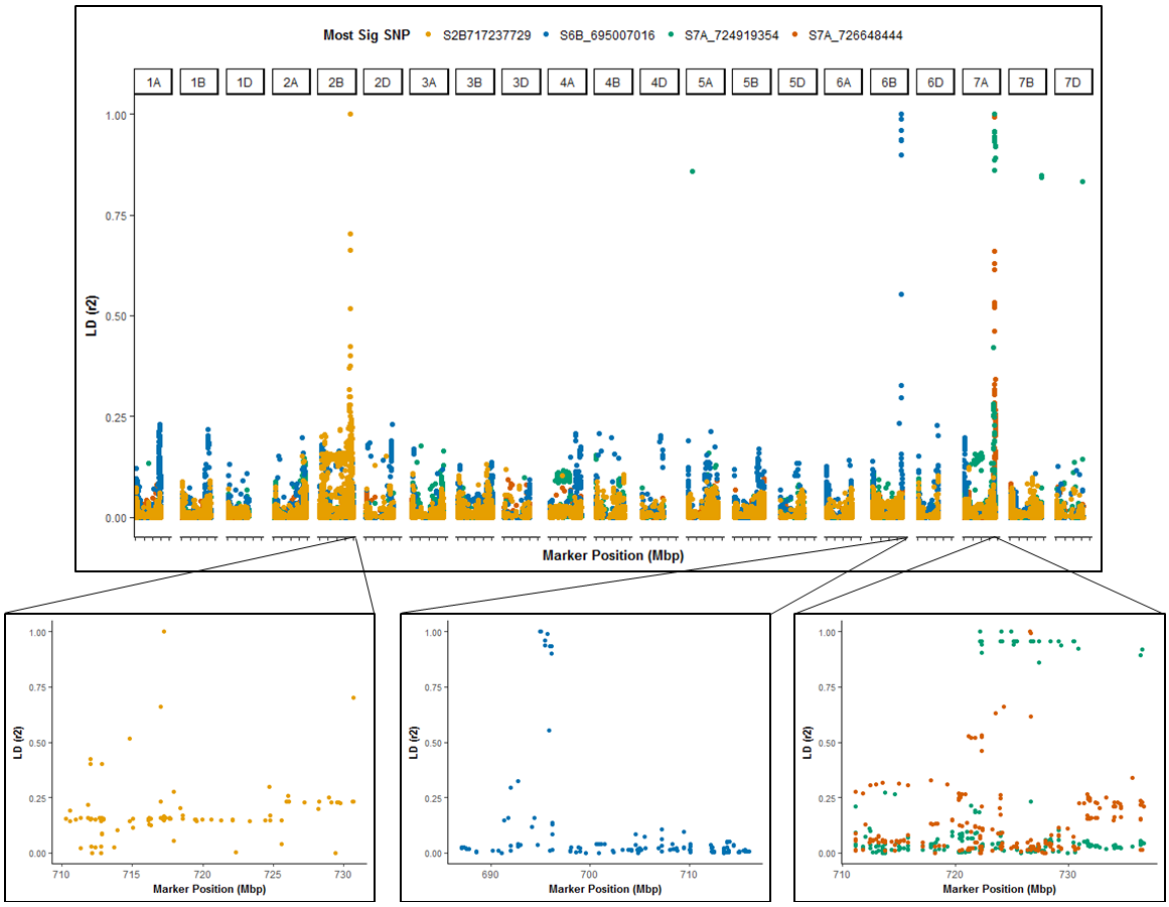


Figure 4. 4: Pattern of pairwise linkage disequilibrium (LD) measured as r^2 for the markers utilized as covariates in the combined population evaluated for powdery mildew resistance under natural infections. Plots include pairwise LD genome wide and zoom in the region were the significant marker-trait association were detected.

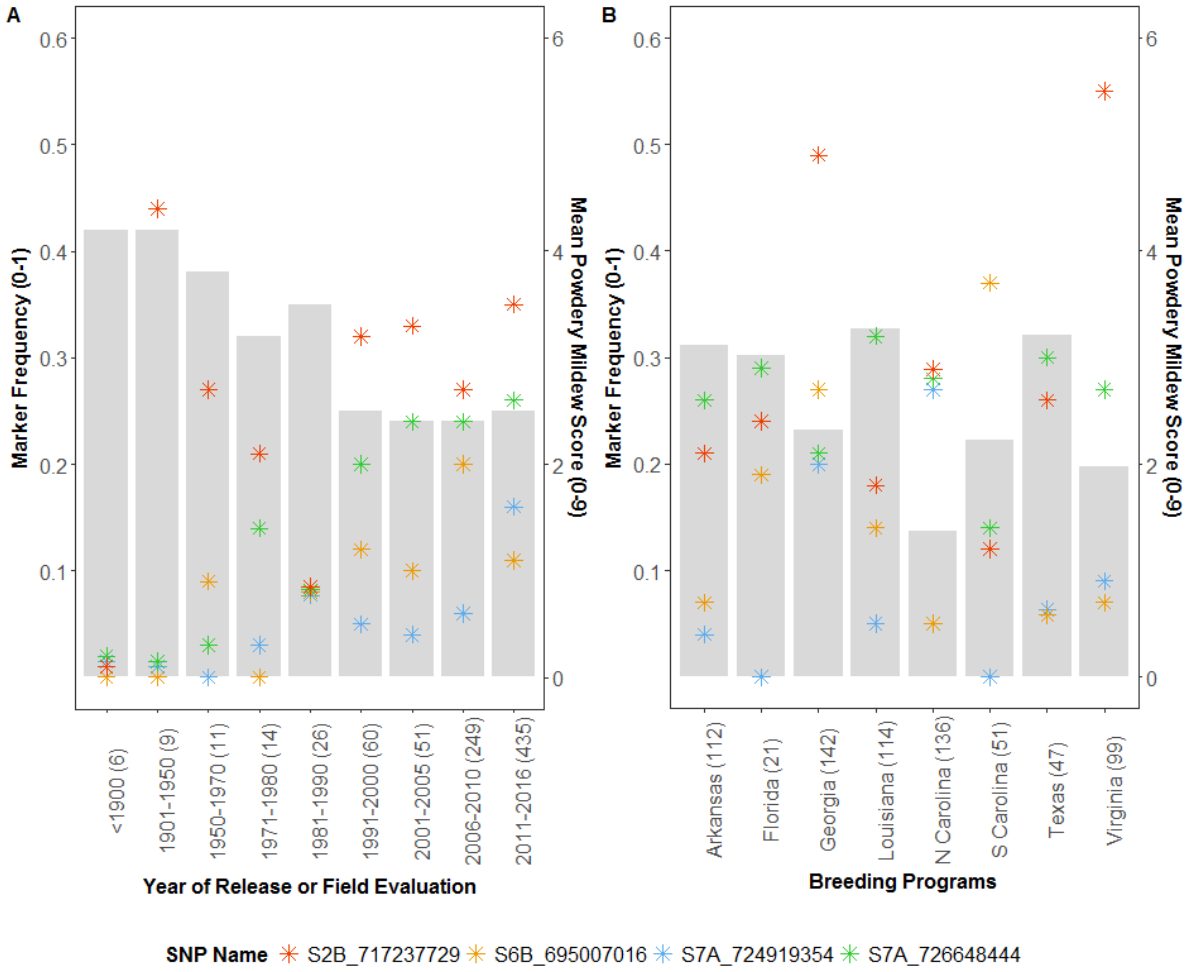


Figure 4. 5: Combined plots with frequency of favorable alleles for the most significant markers detected (color dots) and mean Powdery mildew score (0-9) evaluated under natural infection in the field (bars). A) Data grouped based on the year of cultivar release or first year of evaluation in the SUNGRAINS cooperative nurseries, B) Data grouped based on southern public breeding program where the material was originated. The number of genotypes in each group for each plot is within parenthesis in the x axis in combination with the name of the group.

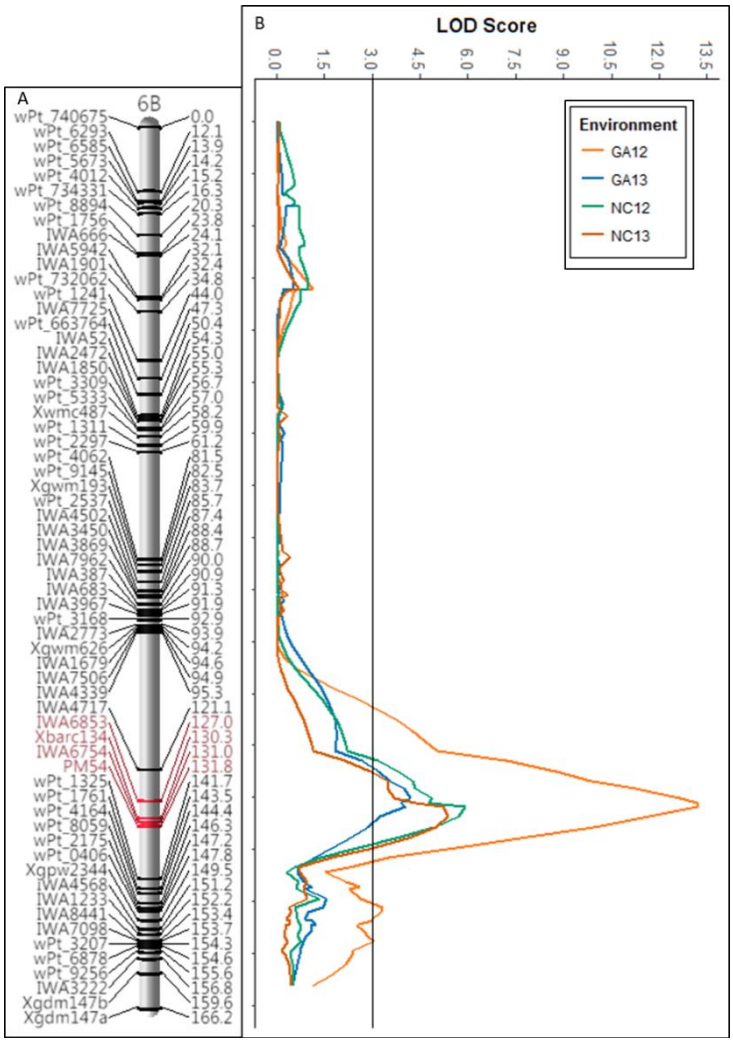


Figure 4. 6: (A) Genetic linkage map of chromosome 6B showing the SSR, Dart and SNP markers linked to the powdery mildew resistance gene *Pm54* in the AGS 2000 x Pioneer brand 26R61 recombinant inbred lines mapping population. Molecular markers with the prefixes IWA, wPt are SNPs and DArT markers respectively while all other prefixes refer to SSR markers. Kosambi map distances (cM) between markers are shown. (B) Logarithm of odds (LOD) scores for powdery mildew resistance under field evaluation in four environments: Plains, GA and Raleigh, NC during season 2012 and 2013.

Supplemental Material for Chapter 4

S. Table 4. 1: Differential powdery mildew reaction for four check cultivars after inoculation with the isolates KEN-2-5-B and KEN-2-5-D in greenhouse screening using a scale from 0 (resistance) to 4 (susceptible).

Isolate	Cultivar			
	Cocker-6815	Saluda	USG3209	NC-Neuse
KEN 2-5-B	3.7	3.6	3.2	0.3
KEN2-5-D	2.5	1.95	2.29	0

S. Table 4. 2: Summary phenotypic information of powdery mildew severity evaluated in the field under natural infection using a scale from 0 (resistance) to 9 (susceptible) for each environment evaluated. The number of different genotypes evaluated (No. Geno), overall mean (Mean), standard deviation (Std Dev), Minimum (Min) and Maximum (Max) for the environment (Mean) were included in the table.

Field Severity Powdery Mildew					
Environment	No. Geno	Mean	Std Dev	Min	Max
ALLRALEIGH2016	577	2.1	1.1	1	6
ALLRALEIGH2017	472	2.2	1.6	0	7
DIVPANELFAYETVILLE2013	365	1.7	1.5	0	6
DIVPANELKINSTON2014	365	2.1	1.5	0	6
DIVPANELRALEIGH2012	364	3.9	1.7	0	8
DIVPANELRALEIGH2013	364	4.9	2.0	0	9
GAWNKINSTON2008	82	2.3	1.5	0	7
GAWNKINSTON2009	74	3.6	1.8	0	8
GAWNKINSTON2010	75	2.2	2.3	0	9
GAWNKINSTON2011	56	2.4	2.6	0	8
GAWNKINSTON2012	76	4.0	2.1	0	7
GAWNKINSTON2013	64	3.4	2.3	0	8
GAWNPLAINS2009	74	1.1	1.8	0	8
GAWNQUINCY2008	59	0.8	1.4	0	6
GAWNQUINCY2012	76	2.1	1.6	0	6
GAWNQUINCY2013	64	2.0	2.1	0	6
GAWNWARSAW2008	82	2.5	2.1	0	7
GAWNWARSAW2009	74	3.0	1.8	0	9
GAWNWARSAW2010	39	2.8	1.7	1	8
GAWNWARSAW2011	56	1.6	1.8	0	7
GAWNWARSAW2012	76	1.4	2.3	0	8
GAWNWARSAW2016	45	2.9	1.5	1	7
SUNWHEATKINSTON2014	79	2.4	1.7	0	7
SUNWHEATKINSTON2015	79	3.1	2.7	0	8
SUNWHEATQUINCY2014	79	1.3	1.9	0	7

S. Table 4. 3: Most significant markers (SNP Name) detected in population A through association analysis for powdery mildew resistance under greenhouse evaluation with *Bgt* isolate Ken-2-5-B. Information regarding the chromosome number (Chr), physical position in Megabase pair (Pos Mbp), allele for each SNP (Allele), minor allele frequency (maf), unadjusted p-value, and Allelic Effect were included.

SNP Name	Chr	Pos (Mbp)	Allele (u/f)	maf	p-value	Allelic Effect
S2B_632449810	2B	632.45	G/A	0.05	2.28E-11	-1.18
S7A_711172059	7A	711.17	G/A	0.15	8.87E-06	-0.41
S7A_711184393	7A	711.18	G/A	0.15	9.82E-06	-0.41
S7A_724919354	7A	724.92	A/T	0.05	2.80E-16	-1.30
S7A_726681246	7A	726.68	T/C	0.23	1.94E-08	-0.48

S. Table 4. 4: Most significant markers (SNP Name) detected in population B through association analysis for powdery mildew resistance under greenhouse evaluation with *Bgt* isolate Ken-2-5-D. Information regarding the chromosome number (Chr), physical position in Megabase pair (Pos Mbp), allele for each SNP (Allele), minor allele frequency (maf), unadjusted p-value, and Allelic Effect were included.

SNP Name	Chr	Pos (Mbp)	Allele (u/f)	maf	p-value	Allelic Effect
S1A_1236236	1A	1.2362	A/G	0.04	3.02E-10	-0.68
S1A_5149784	1A	5.1498	C/T	0.10	2.91E-06	-0.26
S1A_14520376	1A	14.52	T/G	0.18	2.70E-04	-0.19
S1A_219369883	1A	219.37	A/G	0.05	7.45E-07	-0.47
S1A_220176480	1A	220.18	T/C	0.05	7.23E-06	-0.42
S1A_221795939	1A	221.8	G/C	0.05	3.98E-07	-0.48
S1A_238691007	1A	238.69	T/C	0.05	7.46E-07	-0.45
S1A_238866829	1A	238.87	G/A	0.05	9.35E-07	-0.47
1RS1AL	1A	251.18	A/C	0.05	2.04E-07	-0.50
S1A_255874119	1A	255.87	T/A	0.08	2.64E-04	-0.29
S1A_256921241	1A	256.92	C/T	0.08	2.39E-04	-0.28
S1A_259478516	1A	259.48	C/T	0.05	7.50E-06	-0.44
S1A_261835592	1A	261.84	A/C	0.05	2.99E-06	-0.46
S1A_262443087	1A	262.44	C/T	0.05	4.95E-06	-0.45
S1A_264984531	1A	264.98	T/C	0.05	4.85E-06	-0.45
S1A_268120982	1A	268.12	T/G	0.05	9.29E-07	-0.45
S1A_270575281	1A	270.58	C/G	0.05	4.85E-06	-0.45
S1A_278230314	1A	278.23	A/T	0.05	1.24E-05	-0.43
S1A_283847888	1A	283.85	T/G	0.05	2.66E-06	-0.46
S1A_283855283	1A	283.86	C/T	0.05	6.55E-07	-0.47
S1A_284321998	1A	284.32	T/C	0.05	1.47E-06	-0.45
S1A_285351710	1A	285.35	T/C	0.05	1.92E-06	-0.45
S1A_285635261	1A	285.64	G/A	0.05	2.52E-06	-0.43
S1A_289268398	1A	289.27	G/T	0.05	1.07E-05	-0.42
S1A_290369635	1A	290.37	A/G	0.05	1.03E-06	-0.46
S1A_292761222	1A	292.76	T/C	0.05	1.16E-06	-0.45
S1A_293734502	1A	293.73	T/C	0.05	1.40E-06	-0.45
S1A_294513533	1A	294.51	C/A	0.14	2.47E-04	-0.22
S1A_294734999	1A	294.73	A/G	0.05	1.31E-06	-0.45
S1A_294735024	1A	294.74	A/C	0.05	1.31E-06	-0.45
S1A_295537776	1A	295.54	T/C	0.05	6.09E-07	-0.46
S1A_298116671	1A	298.12	C/T	0.05	8.36E-05	-0.38
S1A_364223804	1A	364.22	G/A	0.04	6.21E-06	-0.47
S1A_368746801	1A	368.75	T/C	0.05	3.03E-05	-0.42
S1A_392044194	1A	392.04	G/A	0.04	3.60E-08	-0.56
S1A_396455873	1A	396.46	A/G	0.05	6.53E-05	-0.37
S3A_634947427	3A	634.95	A/C	0.33	1.28E-04	-0.17
S3D_170070173	3D	170.07	T/C	0.08	4.28E-11	-0.71
S4D_373813222	4D	373.81	A/G	0.06	6.14E-05	-0.41

S. Table 4.4 cont.

SNP Name	Chr	Pos (Mbp)	Allele (u/f)	maf	p-value	Allelic Effect
S5A_103749173	5A	103.75	G/T	0.09	1.19E-13	-0.83
S5A_665972392	5A	665.97	C/G	0.10	1.81E-04	0.26
S5B_15543996	5B	15.544	T/G	0.48	1.69E-04	-0.16
S5B_653428582	5B	653.43	A/G	0.07	2.39E-04	0.29
S7A_347569105	7A	347.57	A/G	0.04	2.07E-04	-0.39
S7A_675526339	7A	675.53	G/C	0.09	8.21E-05	0.30
S7A_697829646	7A	697.83	G/A	0.10	3.35E-05	-0.30
S7A_705365587	7A	705.37	T/G	0.13	1.94E-06	-0.38
S7A_705417665	7A	705.42	A/G	0.14	1.23E-06	-0.37
S7A_705490401	7A	705.49	T/G	0.13	1.24E-05	-0.35
S7A_705490416	7A	705.49	C/T	0.13	4.83E-06	-0.36
S7A_705490490	7A	705.49	T/C	0.13	1.50E-06	-0.37
S7A_705556719	7A	705.56	A/C	0.13	8.82E-07	-0.39
S7A_705561437	7A	705.56	G/C	0.13	1.38E-05	-0.35
S7A_705593667	7A	705.59	G/A	0.13	3.28E-06	-0.37
S7A_705666330	7A	705.67	G/A	0.13	5.01E-06	-0.37
S7A_705671050	7A	705.67	T/C	0.09	1.87E-09	-0.58
S7A_708465599	7A	708.47	G/T	0.44	3.75E-04	-0.18
S7A_708778441	7A	708.78	C/T	0.42	1.98E-05	-0.22
S7A_709785025	7A	709.79	G/C	0.47	1.67E-04	0.21
S7A_709814998	7A	709.81	G/A	0.24	2.36E-04	0.25
S7A_711171597	7A	711.17	G/A	0.48	4.15E-05	0.23
S7A_711172059	7A	711.17	G/A	0.22	5.37E-04	-0.21
S7A_712366845	7A	712.37	G/A	0.26	4.27E-06	-0.26
S7A_712491983	7A	712.49	C/T	0.26	5.93E-04	0.23
S7A_712491985	7A	712.49	A/G	0.36	4.68E-06	-0.24
S7A_713758436	7A	713.76	C/A	0.20	2.61E-04	-0.24
S7A_714662345	7A	714.66	G/A	0.23	2.31E-04	-0.22
S7A_720585892	7A	720.59	G/C	0.46	1.17E-04	0.22
S7A_720585903	7A	720.59	C/T	0.26	2.16E-04	-0.21
S7A_720585972	7A	720.59	G/A	0.46	3.47E-04	0.20
S7A_720999591	7A	721	C/T	0.48	4.49E-04	0.19
S7A_721386071	7A	721.39	G/C	0.33	2.82E-05	-0.23
S7A_721773917	7A	721.77	C/T	0.37	2.79E-05	-0.23
S7A_721773997	7A	721.77	C/T	0.36	6.92E-06	-0.25
S7A_722081781	7A	722.08	G/A	0.36	9.12E-05	-0.21
S7A_722211697	7A	722.21	A/G	0.13	8.86E-19	-0.79
S7A_722330211	7A	722.33	G/C	0.37	2.83E-04	0.19
S7A_722330754	7A	722.33	A/C	0.12	2.78E-17	-0.79
S7A_722330781	7A	722.33	T/G	0.12	2.78E-17	-0.79
S7A_722332364	7A	722.33	C/T	0.09	6.59E-12	-0.70
S7A_724031266	7A	724.03	C/T	0.12	2.45E-18	-0.81

S. Table 4.4 cont.

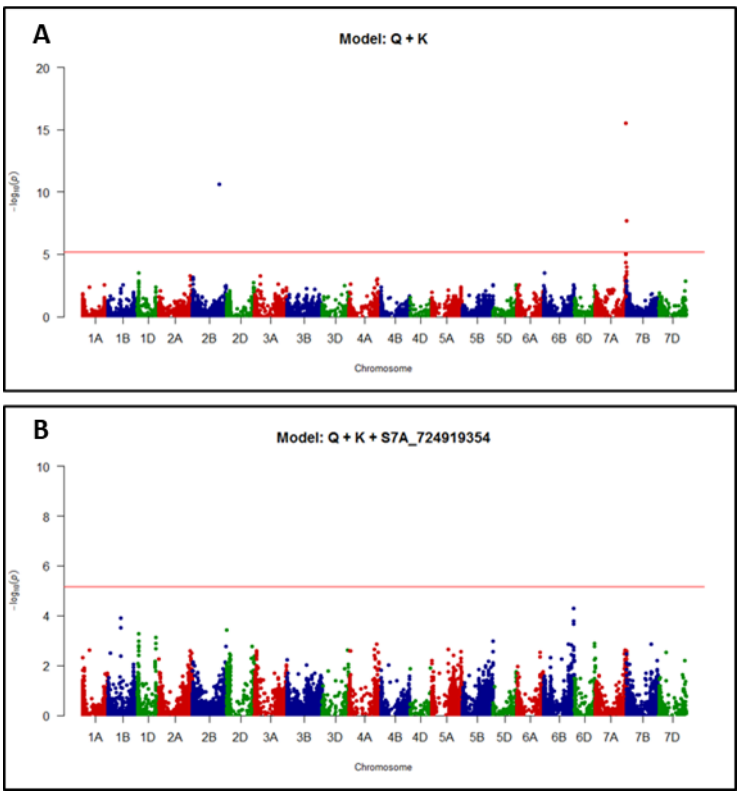
SNP Name	Chr	Pos (Mbp)	Allele (u/f)	maf	p-value	Allelic Effect
S7A_724031294	7A	724.03	G/A	0.46	5.22E-04	0.19
S7A_724031303	7A	724.03	C/T	0.12	2.45E-18	-0.81
S7A_724113839	7A	724.11	T/C	0.12	2.71E-18	-0.81
S7A_724919354	7A	724.92	A/T	0.13	6.33E-19	-0.79
S7A_725136016	7A	725.14	T/C	0.12	4.48E-18	-0.80
S7A_725459184	7A	725.46	G/C	0.12	4.06E-18	-0.81
S7A_726681222	7A	726.68	C/A	0.12	6.75E-19	-0.82
S7A_726681246	7A	726.68	T/C	0.37	5.08E-05	-0.21
S7A_726686008	7A	726.69	T/C	0.12	2.24E-18	-0.81
S7A_727430621	7A	727.43	A/G	0.10	1.00E-12	-0.66
S7A_727443660	7A	727.44	A/G	0.13	1.63E-16	-0.71
S7A_728232085	7A	728.23	A/G	0.42	2.22E-05	0.23
S7A_728427930	7A	728.43	C/G	0.11	9.51E-17	-0.75
S7A_728427937	7A	728.43	C/A	0.40	2.86E-05	0.23
S7A_728501883	7A	728.5	A/C	0.41	2.62E-04	0.20
S7A_729130753	7A	729.13	C/A	0.38	1.23E-04	0.22
S7A_729130806	7A	729.13	C/G	0.12	1.15E-17	-0.80
S7A_729361814	7A	729.36	C/T	0.10	1.63E-15	-0.77
S7A_730488733	7A	730.49	G/A	0.12	9.60E-19	-0.81
S7A_730488740	7A	730.49	G/C	0.39	1.88E-04	0.20
S7A_730488763	7A	730.49	C/G	0.12	9.60E-19	-0.81
S7A_730859470	7A	730.86	G/A	0.10	2.30E-16	-0.84
S7A_736435015	7A	736.44	T/C	0.09	3.31E-13	-0.76
S7A_736526752	7A	736.53	T/C	0.09	3.14E-13	-0.79
S7B_744996923	7B	745	T/C	0.10	4.24E-12	-0.67
S7B_748713164	7B	748.71	T/C	0.09	9.38E-13	-0.72
S7D_593236919	7D	593.24	G/A	0.10	1.47E-06	-0.32
S7D_634582986	7D	634.58	G/T	0.10	5.65E-12	-0.67

S. Table 4. 5: Most significant markers (SNP Name) detected in the combined panel through association analysis for powdery mildew resistance in field screening under natural infection. Information regarding the chromosome number (Chr), physical position in Megabase pair (Pos Mbp), allele for each SNP (Allele), minor allele frequency (maf), lowest unadjusted p-value from the different models considered (p-value), and average allelic effect (Allelic Effect) were included.

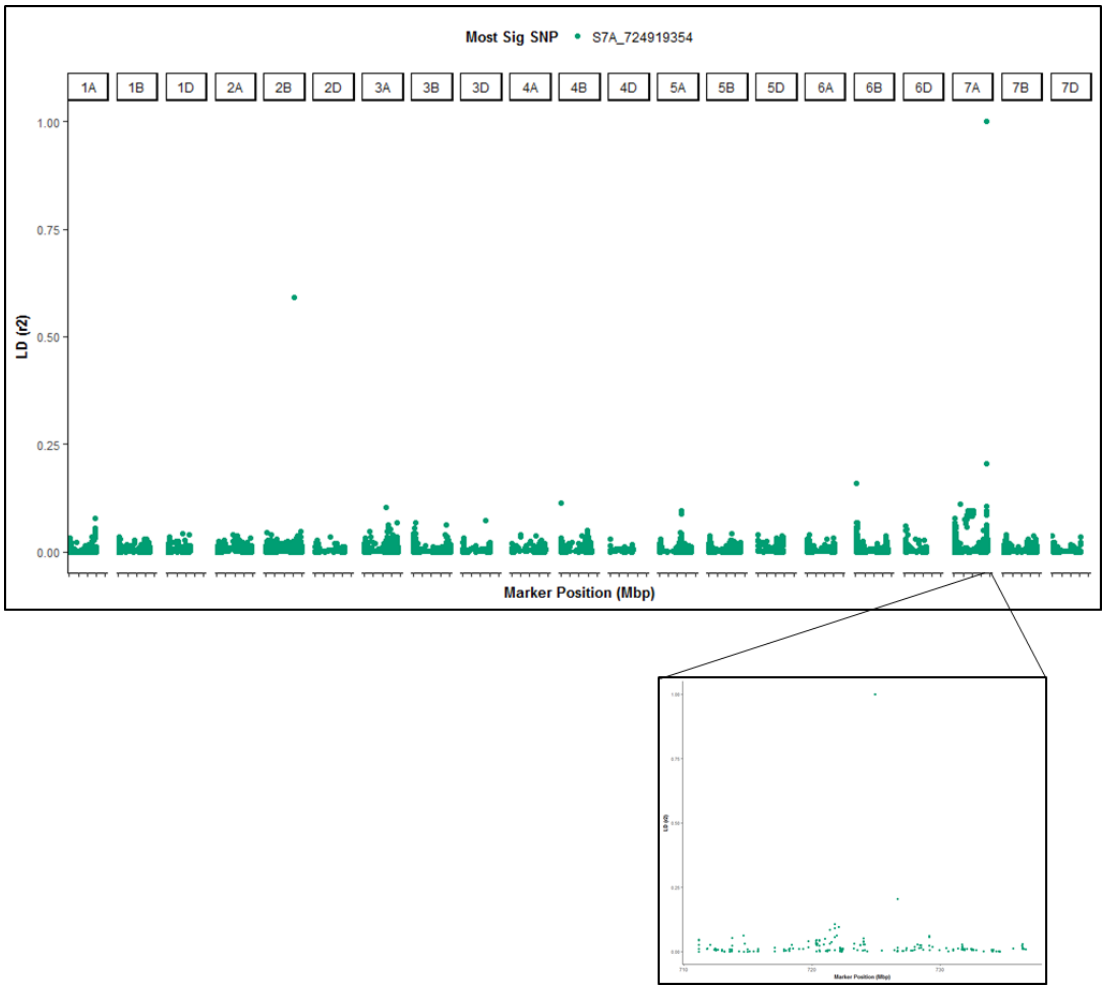
SNP Name	Chr	Pos (Mbp)	Allele (u/f)	maf	p-value	Allelic Effect
S2A_755046447	2A	755.05	A/G	0.40	1.26E-04	0.23
S2A_755068015	2A	755.07	G/T	0.40	2.51E-05	0.23
S2A_757964031	2A	757.96	G/T	0.06	1.21E-05	-0.47
S2B_714778646	2B	714.78	C/T	0.41	1.10E-05	-0.32
S2B_714778647	2B	714.78	A/G	0.32	3.01E-05	0.30
S2B_714778648	2B	714.78	G/T	0.41	1.10E-05	-0.32
S2B_716180576	2B	716.18	G/A	0.31	3.06E-06	0.30
S2B_716180619	2B	716.18	T/A	0.26	1.17E-04	0.27
S2B_716290118	2B	716.29	G/C	0.25	5.36E-05	0.28
S2B_716978432	2B	716.98	G/A	0.35	6.74E-06	-0.32
S2B_717237729	2B	717.24	T/C	0.29	2.61E-06	-0.32
S2B_730676924	2B	730.68	T/C	0.29	5.71E-05	-0.29
S3A_20557241	3A	20.56	G/C	0.49	2.47E-04	0.24
S3A_21787498	3A	21.79	A/G	0.45	1.23E-04	-0.24
S3A_21799570	3A	21.80	C/T	0.43	5.51E-05	-0.23
S3A_22234992	3A	22.23	C/T	0.43	2.19E-05	-0.24
S3A_22244424	3A	22.24	G/A	0.37	6.67E-05	-0.23
S3D_158450587	3D	158.45	C/G	0.43	1.10E-04	-0.21
S3D_206028658	3D	206.03	G/A	0.42	1.20E-04	-0.21
S4D_373813222	4D	373.81	A/G	0.07	4.34E-05	-0.45
S5A_103749173	5A	103.75	G/T	0.07	2.43E-10	-0.85
S5B_547983737	5B	547.98	T/C	0.06	4.70E-05	-0.54
S6B_695007016	6B	695.01	C/A	0.12	2.82E-14	-0.66
S6B_695489247	6B	695.49	G/T	0.09	5.21E-09	-0.61
S6B_695490151	6B	695.49	G/T	0.09	5.11E-10	-0.69
S6B_695676537	6B	695.68	A/G	0.12	3.25E-14	-0.65
S6B_695913077	6B	695.91	G/A	0.19	6.00E-07	-0.34
S6B_695918445	6B	695.92	G/A	0.13	6.17E-12	-0.57
S6B_696099425	6B	696.10	A/G	0.13	1.26E-10	-0.54
S6B_696144281	6B	696.14	G/T	0.09	6.54E-10	-0.62

S. Table 4.5 cont.

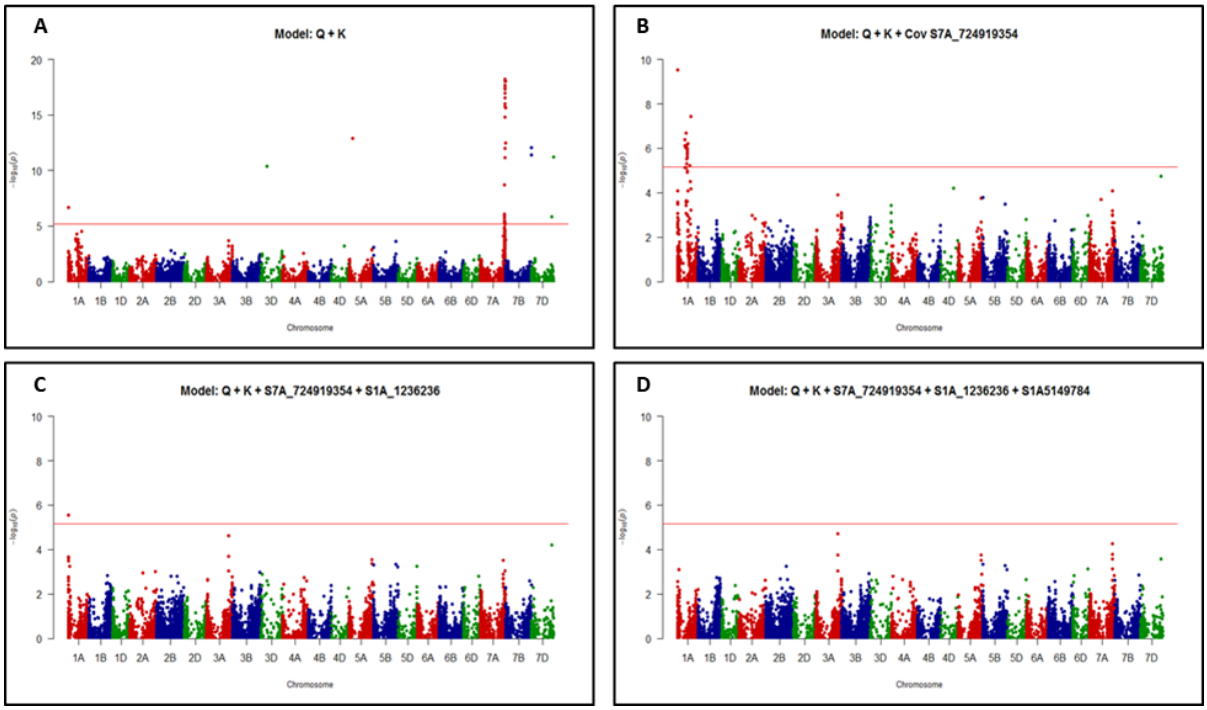
SNP Name	Chr	Pos (Mbp)	Allele (u/f)	maf	p-value	Allelic Effect
S7A_711172059	7A	711.17	G/A	0.20	3.22E-04	-0.26
S7A_722211697	7A	722.21	A/G	0.10	7.56E-12	-0.76
S7A_722211715	7A	722.21	G/C	0.11	2.89E-12	-0.73
S7A_722211723	7A	722.21	C/T	0.10	2.02E-11	-0.75
S7A_722330754	7A	722.33	A/C	0.10	7.12E-14	-0.85
S7A_722330781	7A	722.33	T/G	0.10	7.12E-14	-0.85
S7A_722332364	7A	722.33	C/T	0.08	9.49E-10	-0.77
S7A_724031266	7A	724.03	C/T	0.10	1.20E-13	-0.84
S7A_724031303	7A	724.03	C/T	0.10	1.20E-13	-0.84
S7A_724113839	7A	724.11	T/C	0.10	5.63E-13	-0.82
S7A_724919354	7A	724.92	T/A	0.10	2.60E-15	-0.86
S7A_725136016	7A	725.14	T/C	0.10	2.34E-13	-0.83
S7A_725459184	7A	725.46	G/C	0.10	1.49E-12	-0.82
S7A_726648444	7A	726.65	G/A	0.23	4.90E-07	-0.31
S7A_726681216	7A	726.68	T/C	0.21	2.06E-05	-0.29
S7A_726681222	7A	726.68	C/A	0.10	4.04E-13	-0.83
S7A_726681246	7A	726.68	T/C	0.34	1.06E-11	-0.33
S7A_726686008	7A	726.69	T/C	0.10	3.76E-12	-0.79
S7A_727430621	7A	727.43	A/G	0.08	1.54E-09	-0.71
S7A_727443660	7A	727.44	A/G	0.11	3.42E-12	-0.73
S7A_728427930	7A	728.43	C/G	0.09	1.96E-12	-0.79
S7A_729130806	7A	729.13	C/G	0.10	2.05E-13	-0.85
S7A_729361814	7A	729.36	C/T	0.08	3.76E-08	-0.68
S7A_730488733	7A	730.49	G/A	0.10	4.57E-14	-0.86
S7A_730488763	7A	730.49	C/G	0.10	4.57E-14	-0.86
S7A_730859470	7A	730.86	G/A	0.08	2.49E-10	-0.79
S7A_736435015	7A	736.44	T/C	0.07	8.41E-10	-0.83
S7A_736526752	7A	736.53	T/C	0.07	1.82E-08	-0.75
S7B_744996923	7B	745.00	T/C	0.08	8.14E-10	-0.77
S7B_748713164	7B	748.71	T/C	0.07	2.30E-08	-0.72
S7D_634582986	7D	634.58	G/T	0.08	2.68E-08	-0.68



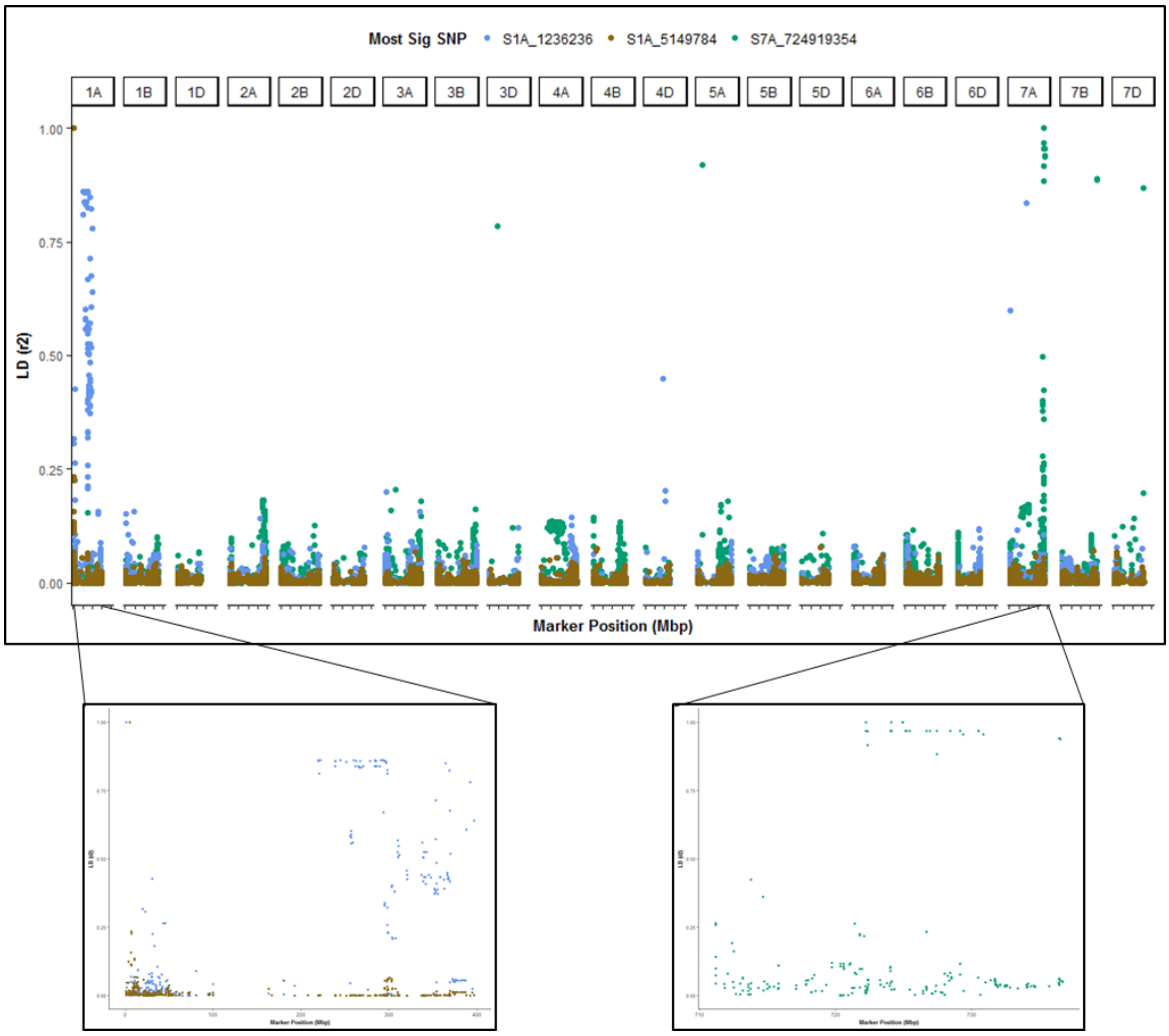
S. Figure 4. 1: Manhattan plot for each step of the association analysis for resistance to powdery mildew using the *Bgt* isolate 'KEN2-5-B' in greenhouse seedling screening with population A. The most significant markers were sequentially added as fixed covariates in different steps. Each analysis was represented in a different plot: (A) mixed model included the first three principal components and the polygenic effect of genotypes named Q + K model, (B): Q + K + 7A_724919354. The y-axis represents the p-value of the marker-trait association on a -log10 scale. Markers are ordered by map physical position and grouped by chromosome (x-axis). The three different wheat genomes are represented with different colors: A genome (red), B genome (blue), and D genome (green). Solid horizontal line represents a significance threshold at a Bonferroni corrected alpha of 0.1.



S. Figure 4. 2: Pattern of pairwise linkage disequilibrium (LD) measured as r^2 for the markers utilized as covariates in population A evaluated in the greenhouse for powdery mildew resistance with the *Bgt* isolate 'KEN2-5-B'. Plots include pairwise LD genome wide and in the region, were the significant marker-trait association was detected.



S. Figure 4. 3: Manhattan plot for each step of the association analysis for resistance to powdery mildew using the *Bgt* isolate 'KEN2-5-D' in greenhouse seedling screening with population B. The most significant markers were sequentially added as fixed covariates in different steps. Each analysis was represented in a different plot: (A) mixed model included the first three principal components and the polygenic effect of genotypes named Q + K model, (B): Q + K + 7A_724919354, (C) Q + K + 7A_724919354 + 1A_1236236, (D) Q + K + 7A_724919354 + 1A_1236236 + 1A_5149784. The y-axis represents the p-value of the marker-trait association on a -log10 scale. Markers are ordered by map physical position and grouped by chromosome (x-axis). The three different wheat genomes are represented with different colors: A genome (red), B genome (blue), and D genome (green). Solid horizontal line represents a significance threshold at a Bonferroni corrected alpha of 0.1.



S. Figure 4. 4: Pattern of pairwise linkage disequilibrium (LD) measured as r^2 for the markers utilized as covariates in population B evaluated in the greenhouse for powdery mildew resistance with the *Bgt* isolate 'KEN2-5-D'. Plots include pairwise LD genome wide and in the region, were the significant marker-trait association was detected.

APPENDICES

APPENDIX A. Information about the complete set of genotypes evaluated in the GAWN and SUNWHEAT nurseries from 2008 to 2016. We include information about the name of each genotype (Variety), number of times each genotype was evaluated (No. Year) nursery and year each genotype was evaluated (Nursery), breeding program where the germplasm was developed (Breeding program), pedigree information if available (Pedigree), and status of the germplasm in the breeding pipeline as commercial cultivar (CC) or breeding line (BL) (Type). The name of the nursery and the year each genotype was evaluated was abbreviated as the initials of the nursery GAWN (G) and SUNWHEAT (S) while the year was represented by the last two characters.

Variety	No. Year	Nursery	Breeding Program	Pedigree	Type
AGS2000	1	G08	Georgia	PIONEER-2555/PF-84301//FL-302	CC
AGS2060	5	G09 to 13	Georgia	FL302/FR8119//COKER9663	CC
AR00036-5-1	1	G10	Arkansas	SABBE/USG3209	BL
AR00039-5-2	1	G10	Arkansas	SABBE/AR684-22-1-1	BL
AR00082-13-2	1	G10	Arkansas	PAT/AR647-1-6	BL
AR00090-1-1	1	G10	Arkansas	PAT/PIONEER2684	BL
AR00120-11-1	1	G09	Arkansas	CEZANNE/USG3209	BL
AR00134-3-4	1	G10	Arkansas	COKER9663/PAT	BL
AR00179-4-1	1	G12	Arkansas	IL94-6727/ROANE	BL
AR00196-10-1	1	G10	Arkansas	OK91P648-41/SABBE	BL
AR00255-16-1	1	G10	Arkansas	ROANE/LA9070G45-3-3-1	BL
AR00343-5-1	1	G12	Arkansas	AR97052/ROANE	BL
AR00380-3-3	1	G12	Arkansas	AR97054/97201	BL
AR01008-12-2-C	1	G09	Arkansas	AR85411/PAT	BL
AR01039-4-1	1	G12	Arkansas	AR800-1-3-1/AR839-28-1-2	BL
AR01040-4-1	1	G12	Arkansas	AR800-1-3-1/AR910-12-1	BL
AR01044-1-1	1	G12	Arkansas	AR800-1-3-1/AR92145E8-7-7-1-0	BL
AR01058-1	1	G12	Arkansas	AR839-27-1-3/ROANE	BL
AR01156-2-1	1	G12	Arkansas	GA901146/AR839-27-1-3	BL
AR01168-3-1	1	G12	Arkansas	VA98W-593/AR839-28-1-2	BL
AR01177-2-1	1	G12	Arkansas	AR92145E8-7-7-1-0/AR9035-4-2	BL
AR01205-1-1	1	G12	Arkansas	PI155271/ARLA85411	BL
AR01209-2-1	1	G12	Arkansas	AGS2000/PI531193	BL
AR04001-3	2	G13-14	Arkansas	AR800-1-3-1/AR908-8-2	BL
AR04002-3	1	G13	Arkansas	AR800-1-3-1/AR93005-6-5	BL
AR04006-1	1	G13	Arkansas	AR857-1-1/KS00WGRC44	BL
AR04008-5	1	G13	Arkansas	AR93005-6-5/COKER9375	BL
AR04015-5	1	G13	Arkansas	BERETTA/COKER9375	BL
AR04016-4	1	G13	Arkansas	BERETTA/DK9410	BL
AR04025-3	1	G13	Arkansas	COKER9375/DK9410	BL
AR04029-4	1	G13	Arkansas	COKER9375/GF931241E16	BL
AR04029-4-5	1	G13	Arkansas	COKER9375/GF931241E16	BL

APPENDIX A cont.

Variety	No. Year	Nursery	Breeding Program	Pedigree	Type
AR04032-2	1	G13	Arkansas	COKER9375/LA94150-104-5-2	BL
AR04084-1-2	1	G14	Arkansas	LA9415D-104-5-2/DIXIE900	BL
AR04084-1-3	2	G13-14	Arkansas	LA9415D-104-5-2/DIXIE900	BL
AR04119-3	2	G13-14	Arkansas	AR857-1-1/LA841	BL
AR05009-12-1	1	S14	Arkansas	AR800-1-3-1/ENDURANCE	BL
AR05043-7-1	1	S14	Arkansas	BERETTA/VA00W-526	BL
AR05055-1-1	2	S14_G15	Arkansas	CO2W-32/P961341A3-1	BL
AR05055-12-2	1	S14	Arkansas	CO2W-32/P961341A3-1	BL
AR05067-1-2	1	S14	Arkansas	GA971127-14-6-6/PAT	BL
AR05067-2-2	1	S14	Arkansas	GA971127-14-6-6/PAT	BL
AR05074-12-1	1	G15	Arkansas	MO981020/P961341A-1-2	BL
AR05079-1-2	1	S14	Arkansas	P961341A3-1-2/AR93035-4-1	BL
AR05079-2-1	2	S14_G15	Arkansas	P961341A3-1-2/AR93035-4-1	BL
AR05079-2-2	1	G15	Arkansas	P961341A3-1-2/AR93035-4-1	BL
AR05080-4-1	1	S14	Arkansas	P961341A3-1-2/CO2W-22	BL
AR05080-6-1	1	S14	Arkansas	P961341A3-1-2/CO2W-22	BL
AR05085-1-1	1	S14	Arkansas	PAT/AR96141-4-1	BL
AR05094-4-1	2	S14_G15	Arkansas	TERRALTV8450/BERETTA	BL
AR05103-7-1	1	S14	Arkansas	VA00W-526/AR96052-4-2	BL
AR06009-3-4	1	S15	Arkansas	AR02136/AR930035-4-1	BL
AR06012-6-3	1	S15	Arkansas	AR02136/GA951216-2E26	BL
AR06017-6-2	2	S15_G16	Arkansas	AR800-1-3-1/COKER9663	BL
AR06024-7-2	1	S15	Arkansas	AR800-1-3-1/VA01W-476	BL
AR06031-7-4	1	S15	Arkansas	AR96077-7-2/AR01135	BL
AR06037-17-2	2	S15_G16	Arkansas	AR96077-7-2/VA00W-526	BL
AR06040-8-1	1	S15	Arkansas	AR97124-4-1/AR930035-4-1	BL
AR06045-16-4	1	S15	Arkansas	BESS/AR97124-4-1	BL
AR06046-10-3	2	S15_G16	Arkansas	BESS/PAT	BL
AR06050-7-2	2	S15_G16	Arkansas	COKER9553/AR98084-4-1	BL
AR06061-11-1	1	S15	Arkansas	P961341A3-1-2/VA01W-476	BL
AR06066-5-4	1	S15	Arkansas	PAT/GA971127-14-6-6	BL
AR07037-15-4	1	S16	Arkansas	AR97139-11-2/TERRALLA482	BL
AR07114-3-4	1	S16	Arkansas	AGS2020/LA978UC-36-1-1-B	BL
AR07119-9-1	1	S16	Arkansas	LA978UC-36-1-1-B/AR98001-5-1	BL
AR07122-16-1	1	S16	Arkansas	LA978UC-36-1-1-B/DK9577	BL
AR07122-9-1	1	S16	Arkansas	LA978UC-36-1-1-B/DK9577	BL
AR07139-11-1	1	S16	Arkansas	PAT/LA978UC-36-1-1-B	BL
AR98068-4-1	1	G09	Arkansas	JAYPEE/COKER9663	BL
AR98088-1-2	1	G09	Arkansas	PIONEER2580/JAYPEE	BL
AR98097-4-1	1	G10	Arkansas	PIONEER26R46/COKER9803	BL
AR990044-3-1	1	G09	Arkansas	AR664-21-1/WGRC35	BL
AR99009-3-2	1	G09	Arkansas	AR494B-2-2/DK9027	BL
AR99015-2-1	1	G09	Arkansas	AR584A-3-1/AR494B-2-2	BL

APPENDIX A cont.

Variety	No. Year	Nursery	Breeding Program	Pedigree	Type
AR99015-3-1	1	G10	Arkansas	AR584A-3-1/AR494B-2-2	BL
AR99015-3-3	1	G09	Arkansas	AR584A-3-1/AR494B-2-2	BL
AR99016-1-1	1	G08	Arkansas	AR584A-3-1/SABBE	BL
AR99016-1-2	1	G10	Arkansas	AR584A-3-1/SABBE	BL
AR99033-5-1	1	G08	Arkansas	AR664-21-1/AR835-21-1-2	BL
AR99033-6-2	1	G08	Arkansas	AR664-21-1/AR835-21-1-2	BL
AR99037-3-1	1	G08	Arkansas	AR664-21-1/CAH-106	BL
AR99039-2-1	1	G09	Arkansas	AR664-21-1/NC97BGTD7	BL
AR99093-1-1	1	G08	Arkansas	PAT/BETTY	BL
AR99095-10-2	1	G08	Arkansas	PAT/COKER9704	BL
AR99095-18-1	1	G08	Arkansas	PAT/COKER9704	BL
AR99110-11-2	1	G08	Arkansas	BETTY/JAYPEE	BL
AR99110-11-4	2	G08-09	Arkansas	BETTY/JAYPEE	BL
AR99110W-13-1	1	G08	Arkansas	BETTY/JAYPEE	BL
AR99114-2-1	1	G08	Arkansas	CAH-106/PAT	BL
AR99136-13-2	1	G08	Arkansas	COKER9663/WGRC40	BL
AR99138-7-1	1	G09	Arkansas	COKER9704/JAYPEE	BL
AR99160-4-1B	1	G10	Arkansas	ERNIE/PI590277//ERNIE	BL
AR99174-5-1	1	G09	Arkansas	FL8868/AR679-1-2	BL
AR99238-4-1	1	G09	Arkansas	NC97BGTD8/JAYPEE	BL
AR99263-7-1	1	G10	Arkansas	P92118/JAYPEE	BL
ARGA04494-11E49	1	G15	Arkansas	AWD99-5261/SS8641	BL
ARGA051160-14LE31	1	S16	Arkansas	SC996284/SS8641//SS8641	BL
ARGA061147-23-6-2	1	S15	Arkansas	SS8641/AGS2035	BL
ARGA06411-9-3-4	1	S15	Arkansas	GA96229-3E39/AGS2031	BL
ARGA06473-9-4-4	1	S15	Arkansas	01063-1-3-2/MCINTOSH//SS8641	BL
ARGA071614-14E34	1	S16	Arkansas	GA991371-6E13*2/AGS2031	BL
ARGE07-1374-17-5-5	1	S14	Arkansas	AGS2060/VA05W-500//VA01-205	BL
ARGE07-1380-4-2-4	1	S14	Arkansas	LA482/LW00058-44//VA01-205	BL
ARLA05009F-1-4	1	S16	Arkansas	DK9577/LA841	BL
ARLA06146E-1-4	1	S16	Arkansas	JAMESTOWN/AGS2060	BL
ARLA07019C-20-4	1	S16	Arkansas	AR98003-7-1/AGS2060	BL
ARLA07053C-14-4	1	S16	Arkansas	GA98244-1-14-5-4/MAGNOLIA	BL
ARLA07084C-10-1	1	S16	Arkansas	LA98094BUB-58-5/AGS2060//LA99005UC-31-3	BL
ARLA07133C-19-4	1	S16	Arkansas	LA99005UC-31-3/VA05W-500	BL
ARLA07133C-3-4	1	S16	Arkansas	LA99005UC-31-3/VA05W-500	BL
ARNC09-22402	1	G14	Arkansas	NC99-18235/NC00-16203//DOMINION	BL
DH11SRW070-14	1	G16	Virginia	GA00067-8E35/SHIRLEY	BL
DH11SRW070-28	1	G16	Virginia	GA00067-8E35/SHIRLEY	BL
FL01005-K5	1	G08	Florida	LA841/GA92601-17-4-9	BL
FL01108C-K2	1	G08	Florida	PIONEER26R61/LA92283C64-1	BL
FL02006C-K1	1	G08	Florida	NC98-24710/AGS2000//PIONEER26R61	BL
FL02006C-K4	1	G08	Florida	NC98-24710/AGS2000//PIONEER26R61	BL
FL02036C-K6	1	G08	Florida	95158BUB69-3/AGS2000	BL

APPENDIX A cont.

Variety	No. Year	Nursery	Breeding Program	Pedigree	Type
FL02154C-K3	1	G08	Florida	ND2928/FL95IN2762	BL
FL03169D-58	1	G10	Florida	PIONEER26R61/GA951079A25	BL
FL04363E-P19	1	G12	Florida	FL03155/AGS2060	BL
FL04363E-P23	1	G13	Florida	FL03155/AGS2060	BL
FL99077D-E29-K4	1	G08	Florida	BL940026/AGS2000	BL
FLLA09061C-8	1	S16	Florida	LA01-425/LA95135	BL
FLLA09073C-19	1	S16	Florida	LA01034D-235-1-C/GA991371-6E13	BL
FLLA09089C-12	1	S16	Florida	GA061654-G14-G2/AGS2060//LA01140D-70	BL
FLLA09154C-17	1	S16	Florida	NC05-21642/GA031238-DH7-7A28	BL
FLLA09167C-35	1	S16	Florida	NC06-27/LA95135	BL
FLLA09180C-31	1	S16	Florida	GA05467-4-G1-G2/PIONEER26R61//VA01-205	BL
FLLA09184C-24	1	S16	Florida	VA03W-509/LA01140D-70	BL
FLLA09189C-41	1	S16	Florida	VA03W-509/LA95135	BL
FLLA09298C-48	1	S16	Florida	P04287A1-10/LA95135	BL
FLLW08145D-20	1	S16	Florida	CK9700/LA95135//LA98094BUB-58-5	BL
FLLW08184D-1	1	S16	Florida	NC03-5921/SS8641//LA841	BL
FLLW08195D-44	1	S16	Florida	PIONEER26R61/LA99005UC-31-3//AGS2026/LA99005UC-31-3	BL
FLLW08219D-35	1	S16	Florida	SS8641/NC03-5921//LA841	BL
GA00034-7A17	1	G08	Georgia	93132/921204	BL
GA00067-8E35	1	G09	Georgia	921204/AGS2000	BL
GA001138-8E36	1	G09	Georgia	GA961581/PIONEER26R61	BL
GA001138-8E37	1	G09	Georgia	GA961581/PIONEER26R61	BL
GA001142-9E23	1	G10	Georgia	GA931520/PIONEER26R61	BL
GA001142-9E24	1	G10	Georgia	GA931520/PIONEER26R61	BL
GA001169-7E15	1	G08	Georgia	96667/AGS2000//96667	BL
GA001169-G1-10-6-3	1	G08	Georgia	96667/AGS2000//96667	BL
GA001170-7E26	1	G08	Georgia	PIONEER26R61/96667//AGS2000	BL
GA00138-7A6	1	G08	Georgia	AGS2000/2*GA881130//GA931520	BL
GA001492-7E9	1	G08	Georgia	941396/AGS2000//AGS2000	BL
GA00190-7A14	1	G08	Georgia	931298/GA92601	BL
GA011027-8LE24	1	G09	Georgia	IN92201/AGS2000//PIONEERXW692	BL
GA011124-8LE28	1	G09	Georgia	93322/VA270//93322	BL
GA011124-8LE32	1	G09	Georgia	93322/VA270//93322	BL
GA011174-8A9	1	G09	Georgia	961526/961565	BL
GA011264-7E13	1	G08	Georgia	AGS2000*3/GA931433	BL
GA011373-10E36	1	G11	Georgia	PIONEERXW692/AGS2000//GA961565-2E46	BL
GA011446-9LE35	1	G10	Georgia	GA941365-D23/GA941238	BL
GA011493-8E18	1	G09	Georgia	IN92201/AGS2000//PIONEERXW692	BL
GA011636-2	1	G08	Georgia	AGS2000*3/93322	BL
GA021087-9LE33	1	G10	Georgia	PIONEER26R61/AGS2010	BL
GA021245-9E16	1	G10	Georgia	FL93024-1/PIONEER26R61//AGS2485	BL
GA021282-8A2	1	G09	Georgia	94261/CHARTER	BL
GA021338-9E11	1	G10	Georgia	PIONEER26R38/GA941238//AGS2000	BL
GA021338-9E15	1	G10	Georgia	PIONEER26R38/GA941238//AGS2000	BL

APPENDIX A cont.

Variety	No. Year	Nursery	Breeding Program	Pedigree	Type
GA021338-9E4	1	G10	Georgia	PIONEER26R38/GA941238//AGS2000	BL
GA02178-9E25	1	G10	Georgia	GA941238/AGS2485	BL
GA02264-8LE17	1	G09	Georgia	USG3592/GA951231-1-1-1	BL
GA02328-8A21	1	G09	Georgia	P2684/3*AGS2000//AGS2000*2/GA84202	BL
GA02343-9LE5	1	G10	Georgia	AGS2000*3/GA941433//GA931520/AGS2000	BL
GA03060-9E29	1	G10	Georgia	AGS2000*3/GA84202	BL
GA031086-10E26	1	G11	Georgia	PIONEER26R24/GA961565-2E46//GA941208-2E35	BL
GA031134-10E29	1	G11	Georgia	PIONEER26R38/2*GA961565-2E46	BL
GA031215-10E42	1	G11	Georgia	GA971507/GA97342	BL
GA031238-DH7-7A28	1	G08	Georgia	GA96229-3E39/GA951395-3E25	BL
GA031238-DH7-7E34	1	G08	Georgia	GA96229-3E39/GA951395-3E25	BL
GA031238-LE33	1	G11	Georgia	SS8641/AGS2031	BL
GA031257-10E34	1	G11	Georgia	SS8641/GA961591	BL
GA031257-10E41	1	G11	Georgia	SS8641/GA961591	BL
GA03185-12LE29	2	G13_S14	Georgia	AGS2000*2/GA84202//GA97173-1	BL
GA03389-10E36	1	G11	Georgia	SS8641/AGS2010	BL
GA03437-10E33	1	G11	Georgia	SC996284/GA961581	BL
GA03564-10E25	1	G11	Georgia	SS8641/4/AGS2000*3/GA931433//PIONEER2684/3*AGS2000/3/AGS2000	BL
GA03564-12E6	1	G13	Georgia	SS8641/4/AGS2000*3/GA931433//PIONEER2484/3*AGS2000	BL
GA03564-9EE42	1	G10	Georgia	SS8641/4/AGS2000*3//PIONEER2684/*3AGS2000/3/AGS2000	BL
GA041052-11E51	1	G12	Georgia	931233-28-2-2/USG3592	BL
GA041229-13E55	1	S14	Georgia	GA971061-59-3-6*2/961171-2-2-2	BL
GA041271-10E39	1	G11	Georgia	MCCORMICK/GA951216-2E26//SS8641	BL
GA041272-12E42	1	G13	Georgia	ARK839-25-8/SS8641//SS8641	BL
GA041293-11E54	1	G12	Georgia	PIONEER26R61/GA96229-3E39//SS8641	BL
GA041293-11LE37	1	G12	Georgia	PIONEER26R61/SS8641//SS8641	BL
GA041296-11LE39	1	G12	Georgia	MCCORMICK/GA961591-17-1-5//GA951395-3A31	BL
GA041323-11E63	1	G12	Georgia	95652-2E56/GA961591-3E42	BL
GA04151-11E26	1	G12	Georgia	98302-17-1-4/SC996284	BL
GA04244-11E1	1	G12	Georgia	GA96229-3E39/AGS2000	BL
GA04244-12LE16	1	G13	Georgia	SS8641/AGS2000	BL
GA04268-12E4	1	G13	Georgia	GA99231-1/GA971061-59	BL
GA04417-11E21	1	G12	Georgia	GA961565-2E46/AGS2485//SS8641	BL
GA04417-12E33	1	G13	Georgia	GA961565-2E46/AGS2485//SS8641	BL
GA04434-11E44	1	G12	Georgia	GA961565-2E46/AGS2485//SS8641	BL
GA04434-12LE28	1	G13	Georgia	GA961565-2E46/AGS2485//SS8641	BL
GA04434-13E52	3	G14-15_S14	Georgia	GA961565-2E46/AGS2485//SS8641	BL
GA04444-11LE25	1	G12	Georgia	SS8641/GA951395-3E27	BL
GA04500-11LE11	1	G12	Georgia	97531-2-11/GA011636-G1-G5-G2	BL
GA04510-11LE24	2	G12-14	Georgia	GA961591-3E42/SS8641	BL
GA04570-10E46	1	G11	Georgia	00440/3/PIONEER2684/3*AGS2000//AGS2000*2/GA84202	BL
GA051033-13LE14	2	G14_S14	Georgia	SS8641/GA941208-3E35//SS8641	BL
GA051102-13LE43	2	S14_G15	Georgia	SS8641/GA941208-3E35//SS8641	BL
GA051207-14E53	1	S15	Georgia	AGS2000/SC996284//IN981359C1	BL

APPENDIX A cont.

Variety	No. Year	Nursery	Breeding Program	Pedigree	Type
GA051304-12E28	1	G13	Georgia	SS8641/USG3295	BL
GA051335-13E13	1	S14	Georgia	SS8641/96562//SS8641	BL
GA051335-13LE19	1	S14	Georgia	SS8641/96562//SS8641	BL
GA051754-12LE13	1	G13	Georgia	NC98-24710/3*AGS2000//00440	BL
GA05304-12E35	2	G13_S14	Georgia	GA011450/AGS2010//OGLETHORPE	BL
GA05450-EL52	1	S16	Georgia	GA96229-3E39/C9553	BL
GA05450-LE41	1	S16	Georgia	GA96229-3E39/C9553	BL
GA061082-13E24	1	S14	Georgia	GA951231-4A15/SS8641//GA951231-4A15	BL
GA061086-14LE23	2	S15_G16	Georgia	SS8641/961176-3A48	BL
GA061096-14E3	2	S15_G16	Georgia	961171-4E21/TRUMAN//SS8641	BL
GA06112-13EE16	1	G16	Georgia	96693-4E15/96629-3E39//AGS2020	BL
GA061157-LE44	1	S16	Georgia	AGS2020*2/GA96229-3E39	BL
GA061158-14LE11	2	S15_G16	Georgia	GA011341-1/AGS2020	BL
GA061349-13E4	2	G14_S14	Georgia	SS8641/981622-1-4-4//SS8641	BL
GA061349-13E5	3	G14-15_S14	Georgia	SS8641/981622-1-4-4//SS8641	BL
GA061349-13LE29	2	S14_G15	Georgia	96629-3E39/981622-1-4-4//SS8641	BL
GA061349-13LE31	1	G15	Georgia	96629-3E39/981622-1-4-4//SS8641	BL
GA061349-14LE1	1	S15	Georgia	SS8641/981622-1-4-4//SS8641	BL
GA061349-LE31	2	G14_S14	Georgia	96629-3E39/981622-1-4-4//SS8641	BL
GA061471-LE38	1	S16	Georgia	AGS2020/6/061636/5/Yr15//99406/3/AGS2000/4/97186	BL
GA06283-LE25	1	S16	Georgia	OGLETHORPE/981592-8//AGS2035	BL
GA06474-EL56	1	S16	Georgia	AGS2035/98302-1//SS8641	BL
GA06478-13E23	1	S14	Georgia	981622-10/001240-7//SS8641	BL
GA06489-14LE8	1	S15	Georgia	PANOLA/SS8641//SS8641	BL
GA06493-13LE6	3	G14-15_S14	Georgia	981394-16-2-1/981622-10-2-3	BL
GA07026-14LE4	1	S15	Georgia	AGS2020/USG3555//AGS2020	BL
GA071012-14E6	1	S15	Georgia	GA991371-6E12/SS8641	BL
GA071171-EL64ES8	1	S16	Georgia	JAMESTOWN/GA991371-6E12	BL
GA07144-LE16	1	S16	Georgia	AGS2035/011264	BL
GA071630-12LE9	1	G13	Georgia	GA011636-G1-G5/GA991371-12//GA991371-12	BL
GA07169-14LE24	2	S15_G16	Georgia	AGS2020/021737-B-5//OGLETHORPE	BL
GA07192-14E9	2	S15_G16	Georgia	D00-6874-9/SS8641//AGS2020	BL
GA07248-14E18	1	S15	Georgia	AGS2020/021737-B-5//SS8641	BL
GA07270-12E15	1	G13	Georgia	GA98249/2*AGS2035//GA001169-G1	BL
GA07353-14E19	1	S15	Georgia	SS8641/OGLETHORPE//GA991371-6E13	BL
GA07592-14E8	1	S15	Georgia	051396/GA991371-6E11	BL
GA081104-EL23	1	S16	Georgia	GA991371-6E13/011177-9-4	BL
GA081113-EL8	1	S16	Georgia	BALDWIN/00190-1-5-1//OGLETHORPE	BL
GA081446-EL47	1	S16	Georgia	02655-7-7/02328-G1-G1	BL
GA08261-EL7	1	S16	Georgia	JAMESTOWN/SS8641//BALDWIN	BL
GA08391-EL19	1	S16	Georgia	GA991227-6A33/GA001169-G1	BL
GA08510-EL9	1	S16	Georgia	USG3120/SS8641//USG3120	BL
GA08535-LE29	1	S16	Georgia	BALDWIN/GA061654BDV3//BALDWIN	BL
GA981131-7E33	1	G08	Georgia	91215/GA901146	BL

APPENDIX A cont.

Variety	No. Year	Nursery	Breeding Program	Pedigree	Type
GA981394-8A37	1	G09	Georgia	861278/85410//PIONEER2684	BL
GAJT020-14E47	1	S15	Georgia	JAMESTOWN/AGS2026	BL
GAJT141-14E45	1	S15	Georgia	JAMESTOWN/AGS2026	BL
HILLIARD	2	G16_S16	Virginia	PIONEER25R47/JAMESTOWN	CC
LA01005D-2-2-C	1	G09	Louisiana	LA841/GA92601-17-4-5	BL
LA01029D-139-3-C	1	G08	Louisiana	TRIBUTE/LA85411D4-1-1-2	BL
LA01034D-235-1-C	1	G08	Louisiana	TRIBUTE/LA94242D4-1	BL
LA01034D-42-3-C	1	G10	Louisiana	LA94214D200/LA841	BL
LA01035D-207-3-B	1	G08	Louisiana	TRIBUTE/AGS2000	BL
LA01059D-127-3-2	1	G10	Louisiana	GA92412-15-5-4/LA841	BL
LA01069D-23-4-4	1	G10	Louisiana	INW9811/PIONEER26R61	BL
LA01108D-71-1-B	1	G08	Louisiana	PIONEER26R61/LA92283C64-1	BL
LA01110C-J10	1	G09	Louisiana	PIONEER26R61/LA94242D24-4	BL
LA01110D-100-6-4	1	G09	Louisiana	PIONEER26R61/LA94242D24-4	BL
LA01110D-150	1	G08	Louisiana	PIONEER26R61/LA94242D24-4	BL
LA01110D-181-6-B	1	G09	Louisiana	PIONEER26R61/LA94242D24-4	BL
LA01110D-208-5-C	1	G09	Louisiana	PIONEER26R61/LA94242D24-4	BL
LA01110D-251	1	G08	Louisiana	PIONEER26R61/LA94242D24-4	BL
LA01110D-81-1-B	1	G08	Louisiana	PIONEER26R61/LA94242D24-4	BL
LA01110D-84-1-C	1	G08	Louisiana	PIONEER26R61/LA94242D24-4	BL
LA01110D-84-2-C	1	G09	Louisiana	PIONEER26R61/LA94242D24-4	BL
LA01113D-44	1	G08	Louisiana	PIONEER26R61/NC98-24710	BL
LA01138D-55	1	G08	Louisiana	LA00056/AGS2000	BL
LA01139D-116	1	G08	Louisiana	LA841/LA422//PIONEER26R61	BL
LA01139D-56-7-3	1	G10	Louisiana	LA841/LA422//PIONEER26R61	BL
LA01139D-86-6-2	1	G09	Louisiana	LA841/LA422//PIONEER26R61	BL
LA01140D-163	1	G08	Louisiana	LA00056/PIONEER26R61	BL
LA01145D-123-5-C	1	G10	Louisiana	LA841/U1254-6-2-7//LA841	BL
LA01164D-43-7-B	1	G09	Louisiana	LA422/FUTAI8944//PIONEER26R61	BL
LA01172D-27-5-4	1	G09	Louisiana	MCNAIR1003/GA90524E35//LA841	BL
LA02006E239	1	G10	Louisiana	NC98-24710/AGS2000//PIONEER26R61	BL
LA02007E227	1	G10	Louisiana	NC98-24710/LA422//VA98W-590	BL
LA02015E201	1	G10	Louisiana	VA99W-169/PIONEER26R61//LA94242D4-4	BL
LA02015E58	1	G11	Louisiana	VA99W-169/PIONEER26R61//LA94242D4-4	BL
LA02024E12	1	G11	Louisiana	LA85411/FL95IN2762	BL
LA02024E7	1	G11	Louisiana	LA85411/FL95IN2762	BL
LA02150E-35	1	G10	Louisiana	NC99-13022/LA9354D9-3-1	BL
LA03012E-27	1	G11	Louisiana	LA92283C64-1/ARLA97-1047-4-2	BL
LA03045E-4	1	G12	Louisiana	LA95361CA18-1/LA95176D56-2	BL
LA03091E-63	1	G12	Louisiana	LA97113UC-124-3/PIONEER26R61	BL
LA03118E117	1	G11	Louisiana	APD99-5627/LA9546D5-1-3-B	BL
LA03136E71	1	G11	Louisiana	ARLA97-1047-4-2/LA95125BUB73-1	BL
LA03148E12	1	G11	Louisiana	AWD99-5528/AGS2060	BL
LA03155D-P13	1	G11	Louisiana	DK9410/LA95361CA8-2-2	BL

APPENDIX A cont.

Variety	No. Year	Nursery	Breeding Program	Pedigree	Type
LA03161D-P1	1	G11	Louisiana	FL95IN2762/AGS2060	BL
LA03200E-23	2	G12-13	Louisiana	NC98-24710/PIONEER26R61	BL
LA03217E2	1	G11	Louisiana	RO23/PIONEER26R61	BL
LA03224E-39	1	G14	Louisiana	SC996284/LA97113UC-124-3	BL
LA04013D-111	1	G12	Louisiana	LA95135/B990399	BL
LA04013D-142	1	G11	Louisiana	LA95135/B990399	BL
LA04041D-10	1	G11	Louisiana	AGS2060/GA951079A25	BL
LA04041D-85	1	G12	Louisiana	AGS2060/GA951079A25	BL
LA04089D-P10	1	G10	Louisiana	LA03205/AGS2060	BL
LA05009D-35	1	G14	Louisiana	DK9577/LA841	BL
LA05027D-26	1	G13	Louisiana	GA951395-3E27/PIONEER26R61	BL
LA05032D-10	1	S14	Louisiana	SS8641/LA841	BL
LA05032D-136	2	G13_S14	Louisiana	SS8641/LA841	BL
LA05038D-105	1	G12	Louisiana	SS8641/PIONEER26R61	BL
LA05038D-51	1	G12	Louisiana	SS8641/PIONEER26R61	BL
LA05038F-P1	1	S15	Louisiana	SS8641/PIONEER26R61	BL
LA05079F-P01	1	G13	Louisiana	GA971127-14-6-6/LA841//ARLA97-1033-3-5	BL
LA05079F-P05	1	G13	Louisiana	GA971127-14-6-6/LA841//ARLA97-1033-3-5	BL
LA05120F-P12	1	G13	Louisiana	LA98094BUB-58-5/LA841	BL
LA05130D-P5	1	G12	Louisiana	LA98149BUB-3-4-B/SS8641	BL
LA05132F-P09	1	G12	Louisiana	LA98149BUB-3-4-B/AGS2060	BL
LA05145D-118	1	S14	Louisiana	JAMESTOWN/LA97113UC-124	BL
LA05145D-16	1	S14	Louisiana	JAMESTOWN/LA97113UC-124	BL
LA05145D-17	1	S14	Louisiana	JAMESTOWN/LA97113UC-124	BL
LA05145D-5	1	S14	Louisiana	JAMESTOWN/LA97113UC-124	BL
LA05145D-66	1	S14	Louisiana	JAMESTOWN/LA97113UC-124	BL
LA06007E-P04	1	G12	Louisiana	APCKB02-8443/LA95135	BL
LA06020E-P16	2	G13_S14	Louisiana	DK9577/PIONEER26R61	BL
LA06036E-P04	1	G13	Louisiana	GA951216-2E26/AGS2060	BL
LA06052E-P07	1	G12	Louisiana	SS8641/LA95135	BL
LA06069E-P01	1	G13	Louisiana	LA95135/VA02W-713	BL
LA06146E-P4	5	G13 to 16_S16	Louisiana	JAMESTOWN/AGS2060	BL
LA06149C-P7	1	S14	Louisiana	JAMESTOWN/LA98094BUB-58-5	BL
LA07040D-P01	2	G13_S14	Louisiana	SS8641/APCB02-8486	BL
LA07085CW-P4	1	S15	Louisiana	LA98094BUB-58-5/AGS2060//VA02W-713	BL
LA07102CW-P10	1	G14	Louisiana	JAMESTOWN/AGS2060//LA841	BL
LA07102CW-P3	1	G14	Louisiana	JAMESTOWN/AGS2060//LA841	BL
LA07128C-91	1	S14	Louisiana	LA98205D-17-2-4/LA01110D-88	BL
LA07178C-44	1	S14	Louisiana	JAMESTOWN/LA95135	BL
LA07599E-21	1	S14	Louisiana	LA96140BUA70-2/GA98401-1B-22-3-3	BL
LA08062C-P5	1	S15	Louisiana	GA00138-31-1-3/LA821	BL
LA08090C-26-3	1	S16	Louisiana	AGS2020/AGS2060	BL
LA08090C-9-1	1	S16	Louisiana	AGS2020/AGS2060	BL
LA08090C-9-2	1	S15	Louisiana	AGS2020/AGS2060	BL

APPENDIX A cont.

Variety	No. Year	Nursery	Breeding Program	Pedigree	Type
LA08095C-23	1	S15	Louisiana	AGS2020/LA821	BL
LA08095C-37	2	S15_G16	Louisiana	AGS2020/LA821	BL
LA08096C-P10	1	G14	Louisiana	AR98003-7-1/LA841//JAMESTOWN	BL
LA08115C-30	1	G15	Louisiana	GA98244-1-14-5-4/LA841//LA01113D-44	BL
LA08155C-67	1	S15	Louisiana	SS8641/CK9700//LA841	BL
LA08218C-57	2	S15_G16	Louisiana	LA841/GA05467-4-G1-G2	BL
LA08221C-23	1	S14	Louisiana	GA9811622-1-40-1/APCB02-8486//LA841	BL
LA08234D-18	1	S16	Louisiana	APCB02-8486/AGS2026//LA95135	BL
LA08240C-23	1	S15	Louisiana	LA95135/LA98094BUB-58-5	BL
LA08265C-50	2	S15_G16	Louisiana	JAMESTOWN/SS8641//PIONEER26R61	BL
LA09011UB-2	1	G15	Louisiana	AGS2026/VA05W-510	BL
LA09048C-P7	1	S15	Louisiana	LA821/MO050146//BALDWIN	BL
LA09050C-P2	1	S15	Louisiana	BALDWIN/LA841	BL
LA09056C-P10	1	S15	Louisiana	GA991371-6E13/AGS2060	BL
LA09122UB-43	1	S15	Louisiana	LA841/GA991371-6E13	BL
LA09179C-5	1	S16	Louisiana	VA01-205/LA01140D-70	BL
LA09202C-34	1	S16	Louisiana	AGS2031/AGS2060	BL
LA09225C-33	1	G16	Louisiana	LA01139D-56-1/GA001492-7E9	BL
LA09263C-P2	1	S15	Louisiana	LA841/VA02W-713//GA051754-G1-1-8-1-G1-G1	BL
LA09264C-P2	2	G15-16	Louisiana	LA841/VA02W-713//LA01139D-56-1	BL
LA09264C-P5	2	G15-16	Louisiana	LA841/VA02W-713//LA01139D-56-1	BL
LA9050C-P4	1	S16	Louisiana	BALDWIN/LA841	BL
LA95135	2	G09-10	Louisiana	CL850643/PIONEER2548//COKER-9877/3/F-302/COKER-762	BL
LANC8170-41-2	1	G15	Louisiana	NC03-11458/BESS//SS8641	BL
NC04-20417	1	G10	North Carolina	P92118/VA94-52-25//NC96BGTD2	BL
NC04-22866	1	G08	North Carolina	P86958/C9835//NC94-7197	BL
NC05-19684	1	G08	North Carolina	P92118/VA94-52-25//NC96BGTD2	BL
NC05-19896	1	G09	North Carolina	BURR/NC96BGTA6//NATCHEZ	BL
NC05-20276	1	G08	North Carolina	P86300/NC95-22426	BL
NC05-20671	2	G08-09	North Carolina	P92188/NC95-22365//ROANE	BL
NC05-21090	1	G08	North Carolina	BURR/NC96BGTA6//NATCHEZ	BL
NC05-21642	1	G08	North Carolina	NC96-14629/PIONEER2643	BL
NC05-21937	1	G08	North Carolina	SHAAN85-15/SS520//NC-NEUSE	BL
NC05-22804	2	G08-09	North Carolina	P92118/VA94-52-25//NC96BGTD2	BL
NC05-22975	1	G08	North Carolina	ROBERTS/NC96BGTA6//NC95-29317/VA96-54-225	BL
NC05-23015	1	G09	North Carolina	BURR/NC96BGTA6//NATCHEZ	BL
NC05-23945	1	G08	North Carolina	NC94-7197/PIONNER2643	BL
NC05-23993	1	G09	North Carolina	NC96-14629/PIONEER2643	BL
NC05-24112	1	G08	North Carolina	TRIBUTE/NC98-25388	BL
NC05-24757	1	G08	North Carolina	NC96-14629/PIONEER2643	BL
NC06-19556	1	G09	North Carolina	B931167/ROANE//CHOPTANK	BL
NC06-20244	1	G09	North Carolina	JACKSON/NC95-22365//NC96-13965	BL

APPENDIX A cont.

Variety	No. Year	Nursery	Breeding Program	Pedigree	Type
NC06-20288	1	G09	North Carolina	JACKSON/NC95-25305//NC96-13965	BL
NC06-20359	1	G10	North Carolina	NC96-13848/B951008//NC96-14629	BL
NC06-20401	2	G09-11	North Carolina	NC95CLB-14543/REN3260//NC96-14629	BL
NC06-21245	1	G09	North Carolina	NC94-7197/NC99BGTAG11//NC96-13374	BL
NC06-22003	2	G09-10	North Carolina	NC95CLB-14543/REN3260//NC96-14629	BL
NC06-22379	1	G09	North Carolina	C9704/NC-NEUSE//VA96W-348	BL
NC07-20850	2	G10-11	North Carolina	UW85-1352/C9704//NC96-13155	BL
NC07-21020	1	G10	North Carolina	NC96-13155/ROANE//TRIBUTE	BL
NC07-21172	1	G10	North Carolina	NC96-13965/C9184//PIONEER2643	BL
NC07-22432	2	G10-11	North Carolina	NC96-14094/PIONEER26R24//TRIBUTE	BL
NC07-22517	1	G10	North Carolina	TRIBUTE/NC98-25380//NC96-13155	BL
NC07-23880	2	G10-11	North Carolina	PIONEER26R24/VA97W-375//NC96-13965	BL
NC07-24337	1	G10	North Carolina	NC98-24710/NC98-26541	BL
NC07-24445	2	G10-11	North Carolina	USG3209/NC98-26541	BL
NC07-25169	2	G10-11	North Carolina	TRIBUTE/NC98-25380//NC96-13155	BL
NC08-140(Bdv2)	1	G12	North Carolina	PIONEER26R61/TC14Spear2289B//NC00-16203	BL
NC08-21273	1	G11	North Carolina	B970499/NC98-26541//USG3209	BL
NC08-23089	1	G11	North Carolina	NC97-10076/C9704//PIONEER26R61	BL
NC08-23090	1	G11	North Carolina	NC97-10076/C9704//PIONEER26R61	BL
NC08-23323	1	G11	North Carolina	B960164/NC94-7197//MCCORMICK	BL
NC08-23324	1	G11	North Carolina	B960164/NC94-7197//MCCORMICK	BL
NC08-23383	1	G11	North Carolina	PATTON/NC96-13374//MCCORMICK	BL
NC09-19946	1	G12	North Carolina	PIONEER26R24/NC96-13965//NC00-16203	BL
NC09-19966	1	G12	North Carolina	PIONEER26R24/NC96-13965//NC00-16203	BL
NC09-20036	1	G13	North Carolina	NC00-14622/NC98-26143//USG3209	BL
NC09-20765	2	G12-14	North Carolina	PIONEER26R24/NC96-13965//NC00-16203	BL
NC09-20768	1	G12	North Carolina	PIONEER26R24/NC96-13965//NC00-16203	BL
NC09-20932	1	G12	North Carolina	TREGO/NC99BGTAG11//NC98-13296W	BL
NC09-20986(Fhb1)	2	G13-14	North Carolina	NC00-15332/VA01-476//DOMINION	BL
NC09-21230	1	G12	North Carolina	PIONEER26R24/NC96-13965//NC00-16203	BL
NC09-21251	1	G12	North Carolina	PIONEER26R24/NC96-13965//NC00-16203	BL
NC09-21256	1	G12	North Carolina	PIONEER26R24/NC96-13965//NC00-16203	BL
NC09-21916	1	G13	North Carolina	B990081/NC96BGTA6//MCCORMICK	BL
NC09-21953	1	G12	North Carolina	NC99-18235/NC00-16203//DOMINION	BL
NC09-22206	1	G12	North Carolina	PIONEER26R24/NC96-13965//NC00-16203	BL
NC09-22368	1	G14	North Carolina	TREGO/NC99BGTAG11//NC98-13296W	BL
NC09-22402	1	G12	North Carolina	NC99-18235/NC00-16203//DOMINION	BL
NC10014-9B	1	S16	North Carolina	NC06-19896/NC08-140	BL
NC10034-11	1	S15	North Carolina	NC-YADKIN/SHIRLEY	BL

APPENDIX A cont.

Variety	No. Year	Nursery	Breeding Program	Pedigree	Type
NC10034-26	1	S16	North Carolina	NC-YADKIN/SHIRLEY	BL
NC10034-43	1	S16	North Carolina	NC-YADKIN/SHIRLEY	BL
NC10034-47	2	G16_S16	North Carolina	NC-YADKIN/SHIRLEY	BL
NC10034-50	2	G16_S16	North Carolina	NC-YADKIN/SHIRLEY	BL
NC10034-86	1	S16	North Carolina	NC-YADKIN/SHIRLEY	BL
NC10080-122	1	S15	North Carolina	GA05052-G1-21/JAMESTOWN	BL
NC10-22592	2	G13-14	North Carolina	NC97BGTAB9/NC00-14622//C9184	BL
NC10-22614	1	G13	North Carolina	NC09MDD14/NC00-16203//MCCORMICK	BL
NC10-22642	1	G13	North Carolina	MCCORMICK/NC00-16203//NC-NEUSE	BL
NC10-23407	1	G13	North Carolina	NC96-13296W/USG3209//GA96229-3E39	BL
NC10-23663(Bdv2)	1	G13	North Carolina	PIONEER2691/TC14Spear289B//NC00-16203	BL
NC10-23720	1	G13	North Carolina	DOMINION/ARGE97-1042//GA96229-3E39	BL
NC10-23730	1	G13	North Carolina	NC02-14897/PIONEER26R61//SS8641	BL
NC10-24889	1	G14	North Carolina	C9184/NC00-14622//C9553	BL
NC10-25196	1	G13	North Carolina	USG3209/NC02-14897//SS8641	BL
NC11-20369	1	S14	North Carolina	NC98-24182/NC-NEUSE//NC01-27764	BL
NC11-20553	2	S14_G15	North Carolina	MCCORMICK/P961341A3//SS8641	BL
NC11-21307	1	S14	North Carolina	C9553/GA96229-3E39	BL
NC11-21447	1	S14	North Carolina	NC03-11457/NC-NEUSE//SS8641	BL
NC11-21899	2	S14_G15	North Carolina	C9184/DOMINION//SS8641	BL
NC11-21982	2	S14_G15	North Carolina	DOMINION/NC-NEUSE//AGS2000	BL
NC11-22289	2	S14_G15	North Carolina	NC97BGTD7/NC-NEUSE//C9511	BL
NC11-22291	1	S14	North Carolina	NC97BGTD7/NC-NEUSE//C9511	BL
NC11-22385	1	S14	North Carolina	ISIDOR/MCCORMICK//SS8641	BL
NC11-22715	1	S14	North Carolina	NC-CAPEFEAR/SS8641	BL
NC11-23084	1	G15	North Carolina	CO25-17/VA04W-478//NC-NEUSE	BL
NC11-23321	1	S14	North Carolina	NC96BGTD3/AGS2010//SS8641	BL
NC12-20785	1	S15	North Carolina	GA951395-3E27/MCCORMICK//SS8641	BL
NC12-20835	1	S15	North Carolina	GA951395-3E27/NC01-27750//USG3555	BL
NC12-20850	1	S15	North Carolina	GA951395-3E27/NC01-27750//USG3555	BL
NC12-21164	1	S15	North Carolina	NC-NEUSE/C9511//NC02-11158	BL
NC12-21166	1	S15	North Carolina	NC-NEUSE/C9511//NC02-11158	BL
NC12-21224	1	S15	North Carolina	NC00-15332/NCA4G//AGS2031	BL
NC12-21568	1	S15	North Carolina	IL96-6472/VA04W-561//SS8641	BL
NC12-22686	1	S15	North Carolina	MCCORMICK/SS8641//C9553	BL
NC12-22844	1	S15	North Carolina	NC-NEUSE/C9511//NC02-11158	BL
NC12-22848	1	S15	North Carolina	NC-NEUSE/C9511//NC02-11158	BL
NC12-23573	1	S15	North Carolina	BESS/SS8641	BL
NC12-23576	1	S15	North Carolina	BESS/SS8641	BL

APPENDIX A cont.

Variety	No. Year	Nursery	Breeding Program	Pedigree	Type
NC13-20076	1	S16	North Carolina	GA951231-4E29/NCAG11//JAMESTOWN	BL
NC13-20227	1	S16	North Carolina	NC06BGTAG12/USG3209//JAMESTOWN	BL
NC13-20278	2	G16_S16	North Carolina	USG3555/NC-CAPEFEAR//NC04-15460	BL
NC13-20332	1	G16	North Carolina	D1AG11/VA02W-713//NC04-15460	BL
NC13-20539	1	G16	North Carolina	NC03-11158/MCCORMICK//NC-NEUSE	BL
NC13-21213	2	G16_S16	North Carolina	OGLETHORPE/JAMESTOWN	BL
NC13-21217	1	S16	North Carolina	OGLETHORPE/JAMESTOWN	BL
NC13-21445	1	S16	North Carolina	VA04W-259/JAMESTOWN	BL
NC13-21987	2	G16_S16	North Carolina	NC06BGTAG12/USG3209//JAMESTOWN	BL
NC13-22649	1	G16	North Carolina	VA02W-713/BESS//NC04-15533	BL
NC13-22836	1	S16	North Carolina	NC03-11465/NC02-1957//JAMESTOWN	BL
NC13-23443	2	G16_S16	North Carolina	NC04-15533/VA05W-500//VA05W-108	BL
NC8170-4-3(Fhb1)	1	G13	North Carolina	NC03-11458/BESS//SS8641	BL
NC8248-1	1	S14	North Carolina	No Record	BL
NC8401-5	1	S14	North Carolina	NC03-11465/VA04W-264	BL
NC8932-12	1	G14	North Carolina	NC04-15460/OGLETHORPE//NC05-21937	BL
NC8932-16	1	S15	North Carolina	NC04-15460/OGLETHORPE//NC05-21937	BL
NC9305-7	2	S14_G15	North Carolina	BESS/NC-YADKIN	BL
NC9341-10	1	S14	North Carolina	NC-CAPEFEAR/BESS	BL
NC9485-14	1	S14	North Carolina	VA05W-510/BESS	BL
PIONEER26R41	1	G16	Pioneer	No Record	CC
SAVOY	1	S16	Georgia	931233-28-2-2/USG3592	CC
SCAR99050B1	1	G09	South Carolina	AR679-9-1-2/NC97BGTD8	BL
SCAR99080E1	1	G09	South Carolina	AR835-21-1-2/COKER9663	BL
SCAR99103N1	1	G09	South Carolina	AR839-27-1-3/WGRC34	BL
SCAR99143A1	1	G09	South Carolina	COKER9704/NC97BGTD8	BL
SCAR99175B1	1	G09	South Carolina	FL8868/COKER9663	BL
SCAR99180A1	1	G09	South Carolina	FL8868/FFR522W	BL
SCLA01111C-J7	2	G10-11	South Carolina	PIONEER26R61/FUTAI8944	BL
SCLA1030J1	1	G12	South Carolina	LA94162D157-1/AGS2000	BL
SCLA1067A1	1	G12	South Carolina	KS94U275/AGS2000	BL
SCLA1084A1	1	G12	South Carolina	NC98-24710/AGS2000	BL
SCLA1084B1	1	G12	South Carolina	NC98-24710/AGS2000	BL
SCLA1084C1	1	G12	South Carolina	NC98-24710/AGS2000	BL
SCLA1084K1	1	G12	South Carolina	NC98-24710/AGS2000	BL
SCLA1102D1	1	G12	South Carolina	PIONEER26R61/LA841	BL
SCLA1102G1	1	G12	South Carolina	PIONEER26R61/LA841	BL
SCLA1102G3	1	G12	South Carolina	PIONEER26R61/LA841	BL
SCLA1102H1	1	G12	South Carolina	PIONEER26R61/LA841	BL

APPENDIX A cont.

Variety	No. Year	Nursery	Breeding Program	Pedigree	Type
SCLA1110A1	1	G11	South Carolina	PIONEER26R61/NC98-24710	BL
SCLA1110B1	1	G11	South Carolina	PIONEER26R61/NC98-24710	BL
SCLA1110P1	1	G12	South Carolina	PIONEER26R61/NC98-24710	BL
SCLA1110R1	1	G12	South Carolina	PIONEER26R61/NC98-24710	BL
SCLA99049D-E1-J1	2	G10-11	South Carolina	FL931339AS/LA87167D8-10-2//NC96BGTA5	BL
SCTX98-17A1	1	G09	South Carolina	MOREY/VA94-54-479	BL
SCTX98-20-J10	1	G10	South Carolina	LA87107-08-10-2/P88288C1-6-1-2	BL
SCTX98-27A1	1	G09	South Carolina	VA94-54-479/A93-7162	BL
SCTX98-27B1	1	G09	South Carolina	VA94-54-479/A93-7162	BL
SCTX98-27C1	1	G09	South Carolina	VA94-54-479/A93-7162	BL
SCTX98-27-J1	1	G10	South Carolina	VA94-54-479/A93-7162	BL
SCTX98-27-J7	1	G10	South Carolina	VA94-54-479/A93-7162	BL
SCTX98-56G1	1	G09	South Carolina	VA94-54-479/TX97-89	BL
SCTX98-5B1	1	G09	South Carolina	VA94-52-25/LA87107-08-10-2	BL
SCW010025D1	1	G10	South Carolina	SC957755/SC967725//PI531193	BL
SCW010025G1	1	G10	South Carolina	SC957755/SC967725//PI531193	BL
SCW010025G2	1	G10	South Carolina	SC957755/SC967725//PI531193	BL
SCW010025H1	1	G10	South Carolina	SC957755/SC967725//PI531193	BL
SCW010025K1	1	G10	South Carolina	SC957755/SC967725//PI531193	BL
SCW010025L1	1	G10	South Carolina	SC957755/SC967725//PI531193	BL
SCW010025T1	1	G10	South Carolina	SC957755/SC967725//PI531193	BL
SCW990002K1	1	G08	South Carolina		BL
SCW990002V1	1	G08	South Carolina		BL
SCW990002W1	1	G08	South Carolina		BL
SCW990013D1	1	G08	South Carolina		BL
SCW990013H1	1	G08	South Carolina		BL
SCW990013J1	1	G08	South Carolina		BL
SCW990013K1	1	G08	South Carolina		BL
SCW990013N1	1	G08	South Carolina		BL
SCW990013V1	1	G08	South Carolina		BL
SCW990022A1	1	G08	South Carolina		BL
SCW990022B1	1	G08	South Carolina		BL
SCW990022C1	1	G08	South Carolina		BL
SHIRLEY	7	G11 to 15_S14-15	Virginia	VA94-52-25/COKER9835//VA96-54-234	CC
SS8641	12	G08 to16_S14 to 16	Georgia	GA881130/GA881582//GA881582	CC
TX12D4603	1	S14	Texas	COKER9553/LA841	BL
TX12D4625	2	S14_G15	Texas	SS8641/CK9553	BL
TX12D4700	1	S14	Texas	JAMESTOWN/LA97113UC-124	BL
TX12D4733	1	S14	Texas	DKGR9108/LA841	BL

APPENDIX A cont.

Variety	No. Year	Nursery	Breeding Program	Pedigree	Type
TX12D4741	1	S14	Texas	DK9577/GA96229-3E39	BL
TX12D4768	2	S14_G15	Texas	GA951216-2E26/AGS2060	BL
TX12D4788	1	S14	Texas	GA951216-2E26/PIONEER26R61	BL
TX12D4791	2	S14_G15	Texas	SS8641/PIONEER26R61	BL
TX12D4845	1	S14	Texas	LA482/LA98094BUB-58-5	BL
TX12D4858	1	S14	Texas	LA97113UC-124-3/GA951216-2E26	BL
TX12D4860	1	S14	Texas	LA97113UC-124-3/GA951216-2E26	BL
TX12D4896	1	S14	Texas	PIONEER26R61/SS8641	BL
TX12D4898	2	S14_G15	Texas	PIONEER26R61/SS8641	BL
TX12D4927	1	S14	Texas	VA02W-713/AGS2010	BL
TX12D4930	1	S14	Texas	VA02W-713/AGS2010	BL
TX13D5026	1	S15	Texas	Sturdy2K/CRAWFORD	BL
TX13D5129	1	S15	Texas	CUTTER/C9553	BL
TX13D5137	2	S15_G16	Texas	AGS2060/SS8641	BL
TX13D5157	1	S15	Texas	GR9108/SS8641//CK9700	BL
TX13D5161	2	S15_G16	Texas	CK9700/LA95135	BL
TX13D5169	2	S15_G16	Texas	SS8641/APCB02-8486	BL
TX13D5191	1	S15	Texas	LA01110D-88/AGS2060	BL
TX13D5193	1	S15	Texas	LA06012,F1/CK9700	BL
TX13D5217	2	S15_G16	Texas	LA98205D-14-2-3/LA482//AGS2060	BL
TX13D5234	2	S15_G16	Texas	LA98205D-17-2-4/LA01110D-88	BL
TX13D5237	2	S15_G16	Texas	VA02W-713/GA951079-2E31//LA99005UC-31-3	BL
TX13D5245	1	S15	Texas	NC03-5921/COKER9553	BL
TX13D5252	1	S15	Texas	NC03-5921/LA99042E-117	BL
TX13D5259	1	S15	Texas	NC03-5921/LA99042E-117	BL
TX13D5261	1	S15	Texas	PIONEER26R61/LA99005UC-31-3	BL
TX14D8130	1	S16	Texas	JAMESTOWN/LA482//GA991227-6A33	BL
TX14D8142	1	S16	Texas	P961341A3-2-2/LA98205D-17-2-4//AGS2020	BL
TX14D8160	1	S16	Texas	LA01110D-111/LA841	BL
TX14D8237	1	S16	Texas	CK9700/LA95135//LA98094BUB-58-5	BL
TX14D8282	1	S16	Texas	SS8641/NC03-5921//LA98094BUB-58-5	BL
TX14D8283	1	S16	Texas	SS8641/NC03-5921//TX98-71D1	BL
TX14D8306	1	S16	Texas	LA99042E-117/LA98113D-41-1-C//ARGE97-1060-5-5/LA98113D-41-1-C	BL
TX14D8331	1	S16	Texas	JAMESTOWN/LA95135//GA9811622-1-40-1	BL
TX14D8337	1	S16	Texas	JAMESTOWN/LA95135//GA991227-6A33	BL
TX14D8343	1	S16	Texas	VA02W-713/LA95135//LA841	BL
TX14D8409	1	S16	Texas	PIONEER26R22/AGS2060//LA841	BL
TX14D8440	1	S16	Texas	VA05W-500/LA841//LA95135	BL
TX14D8444	1	S16	Texas	VA05W-500/LA98205D-17-2-4//LA95135	BL
TX14D8445	1	S16	Texas	VA05W-500/LA98205D-17-2-4//LA95135	BL
TX14D8488	1	S16	Texas	NC04-27617/LA01110D-88//PIONEER26R61	BL
TXGA051407-2-15-6-EL61	1	G15	Texas	GA98401-1B-22-3-3/97531-1	BL
TXGA06343-17-3-5-EL2	1	G15	Texas	011638-G1-G1/981592-8-8-1//AGS2020	BL
USG3120	4	G14-15_S14-15	Georgia	GA901146/GA9006//AGS2000	CC

APPENDIX A cont.

Variety	No. Year	Nursery	Breeding Program	Pedigree	Type
USG3209	1	G08	Georgia	SALUDA/4/MASSEY*2/3/MASSEY*3/BALKAN//SALUDA	CC
USG3555	9	G09 to 15_S14-15	Georgia	VA94-52-60/PIONEER2643//USG3209	CC
VA03W-509	1	G08	Virginia	96-52-69/96-52-82//96-52-68/PIONEER2643	BL
VA04W-306	1	G09	Virginia	FFR518W/SS520	BL
VA05W-139	1	G08	Virginia	PIONEER26R24/MCCORMICK	BL
VA05W-22	1	G08	Virginia	SISSON/PIONEER2552//96W-391	BL
VA05W-376	1	G08	Virginia	SISSON/TRIBUTE	BL
VA06W-112	1	G08	Virginia	VA97W-533/RC-STRATEGY	BL
VA06W-146	1	G10	Virginia	RENWOOD3260/LA87167D8-10-2	BL
VA06W-194	2	G08-09	Virginia	RC-STRATEGY/VA98W-179	BL
VA06W-215	1	G08	Virginia	MCCORMICK/GA881130	BL
VA06W-237	1	G08	Virginia	COKER9025/VA98W-430	BL
VA06W-256	2	G08-09	Virginia	MCCORMICK/VAN99W-97	BL
VA06W-392	1	G08	Virginia	VA96W-49/AGS2000//RC-STRATEGY	BL
VA06W-412	1	G09	Virginia	TRIBUTE/AGS2000//VAN99W-20	BL
VA06W-423	1	G09	Virginia	PATTON/AGS2000//VA99W-175	BL
VA06W-6	2	G08-09	Virginia	TRIBUTE/AGS2000	BL
VA06W-93	1	G08	Virginia	SS550/RC-STRATEGY	BL
VA07MAS14-9260-8-2-2	1	G15	Virginia	NC03-11458/IL99-15867//VA05W-436	BL
VA07MAS3-7304-3-1-2-3	1	G16	Virginia	SHIRLEY/AGS2060//SS8404	BL
VA07MAS3-7304-3-2-4-3	1	G16	Virginia	SHIRLEY/AGS2060//SS8404	BL
VA07MAS4-7417-1-3-3	1	G15	Virginia	GA951231-4E25/SS8404//SHIRLEY	BL
VA07W-138	1	G09	Virginia	PIONEER26R24/MCCORMICK	BL
VA07W-214	1	G09	Virginia	VAN99W-20/VA99W-187	BL
VA07W-27	1	G09	Virginia	TRIBUTE/AGS2000	BL
VA07W-347	1	G09	Virginia	VA98W-430/PIONEERXW681//AGS2000	BL
VA07W-415	3	G09-10-12	Virginia	VA98W-895/GA881130//VA98W-627RS	BL
VA07W-83	1	G09	Virginia	TRIBUTE/LA9070G45-3-3-1	BL
VA08MAS1-188-6-4	2	G15-16	Virginia	VA05W-640/VA05W-693//SHIRLEY	BL
VA08W-165	1	G10	Virginia	VA00W-366/TRIBUTE	BL
VA08W-176	1	G10	Virginia	KY96C-0079-5/MCCORMICK	BL
VA08W-193	1	G10	Virginia	AGS2000/USG3706//DOMINION	BL
VA08W-196	1	G10	Virginia	AGS2000/USG3706//DOMINION	BL
VA08W-223	1	G10	Virginia	GF90524E1/USG3706//SS520	BL
VA08W-232	1	G10	Virginia	GF921221E16/MCCORMICK//VA99W-200	BL
VA08W-286	1	G10	Virginia	PIONEER25R38/TRIBUTE	BL
VA08W-294	1	G10	Virginia	SS520/VA99W-188//TRIBUTE	BL
VA08W-295	1	G10	Virginia	SS520/VA99W-188//TRIBUTE	BL
VA08W-613	1	G12	Virginia	FREEDOM/NC-NEUSE//VA98W-688	BL
VA08W-630	1	G11	Virginia	OH552/SS550//RC-STRATEGY	BL
VA08W-632	1	G11	Virginia	OH552/SS550//RC-STRATEGY	BL
VA08W-92	1	G10	Virginia	RENWOOD3260/LA9070G45-3-3-1//TRIBUTE	BL
VA09MAS1-12-8-4	1	G16	Virginia	GA991371-6E13/USG3555//OAKES	BL
VA09MAS6-122-7-1	1	G16	Virginia	SHIRLEY/GA991371-6E13//SS5205	BL

APPENDIX A cont.

Variety	No. Year	Nursery	Breeding Program	Pedigree	Type
VA09MAS7-61-2-1	1	G16	Virginia	VA06W-256/SS8641//12V51	BL
VA09W-110	2	G11-12	Virginia	USG3592/VA01W-303	BL
VA09W-112	2	G11-12	Virginia	USG3592/VA01W-303	BL
VA09W-114	1	G12	Virginia	USG3592/VA01W-303	BL
VA09W-45	1	G11	Virginia	GF921221E16/MCCORMICK//VA99W-200	BL
VA09W-46	2	G11-12	Virginia	GF921221E16/MCCORMICK//VA99W-200	BL
VA09W-52	2	G11-12	Virginia	GF921221E16/MCCORMICK//VA99W-200	BL
VA09W-656	1	G11	Virginia	NC-NEUSE/VA99W-200//MCCORMICK	BL
VA09W-657	1	G11	Virginia	NC-NEUSE/VA99W-200//MCCORMICK	BL
VA09W-67	1	G11	Virginia	VA99W-188/TRIBUTE	BL
VA09W-69	1	G12	Virginia	SS520/VA99W-188//TRIBUTE	BL
VA09W-73	1	G11	Virginia	SS520/VA99W-188//TRIBUTE	BL
VA09W-75	1	G11	Virginia	SS520/VA99W-188//TRIBUTE	BL
VA10W-112	1	G13	Virginia	KY97C-0540-04/GF951079-2E31	BL
VA10W-118	1	G13	Virginia	KY97C-0540-04/GF951079-2E31	BL
VA10W-123	1	G12	Virginia	PIONEER25R47/GF951079-2E31	BL
VA10W-125	1	G12	Virginia	PIONEER25R47/JAMESTOWN	BL
VA10W-140	1	G12	Virginia	VA01W-210/SS520//TRIBUTE	BL
VA10W-663	1	G12	Virginia	P97397B1-4-5/MCCORMICK//COKER9511	BL
VA10W-96	1	G13	Virginia	FG95195/JAMESTOWN	BL
VA11MAS-7520-2-3-255	1	G14	Virginia	OGLETHORPE/SS8404//SHIRLEY	BL
VA11W-106	1	G13	Virginia	PIONEER26R47/JAMESTOWN	BL
VA11W-108	1	G13	Virginia	PIONEER26R47/JAMESTOWN	BL
VA11W-111	1	G14	Virginia	PIONEER25R47/JAMESTOWN	BL
VA11W-165	1	G13	Virginia	VA03W-192/VA03W-436	BL
VA11W-195	1	G13	Virginia	MPV57/M99-3098//VA03W-434	BL
VA11W-196	1	G13	Virginia	MPV57/M99-3098//VA03W-434	BL
VA11W-230	1	G13	Virginia	SS520/GF951208-2E35//JAMESTOWN	BL
VA11W-278	1	G13	Virginia	NC00-15389/GF951079-2E31//USG3555	BL
VA11W-279	1	G14	Virginia	NC00-15389/GF951079-2E31//USG3555	BL
VA11W-301	1	G13	Virginia	PIONEER25R47/NC00-15389//JAMESTOWN	BL
VA11W-31	1	G13	Virginia	FG95195/JAMESTOWN	BL
VA11W-313	1	G14	Virginia	PIONEER25R47/GF951079-2E31//USG3555	BL
VA11W-95	1	G14	Virginia	PIONEER25R47/JAMESTOWN	BL
VA12W-102	1	G14	Virginia	VA03W-436/IL99-15867	BL
VA12W-150	1	G14	Virginia	IL99-15867/JAMESTOWN	BL
VA12W-22	1	G15	Virginia	KY93C-1238-17-1/VA03W-436	BL
VA12W-26	1	G14	Virginia	MPV57/M99-3098//VA03W-434	BL
VA12W-31	1	G15	Virginia	MPV57/M99-3098//VA03W-434	BL
VA12W-54	1	G14	Virginia	NC00-15389/GF951079-2E31//USG3555	BL
VA12W-72	1	G14	Virginia	PIONEER25R47/GF951079-2E31//USG3555	BL
VA12W-97	1	G15	Virginia	MERL/AGS2026	BL
VA13W-124	1	G15	Virginia	12V51/AGS2026	BL
VA13W-177	1	G15	Virginia	SHIRLEY/BRANSON//JAMESTOWN	BL

APPENDIX A cont.

Variety	No. Year	Nursery	Breeding Program	Pedigree	Type
VA13W-38	1	G15	Virginia	IL99-15867/JAMESTOWN	BL
VA13W-56	1	G15	Virginia	USG3555/SHIRLEY//JAMESTOWN	BL
VA14FHB-28	1	G16	Virginia	RecurrentSelectionPop2	BL
VA14W-29	1	G16	Virginia	VA05W-139/SS5205	BL

APPENDIX B. Information about the genotypes evaluated in the diversity panel (population A). We include information about the name of each genotype (Variety), Year of release for cultivars and first year of evaluation for breeding lines (Year), breeding program where the germplasm was developed (Breeding program), and status of the germplasm in the breeding pipeline as cultivar or breeding line (Type).

Variety	Year	Breeding Program	Type
001169-7E15	2000	GA AES	Breeding Line
01063-1-3-6-2-G2	2001	GA AES	Breeding Line
011124-1-42-13	2001	GA AES	Breeding Line
011388-8-4-5	2001	GA AES	Breeding Line
031086-44-4-2	2003	GA AES	Breeding Line
051336-B-B-1	2005	GA AES	Breeding Line
071628-G3-G1-G4-G1	2007	GA AES	Breeding Line
071694-G5-G5-G1PUB	2007	GA AES	Breeding Line
081515-G1-G2	2008	GA AES	Breeding Line
09283-G1-G1	2009	GA AES	Breeding Line
222-22-5		Cornell	Breeding Line
991227-6A33	1999	GA AES	Breeding Line
ADENA	1985	Ohio State	Cultivar
AGS_2000	2002	GA AES	Cultivar
AGS_2010	2006	GA AES	Cultivar
AGS_2020	2008	GA AES	Cultivar
AGS_2026	2007	GA AES	Cultivar
AGS_2031	2006	GA AES	Cultivar
AGS_2035	2008	GA AES	Cultivar
AGS_2060	2011	LSU	Cultivar
AGS_CL7	2007	GA AES	Cultivar
AMERICAN-BANNER	1894	Michigan State	Cultivar
ARGEE	1977	Univ. Wisconsin	Cultivar
ARS05-0074	2005	NC ARS	Breeding Line
ARS05-0241	2005	NC ARS	Breeding Line
ARS05-0401	2005	NC ARS	Breeding Line

APPENDIX B cont.

Variety	Year	Breeding Program	Type
ARS07-0203	2007	NC ARS	Breeding Line
ARS07-0404	2007	NC ARS	Breeding Line
ARS07-0558	2007	NC ARS	Breeding Line
ARS07-0815	2007	NC ARS	Breeding Line
ARS07-0912	2007	NC ARS	Breeding Line
ARS07-1208	2007	NC ARS	Breeding Line
ARS07-1243	2007	NC ARS	Breeding Line
ARS08-0111	2008	NC ARS	Breeding Line
ARS09-776	2009	NC ARS	Breeding Line
ARTHUR	1968	Purdue	Cultivar
BALDROCK	1932	MichiganState	Cultivar
BALDWIN	2009	GA AES	Cultivar
BECKER	1985	Ohio State	Cultivar
BENHUR	1966	Purdue	Cultivar
BLUEBOY	1967	NCSU	Cultivar
BOONE	1993	Purdue	Cultivar
BRANSON	2005	AgriPro	Cultivar
CALDWELL	1982	Purdue	Cultivar
CALEDONIA	2004	Cornell	Cultivar
CARDINAL	1986	Ohio State	Cultivar
CHANCELLOR	1958	GA AES	Cultivar
CHARMANY	1984	WI AES	Cultivar
CHESAPEAKE	2007	MD AES	Cultivar
CLARK	1987	Purdue	Cultivar
CLEMSON-201	1994	SC AES	Cultivar
COKER_65-20	1967	Coker Pedigree Seed	Cultivar
COKER_68-15	1971	Coker Pedigree Seed	Cultivar
COKER_747	1975	Coker Pedigree Seed	Cultivar
COKER_762	1980	Coker Pedigree Seed	Cultivar
COKER_797	1980	Coker Pedigree Seed	Cultivar
COKER_9134_SYN	1992	Coker Pedigree Seed	Cultivar
COKER_9152	2002	Coker Pedigree Seed	Cultivar
COKER_916	1980	Coker Pedigree Seed	Cultivar
COKER_9375	2004	Agripro/Syngenta	Cultivar
COKER_9553	2006	Agripro/Syngenta	Cultivar

APPENDIX B cont.

Variety	Year	Breeding Program	Type
COKER_9663	1997	Northrup King	Cultivar
COKER_9766	1987	Coker Pedigree Seed	Cultivar
COKER_9803	1990	Northrup King	Cultivar
COKER_9835	1990	Northrup King	Cultivar
DELTA_KING_GR9108	2004	AR AES	Cultivar
DOMINION	2006	VA Tech	Cultivar
DOUBLECROP	1975	AR AES	Cultivar
ELKHART	1996	AgriPro	Cultivar
ERNIE	1994	MO AES	Cultivar
EXCEL	1990	Ohio State	Cultivar
FAIRFIELD	1942	Purdue	Cultivar
FFR_555W	1991	VA Tech	Cultivar
FG95195	1995	GA AES	Breeding Line
FL_302	1984	FloridaState	Cultivar
FLINT	1919	GA AES	Cultivar
FOSTER	1996	AgriPro	Cultivar
FRANKENMUTH	1979	MichiganState	Cultivar
FREEDOM	1991	Ohio State	Cultivar
FULCASTER	1886	Schindel	Cultivar
GA_1123	1961	GA AES	Cultivar
GA00067-8E35	2000	GA AES	Breeding Line
GA001138-8E36	2000	GA AES	Breeding Line
GA001142-9E23	2000	GA AES	Breeding Line
GA001170-7E26	2000	GA AES	Breeding Line
GA011493-8E18	2001	GA AES	Breeding Line
GA021245-9E16	2002	GA AES	Breeding Line
GA021338-9E15	2002	GA AES	Breeding Line
GA031238-7E34	2003	GA AES	Breeding Line
GLACIER	1991	WI AES	Cultivar
GLORY	1992	Ohio State	Cultivar
GOENS	1808	J.Dent, Ohio	Cultivar
GORE	1990	GA AES	Cultivar
GR_860	1986	Ohio State	Cultivar
GRANDPRIZE	1900	Jones	Cultivar
HART	1976	Penn State	Cultivar

APPENDIX B cont.

Variety	Year	Breeding Program	Type
HARUS	1985	Ag Canada	Cultivar
HILLSDALE	1983	MichiganState	Cultivar
HOLLEY	1970	GA AES	Cultivar
HOPEWELL	1995	Ohio State	Cultivar
HOWELL	1988	Univ. Illinois	Cultivar
HUNTER	1982	Coker Pedigree Seed	Cultivar
IL00-8061	2006	Univ. Illinois	Breeding Line
IL00-8530	2008	Univ. Illinois	Breeding Line
IL00-8633	2006	Univ. Illinois	Breeding Line
IL00-8641	2000	Univ. Illinois	Breeding Line
IL02-18228	2010	Univ. Illinois	Breeding Line
IL05-4236	2005	Univ. Illinois	Breeding Line
IL06-13721	2006	Univ. Illinois	Breeding Line
IL06-23571	2006	Univ. Illinois	Breeding Line
IL08-24578	2008	Univ. Illinois	Breeding Line
IL96-6472	1996	Univ. Illinois	Breeding Line
IL97-1828	1997	Univ. Illinois	Breeding Line
ILLINI_CHIEF	1915	Gillham-IL	Cultivar
INW0304	2003	Purdue	Cultivar
INW0411	2005	Purdue	Cultivar
INW0412	2004	Purdue	Cultivar
INW0731	2007	Purdue	Cultivar
JACKSON	1993	VA Tech	Cultivar
JAMESTOWN	2007	VA Tech	Cultivar
JAYPEE	1995	AR AES	Cultivar
KASKASKIA	1998	Univ. Illinois	Cultivar
KENOSHA	1968	WI AES	Cultivar
KEY	1976	Purdue	Cultivar
KNOX-62	1962	Purdue	Cultivar
KRISTY	2001	Univ. Kentucky	Cultivar
KY02C-1043-04	2002	Univ. Kentucky	Breeding Line
KY02C-1058-03	2002	Univ. Kentucky	Breeding Line
KY02C-1076-07	2002	Univ. Kentucky	Breeding Line
KY02C-1121-11	2002	Univ. Kentucky	Breeding Line
KY02C-2215-02	2002	Univ. Kentucky	Breeding Line

APPENDIX B cont.

Variety	Year	Breeding Program	Type
KY03C-1002-02	2003	Univ. Kentucky	Breeding Line
KY03C-1237-39	2003	Univ. Kentucky	Breeding Line
KY93C-1238-17-1	1993	Univ. Kentucky	Breeding Line
LA01069D-23-4-4	2001	LSU	Breeding Line
LA0110D-150	2001	LSU	Breeding Line
LA01139D-56-1	2001	LSU	Breeding Line
LA01164D-94-2-B	2001	LSU	Breeding Line
LA02015E201	2002	LSU	Breeding Line
LA02015E42	2002	LSU	Breeding Line
LA02015E58	2002	LSU	Breeding Line
LA02024E12	2002	LSU	Breeding Line
LA02024E7	2002	LSU	Breeding Line
LA03012E-27	2003	LSU	Breeding Line
LA03118E117	2003	LSU	Breeding Line
LA03136E71	2003	LSU	Breeding Line
LA03148E12	2003	LSU	Breeding Line
LA03155D-P13	2003	LSU	Breeding Line
LA03161D-P1	2003	LSU	Breeding Line
LA03217D-P2	2003	LSU	Breeding Line
LA03217E2	2003	LSU	Breeding Line
LA04013D-142	2004	LSU	Breeding Line
LA04041D-10	2004	LSU	Breeding Line
LA821	2010	LSU	Cultivar
LA841	2002	LSU	Cultivar
LA95135	1995	LSU	Breeding Line
LA97113UC-124	1997	LSU	Breeding Line
LOGAN	1968	Ohio State	Cultivar
MADISON	1990	VA Tech	Cultivar
MAGNOLIA	2007	Agripro/Coker	Cultivar
MALLARD	1991	Agripro	Cultivar
MASSEY	1981	VA Tech	Cultivar
MCCORMICK	2002	VA Tech	Cultivar
MCNAIR_1813	1975	McNair Seed Company	Cultivar
MCNAIR-701	1973	McNair Seed Company	Cultivar
MD00W16-07-3	2000	MD AES	Breeding Line

APPENDIX B cont.

Variety	Year	Breeding Program	Type
MD01W28-08-11	2001	MD AES	Breeding Line
MD99W64-05-11	1999	MD AES	Breeding Line
MEDITERRANEAN	1837	Gordon	Cultivar
MERL	2009	VA Tech	Cultivar
MILTON	2009	MO AES	Cultivar
MO_011126	2001	MO AES	Breeding Line
MO_050921	2005	MO AES	Breeding Line
MO_080104	2008	MO AES	Breeding Line
MO_081652	2008	MO AES	Breeding Line
MO_980525	1998	MO AES	Breeding Line
MO_980829	1998	MO AES	Breeding Line
MONON	1949	Purdue	Cultivar
MPV_57	2005	VA Tech	Cultivar
NC-CAPE FEAR	2010	NCSU	Cultivar
NC-NEUSE	2003	NCSU	Cultivar
NC-YADKIN	2011	NCSU	Cultivar
NC06-19896	2006	NCSU	Breeding Line
NC06-20401	2006	NCSU	Breeding Line
NC06BGTAG12	2006	NCSU	Breeding Line
NC07-20850	2007	NCSU	Breeding Line
NC07-22432	2007	NCSU	Breeding Line
NC07-23880	2007	NCSU	Breeding Line
NC07-24445	2007	NCSU	Breeding Line
NC07-25169	2007	NCSU	Breeding Line
NC08-21273	2008	NCSU	Breeding Line
NC08-23089	2008	NCSU	Breeding Line
NC08-23090	2008	NCSU	Breeding Line
NC08-23323	2008	NCSU	Breeding Line
NC08-23324	2008	NCSU	Breeding Line
NC08-23383	2008	NCSU	Breeding Line
NC08-23925	2008	NCSU	Breeding Line
NC09BGTS16	2009	NCSU	Breeding Line
NC09BGTUM15	2009	NCSU	Breeding Line
NC96BGTA4	1996	NCSU	Breeding Line
NC96BGTA5	1996	NCSU	Breeding Line

APPENDIX B cont.

Variety	Year	Breeding Program	Type
NC96BGTA6	1996	NCSU	Breeding Line
NC96BGTD1	1996	NCSU	Breeding Line
NC96BGTD2	1996	NCSU	Breeding Line
NC96BGTD3	1996	NCSU	Breeding Line
NC97BGTAB10	1997	NCSU	Breeding Line
NC97BGTAB9	1997	NCSU	Breeding Line
NC97BGTD7	1997	NCSU	Breeding Line
NC97BGTD8	1997	NCSU	Breeding Line
NC99BGTAG11	1999	NCSU	Breeding Line
OAKES	2006	Syngenta	Cultivar
OASIS	1973	Purdue	Cultivar
OGLETHORPE	2007	GA AES	Cultivar
P03528A1-10	2003	Purdue	Breeding Line
P0570A1-2	2005	Purdue	Breeding Line
P07290A1-12	2007	Purdue	Breeding Line
P99840C4-8	1999	Purdue	Breeding Line
PANOLA	2005	Syngenta	Cultivar
PAT	2001	AR AES	Cultivar
PATTERSON	1995	Purdue	Cultivar
PATTON	1999	AgriPro	Cultivar
PEMBROKE	2008	Univ. Kentucky	Cultivar
PIONEER_2548	1988	Pioneer	Cultivar
PIONEER_2555	1986	Pioneer	Cultivar
PIONEER_2568	1995	Pioneer	Cultivar
PIONEER_2580	1993	Pioneer	Cultivar
PIONEER_25W60	1998	Pioneer	Cultivar
PIONEER_2643	1994	Pioneer	Cultivar
PIONEER_2684	1993	Pioneer	Cultivar
PIONEER_26R15	2003	Pioneer	Cultivar
PIONEER_26R24	1999	Pioneer	Cultivar
PIONEER_26R31	2004	Pioneer	Cultivar
PIONEER_26R46	1998	Pioneer	Cultivar
PIONEER_26R61	1998	Pioneer	Cultivar
POTOMAC	1975	VA Tech	Cultivar
RED-MAY	1845		Cultivar

APPENDIX B cont.

Variety	Year	Breeding Program	Type
REDCOAT	1960	Purdue	Cultivar
ROANE	1999	VA Tech	Cultivar
ROY	1979	NCSU	Cultivar
ROYAL	1947	Univ. Illinois	Cultivar
RUDY	1871	Rudy	Cultivar
SABBE	2000	AR AES	Cultivar
SALUDA	1983	VA Tech	Cultivar
SCOTTY	1982	Univ. Illinois	Cultivar
SENECA	1950	Ohio State	Cultivar
SEVERN	1981	MD AES	Cultivar
SHIRLEY	2008	VA Tech	Cultivar
SISSON	2000	VA Tech	Cultivar
SS_520	2001	Southern States	Cultivar
SS_5205	2008	Southern States	Cultivar
SS8641	2006	GA AES	Cultivar
TECUMSEH	1973	MichiganState	Cultivar
TITAN	1978	Ohio State	Cultivar
TRIBUTE	2002	VA Tech	Cultivar
TYLER	1980	VA Tech	Cultivar
USG_3120	2009	Unisouth Genetics	Cultivar
USG_3209	2003	Unisouth Genetics	Cultivar
USG_3295	2006	Unisouth Genetics	Cultivar
USG_3555	2007	Unisouth Genetics	Cultivar
USG_3592	2003	Unisouth Genetics	Cultivar
VA-259		VA Tech	Breeding Line
VA-90	1990	VA Tech	Breeding Line
VA-96W-247	1996	VA Tech	Breeding Line
VA00W-38	2000	VA Tech	Breeding Line
VA01W-21	2001	VA Tech	Breeding Line
VA01W713	2001	VA Tech	Breeding Line
VA03W-211	2003	VA Tech	Breeding Line
VA03W-235	2003	VA Tech	Breeding Line
VA05W-139	2011	VA Tech	Breeding Line
VA05W-151	2005	VA Tech	Breeding Line
VIGO	1946	Purdue	Cultivar

APPENDIX B cont.

Variety	Year	Breeding Program	Type
WAKEFIELD	1990	VA Tech	Cultivar
WAKELAND	1959	NCSU	Cultivar
WHEELER	1980	VA Tech	Cultivar