GENOMICS-ASSISTED APPROACHES TO IMPROVE GRAIN YIELD AND END-USE QUALITY IN HARD WINTER WHEAT (*Triticum Aestivum L.*)

BY

HARSIMARDEEP SINGH GILL

A dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Major in Plant Science

South Dakota State University

2023

DISSERTATION ACCEPTANCE PAGE
Harsimardeep Singh Gill

This dissertation is approved as a creditable and independent investigation by a candidate for the Doctor of Philosophy degree and is acceptable for meeting the dissertation requirements for this degree. Acceptance of this does not imply that the conclusions reached by the candidate are necessarily the conclusions of the major department.

Sunish Sehgal
Advisor Date

Dr. David Wright
Department Head Date

Nicole Lounsbery, PhD
Director, Graduate School Date

Dedicated to my mother and father for their love, support, and sacrifices.

ACKNOWLEDGEMENT

I would first like to express my appreciation to my advisor, Dr. Sunish K Sehgal, for his love, support, motivation, and excellent guidance for my doctoral research. Without him, this work would not have been possible. I would also like to thank my committee members, Dr. Brent Turnipseed, Dr. Gazala Ameen, and Dr. Prafulla Salunke for their useful suggestions and for being a part of my advisory committee. I am also thankful to Dr. Shaukat Ali for his help and support throughout my Ph.D. journey.

I would like to say thanks to the current and past members of the winter wheat breeding group at SDSU including Navreet Brar, Jagdeep Sidhu, Jyotirmoy Halder, Rami Altameemi, Jinfeng Zhang, and Cody Hall for all their unconditional support. I am also thankful to all my friends for being there in good and bad times. I would like to thank Bhupinder, Armaan, Jashan, Navdeep, Teerath, Jasdeep, Navreet, and all my friends for being part of this amazing journey and for countless memories.

I would like to thank South Dakota State University for providing the study site for my research. My sincere thanks also go to the faculty and staff of the Department of Agronomy, Horticulture and Plant Science at South Dakota State University for their support and for sharing their knowledge. I am very grateful to all the funding agencies; USDA-NIFA (Wheat-CAP 2017-67007-25939) and South Dakota Wheat Commission for supporting my research.

CONTENTS

LIST OF FIGURES

Chapter 3

Chapter 4

Chapter 5

LIST OF TABLES

Chapter 3

Chapter 4

Chapter 5

ABSTRACT

GENOMICS-ASSISTED APPROACHES TO IMPROVE GRAIN YIELD AND END-USE QUALITY IN HARD WINTER WHEAT (*Triticum Aestivum L.*)

HARSIMARDEEP SINGH GILL

2023

Global wheat production needs to be increased by 60% to meet the future demand of feeding nine billion people by 2050. Simultaneously, it is important to improve the end-use quality to meet the requirements of producers, grain markets, processors, and consumers. Thus, the development of more productive wheat varieties with better end-use quality remains the primary focus for all wheat breeding programs. However, direct phenotypic selection for improving grain yield and end-use quality is difficult as it is highly influenced by environmental factors. This dissertation focuses on harnessing advancements in genomics applications, including genome-wide association studies (GWAS), for the genetic characterization of yield component traits and utilizing it in marker-assisted selection for grain yield. Further, we investigated the efficacy of genomic selection GS and assessed the performance of various statistical models in predicting agronomic and end-use quality traits in the South Dakota hard winter wheat (HWW) breeding program.

In the first study, GWAS was used to identify genetic determinants for yield-component traits in HWW, which exhibits higher heritability compared to grain yield per se. We assembled a population of breeding lines and well-adapted cultivars, genotyped using genotyping-by-sequencing (GBS), and evaluated over four environments for phenotypic analysis of spike and kernel traits. GWAS using 8,030 single nucleotide polymorphisms (SNPs) identified 17 significant and multi-environment marker-trait

associations (MTAs) for various traits, representing 12 putative quantitative trait loci (QTLs), with five QTLs affecting multiple traits. Further, a highly significant QTL was detected on chromosome 7AS that has not been previously associated with the number of spikelets/spike and putative candidate genes were identified in this region. The allelic frequencies of important QTLs were deduced in a larger set of 1,124 accessions which revealed the importance of identified MTAs in the U.S. HWW breeding programs.

In the second strategy, we studied to evaluate the potential of genomic selection in predicting complex traits at earlier stages of the breeding program. Here, we used multi-trait genomic prediction (GP) models to predict multiple agronomic traits using 314 advanced and elite breeding lines of HWW evaluated at ten site-year environments. Extensive data from multi-environment trials was used to cross-validate the multivariate machine learning (ML) models that integrate the analysis of multiple traits and/or include GxE interaction. The multivariate ML models performed better for all traits, with average improvement over the ST-CV1 reaching up to 19%, 71%, 17%, 48%, and 51% for grain yield, grain protein content, test weight, plant height, and days to heading, respectively. Next, we evaluated the efficacy of multivariate GP using a set of advanced breeding lines from 2015-2021 to predict various end-use quality traits that are otherwise difficult to phenotype in earlier generations. The multivariate GP model outperformed the univariate model with up to a two-fold increase in prediction accuracy (PA). For instance, PA was improved from 0.38 to 0.75 for bake absorption and from 0.32 to 0.52 for loaf volume. Further, we compared multi-trait GP models by including different combinations of easy-to-score traits as model covariates to predict end-use quality traits and observed that the

incorporation of simple traits such as flour protein and flour sedimentation weight value can substantially improve the PA for baking traits.

Overall, the findings of these studies elucidate the potential of multivariate GP for agronomic traits when advanced breeding lines are used as training population to predict preliminary breeding lines. The results also showed the application of multivariate GP models in the breeding program can reduce phenotyping costs by facilitating a sparse testing design. Furthermore, we observed that the inclusion of rapid low-cost traits like flour protein and flour sedimentation weight value in MT genomic prediction models can facilitate the use of GS to predict baking traits in earlier generations and provide breeders an opportunity for selection on end-use quality traits by culling inferior lines to increase selection accuracy and genetic gains.

CHAPTER 1

INTRODUCTION

Wheat (*Triticum aestivum* L.) provides an adequate and affordable intake of calories and proteins in human diets and plays a critical role in food security (FAO, 2017; Grote et al., 2021). Global wheat production needs to be increased by 60% to meet the future demand of feeding 9 billion people by 2050 (Fischer et al., 2014); however, a gradual decrease in arable land and climate change is predicted to make this increase more challenging for the breeders (Grote et al., 2021; Wheeler & Von Braun, 2013). Thus, continued research efforts to understand the genetic basis of grain yield and the development of more productive wheat varieties remain the primary focus for all wheat breeding programs. Further, hard winter wheat is the major wheat class grown in the US (USDA NASS, 2021) which is known for its excellent milling and baking characteristics suitable for a variety of wheat foods, especially bread. Thus, wheat breeders must improve end-use quality traits while simultaneously breeding for increased yield to meet projected demand.

In conventional wheat breeding, the selection of progeny with desirable agronomic and end-use quality traits is a resource-intensive process and could take up to 10-15 years to develop a new cultivar (Haile et al., 2020). Further, in traits with complex genetic architecture such as grain yield and end-use quality, the genotype-by-environment interactions play a paramount role and impose additional challenges in selection. Nevertheless, Nevertheless, deployment of molecular markers for marker-assisted selection (MAS) has been used to increase selection accuracy and accumulation of desired traits in an efficient way (Randhawa et al., 2013).

Wheat grain yield is a complex quantitative trait involving many QTLs with small effects and is highly influenced by environmental factors, which makes it difficult to improve yield by direct phenotypic selection (K. Liu et al., 2018). Nevertheless, grain yield is mainly determined by several component traits exhibiting high heritability (J. Liu et al., 2018). Thus, a promising strategy to improve grain yield in wheat is to characterize individual yield component traits and exploit them for improving the yield potential (Kuzay et al., 2019; Würschum et al., 2018). In past decades, QTL mapping approaches including linkage mapping and genome-wide association study (GWAS) have been extensively used to identify QTLs governing these yield-related traits. However, the exploitation of GWAS to characterize yield component traits in winter wheat had been relatively limited (Ward et al., 2019; Zanke et al., 2015; Zhai et al., 2016). Moreover, GWAS have not been reported to date on spike and kernel traits in hard winter wheat (HWW). This necessitates the need to explore the phenotypic and genetic variation for yield-related traits in the important class of wheat to provide a useful resource to the breeders.

Genomic selection (GS) is another approach that utilizes genome-wide marker data to select individuals superior for complex traits in the early breeding cycle to increase the genetic gain per unit of time (Heffner et al., 2009; Meuwissen et al., 2001). Several studies have reported the successful implementation of GS in different crops resulting in an accelerated rate of genetic gain compared to traditional breeding (Bassi et al., 2015; Battenfield et al., 2016; Bhat et al., 2016). Moreover, GS has shown to be particularly useful in traits where phenotyping is cumbersome, such as quality traits and complex agronomic traits (Battenfield et al., 2016; Dong et al., 2018). The widespread

availability of genome-wide markers attributed to low-cost genotyping technologies has facilitated the adaptability of GS in wheat breeding programs (Bhat et al., 2016; Poland et al., 2012). Despite the successful evaluations of GS in wheat breeding programs, there is a continuous scope to improve the prediction accuracy/ability of genomic prediction (GP) models for quantitative traits to achieve higher genetic gains that will lead to the routine implementation of GS in various wheat breeding schemes. In recent years, multi-trait (MT) genomic prediction models have been suggested to improve the PA for a primary trait when secondary traits correlated to the primary trait are available (Jia & Jannink, 2012). Thus, there is a need to explore the usefulness of GS for the improvement of grain yield and end-use quality traits, which otherwise exhibit complex quantitative inheritance or are difficult to phenotype.

Based on this, the objectives of this study were:

1. Phenotypic and genetic characterization of yield-component traits in a diverse population of hard winter lines from the Great Plains region of the US.
2. Evaluation of univariate and multivariate genomic prediction models for predicting agronomic traits in advanced breeding lines of hard winter wheat.
3. Explore the usability of multi-trait genomic prediction models with rapid, small-scale, and NIRS-based traits as covariates for the prediction of processing and end-use quality traits in hard winter wheat.

References

Bassi, F. M., Bentley, A. R., Charmet, G., Ortiz, R., & Crossa, J. (2015). Breeding

schemes for the implementation of genomic selection in wheat (Triticum spp.). *Plant Science, 242*, 23–36. https://doi.org/10.1016/j.plantsci.2015.08.021

Battenfield, S. D., Guzmán, C., Gaynor, R. C., Singh, R. P., Peña, R. J., Dreisigacker, S., Fritz, A. K., & 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*(2). https://doi.org/10.3835/plantgenome2016.01.0005

Bhat, J. A., Ali, S., Salgotra, R. K., Mir, Z. A., Dutta, S., Jadon, V., Tyagi, A., Mushtaq, M., Jain, N., Singh, P. K., Singh, G. P., & Prabhu, K. V. (2016). Genomic selection in the era of next generation sequencing for complex traits in plant breeding. In *Frontiers in Genetics* (Vol. 7, Issue DEC). Frontiers Media S.A. https://doi.org/10.3389/fgene.2016.00221

Dong, H., Wang, R., Yuan, Y., Anderson, J., Pumphrey, M., Zhang, Z., & Chen, J. (2018). Evaluation of the Potential for Genomic Selection to Improve Spring Wheat Resistance to Fusarium Head Blight in the Pacific Northwest. *Frontiers in Plant Science*, *9*, 911. https://doi.org/10.3389/fpls.2018.00911

FAO. (2017). *The future of food and agriculture – Trends and challenges.* FAO.

Fischer, R., Byerlee, D., & Edmeades, G. (2014). Crop yields and global food security. In *academia.edu*. ACIAR: Canberra, ACT. http://www.academia.edu/download/35887178/Crop_yields_and_global_food_security___a_book_by_T.Fischer_et_al__2014.pdf

Grote, U., Fasse, A., Nguyen, T. T., & Erenstein, O. (2021). Food Security and the Dynamics of Wheat and Maize Value Chains in Africa and Asia. In *Frontiers in Sustainable Food Systems* (Vol. 4, p. 317). Frontiers Media S.A.

https://doi.org/10.3389/fsufs.2020.617009

Haile, T. A., Walkowiak, S., N'Diaye, A., Clarke, J. M., Hucl, P. J., Cuthbert, R. D., Knox, R. E., & Pozniak, C. J. (2020). Genomic prediction of agronomic traits in wheat using different models and cross-validation designs. *Theoretical and Applied Genetics*, *1*, 3. https://doi.org/10.1007/s00122-020-03703-z

Heffner, E. L., Sorrells, M. E., & Jannink, J. L. (2009). Genomic selection for crop improvement. In *Crop Science* (Vol. 49, Issue 1, pp. 1–12). https://doi.org/10.2135/cropsci2008.08.0512

Jia, Y., & Jannink, J. L. (2012). Multiple-trait genomic selection methods increase genetic value prediction accuracy. *Genetics*, *192*(4), 1513–1522. https://doi.org/10.1534/genetics.112.144246

Kuzay, S., Xu, Y., Zhang, J., Katz, A., Pearce, S., Su, Z., Fraser, M., Anderson, J. A., Brown-Guedira, G., DeWitt, N., Peters Haugrud, A., Faris, J. D., Akhunov, E., Bai, G., & Dubcovsky, J. (2019). Identification of a candidate gene for a QTL for spikelet number per spike on wheat chromosome arm 7AL by high-resolution genetic mapping. *Theoretical and Applied Genetics*, *132*(9), 2689–2705. https://doi.org/10.1007/s00122-019-03382-5

Liu, J., Xu, Z., Fan, X., Zhou, Q., Cao, J., Wang, F., Ji, G., Yang, L., Feng, B., & Wang, T. (2018). A genome-wide association study of wheat spike related traits in China. *Frontiers in Plant Science*, *871*, 1584. https://doi.org/10.3389/fpls.2018.01584

Liu, K., Sun, X., Ning, T., Duan, X., Wang, Q., Liu, T., An, Y., Guan, X., Tian, J., & Chen, J. (2018). Genetic dissection of wheat panicle traits using linkage analysis and a genome-wide association study. *Theoretical and Applied Genetics*, *131*(5), 1073–

1090. https://doi.org/10.1007/s00122-018-3059-9

Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. In *Genetics Soc America*. https://www.genetics.org/content/157/4/1819.short

Poland, J., Endelman, J., Dawson, J., Rutkoski, J., Wu, S., Manes, Y., Dreisigacker, S., Crossa, J., Sánchez-Villeda, H., Sorrells, M., & Jannink, J. (2012). Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing. *The Plant Genome*, *5*(3), plantgenome2012.06.0006. https://doi.org/10.3835/plantgenome2012.06.0006

Randhawa, H. S., Asif, M., Pozniak, C., Clarke, J. M., Graf, R. J., Fox, S. L., Humphreys, D. G., Knox, R. E., DePauw, R. M., Singh, A. K., Cuthbert, R. D., Hucl, P., & Spaner, D. (2013). Application of molecular markers to wheat breeding in Canada. *Plant Breeding*, *132*(5), n/a-n/a. https://doi.org/10.1111/pbr.12057

USDA NASS. (2021). *Small Grains 2021 Summary*. https://www.nass.usda.gov/Publications/Todays_Reports/reports/smgr0921.pdf

Ward, B. P., Brown-Guedira, G., Kolb, F. L., Van Sanford, D. A., Tyagi, P., Sneller, C. H., & Griffey, C. A. (2019). Genome-wide association studies for yield-related traits in soft red winter wheat grown in Virginia. *PLoS ONE*, *14*(2), e0208217. https://doi.org/10.1371/journal.pone.0208217

Wheeler, T., & Von Braun, J. (2013). Climate change impacts on global food security. In *Science* (Vol. 341, Issue 6145, pp. 508–513). American Association for the Advancement of Science. https://doi.org/10.1126/science.1239402

Würschum, T., Leiser, W. L., Langer, S. M., Tucker, M. R., & Longin, C. F. H. (2018). Phenotypic and genetic analysis of spike and kernel characteristics in wheat reveals

long-term genetic trends of grain yield components. *Theoretical and Applied Genetics*, *131*(10), 2071–2084. https://doi.org/10.1007/s00122-018-3133-3

Zanke, C. D., Ling, J., Plieske, J., Kollers, S., Ebmeyer, E., Korzun, V., Argillier, O., Stiewe, G., Hinze, M., Neumann, F., Eichhorn, A., Polley, A., Jaenecke, C., Ganal, M. W., & Röder, M. S. (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*(september), 644. https://doi.org/10.3389/fpls.2015.00644

Zhai, H., Feng, Z., Li, J., Liu, X., Xiao, S., Ni, Z., & Sun, Q. (2016). QTL analysis of spike morphological traits and plant height in winter wheat (Triticum aestivum L.) using a high-density SNP and SSR-based linkage map. *Frontiers in Plant Science*, *7*(November 2016). https://doi.org/10.3389/fpls.2016.01617

CHAPTER 2

LITERATURE REVIEW

2.1 General introduction

Wheat (*Triticum aestivum* L.) provides an adequate and affordable intake of calories and proteins in human diets and plays a critical role in food security (FAO, 2017; Grote et al., 2021). In the United States, six different classes of wheat are grown, which are designated by color, hardness, and growing season. The six classes of wheat include Hard Red Winter (HWW), Hard Red Spring (HRS), Hard White (HW), Soft White (SW), Soft Red Winter (SRW), and Durum wheat (https://www.uswheat.org/working-with-buyers/wheat-classes). Hard winter wheat (*Triticum aestivum* L.; HWW) is the major wheat class grown in the US and accounts for about 46 percent of the total wheat production in the country (USDA NASS, 2021). Further, HWW was the major class of wheat grown in South Dakota based on acreage and production https://www.nass.usda.gov/Quick_Stats/Ag_Overview/stateOverview.php?state=SOUTH%20DAKOTA). This versatile class of wheat exhibits excellent milling and baking characteristics suitable for a variety of wheat foods, especially bread. Owing to high demand, most of the US-produced HWW is exported. For instance, 52 percent of the total HWW produced in 2020 was exported worldwide (USDA ERS, 2022). Throughout the wheat supply chain, end-use quality characteristics play an important role in the marketing and pricing of HWW (Roberts et al., 2022).

2.1.1 Origin and Domestication

Bread wheat or common wheat (*Triticum aestivum* L.) is a member of the Triticeae tribe within the Poaceae family. The *Triticum* genus encompasses about 25 distinct species, which include species containing a single genome (diploid wheat) or others having multiple homoeologous genomes resulting from hybridization (tetraploid or hexaploid species). For example, diploid forms include wild species (*Triticum urartu* with AA genome), cultivates einkorn wheat (*Triticum monococcum* with AA genome), allotetraploid emmer wheat (*Triticum turgidum* var. *durum* with AABB genome), and the allohexaploid common wheat (*Triticum aestivum* L. with AABBDD genome). (Kim et al., 2017).

Common Bread Wheat is an allohexaploid (6x) species with three sets of homeologous chromosomes designated as A, B, and D subgenomes, with a large genome size of about 17 gigabases (Gb) (William et al., 2007). The hexaploid bread wheat (AABBDD, 2n = 6x = 42) is believed to have originated from hybridization between the diploid (DD) genome of grass species *Aegilops tauschii* and the domesticated emmer wheat *T. turgidum* ssp. *dicoccum*, a tetraploid with AABB genome (Dubcovsky & Dvorak, 2007). Further, *Triticum urartu* is considered as the progenitor of the A genome of bread wheat, while the B genome of bread wheat is believed several were believed to have originated from an annual diploid S genome species in the genus *Aegilops* sect. *sitopsis* (Feldman & Levy, 2015). The hybridization event between tetraploid (*T. turgidum* subsp. *dicoccum*) wheat and the D subgenome donor species (*Ae. tauschii*; 2n=2x=14, DD) is believed to have happened spontaneously in the Caspian Sea region about 9,000 years that gave rise to the modern bread wheat, *T. aestivum* (2n=6x=42, AABBDD). *Triticum aestivum* ssp. *vulgare* is commonly called bread wheat, while other

sub species in this group include *compactum*, *sphaerococcum*, *spelta*, *macha* and *vavilovii* (Shewry, 2009).

The transition from the tetraploid form to the modern hexaploid resulted in enhanced geographic and environmental adaptability, along with increased grain yield and quality. The event of wheat domestication is believed to have occurred in the present-day Middle East, and had a profound impact on the development and evolution of human civilization, as this transition led human civilization to a more agrarian society (Eckardt, 2010).

2.2 Challenges and opportunities in wheat yield improvement

Wheat grain yield is a complex quantitative trait influenced by various factors such as morphological characteristics, physiological indices, grain-related traits, and different environmental conditions, involving many quantitative trait loci (QTLs) with small effects and is highly influenced by environmental factors, which makes it difficult to improve yield by direct phenotypic selection (K. Liu et al., 2018). Understanding the genetic basis of grain yield is further hampered by the low heritability of this trait (Kuzay et al., 2019). However, grain yield is a collective output of several component traits namely spikelet number per spike (NSPS), spike length (SL), spike number, kernels per spike (KPS), kernel size (KS), and thousand kernel weight (TKW), which are less sensitive to the environment and exhibit higher heritability than that of grain yield per se (Hai et al., 2008; Kato et al., 2000). Among these component traits, three traits play a major role in determining final grain yield i.e. spike number per unit area, kernel number per spike, and thousand kernel weight (TKW) (J. Liu et al., 2018). The final kernel number per spike is affected by spike-related traits such as spike length (SL), number of

spikelets per spike (NSPS), and spikelet density (SD) (Z. Guo et al., 2017). Similarly, TKW is influenced by several kernel morphology traits such as kernel length (KL), kernel width (KW), and kernel area (KA) (Gegas et al., 2010). Most of these yield component traits are controlled by several QTLs; however, these traits are less sensitive to the environment and exhibit higher heritability compared to grain yield per se (Zhang et al., 2018). Therefore, a promising strategy to improve grain yield in wheat is to understand the genes and gene networks controlling individual yield component traits and exploiting them for improving the yield potential (Kuzay et al., 2019; Würschum et al., 2018).

2.3 Breeding for end-use quality in hard winter wheat

Hard winter wheat is the major wheat class grown in the US and exhibits excellent milling and baking characteristics suitable for a variety of wheat foods, especially bread. Owing to high demand, most of the US-produced HWW is exported. For instance, 52 percent of the total HWW produced in 2020 was exported worldwide (USDA ERS, 2022). Throughout the wheat supply chain, end-use quality characteristics play an important role in the marketing and pricing of HWW (Roberts et al., 2022). Moreover, consumers' preferences for healthier food necessitate an emphasis on the selection of desirable end-use quality traits. Henceforth, in HWW breeding, end-use quality and processing traits are important factors in varietal development and determining acceptance by the industry.

The high gluten strength and damaged starch in HWW make it very suitable for baking, and yeast-leavened bread is a major end-use product. Bread quality is an important but complex trait that is defined by a combination of many parameters

(Battenfield et al., 2016). Several important factors including kernel characteristics, the milled flour quality, protein and starch strength, and dough properties all play a crucial role in determining the end-use quality of the final product. Hence, several assays are used to profile these factors and inform the selection for end-use quality. Nevertheless, most of the assays for the evaluation of end-use products are expensive, time-consuming, and require large quantities of flour. Therefore, breeders mostly prioritize the selection of agronomic traits and disease resistance in earlier generations and quality traits in advanced generations in most breeding programs (Battenfield et al., 2016). Previous studies have shown that end-use quality traits are controlled by a few major genes and a large number of quantitative trait loci with minor effects (Carter et al., 2012; Jernigan et al., 2018; Kiszonas & Morris, 2017; Sandhu et al., 2021). Though available major genes have been exploited in breeding programs, the majority of minor genes are highly influenced by the environment and remain uncharacterized (Jernigan et al., 2018; Kiszonas & Morris, 2017).

2.4 Genetic characterization of agronomic traits for marker-assisted selection

Traditional wheat breeding involves creating novel genetic variation by different methods, followed by extensive selection and advancement of generations. The selection of progeny with desirable agronomic and end-use quality traits is a resource-intensive process and could take up to 10-15 years to develop a new cultivar (Haile et al., 2020). Further, in traits with complex genetic architecture such as grain yield, the genotype-by-environment interactions play a paramount role and impose additional challenges in selection. In recent years, genetic characterization of complex traits using molecular markers followed by deployment of markers linked/associated with a trait of interest for

marker-assisted selection (MAS) has been used to increase selection accuracy and accelerate genetic gain (Randhawa et al., 2013). Moreover, recent advents in sequencing technologies and advanced molecular marker technologies have facilitated the genetic characterization of complex traits with improved resolutions. For example, Single nucleotide polymorphisms (SNPs) have emerged as powerful molecular markers in the recent past and become extremely popular in plant breeding research owing to their genome-wide abundance and ability to capture variations quickly (Korte & Ashley, 2013). Further, the discovery of SNPs has been significantly improved by rapid advances in sequencing technologies (Allen et al., 2011; Berkman et al., 2012; J. A. Poland et al., 2012). (Thomson, 2014). Recently, SNPs have been widely employed for the identification of quantitative trait locus (QTL) for important agronomic traits in various crop species. (Cook et al., 2012; Chen et al., 2016; Halder et al., 2019; Kuzay et al., 2019; Sidhu et al., 2020; Yang et al., 2020)

There are two common methods used for genetic characterization of complex traits in plants including linkage analysis (QTL mapping) and association mapping, also known as LD mapping (Gupta et al., 2014). Linkage mapping or QTL mapping is based on genetic recombination events for a specific trait by establishing a segregation population such as an F_2 population, Doubled Haploid (DH), or a recombinant inbred lines (RIL) population; and has been used to characterize both qualitative and quantitative traits(Collard & Mackill, 2008; Gupta et al., 2014). Though linkage mapping is a powerful tool and has been widely used to characterize important traits in a variety of crops, there are several shortcomings of this approach. It involves the development of segregation populations which could be time-consuming and detects only those QTL that

are polymorphic in the given population (Ayana et al., 2018). Further, it suffers from limited genetic variation and is able to exploit only one or few meiotic generations resulting in low resolution of QTL mapping (Gupta et al., 2014).

Genome-wide association mapping also known as linkage-disequilibrium mapping, is another approach that uses genome-wide dense markers and relies on linkage disequilibrium to uncover the association between genotype and phenotype (Myles et al., 2009; Randhawa et al., 2013; Zhu et al., 2008). Association mapping, unlike linkage mapping, utilizes genetic diversity across natural populations to detect molecular polymorphisms that associate with the phenotypic variation for the trait of interest, thus offering higher mapping resolution by exploiting historic recombination events in broad-based diversity panels (Gupta et al., 2014; Zhu et al., 2008). In the recent past, Genome-wide association studies (GWAS) have been used in a variety of economically important crops, including Arabidopsis, wheat, maize, rice, barley, soybean, tomato, etc. (Wang et al., 2014; Xu et al., 2017). Despite having several merits over linkage mapping, association mapping suffers from a few limitations, as it leads to a high frequency of false-positive associations due to the kinship and population structure that may exist among the population (Neumann et al., 2011; Korte and Farlow, 2013). Nevertheless, there have been recent developments in the statistical methods underlying association mapping to overcome these limitations and novel single or multi-locus models have been developed to control for false-positives and increase the power of mapping. In comparison to the naïve General Linear Model (GLM) method (Price et al., 2006), Mixed Linear Model (MLM) takes account of kinship and population structure in association analysis reducing type I error, and controlling for false-positive associations (Yu et al.,

2006). Further, the introduction of multi-locus models, including Multi-locus analysis such as Multiple Loci Mixed Linear Model (MLMM) (Segura et al., 2012) and Fixed and random model Circulating Probability Unification (FarmCPU) (Liu et al., 2016) have increased the power of association mapping and provides better control of false-positive associations. Recently, s multi-locus Bayesian-information and Linkage-disequilibrium Iteratively Nested Keyway (BLINK) method was developed and reported to perform better than the popular multi-locus FarmCPU approach (Huang et al., 2019; X. Liu et al., 2016) because it overcomes the limitations of the FarmCPU and identifies more true positives and produces fewer false positives when compared to other GWAS methods (Huang et al., 2019). In BLINK, the bin method of FarmCPU is replaced by LD information, eliminating the requirement that causal genes are evenly distributed (Huang et al., 2019). This method showed better performance than other models using simulated data as well as in several empirical studies from different crop species (Habyarimana et al., 2020; Juliana et al., 2021; L. Liu et al., 2020).

2.5 Genomic Selection (GS)

Though MAS has shown good potential in wheat breeding for the deployment of QTLs with large effects, its application has been limited to improving complex traits governed by many QTLs with small effects (Heffner et al., 2009). Genomic selection (GS) is a recent approach that utilizes genome-wide marker data to select individuals superior for complex traits in the early breeding cycle to increase the genetic gain per unit of time (Heffner et al., 2009; Meuwissen et al., 2001). Unlike MAS, GS does not require prior identification of QTLs for the traits of interest; instead, it employs all available markers across the genome to predict individuals' breeding values (Bassi et al., 2015). Briefly, GS

requires a training population (TP), which is genotyped with genome-wide markers and phenotyped for a given trait(s) of interest. GS involves the calibration of a prediction model using TP to estimate marker effects and evaluate the predictive ability of the model through cross-validation. Finally, the developed model is used to calculate genome-estimated breeding values (GEBVs) and rank the lines from a breeding or testing population (BP) that consists of lines with only genotypic information. Thus, the early selection or culling of individuals based on the GEBVs permits greater genetic gain per breeding cycle, facilitating an increase in the efficacy of breeding programs and resulting in reduced varietal development costs. Several studies have reported the successful implementation of GS in different crops resulting in an accelerated rate of genetic gain compared to traditional breeding (Bassi et al., 2015; Battenfield et al., 2016; Bhat et al., 2016). Moreover, GS has shown to be particularly useful in traits where phenotyping is cumbersome, such as quality traits and complex resistance to diseases (Battenfield et al., 2016; Dong et al., 2018).

The widespread availability of genome-wide markers attributed to low-cost genotyping technologies has facilitated the adaptability of GS in wheat breeding programs (Bhat et al., 2016; J. Poland et al., 2012). Thus, there is growing interest in recent years to complement phenotyping selection and genomic selection in wheat breeding. GS has been evaluated for many complex traits in wheat, including but not limited to grain yield and yield-related traits (Guo et al., 2020; Haile et al., 2020; Juliana et al., 2020; Rutkoski et al., 2016; Ward et al., 2019), wheat resistance to rusts (Juliana et al., 2017; J. E. Rutkoski et al., 2014) and Fusarium head blight (Arruda et al., 2015; Dong et al., 2018; Rutkoski et al., 2012), and end-use quality traits (Battenfield et al., 2016;

Ibba et al., 2020; Lado et al., 2018). Despite the successful evaluations of GS in wheat breeding programs, there is a continuous scope to improve the prediction accuracy/ability of GS models for quantitative traits to achieve higher genetic gains that will lead to the routine implementation of GS in various wheat breeding schemes.

2.6 Genomic prediction models and predictive ability

The predictive ability (PA) of the GS model refers to the correlation between estimated GEBVs and the actual phenotypic values of the individuals in the validation set and is generally calculated through a cross-validation approach. Along with TP size, the extent of linkage disequilibrium (LD), and the heritability of the traits, the PA also depends on the choice and optimization of the statistical models (de los Campos et al., 2013; J. Guo et al., 2020; J. Rutkoski et al., 2016). In most studies, penalized genomic prediction models, including ridge-regression best linear unbiased prediction (rrBLUP) and genomic best linear unbiased prediction (GBLUP), have been standard GS approaches (Endelman, 2011; VanRaden et al., 2009). In addition, several Bayesian methods with different prior distributions and relying on Markov-Chain Monte Carlo (MCMC) for the estimation of parameters have proven useful for genomic prediction (Habier et al., 2011; Wang et al., 2018). Among Bayesian models, Bayes A (BA) and Bayes B (BB) are commonly used genomic prediction models, which assume different prior distributions for estimating marker effects and variances (Pérez & De Los Campos, 2014). The Bayes A model uses the scaled inverse chi-squared probability distribution for estimating marker variances. Bayes B is an extension of the Bayes A model (Meuwissen et al., 2001) and employs an inverse chi-square distribution for marker effects and assumes that some markers have no

effect. However, most of these models implement a univariate linear mixed model and are helpful to predict one dependent variable at a time.

In recent years, multi-trait (MT) genomic prediction models have been suggested to improve the PA for a primary trait when secondary traits correlated to the primary trait are available (Jia & Jannink, 2012). The use of genetically correlated traits is of particular importance when the primary trait is difficult or expensive to phenotype and has low heritability. Several empirical studies have successfully evaluated MT approaches for different agronomic traits in wheat breeding (Hayes et al., 2017; Lado et al., 2018; Rutkoski et al., 2012). Improvement of 70% in the PA for grain yield was observed by including canopy temperature (CT) and normalized difference vegetation index as secondary traits using the MT approach (Rutkoski et al., 2016; Sun et al., 2017). Similarly, Hayes et al., (2017) and Lado et al., (2018) observed an increase in PA using multivariate approaches (MT) over single trait (ST) models in end-use quality traits.

For complex traits, genotype-by-environment interactions (G x E) necessitate the evaluation of breeding lines for multiple traits over multiple environments. Thus, the extension of the MT approaches to account for the trait x genotype x environment (T x G x E) interaction could improve the model for genomic prediction accuracy in breeding programs. Montesinos-López et al. (2016) proposed a Bayesian multi-trait and multi-environment (BMTME) model that integrates the analysis of multi-traits recorded over multi-environments and accounts for T x G x E interaction in a unified approach. Recently, an improved BMTME model has been introduced that estimates the variance-covariance structure among trait, genotype, and environment to predict multiple traits

evaluated in various environments (Montesinos-López et al., 2019; Montesinos-López et al., 2019). Few studies using simulated and empirical data found that the BMTME model outperforms ST models in agronomic and end-use quality traits in wheat (Guo et al., 2020; Ibba et al., 2020; Montesinos-López et al., 2016). Better performance of multivariate GS approaches stimulates us to evaluate these models in an actual breeding pipeline, where several traits are evaluated over diverse environments.

2.7 References

Allen, A. M., Barker, G. L. A., Berry, S. T., Coghill, J. A., Gwilliam, R., Kirby, S., Robinson, P., Brenchley, R. C., D'Amore, R., McKenzie, N., Waite, D., Hall, A., Bevan, M., Hall, N., & Edwards, K. J. (2011). Transcript-specific, single-nucleotide polymorphism discovery and linkage analysis in hexaploid bread wheat (Triticum aestivum L.). *Plant Biotechnology Journal*, *9*(9), 1086–1099. https://doi.org/10.1111/j.1467-7652.2011.00628.x

AlTameemi, R., Gill, H. S., Ali, S., Ayana, G., Halder, J., Sidhu, J. S., Gill, U. S., Turnipseed, B., Hernandez, J. L. G., & Sehgal, S. K. (2021). Genome-wide association analysis permits characterization of Stagonospora nodorum blotch (SNB) resistance in hard winter wheat. *Scientific Reports*, *11*(1), 12570. https://doi.org/10.1038/s41598-021-91515-6

Arruda, M. P., Brown, P. J., Lipka, A. E., Krill, A. M., Thurber, C., & Kolb, F. L. (2015). Genomic Selection for Predicting *Fusarium* Head Blight Resistance in a Wheat Breeding Program. *The Plant Genome*, *8*(3).

https://doi.org/10.3835/plantgenome2015.01.0003

Ayana, G. T., Ali, S., Sidhu, J. S., Gonzalez Hernandez, J. L., Turnipseed, B., & Sehgal, S. K. (2018). Genome-Wide Association Study for Spot Blotch Resistance in Hard Winter Wheat. *Frontiers in Plant Science*, *9*(July), 1–15. https://doi.org/10.3389/fpls.2018.00926

Bassi, F. M., Bentley, A. R., Charmet, G., Ortiz, R., & Crossa, J. (2015). Breeding schemes for the implementation of genomic selection in wheat (Triticum spp.). *Plant Science*, *242*, 23–36. https://doi.org/10.1016/j.plantsci.2015.08.021

Battenfield, S. D., Guzmán, C., Gaynor, R. C., Singh, R. P., Peña, R. J., Dreisigacker, S., Fritz, A. K., & 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*(2). https://doi.org/10.3835/plantgenome2016.01.0005

Berkman, P. J., Lai, K., Lorenc, M. T., & Edwards, D. (2012). Next-generation sequencing applications for wheat crop improvement. *American Journal of Botany*, *99*(2), 365–371. https://doi.org/10.3732/ajb.1100309

Bhat, J. A., Ali, S., Salgotra, R. K., Mir, Z. A., Dutta, S., Jadon, V., Tyagi, A., Mushtaq, M., Jain, N., Singh, P. K., Singh, G. P., & Prabhu, K. V. (2016). Genomic selection in the era of next generation sequencing for complex traits in plant breeding. In *Frontiers in Genetics* (Vol. 7, Issue DEC). Frontiers Media S.A. https://doi.org/10.3389/fgene.2016.00221

Chen, G., Zhang, H., Deng, Z., Wu, R., Li, D., Wang, M., & Tian, J. (2016). Genome-wide association study for kernel weight-related traits using SNPs in a Chinese

winter wheat population. *Euphytica*, *212*(2), 173–185. https://doi.org/10.1007/s10681-016-1750-y

Collard, B. C. Y., & Mackill, D. J. (2008). Marker-assisted selection: An approach for precision plant breeding in the twenty-first century. In *Philosophical Transactions of the Royal Society B: Biological Sciences* (Vol. 363, Issue 1491, pp. 557–572). Royal Society. https://doi.org/10.1098/rstb.2007.2170

Cook, J. P., McMullen, M. D., Holland, J. B., Tian, F., Bradbury, P., Ross-Ibarra, J., Buckler, E. S., & Flint-Garcia, S. A. (2012). Genetic architecture of maize kernel composition in the nested association mapping and inbred association panels. *Plant Physiology*, *158*(2), 824–834. https://doi.org/10.1104/pp.111.185033

de los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D., & Calus, M. P. L. (2013). Whole-genome regression and prediction methods applied to plant and animal breeding. In *Genetics* (Vol. 193, Issue 2, pp. 327–345). https://doi.org/10.1534/genetics.112.143313

Dong, H., Wang, R., Yuan, Y., Anderson, J., Pumphrey, M., Zhang, Z., & Chen, J. (2018). Evaluation of the Potential for Genomic Selection to Improve Spring Wheat Resistance to Fusarium Head Blight in the Pacific Northwest. *Frontiers in Plant Science*, *9*, 911. https://doi.org/10.3389/fpls.2018.00911

Dubcovsky, J., & Dvorak, J. (2007). Genome Plasticity a Key Factor in the Success of Polyploid Wheat Under Domestication. *Science*, *316*(5833), 1862–1866. https://doi.org/10.1126/science.1143986

Eckardt, N. A. (2010). Evolution of domesticated bread wheat. *The Plant Cell*, *22*(4),

993. https://doi.org/10.1105/tpc.110.220410

Endelman, J. B. (2011). Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *The Plant Genome*, *4*(3), 250–255. https://doi.org/10.3835/plantgenome2011.08.0024

FAO. (2017). *The future of food and agriculture – Trends and challenges.* FAO.

Feldman, M., & Levy, A. A. (2015). Origin and Evolution of Wheat and Related Triticeae Species. *Alien Introgression in Wheat: Cytogenetics, Molecular Biology, and Genomics*, 21–76. https://doi.org/10.1007/978-3-319-23494-6_2

Gegas, V. C., Nazari, A., Griffiths, S., Simmonds, J., Fish, L., Orford, S., Sayers, L., Doonan, J. H., & Snape, J. W. (2010). A Genetic Framework for Grain Size and Shape Variation in Wheat . *The Plant Cell*, *22*(4), 1046–1056. https://doi.org/10.1105/tpc.110.074153

Gill, H. S., Halder, J., Zhang, J., Brar, N. K., Rai, T. S., Hall, C., Bernardo, A., Amand, P. S., Bai, G., Olson, E., Ali, S., Turnipseed, B., & Sehgal, S. K. (2021). Multi-Trait Multi-Environment Genomic Prediction of Agronomic Traits in Advanced Breeding Lines of Winter Wheat. *Frontiers in Plant Science*, *12*. https://doi.org/10.3389/fpls.2021.709545

Grote, U., Fasse, A., Nguyen, T. T., & Erenstein, O. (2021). Food Security and the Dynamics of Wheat and Maize Value Chains in Africa and Asia. In *Frontiers in Sustainable Food Systems* (Vol. 4, p. 317). Frontiers Media S.A. https://doi.org/10.3389/fsufs.2020.617009

Guo, J., Khan, J., Pradhan, S., Shahi, D., Khan, N., Avci, M., Mcbreen, J., Harrison, S., Brown-Guedira, G., Murphy, J. P., Johnson, J., Mergoum, M., Esten Mason, R., Ibrahim, A. M. H., Sutton, R., Griffey, C., & Babar, M. A. (2020). Multi-Trait Genomic Prediction of Yield-Related Traits in US Soft Wheat under Variable Water Regimes. *Genes*, *11*(11), 1270. https://doi.org/10.3390/genes11111270

Guo, Z., Chen, D., Alqudah, A. M., Röder, M. S., Ganal, M. W., & Schnurbusch, T. (2017). Genome-wide association analyses of 54 traits identified multiple loci for the determination of floret fertility in wheat. *New Phytologist*, *214*(1), 257–270. https://doi.org/10.1111/nph.14342

Gupta, P. K., Kulwal, P. L., & Jaiswal, V. (2014). Association mapping in crop plants: Opportunities and challenges. In *Advances in Genetics* (Vol. 85, pp. 109–147). Academic Press Inc. https://doi.org/10.1016/B978-0-12-800271-1.00002-0

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

Habyarimana, E., De Franceschi, P., Ercisli, S., Baloch, F. S., & Dall'Agata, M. (2020). Genome-Wide Association Study for Biomass Related Traits in a Panel of Sorghum bicolor and S. bicolor × S. halepense Populations. *Frontiers in Plant Science, 11*, 1796. https://doi.org/10.3389/fpls.2020.551305

Hai, L., Guo, H., Wagner, C., Xiao, S., & Friedt, W. (2008). *Plant Science Genomic regions for yield and yield parameters in Chinese winter wheat (Triticum aestivum L .) genotypes tested under varying environments correspond to QTL in widely*

different wheat materials. *175*, 226–232. https://doi.org/10.1016/j.plantsci.2008.03.006

Haile, T. A., Walkowiak, S., N'Diaye, A., Clarke, J. M., Hucl, P. J., Cuthbert, R. D., Knox, R. E., & Pozniak, C. J. (2020). Genomic prediction of agronomic traits in wheat using different models and cross-validation designs. *Theoretical and Applied Genetics*, *1*, 3. https://doi.org/10.1007/s00122-020-03703-z

Halder, J., Zhang, J., Ali, S., Sidhu, J. S., Gill, H. S., Talukder, S. K., Kleinjan, J., Turnipseed, B., & Sehgal, S. K. (2019). Mining and genomic characterization of resistance to tan spot, Stagonospora nodorum blotch (SNB), and Fusarium head blight in Watkins core collection of wheat landraces. *BMC Plant Biology*, *19*(1), 1–15. https://doi.org/10.1186/s12870-019-2093-3

Hayes, B. J., Panozzo, J., Walker, C. K., Choy, A. L., Kant, S., Wong, D., Tibbits, J., Daetwyler, H. D., Rochfort, S., Hayden, M. J., & Spangenberg, G. C. (2017). Accelerating wheat breeding for end-use quality with multi-trait genomic predictions incorporating near infrared and nuclear magnetic resonance-derived phenotypes. *Theoretical and Applied Genetics*, *130*(12), 2505–2519. https://doi.org/10.1007/s00122-017-2972-7

Heffner, E. L., Sorrells, M. E., & Jannink, J. L. (2009). Genomic selection for crop improvement. In *Crop Science* (Vol. 49, Issue 1, pp. 1–12). https://doi.org/10.2135/cropsci2008.08.0512

Huang, M., Liu, X., Zhou, Y., Summers, R. M., & Zhang, Z. (2019). BLINK: A package for the next level of genome-wide association studies with both individuals and

markers in the millions. *GigaScience*, *8*(2), 1–12. https://doi.org/10.1093/gigascience/giy154

Ibba, M. I., Crossa, J., Montesinos-López, O. A., Montesinos-López, A., Juliana, P., Guzman, C., Delorean, E., Dreisigacker, S., & Poland, J. (2020). Genome-based prediction of multiple wheat quality traits in multiple years. *The Plant Genome*, *13*(3). https://doi.org/10.1002/tpg2.20034

Jia, Y., & Jannink, J. L. (2012). Multiple-trait genomic selection methods increase genetic value prediction accuracy. *Genetics*, *192*(4), 1513–1522. https://doi.org/10.1534/genetics.112.144246

Juliana, P., Singh, R. P., Braun, H.-J., Huerta-Espino, J., Crespo-Herrera, L., Govindan, V., Mondal, S., Poland, J., & Shrestha, S. (2020). Genomic Selection for Grain Yield in the CIMMYT Wheat Breeding Program—Status and Perspectives. *Frontiers in Plant Science*, *11*, 1. https://doi.org/10.3389/fpls.2020.564183

Juliana, P., Singh, R. P., Poland, J., Shrestha, S., Huerta-Espino, J., Govindan, V., Mondal, S., Crespo-Herrera, L. A., Kumar, U., Joshi, A. K., Payne, T., Bhati, P. K., Tomar, V., Consolacion, F., & Campos Serna, J. A. (2021). Elucidating the genetics of grain yield and stress-resilience in bread wheat using a large-scale genome-wide association mapping study with 55,568 lines. *Scientific Reports*, *11*(1), 1–15. https://doi.org/10.1038/s41598-021-84308-4

Juliana, P., Singh, R. P., Singh, P. K., Crossa, J., Huerta-Espino, J., Lan, C., Bhavani, S., Rutkoski, J. E., Poland, J. A., Bergstrom, G. C., & Sorrells, M. E. (2017). Genomic and pedigree-based prediction for leaf, stem, and stripe rust resistance in wheat.

Theoretical and Applied Genetics, *130*(7), 1415–1430. https://doi.org/10.1007/s00122-017-2897-1

Kato, K., Miura, H., & Sawada, S. (2000). Mapping QTLs controlling grain yield and its components on chromosome 5A of wheat. *Theoretical and Applied Genetics*, *101*(7), 1114–1121. https://doi.org/10.1007/s001220051587

Kim, S. K., Kim, J.-H., & Jang, W.-C. (2017). Past, Present and Future Molecular Approaches to Improve Yield in Wheat. *Wheat Improvement, Management and Utilization*. https://doi.org/10.5772/67112

Korte, A., & Ashley, F. (2013). The advantages and limitations of trait analysis with GWAS : a review Self-fertilisation makes Arabidopsis particularly well suited to GWAS. *Plant Methods*, *9*(1), 29.

Kuzay, S., Xu, Y., Zhang, J., Katz, A., Pearce, S., Su, Z., Fraser, M., Anderson, J. A., Brown-Guedira, G., DeWitt, N., Peters Haugrud, A., Faris, J. D., Akhunov, E., Bai, G., & Dubcovsky, J. (2019). Identification of a candidate gene for a QTL for spikelet number per spike on wheat chromosome arm 7AL by high-resolution genetic mapping. *Theoretical and Applied Genetics*, *132*(9), 2689–2705. https://doi.org/10.1007/s00122-019-03382-5

Lado, B., Vázquez, D., Quincke, M., Silva, P., Aguilar, I., & Gutiérrez, L. (2018). Resource allocation optimization with multi-trait genomic prediction for bread wheat (Triticum aestivum L.) baking quality. *Theoretical and Applied Genetics*, *131*(12), 2719–2731. https://doi.org/10.1007/s00122-018-3186-3

Liu, J., Xu, Z., Fan, X., Zhou, Q., Cao, J., Wang, F., Ji, G., Yang, L., Feng, B., & Wang,

T. (2018). A genome-wide association study of wheat spike related traits in China. *Frontiers in Plant Science*, *871*, 1584. https://doi.org/10.3389/fpls.2018.01584

Liu, K., Sun, X., Ning, T., Duan, X., Wang, Q., Liu, T., An, Y., Guan, X., Tian, J., & Chen, J. (2018). Genetic dissection of wheat panicle traits using linkage analysis and a genome-wide association study. *Theoretical and Applied Genetics*, *131*(5), 1073–1090. https://doi.org/10.1007/s00122-018-3059-9

Liu, L., Wang, M., Zhang, Z., See, D. R., & Chen, X. (2020). Identification of Stripe Rust Resistance Loci in U.S. Spring Wheat Cultivars and Breeding Lines Using Genome-Wide Association Mapping and Yr Gene Markers. *Plant Disease*, *104*(8), 2181–2192. https://doi.org/10.1094/PDIS-11-19-2402-RE

Liu, X., Huang, M., Fan, B., Buckler, E. S., & Zhang, Z. (2016). Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient Genome-Wide Association Studies. *PLOS Genetics*, *12*(2), e1005767. https://doi.org/10.1371/journal.pgen.1005767

Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. In *Genetics Soc America*. https://www.genetics.org/content/157/4/1819.short

Montesinos-López, O. A., Montesinos-López, A., Crossa, J., Toledo, F. H., Pérez-Hernández, O., Eskridge, K. M., & Rutkoski, J. (2016). A genomic bayesian multi-trait and multi-environment model. *G3: Genes, Genomes, Genetics*, *6*(9), 2725–2774. https://doi.org/10.1534/g3.116.032359

Montesinos-López, O. A., Montesinos-López, A., Luna-Vázquez, F. J., Toledo, F. H.,

Pérez-Rodríguez, P., Lillemo, M., & Crossa, J. (2019). An R package for Bayesian analysis of multi-environment and multi-trait multi-environment data for genome-based prediction. *G3: Genes, Genomes, Genetics*, *9*(5), 1355–1369. https://doi.org/10.1534/g3.119.400126

Montesinos-López, O. A., Montesinos-López, A., Hernández, M. V., Ortiz-Monasterio, I., Pérez-Rodríguez, P., Burgueño, J., & Crossa, J. (2019). Multivariate Bayesian Analysis of On-Farm Trials with Multiple-Trait and Multiple-Environment Data. *Agronomy Journal*, *111*(6), 2658–2669. https://doi.org/10.2134/agronj2018.06.0362

Myles, S., Peiffer, J., Brown, P. J., Ersoz, E. S., Zhang, Z., Costich, D. E., & Buckler, E. (2009). Association mapping: Critical considerations shift from genotyping to experimental design. *Plant Cell*, *21*(8), 2194–2202. https://doi.org/10.1105/tpc.109.068437

Pérez, P., & De Los Campos, G. (2014). Genome-wide regression and prediction with the BGLR statistical package. *Genetics*, *198*(2), 483–495. https://doi.org/10.1534/genetics.114.164442

Poland, J. A., Brown, P. J., Sorrells, M. E., & 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*(2). https://doi.org/10.1371/journal.pone.0032253

Poland, J., Endelman, J., Dawson, J., Rutkoski, J., Wu, S., Manes, Y., Dreisigacker, S., Crossa, J., Sánchez-Villeda, H., Sorrells, M., & Jannink, J. (2012). Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing. *The Plant Genome*,

5(3), plantgenome2012.06.0006. https://doi.org/10.3835/plantgenome2012.06.0006

Randhawa, H. S., Asif, M., Pozniak, C., Clarke, J. M., Graf, R. J., Fox, S. L., Humphreys, D. G., Knox, R. E., DePauw, R. M., Singh, A. K., Cuthbert, R. D., Hucl, P., & Spaner, D. (2013). Application of molecular markers to wheat breeding in Canada. *Plant Breeding*, *132*(5), n/a-n/a. https://doi.org/10.1111/pbr.12057

Roberts, S., Brooks, K., Nogueira, L., & Walters, C. G. (2022). The role of quality characteristics in pricing hard red winter wheat. *Food Policy, 108*, 102246. https://doi.org/10.1016/j.foodpol.2022.102246

Rutkoski, J., Benson, J., Jia, Y., Brown-Guedira, G., Jannink, J.-L., & Sorrells, M. (2012). Evaluation of Genomic Prediction Methods for Fusarium Head Blight Resistance in Wheat. *The Plant Genome, 5*(2), 51–61. https://doi.org/10.3835/plantgenome2012.02.0001

Rutkoski, J. E., Poland, J. A., Singh, R. P., Huerta-Espino, J., Bhavani, S., Barbier, H., Rouse, M. N., Jannink, J., & Sorrells, M. E. (2014). Genomic Selection for Quantitative Adult Plant Stem Rust Resistance in Wheat. *The Plant Genome*, *7*(3). https://doi.org/10.3835/plantgenome2014.02.0006

Rutkoski, J., Poland, J., Mondal, S., Autrique, E., Pérez, L. G., Crossa, J., Reynolds, M., & Singh, R. (2016). Canopy temperature and vegetation indices from high-throughput phenotyping improve accuracy of pedigree and genomic selection for grain yield in wheat. *G3: Genes, Genomes, Genetics*, *6*(9), 2799–2808. https://doi.org/10.1534/g3.116.032888

Shewry, P. R. (2009). Wheat. *Journal of Experimental Botany, 60*(6), 1537–1553.

https://doi.org/10.1093/jxb/erp058

Sidhu, J. S., Singh, D., Gill, H. S., Brar, N. K., Qiu, Y., Halder, J., Al Tameemi, R., Turnipseed, B., & Sehgal, S. K. (2020). Genome-Wide Association Study Uncovers Novel Genomic Regions Associated With Coleoptile Length in Hard Winter Wheat. *Frontiers in Genetics*, *10*, 1345. https://doi.org/10.3389/fgene.2019.01345

Sun, J., Rutkoski, J. E., Poland, J. A., Crossa, J., Jannink, J., & 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*(2). https://doi.org/10.3835/plantgenome2016.11.0111

Thomson, M. J. (2014). High-Throughput SNP Genotyping to Accelerate Crop Improvement. *Plant Breeding and Biotechnology*, *2*, 195–212.

USDA ERS. (2022). *USDA Economic Reseasrch Service: Wheat Data*. https://www.ers.usda.gov/data-products/wheat-data/

USDA NASS. (2021). *Small Grains 2021 Summary*. https://www.nass.usda.gov/Publications/Todays_Reports/reports/smgr0921.pdf

VanRaden, P. M., Van Tassell, C. P., Wiggans, G. R., Sonstegard, T. S., Schnabel, R. D., Taylor, J. F., & Schenkel, F. S. (2009). Invited review: Reliability of genomic predictions for North American Holstein bulls. In *Journal of Dairy Science* (Vol. 92, Issue 1, pp. 16–24). American Dairy Science Association. https://doi.org/10.3168/jds.2008-1514

Wang, X., Xu, Y., Hu, Z., & Xu, C. (2018). Genomic selection methods for crop

improvement: Current status and prospects. In *Crop Journal* (Vol. 6, Issue 4, pp. 330–340). Crop Science Society of China/ Institute of Crop Sciences. https://doi.org/10.1016/j.cj.2018.03.001

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

William, H. M., Trethowan, R., & Crosby-Galvan, E. M. (2007). Wheat breeding assisted by markers: CIMMYT's experience. *Euphytica*, *157*(3), 307–319.

Würschum, T., Leiser, W. L., Langer, S. M., Tucker, M. R., & Longin, C. F. H. (2018). Phenotypic and genetic analysis of spike and kernel characteristics in wheat reveals long-term genetic trends of grain yield components. *Theoretical and Applied Genetics*, *131*(10), 2071–2084. https://doi.org/10.1007/s00122-018-3133-3

Yang, L., Zhao, D., Meng, Z., Xu, K., Yan, J., Xia, X., Cao, S., Tian, Y., He, Z., & Zhang, Y. (2020). QTL mapping for grain yield-related traits in bread wheat via SNP-based selective genotyping. *Theoretical and Applied Genetics*, *133*(3), 857–872. https://doi.org/10.1007/s00122-019-03511-0

Zhang, J., Gizaw, S. A., Bossolini, E., Hegarty, J., Howell, T., Carter, A. H., Akhunov, E., & Dubcovsky, J. (2018). Identification and validation of QTL for grain yield and plant water status under contrasting water treatments in fall-sown spring wheats. *Theoretical and Applied Genetics*, *131*(8), 1741–1759. https://doi.org/10.1007/s00122-018-3111-9

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

CHAPTER 3

Whole genome analysis of hard winter wheat germplasm identifies genomic regions associated with spike and kernel traits

This chapter has been published in the *Theoretical and Applied Genetics* journal.

Citation: Gill, H. S., Halder, J., Zhang, J., Rana, A., Kleinjan, J., Amand, P. S., ... & Sehgal, S. K. (2022). Whole-genome analysis of hard winter wheat germplasm identifies genomic regions associated with spike and kernel traits. *Theoretical and Applied Genetics*, *135*(9), 2953-2967.

3.1 Abstract

Genetic dissection of yield-component traits including spike and kernel characteristics is essential for the continuous improvement of wheat yield. Genome-wide association studies (GWAS) have been frequently used to identify genetic determinants for spike and kernel-related traits in wheat, though none have been employed in hard winter wheat (HWW) which represents a major class in U.S. wheat acreage. Further, most studies relied on assembled diversity panels instead of adapted breeding lines, limiting the transferability of results to practical breeding. Here we assembled a population of advanced/elite breeding lines and well-adapted cultivars and evaluated over four environments for phenotypic analysis of spike and kernel traits. GWAS identified 17 significant and multi-environment marker-trait associations (MTAs) for various traits, representing 12 putative quantitative trait loci (QTLs), with five QTLs affecting multiple traits. Four of these QTLs mapped on three chromosomes 1A, 5B, and 7A for spike length, number of spikelets per spike (NSPS), and kernel length are likely novel. Further, a highly significant QTL was detected on chromosome 7AS that has not been previously associated with NSPS and putative candidate genes were identified in this region. The allelic frequencies of important quantitative trait nucleotides (QTNs) were deduced in a larger set of 1,124 accessions which revealed the importance of identified MTAs in the U.S. HWW breeding programs. The results from this study could be directly used by breeders to select the lines with favorable alleles for making crosses and reported markers will facilitate marker-assisted selection of stable QTLs for yield components in wheat breeding.

3.2 Introduction

Wheat (*Triticum aestivum* L.) provides an adequate and affordable intake of calories and proteins in human diets and plays a critical role in food security (FAO, 2017; Grote et al., 2021). Global wheat production needs to be increased by 60% to meet the future demand of feeding 9 billion people by 2050 (Fischer et al., 2014); however, a gradual decrease in arable land and climate change is predicted to make this increase more challenging for the breeders (Grote et al., 2021; Wheeler & Von Braun, 2013). Thus, continued research efforts to understand the genetic basis of grain yield and the development of more productive wheat varieties remain the primary focus for all wheat breeding programs.

Wheat grain yield is a complex quantitative trait involving many quantitative trait loci (QTLs) with small effects and is highly influenced by environmental factors, which makes it difficult to improve yield by direct phenotypic selection (K. Liu et al., 2018). Understanding the genetic basis of grain yield is further hampered by the low heritability of this trait (Kuzay et al., 2019). Nevertheless, grain yield is mainly determined by three component traits including spike number per unit area, kernel number per spike, and thousand kernel weight (TKW) (J. Liu et al., 2018). Final kernel number per spike is affected by spike-related traits such as spike length (SL), number of spikelets per spike (NSPS), and spikelet density (SD) (Guo et al., 2017). Similarly, TKW is influenced by several kernel morphology traits such as kernel length (KL), kernel width (KW), and kernel area (KA) (Gegas et al., 2010). Most of these yield component traits are controlled by several QTLs; however, these traits are less sensitive to the environment and exhibit higher heritability compared to grain yield per se (Zhang et al., 2018). Therefore, a promising strategy to improve grain yield in wheat is to understand the genes and gene

networks controlling individual yield component traits and exploiting them for improving the yield potential (Kuzay et al., 2019; Würschum et al., 2018).

In past decades, QTL mapping approaches including linkage mapping and genome-wide association study (GWAS) have been extensively used to identify QTLs governing these yield-related traits. These studies reported QTLs on different wheat chromosomes for spike and kernel traits such as SL (Alqudah et al., 2020; Gao et al., 2015; J. Liu et al., 2018; Wu et al., 2012; Würschum et al., 2018; M. Yu et al., 2014), NSPS (Alqudah et al., 2020; Kuzay et al., 2019; J. Liu et al., 2018; Muqaddasi et al., 2019; Wu et al., 2012; Zhai et al., 2016; Zhou et al., 2017), SD (Faris et al., 2014; Sourdille et al., 2000; Wu et al., 2012; Würschum et al., 2018; Zhou et al., 2017), kernel morphology (G. Chen et al., 2016; Z. Chen et al., 2020; H. Liu et al., 2020; K. Liu et al., 2018; Würschum et al., 2018), and TKW (Alqudah et al., 2020; Börner et al., 2002; Z. Chen et al., 2020; Dhakal et al., 2021; H. Liu et al., 2020; Pang et al., 2020; Ward et al., 2019; Zanke et al., 2015). Further, these approaches uncovered several genes or major-effect QTLs affecting grain weight or spike morphology, such as *TaGW2* homeologous genes for grain weight (Z. Su et al., 2011), *TaTGW6* for grain size and weight (M. J. Hu et al., 2016), *TaSus1* and *TaSus2* affecting TKW (Hou et al., 2014), and *TaAPO1* for spike morphology (Kuzay et al., 2019; Muqaddasi et al., 2019). Though majority of these studies employed linkage mapping, GWAS has also been successfully used to dissect various agronomic traits in wheat in recent years (Alqudah et al., 2020; Ward et al., 2019; Würschum et al., 2018). Furthermore, the development of more powerful methods like FarmCPU and BLINK has increased the ability of GWAS to detect loci of smaller effects, making it useful to dissect yield-related traits (Huang et al., 2019).

The exploitation of GWAS to characterize yield component traits in winter wheat had been relatively limited (Ward et al., 2019; Zanke et al., 2015; Zhai et al., 2016). Moreover, GWAS have not been reported to date on spike and kernel traits in hard winter wheat (HWW) which is the major class grown in the U.S. Great Plains region and accounted for ~58% of U.S. wheat acreage in 2020 (USDA, 2021). This necessitates the need to explore the phenotypic and genetic variation for yield-related traits in the important class of wheat to provide a useful resource to the breeders. Secondly, most of the GWAS studies in different crop species make use of assembled diversity panels or landraces (Halder et al., 2019; Sidhu et al., 2020), rather than using the breeding materials of a program (Ward et al., 2019). The use of such panels of elite breeding lines is advantageous for identifying novel genomic regions underlying the trait(s) of interest. Several studies used a set of elite breeding lines to perform GWAS for different traits (Begum et al., 2015; Sukumaran et al., 2014; Ward et al., 2019), which provides an opportunity to harness the historical recombination along with the recombination events arising after crosses in the breeding programs. Another advantage of using advanced breeding lines is its relevance to the process of cultivar development (Ward et al., 2019) as the use of such a panel allows direct transfer of the identified QTLs to new germplasm and cultivars without linkage drag in the breeding programs through marker-assisted selection.

The present study aimed to characterize the genetic basis of spike and kernel traits in U.S. hard winter wheat. We assembled a population of advanced and elite lines from the South Dakota State University (SDSU) breeding program representing most of the diversity of the program. To further improve the resolution, a set of released cultivars and

breeding lines from different breeding programs in the U.S. Great Plains region which were frequently used as parents in hybridizations were added to this panel. This panel was phenotyped for seven spike and kernel-related traits under four field environments over two years and genotyped via genotyping-by-sequencing (GBS) approach. The objectives of the current study were to (a) assess the phenotypic and genetic variation for the spike and kernel traits, (b) dissect the genetic architecture of these traits using GWAS, and (c) identify putative candidate genes responsible for the traits of interest using the wheat reference genome, and (d) study the distribution of the identified quantitative trait nucleotides (QTNs) in the SDSU winter wheat breeding program.

3.3 Materials and Methods

3.3.1 Plant material and field experiments

A panel of 314 hard winter wheat breeding lines and released cultivars was assembled based on their pedigree to represent most of the genetic diversity in the SDSU breeding program. Majority of the panel (referred to as SD-Panel) included the hard winter wheat breeding lines developed and evaluated in advanced yield trials (AYT) and elite yield trials (EYT) over the past decade at SDSU winter wheat breeding program. The panel comprised 243 SDSU breeding lines including 16 doubled haploids (DHs), 40 widely adapted and grown hard winter wheat cultivars, and 31 elite breeding lines from regional performance nurseries developed by other state breeding programs in the U.S. Great Plains including Colorado, Kansas, Nebraska, Oklahoma, and Texas. The selected hard winter wheat cultivars have been frequently used as parents in the regional breeding programs.

The panel was phenotyped in four field environments with three trials conducted in the 2019-20 winter wheat growing season (referred to as E1, E2, and E3 from now on) and one trial conducted in the 2020-21 growing season (referred to as E4). In the 2019-20 growing season, the panel was planted at three SDSU experimental stations at Aurora, Brookings, and Volga, South Dakota, respectively, while the 2020-21 trial was conducted at the SDSU experimental station in Brookings, South Dakota. In each environment, trials were planted using a randomized complete block design (RCBD) with two replications. Each experimental unit consisted of a 1.25-m-long row plot with an inter-row spacing of 20 cm. The experiments were managed using the regional standard cultural practices for proper growth and development of wheat plants.

3.3.2 Phenotypic evaluations and statistical analysis

The spike-related traits were evaluated by measuring 10 representative spikes from each accession per replication at physiological maturity. Spike length (SL) was measured from the base of the first spikelet to the apex of the last spikelet excluding awns using a ruler (in cm). Spikelet number per spike (NSPS) was counted and subsequently averaged across ten spikes. The spikelet density (SD), referring to the spikelet number per unit of spike length, was estimated by dividing NSPS by the spike length. In addition, whole rows were manually harvested, threshed, and cleaned to obtain a seed sample for measuring KL (in mm), kernel width (KW, in mm), and kernel area (KA, in mm^2) using an automatic grain analyzer Vibe QM3 (Vibe Imaging Analytics, CA, USA). Subsequently, TKW was estimated by counting 1,000 kernels using a Contrad 2 Seed Counter (Hoffman Manufacturing Inc, OR, USA) and weighing the counted kernels, with three repetitions. In each environment, the lines with missing replications or lines

subjected to any damage or contamination during harvesting were removed from the phenotypic analyses.

Phenotypic data were analyzed as described previously in Gill et al. (2021). Briefly, we estimated the best linear unbiased estimates (BLUEs) for all the traits within each environment as well as across the four environments. The BLUEs and variance components were estimated using META-R (Alvarado et al., 2020), which employs LME4 R-package (Bates et al., 2015) for linear mixed model analysis. The broad-sense heritability (H^2) of a trait of interest in a combined environment analysis was assessed based on the variance estimates from the linear mixed model as follows:

$$H^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_{ge}^2/nLoc + \sigma_e^2/(nLoc \times nRep)}$$

where σ_g^2 and σ_e^2 , are the genotype and error variance components, σ_{ge}^2 is the G × E interaction variance component and *nLoc* is the number of environments and *nRep* refers to the replications within an environment, respectively. The estimated BLUEs were used to visualize the correlation matrices between the studied traits. The correlation and network plots were obtained using R packages 'psych' (William, 2013) and 'qgraph' (Epskamp et al., 2012). The summary statistics and pairwise comparisons were performed in the R (R Core Team 2014).

3.3.3 Genotyping analysis

The SD-Panel was genotyped using the GBS approach at the USDA Central Small Grain Genotyping Lab, Manhattan, KS. The lines were grown in small pots and fresh leaf tissues were collected from each line for DNA isolation using the hexadecyltrimethylammonium bromide (CTAB) method (Doyle & Doyle, 1987).

Genotyping-by-sequencing (GBS) libraries were prepared by double restriction digestion with HF-*PstI* and *MspI* enzymes (Poland et al., 2012) and sequenced using an Ion Proton sequencer (Thermo Fisher Scientific, Waltham, MA, USA). The GBS discovery pipeline v2.0 in TASSEL v5.0 (Trait Analysis by aSSociation, Evolution and Linkage) was used to call single-nucleotide polymorphisms (SNPs) (Bradbury et al., 2007). The GBS reads were aligned to the Chinese Spring reference genome RefSeq v2.0 (IWGSC, 2018) using the default settings of Burrows-Wheeler Aligner v0.6.1. The SNP data for the 314 lines was extracted from the pooled data of 1,124 lines and subjected to quality control in TASSEL v5.0. For quality control, SNPs with more than 25% missing data points and more than 15% heterozygosity, and the SNPs that were unmapped on any wheat chromosome were removed. The missing data points in the selected SNP set were imputed using BEAGLE v4.1 (beagle.27Jan18.7e1.jar; https://faculty.washington.edu/browning/beagle/b4_1.html) (Browning & Browning, 2007). The imputed set was filtered to remove SNPs with minor allele frequency (MAF) of less than 0.05. The nomenclature of SNPs was based on the chromosome and the physical position on IWGSC RefSeq v2.0 (IWGSC, 2018; Zhu et al., 2021), such as *S1A_1230000* indicates SNP on chromosome 1A mapped at 1.23 Mbp.

3.3.4 Population structure and linkage disequilibrium

To examine the population structure, principal component analysis (PCA) of the filtered and imputed genotypic data was conducted using Genomic Association and Prediction Integrated Tool (GAPIT) v3.0 (J. Wang & Zhang, 2021) in R (R Core Team 2014). We also assessed the population stratification using a Bayesian model-based clustering program, STRUCTURE v2.3.4 assuming an Admixture model (Pritchard et al., 2000).

We used ten subgroups (*K =1-10*) with ten independent runs for each subgroup using a burn-in period of 10,000 iterations followed by 10,000 Monte-Carlo iterations. An ad-hoc statistic (DeltaK) was used to infer the most likely number of subpopulations, which uses the rate of change in the log probability between runs using successive K-values (Evanno et al., 2005) using STRUCTURE HARVESTER (Earl & vonHoldt, 2012). Linkage disequilibrium (LD) analysis was performed for the whole genome as well as each sub-genome by computing r^2 values for all pairwise marker comparisons using a sliding window size of 50 markers in TASSEL v5.0. LD decay over genetic distance was estimated by fitting a non-linear model using the modified Hill and Weir method (Hill & Weir, 1988) with the r^2 threshold set at 0.2 and r^2 equals half decay distance. The LD decay distance for the whole genome and each sub-genomes was plotted using R (R Core Team 2014).

3.3.5 Association mapping and candidate gene analysis

We performed GWAS using the 8,030 high-quality SNPs and BLUEs for seven traits from four individual environments (E1, E2, E3, and E4) as well as BLUEs from the combined analysis across all environments (CEnv). Two methods, single-locus mixed linear model and multi-locus mixed model, were compared to select the appropriate algorithm for GWAS on each trait. The single-locus method used the mixed linear model (MLM) by considering the kinship and population structure to adjust for population stratification (K-PC model) (J. Yu et al., 2006), whereas the multi-locus mixed model used the Bayesian-information and Linkage-disequilibrium Iteratively Nested Keyway (BLINK) method. The BLINK model was developed more recently and reported to perform better than the popular multi-locus FarmCPU approach (Huang et al., 2019; X.

Liu et al., 2016) because it overcomes the limitations of the FarmCPU and identifies more true positives and produces fewer false positives when compared to other GWAS methods (Huang et al., 2019). Both the models were implemented through Genomic Association and Prediction Integrated Tool (GAPIT) version 3.0 in the R environment (J. Wang & Zhang, 2021), and included the first two principal components to account for the population structure, based upon visual examination of the scree plot and DeltaK statistic from STRUCTURE analysis.

The two approaches were compared based on quantile-quantile (QQ) plots and the power of the models to detect known loci. The BLINK model performed better than the MLM model for all trait-environment combinations and was used to report the GWAS results. The Bonferroni-corrected threshold of $P < 0.1$ was estimated as $-log_{10}(P) = 4.90$ to declare any association as significant, however, this threshold proves too stringent as it accounts for all the SNPs in the dataset rather than independent tests. Thus, most studies use an exploratory threshold or a corrected Bonferroni threshold based on independent tests (D. Kumar et al., 2021; Pang et al., 2020). In this study, we used an exploratory threshold of $-log_{10}(P) = 4.00$ to declare any MTA as significant in individual environments. Nevertheless, only those MTAs were reported as stable MTAs, which surpassed this threshold and were detected for the same trait over two years/growing seasons (not in two locations in one season) or for more than one trait. The proportion of the phenotypic variance explained by stable MTAs was deduced using the 'random.model' attribute in GAPIT v3.0.

The stable MTAs were subjected to a pairwise comparison of different alleles on the respective traits. For each stable MTA, mean trait values for two groups of alleles (favorable v/s unfavorable) were compared using a *t-test* and visualized using boxplots with R package 'ggplot2' (Wickham, 2016). The allelic frequencies of significant and stable MTAs in the panel were analyzed to compare the effect of stacking favorable alleles for SL and NSPS. The alleles increasing SL or NSPS were defined as favorable alleles. The whole panel was grouped by accessions carrying the favorable alleles for each trait. These groups were compared using an FDR-adjusted *pairwise t-test* to verify the additive effect of the favorable alleles on SL and NSPS. Furthermore, the highly significant MTAs for selected traits were used for haplotype analysis. The LD-based haplotypes for selected regions were generated and visualized using Haploview (Barrett et al., 2005). The accessions from the panel were grouped based on identified haplotypes and trait means for each haplotype were compared using analysis of variance (ANOVA) in the R package 'agricolae' (Mendiburu Felipe de, 2021). We also performed the candidate gene analysis for a highly significant and stable region on wheat chromosome 7AS. The high-confidence genes within the haplotype block of respective MTA were extracted from the IWGSC reference genome and identified using IWGSC v2.1 RefSeq annotation (IWGSC, 2018; Zhu et al., 2021). Gene expression browser (http://www.wheat-expression.com/) was used to exclude the unlikely candidates and the remaining genes were annotated manually using Blast2GO (Conesa et al., 2005) for the identification of likely candidates.

3.3.6 Allelic frequencies of important QTNs in breeding material

To investigate the allelic frequencies of several QTNs for SL, NSPS, and TKW in the breeding materials used in the U.S. Great Plains, an additional set of 810 winter wheat accessions including advanced breeding lines of SDSU and other breeding programs, and released cultivars from the U.S. Great Plains region were analyzed for the selected markers. We also included the 314 accessions of the SD-Panel for this analysis, making a large set of 1,124 accessions in total. Out of 1,124 accessions, 204 genotypes that were evaluated in advanced trials or released as cultivars were categorized as 'elite' genotypes. The GBS data for the set of 810 accessions was available (as described in earlier sections) and genotypic information for the required quantitative trait nucleotides (QTNs) was extracted for further analysis. Based on this data, the allelic frequencies for selected QTNs in the complete set (1,124 accessions) as well as in 204 'elite' genotypes were estimated and visualized using the R package 'ggplot2' (Wickham, 2016).

3.4 Results

3.4.1 Variation for spike and kernel traits

A significant variation (Table 3.1) was observed for all the spike and kernel traits (SL, NSPS, SD, KL, KW, KA, and TKW) and the phenotypic distribution of studied traits was found to be consistent in all four environments (Figure 3.1). Broad-sense heritability (H^2) was high for most of the studied traits, ranging from 0.54 to 0.94 (Table 3.1) with the highest for SL and NSPS (0.94) and the lowest for spikelet density (H^2 = 0.54). For the kernel traits, KL had much higher heritability (0.91) than KW (0.76).

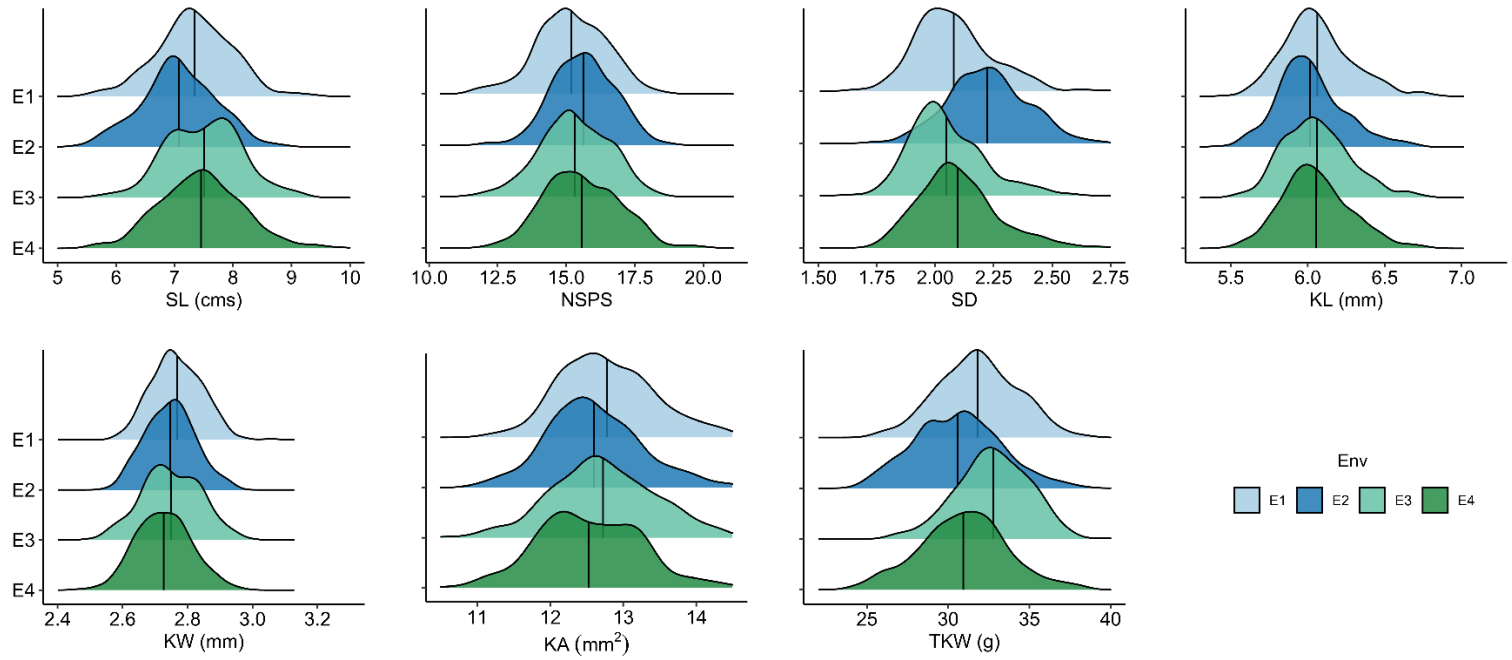

Figure 3.1 Phenotypic distribution of the investigated spike and kernel traits in a panel of 314 genotypes evaluated in four different environments (E1, E2, E3, and E4). SL, spike length; SPS, spikelet number per spike; SD, spikelet density; TKW, thousand kernel weight; KL, kernel length; KW, kernel width; KA, kernel area. The vertical black lines represent the mean trait value in respective environments.

Table 3.1 Descriptive statistics for spike and kernel traits and broad-sense heritability estimates obtained using a combined analysis of four environments.

Trait[a]	Unit(s)	Mean	Min	Max	LSD	CV	Heritability
NSPS	Count	15.42	12.25	18.58	0.80	4.50	0.94
SL	Centimeters (cm)	7.35	5.76	9.10	0.43	5.40	0.94
SD	Ratio	2.07	1.26	2.56	0.36	12.57	0.54
KW	Millimeters (mm)	2.75	2.58	3.20	0.10	3.15	0.76
KL	Millimeters (mm)	6.04	5.47	6.89	0.18	2.57	0.91
KA	Square millimeter (mm^2)	12.66	11.16	14.25	0.56	1.62	0.88
TKW	Grams (g)	31.54	25.86	36.56	2.28	3.45	0.84

[a]NSPS, number of spikelets per spike; SL, spike length; SD, spikelet density; KW, kernel width; KL, kernel length; KA, kernel area; TKW, thousand kernel weight
CV, coefficient of variation; LSD, least significant difference; Min, minimum; Max, maximum

3.4.2 Relationship between traits

Pearson's correlation coefficients estimated using the phenotypic BLUEs obtained from the combined environment analysis were significant among several pairs of traits (Figure 3.2a). The strongest positive correlation ($r = 0.81$) was observed between the TKW and KA, whereas the strongest negative correlation ($r = -0.31$) was observed between SL and SD. TKW also showed a significant positive correlation with KW ($r = 0.70$) and KL ($r = 0.47$). Contrarily, TKW showed a negative correlation with spike traits including SD ($r = -0.11$) and NSPS ($r = -0.10$). Further, a correlation-based network analysis was performed to visualize a pattern of association among spike and kernel traits. The network analysis revealed a moderate to strong association within the kernel traits (Figure 3.2b). In addition to a strong association between KL, KW, and KA, the kernel traits were

relatively more associated with TKW as compared to spike traits, with the highest positive association between KA and TKW. Among the spike traits, NSPS and SD were negatively associated with all the kernel traits (KL, KW, and KA) as well as TKW. To validate these relationships in individual environments, correlation coefficients were calculated among seven traits in individual environments. Overall, we observed a consistent relationship among different trait pairs in all individual environments (Figure 3.3).

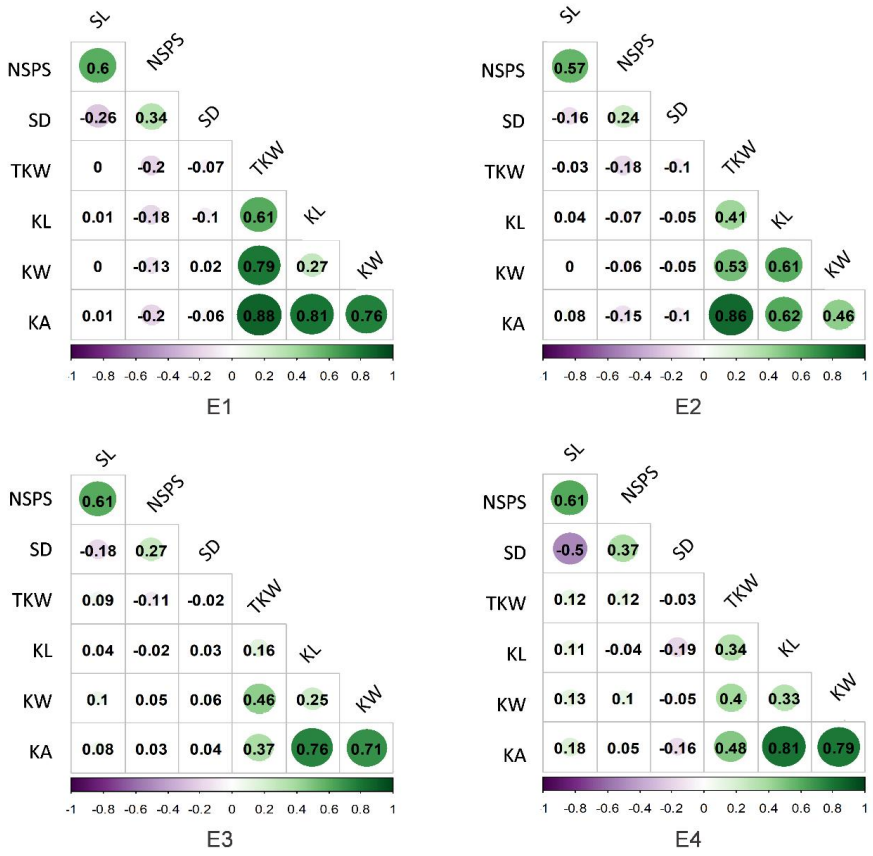

Figure 3.3 Correlation coefficients among various spike and kernel traits in individual environments. SL, spike length; NSPS, number of spikelets per spike; SD, spikelet density; TKW, thousand kernel weight; KL, kernel length; KW, kernel width; KA, kernel area.

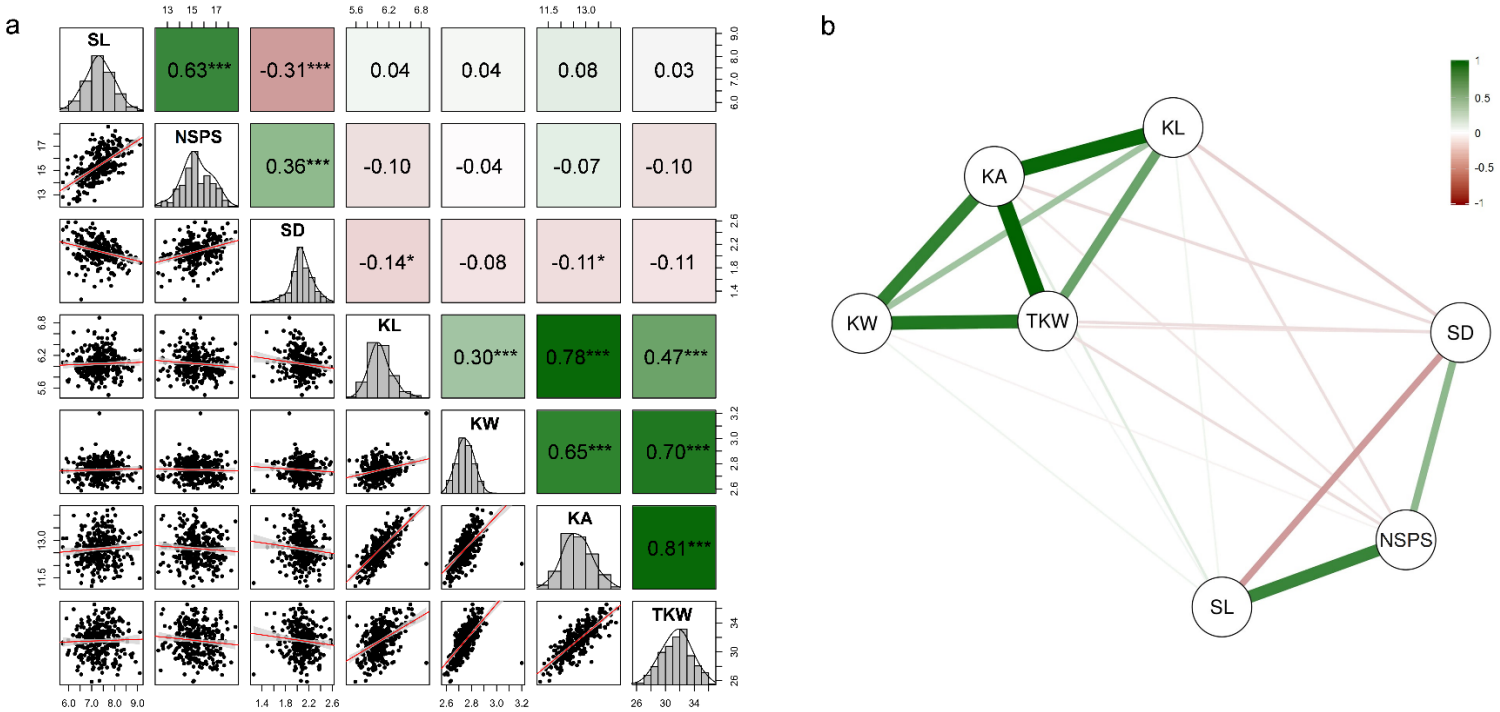

Figure 3.2 (a) Correlation coefficients among investigated spike and kernel traits calculated by using the best linear unbiased estimates (BLUEs) from combined analysis of four environments. (b) Correlation-based network analysis plot depicting the association between studied traits. SL, spike length; NSPS, number of spikelets per spike; SD, spikelet density; KW, kernel width; KL, kernel length; KA, kernel area; and TKW, thousand kernel weight. Statistically significant correlations are denoted by an asterisk (*) where * $P \leq 0.05$, ** $P \leq 0.01$, and *** $P \leq 0.001$.

3.4.3 Genotypic analysis, population structure, and LD

Among 8,030 high-quality GBS SNPs, the numbers of SNPs from the A and B sub-genomes were more than double that from the D sub-genome, with the highest number in the B sub-genome (3,627; 45.2%) and the lowest in the D sub-genome (1,328; 16.5%) (Table 3.2). Chromosome 3B had the most SNP markers (649 SNPs), while chromosome 4D had the lowest number of markers (34 SNPs). The average LD decay distance for the whole genome was approximately 3.6 Mbp (Figure 3.4a). The LD decay for the three sub-genomes A, B, and D revealed different patterns for individual genomes (Figures 3.4a and 3.4b). For instance, sub-genomes A and B showed smaller LD decay distance than sub-genome D. Principal component analysis showed substantial admixture among genotypes, with the first two principal components explaining around only 6.5% and 3.4% of the total variance, respectively (Figure 3.5). The DeltaK statistic from STRUCTURE analysis showed a single peak at K = 2, suggesting only two sub-groups in the panel (Figure 3.4c).

Table 3.2 The distribution of 8,030 SNPs across 21 wheat chromosomes in the panel of 314 accessions.

Sub-genome	Chromosome	Number of SNPs	% SNPs
A	1	458	
	2	380	
	3	454	
	4	286	
	5	410	
	6	385	
	7	702	
Subtotal A		3,075	38.3
B	1	554	
	2	593	
	3	649	
	4	163	
	5	509	
	6	610	
	7	549	
Subtotal B		3,627	45.2
D	1	191	
	2	327	
	3	257	
	4	34	
	5	153	
	6	152	
	7	214	
Subtotal D		1,328	16.5
Total (A, B, and D)		8,030	100

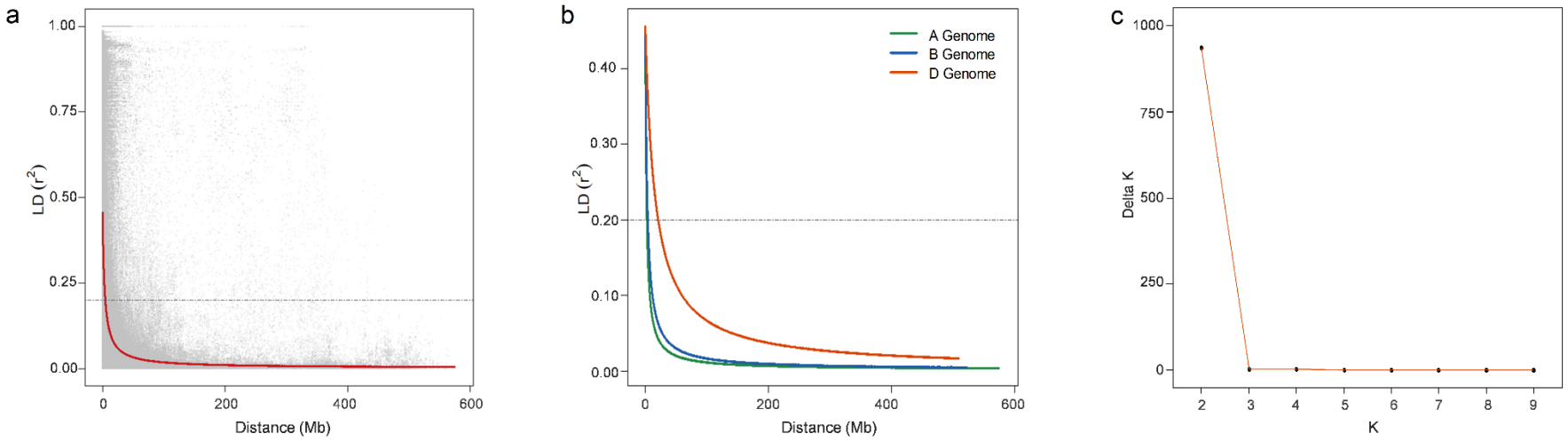

Figure 3.4 Intra-chromosomal linkage disequilibrium (LD) in the SD-Panel for (a) for the whole genome, and (b) for A, B, and D sub-genomes. (c) Evanno plot of Delta-K statistic from the STRUCTURE analysis.

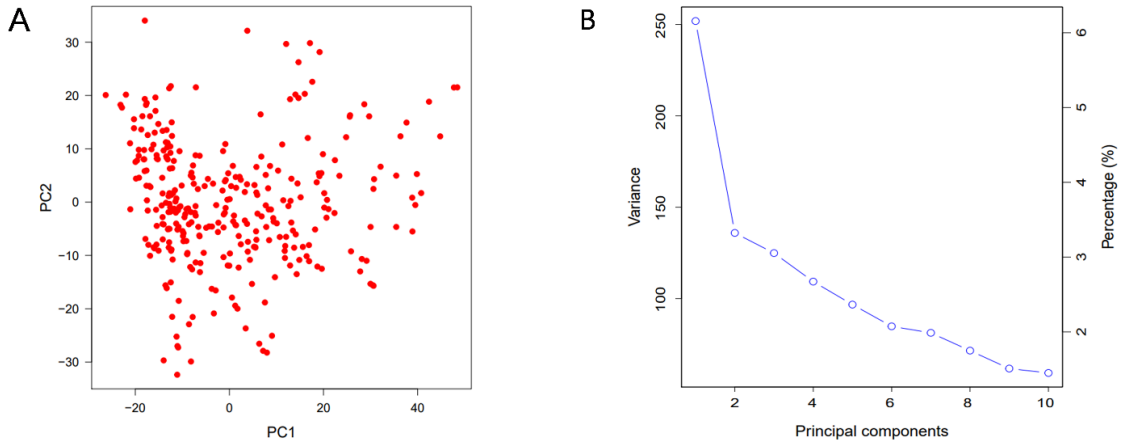

Figure 3.5 Principal component analysis (PCA) of 314 wheat accessions using 8,030 SNPs (A) PCA scatterplot showing the first two principal components, and (B) The scree plot was generated to illustrate the changes in each principal component.

3.4.4 Marker trait associations

GWAS using BLUEs for each trait obtained from analyses of phenotypic data from four individual environments (namely E1, E2, E3, E4) and for environments combined (CEnv) identified 69 significant MTAs for six traits except KA, based on the exploratory threshold of $-log_{10}(P) = 4.0$ (Appendix 3.1). Out of the 69 MTAs, 49 unique SNPs were associated with the six different traits (Appendix 3.1, Figure 3.6). The identified MTAs were distributed on 18 wheat chromosomes, except chromosomes 4D, 6A, and 6D. Among the 69 MTAs, the highest number of MTAs were detected for NSPS (18), followed by SL (13), and SD had the lowest MTAs (6).

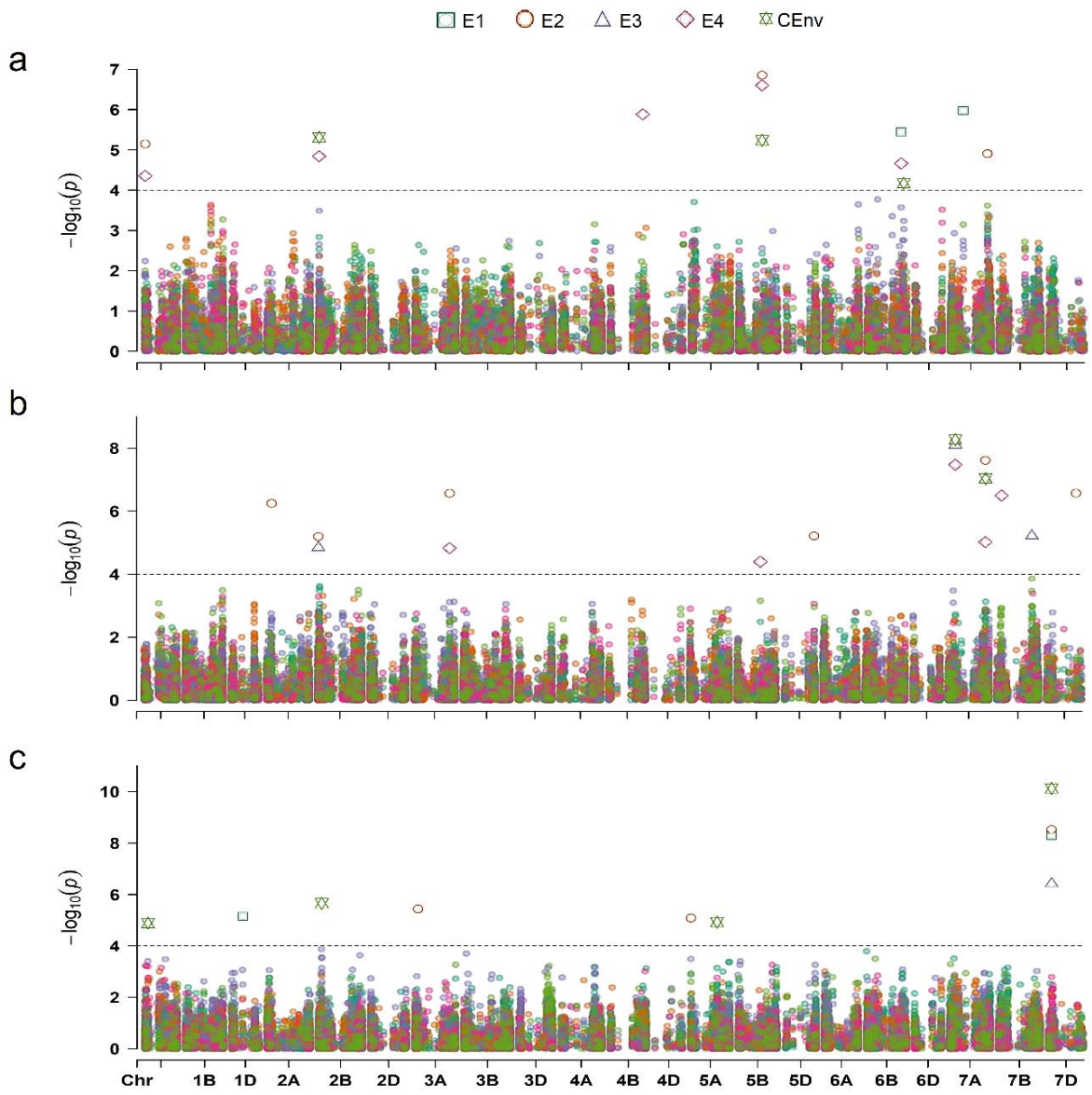

Figure 3.6 Manhattan plot summarizing the significant MTAs reported for (a) spike length, SL (b) number of spikelets per spike, NSPS and (c) thousand kernel weight, TKW in four individual environments (E1 – E4) and combined analysis (CEnv).

Table 3.3 Details of stable significant marker-trait associations (MTAs) identified by genome-wide association studies (GWAS) for spike and kernel traits.

Trait[a]	SNP[b]	Chr	Pos[c]	Allele	-log_{10}(*P*-value)[d]	Env[e]	Other trait(s)[f]
SL	S1A_13099591	1A	13,099,591	T/G	4.36 - 5.14	E2, E4	-
	S2B_16305395	2B	16,305,395	G/A	4.84 - 5.31	E4, CEnv	NSPS
	S5B_432612793	5B	432,612,793	C/G	5.24 - 6.86	E2, E4, CEnv	NSPS
	S6B_619882604	6B	619,882,604	C/G	4.66 - 5.45	E1, E4	-
	S7A_676732614	7A	676,732,614	A/G	4.91	E2	NSPS
NSPS	S2B_16305395	2B	16,305,395	G/A	4.84 - 5.20	E2, E3	SL
	S3A_647983369	3A	647,983,369	T/C	4.43 - 6.57	E2, E4	-
	S5B_432612793	5B	432,612,793	C/G	4.39	E4	SL
	S7A_132414615	7A	132,414,615	C/A	7.48 - 8.27	E2, E3, E4, CEnv	-
	S7A_676732614	7A	676,732,614	A/G	5.02 - 7.61	E2, E3, E4, CEnv	SL
TKW	S5A_476847493[g]	5A	476,847,493	C/T	4.94	CEnv	KL
	S7D_60662020	7D	60,662,020	T/G	6.42 - 10.12	E1, E2, E3, CEnv	KW
KL	S1A_299864277	1A	299,864,277	A/G	4.56 - 5.54	E3, E4	-
	S5A_476898590[g]	5A	476,898,590	A/C	4.54	E3	TKW
	S7A_717859384	7A	717,859,384	A/G	5.39 - 7.35	E1, E4	-
KW	S4A_619197841	4A	619,197,841	A/G	7.29 - 7.50	E1, E3	-
	S7D_60662020	7D	60,662,020	T/G	5.92	E1	TKW

[a]SL, spike length; NSPS, number of spikelets per spike; TKW, thousand kernel weight; KL, kernel length; KW, kernel width
[b]SNP, single nucleotide polymorphism with the peak threshold value
[c]Physical position is based on IWGSC RefSeq v2.0 (IWGSC, 2018)
[d]The range for threshold depicts the minimum to maximum -log_{10}(*P*) values obtained by GWAS in different environments
[e]The environment(s) where the MTA was declared significant based on described threshold
[f] Pleiotropic effect of the respective MTA on other traits(s) if any
[g] Different SNPs representing same genomic region/QTL for KL

Of the 69 MTAs, 17 were considered ‘stable’ MTAs based on their repeatability among environments, with five MTAs each for SL and NSPS, three for KL, and two each for KW and TKW (Table 3.3, Figure 3.7). Several MTAs were associated with more than one trait, and hence the 17 MTAs representing 12 putative QTLs, with five of them affecting multiple traits (Figure 3.7). Five stable MTAs identified for spike length (SL) were distributed on chromosomes 1A, 2B, 5B, 6B, and 7A (Table 3.3). The most significant MTA for SL (*S5B_432612793*; $-log_{10}(P)$ = 5.24 - 6.86) was detected on chromosome 5B, which explained around 14% of the phenotypic variation. For NSPS, five MTAs were located on chromosomes 2B, 3A, 5B, and 7A. Interestingly, two significant MTAs (*S7A_132414615* and *S7A_676732614*) were observed for NSPS on chromosome 7A and both the MTAs were consistent in four individual GWAS analyses (E2, E3, E4, and CEnv). One of these MTAs (*S7A_676732614*) localized on the long arm of chromosome 7A at 676 Mbp, the region harboring a major locus governing NSPS, whereas another locus (*S7A_132414615*) was on the short arm of chromosome 7A at 132 Mbp. The phenotypic variance explained by *S7A_132414615* and *S7A_676732614* was around 5.5% and 7.1%, respectively. Out of the ten MTAs identified for SL or NSPS, three MTAs (*S2B_16305395, S5B_432612793,* and *S7A_676732614*) exhibited a pleiotropic effect on SL and NSPS (Table 3.3), suggesting seven unique QTLs for these traits.

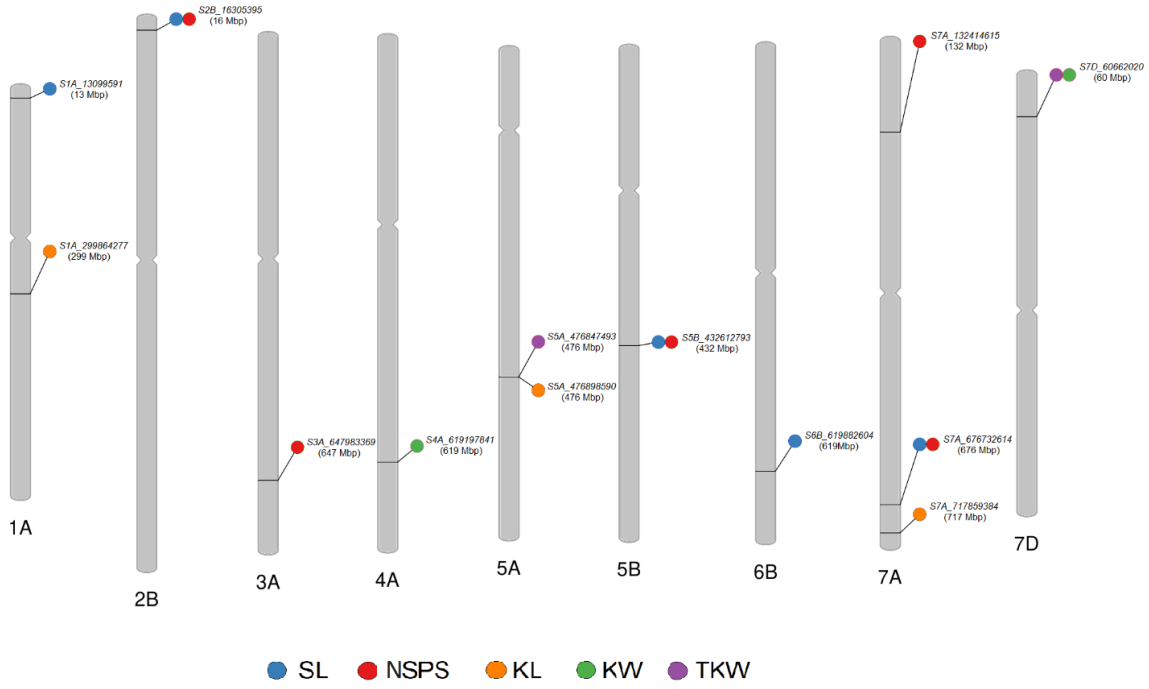

Figure 3.7 A Phenogram representing the distribution of stable MTAs identified on different wheat chromosomes.

Seven stable MTAs were identified for three kernel traits, KL, KW, and TKW (Table 3.3). Three MTAs for KL were mapped on chromosomes 1A, 5A, and 7A. The MTA *S7A_717859384* on chromosome 7A explained about 13% of the phenotypic variation for the KL. Two MTAs (*S5A_476847493* and *S7D_60662020*) for TKW were observed on chromosomes 5A and 7D, and these MTAs had a pleiotropic effect on KL and KW, respectively (Table 3.3). The MTA for TKW on chromosome 7D (*S7D_60662020*) was highly significant in four environments explaining around 7.9% of the phenotypic variation on average.

3.4.5 Evaluation of allelic effects for major QTNs

A pairwise comparison was made to assess the allelic effects of the highly significant QTNs on SL, NSPS, and TKW by comparing the trait means (using BLUEs from combined environments) between the favorable and unfavorable alleles using an FDR-adjusted *pairwise t-test*. The results showed significant differences between the two allelic groups for all three traits (Table 3.4; Figures 3.8a, 3.8b, and 3.8c).

Table 3.4 Pairwise comparison for the effect of two alleles of stable marker-trait associations (MTAs) identified for various traits.

Trait[a]	**SNP**[b]	**Alleles**[c]	**Mean for Allele 1**	**Mean for Allele 2**	***P*-Value**[d]
SL	S1A_13099591	**G**/T	7.73	7.30	0.0269
	S2B_16305395	A/**G**	6.85	7.39	0.0010
	S5B_432612793	C/**G**	7.28	7.92	1.94E-05
	S6B_619882604	**C**/G	7.39	6.92	0.0050
NSPS	S2B_16305395	A/**G**	14.38	15.50	0.0006
	S5B_432612793	C/**G**	15.30	16.34	0.0001
	S7A_132414615	**A**/C	16.07	15.23	8.14E-06
	S7A_676621121	**A**/G	15.53	14.71	4.51E-05
KL	S1A_299864277	A/**G**	6.03	6.22	0.0083
	S5A_476847493	**C**/T	6.07	5.99	0.0029
	S7A_717859384	A/**G**	6.03	6.16	0.0043
KW	S7D_60662020	G/**T**	2.74	2.76	0.0036
TKW	S5A_476847493	**C**/T	31.84	31.06	0.0034
	S7D_60662020	G/**T**	30.82	32.16	2.78E-08

[a]SL, spike length; NSPS, number of spikelets per spike; KL, kernel length; KW, kernel width; TKW, thousand kernel weight
[b]The most significant SNP representing respective QTLs
[c]Allele1/Allele2 notation. The favorable allele has been depicted using bold font
[d]*P*-value from the *t-test*

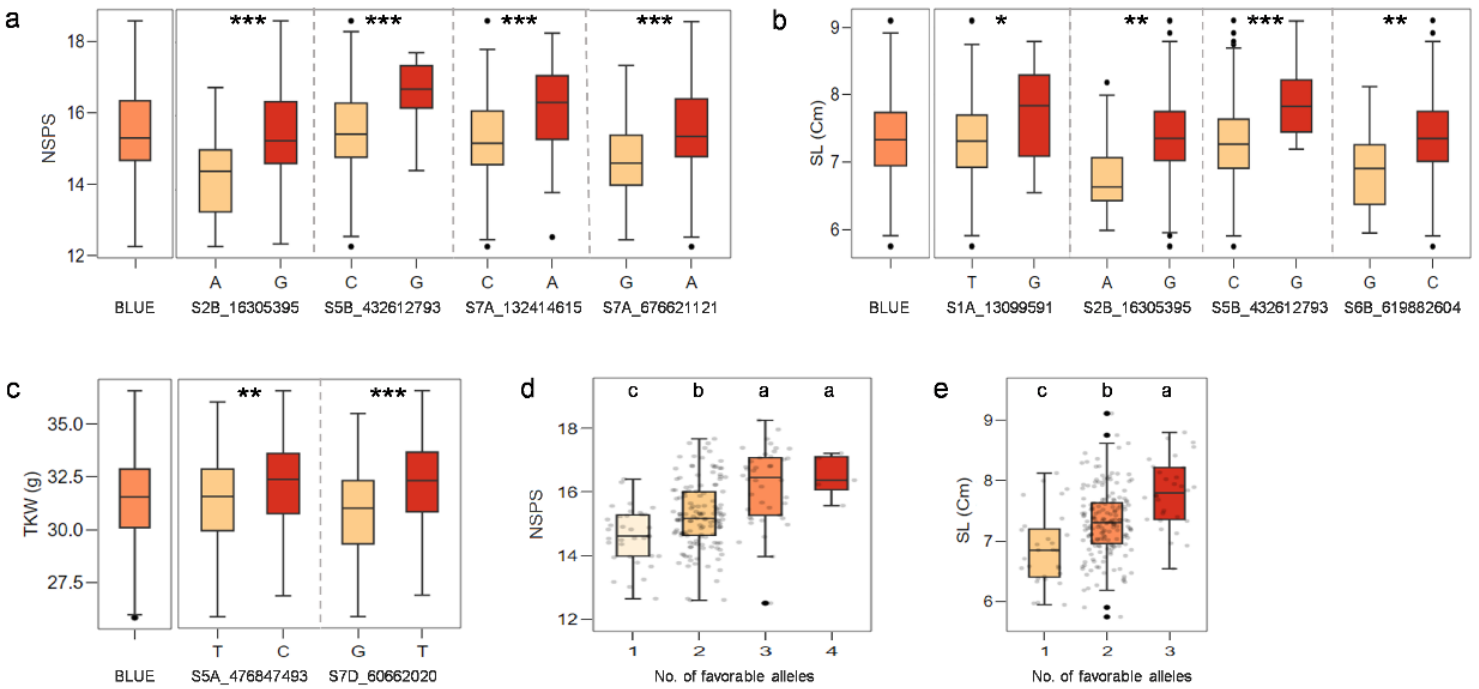

Figure 3.8 Boxplots showing the effect of two alleles (favorable v/s unfavorable) of the stable MTAs (enlisted in Table 2) on the trait means for (a) number of spikelets per spike (NSPS), (b) spike length (SL), and (c) thousand kernel weight (TKW). Trait performance of the lines carrying different numbers of favorable alleles for (d) number of spikelets per spike (NSPS) and (e) spike length (SL), compared using an FDR adjusted Least Significance Difference (LSD) test. Statistically significant differences are denoted by an asterisk (*) where * $P \leq 0.05$, ** $P \leq 0.01$, and *** $P \leq 0.001$.

Interestingly, two MTAs (*S2B_16305395, S5B_432612793*) with a pleiotropic effect on SL and NSPS showed a positive effect on both traits (Figures 3.8a and 3.8b). We further investigated the effect of accumulating favorable alleles on associated phenotypes. The panel of 314 lines was divided into groups based on the number of favorable alleles for SL and NSPS carried by each accession. Four accession groups (each group carrying one to four favorable alleles) were identified for NSPS, while only three groups were detected for SL carrying one to three favorable alleles (Figures 3.8d and 3.8e). NSPS were increased significantly with the increase in the numbers of favorable alleles (P = 4e-10), ranging from 14.6 spikelets/spike for the group with one favorable allele to 16.5 spikelets/spike for the group with four favorable alleles. Similarly, a significant increase in SL was observed with the accumulation of favorable alleles (P = 9.82e-09), indicating additive effects of these MTAs for these traits (Figures 3.8d and 3.8e).

3.4.6 Haplotype and candidate gene analysis for 7AS region associated with NSPS

The NSPS QTL on chromosome 7A associated with SNP *S7A_132414615* showed a high significance and stability across different environments, which appears to be a novel region associated with this trait. Thus, haplotype comparison and candidate gene analysis were performed for this genomic region. The LD block harboring *S7A_132414615* was 2.95 Mbp long and the region contained a total of 15 SNPs (Figure 3.9a). Based on the allelic distribution of these SNPs in the SD-Panel, we identified four major haplotypes (*Hap1*, *Hap2*, *Hap3*, and *Hap4*) with a frequency of 0.76, 0.15, 0.04, and 0.01, respectively (Figure 3.9b). As *Hap4* had very low frequency, the remaining three haplotypes were compared for differences in trait means. The ANOVA revealed

significant differences (P = 1.28e-06) among three haplotypes for NSPS, where *Hap1* had the lower NSPS compared to *Hap2* and *Hap3* (Figure 3.9c).

Candidate gene analysis for the 7AS region by blastN searching of the marker sequences in the 2.95 Mbp LD block against IWGSC RefSeq v2.0 (IWGSC, 2018) identified 41 high confidence (HC) genes. Further analysis using the wheat expression browser (http://www.wheat-expression.com) removed 24 tissue-specifically expressed genes. The remaining 17 HC genes were annotated manually, and several of them showed putative functions of interest (Table 3.5), including the gene *TraesCS7A03G0411400* which encodes a MADS-box transcription factor belonging to the SHORT VEGETATIVE PHASE family. Intriguingly, this gene was present in a region within two SNPs (*S7A_132414615* and *S7A_132532523*) that were significantly associated with NSPS in three different environments and combined analysis, therefore could be a putative candidate gene underlying the QTL for NSPS.

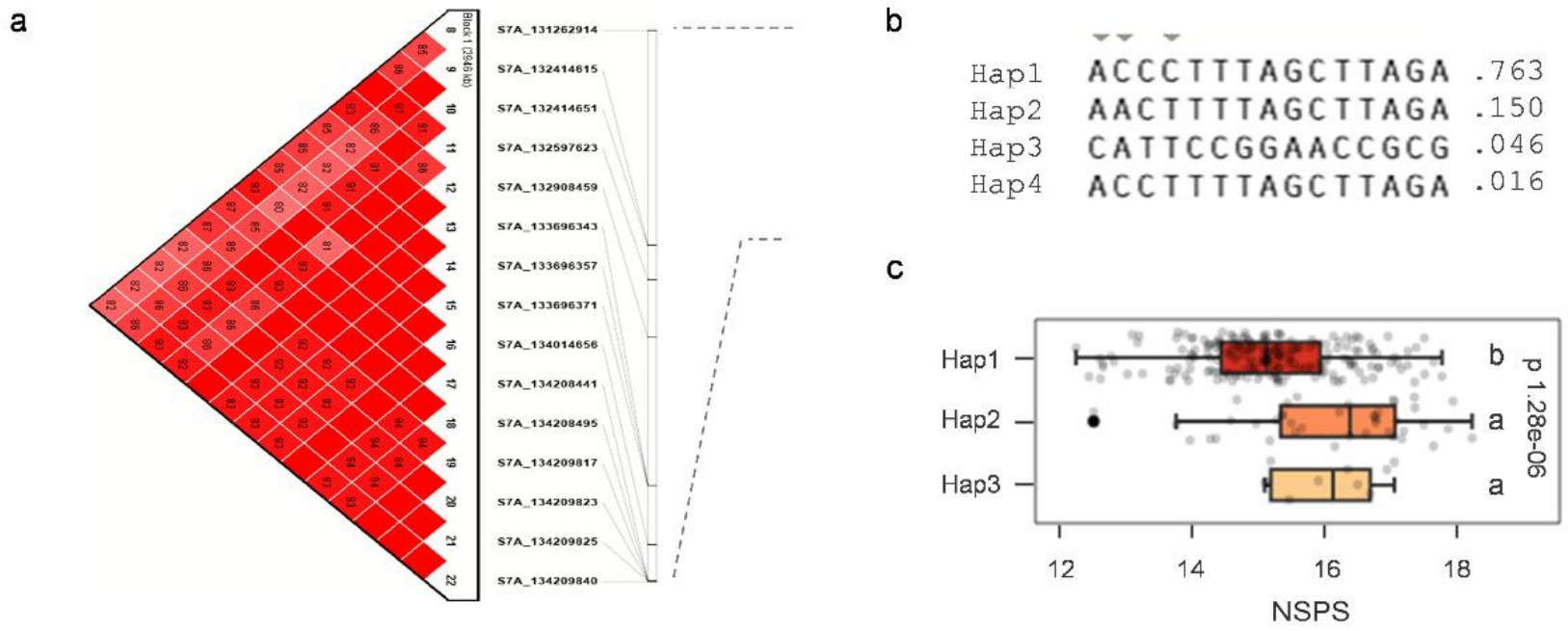

Figure 3.9 (a) Local linkage disequilibrium (LD) block for the 2.9 Mbp region harboring QTL for NSPS on chromosome 7A represented by S7A_132414615. (b) Four allelic haplotypes identified in the SD-Panel based on 15 SNPs present in the LD block along with frequencies for each haplotype, and (c) Differences in NSPS among three major haplotypes using analysis of variance (ANOVA) and an FDR adjusted Least Significance Difference (LSD) test. The fourth haplotype was excluded from ANOVA due to very low frequency in the studies panel.

Table 3.5 List of selected candidate genes with putative functions identified in the genomic region harboring QTL for NSPS (represented by S7A_132414615) on chromosome 7A.

Gene ID[a]	Start[b]	End	Previous ID[c]	Annotation
TraesCS7A03G0408100	131,270,859	131,272,382	*TraesCS7A02G173100*	protein transport protein SEC31-like
TraesCS7A03G0408200	131,278,636	131,282,169	*TraesCS7A02G173200*	zinc finger protein ZOP1
TraesCS7A03G0408300	131,403,952	131,412,808	*TraesCS7A02G173300*	dipeptidyl peptidase family member 6-like
TraesCS7A03G0408500	131,469,863	131,476,058	*TraesCS7A02G173500*	putative aminopeptidase C
TraesCS7A03G0410400	132,037,610	132,038,925	*TraesCS7A02G174500*	12-oxophytodienoate reductase 1-like
TraesCS7A03G0411100	132,372,770	132,387,955	*TraesCS7A02G174900*	kinesin-like protein KIN-14L
TraesCS7A03G0411200	132,405,506	132,407,972	*TraesCS7A02G175000*	putative laccase-9
TraesCS7A03G0411300	132,412,108	132,415,806	*TraesCS7A02G175100*	mediator of RNA polymerase II transcription subunit 6-like
TraesCS7A03G0411400	132,457,430	132,464,217	*TraesCS7A02G175200*	MIKC-type MADS-box transcription factor VRT-A2
TraesCS7A03G0411700	132,525,848	132,531,969	*TraesCS7A02G175300*	peptidyl-prolyl cis-trans isomerase CYP40-like
TraesCS7A03G0411800	132,541,926	132,543,485	*TraesCS7A02G175400*	blue copper protein 1b-like
TraesCS7A03G0411900	132,543,489	132,545,593	*TraesCS7A02G175500*	pentatricopeptide repeat-containing protein At5g18475-like
TraesCS7A03G0412600	132,688,877	132,690,801	*TraesCS7A02G176100*	KAT8 regulatory NSL complex subunit 3
TraesCS7A03G0413100	132,886,443	132,890,382	*TraesCS7A02G176200*	transport inhibitor response 1-like protein
TraesCS7A03G0413300	133,233,374	133,237,039	*TraesCS7A02G176300*	GEM-like protein 1
TraesCS7A03G0413900	133,839,827	133,841,015	*TraesCS7A02G176800*	blue copper protein-like
TraesCS7A03G0414600	134,014,983	134,016,747	*TraesCS7A02G177300*	patatin-like protein 1

[a]Gene ID based on the IWGSC RefSeq Annotation v2.1 (IWGSC 2018; Zhu et al. 2021)
[b]Physical position of start and end points for respective genes are based on IWGSC RefSeq v2.0 (IWGSC 2018)
[c]Previous IDs for respective genes to the IDs used in IWGSC RefSeq Annotation v1.1 (IWGSC 2018)

3.4.7 Allelic frequencies of significant QTNs in the HWW breeding programs

Allelic frequencies were extracted for 314 accessions of SD-Panel and another set of 810 accessions, making a total of 1,124 accessions, (Appendix 3.2) to study the effect of selection on these QTNs in the HWW breeding programs. The favorable alleles for three QTNs, *S2B_16305395* (SL and NSPS), *S6B_619882604* (SL), and *S7A_676732614* (SL and NSPS) had a high frequency (0.94, 0.91, and 0.88, respectively) in the breeding materials (Figure 3.10). *S5B_432612793*, a pleiotropic QTN for SL and NSPS on chromosome 5B, had the lowest frequency of the favorable allele (0.07) in the panel. Interestingly, two important QTNs (*S7A_132414615* for NSPS) and (*S7D_60662020* for TKW) showed low to moderate frequencies (28% and 46%, respectively) of the favorable alleles (Figure 3.10), suggesting a possibility of exploring these important genomic regions to improve wheat yield. Out of the 1,124 accessions, 204 accessions were categorized as 'elite' as they were either released cultivars or evaluated in regional nurseries. We assessed the frequencies of favorable alleles for these QTNs in 'elite' material and observed that the frequencies were close in the elite lines and the whole breeding materials (panel of 1,124 accessions), except for *S7A_132414615* (Figure 3.10). For *S7A_132414615*, the frequency of favorable allele was slightly lower in the elite material compared to combined breeding material (Figure 3.10).

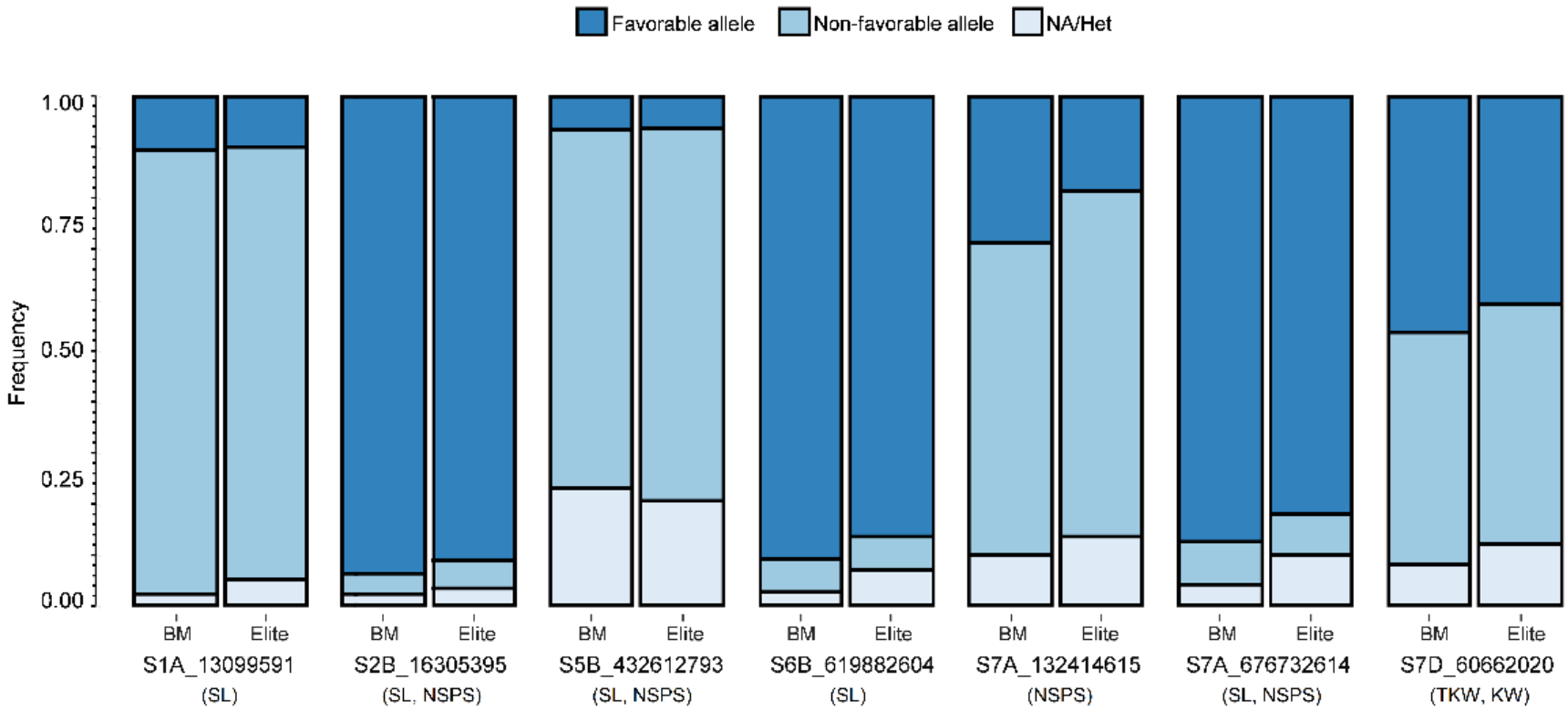

Figure 3.10 Barplots showing the allelic frequencies of stable MTAs identified for different traits in a panel of 1,124 accessions. The bars for 'BM' represent the distribution of favorable/unfavorable alleles in the complete set of breeding material (1,124 accessions) while the bars for 'Elite' represent the distribution in a subset of only elite accessions from the complete panel. A detailed account of the represented MTAs can be found in Table 2.

3.5 Discussion

In the past few decades, continuous improvement in wheat yield throughout the globe has been achieved through improved genetics, breeding, agronomics, and mechanization. Nevertheless, various predictions have indicated that the current rate of increase in wheat improvement might not be sufficient to achieve the yields for future needs. Thus, incessant research efforts are required to improve the yield potential of wheat. Grain yield in wheat is a complex and highly polygenic trait that exhibits a low heritability, making it challenging to improve yield per se. Contrarily, grain yield is largely affected by several components including spike number per unit area, kernel number per spike (KNS), and TKW (J. Liu et al., 2018), which are significantly associated with several spike- and kernel-related traits with high heritability. Hence, a better understanding of the genetic architecture of these yield component traits is key to improving the yield potential of wheat.

A wide variability for three spike traits (SL, NSPS, and SD) and four kernel traits (KL, KW, KA, and TKW) was observed in this study (Table 3.1). The heritability estimates were higher for SL (0.94), NSPS (0.94), and TKW (0.84) in this study than those reported in previous studies (F. Li et al., 2019; Würschum et al., 2018). Moderate (0.76 for KW) to high heritability (0.88 for KA and 0.91 for KL) was observed for kernel traits, in corroboration with previous studies (F. Li et al., 2019; Pang et al., 2020). Overall, high broad-sense heritability estimates across four environments for most of the traits suggest that it could be useful to employ these traits for a better understanding of the genetics underlying the yield potential of wheat. Further, very low coefficients of

variation for the kernel traits across four environments suggest the high repeatability of image-based phenotyping of seed traits.

A significant pairwise correlation was observed among different pairs of the traits (Figure 3.2). Among the spike traits, NSPS was positively correlated to SL (r = 0.63) and SD (r = 0.36) whereas a significant negative correlation was observed between SL and SD (r = - 0.31), consistent with previous studies (J. Liu et al., 2018; Pang et al., 2020; Würschum et al., 2018). A consistent negative association between SL and SD indicates that longer spikes tend to be sparser. Strong positive correlations among KL, KW, and KA are in harmony with previous studies (F. Li et al., 2019; Pang et al., 2020). Further, negative correlation coefficients were observed between NSPS and KL, suggesting a negative impact of increased NSPS on kernel traits. Negative associations of SD with KL and KW indicate that increased SD may negatively impact kernel size in both dimensions. Finally, positive correlations of TKW with KL, KW, and KA and negative correlations of TKW with NSPS and SL agreed with previous reports (Pang et al., 2020; Würschum et al., 2018). Higher effect of KW (r = 0.70) than KL (r = 0.47) on TKW suggests that KW plays a more important role in determining grain weight. Overall, the correlations among different traits show the complex nature of yield component traits, which suggests that genetic progress in yield can be achieved by indirect selection of the component traits with consideration of relationships between individual traits to identify the genotypes that break negative correlations (Würschum et al., 2018). For instance, selecting for higher NSPS with a simultaneous increase in SL can lower the negative impact on SD.

In the current study, BLINK algorithm (Huang et al., 2019) was used to perform GWAS for all traits due to its higher statistical power than MLM (a single-locus method) and FarmCPU (a popular multi-locus method). In BLINK, the bin method of FarmCPU is replaced by LD information, eliminating the requirement that causal genes are evenly distributed (Huang et al., 2019). This method showed better performance than other models using simulated data as well as in several empirical studies from different crop species (Habyarimana et al., 2020; Juliana et al., 2021; L. Liu et al., 2020). We used this method for GWAS on the trait data collected from individual environments as well as combined data from four environments. However, only those MTAs which surpassed the defined threshold and were identified in at least two environments from different years or were associated with more than one trait were reported as stable MTAs. These MTAs could be more reliable and useful in breeding as they are observed over diverse environments.

A total of 10 stable MTAs were identified for SL and NSPS, with three showing pleiotropic effects on both traits (Table 3.3). Two MTAs for SL were present on chromosomes 1A (*S1A_13099591*) and 6B (*S6B_619882604*). A comparison of MTAs identified in the current study with those from previous studies found that QTL for SL has not been reported in the vicinity of *S1A_13099591* in winter wheat although Li et al. (2019) reported a QTL for kernel morphology in the same region (10 – 12 Mbp). In our study, a week MTA ($-\log_{10}(P) > 3.48$) in this region (6 – 11 Mbp) for KL was identified in two individual environments suggesting the presence of a putative QTL with a pleiotropic effect in this region. *S6B_619882604* for SL was co-localized with several QTLs from different studies including plant height at ~620 Mbp (Pang et al., 2020),

NSPS at ~607 Mbp (F. Li et al., 2019), and SL at ~630 Mbp (F. Li et al., 2018), which suggests that this may be an important genomic region for improving spike-related traits.

Two stable MTAs (*S3A_647983369* and *S7A_132414615*) were identified for NSPS on chromosomes 3A and 7A (Table 3.3). The MTA on chromosome 3A was present in a similar location to QTLs for NSPS and grain size (646 – 659 Mbp) in two previous studies (J. Hu et al., 2020; Pang et al., 2020) and these are more likely the same QTL. The second stable MTA (*S7A_132414615*) on chromosome arm 7AS was highly significant in multiple environments and no QTLs for NSPS or SL have been reported in this region, therefore, it is likely a new QTL for NSPS. A strong QTL for NSPS has been recently reported at 47 Mbp region on 7AS (Kuzay et al., 2019) which is about 85 Mbp away from the region identified in the current study.

Three MTAs including *S2B_16305395, S5B_432612793,* and *S7A_676732614* were identified with a pleiotropic effect on both SL and NSPS (Table 3.3, Figure 3.7). The *S2B_16305395* was located at ~16 Mbp on chromosome arm 2BS, where numerous QTLs for SL or NSPS have been identified in several studies (Katkout et al., 2014; Zhai et al., 2016), suggesting this previously reported QTL is present in U.S. hard winter wheat breeding programs. On the other hand, *S5B_432612793* on chromosome 5B has not been reported in previous studies, thus it may be a novel QTL for SL and NSPS. The third MTA *S7A_676732614* on chromosome arm 7AL was co-localized with *WAPO-A1*, a causal gene for NSPS (Kuzay et al., 2019, 2022; Muqaddasi et al., 2019), suggesting these are the same gene for NSPS. Identification of *WAPO-A1* indicates the importance and widespread distribution of this gene in modern wheat cultivars including hard winter wheat breeding material from the U.S. Great Plains.

Seven stable MTAs were identified for KL, KW, and TKW (Table 3.3). *S1A_299864277* for KL is likely associated with a novel QTL because no QTL has been identified previously for kernel morphology in this region. Another MTA, *S7A_717859384* was 15 Mbp away from a previously identified QTL for the same trait (A. Kumar et al., 2016), suggesting they may be the same QTL. Two additional MTAs *S5A_476847493* and *S7D_60662020* for TKW showed a pleiotropic effect on KL and KW, respectively. *S5A_476847493* associated with TKW and KL was in a similar region for a previously reported QTL for KL (G. Liu et al., 2014; R. X. Wang et al., 2009) and a QTL for SL and grain yield (Hu et al. 2020). The MTA represented by *S7D*_60662020 was consistently identified for TKW in multiple environments and was co-localized with an important yield QTL identified in a linkage mapping study using hard winter wheat cultivars from the same region (Dhakal et al., 2021). Apart from this, *S4A_ 619197841* was located near a previously reported genomic region for KL (610-616 Mbp) (Mohler et al., 2016) and kernel weight (Q. Su et al., 2018), and they are more likely to be the same QTL. However, four MTAs (*S1A_13099591, S5B_432612793, S7A_132414615*, and *S1A_299864277*) for SL, NSPS, and KL on chromosomes 1A, 5B, and 7A are likely novel (Table2, Figure 3.7) because QTLs for these traits on these chromosome positions have not been documented to date. Overall, our study not only validated several previously identified QTLs for the spike and kernel traits but also identified putative novel genomic regions associated with those traits. The combined allele analysis showed the considerable additive effects of the identified QTLs on SL and NSPS (Figures 3.8d and 3.8e). Thus, the identified QTLs have the potential to be deployed in winter wheat breeding programs by genomic breeding.

The putative novel QTL on chromosome 7AS for NSPS was stable across multiple environments and showed a similar size of effect to *WAPO-A1* for NSPS. The LD analysis of this region delimited the QTL to a 2.9 Mbp region, and characterization of the 15 SNPs in this region revealed four haplotypes, with *Hap1* being predominant in the SD-Panel (Figure 3.9). The *Hap1* was found to be associated with lower NSPS as significantly higher NSPS were observed in *Hap2* and *Hap3* than in *Hap1* (Figure 3.9c). Though only a few lines carried *Hap2*, many of the *Hap2* lines were elite breeding lines or released cultivars. For example, 'Arapahoe', 'Robidoux', and 'Thompson' carry *Hap2* haplotype and have been important cultivars in the northern Great Plains, suggesting the importance of this QTL in the improvement of the wheat grain yield potential in this region.

The candidate gene analysis for the 2.9 Mbp 7AS region identified 17 putative high confidence candidate genes underlying this QTL. Functional annotation of these genes found that *TraesCS7A03G0411400*, a recently identified *VEGETATIVE TO REPRODUCTIVE TRANSITION* 2 (*VRT2*) gene, encodes a MADS-box transcription factor belonging to the SHORT VEGETATIVE PHASE (SVP) family. Recently, *VRT-A2* has been reported as a causal gene underlying the well-known P1 locus for the specific long-glume trait in Polish wheat (*Triticum polonicum*) (Adamski et al., 2021; J. Liu et al., 2021). The expression levels of *VRT-A2* were correlated with glume length, grain length, and floral organ size (Adamski et al., 2021). One more study suggested *VRT-A2* as a major gene for KL and KL-PW at the *P1* locus in Polish wheat with pleiotropic effects on KL, glume length, and flowering time (Chai et al., 2021). Recently, Li et al. (2021) investigated the interactions between *VRN1* and *FUL2* genes from SQUAMOSA-clade

with *VRT* and *SVP* genes for their effects on the regulation of spike and spikelet development, where *VRT* and *SVP* mutants (*vrt2* and *svp1*) significantly reduced the NSPS, with differences for *VRT* alleles being more predominant (K. Li et al., 2021). Another recent study reported that VRT2 is involved in increasing the number of rudimentary basal spikelets in wheat (Backhaus et al., 2022). Based on the findings from the above-discussed studies, *VRT-A2* could be a likely candidate gene underlying the 7AS QTL (*S7A_132414615*) for NSPS.

Further, we analyzed 1,124 accessions for the distribution of favorable alleles of the stable MTAs in the breeding materials from the SDSU breeding program and other hard winter wheat breeding programs in the Great Plains (Figure 3.10). A high frequency (94%, 91%, and 88%) of the favorable alleles of the three genomic regions (*S2B_16305395* for SL and NSPS, *S6B_619882604* for SL, and *S7A_676732614* for SL and NSPS), respectively, were observed indicating the importance of these regions to yield improvement of the U.S. hard winter wheat. As expected, *WAPO-A1* (*S7A_676732614* region) seems to be an important gene in the HWW germplasm, because 37 released hard winter wheat cultivars used in this study all carry the favorable allele with only two exceptions (Appendix 3.2). The favorable allele of another QTN (*S5B_432612793*) for SL and NSPS was found in only 7% of the lines studied. Intriguingly, three released cultivars from different breeding programs (‘Emerson’, ‘Flourish’, and ‘Oahe’) had favorable alleles for *S5B_432612793*. For *S7A_676732614*, only ‘Oahe’ carries the positive allele among the three cultivars carrying the favorable allele of *S5B_432612793* (Appendix 3.2). Thus, QTL on chromosome 5B could be

another useful QTL for improving NSPS in the breeding programs from the northern Great Plains.

Favorable alleles for two additional QTNs, *S7A_132414615* for NSPS and *S7D_60662020* for TKW, showed low to moderate frequencies in the breeding materials. Among the 1,124 accessions, although only 46% of the germplasm carried the favorable allele of *S7D_60662020*, 24 released cultivars from the Great Plains carry the favorable allele (Appendix 3.2). Similarly, the allelic frequency of the favorable allele for *S7A_132414615* was only 29% in the complete set of 1,124 lines and slightly lower in the elite subset (Figure 3.10). However, several cultivars including 'Arapahoe', 'Art', 'Robidoux', 'Smoky Hill', and 'Thompson' carry the favorable allele for this QTN (Appendix 3.2). Thus, it will be interesting to study if these regions are associated with factors affecting adaptability and hinder the selection for certain loci. Overall, increasing the frequency of the favorable allele at the QTLs with low frequency through genomic-assisted breeding may improve the yield potential of hard winter wheat.

In conclusion, a significant variation for various yield component traits exists in hard winter wheat breeding programs from the U.S. Great Plains. Among 17 stable MTAs identified in this study, four represent putative novel genomic regions. Development of breeder-friendly Kompetitive allele-specific PCR (KASP) assays for these MTAs using the provided information (Appendix 3.3) will be useful to facilitate the deployment of these QTLs through marker-assisted selection in the early stages of the breeding process (Gill et al., 2019). Unlike a diversity panel, this study used breeding materials as the panels for GWAS and the results could be directly used to select the parental lines with more favorable alleles for making crosses by the breeders. Moreover,

the allelic frequencies of identified MTAs could be used to accumulate the useful QTLs at low frequency in the breeding materials. The QTL for NSPS on chromosome arm 7AS (*S7A_132414615*) identified in this study appears to be an important yield-related QTL for U.S. hard winter wheat and can be investigated further. Finally, the genomic information for MTAs reported in this study can be incorporated into the genomic prediction models to evaluate their potential for the selection of future winter wheat varieties with higher grain yield potential.

3.6 References

Adamski, N. M., Simmonds, J., Brinton, J. F., Backhaus, A. E., Chen, Y., Smedley, M., Hayta, S., Florio, T., Crane, P., Scott, P., Pieri, A., Hall, O., Barclay, J. E., Clayton, M., Doonan, J. H., Nibau, C., & Uauy, C. (2021). Ectopic expression of *Triticum polonicum VRT-A2* underlies elongated glumes and grains in hexaploid wheat in a dosage-dependent manner. *The Plant Cell*, *33*(7), 2296–2319. https://doi.org/10.1093/plcell/koab119

Alqudah, A. M., Haile, J. K., Alomari, D. Z., Pozniak, C. J., Kobiljski, B., & Börner, A. (2020). Genome-wide and SNP network analyses reveal genetic control of spikelet sterility and yield-related traits in wheat. *Scientific Reports*, *10*(1), 1–12. https://doi.org/10.1038/s41598-020-59004-4

Alvarado, G., Rodríguez, F. M., Pacheco, A., Burgueño, J., Crossa, J., Vargas, M., Pérez-Rodríguez, P., & Lopez-Cruz, M. A. (2020). META-R: A software to analyze data from multi-environment plant breeding trials. *Crop Journal*, *8*(5), 745–756. https://doi.org/10.1016/j.cj.2020.03.010

Backhaus, A. E., Lister, A., Tomkins, M., Adamski, N. M., Macaulay, I., Morris, R. J., Haerty, W., & Uauy, C. (2022). High expression of VRT2 increases the number of rudimentary basal 2 spikelets in wheat. *BioRxiv*, 2021.08.03.454952. https://doi.org/10.1101/2021.08.03.454952

Barrett, J. C., Fry, B., Maller, J., & Daly, M. J. (2005). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, *21*(2), 263–265. https://doi.org/10.1093/bioinformatics/bth457

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Begum, H., Spindel, J. E., Lalusin, A., Borromeo, T., Gregorio, G., Hernandez, J., Virk, P., Collard, B., & McCouch, S. R. (2015). Genome-Wide Association Mapping for Yield and Other Agronomic Traits in an Elite Breeding Population of Tropical Rice (Oryza sativa). *PLOS ONE*, *10*(3), e0119873. https://doi.org/10.1371/journal.pone.0119873

Börner, A., Schumann, E., Fürste, A., Cöster, H., Leithold, B., Röder, M. S., & Weber, W. E. (2002). Mapping of quantitative trait loci determining agronomic important characters in hexaploid wheat (Triticum aestivum L.). *Theoretical and Applied Genetics*, *105*(6–7), 921–936. https://doi.org/10.1007/s00122-002-0994-1

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics*, *23*(19), 2633–2635. https://doi.org/10.1093/bioinformatics/btm308

Browning, S. R., & Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics*, *81*(5), 1084–1097. https://doi.org/10.1086/521987

Chai, S., Yao, Q., Liu, R., Xiang, W., Xiao, X., Fan, X., Zeng, J., Sha, L., Kang, H., Zhang, H., Long, D., Wu, D., Zhou, Y., & Wang, Y. (2021). Identification and validation of a major gene for kernel length at the P1 locus in Triticum polonicum. *Crop Journal*. https://doi.org/10.1016/j.cj.2021.07.006

Chen, G., Zhang, H., Deng, Z., Wu, R., Li, D., Wang, M., & Tian, J. (2016). Genome-wide association study for kernel weight-related traits using SNPs in a Chinese winter wheat population. *Euphytica*, *212*(2), 173–185. https://doi.org/10.1007/s10681-016-1750-y

Chen, Z., Cheng, X., Chai, L., Wang, Z., Bian, R., Li, J., Zhao, A., Xin, M., Guo, W., Hu, Z., Peng, H., Yao, Y., Sun, Q., & Ni, Z. (2020). Dissection of genetic factors underlying grain size and fine mapping of QTgw.cau-7D in common wheat (Triticum aestivum L.). *Theoretical and Applied Genetics*, *133*(1), 149–162. https://doi.org/10.1007/s00122-019-03447-5

Conesa, A., Gotz, S., Garcia-Gomez, J. M., Terol, J., Talon, M., & Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, *21*(18), 3674–3676. https://doi.org/10.1093/bioinformatics/bti610

Dhakal, S., Liu, X., Chu, C., Yang, Y., Rudd, J. C., Ibrahim, A. M. H., Xue, Q., Devkota, R. N., Baker, J. A., Baker, S. A., Simoneaux, B. E., Opena, G. B., Sutton, R., Jessup,

K. E., Hui, K., Wang, S., Johnson, C. D., Metz, R. P., & Liu, S. (2021). Genome-wide QTL mapping of yield and agronomic traits in two widely adapted winter wheat cultivars from multiple mega-environments. *PeerJ, 9*, e12350. https://doi.org/10.7717/peerj.12350

Doyle, J. J., & Doyle, J. L. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *PHYTOCHEMICAL BULLETIN*. https://worldveg.tind.io/record/33886

Earl, D. A., & vonHoldt, B. M. (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, *4*(2), 359–361. https://doi.org/10.1007/s12686-011-9548-7

Epskamp, S., Cramer, A. O. J., Waldorp, L. J., Schmittmann, V. D., & Borsboom, D. (2012). Qgraph: Network visualizations of relationships in psychometric data. *Journal of Statistical Software*, *48*(4). https://doi.org/10.18637/jss.v048.i04

Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Molecular Ecology*, *14*(8), 2611–2620. https://doi.org/10.1111/j.1365-294X.2005.02553.x

FAO. (2017). *The future of food and agriculture – Trends and challenges.* FAO.

Faris, J. D., Zhang, Z., Garvin, D. F., & Xu, S. S. (2014). Molecular and comparative mapping of genes governing spike compactness from wild emmer wheat. *Molecular Genetics and Genomics*, *289*(4), 641–651. https://doi.org/10.1007/s00438-014-0836-2

Fischer, R., Byerlee, D., & Edmeades, G. (2014). Crop yields and global food security. In

academia.edu. ACIAR: Canberra, ACT. http://www.academia.edu/download/35887178/Crop_yields_and_global_food_security___a_book_by_T.Fischer_et_al__2014.pdf

Gao, F., Wen, W., Liu, J., Rasheed, A., Yin, G., Xia, X., Wu, X., & He, Z. (2015). Genome-wide linkage mapping of QTL for yield components, plant height and yield-related physiological traits in the Chinese wheat cross Zhou 8425B/Chinese spring. *Frontiers in Plant Science*, *6*(DEC). https://doi.org/10.3389/fpls.2015.01099

Gegas, V. C., Nazari, A., Griffiths, S., Simmonds, J., Fish, L., Orford, S., Sayers, L., Doonan, J. H., & Snape, J. W. (2010). A Genetic Framework for Grain Size and Shape Variation in Wheat . *The Plant Cell*, *22*(4), 1046–1056. https://doi.org/10.1105/tpc.110.074153

Gill, H. S., Halder, J., Zhang, J., Brar, N. K., Rai, T. S., Hall, C., Bernardo, A., Amand, P. S., Bai, G., Olson, E., Ali, S., Turnipseed, B., & Sehgal, S. K. (2021). Multi-Trait Multi-Environment Genomic Prediction of Agronomic Traits in Advanced Breeding Lines of Winter Wheat. *Frontiers in Plant Science*, *12*. https://doi.org/10.3389/fpls.2021.709545

Gill, H. S., Li, C., Sidhu, J. S., Liu, W., Wilson, D., Bai, G., Gill, B. S., & Sehgal, S. K. (2019). Fine Mapping of the Wheat Leaf Rust Resistance Gene Lr42. *International Journal of Molecular Sciences*, *20*(10). https://doi.org/10.3390/ijms20102445

Grote, U., Fasse, A., Nguyen, T. T., & Erenstein, O. (2021). Food Security and the Dynamics of Wheat and Maize Value Chains in Africa and Asia. In *Frontiers in Sustainable Food Systems* (Vol. 4, p. 317). Frontiers Media S.A. https://doi.org/10.3389/fsufs.2020.617009

Guo, Z., Chen, D., Alqudah, A. M., Röder, M. S., Ganal, M. W., & Schnurbusch, T. (2017). Genome-wide association analyses of 54 traits identified multiple loci for the determination of floret fertility in wheat. *New Phytologist*, *214*(1), 257–270. https://doi.org/10.1111/nph.14342

Habyarimana, E., De Franceschi, P., Ercisli, S., Baloch, F. S., & Dall'Agata, M. (2020). Genome-Wide Association Study for Biomass Related Traits in a Panel of Sorghum bicolor and S. bicolor × S. halepense Populations. *Frontiers in Plant Science, 11*, 1796. https://doi.org/10.3389/fpls.2020.551305

Halder, J., Zhang, J., Ali, S., Sidhu, J. S., Gill, H. S., Talukder, S. K., Kleinjan, J., Turnipseed, B., & Sehgal, S. K. (2019). Mining and genomic characterization of resistance to tan spot, Stagonospora nodorum blotch (SNB), and Fusarium head blight in Watkins core collection of wheat landraces. *BMC Plant Biology*, *19*(1), 1–15. https://doi.org/10.1186/s12870-019-2093-3

Hill, W. G., & Weir, B. S. (1988). Variances and covariances of squared linkage disequilibria in finite populations. *Theoretical Population Biology*, *33*(1), 54–78. https://doi.org/10.1016/0040-5809(88)90004-4

Hou, J., Jiang, Q., Hao, C., Wang, Y., Zhang, H., & Zhang, X. (2014). Global Selection on Sucrose Synthase Haplotypes during a Century of Wheat Breeding. *Plant Physiology*, *164*(4), 1918–1929. https://doi.org/10.1104/pp.113.232454

Hu, J., Wang, X., Zhang, G., Jiang, P., Chen, W., Hao, Y., Ma, X., Xu, S., Jia, J., Kong, L., & Wang, H. (2020). QTL mapping for yield-related traits in wheat based on four RIL populations. *Theoretical and Applied Genetics*, *133*(3), 917–933. https://doi.org/10.1007/s00122-019-03515-w

Hu, M. J., Zhang, H. P., Cao, J. J., Zhu, X. F., Wang, S. X., Jiang, H., Wu, Z. Y., Lu, J., Chang, C., Sun, G. Lou, & Ma, C. X. (2016). Characterization of an IAA-glucose hydrolase gene TaTGW6 associated with grain weight in common wheat (Triticum aestivum L.). *Molecular Breeding*, *36*(3), 1–11. https://doi.org/10.1007/s11032-016-0449-z

Huang, M., Liu, X., Zhou, Y., Summers, R. M., & Zhang, Z. (2019). BLINK: A package for the next level of genome-wide association studies with both individuals and markers in the millions. *GigaScience*, *8*(2), 1–12. https://doi.org/10.1093/gigascience/giy154

IWGSC. (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science (New York, N.Y.)*, *361*(6403), eaar7191. https://doi.org/10.1126/science.aar7191

Juliana, P., Singh, R. P., Poland, J., Shrestha, S., Huerta-Espino, J., Govindan, V., Mondal, S., Crespo-Herrera, L. A., Kumar, U., Joshi, A. K., Payne, T., Bhati, P. K., Tomar, V., Consolacion, F., & Campos Serna, J. A. (2021). Elucidating the genetics of grain yield and stress-resilience in bread wheat using a large-scale genome-wide association mapping study with 55,568 lines. *Scientific Reports*, *11*(1), 1–15. https://doi.org/10.1038/s41598-021-84308-4

Katkout, M., Kishii, M., Kawaura, K., Mishina, K., Sakuma, S., Umeda, K., Takumi, S., Nitta, M., Nasuda, S., & Ogihara, Y. (2014). QTL analysis of genetic loci affecting domestication-related spike characters in common wheat. *Genes and Genetic Systems*, *89*(3), 121–131. https://doi.org/10.1266/ggs.89.121

Kumar, A., Mantovani, E. E., Seetan, R., Soltani, A., Echeverry-Solarte, M., Jain, S.,

Simsek, S., Doehlert, D., Alamri, M. S., Elias, E. M., Kianian, S. F., & Mergoum, M. (2016). Dissection of Genetic Factors underlying Wheat Kernel Shape and Size in an Elite × Nonadapted Cross using a High Density SNP Linkage Map. *The Plant Genome*, *9*(1), plantgenome2015.09.0081. https://doi.org/10.3835/plantgenome2015.09.0081

Kumar, D., Sharma, S., Sharma, R., Pundir, S., Singh, V. K., Chaturvedi, D., Singh, B., Kumar, S., & Sharma, S. (2021). Genome-wide association study in hexaploid wheat identifies novel genomic regions associated with resistance to root lesion nematode (Pratylenchus thornei). *Scientific Reports*, *11*(1), 3572. https://doi.org/10.1038/s41598-021-80996-0

Kuzay, S., Lin, H., Li, C., Chen, S., Woods, D. P., Zhang, J., Lan, T., von Korff, M., & Dubcovsky, J. (2022). WAPO-A1 is the causal gene of the 7AL QTL for spikelet number per spike in wheat. *PLOS Genetics*, *18*(1), e1009747. https://doi.org/10.1371/journal.pgen.1009747

Kuzay, S., Xu, Y., Zhang, J., Katz, A., Pearce, S., Su, Z., Fraser, M., Anderson, J. A., Brown-Guedira, G., DeWitt, N., Peters Haugrud, A., Faris, J. D., Akhunov, E., Bai, G., & Dubcovsky, J. (2019). Identification of a candidate gene for a QTL for spikelet number per spike on wheat chromosome arm 7AL by high-resolution genetic mapping. *Theoretical and Applied Genetics*, *132*(9), 2689–2705. https://doi.org/10.1007/s00122-019-03382-5

Li, F., Wen, W., He, Z., Liu, J., Jin, H., Cao, S., Geng, H., Yan, J., Zhang, P., Wan, Y., & Xia, X. (2018). Genome-wide linkage mapping of yield-related traits in three Chinese bread wheat populations using high-density SNP markers. *Theoretical and*

Applied Genetics, *131*(9), 1903–1924. https://doi.org/10.1007/s00122-018-3122-6

Li, F., Wen, W., Liu, J., Zhang, Y., Cao, S., He, Z., Rasheed, A., Jin, H., Zhang, C., Yan, J., Zhang, P., Wan, Y., & Xia, X. (2019). Genetic architecture of grain yield in bread wheat based on genome-wide association studies. *BMC Plant Biology*, *19*(1), 168. https://doi.org/10.1186/s12870-019-1781-3

Li, K., Debernardi, J. M., Li, C., Lin, H., Zhang, C., Jernstedt, J., Korff, M. von, Zhong, J., & Dubcovsky, J. (2021). Interactions between SQUAMOSA and SHORT VEGETATIVE PHASE MADS-box proteins regulate meristem transitions during wheat spike development. *The Plant Cell*, *33*(12), 3621–3644. https://doi.org/10.1093/plcell/koab243

Liu, G., Jia, L., Lu, L., Qin, D., Zhang, J., Guan, P., Ni, Z., Yao, Y., Sun, Q., & Peng, H. (2014). Mapping QTLs of yield-related traits using RIL population derived from common wheat and Tibetan semi-wild wheat. *Theoretical and Applied Genetics*, *127*(11), 2415–2432. https://doi.org/10.1007/s00122-014-2387-7

Liu, H., Zhang, X., Xu, Y., Ma, F., Zhang, J., Cao, Y., Li, L., & An, D. (2020). Identification and validation of quantitative trait loci for kernel traits in common wheat (Triticum aestivum L.). *BMC Plant Biology*, *20*(1), 529. https://doi.org/10.1186/s12870-020-02661-4

Liu, J., Chen, Z., Wang, Z., Zhang, Z., Xie, X., Wang, Z., Chai, L., Song, L., Cheng, X., Feng, M., Wang, X., Liu, Y., Hu, Z., Xing, J., Su, Z., Peng, H., Xin, M., Yao, Y., Guo, W., … Ni, Z. (2021). Ectopic expression of VRT-A2 underlies the origin of Triticum polonicum and Triticum petropavlovskyi with long outer glumes and grains. *Molecular Plant*, *14*(9), 1472–1488.

https://doi.org/10.1016/j.molp.2021.05.021

Liu, J., Xu, Z., Fan, X., Zhou, Q., Cao, J., Wang, F., Ji, G., Yang, L., Feng, B., & Wang, T. (2018). A genome-wide association study of wheat spike related traits in China. *Frontiers in Plant Science*, *871*, 1584. https://doi.org/10.3389/fpls.2018.01584

Liu, K., Sun, X., Ning, T., Duan, X., Wang, Q., Liu, T., An, Y., Guan, X., Tian, J., & Chen, J. (2018). Genetic dissection of wheat panicle traits using linkage analysis and a genome-wide association study. *Theoretical and Applied Genetics*, *131*(5), 1073–1090. https://doi.org/10.1007/s00122-018-3059-9

Liu, L., Wang, M., Zhang, Z., See, D. R., & Chen, X. (2020). Identification of Stripe Rust Resistance Loci in U.S. Spring Wheat Cultivars and Breeding Lines Using Genome-Wide Association Mapping and Yr Gene Markers. *Plant Disease*, *104*(8), 2181–2192. https://doi.org/10.1094/PDIS-11-19-2402-RE

Liu, X., Huang, M., Fan, B., Buckler, E. S., & Zhang, Z. (2016). Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient Genome-Wide Association Studies. *PLOS Genetics*, *12*(2), e1005767. https://doi.org/10.1371/journal.pgen.1005767

Mendiburu Felipe de. (2021). *"agricolae": Statistical Procedures for Agricultural Research*. https://cran.r-project.org/web/packages/agricolae/agricolae.pdf

Mohler, V., Albrecht, T., Castell, A., Diethelm, M., Schweizer, G., & Hartl, L. (2016). Considering causal genes in the genetic dissection of kernel traits in common wheat. *Journal of Applied Genetics*, *57*(4), 467–476. https://doi.org/10.1007/s13353-016-0349-2

Muqaddasi, Q. H., Brassac, J., Koppolu, R., Plieske, J., Ganal, M. W., & Röder, M. S.

(2019). TaAPO-A1, an ortholog of rice ABERRANT PANICLE ORGANIZATION 1, is associated with total spikelet number per spike in elite European hexaploid winter wheat (Triticum aestivum L.) varieties. *Scientific Reports*, *9*(1), 1–12. https://doi.org/10.1038/s41598-019-50331-9

Pang, Y., Liu, C., Wang, D., St. Amand, P., Bernardo, A., Li, W., He, F., Li, L., Wang, L., Yuan, X., Dong, L., Su, Y., Zhang, H., Zhao, M., Liang, Y., Jia, H., Shen, X., Lu, Y., Jiang, H., … Liu, S. (2020). High-Resolution Genome-wide Association Study Identifies Genomic Regions and Candidate Genes for Important Agronomic Traits in Wheat. *Molecular Plant*, *13*(9), 1311–1327. https://doi.org/10.1016/j.molp.2020.07.008

Poland, J., Endelman, J., Dawson, J., Rutkoski, J., Wu, S., Manes, Y., Dreisigacker, S., Crossa, J., Sánchez-Villeda, H., Sorrells, M., & Jannink, J. (2012). Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing. *The Plant Genome*, *5*(3), plantgenome2012.06.0006. https://doi.org/10.3835/plantgenome2012.06.0006

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, *155*(2), 945 LP – 959. http://www.genetics.org/content/155/2/945.abstract

Sidhu, J. S., Singh, D., Gill, H. S., Brar, N. K., Qiu, Y., Halder, J., Al Tameemi, R., Turnipseed, B., & Sehgal, S. K. (2020). Genome-Wide Association Study Uncovers Novel Genomic Regions Associated With Coleoptile Length in Hard Winter Wheat. *Frontiers in Genetics*, *10*, 1345. https://doi.org/10.3389/fgene.2019.01345

Sourdille, P., Tixier, M. H., Charmet, G., Gay, G., Cadalen, T., Bernard, S., & Bernard, M. (2000). Location of genes involved in ear compactness in wheat (Triticum

aestivum) by means of molecular markers. *Molecular Breeding*, *6*(3), 247–255. https://doi.org/10.1023/A:1009688011563

Su, Q., Zhang, X., Zhang, W., Zhang, N., Song, L., Liu, L., Xue, X., Liu, G., Liu, J., Meng, D., Zhi, L., Ji, J., Zhao, X., Yang, C., Tong, Y., Liu, Z., & Li, J. (2018). QTL Detection for Kernel Size and Weight in Bread Wheat (Triticum aestivum L.) Using a High-Density SNP and SSR-Based Linkage Map. *Frontiers in Plant Science*, *9*, 1484. https://doi.org/10.3389/fpls.2018.01484

Su, Z., Hao, C., Wang, L., Dong, Y., & Zhang, X. (2011). Identification and development of a functional marker of TaGW2 associated with grain weight in bread wheat (Triticum aestivum L.). *Theoretical and Applied Genetics*, *122*(1), 211–223. https://doi.org/10.1007/s00122-010-1437-z

Sukumaran, S., Dreisigacker, S., Lopes, M., Chavez, P., & Reynolds, M. P. (2014). Genome-wide association study for grain yield and related traits in an elite spring wheat population grown in temperate irrigated environments. *Theoretical and Applied Genetics*, *128*(2), 353–363. https://doi.org/10.1007/s00122-014-2435-3

Team, R. C. (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.r-project.org/

USDA. (2021). *Acreage (June 2021): USDA National Agricultural Statistics Service*. https://www.nass.usda.gov/Publications/Todays_Reports/reports/acrg0621.pdf

Wang, J., & Zhang, Z. (2021). GAPIT Version 3: Boosting Power and Accuracy for Genomic Association and Prediction. *Genomics, Proteomics & Bioinformatics*. https://doi.org/10.1016/j.gpb.2021.08.005

Wang, R. X., Hai, L., Zhang, X. Y., You, G. X., Yan, C. S., & Xiao, S. H. (2009). QTL

mapping for grain filling rate and yield-related traits in RILs of the Chinese winter wheat population Heshangmai x Yu8679. *Theoretical and Applied Genetics*, *118*(2), 313–325. https://doi.org/10.1007/s00122-008-0901-5

Ward, B. P., Brown-Guedira, G., Kolb, F. L., Van Sanford, D. A., Tyagi, P., Sneller, C. H., & Griffey, C. A. (2019). Genome-wide association studies for yield-related traits in soft red winter wheat grown in Virginia. *PLoS ONE*, *14*(2), e0208217. https://doi.org/10.1371/journal.pone.0208217

Wheeler, T., & Von Braun, J. (2013). Climate change impacts on global food security. In *Science* (Vol. 341, Issue 6145, pp. 508–513). American Association for the Advancement of Science. https://doi.org/10.1126/science.1239402

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York. https://cran.r-project.org/web/packages/ggplot2/citation.html

William, R. (2013). *psych: Procedures for Personality and Psychological Research*. http://cran.r-project.org/package=psych

Wu, X., Chang, X., & Jing, R. (2012). Genetic Insight into Yield-Associated Traits of Wheat Grown in Multiple Rain-Fed Environments. *PLoS ONE, 7*(2), e31249. https://doi.org/10.1371/journal.pone.0031249

Würschum, T., Leiser, W. L., Langer, S. M., Tucker, M. R., & Longin, C. F. H. (2018). Phenotypic and genetic analysis of spike and kernel characteristics in wheat reveals long-term genetic trends of grain yield components. *Theoretical and Applied Genetics*, *131*(10), 2071–2084. https://doi.org/10.1007/s00122-018-3133-3

Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., Kresovich, S., & Buckler, E. S.

(2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, *38*(2), 203–208. https://doi.org/10.1038/ng1702

Yu, M., Mao, S. L., Chen, G. Y., Pu, Z. E., Wei, Y. M., & Zheng, Y. L. (2014). QTLs for uppermost internode and spike length in two wheat RIL populations and their affect upon plant height at an individual QTL level. *Euphytica*, *200*(1), 95–108. https://doi.org/10.1007/s10681-014-1156-7

Zanke, C. D., Ling, J., Plieske, J., Kollers, S., Ebmeyer, E., Korzun, V., Argillier, O., Stiewe, G., Hinze, M., Neumann, F., Eichhorn, A., Polley, A., Jaenecke, C., Ganal, M. W., & Röder, M. S. (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*(september), 644. https://doi.org/10.3389/fpls.2015.00644

Zhai, H., Feng, Z., Li, J., Liu, X., Xiao, S., Ni, Z., & Sun, Q. (2016). QTL analysis of spike morphological traits and plant height in winter wheat (Triticum aestivum L.) using a high-density SNP and SSR-based linkage map. *Frontiers in Plant Science*, *7*(November 2016). https://doi.org/10.3389/fpls.2016.01617

Zhang, J., Gizaw, S. A., Bossolini, E., Hegarty, J., Howell, T., Carter, A. H., Akhunov, E., & Dubcovsky, J. (2018). Identification and validation of QTL for grain yield and plant water status under contrasting water treatments in fall-sown spring wheats. *Theoretical and Applied Genetics*, *131*(8), 1741–1759. https://doi.org/10.1007/s00122-018-3111-9

Zhou, Y., Conway, B., Miller, D., Marshall, D., Cooper, A., Murphy, P., Chao, S.,

Brown-Guedira, G., & Costa, J. (2017). Quantitative Trait Loci Mapping for Spike Characteristics in Hexaploid Wheat. *The Plant Genome*, *10*(2), plantgenome2016.10.0101. https://doi.org/10.3835/plantgenome2016.10.0101

Zhu, T., Wang, L., Rimbert, H., Rodriguez, J. C., Deal, K. R., De Oliveira, R., Choulet, F., Keeble-Gagnère, G., Tibbits, J., Rogers, J., Eversole, K., Appels, R., Gu, Y. Q., Mascher, M., Dvorak, J., & Luo, M. (2021). Optical maps refine the bread wheat *Triticum aestivum* cv. Chinese Spring genome assembly. *The Plant Journal*, *107*(1), 303–314. https://doi.org/10.1111/tpj.15289

CHAPTER 4

Multi-trait multi-environment genomic prediction of agronomic traits in advanced breeding lines of winter wheat

This chapter has been published as an original research article in the journal '*Frontiers in Plant Science'*.

Citation: Gill, Harsimardeep S., Jyotirmoy Halder, Jinfeng Zhang, Navreet K. Brar, Teerath S. Rai, Cody Hall, Amy Bernardo et al. "Multi-trait multi-environment genomic prediction of agronomic traits in advanced breeding lines of winter wheat." Frontiers in Plant Science 12 (2021): 709545.

4.1 Abstract

Genomic prediction (GP) is a promising approach for accelerating the genetic gain of complex traits in wheat breeding. However, increasing the prediction accuracy (PA) of GP models remains a challenge in the successful implementation of this approach. Multivariate approaches can leverage the simultaneous evaluation of several traits under multiple environments by exploring correlations to improve GS performance in breeding programs. Though, these models have been mostly evaluated using diverse panels of unrelated accessions. Here, we used multivariate GP models to predict multiple agronomic traits using 314 advanced and elite breeding lines of winter wheat evaluated at ten site-year environments. We evaluated a multi-trait (MT) model with two cross-validation schemes representing different breeding scenarios (CV1, prediction of completely unphenotyped lines; and CV2, prediction of lines partially phenotyped for correlated traits). Moreover, extensive data from multi-environment trials (METs) was used to cross-validate the Bayesian multi-trait multi-environment (MTME) model that integrates the analysis of multiple-traits including GxE interaction. The MT-CV2 model outperformed all other models for predicting grain yield with significant improvement in PA over the single-trait (ST-CV1) model. The MTME model performed better for all traits, with average improvement over the ST-CV1 reaching up to 19%, 71%, 17%, 48%, and 51% for grain yield, grain protein content, test weight, plant height, and days to heading, respectively. Overall, our empirical analyses elucidate the potential of both MT-CV2 and MTME models when advanced breeding lines are used as training population to predict related preliminary breeding lines. Further, we also evaluated the practical application of MTME model in our breeding program to reduce phenotyping cost by

using a sparse testing design. This showed that complementing METs with GP can substantially enhance resource efficiency. Our results demonstrate that the multivariate GS models hold great potential in implementing GS in breeding programs.

4.2 Introduction

Global wheat production needs to be increased by 60% to meet the demand of a projected population of 9 billion by 2050 (Fischer et al., 2014; Tester & Langridge, 2010). In the past few decades, wheat breeding successfully achieved a significant increase in grain yield owing to significantly improved genetic resources, implementation of modern agronomic practices, accurate experimental designs, and other improved technology packages (Tadesse et al., 2019), which translates into an annual increase of 1% in terms of genetic gain in grain yield. However, this increase is still far from the expected yearly growth of 1.7% to meet the future wheat demand (Oury et al., 2012; Tadesse et al., 2019). Thus, new and innovative breeding technologies are essential to achieve a two-fold increase in annual yield to avoid potential food crises in the coming decades.

Traditional wheat breeding involves creating novel genetic variation by different methods, followed by extensive selection and advancement of generations. The selection of progeny with desirable agronomic and end-use quality traits is a resource-intensive process and could take up to 10-15 years to develop a new cultivar (Haile et al., 2020). Further, in traits with complex genetic architecture such as grain yield, the genotype-by-environment interactions play a paramount role and impose additional challenges in selection. In recent years, the deployment of molecular markers for marker-assisted selection (MAS) has been used to increase selection accuracy and accelerate genetic gain (Randhawa et al., 2013). Though MAS has shown good potential in wheat breeding for

the deployment of QTLs with large effects, its application has been limited to improve complex traits governed by many QTLs with small effects (Heffner et al., 2009).

Genomic selection (GS) is a recent approach that utilizes genome-wide marker data to select individuals superior for complex traits in the early breeding cycle to increase the genetic gain per unit of time (Heffner et al., 2009; Meuwissen et al., 2001). Unlike MAS, GS does not require prior identification of QTLs for the traits of interest; instead, it employs all available markers across the genome to predict individuals' breeding values (Bassi et al., 2015). Briefly, GS requires a training population (TP), which is genotyped with genome-wide markers and phenotyped for a given trait(s) of interest. GS involves calibration of a prediction model using TP to estimate marker effects and evaluate the predictive ability of the model through cross-validation. Finally, the developed model is used to calculate genome-estimated breeding values (GEBVs) and rank the lines from a breeding or testing population (BP) that consists of lines with only genotypic information. Thus, the early selection or culling of individuals based on the GEBVs permits greater genetic gain per breeding cycle, facilitating an increase in the efficacy of breeding programs and resulting in reduced varietal development costs. Several studies have reported successful implementation of GS in different crops resulting in an accelerated rate of genetic gain compared to traditional breeding (Bassi et al., 2015; Battenfield et al., 2016; Bhat et al., 2016). Moreover, GS has shown to be particularly useful in traits where phenotyping is cumbersome, such as quaality traits and complex resistance to diseases (Battenfield et al., 2016; Dong et al., 2018).

The widespread availability of genome-wide markers attributed to low-cost genotyping technologies has facilitated the adaptability of GS in wheat breeding

programs (Bhat et al., 2016; J. Poland et al., 2012). Thus, there is growing interest in recent years to complement phenotyping selection and genomic selection in wheat breeding. GS has been evaluated for many complex traits in wheat, including but not limited to grain yield and yield-related traits (Guo et al., 2020; Haile et al., 2020; Juliana et al., 2020; Rutkoski et al., 2016; Ward et al., 2019), wheat resistance to rusts (Juliana et al., 2017; J. E. Rutkoski et al., 2014) and Fusarium head blight (Arruda et al., 2015; Dong et al., 2018; Rutkoski et al., 2012), and end-use quality traits (Battenfield et al., 2016; Ibba et al., 2020; Lado et al., 2018). Despite the successful evaluations of GS in wheat breeding programs, there is a continuous scope to improve the prediction accuracy/ability of GS models for quantitative traits to achieve higher genetic gains that will lead to the routine implementation of GS in various wheat breeding schemes.

Predictive ability (PA) of the GS model refers to the correlation between estimated GEBVs and the actual phenotypic values of the individuals in the validation set and is generally calculated through a cross-validation approach. Along with TP size, the extent of linkage disequilibrium (LD), and the heritability of the traits, the PA also depends on the choice and optimization of the statistical models (de los Campos et al., 2013; J. Guo et al., 2020; J. Rutkoski et al., 2016). In most studies, penalized genomic prediction models, including ridge-regression best linear unbiased prediction (rrBLUP) and genomic best linear unbiased prediction (GBLUP), have been standard GS approaches (Endelman, 2011; VanRaden et al., 2009). In addition, several Bayesian methods with different prior distributions and relying on Markov-Chain Monte Carlo (MCMC) for the estimation of parameters have proven useful for genomic prediction

(Habier et al., 2011; Xin Wang et al., 2018). However, most of these models implement a univariate linear mixed model and are helpful to predict one dependent variable at a time.

In recent years, multi-trait (MT) genomic prediction models have been suggested to improve the PA for a primary trait when secondary traits correlated to the primary trait are available (Jia & Jannink, 2012). The use of genetically correlated traits is of particular importance when the primary trait is difficult or expensive to phenotype and has low heritability. Several empirical studies have successfully evaluated MT approaches for different agronomic traits in wheat breeding (Hayes et al., 2017; Lado et al., 2018; Rutkoski et al., 2012). Improvement of 70% in the PA for grain yield was observed by including canopy temperature (CT) and normalized difference vegetation index as secondary traits using the MT approach (Rutkoski et al., 2016; Sun et al., 2017). Similarly, Hayes et al., (2017) and Lado et al., (2018) observed an increase in PA using multivariate approaches (MT) over single trait (ST) models in end-use quality traits.

For complex traits, genotype-by-environment interactions (G × E) necessitate the evaluation of breeding lines for multiple traits over multiple environments. Thus, the extension of the MT approaches to account for G × E interaction could improve the model for genomic prediction accuracy in breeding programs. Montesinos-López et al. (2016) proposed a Bayesian multi-trait and multi-environment (BMTME) model that integrates the analysis of multi-traits recorded over multi-environments and accounts for T x G x E interaction in a unified approach. Recently, an improved BMTME model has been introduced that estimates the variance-covariance structure among trait, genotype, and environment to predict multiple traits evaluated in various environments (Montesinos-López et al., 2019). Few studies using simulated and empirical data found

that the BMTME model outperforms ST models in agronomic and end-use quality traits in wheat (Guo et al., 2020; Ibba et al., 2020; Montesinos-López et al., 2016). Better performance of multivariate GS approaches stimulates us to evaluate these models in an actual breeding pipeline, where several traits are evaluated over the diverse environments.

Although different GS approaches have been tested for predicting complex traits in wheat breeding programs, only a few studies have reported the application of GS in actual yield trials where lines are evaluated over several environments (Belamkar et al., 2018). GS has great potential in the early selection or culling in preliminary trials using information from advanced trials and accelerate the genetic gain. Furthermore, GS can complement the phenotypic selection in practical scenarios such as loss of complete/partial trials due to weather extremes. In the present study, we focused on the use of advanced breeding lines evaluated over multiple environments as training sets to predict untested genotypes using univariate and multivariate GS approaches. The specific objectives of this study were to (1) estimate the PA of various agronomic traits in advanced breeding lines using univariate and multivariate GP models and different cross-validation schemes, (2) assess the reliability of multivariate GP models in predicting complex traits over different years and locations, and (3) investigate the application of multi-trait multi-environment GP models in sparse testing of breeding lines.

4.3 Materials and methods

4.3.1 Plant Materials

The experiment was conducted over two growing seasons (2018-19 and 2019-20) using a total of 314 winter wheat genotypes. The genotypes included breeding lines from 2018-

19 and 2019-20 wheat advanced yield trials (AYT) and elite yield trials (EYT) from the South Dakota State University (SDSU) winter wheat breeding program and well-adapted check cultivars. The majority of genotypes were either $F_{4:7}$ and $F_{4:8}$ filial generation. Of the 314 genotypes, 157 were evaluated in the growing season of 2019 and another 157 in 2020. Forty-four genotypes were shared between the two sets of wheat materials, leaving 270 unique genotypes in the study. We removed seven genotypes from genomic prediction analyses attributing to low-quality genotypic data. Thus, 151 and 156 genotypes were used for further analyses in the 2018-19 and 2019-20 growing seasons, respectively.

4.3.2 Experimental Design and Trait Measurement

The experimental plots were planted under no-till system at five locations in South Dakota State (Table 4.1) in both seasons. The experimental unit at each of the five locations consisted of 1.5m wide and 4m long plots with seven rows spaced 20 cm apart. A seeding rate for plots was 300 seeds m^{-2} at all the locations. The recommended agronomic practices were followed for proper growth and yield.

Five agronomic traits measured in this study were grain yield (bushels $acre^{-1}$), grain protein content (%), test weight (kg hL^{-1}), plant height (cm), and days to heading (Julian days). Grain yield (YLD) was weighed after harvesting the plots at maturity using a plot combine (Zurn, Germany). Grain protein content (PROT), test weight (TW), and moisture content were measured using Infratec^{TM} 1241 Grain Analyzer (FOSS North America, USA). Grain yield from plot and grain protein content were adjusted to 13% moisture content equivalence. Plant height (HT) was recorded as the distance from the soil surface to the tip of the fully emerged spike, excluding any awns if present. Days to

heading (HD) were recorded as the Julian days required for 50% of heads to emerge from the boot in each plot.

Table 4.1 Information of the experimental sites used in the growing seasons of 2018-19 and 2019-20.

Site	Coordinates	2018-19		2019-20	
		Date seeded	Date harvested	Date seeded	Date harvested
Brookings (BRK)	44°18'35.3"N 96°40'14.5"W	9/16/2018	8/6/2019	9/20/2019	7/20/2020
Dakota Lakes (DL)	44°17'34.2"N 99°59'40.6"W	9/28/2018	7/23/2019	9/19/2019	7/17/2020
Hayes (HYS)	44°22'24.8"N 101°02'45.1"W	9/14/2018	7/31/2019	9/17/2019	7/21/2020
Onida (OND)	44°42'57.5"N 100°23'04.2"W	9/25/2018	8/1/2019	9/18/2019	7/28/2020
Winner (WIN)	43°29'57.0"N 99°51'58.4"W	10/2/2018	7/25/2019	9/27/2019	7/15/2020

4.3.3 Phenotypic Data Analysis

The phenotypic data for all five agronomic traits were analyzed using best linear unbiased estimates (BLUEs) for individual environments. The model used for estimation of the genotypic BLUEs for individual environments was as follows:

$$y_{ij} = \mu + R_i + G_j + e_{ij}$$

where y_{ij} is the trait of interest, μ is the overall mean, R_i is the effect of the ith replicate, G_j is the effect of the j^{th} genotype, and e_{ij} is the residual error effect associated with the i^{th} replication and j^{th} genotype. The replicates correspond to the complete blocks.

For across environment estimation of BLUEs and best linear unbiased predictions (BLUPs), the statistical model was modified as below:

$$y_{ijk} = \mu + E_i + R_{j(i)} + G_k + GE_{ik} + e_{ijk}$$

where y_{ijk} is the trait of interest, μ is the overall mean, E_i is the effect of the i^{th} environment, $R_{j(i)}$ is the effect of the j^{th} replicate nested within the ith environment, G_k is the effect of the k^{th} genotype, GE_{ik} is the effect of the genotype x environment (G x E) interaction, and e_{ijk} is the residual error effect associated with the i^{th} replication and j^{th} genotype. The environment corresponds to the individual locations and replicates correspond to the complete blocks. The genotype was assumed as a fixed effect, whereas environment and block nested within the environment were assumed as random effects. The broad-sense heritability (H^2) of a trait of interest in an independent environment was assessed as follow:

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

where σ_g^2 and σ_e^2 , are the genotype and error variance components, respectively. The BLUEs and variance components were estimated using META-R (Alvarado et al., 2020), which employs LME4 R-package (Bates et al., 2015) for linear mixed model analysis. The Pearson correlations among traits and environments were estimated based on the BLUEs and BLUPs using the 'psych' package in the R environment (R Core Team, 2018). The genetic correlations between five traits were estimated for individual years using the 'BMTME' R package (Montesinos-López et al., 2019).

4.3.4 SNP Genotyping

Fresh leaf tissues were collected from each line for DNA isolation using the hexadecyltrimethylammonium bromide (CTAB) method (Doyle & Doyle, 1987). Genotyping-by-sequencing (GBS) was performed following the double digestion with HF-*PstI* and *MspI* restriction enzymes for library preparation (Poland et al., 2012). GBS libraries were sequenced using an Ion Proton sequencer (Thermo Fisher Scientific, Waltham, MA, USA) at the USDA Central Small Grain Genotyping Lab, Manhattan, KS, USA. TASSEL v5.0 was used to call single-nucleotide polymorphisms (SNPs) using the GBS v2.0 discovery pipeline (Bradbury et al., 2007). The reads were aligned to the Chinese Spring wheat genome reference RefSeq v1.1 (IWGSC, 2018) using the default settings of Burrows-Wheeler Aligner v0.6.1.

For quality control, SNPs with more than 20% missing data points and minor allele frequency (MAF) of less than 0.05 were removed. Additionally, we obtained 10,290 high-quality SNPs after removing the SNPs that were unmapped on any wheat chromosome. The missing data points in the selected SNP set were imputed using BEAGLE v4.1 (Browning & Browning, 2007). The additive relationship matrix for GP

models was estimated using the *A.mat* function in the 'rrBLUP' package in R (Endelman, 2011). The Kinship (K)-based marker matrix was estimated using the Centered IBS (identity by state) method (Endelman & Jannink, 2012) implemented through Genomic Association and Prediction Integrated Tool (GAPIT) (Tang et al., 2016).

4.3.5 Genomic Prediction Models and Cross-validation

We evaluated one univariate and two multivariate GP models for predicting five agronomic traits. Different cross-validation schemes that mimic actual scenarios in a breeding program were used to estimate the PA of these traits and compare the performance of different models.

Single-trait model

The ridge regression (rrBLUP) model (Endelman, 2011) is the commonly used GS model in plant breeding. Like the genomic best linear unbiased prediction (GBLUP) model, rrBLUP assumes the normal distribution of marker effects with equal variance. We used rrBLUP as a baseline GS model for all the traits to evaluate the performance of multivariate models. The within-environment trait BLUEs were calculated and then used as input to perform rrBLUP within each environment. A linear mixed model was implemented using the following model:

$$y = 1\mu + Zu + \varepsilon$$

where y is the vector ($n \times 1$) of adjusted means (BLUEs) from n genotypes for a given trait; μ is the overall mean; Z is the design matrix $(n \times p)$ with known values of p markers

for *n* genotypes; *u* is a genotypic predictor with $u \sim N(0, G_{nxn}\sigma_g^2)$, where *G* is positive semidefinite matrix, obtained from markers using *'A.mat'* which is an additive relation matrix function and σ_g^2 is the additive genetic variance; ε is the residual error with $e \sim N(0, \sigma_e^2)$.

Multi-trait model

A Bayesian Multivariate Gaussian model with an unstructured variance-covariance matrix was used for the multi-trait model (MT) (Lado et al., 2018). The MT model can be described as:

$$y = 1\mu + Zu + \varepsilon$$

where *y* is the vector with a length of $n \times t$ (*n* genotypes and *t* traits); μ is the means vector; *Z* represents the incidence matrix of order $[(n \times t)p]$, ; $u_{[(n \times t)p]}$ is genotypic predictor for all individuals and traits with $u \sim N(0, \sum \otimes G)$. The matrix *G* represents the positive semidefinite matrix obtained from markers. The residuals of the MT model are represented by the vector ε, with $\varepsilon \sim N(0, R \otimes I)$. The matrices $\sum$ and *R* are the variance-covariance matrices for depicting the genetic and residual effects for each individual in all traits, respectively, estimated by the Gibbs sampler with 5,000 burn-in and 25,000 iterations in R package 'MTM' (de los Campos & Grüneberg, 2016). $\sum$ was estimated as an unstructured matrix and R as a diagonal matrix following Lado et. al., 2018.

Bayesian multi-trait multi-environment model

The Bayesian multi-trait multi-environment (BMTME) model for genomic predictions (Montesinos-López et al., 2016, 2019) can be briefly described as:

$$y = X\beta + Z_1 b_1 + Z_2 b_2 + \varepsilon$$

where y is the response matrix of order $j \times t$ (where t is the number of traits and $j = n \times l$, where n denotes the number of genotypes and l denotes number of environments); X is the design matrix for environmental effects of order $n \times l$, whereas β is the matrix of beta coefficients of order $l \times t$. Z_1 is the incidence matrix of genotypes of order $j \times n$, and b_1 is the matrix of genotypic random effects of order $n \times t$. Z_2 is the incidence matrix of genotype × environment interaction of order $j \times ln$ and b_2 is the random effect of genotype × environment × traits of order $ln \times t$. We assume b_1 is distributed under a matrix variate normal distribution as $b_1 \sim MN\ (0, G, \sum_t)$, where G is of order $n \times n$, obtained from SNP markers using '*A.mat*' which is an additive relation matrix function in rrBLUP, and is the $\sum_t$ is the unstructured variance-covariance matrix of traits of order $t \times t$. The b_2 is assumed to be distributed under matrix variate normal distribution as $b_2 \sim MN\ (0, \sum_E \otimes G, \sum_t)$, where $\otimes$ denotes Kronecker product and $\sum_E$ is the unstructured variance-covariance matrix of $l \times l$. The matrix ε is the matrix of residuals of order $j \times t$ distributed as $\varepsilon \sim MN\ (0, l_j, R_e)$. A detailed account of this model and prior distributions can be found in Montesinos-López et al. (2019). Model simulations were carried out using the R package 'BMTME' (Montesinos-López et al., 2019) with 5,000 burn-in and 25,000 iterations.

Assessment of prediction ability

Predictive ability was estimated as Pearson correlation coefficient between GEBVs and observed phenotypes for the testing set of breeding lines. The PA for the rrBLUP model was estimated using a cross-validation scheme 1 (CV1), where the population was equally divided into five subpopulations, with four subpopulations (80%) as the training population (phenotyped and genotyped) to train the model and one subpopulation (20%) as the testing population (genotyped only) for prediction. The single-trait model with cross-validation scheme 1 (designated as ST-CV1 hereafter) was implemented in the 'rrBLUP' R package (Endelman, 2011) for one trait at a time. The cross-validation process was repeated 1,000 times, and each iteration included different lines in the training and testing sets.

The prediction accuracy of the MT model was estimated using two cross-validation schemes as described in Lado et al. (2018) (Figure 4.1). Similar to the ST-CV1 scheme, the first cross-validation scheme (MT-CV1) used a random set of lines (80%) as a training set and the remaining lines (20%) as a testing set. The model was trained using genotypic and phenotypic data of these lines in the training set, and only genotypic data were used to predict the performance of the testing set lines based on the model built from the training set. This process of splitting the data into training and testing sets was repeated 50 times. Hence, a different set of lines were selected into the training and testing dataset for each iteration. The CV1 scheme mocks the breeding situation where a set of lines that are evaluated for given traits could be used to predict an unphenotyped set of lines that only have genotypic information. In the second cross-validation scheme (MT-CV2), the lines were randomly split into a training set (80%) and a testing set (20%). To train the model, MT-CV2 used genotypic data and phenotypic data of

secondary traits from both the training and testing sets, but the phenotypic data of the target trait (primary trait) only from the training set. The BMTME model used a cross-validation scheme similar to MT-CV1 to estimate the model's PA by randomly splitting the lines into 80% training set and 20% testing set. Since the BMTME model employs a Gibbs sampler with multiple iterations and is computationally expensive, the cross-validation scheme was repeated only 25 times.

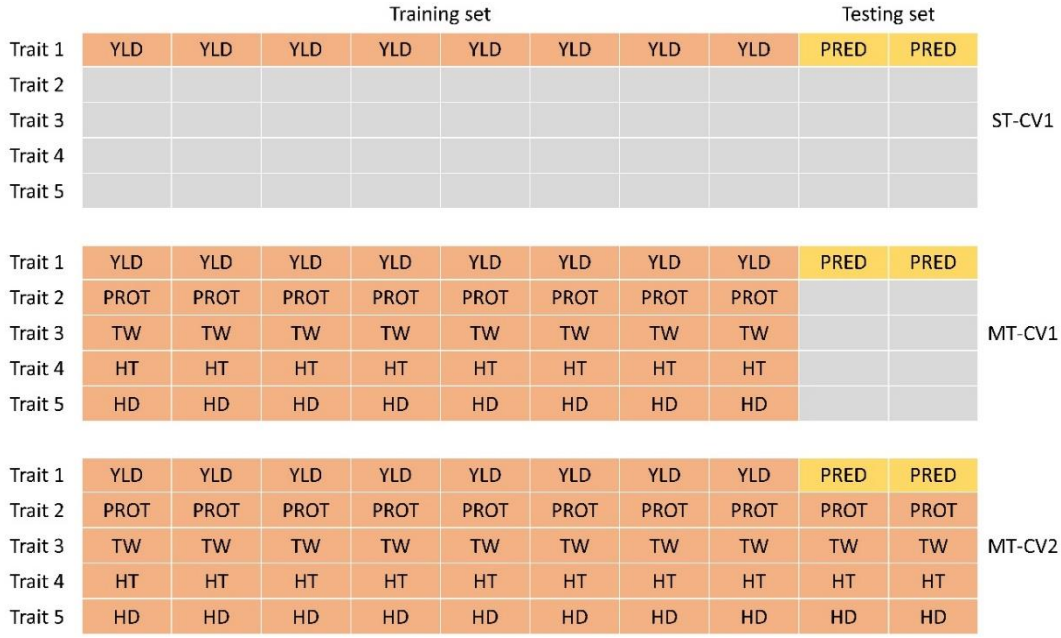

Figure 4.1 Illustration of different cross-validation schemes used to evaluate different genomic prediction models.

4.3.6 Application of MTME genomic prediction in the breeding program

As the MTME model showed promise in predicting different agronomic traits using a cross-validation approach, we evaluated the possible application of this method in our breeding program to reduce the phenotyping efforts and per-plot costs. As discussed earlier, we evaluate ~40 elite lines and ~110 advanced lines each year under multiple environments. The per-plot costs and phenotyping efforts could be reduced if we can

successfully determine genomic estimation of breeding values (GEBVs) of the advanced lines at fewer locations rather than testing these lines at all available locations. The MTME model can estimate the environmental effect based on elite lines evaluated at all locations and genotypic effect of advanced lines from fewer locations. To test this, we used MTME model in an allocation design where we used the phenotypic data of elite lines from five tested environments; however, we used phenotypic records of advanced lines from three environments only. We predicted five traits in the remaining two environments in both the growing seasons. The model was fitted using the R package 'BMTME' (Montesinos-López et al., 2016, 2019) with 5,000 burn-in and 15,000 iterations. The observed phenotypic records from the remaining two environments were used to assess the predictive accuracy of the design.

4.4 Results

4.4.1 Descriptive Statistics

The phenotypic BLUEs for grain yield, grain protein content, test weight, plant height, and days to heading varied significantly among different environments (Table 4.2). HYS produced the highest mean grain yield in both years, whereas BRK and WIN produced the lowest grain yield in 2018-19 and 2019-20, respectively. Broad-sense heritability (H^2) was estimated for all five agronomic traits in each environment (Table 4.2). Differences in heritability estimates (0.63 to 0.96) describe the different genetic architecture of traits and contrasting environmental effects. Among the five traits evaluated in the study, test weight, plant height, and days to heading had moderate to high heritability values in most environments and over both years. Relatively, grain yield (0.64 – 0.84) and grain protein content (0.63 – 0.96) had comparatively lower heritability than other traits. Among the

five environments, heritability for all the traits was high in both experimental years in DL. For grain yield heritability, HYS (2019-20) had the highest (0.84), whereas BRK (2019-20) had the lowest (Table 4.2).

Pearson correlations among agronomic traits were calculated using BLUEs by combining phenotypic data from all environments in each of the two growing seasons (Figure 4.2). As expected, significant negative correlation values (-0.28 and -0.54) were observed between grain yield and grain protein content in both years. Grain yield was also negatively correlated with days to heading (in both years) and plant height (2019-20) (Figure 4.2). Similarly, test weight was positively correlated with grain protein content and plant height in both growing seasons. Genetic correlations between five traits were estimated by fitting the BMTME model for individual growing seasons and presented in Tables 4.3 and 4.4. Similar to the phenotypic correlation estimates, we observed a higher genetic correlation in 2019-20 compared to 2018-19.

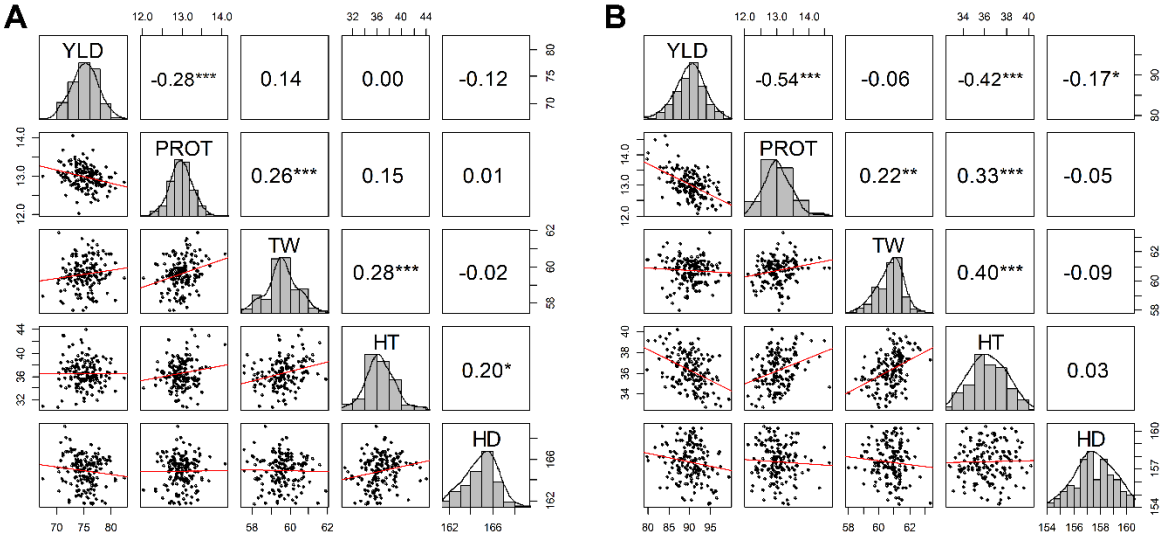

Figure 4.2 Scatter plot matrix with phenotypic distributions and Pearson correlations between agronomic traits using best linear unbiased predictions (BLUPs) by combining

five experimental sites (BRK, DL, HYS, OND, and WIN) (A) from the growing season of 2018-19 and (B) from the growing season of 2019-20. YLD, grain yield; PROT, grain protein content; TW, test weight; HT, plant height; and HD, days to heading.

Table 4.2. Trait descriptive statistics and broad-sense heritability estimate for individual site-year environments for lines grown over five locations (Env) in 2018-19 and 2019-20 growing seasons. BRK, Brookings; DL, Dakota Lakes; HYS, Hayes; OND, Onida; and WIN, Winner. 19, the growing season of 2019-19; and 20, the growing season of 2019-20.

Year	Env[a]	Yield (bu ac^{-1})			Protein content (%)			Test weight (kg hL^{-1})			Plant height (cm)			Days to heading (j days)		
		GM[b]	CV	H^2	GM	CV	H^2	GM	CV	H^2	GM	CV	H^2	GM	CV	H^2
2018-19	BRK	64.69	8.96	0.80	12.16	5.45	0.69	56.34	1.42	0.91	94.78	4.25	0.89	163.22	0.48	0.92
	DL	77.44	6.48	0.77	14.25	1.55	0.94	59.90	1.35	0.78	86.06	3.99	0.89	164.65	0.76	0.92
	HYS	81.98	6.46	0.73	12.04	3.66	0.72	59.48	1.06	0.92	99.12	3.18	0.90	163.84	0.70	0.74
	OND	71.21	7.27	0.76	13.25	3.30	0.85	60.77	1.27	0.82	89.91	3.62	0.91	168.73	0.65	0.87
	WIN	81.27	5.89	0.79	13.17	4.27	0.63	61.48	1.00	0.88	93.63	2.90	0.95	164.19	0.80	0.89
2019-20	BRK	84.26	6.25	0.64	12.49	3.65	0.80	60.13	0.90	0.89	86.67	4.42	0.75	156.18	0.63	0.89
	DL	93.31	4.14	0.78	13.55	1.40	0.96	61.46	0.64	0.95	85.80	3.45	0.83	155.74	0.40	0.94
	HYS	96.64	4.66	0.84	13.87	2.02	0.90	60.03	1.30	0.91	102.8	3.95	0.82	159.36	0.51	0.85
	OND	92.21	4.40	0.81	11.99	4.97	0.59	61.11	1.08	0.87	92.53	3.59	0.85	157.09	0.63	0.87
	WIN	84.16	4.73	0.80	13.24	2.99	0.84	60.79	0.90	0.89	92.70	3.47	0.85	158.75	0.63	0.91

[a]: Env, refers to different trial location. BRK, Brookings; DL, Dakota Lakes; HYS, Hayes; OND, Onida; and WIN, Winner.
[b]: GM, general mean for respective trait; CV, coefficient of variation; H^2, broad sense heritability

Table 4.3. Genetic correlation between five agronomic traits evaluated in 2018-19 estimated using the BMTME model. Evaluated traits include grain yield (YLD); grain protein content (PROT); test weight (TW); plant height (HT); and days to heading (HD).

Trait	**YLD**	**PROT**	**TW**	**HT**	**HD**
YLD	1	-0.15	0.29	0.00	-0.13
PROT	-0.15	1	0.35	-0.01	0.09
TW	0.29	0.35	1	-0.02	0.07
HT	0.00	-0.01	-0.02	1	-0.01
HD	-0.13	0.09	0.07	-0.01	1

Table 4.4. Genetic correlation between five agronomic traits evaluated in 2019-20 estimated using the BMTME model. Evaluated traits include grain yield (YLD); grain protein content (PROT); test weight (TW); plant height (HT); and days to heading (HD).

Trait	**YLD**	**PROT**	**TW**	**HT**	**HD**
YLD	1	-0.44	-0.14	-0.43	-0.18
PROT	-0.44	1	0.25	0.38	-0.14
TW	-0.14	0.25	1	0.39	-0.09
HT	-0.43	0.38	0.39	1	0.18
HD	-0.18	-0.14	-0.09	0.18	1

We further estimated the Pearson correlations among the five environments in 2018-19 and 2019-20 using data of all five agronomic traits (Figure 4.3). Significantly higher correlation values were observed for grain yield among five environments in

2019-20 than those in 2018-19. A similar trend was observed for grain protein content, test weight, and days to heading; however, correlations were comparable for plant height among the two growing seasons (Figure 4.3). Moreover, the principal component analysis (PCA) on grain yield validated strong correlations among testing locations, in particular between HYS and OND and between DL and WIN, in the 2019-20 growing season (Figure 4.4); however, only a weak correlation was observed between DL and BRK in the 2018-19 growing season. The varying degrees of correlation among locations in different growing seasons provide an opportunity to compare the performance of MTME model in different growing environments.

4.4.2 Genetic Relationship Among Lines

The kinship-based marker relationship matrix was derived using 10,290 SNPs from 151 lines evaluated in the 2018-19 growing season and 156 lines evaluated in the 2019-20 growing season (Figure 4.5). The relationship matrix's positive values signify an increased likelihood of the allele from one line being detected in the other lines. The heatmaps of both the relationship matrices elucidate several small groups of closely related individuals over both the growing seasons. Most of the lines seems genetically related to several other lines. However, the heatmaps did not reveal any large genetically structured sub-populations in either set of 151 or 156 lines, respectively. Thus, the absence of a strong structure suggests no advantage of using stratified sampling for the cross-validation schemes to estimate the prediction accuracy. Furthermore, the density of heatmaps revealed a closer relationship among 156 lines evaluated in 2019-20 (Figure 4.5A) than among 151 lines evaluated in 2018-19 (Figure 4.5B).

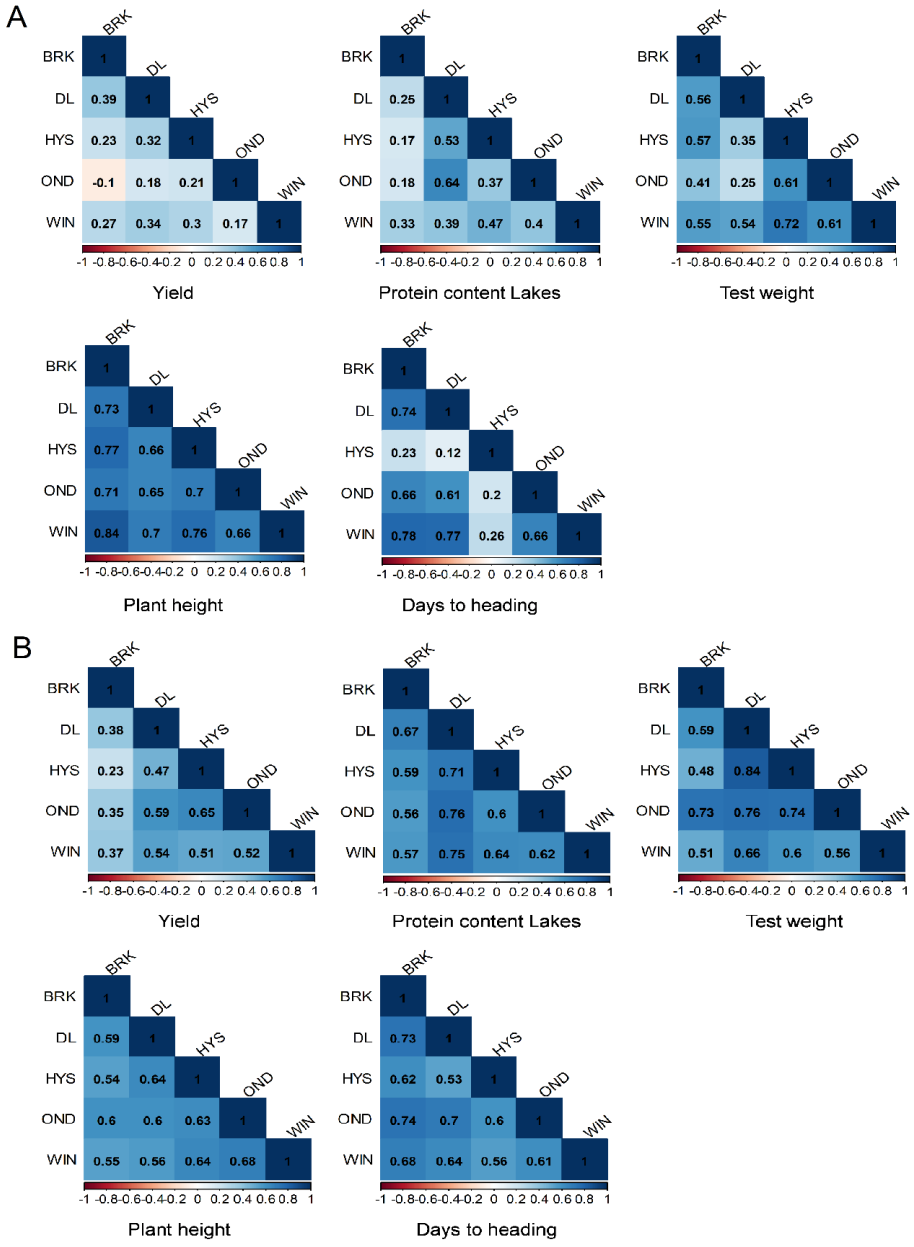

Figure 4.3 Correlation coefficients among five environments (Brookings, BRK; Dakota Lakes, DL; Hayes, HYS; Onida, OND; and Winner, WIN) for five traits evaluated in (A) 2018-19 and (B) 2019-20.

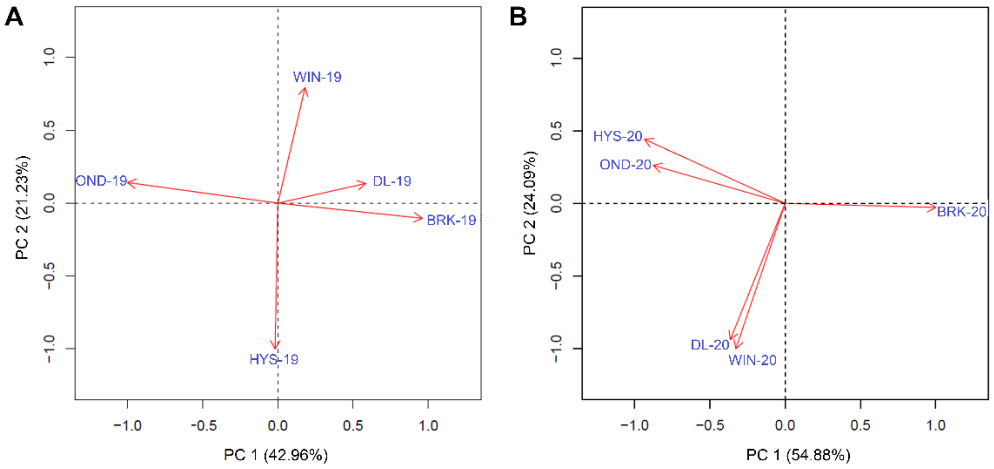

Figure 4.4 Principal component analysis to determine the association of the observed grain yield among five different experimental sites in the 2018-19 growing season (A) and the 2019-20 growing season (B). BRK, Brookings; DL, Dakota Lakes; HYS, Hayes; OND, Onida; and WIN, Winner.

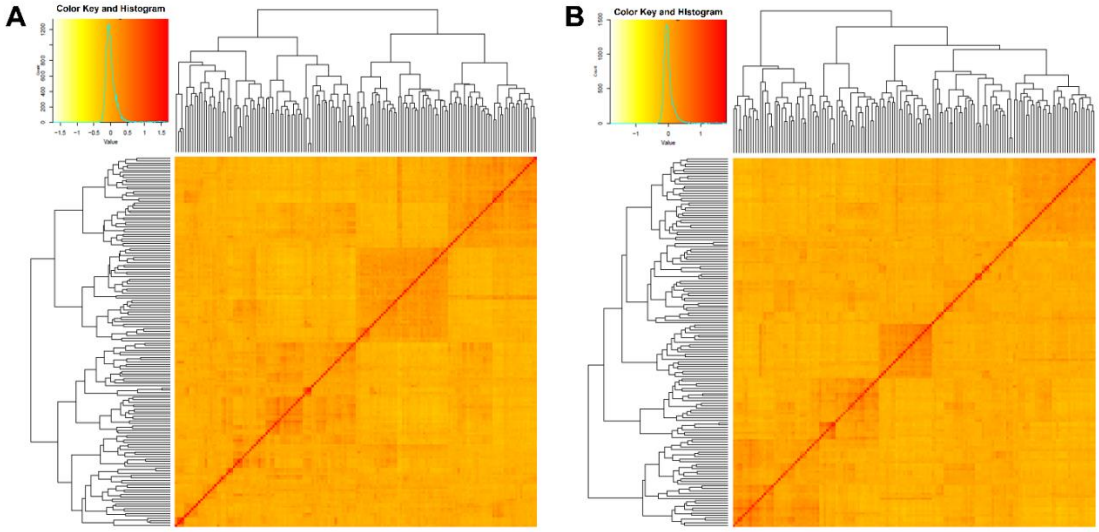

Figure 4.5 Heatmap of the kinship matrix using 10,294 SNPs (A) for 151 lines evaluated in the growing season of 2018-19, and (B) for 156 lines evaluated in the growing season of 2019-20.

4.4.3 Genomic prediction using 2018-19 and 2019-20 datasets

We compared the predicted performance of five traits among four different approaches using two data sets (2018-19 and 2019-20). The PA of various models for five traits is presented in Appendices 4.1 and 4.2. The ST-CV1 model was used as a baseline model to compare the performance of different multivariate models. In 2018-19, the mean PA using ST-CV1 was 0.31, 0.35, 0.36, 0.35, and 0.36 for grain yield, grain protein content, test weight, plant height, and days to heading (Figure 4.6). Slightly better performance was observed in 2019-20 where ST-CV1 yielded an average PA of 0.36, 0.35, 0.54, 0.33, and 0.35 for these traits, respectively. The multi-trait model was tested using two prediction scenarios, MT-CV1 and MT-CV2. The MT-CV1 model did not show improvement in the PA over ST-CV1 for any of the five traits in either growing season (Appendices 4.1 and 4.2).

Multi-trait model, MT-CV2, that includes phenotypic data for secondary agronomic traits from individuals to be predicted showed an overall higher prediction accuracy for grain yield in both growing seasons. In 2018-19, the PA for grain yield using the MT-CV2 model ranged from 0.15 to 0.56, outperforming the single-trait (ST-CV1) model by an average of 26% (Appendices 4.1 and 4.2). Similarly, the mean PA for grain yield in 2019-20 using MT-CV2 was 0.59, showing 63% improvement over the ST-CV1 model. The best PA for grain yield in 2019-20 was observed in HYS (0.71), followed by WIN (0.67) and DL (0.57). The improvement in PA over ST-CV1 reached up to 148% in WIN and 80% in BRK in 2019-20.

Likewise, we observed marginal to moderate improvement in PA for other agronomic traits using MT-CV2 model in both the growing seasons (Figures 4.6 and 4.7;

Appendices 4.1 and 4.2). In 2018-19, the mean PA using MT-CV2 was 0.40, 0.42, 0.34, and 0.38 for grain protein content, test weight, plant height, and days to heading exhibiting an improvement of 14%, 19%, 36%, and 8%, respectively. In comparison, the PA using MT-CV2 was higher in 2019-20, with average PA of 0.54, 0.59, 0.43, and 0.38 for grain protein content, test weight, plant height, and days to heading with an improvement of 54%, 9%, 30%, and 8%, respectively. Overall, the better performance of MT-CV2 model can be attributed to the higher genetic correlation between traits evaluated in 2019-20 over the 2018-19 season (Tables 4.3 and 4.4).

The multi-trait multi-environment MTME model generalizes the multi-trait model to consider the correlation between environments on top the genetic correlation between traits. In 2018-19, the MTME model did not show significantly different PA over the ST-CV1 model for grain yield (0.18 – 0.36) and grain protein content (0.13 – 0.46). The performance of MTME model for these two traits likely relates to the lower genetic trait-correlations and lower correlation between environments for these traits in 2018-19 (Figure 4.3). Analogous to grain yield and grain protein, MTME model resulted in higher prediction accuracy than the ST-CV1 model for the test weight, plant height, and days to heading in 2018-19 (Figure 4.6). For instance, the average PA using MTME for test weight, plant height, and days to heading was 0.42, 0.42, and 0.36, which translates to an improvement of 19%, 68%, and 12%, respectively. Furthermore, the PA using MTME model outstripped ST-CV1 model in all five environments for test weight (0.32 – 0.52) and plant height (0.41 – 0.54), while four environments for days to heading (Figure 4.6).

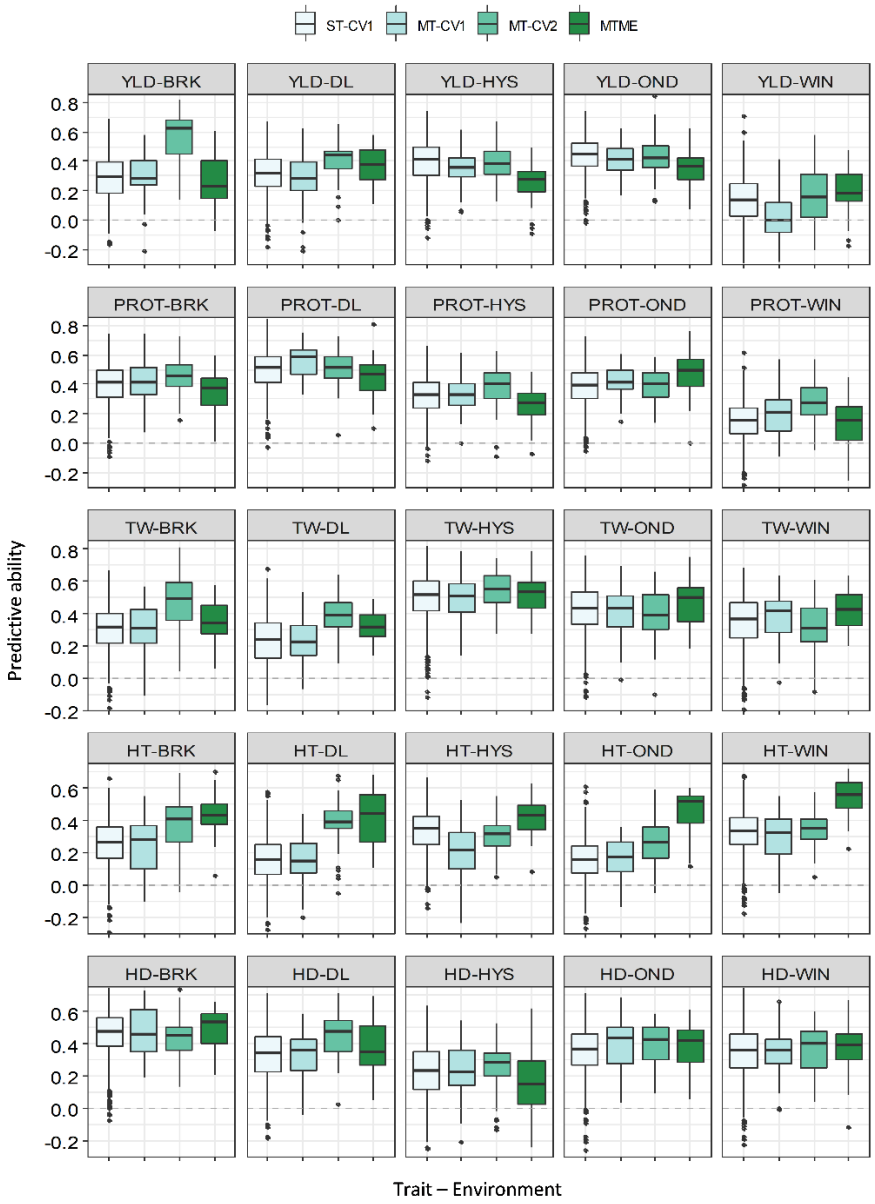

Figure 4.6 The predictive ability (PA) for five agronomic traits evaluated at five environments in the growing season of 2018-19. Boxplots compare the PA using a single-trait prediction model with one cross-validation scheme (ST-CV1), a multi-trait prediction model with two cross-validation schemes (MT-CV1 and MT-CV2), and a Bayesian multi-trait multi-environment prediction model (MTME). Traits include YLD, grain yield; PROT, grain protein content; TW, test weight; HT, plant height; and HD, days to heading.

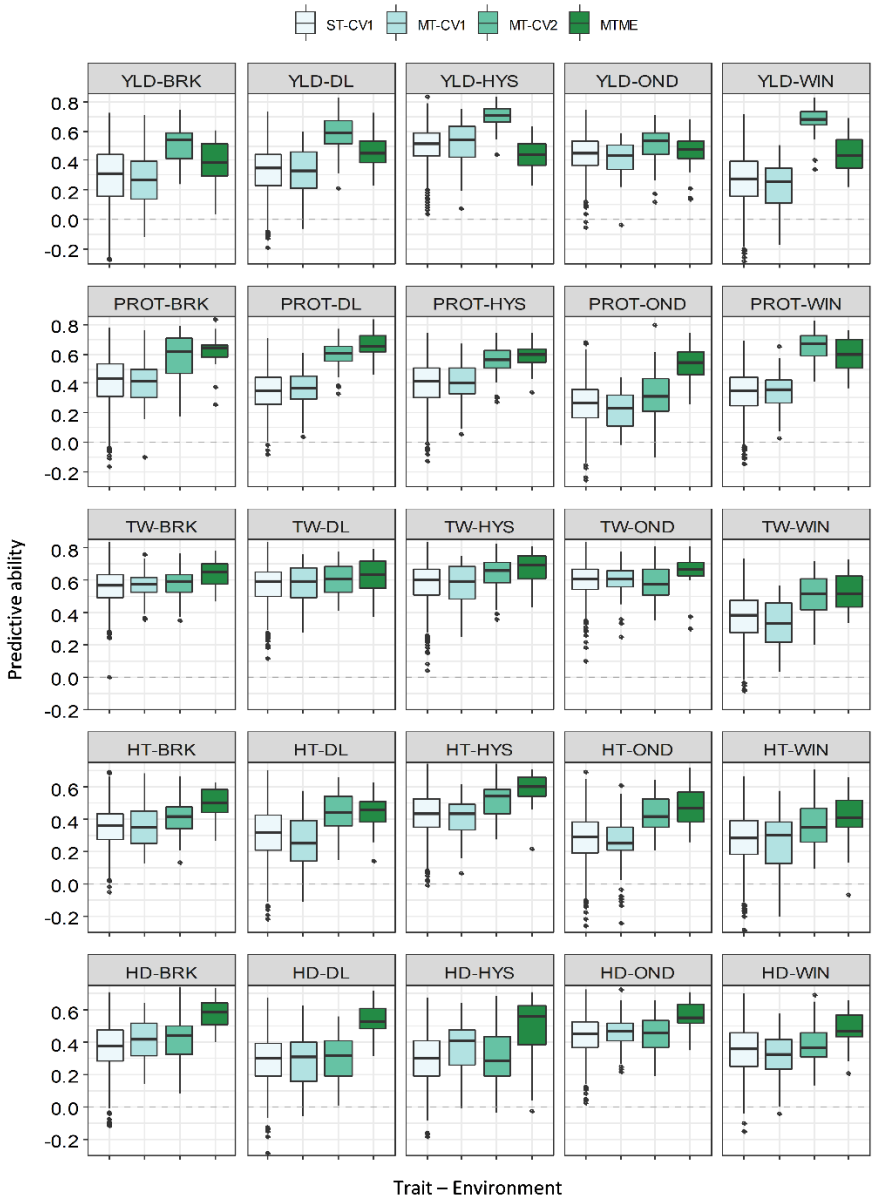

Figure 4.7 The predictive ability (PA) for five agronomic traits evaluated at five environments in the growing season of 2019-20. Boxplots compare the PA using a single-trait prediction model with one cross-validation scheme (ST-CV1), a multi-trait prediction model with two cross-validation schemes (MT-CV1 and MT-CV2), and a Bayesian multi-trait multi-environment prediction model (MTME). Traits include YLD, grain yield; PROT, grain protein content; TW, test weight; HT, plant height; and HD, days to heading.

In contrast 2018-19, we observed higher genetic correlations between five traits and higher environmental correlations in 2019-20 (Tables 4.3 and 4.4; Figure 4.3). As a result of high correlation values, we observed a consistent improvement in the PA of MTME in all the environments for all five traits (Figure 4.7). For grain yield, the MTME model also performed better than the single-trait model in most environments, except HYS. The average PA for grain yield using MTME model was 0.43, which was 22% better than the ST-CV1 model. Further, MTME model appeared to be superior for predicting grain protein content and test weight (Figure 4.7). For grain protein content, the MTME model performed best in all locations, with PA ranging from 0.52 – 0.67. We achieved an improvement in prediction accuracy of up to 100% (OND) using the MTME model (0.52) over the single-trait model (0.26) with 71% improvement on average. The PA for test weight was higher using the MTME model than other models, ranging from 0.53 to 0.67, with a mean improvement of 17% over ST-CV1 model. Similarly, the average PA of the MTME model was the highest for plant height (0.49) and days to heading (0.53), which outstrips the ST-CV1 model by 48% and 51%, respectively.

4.4.4 Application of MTME model in the breeding program

Based on the cross-validation results, we evaluated the efficacy of MTME model in reducing phenotypic efforts in our breeding program. We used the MTME model in estimating GEBV of advanced lines in environments where only elite lines are evaluated. In the tested allocation design, we used phenotypic data of EYTs from five environments and AYTs from three environments to predict GEBVs of AYTs in remaining environments (Figure 4.8). Two environments, OND and WIN, were used as testing environments for predicting AYTs. In 2018-19, we predicted the performance of 96 AYT

lines, whereas 2019-20 comprised a prediction of 114 AYT lines in two environments. Table 4.5 elucidates the predictive ability for five agronomic traits using MTME in an independent prediction scenario. Moderate PA was observed for all the traits in both environments except for WIN in 2019-20. For OND, results showed better prediction accuracy than WIN for grain yield and test weight. Overall, the results suggest that the MTME model could be used by evaluating an overlapping set of lines over multiple environments and lines in early testing could be tested in fewer environments.

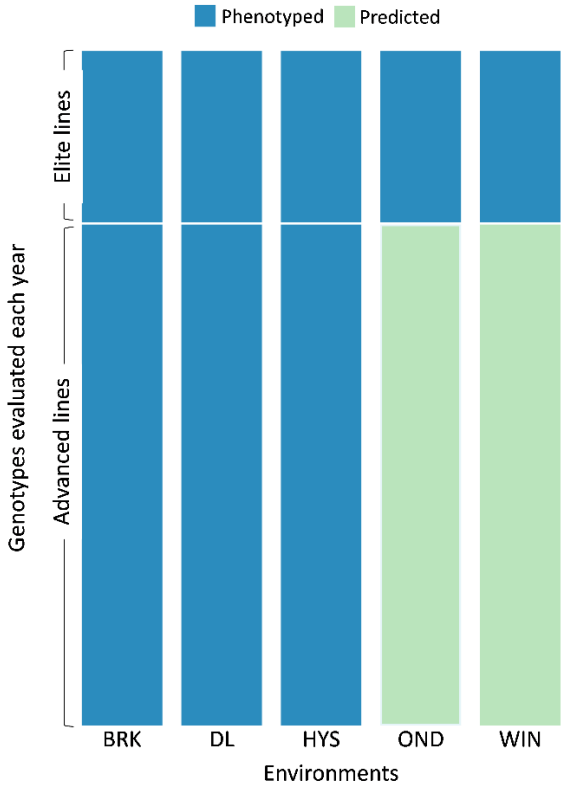

Figure 4.8 Testing design for independent prediction of agronomic traits using the MTME model. Each year a set of elite and advanced lines is evaluated over multiple locations. The sparse testing design proposes phenotyping of elite lines in all environments (five in this scenario) and advanced lines in fewer environments (three in this scenario). For independent prediction, the dataset from 2018-19 comprised 55 elite

lines with checks and 96 advanced lines. The 2019-20 dataset comprised 42 elite lines with checks and 114 advanced lines. Five environments: BRK, Brookings; DL, Dakota Lakes; HYS, Hayes; OND, Onida; and WIN, Winner.

Table 4.5 Predictive ability for independent prediction of advanced breeding lines (AYTs) in new environments using MTME model. Tables shows Pearson correlation between the observed and predictive values of agronomic traits in the AYTs at two different environments over two growing seasons.

Year	**Env[a]**	**Predictive ability[b]**				
		Grain yield	**Grain protein**	**Test weight**	**Plant height**	**Days to heading**
2018-19	OND	0.44	0.37	0.43	0.49	0.27
	WIN	0.30	0.25	0.38	0.30	0.46
2019-20	OND	0.36	0.27	0.44	0.22	0.41
	WIN	0.15	0.32	0.25	0.18	0.24

[a]: Env, refers to different trial location. BRK, Brookings; DL, Dakota Lakes; HYS, Hayes; OND, Onida; and WIN, Winner.
[b]: The predictive ability for five agronomic traits using MTME model in independent prediction of advanced lines. Refer to Figure 4.8 for design of the prediction scheme.

4.5 Discussion

In recent years, genomic prediction has been intensively evaluated in wheat breeding programs to select and advance lines for several traits of interest (Haile et al., 2020; Juliana et al., 2020; Rutkoski et al., 2016; Rutkoski et al., 2014). However, improving the prediction accuracy of complex traits remains a challenge for successfully implementing GS in breeding programs. The choice and optimization of the statistical models are crucial to improve the performance of GS. Most plant breeding programs currently rely on univariate genomic prediction models to target a single trait at a time. An advantage of multivariate prediction approaches over single-trait models that have been demonstrated in some recent studies is utilizing correlations between multiple traits and environments (Ibba et al., 2020; Jia & Jannink, 2012; Lado et al., 2018; Sun et al., 2017; Ward et al., 2019). This study evaluated the application of multi-trait and multi-environment prediction models to predict five key traits of varying genetic architecture across diverse environments in a breeding program.

The ridge-regression best linear unbiased prediction (rrBLUP) is one of the most often used single-trait prediction models. The rrBLUP has an advantage over Bayesian models in predicting complex traits governed by several loci with small effects (Lorenz et al., 2011). We used rrBLUP as a baseline model (ST-CV1) to compare with different multivariate approaches. The PA for agronomic traits using ST-CV1 was comparable with other studies using the same model (Charmet et al., 2014; He et al., 2016; Maulana et al., 2021; Pérez-Rodríguez et al., 2012). For instance, the PA for grain yield were between 0.13 – 0.43 for 2018-19 and 0.27 – 0.50 for 2019-20. The PA for test weight in

both growing seasons was higher than the PA for other traits due to the highly heritable nature of this trait (Figures 4.6 and 4.7).

We evaluated the multi-trait model using two cross-validation schemes. The first scheme (MT-CV1) conducts multi-trait prediction for new un-phenotyped individuals, and the testing set has not been phenotyped for any of the traits. In the second cross-validation scheme (MT-CV2), phenotype information for the predicted trait is missing, whereas phenotype information for the secondary traits is available in the testing set (Bhatta et al., 2020; Lado et al., 2018). In our study, prediction accuracy of the MT-CV1 model was comparable to the ST-CV1 model for most of the trait-environment combinations in both growing seasons (Appendices 4.1 and 4.2). Several studies have reported marginal or no improvement with the MT-CV1, where information from secondary traits is limited to the training set (Arojju et al., 2020; Bhatta et al., 2020; Calus & Veerkamp, 2011; Lado et al., 2018; Schulthess et al., 2018). However, other studies reported an improvement in genomic prediction when the MT-CV1 model included secondary traits with moderate-high heritability (Guo et al., 2014; Jia & Jannink, 2012; Rutkoski et al., 2012). Jia and Jannick (2012) suggested that the MT-CV1 approach might be more useful when the primary trait has very low heritability ($H^2 < 0.2$). In the current study, similar performance of MT-CV1 and ST-CV1 models might be contributed by the moderate to high heritability estimated for most of the traits and the small size of the training population.

In the current study, the MT-CV2 significantly improved the PA for all agronomic traits in all the environments, suggesting that inclusion of secondary traits in the training and testing sets improves the predictive performance for complex traits

(Appendices 4.1 and 4.2). Several studies have reported a similar improvement in the prediction using the MT-CV2 model for agronomic and end-use quality traits in wheat (Lado et al., 2018; Rutkoski et al., 2016; Sun et al., 2017), rice (X Wang et al., 2017), barley (Bhatta et al., 2020), sorghum (Fernandes et al., 2018), and ryegrass (Arojju et al., 2020). The MT-CV2 model outperformed single-trait model for grain yield prediction in all environments. However, the extent of improvement using MT-CV2 model varied with traits and environments tested.). As the multi-trait models rely on the genetic correlation between traits (Calus & Veerkamp, 2011; Jia & Jannink, 2012), the differences in prediction improvements due to the MT-CV2 model can be attributed to the varying degrees of genetic correlations observed in different environments. We observed a high genetic correlation among traits in 2019-20 that resulted in higher prediction accuracy for different traits in this growing season (Figure 4.2; Tables 4.3 and 4.4). Our results suggest that MT-CV2 could likely be very useful if we can include data for plant height, days to heading, and other spectral indices recorded using a high throughput method for predicting grain yield. In addition, MT-CV2 approach could be really useful to predict hard to phenotype end-use quality traits by inclusion of already available agronomic data for the testing set.

We also evaluated the BMTME model (referred to as MTME) that generalizes multi-trait model to consider the correlations among multiple environments. Recently, two studies reported an increase in the prediction accuracy of agronomic and end-use quality traits in wheat using the BMTME approach (Guo et al., 2020; Ibba et al., 2020). Due to different training process, we did not directly compare the MTME model with the MT-CV2 model but compared both against the ST-CV1 model. In 2018-19, the MTME

model proved to be better than the ST-CV1 and MT-CV1 models for all traits except yield and grain protein. However, the MTME model outperformed the ST-CV1 and MT-CV1 models in 2019-20 for all traits in all the environments. The mean improvement in PA (across five environments) using MTME model over the ST-CV1 reached up to 19%, 71%, 17%, 48%, and 51% for grain yield, grain protein content, test weight, plant height, and days to heading, respectively. The differences in performance of the MTME model in 2019-20 compared to 2018-19 relate to the observed genetic correlations among traits as well as among environments in these growing season. As discussed earlier, the genetic correlations between traits and correlation among environments were higher in 2019-20 as compared to 2018-19 were higher in 2019-20 as compared to 2018-19.Thus, a higher PA was observed for the traits showing high correlation among different environments. For example, five environments were highly correlated for grain protein content (0.56 - 0.76) compared to grain yield (0.23 – 0.65), explaining the difference in improvement of PA for these traits. Overall, our results suggest that the MTME could be successfully applied in a program if there is a moderate to high correlation for a trait between environments and overcome the effect of a small training population.

Apart from the statistical model, heritability (H^2) of a trait is another crucial factor for improving PA (Combs & Bernardo, 2013; Lorenz et al., 2011). Several studies have found that low heritability often results in lower prediction accuracy in single-trait genomic prediction (Heffner et al., 2009; Jannink et al., 2010). The application of multi-trait models can improve the prediction accuracy of low-heritability traits by using the information from correlated traits with high heritability (Bhatta et al., 2020; Jia & Jannink, 2012; Jiang et al., 2015; Lado et al., 2018). The heritability estimates for most of

the traits in different environments were moderate to high in our study, with few exceptions. The use of MT-CV2 model significantly improved the predictive ability for grain protein content in WIN (0.15 to 0.29) and test weight in DL (0.23 to 0.39), where highly heritable and moderately correlated traits were included in the model. In contrast, the MT-CV2 model did not improve the PA for days to heading in HYS (0.23 to 0.25) as the primary trait was weakly correlated to the highly heritable secondary traits in the model. The results suggest that the inclusion of highly heritable but weakly correlated secondary traits in the multi-trait model may not improve the PA.

Genomic prediction has been suggested to implement sparse testing in multi-environment trials and reduce the resources involved in phenotyping (Jarquin et al., 2020). Based on the promising cross-validation results using MTME models, we evaluated the application of this model in our breeding program to reduce the phenotyping resources. At SDSU winter wheat breeding program, we evaluate a set of elites (EYTs) and advanced lines (AYTs) each year in multiple environments. However, our results suggest GP models developed using phenotypic data from all locations of EYTs and limited locations of AYTs can predict AYTs in remaining environments (Table 2). This strategy could be useful as we evaluate ~40 EYTs and ~110 AYTs each year in replicated nurseries and testing the AYT plots at two/three locations instead of five can save substantial resources. Though we used this strategy to predict AYTs at two locations, further improved GP models assisted with environics data can help to predict more environments with better accuracy. Moreover, this strategy can be expanded to predict preliminary breeding lines at earlier testing stages.

In conclusion, our study evaluated the PA of univariate and multivariate GP models for five agronomic traits in advanced winter wheat breeding lines. We compared two different cross-validation strategies mocking practical breeding scenarios. Overall, our results supported the practical implementation of multivariate GS models in predicting complex traits. We found a significant advantage of using MT and MTME models when correlated traits and/or environments are included in the models. Our results suggest inclusion of correlated traits and environments in the prediction models can offset the limitation of a small training population, allowing the use of advanced breeding lines to predict preliminary breeding lines in the same year or following year. It will be interesting to further study the inclusion of different combinations of secondary traits in the MT model to increase PA of the grain yield. We envision that evaluation of secondary traits like plant height, tillers/m^2, spike length, and spike density that have high correlations with grain yield using unmanned aerial system (UAS) in winter wheat yield trials could help predict grain yield. This would permit trials on a large number of locations (e.g., > 10) but harvesting only a limited number (e.g., 2-3) of locations. Similarly, evaluating secondary traits (grain protein, flour protein, water absorption, gluten content, and quality) could facilitate predicting other complex traits such as end-use quality. Finally, GS holds tremendous potential for improving the selection accuracy of complex traits in wheat breeding; however, we believe GEBVs will complement the phenotyping efforts rather than replacing them. Future breeding strategies should focus on increasing the efficiency of breeding programs by maximizing the genetic gain.

4.6 References

Alvarado, G., Rodríguez, F. M., Pacheco, A., Burgueño, J., Crossa, J., Vargas, M., Pérez-Rodríguez, P., & Lopez-Cruz, M. A. (2020). META-R: A software to analyze data from multi-environment plant breeding trials. *Crop Journal*, *8*(5), 745–756. https://doi.org/10.1016/j.cj.2020.03.010

Arojju, S. K., Cao, M., Trolove, M., Barrett, B. A., Inch, C., Eady, C., Stewart, A., & Faville, M. J. (2020). Multi-Trait Genomic Prediction Improves Predictive Ability for Dry Matter Yield and Water-Soluble Carbohydrates in Perennial Ryegrass. *Frontiers in Plant Science*, *11*, 1. https://doi.org/10.3389/fpls.2020.01197

Arruda, M. P., Brown, P. J., Lipka, A. E., Krill, A. M., Thurber, C., & Kolb, F. L. (2015). Genomic Selection for Predicting *Fusarium* Head Blight Resistance in a Wheat Breeding Program. *The Plant Genome*, *8*(3). https://doi.org/10.3835/plantgenome2015.01.0003

Bassi, F. M., Bentley, A. R., Charmet, G., Ortiz, R., & Crossa, J. (2015). Breeding schemes for the implementation of genomic selection in wheat (Triticum spp.). *Plant Science*, *242*, 23–36. https://doi.org/10.1016/j.plantsci.2015.08.021

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Battenfield, S. D., Guzmán, C., Gaynor, R. C., Singh, R. P., Peña, R. J., Dreisigacker, S., Fritz, A. K., & 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*(2). https://doi.org/10.3835/plantgenome2016.01.0005

Belamkar, V., Guttieri, M. J., Hussain, W., Jarquín, D., El-basyoni, I., Poland, J., Lorenz, A. J., & Baenziger, P. S. (2018). Genomic selection in preliminary yield trials in a winter wheat breeding program. *G3: Genes, Genomes, Genetics*, *8*(8), 2735–2747. https://doi.org/10.1534/g3.118.200415

Bhat, J. A., Ali, S., Salgotra, R. K., Mir, Z. A., Dutta, S., Jadon, V., Tyagi, A., Mushtaq, M., Jain, N., Singh, P. K., Singh, G. P., & Prabhu, K. V. (2016). Genomic selection in the era of next generation sequencing for complex traits in plant breeding. In *Frontiers in Genetics* (Vol. 7, Issue DEC). Frontiers Media S.A. https://doi.org/10.3389/fgene.2016.00221

Bhatta, M., Gutierrez, L., Cammarota, L., Cardozo, F., Germán, S., Gómez-Guerrero, B., Pardo, M. F., Lanaro, V., Sayas, M., & Castro, A. J. (2020). Multi-trait genomic prediction model increased the predictive ability for agronomic and malting quality traits in barley (Hordeum vulgare L.). *G3: Genes, Genomes, Genetics*, *10*(3), 1113–1124. https://doi.org/10.1534/g3.119.400968

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics*, *23*(19), 2633–2635. https://doi.org/10.1093/bioinformatics/btm308

Browning, S. R., & Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics*, *81*(5), 1084–1097.

https://doi.org/10.1086/521987

Calus, M. P. L., & Veerkamp, R. F. (2011). Accuracy of multi-trait genomic selection using different methods. *Genetics Selection Evolution*, *43*(1), 26. https://doi.org/10.1186/1297-9686-43-26

Charmet, G., Storlie, E., Oury, F. X., Laurent, V., Beghin, D., Chevarin, L., Lapierre, A., Perretant, M. R., Rolland, B., Heumez, E., Duchalais, L., Goudemand, E., Bordes, J., & Robert, O. (2014). Genome-wide prediction of three important traits in bread wheat. *Molecular Breeding*, *34*(4), 1843–1852. https://doi.org/10.1007/s11032-014-0143-y

Combs, E., & Bernardo, R. (2013). Accuracy of Genomewide Selection for Different Traits with Constant Population Size, Heritability, and Number of Markers. *The Plant Genome*, *6*(1), plantgenome2012.11.0030. https://doi.org/10.3835/plantgenome2012.11.0030

de los Campos, G., & Grüneberg, A. (2016). *MTM Package.* https://github.com/QuantGen/MTM/

de los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D., & Calus, M. P. L. (2013). Whole-genome regression and prediction methods applied to plant and animal breeding. In *Genetics* (Vol. 193, Issue 2, pp. 327–345). https://doi.org/10.1534/genetics.112.143313

Dong, H., Wang, R., Yuan, Y., Anderson, J., Pumphrey, M., Zhang, Z., & Chen, J. (2018). Evaluation of the Potential for Genomic Selection to Improve Spring Wheat Resistance to Fusarium Head Blight in the Pacific Northwest. *Frontiers in Plant*

Science, *9*, 911. https://doi.org/10.3389/fpls.2018.00911

Doyle, J. J., & Doyle, J. L. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *PHYTOCHEMICAL BULLETIN*. https://worldveg.tind.io/record/33886

Endelman, J. B. (2011). Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *The Plant Genome*, *4*(3), 250–255. https://doi.org/10.3835/plantgenome2011.08.0024

Endelman, J. B., & Jannink, J.-L. (2012). Shrinkage Estimation of the Realized Relationship Matrix. *G3: Genes|Genomes|Genetics*, *2*(11), 1405–1413. https://doi.org/10.1534/g3.112.004259

Fernandes, S. B., Dias, K. O. G., Ferreira, D. F., & Brown, P. J. (2018). Efficiency of multi-trait, indirect, and trait-assisted genomic selection for improvement of biomass sorghum. *Theoretical and Applied Genetics*, *131*(3), 747–755. https://doi.org/10.1007/s00122-017-3033-y

Fischer, R., Byerlee, D., & Edmeades, G. (2014). Crop yields and global food security. In *academia.edu*. ACIAR: Canberra, ACT. http://www.academia.edu/download/35887178/Crop_yields_and_global_food_security___a_book_by_T.Fischer_et_al__2014.pdf

Guo, G., Zhao, F., Wang, Y., Zhang, Y., Du, L., & Su, G. (2014). Comparison of single-trait and multiple-trait genomic prediction models. *BMC Genetics*, *15*. https://doi.org/10.1186/1471-2156-15-30

Guo, J., Khan, J., Pradhan, S., Shahi, D., Khan, N., Avci, M., Mcbreen, J., Harrison, S., Brown-Guedira, G., Murphy, J. P., Johnson, J., Mergoum, M., Esten Mason, R., Ibrahim, A. M. H., Sutton, R., Griffey, C., & Babar, M. A. (2020). Multi-Trait Genomic Prediction of Yield-Related Traits in US Soft Wheat under Variable Water Regimes. *Genes*, *11*(11), 1270. https://doi.org/10.3390/genes11111270

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

Haile, T. A., Walkowiak, S., N'Diaye, A., Clarke, J. M., Hucl, P. J., Cuthbert, R. D., Knox, R. E., & Pozniak, C. J. (2020). Genomic prediction of agronomic traits in wheat using different models and cross-validation designs. *Theoretical and Applied Genetics*, *1*, 3. https://doi.org/10.1007/s00122-020-03703-z

Hayes, B. J., Panozzo, J., Walker, C. K., Choy, A. L., Kant, S., Wong, D., Tibbits, J., Daetwyler, H. D., Rochfort, S., Hayden, M. J., & Spangenberg, G. C. (2017). Accelerating wheat breeding for end-use quality with multi-trait genomic predictions incorporating near infrared and nuclear magnetic resonance-derived phenotypes. *Theoretical and Applied Genetics*, *130*(12), 2505–2519. https://doi.org/10.1007/s00122-017-2972-7

He, S., Schulthess, A. W., Mirdita, V., Zhao, Y., Korzun, V., Bothe, R., Ebmeyer, E., Reif, J. C., & Jiang, Y. (2016). Genomic selection in a commercial winter wheat population. *Theoretical and Applied Genetics*, *129*(3), 641–651. https://doi.org/10.1007/s00122-015-2655-1

Heffner, E. L., Sorrells, M. E., & Jannink, J. L. (2009). Genomic selection for crop improvement. In *Crop Science* (Vol. 49, Issue 1, pp. 1–12). https://doi.org/10.2135/cropsci2008.08.0512

Ibba, M. I., Crossa, J., Montesinos-López, O. A., Montesinos-López, A., Juliana, P., Guzman, C., Delorean, E., Dreisigacker, S., & Poland, J. (2020). Genome-based prediction of multiple wheat quality traits in multiple years. *The Plant Genome*, *13*(3). https://doi.org/10.1002/tpg2.20034

IWGSC. (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science (New York, N.Y.)*, *361*(6403), eaar7191. https://doi.org/10.1126/science.aar7191

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

Jarquin, D., Howard, R., Crossa, J., Beyene, Y., Gowda, M., Martini, J. W. R., Pazaran, G. C., Burgueño, J., Pacheco, A., Grondona, M., Wimmer, V., & Prasanna, B. M. (2020). Genomic prediction enhanced sparse testing for multi-environment trials. *G3: Genes, Genomes, Genetics*, *10*(8), 2725–2739. https://doi.org/10.1534/g3.120.401349

Jia, Y., & Jannink, J. L. (2012). Multiple-trait genomic selection methods increase genetic value prediction accuracy. *Genetics*, *192*(4), 1513–1522. https://doi.org/10.1534/genetics.112.144246

Jiang, J., Zhang, Q., Ma, L., Li, J., Wang, Z., & Liu, J. F. (2015). Joint prediction of

multiple quantitative traits using a Bayesian multivariate antedependence model. *Heredity*, *115*(1), 29–36. https://doi.org/10.1038/hdy.2015.9

Juliana, P., Singh, R. P., Braun, H.-J., Huerta-Espino, J., Crespo-Herrera, L., Govindan, V., Mondal, S., Poland, J., & Shrestha, S. (2020). Genomic Selection for Grain Yield in the CIMMYT Wheat Breeding Program—Status and Perspectives. *Frontiers in Plant Science*, *11*, 1. https://doi.org/10.3389/fpls.2020.564183

Juliana, P., Singh, R. P., Singh, P. K., Crossa, J., Huerta-Espino, J., Lan, C., Bhavani, S., Rutkoski, J. E., Poland, J. A., Bergstrom, G. C., & Sorrells, M. E. (2017). Genomic and pedigree-based prediction for leaf, stem, and stripe rust resistance in wheat. *Theoretical and Applied Genetics*, *130*(7), 1415–1430. https://doi.org/10.1007/s00122-017-2897-1

Lado, B., Vázquez, D., Quincke, M., Silva, P., Aguilar, I., & Gutiérrez, L. (2018). Resource allocation optimization with multi-trait genomic prediction for bread wheat (Triticum aestivum L.) baking quality. *Theoretical and Applied Genetics*, *131*(12), 2719–2731. https://doi.org/10.1007/s00122-018-3186-3

Lorenz, A. J., Chao, S., Asoro, F. G., Heffner, E. L., Hayashi, T., Iwata, H., Smith, K. P., Sorrells, M. E., & Jannink, J. L. (2011). Genomic Selection in Plant Breeding. Knowledge and Prospects. In *Advances in Agronomy* (Vol. 110, Issue C). Academic Press Inc. https://doi.org/10.1016/B978-0-12-385531-2.00002-5

Maulana, F., Kim, K., Anderson, J. D., Sorrells, M. E., Butler, T. J., Liu, S., Baenziger, P. S., Byrne, P. F., & Ma, X. (2021). Genomic selection of forage agronomic traits in winter wheat. *Crop Science*, *61*(1), 410–421. https://doi.org/10.1002/csc2.20304

Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. In *Genetics Soc America*. https://www.genetics.org/content/157/4/1819.short

Montesinos-López, O. A., Montesinos-López, A., Crossa, J., Toledo, F. H., Pérez-Hernández, O., Eskridge, K. M., & Rutkoski, J. (2016). A genomic bayesian multi-trait and multi-environment model. *G3: Genes, Genomes, Genetics*, *6*(9), 2725–2774. https://doi.org/10.1534/g3.116.032359

Montesinos-López, O. A., Montesinos-López, A., Luna-Vázquez, F. J., Toledo, F. H., Pérez-Rodríguez, P., Lillemo, M., & Crossa, J. (2019). An R package for Bayesian analysis of multi-environment and multi-trait multi-environment data for genome-based prediction. *G3: Genes, Genomes, Genetics*, *9*(5), 1355–1369. https://doi.org/10.1534/g3.119.400126

Oury, F. X., Godin, C., Mailliard, A., Chassin, A., Gardet, O., Giraud, A., Heumez, E., Morlais, J. Y., Rolland, B., Rousset, M., Trottet, M., & Charmet, G. (2012). A study of genetic progress due to selection reveals a negative effect of climate change on bread wheat yield in France. *European Journal of Agronomy*, *40*, 28–38. https://doi.org/10.1016/j.eja.2012.02.007

Pérez-Rodríguez, P., Gianola, D., González-Camacho, J. M., Crossa, J., Manès, Y., & Dreisigacker, S. (2012). Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. *G3: Genes, Genomes, Genetics*, *2*(12), 1595–1605. https://doi.org/10.1534/g3.112.003665

Poland, J. A., Brown, P. J., Sorrells, M. E., & 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*(2), e32253. https://doi.org/10.1371/journal.pone.0032253

Poland, J., Endelman, J., Dawson, J., Rutkoski, J., Wu, S., Manes, Y., Dreisigacker, S., Crossa, J., Sánchez-Villeda, H., Sorrells, M., & Jannink, J. (2012). Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing. *The Plant Genome*, *5*(3), plantgenome2012.06.0006. https://doi.org/10.3835/plantgenome2012.06.0006

R Core Team. (2018). *R: A language and environment for statistical computing; 2015*. https://scholar.google.com/scholar?cluster=9441913529578809097&hl=en&as_sdt=5,42&sciodt=0,42

Randhawa, H. S., Asif, M., Pozniak, C., Clarke, J. M., Graf, R. J., Fox, S. L., Humphreys, D. G., Knox, R. E., DePauw, R. M., Singh, A. K., Cuthbert, R. D., Hucl, P., & Spaner, D. (2013). Application of molecular markers to wheat breeding in Canada. *Plant Breeding*, *132*(5), n/a-n/a. https://doi.org/10.1111/pbr.12057

Rutkoski, J., Benson, J., Jia, Y., Brown-Guedira, G., Jannink, J.-L., & Sorrells, M. (2012). Evaluation of Genomic Prediction Methods for Fusarium Head Blight Resistance in Wheat. *The Plant Genome*, *5*(2), 51–61. https://doi.org/10.3835/plantgenome2012.02.0001

Rutkoski, J. E., Poland, J. A., Singh, R. P., Huerta-Espino, J., Bhavani, S., Barbier, H., Rouse, M. N., Jannink, J., & Sorrells, M. E. (2014). Genomic Selection for Quantitative Adult Plant Stem Rust Resistance in Wheat. *The Plant Genome*, *7*(3). https://doi.org/10.3835/plantgenome2014.02.0006

Rutkoski, J., Poland, J., Mondal, S., Autrique, E., Pérez, L. G., Crossa, J., Reynolds, M., & Singh, R. (2016). Canopy temperature and vegetation indices from high-throughput phenotyping improve accuracy of pedigree and genomic selection for grain yield in wheat. *G3: Genes, Genomes, Genetics*, *6*(9), 2799–2808. https://doi.org/10.1534/g3.116.032888

Schulthess, A. W., Zhao, Y., Longin, C. F. H., & Reif, J. C. (2018). Advantages and limitations of multiple-trait genomic prediction for Fusarium head blight severity in hybrid wheat (Triticum aestivum L.). *Theoretical and Applied Genetics*, *131*(3), 685–701. https://doi.org/10.1007/s00122-017-3029-7

Sun, J., Rutkoski, J. E., Poland, J. A., Crossa, J., Jannink, J., & 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*(2). https://doi.org/10.3835/plantgenome2016.11.0111

Tadesse, W., Sanchez-Garcia, M., Gizaw Assefa, S., Amri, A., Bishaw, Z., Ogbonnaya, F. C., & Baum, M. (2019). Genetic Gains in Wheat Breeding and Its Role in Feeding the World. *Crop Breeding, Genetics and Genomics*. https://doi.org/10.20900/cbgg20190005

Tang, Y., Liu, X., Wang, J., Li, M., Wang, Q., Tian, F., Su, Z., Pan, Y., Liu, D., Lipka, A. E., Buckler, E. S., & Zhang, Z. (2016). GAPIT Version 2: An Enhanced Integrated Tool for Genomic Association and Prediction. *The Plant Genome*, *9*(2), plantgenome2015.11.0120. https://doi.org/10.3835/plantgenome2015.11.0120

Tester, M., & Langridge, P. (2010). Breeding technologies to increase crop production in

a changing world. In *Science* (Vol. 327, Issue 5967, pp. 818–822). American Association for the Advancement of Science. https://doi.org/10.1126/science.1183700

VanRaden, P. M., Van Tassell, C. P., Wiggans, G. R., Sonstegard, T. S., Schnabel, R. D., Taylor, J. F., & Schenkel, F. S. (2009). Invited review: Reliability of genomic predictions for North American Holstein bulls. In *Journal of Dairy Science* (Vol. 92, Issue 1, pp. 16–24). American Dairy Science Association. https://doi.org/10.3168/jds.2008-1514

Wang, X, Li, L., Yang, Z., Zheng, X., Yu, S., Xu, C., & Hu, Z. (2017). Predicting rice hybrid performance using univariate and multivariate GBLUP models based on North Carolina mating design II. *Heredity*, *118*(3). https://doi.org/10.1038/hdy.2016.87

Wang, Xin, Xu, Y., Hu, Z., & Xu, C. (2018). Genomic selection methods for crop improvement: Current status and prospects. In *Crop Journal* (Vol. 6, Issue 4, pp. 330–340). Crop Science Society of China/ Institute of Crop Sciences. https://doi.org/10.1016/j.cj.2018.03.001

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

CHAPTER 5

Multi-trait genomic selection improves the prediction accuracy of end-use quality traits in hard winter wheat

This chapter has been published in *The Plant Genome* Journal.

Citation: Gill, H. S., Brar, N., Halder, J., Hall, C., Seabourn, B. W., Chen, Y. R., ... & Sehgal, S. K. (2023). Multi-trait genomic selection improves the prediction accuracy of end-use quality traits in hard winter wheat. The Plant Genome, e20331.

5.1 Abstract

Improvement of end-use quality remains one of the most important goals in hard winter wheat (HWW) breeding. Nevertheless, evaluation of end-use quality traits is confined to later development generations owing to resource-intensive phenotyping. Genomic selection (GS) has shown promise in facilitating selection for end-use quality, however, lower prediction accuracy (PA) for complex traits remains a challenge in GS implementation. Multi-trait genomic prediction (MTGP) models can improve PA for complex traits by incorporating information on correlated secondary traits, but these models remain to be optimized in HWW. A set of advanced breeding lines from 2015-2021 were genotyped with 8,725 SNPs and was employed to evaluate MTGP to predict various end-use quality traits that are otherwise impossible to phenotype in earlier generations. The MTGP model outperformed the ST model with up to a two-fold increase in PA. For instance, PA was improved from 0.38 to 0.75 for bake absorption and from 0.32 to 0.52 for loaf volume. Further, we compared MTGP models by including different combinations of easy-to-score traits as covariates to predict end-use quality traits. Incorporation of simple traits such as flour protein (FLRPRO) and sedimentation weight value (FLRSDS) substantially improved the PA of MT models. Thus, rapid low-cost measurement of traits like FLRPRO and FLRSDS can facilitate the use of GP to predict Mixograph and baking traits in earlier generations and provide breeders an opportunity for selection on end-use quality traits by culling inferior lines to increase selection accuracy and genetic gains.

5.2 Introduction

Hard winter wheat (*Triticum aestivum* L.; HWW) is the major wheat class grown in the US and accounts for about 46 percent of the total wheat production in the country (USDA NASS, 2021). This versatile class of wheat exhibits excellent milling and baking characteristics suitable for a variety of wheat foods, especially bread. Owing to high demand, most of the US-produced HWW is exported. For instance, 52 percent of the total HWW produced in 2020 was exported worldwide (USDA ERS, 2022). Throughout the wheat supply chain, end-use quality characteristics play an important role in the marketing and pricing of HWW (Roberts et al., 2022). Moreover, consumers' preferences for healthier food necessitate an emphasis on the selection for desirable end-use quality traits. Thus, wheat breeders must improve end-use quality traits while simultaneously breeding for increased yield to meet projected demand.

The high gluten strength and damaged starch in HWW makes it very suitable for baking, and yeast-leavened bread is a major end-use product. Bread quality is an important but complex trait that is defined by a combination of many parameters (Battenfield et al., 2016). Several important factors including kernel characteristics, the milled flour quality, protein and starch strength, and dough properties all play a crucial role in determining end-use quality of the final product. Hence, several assays are used to profile these factors and inform the selection for end-use quality. Nevertheless, most of the assays for evaluation of end-use products are expensive, time-consuming, and require large quantities of flour. Therefore, breeders mostly prioritize the selection for agronomic traits and disease resistance in earlier generations and for quality traits in advanced generations in most breeding programs (Battenfield et al., 2016). Previous studies have

shown that end-use quality traits are controlled by a few major genes and large number of quantitative trait loci with minor effects (Carter et al., 2012; Jernigan et al., 2018; Kiszonas & Morris, 2017; Sandhu et al., 2021). Though available major genes have been exploited in breeding programs, the majority of minor genes are highly influenced by the environment and remain uncharacterized (Jernigan et al., 2018; Kiszonas & Morris, 2017). Thus, genomic selection (GS) is a potentially effective tool to assist in the selection for end-use quality traits in earlier generations.

Genomic selection employs whole genome marker information to predict the breeding value of an individual (Heffner et al., 2009; Meuwissen et al., 2001). Genomic selection has been shown to increase genetic gain per breeding cycle and improve selection accuracy for complex traits (Bassi et al., 2015; Juliana et al., 2019), particularly the traits that are expensive and difficult to phenotype and cannot be evaluated at earlier stages of the breeding program (Battenfield et al., 2016; Gill et al., 2021). In recent years, GS has been evaluated for the prediction of various complex traits in wheat including agronomic traits (Gill et al., 2021; Juliana et al., 2020; Rutkoski et al., 2016; Ward et al., 2019), disease resistance (Juliana et al., 2017; Rutkoski et al., 2012; Zhang et al., 2022), and end-use quality (Battenfield et al., 2016; Ibba et al., 2020; Lado et al., 2018; Sandhu et al., 2021; Zhang-Biehn et al., 2021). Most of these studies showed success in GS and suggested the possibility for use of GS to predict complex traits in wheat breeding. Although numerous studies have evaluated GS for end-use quality in wheat, most employed soft wheat germplasm with only one study using HWW from the US Great Plains (Zhang-Biehn et al., 2021). Since the processing methods and end-use objectives

are quite different among different classes of wheat, it is necessary to further evaluate the usability of GS to predict various milling and baking traits in HWW.

Most of the GS studies primarily used single-trait genomic prediction (STGP) models for predicting individual traits. In recent years, several simulated and empirical studies evaluated multi-trait (MT) genomic prediction models that can leverage genetic correlations among different traits to improve prediction accuracy (PA) for traits(s) of interest (Gaire et al., 2022; Gill et al., 2021; Jia & Jannink, 2012; Lado et al., 2018; Zhang et al., 2022). In most of these studies, MTGP models showed superior performance to the conventional STGP models. Few studies evaluated the MT models to evaluate GS for end-use quality traits in wheat (Hayes et al., 2017; Lado et al., 2018; Michel et al., 2018; Sandhu et al., 2022; Zhang-Biehn et al., 2021). However, only Zhang-Biehn et al. (2021) evaluated MTGP for HWW end-use quality traits and used few pre-baking assays as covariates to predict bread quality. Henceforth, further evaluation of MTGP to predict end-use quality traits in HWW may facilitate its implementation in HWW breeding programs. In addition, several grain quality assays such as hybrid sodium dodecyl sulfate-solvent retention capacity (SDS-SRC) and flour characteristics that require minimal resources and can be evaluated in early breeding generations show a high association with primary end-use quality traits (Seabourn et al., 2012). Evaluating the MTGP models that incorporate those traits to predict baking traits may help to predict end-use traits in the early stages of breeding programs.

Depending on the availability of flour and other resources, different types of quality tests are used to assess the end-use traits and inform selections in different trials in breeding programs including South Dakota State University winter wheat breeding

program (Figure 5.1). In early generations, the grain and flour characteristics are assessed using a near-infrared reflectance (NIR) based analyzer followed by a rapid and small-scale hybrid SDS-SRC test to predict breadmaking quality. Additional tests like Mixograph analysis for rheological properties, and Glutomatic analysis for gluten quantity and quality determination are also included in advanced generations and lastly, complete end-use profiling including baking tests are conducted in the final stages of the breeding programs (Figure 5.1). As it is challenging to perform a Mixograph or actual baking tests in earlier stages, such as preliminary yield trials (PYT), MTGP could be employed to predict baking traits at these stages. The MTGP models can be informed with trait measurements from rapid assays in PYTs along with complete data from more advanced trials to predict end-use quality traits with higher selection accuracy. This necessitates the evaluation of different trait combinations from different assays that can help in improving the PA of baking traits using MTGP. In this study, we used a set of breeding lines from the advanced breeding trials of the SDSU winter wheat program that were evaluated for a variety of quality parameters including milling, processing, and baking characteristics, and genotyped using the genotyping-by-sequencing (GBS) approach. The objectives of this study were to (i) evaluate GS using single-trait and multi-trait models for different end-use traits and (ii) explore the usability of rapid, small-scale, and NIRS-based traits as covariates to inform muti-trait GP models for baking quality.

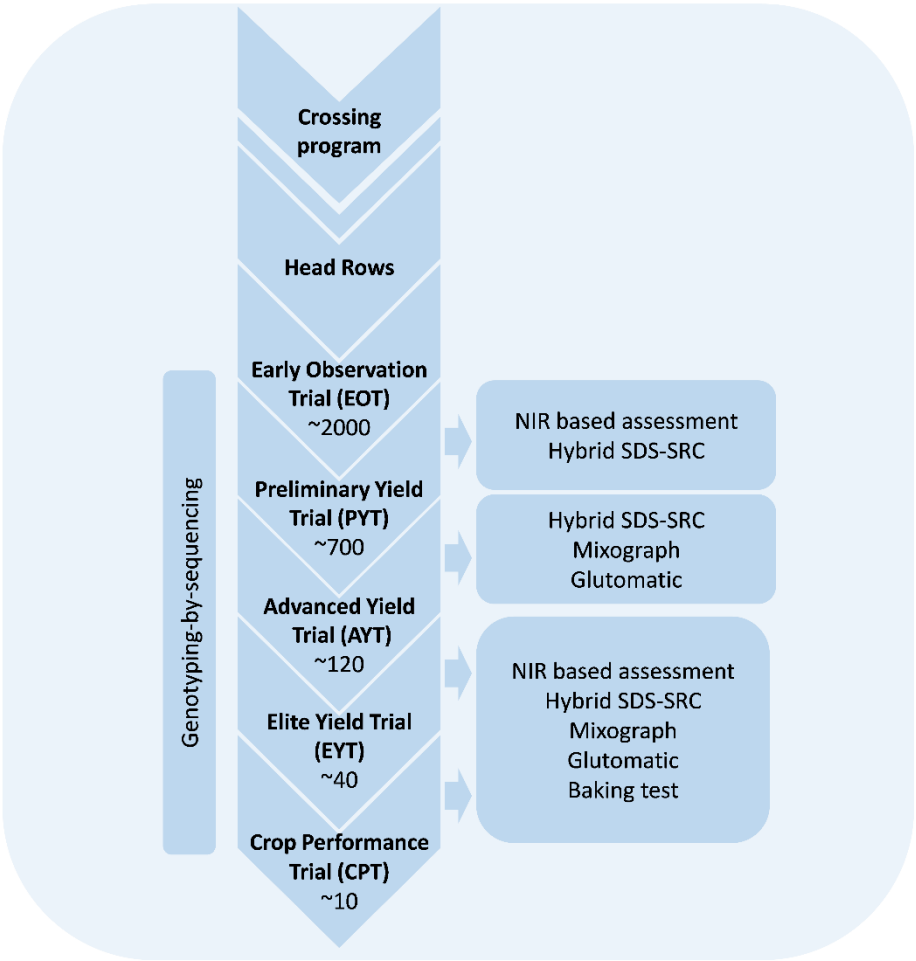

Figure 5.1 Schematic representation of the South Dakota State University winter wheat breeding program. The early stage of yield trials is the Preliminary Yield Trial (PYT, ~700 lines) advanced from Early Observation Trial (EOT) consisting of short rows derived from single selected plants. The PYT is followed by Advanced Yield Trial (AYT), Elite Yield Trial (EYT), and statewide Crop Performance Testing (CPT) nursery. The quality assessment starts from the PYTs, and various quality assays are used at different stages of development owing to the availability of flour and other resources. The different quality assays performed at various stages of the breeding program are elucidated in the figure.

5.3 Materials and Methods

5.3.1 Plant materials and Phenotyping

In this study, we used a set of lines evaluated in SDSU Elite Yield Trials (EYT) and Crop Performance Trials (CPT) nurseries from 2015 to 2021 that were profiled for a variety of end-use quality traits including baking tests. During this period, a total of 300 samples including checks were evaluated in these nurseries. The EYT and CPT nurseries were planted at multiple locations each year in a randomized complete block design with three and four replications, respectively, along with a set of check cultivars (‘Alice’, ‘Expedition’, ‘Lyman’, ‘Overland’, ‘Redfield’ and ‘Winner’). Each year, lines from two locations were used for end-use quality profiling including baking. In addition to checks, a set of about 15 lines was shared among every two years as these lines were advanced from EYT to CPT nurseries. For instance, 18 lines were common between 2015 and 2016, while 19 lines overlapped between 2016 and 2017. Overall, 176 unique lines were evaluated for milling and baking tests from 13 environments (site-year combination) from 2015 to 2021. A detailed description of the lines used in this study is provided in Table 1. Since the short harvest-to-planting interval of less than one month in SD, the quality traits of the lines are analyzed using grain harvested from preceding nursery seasons.

A total of 14 processing and end-use quality traits were assessed using various assays. Owing to limited test capacity, grain from replications within locations was pooled into a single sample for milling and baking test profiles. Grain protein content (GRPROT) was determined using NIR analysis following AACC-approved methods 39-10.01 (AACC International, 2011) and adjusted to a 13% moisture equivalent. Grain

samples were tempered to the required moisture (14%) and milled using a Brabender Quadrumat Senior laboratory mill at the USDA Hard Winter Wheat Quality Laboratory (HWWQL), in Manhattan, KS. The milled product was weighed to determine fraction yield on a total product, thus providing an estimation of flour yield (FLRYLD). Flour characteristics including protein (FLRPROT) and ash (FLRASH) were analyzed using NIR following AACC methods 39-11.01 and 08-21.01 (AACC International, 2011). These traits were further adjusted to 14% moisture content equivalence. Mixograph analysis was performed using Mixgraph (National Manufacturing Company, USA) based on AACC-approved method 54-40.02 (AACC International, 2011) to estimate the water absorption (MIXABS), optimum development time (MIXTIM), and tolerance to overmixing (MIXTOL). Glutomatic analysis was performed using a Perten Glutomatic 2000 system following AACC 38-12.02 (AACC International, 2011) to estimate wet gluten content (WGC) and gluten index (GI) as described in AACC method. Further, we used WGC and GI to determine another index for wet gluten (WGI) using the formula (WGC*GI)/100. These analyses were performed using two replicates per sample. Sodium dodecyl sulfate (SDS) sedimentation was performed using the modified hybrid SDS-SRC method (Seabourn et al., 2012) at SDSU Crop Quality Lab using the residual flour samples for all the entries analyzed for baking. Briefly, the hybrid SDS-SRC method combines the sodium dodecyl sulfate (SDS) sedimentation method (AACC 56-70) (AACC International, 2000) and solvent retention capacity (SRC) method (AACC 56-11) (AACC International, 2000) to estimate flour sedimentation weight value (FLRSDS) in percent using the formula described in Seabourn et. al. (2012). To evaluate the end-use quality of yeast-leavened bread, a pan-bread was baked as pup loaves following the

AACC method 10-10.03 (AACC International, 2011). Optimum water absorption from baking tests was recorded as bake absorption (BAKEABS). Loaf volume (LVOL) in cc and specific loaf volume (SpLVOL) in cc/g were recorded after baking by using the rapeseed displacement method 10-05.01 (AACC International, 2011). Samples were baked in two replicates and means were used in the final analysis.

Table 5.1. Description of trial years for the breeding lines used in the current study.

Year of evaluation	Nurseries[a]	Number of Entries	Number of locations	Entries overlapping with previous year
2015	EYT, CPT	35	1	-
2016	EYT, CPT	39	2	18
2017	EYT, CPT	43	2	19
2018	EYT, CPT	46	2	42
2019	EYT, CPT	47	2	14
2020	EYT, CPT	44	2	15
2021	EYT, CPT	46	2	10
	Total	300	13	-

[a]Nurseries: Elite Yield Trial, EYT; Crop Performance Trial, CPT.

5.3.2 Genotyping

The set of breeding lines used in this study were genotyped using the GBS approach at the USDA Central Small Grain Genotyping Lab, Manhattan, KS as described in Gill et al. (2022). Briefly, the DNA was isolated from each line using fresh leaf tissue at the three-leaf stage using a modified cetyl-trimethyl ammonium bromide (CTAB) method (Bai et al., 1999). The GBS libraries were prepared using double restriction digestion with HF-*PstI* and *MspI* restriction enzymes (Poland et al., 2012) and sequenced on an Ion

Proton sequencer (Thermo Fisher Scientific, Waltham, MA, USA) or NextSeq 500 (Illumina Inc, USA). Single-nucleotide polymorphism (SNP) variants were called using the GBS v2.0 SNP discovery pipeline in TASSEL v5.0 (Bradbury et al., 2007) using the Chinese Spring wheat genome reference RefSeq v2.0 (IWGSC, 2018; Zhu et al., 2021). For quality control, SNPs were filtered to remove the markers with > 30% missing calls, < 5% minor allele frequency (MAF), and > 10% heterozygosity. The remaining SNPs were imputed using BEAGLE v4.1 (beagle.27Jan18.7e1.jar; https://faculty.washington.edu/browning/beagle/b4_1.html) (Browning & Browning, 2007) for further analyses.

5.3.3 Statistical analysis

The best linear unbiased estimates (BLUE)s of each line for majority of the traits were estimated as described by Zhang-Biehn et al (2021) using the following model:

$$Y_{ij} = \mu + Line_i + Env_j + e_{ij}$$

where Y_{ij} is the phenotypic value of the i^{th} line in the j^{th} environment (site-year combination), μ is the overall mean, $Line_i$ is the fixed effect of the i^{th} line, Env_j was the random effect of the j^{th} environment, and e_{ij} is the residual error for genotype the i^{th} line in j^{th} environment. The BLUEs for Glutomatic traits were estimated using the following model:

$$Y_{ijk} = \mu + Line_i + Env_j + Rep(Env)_{jk} + e_{ijk}$$

where Y_{ijk} is the observed phenotypic, μ is the overall mean, $Line_i$ is the fixed effect of i^{th} line, Env_j was the random effect of j^{th} environment, and $Rep(Env)_k$ is the random effect of k^{th} replicate within the j^{th} environment, and e_{ij} is the residual error term.

To extract variance components, best linear unbiased prediction (BLUP), and estimate broad-sense heritability, line effects were fitted as random in the above equations. Broad-sense heritability was estimated with Cullis's method (Cullis et al., 2006) using the following equation:

$$H^2_{Cullis} = 1 - \frac{\bar{v}^{BLUP}_{\Delta..}}{2\sigma^2_g}$$

where $\bar{v}^{BLUP}_{\Delta..}$ is the mean-variance of the difference of two BLUPs for the line effect and σ^2_g is the genotypic variance. The above-described equations were implemented using ASReml-R Version 4.0 (Butler et al., 2018) in the R programming language (R Core Team, 2018). The summary statistics, pairwise comparisons, and principal component analysis (PCA) were performed in R using custom scripts or different packages including psych and ggplot2 (Wickham, 2016; William, 2013).

5.3.4 Genomic prediction models

5.3.4.1 Single-trait GP models

Three different ST models were compared to select the best-performing model as a baseline for comparison with MTGP model. The first ST model was standard genomic best linear unbiased prediction (GBLUP) employing a genomic relationship (G) matrix implemented using the following equation:

$$y = \mu + Zg + e$$

where y is the vector (n × 1) of BLUE values for each trait; μ is the overall mean, Z is the incidence matrix for genotype effects; g is a vector of normally distributed marker predictor effects with $g \sim N(0, \mathrm{G}\sigma^2_g)$, where G is the genomic relationship matrix

(VanRaden, 2008) σ_g^2 is the additive genetic variance; and *e* is the vector of residual errors with $e \sim N(0, \sigma_e^2)$.

In addition to GBLUP, we used two Bayesian models, Bayes A (BA) and Bayes B (BB), which assume different prior distributions for estimating marker effects and variances (Pérez & De Los Campos, 2014). The Bayes A model uses the scaled inverse chi-squared probability distribution for estimating marker variances. Bayes B is an extension of the Bayes A model (Meuwissen et al., 2001) and employs an inverse chi-square distribution for marker effects and assumes that some markers have no effect. The Bayesian models were implemented as follows:

$$y_i = \mu + \sum_{j=1}^{j=p} x_{ij}\,\beta_j + \ e_i$$

where *y* refers to BLUE values for each trait; μ is the overall mean; x_{ij} is the identity of the SNP marker, β_j is the marker effect, and e_i represents the residual error term. The ST models were implemented with 5000 burn-ins and 25000 iterations of the Gibbs sampler in R package BGLR (https://github.com/gdlc/BGLR-R/blob/master/inst/md/GBLUP.md; Pérez & De Los Campos, 2014).

5.3.4.2 Multi-trait GP models

We used a Bayesian Multivariate Gaussian model to implement multivariate GBLUP for various traits. The MT model can be expressed using the following equation:

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} I & 0 \\ \vdots & \vdots \\ 0 & I_n \end{bmatrix} \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix} + \begin{bmatrix} Z & 0 \\ \vdots & \vdots \\ 0 & Z_n \end{bmatrix} \begin{bmatrix} g_1 \\ \vdots \\ g_n \end{bmatrix} + \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}$$

where y is the n-dimensional vector of BLUEs for n traits, I and Z are the design matrices, μ_t, $t = 1 \ldots n$, refers to trait intercepts of n traits, $\begin{bmatrix} g_1 \\ \vdots \\ g_n \end{bmatrix}$ are the predicted genetic values assumed to be distributed as $\sim MVN(0, \sum \otimes G)$ with G representing the genomic relationship matrix obtained following VanRaden (2008) and $\otimes$ refers to the Kronecker product of two matrices. The residuals of the MT model were assumed to be distributed as $\begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix} \sim MVN(0, R \otimes \mathrm{I})$. The matrices $\sum$ and R are the variance-covariance matrices for the genetic and residual effects between traits, respectively, where $\sum$ was estimated as an unstructured variance-covariance matrix and R as a diagonal variance-covariance matrix. The MT GBLUP was implemented in the MTM package in R (de los Campos & Grüneberg, 2016) using the Gibbs sample algorithm with 5000 burn-ins and 25,000 iterations. The abovementioned model was also used to estimate the genetic correlation between different traits.

5.3.4.3 Combination of traits for multi-trait GP models

The MTGP model was evaluated for prediction of Mixograph and Baking traits using various combinations of secondary traits. For the Mixograph predictions, three traits including MIXABS, MIXTIM, and MIXTOL were used as primary traits. Similarly, three baking traits, namely BAKEABS, LVOL, and SpLVOL, were used as primary traits. To predict these primary traits using the MT model, we evaluated the inclusion of different sets of secondary traits from different quality assays (Figure 5.2). For instance, we compared the incorporation of grain characteristics, flour characteristics, Glutomatic assay, flour SDS, or Mixograph traits as secondary traits to predict LVOL. Likewise,

these sets of secondary traits were used to assess the PA of the MT model for different primary traits.

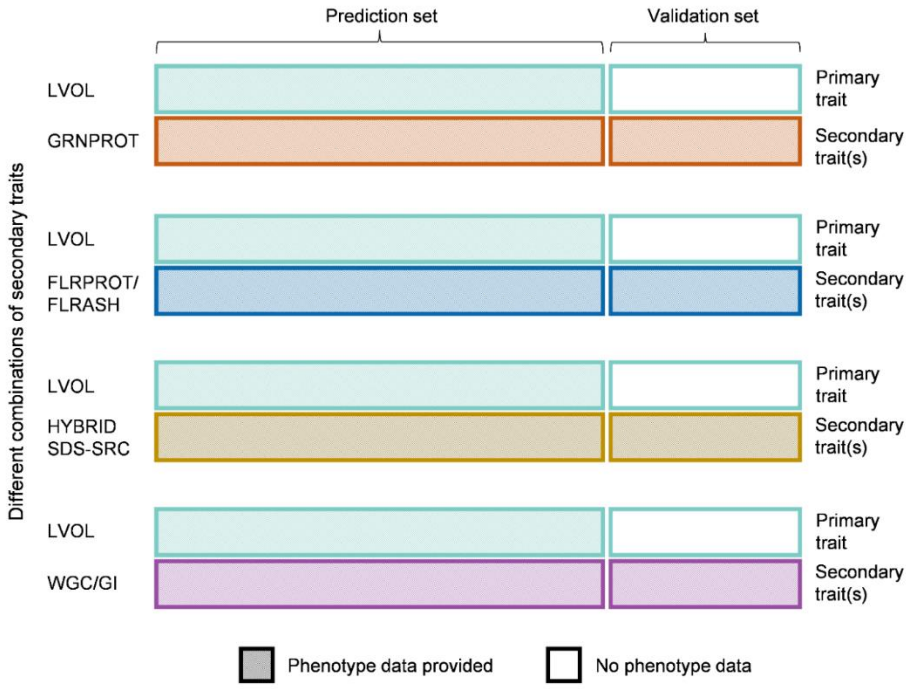

Figure 5.2 Schematic representation of different combinations of secondary traits used in the MT model to predict primary traits of interest. The diagram illustrates a scenario to predict LVOL (primary trait) using the MT model. Various combinations of secondary traits were selected based on different types of pre-baking assays, such as grain/flour characteristics or a flour sedimentation test, performed at various levels in a breeding program. The training set had phenotype data for LVOL while the validation set was not phenotyped for this trait. Contrarily, phenotype data or secondary trait(s) were available for both testing and validation sets.

5.3.5 Cross-validation of the GS models

The PA of the GS models was deduced by calculating the correlation between genome-estimated breeding values (GEBVs) and the actual phenotypic values of individuals in a validation set through a cross-validation approach. We used 100 random sets of five-fold CV approach to evaluate the PA of ST and MT models using two different validation schemes (CV1 and CV2). These validation schemes were designed based on real scenarios observed in plant breeding experiments. The CV1 was used to evaluate the ST models, where four of the five folds (80%) were used as the training set (had genotypic and phenotypic data) to train the model, and the remaining fold (20%) was used as the validation set (only genotypic data) for prediction. The MT model was evaluated using the CV2 scheme, in which lines were split into five folds of equal sizes, with four folds as the training set, and the remaining fold as the validation set. To train the model, we used genotypic data and the phenotypic data of the primary trait for the training set, along with phenotypic data of the secondary traits for training as well as the testing set with the objective of predicting the primary trait of the testing set. Figure 5.3 illustrates various validation schemes used for both models.

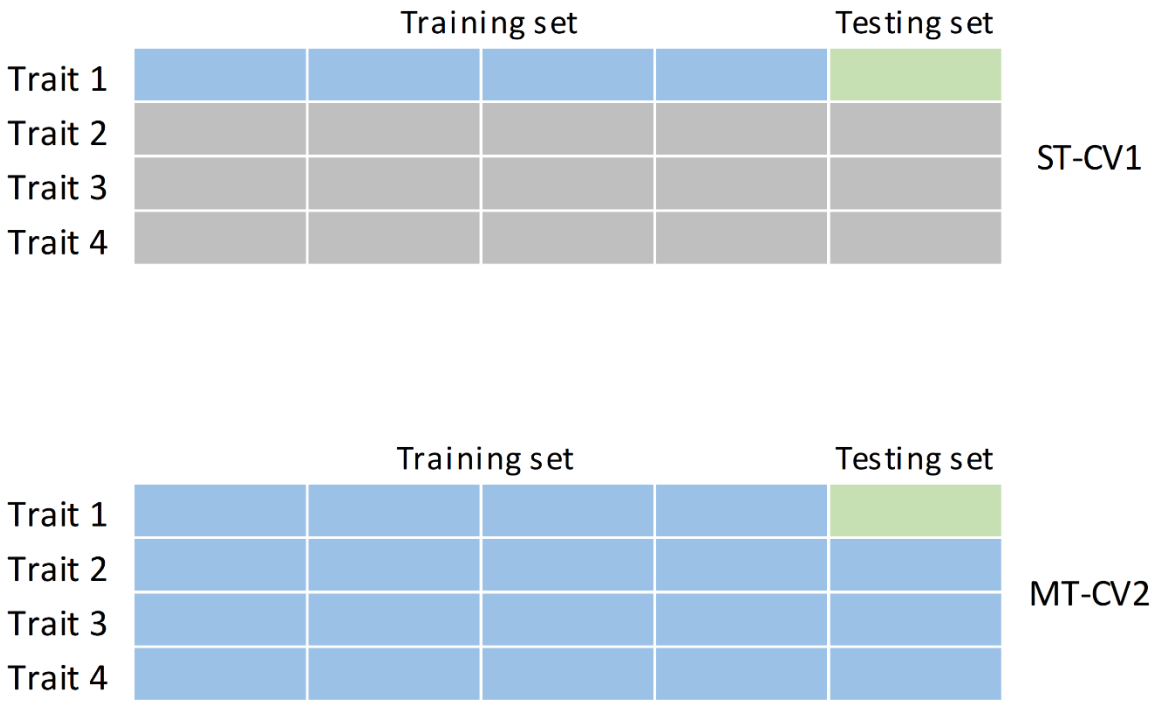

Figure 5.3 Illustration of the different cross-validation (CV) schemes used in this study. The Single trait (ST) Model was evaluated using a CV1 scheme where four sets were used to train the model and the remaining set was used as a testing/validation set. The training set had phenotypic data and genotypic data while the testing data had only genotypic data. The MT model was evaluated using CV2 scheme. In CV2, the training set had phenotyped for the primary trait (trait to be predicted) while the validation set was not phenotyped for this trait. Contrarily, phenotype data or secondary trait(s) was available for both testing and validation sets.

5.4 Results

5.4.1 Phenotypic analyses and trait correlations

The BLUEs for various end-use quality traits were obtained from multi-environment analysis to correct for the environmental effects. An approximate normal distribution of obtained BLUEs was observed for most of the traits (Figure 5.4). Broad-sense heritability estimates were moderate to high for different traits, ranging from 0.44 to 0.88 (Table 2), with low heritability for GRPROT (0.44) and FLRPROT (0.48) and high heritability for MIXTOL (0.84) and MIXTM (0.88).

Significant phenotypic correlations were observed among different quality traits (Figure 5.4). A high positive correlation was observed between GRPROT and FLRPROT (Figure 5.4). FLRPROT also showed a high positive correlation with MIXABS (0.65; $P \leq 0.001$), BAKEABS (0.53; $P \leq 0.001$), and LVOL (0.44; $P \leq 0.001$). FLRSDS was the only trait that exhibited significant correlations with most of the quality traits including LVOL (0.34; $P \leq 0.001$). BAKEABS showed a positive correlation with Mixograph traits including MIXABS (0.76; $P \leq 0.001$), MIXTM (0.47; $P \leq 0.001$), and MIXTOL (0.52; $P \leq 0.001$). Similarly, LVOL was positively correlated with MIXABS (0.40; $P \leq 0.001$) and MIXTM (0.31; $P \leq 0.001$).

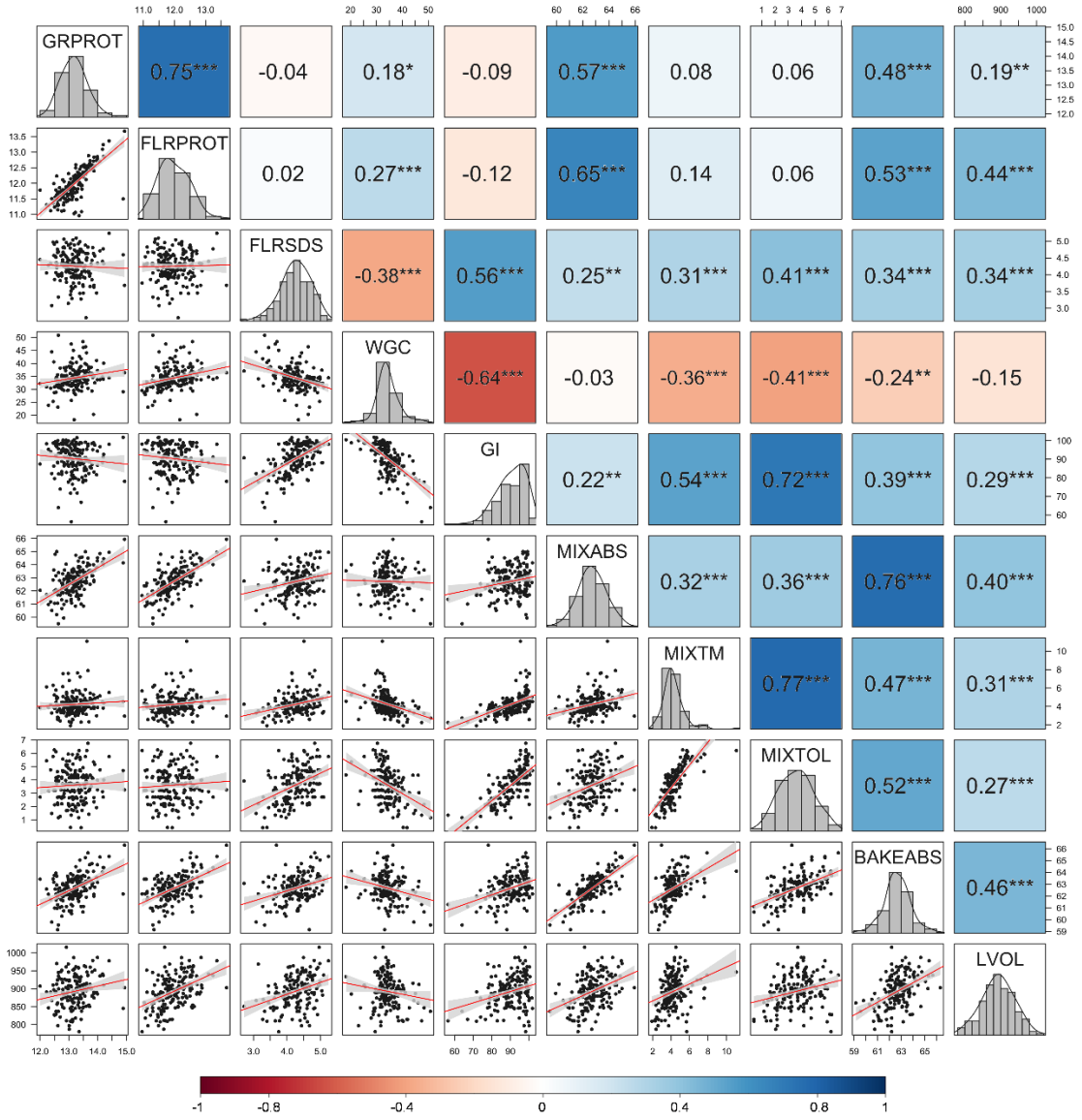

Figure 5.4 Correlation coefficients among investigated traits using the best linear unbiased estimates (BLUEs) obtained from a multi-environment analysis. Statistically significant differences are denoted by an asterisk (*) where * $P \leq 0.05$, ** $P \leq 0.01$, and *** $P \leq 0.001$. GRPROT, grain protein content; FLRPROT, flour protein content; FLRSDS, flour sedimentation weight; WGC, wet gluten content; GI, gluten index; MIXABS, mixing absorption, MIXTM, optimum mix time; MIXTOL, Mixograph mix tolerance; BAKEABS, bake absorption; LVOL, pup loaf volume.

Table 5.2 Descriptive statistics and Broad-sense heritability (H2) for different end-use quality traits.

Trait[a]	Mean	SD	H^2
GRPROT	13.19	0.47	0.44
FLRYLD	68.90	1.29	0.74
FLRASH	0.42	0.03	0.77
FLRPROT	11.93	0.49	0.48
FLRSDS	325.88	43.89	0.70
MIXABS	62.72	1.06	0.55
MIXTM	4.24	1.15	0.88
MIXTOL	3.61	1.27	0.84
WGC	34.07	4.19	0.71
GI	90.10	7.48	0.84
WGI	30.74	2.80	0.72
BAKEABS	62.66	1.12	0.50
LVOL	892.71	44.21	0.59
SpLVOL	5.95	0.30	0.59

[a]GRPROT, grain protein content (%); FLRYLD, flour yield (% recovered); FLRASH, flour ash (%); FLRPROT, flour protein content (%); FLRSDS, Hybrid SDS-SRC sedimentation (weight value %); MIXABS, Mixograph mixing absorption (%), MIXTM, Mixograph mix time (min); MIXTOL, Mixograh mix tolerance score; WGC, wet gluten content; GI, gluten index; BAKEABS, bake absorption; LVOL, pup loaf volume (cm^3); SpLVOL, loaf volume by weight (g/cm^3).

Principal component analysis (PCA) using all the trait data also showed similar groups of correlated traits (Figure 5.6). The first and second components from PCA explained 31.1% and 17.4% of the total phenotypic variance. PCA showed a strong association between GRPROT and FLRPROT. The baking traits were found to be positively associated with MIXABS while another group included MIXTM, MIXTOL, GI, and FLRSDS. Further, we estimated the genetic correlations among different traits using the Bayesian Multivariate Gaussian model (Figure 5.5). Interestingly, FLRSDS

exhibited significant genetic correlations with different quality parameters ($P \leq 0.001$) including GI (0.71), MIXTM (0.49), MIXTOL (0.56), BAKEABS (0.47), and LVOL (0.44). Apart from FLRSDS, FLRPROT showed significant genetic correlations with MIXABS (0.78), BAKEABS (0.62), and LVOL (0.43) (Figure 5.5).

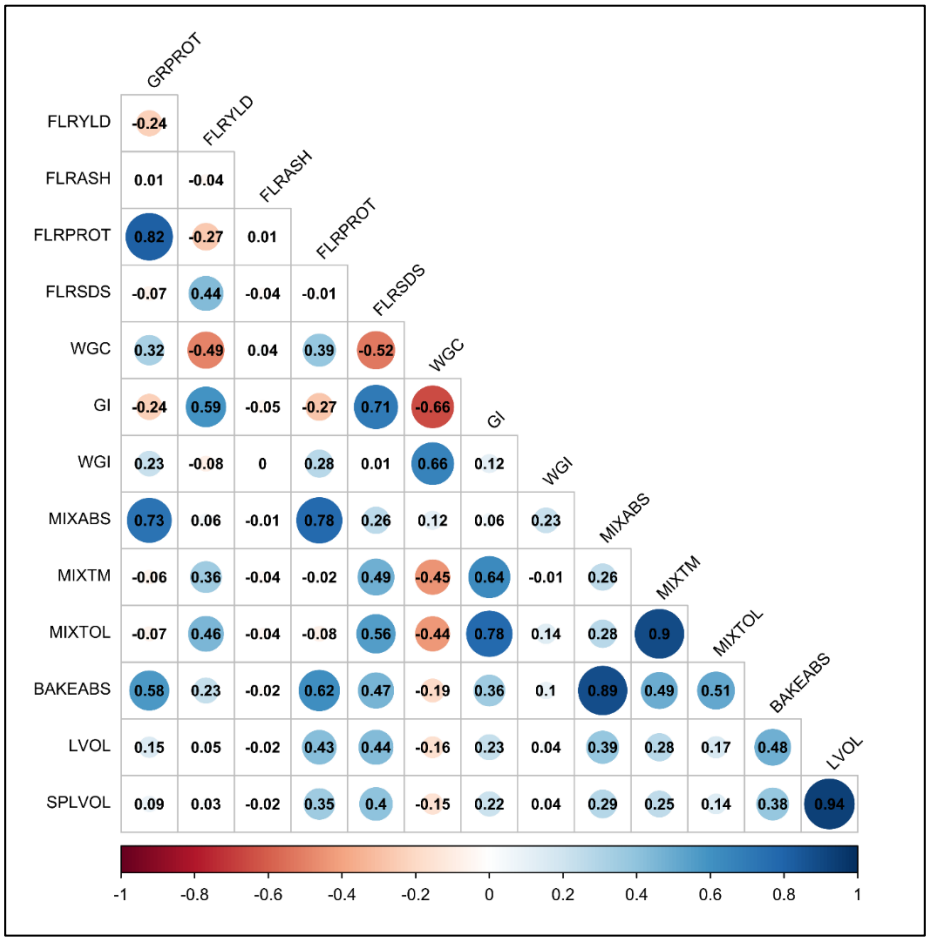

Figure 5.5 The genetic correlation among various end-use quality traits. GRPROT, grain protein content; FLRYLD, flour yield; FLRASH, flour ash content; FLRPROT, flour protein content; FLRSDS, flour sedimentation weight; WGC, wet gluten content; GI, gluten index; MIXABS, Mixograph mixing absorption (%), MXTIM, Mixograph mix time (min); MIXTOL, Mixograph mix tolerance score; BAKEABS, bake absorption; LVOL, pup loaf volume.

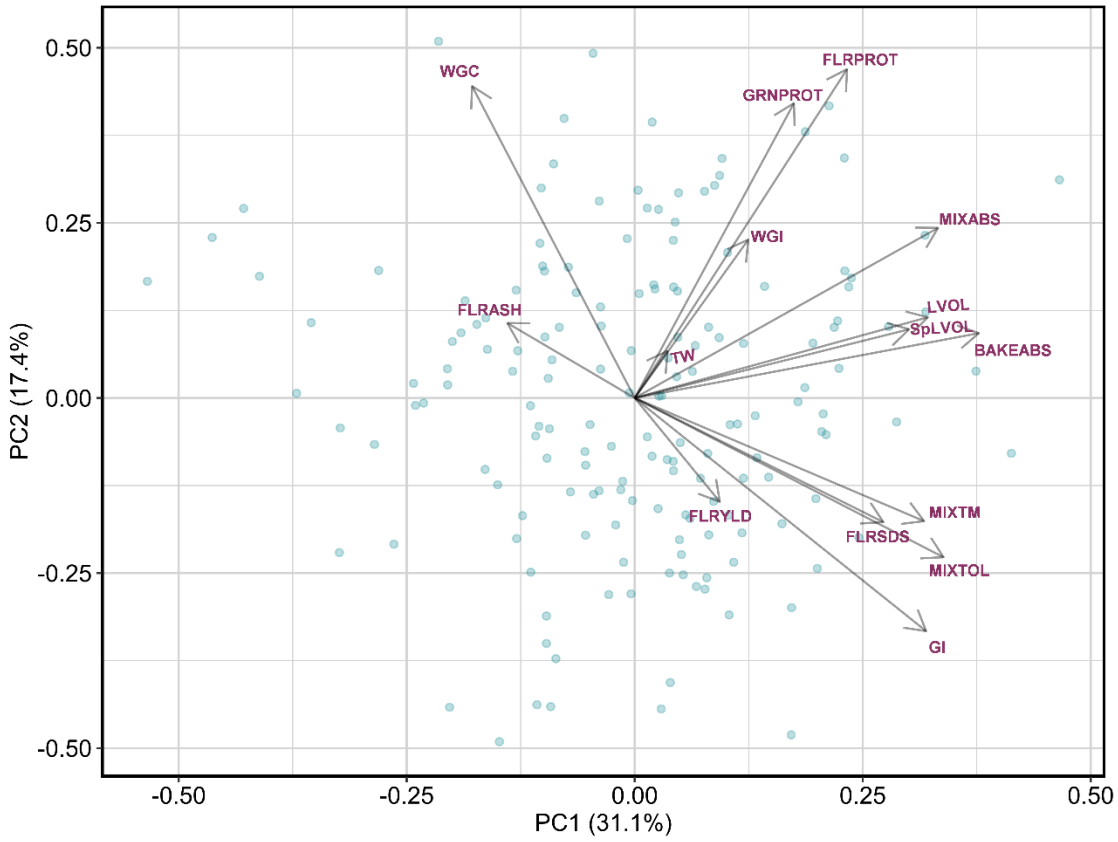

Figure 5.6 Principal component analysis for the studied end-use quality traits based on the phenotypic data. GRPROT, grain protein content; FLRYLD, flour yield; FLRASH, flour ash content; FLRPROT, flour protein content; FLRSDS, flour sedimentation weight value (%); WGC, wet gluten content; GI, gluten index; WGI, wet gluten index; MIXABS, Mixograph mixing absorption (%), MIXTM, Mixograph mix time (min); MIXTOL, Mixograph mix tolerance score; BAKEABS, bake absorption; LVOL, pup loaf volume; SpLVOL, loaf volume by weight, TW, test weight.

5.4.2 Genotypic analyses

The breeding lines were genotyped using the GBS markers. A total of 8,725 high-quality single nucleotide polymorphisms (SNP) markers were obtained covering all 21 wheat chromosomes. PCA was performed using 8,725 SNPs to investigate relationships among studied lines (Figure 5.7). The first and second components explained 6.6% and 5.5% of the total variance, respectively. Overall, the absence of strong clustering based on PCA suggested close relationships between the lines used in this study indicating the suitability of this set of wheat breeding lines for evaluating GP models.

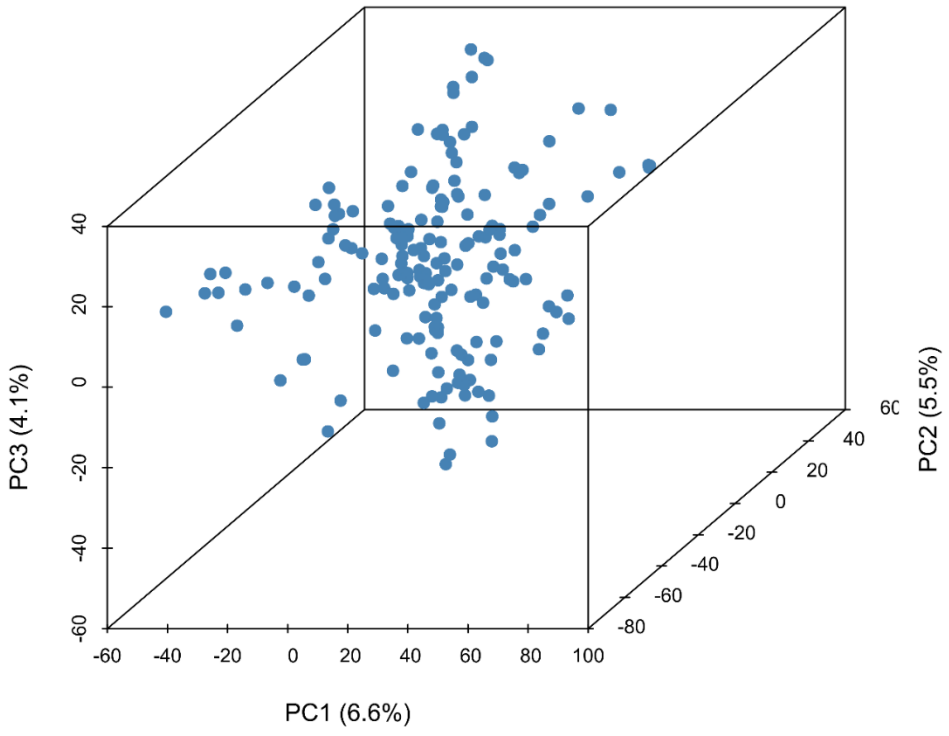

Figure 5.7 Principal component analysis for studied lines based on 8,725 single nucleotide polymorphism (SNP) markers.

5.4.3 Predictive ability of single trait models

We used STGP to assess the PA of end-use quality traits using CV1 scheme of cross-validation (Figure 5.3). Three different single-trait models (GBLUP, BA, and BB) were compared to identify the baseline for evaluation of different MT models. Overall, we did not observe any difference in the performance of these three ST models as all yielded comparable PA for different traits (Figure 5.8). Thus, the univariate GBLUP model (ST-GBLUP hereafter) was selected as a baseline to compare with the MT models. Prediction accuracies using ST-GBLUP varied from 0.23 to 0.54 for different end-use quality traits, with the lowest PA for MIXABS (0.26), and the highest PA for MIXTOL (0.54) (Figure 5.8).

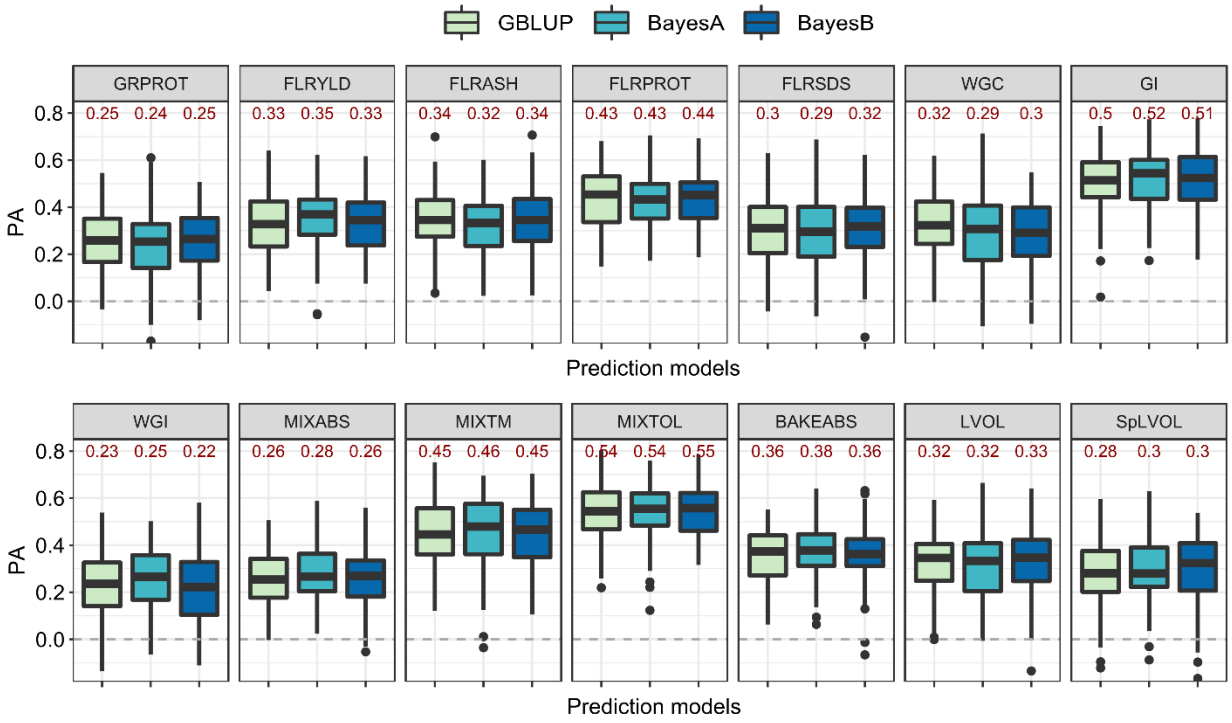

Figure 5.8 Prediction ability (PA) of different single-trait GP models for 14 end-use quality traits in cross-validation. GRPROT, grain protein content; FLRYLD, flour yield;

FLRASH, flour ash content; FLRPROT, flour protein content; FLRSDS, flour sedimentation weight value (%); WGC, wet gluten content; GI, gluten index; MIXABS, Mixograph mixing absorption (%), MIXTM, Mixograph mix time (min); MIXTOL, Mixograph mix tolerance score; BAKEABS, bake absorption; LVOL, pup loaf volume; SpLVOL, loaf volume by weight

5.4.4 MT models to predict Mixograph and Glutomatic traits

We evaluated the ability of MTGP models to predict a primary trait of interest by incorporating less resource-intensive secondary traits as covariates in the MTGP model and to determine the most effective combinations. The MT model was used to predict Mixograph traits, Glutomatic traits, and baking traits, which otherwise are time and resource intensive to phenotype. We evaluated the PA for three important Mixograph traits (MIXABS, MIXTM, and MIXTOL) using the MT model with secondary traits from rapid assays such as FLRPROT. A considerable improvement in PA for MIXABS was observed when MT model was used with different combinations of secondary traits (Figure 5.9; Appendix 5.1). The PA for MIXABS was 0.61 when FLRPROT was used as a secondary trait, and 0.64 when FLRPROT and FLRSDS were used together, which was higher than the PA for MIXABS using ST-GBLUP (0.26). For MIXTM, MT model using FLRSDS as covariate yielded higher PA (0.51) compared to the ST-GBLUP model (0.45). Similarly, the inclusion of FLRSDS in MT model yielded the highest PA (0.61) for MIXTOL. Overall, the inclusion of FLRPROT and FLRSDS as the secondary traits in MT model was effective to predict different Mixograph traits (Figure 5.9).

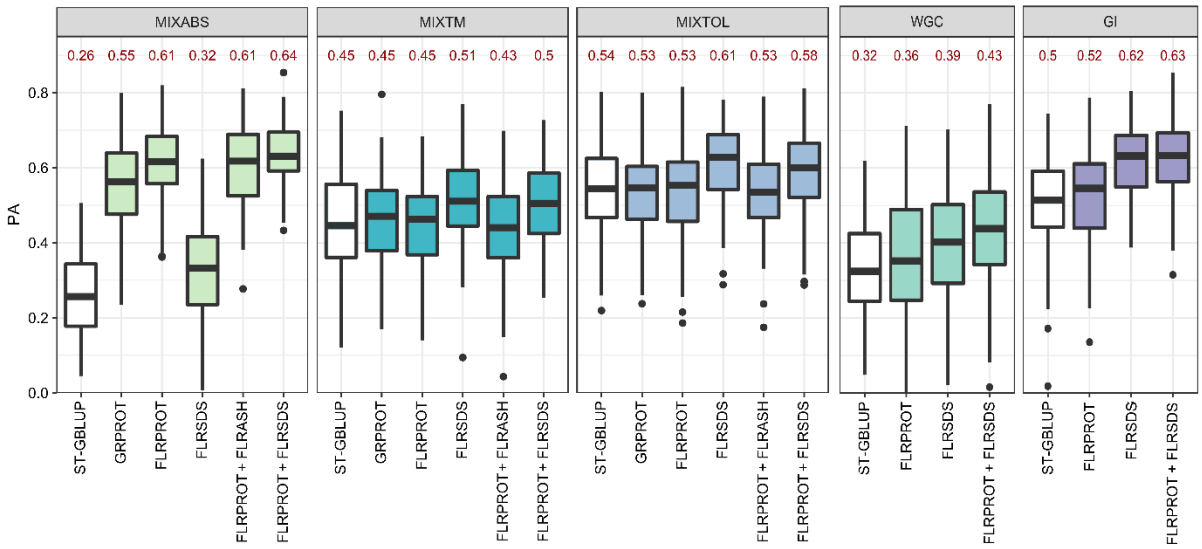

Figure 5.9 Prediction ability (PA) of the MTGP model for various Mixograph and Glutomatic traits using different combinations of secondary traits. The ST-GBLUP refers to the baseline single-trait GP model for the respective trait. GRPROT, grain protein content; FLRASH, flour ash content; FLRPROT, flour protein content; FLRSDS, flour sedimentation weight value (%); WGC, wet gluten content; GI, gluten index; MIXABS, Mixograph mixing absorption (%), MIXTM, Mixograph mix time (min); MIXTOL, Mixograh mix tolerance score.

Further, we evaluated different combinations of secondary traits to identify the best set of covariates to predict Glutomatic traits. We used three combinations (FLRPROT, FLRSDS, or FLRPROT + FLRSDS) of secondary traits to predict WGC and GI using the MT model (Figure 5.9; Appendix 5.2). The PA for WGC was the highest (0.43) when FLRPROT + FLRSDS were used as the covariates in the MT model, which was a substantial improvement over the ST-GBLUP model (0.32). For GI, the inclusion of FLRPROT + FLRSDS in the MT model yielded a PA of 0.63 compared to 0.50 for the ST-GBLUP model (Figure 5.9).

5.4.5 MT models to predict baking traits

BAKEABS and LVOL are important traits used by breeders to assess the end-use quality of yeast-leavened bread. Intending to predict the baking traits (BAKEABS, LVOL, and SpLVOL), we evaluated various combinations of secondary traits in the MT model (Figure 5.10; Appendix 5.3). These combinations were designed by selecting traits from different pre-baking assays performed at various scales. The ST-GBLUP showed a PA of 0.36 for BAKEABS (Figure 5.8), whereas a high PA (0.75) was observed when the MIXABS was included in the MT model (Figure 5.10). Intriguingly, incorporation of two easy-to-score traits, FLRPROT and FLRSDS, showed a high PA of 0.61. However, the addition of Glutomatic traits did not show any substantial improvement (0.66) in PA for BAKEABS (Figure 5.10).

The PA for LVOL ranged from 0.31 to 0.52 using different combinations of covariates in the MT genomic prediction model (Figure 5.10). The highest PA (0.52) for LVOL was obtained when both FLRPROT and FLRSDS were included in the MT model, which was considerably higher than the ST-GBLUP model (0.32) and also higher than the inclusion of only FLRPROT or FLRSDS in the MT model (Figure 5.10). Similar to BAKEABS, including Glutomatic traits in the MT model did not improve the PA for LVOL (Figure 5.10). For SpLVOL, the PA ranged from 0.30 to 0.45 (Figure 5.10) using the MT model with the highest PA from the inclusion of both FLRPROT and FLRSDS.

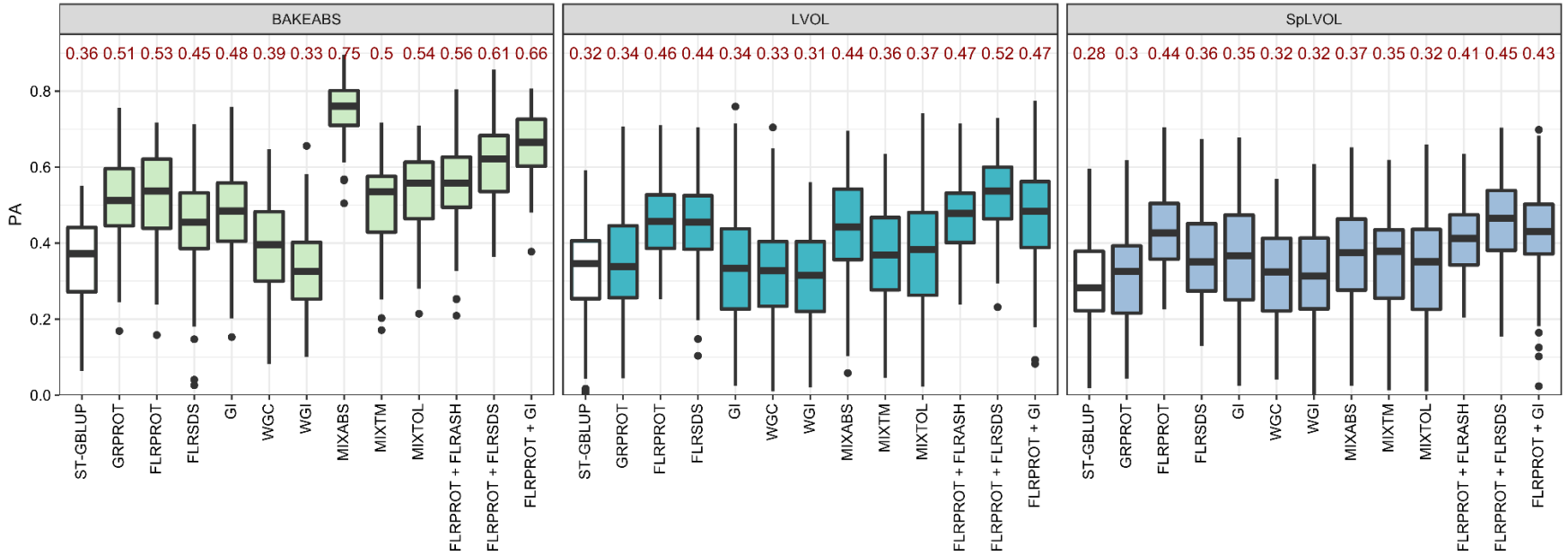

Figure 5.10 Prediction ability of the MT genomic prediction model for various baking traits using different combinations of secondary traits. The ST-GBLUP refers to the baseline single-trait GP model for the respective trait. GRPROT, grain protein content; FLRASH, flour ash content; FLRPROT, flour protein content; FLRSDS, flour sedimentation weight value (%); WGC, wet gluten content; GI, gluten index; WGI, wet gluten index; MIXABS, Mixograph mixing absorption (%), MIXTM, Mixograph mix time (min); MIXTOL, Mixograph mix tolerance score; BAKEABS, bake absorption; LVOL, pup loaf volume; SpLVOL, loaf volume by weight.

5.5 Discussion

In HWW breeding, end-use quality and processing traits are important factors in varietal development and determining acceptance by the industry. However, intensive selection for these traits is restricted to later generations in a breeding program because of expensive and time-consuming phenotyping requirements. Moreover, a short turnaround time in winter wheat breeding program creates another challenge in performing quality evaluations. For instance, winter wheat breeders particularly from the northern Great Plains get only a month between the harvest and planting to make selections, seed preparation and turn-around the breeding cycle, which eliminates the possibility to conduct comprehensive quality analysis on the lines that are to be advanced. This limits the selection decisions of quality traits to be made based on some easily measured traits or preliminary data obtained from prior year nurseries. Genomic selection may provide an alternative strategy to estimate GEBVs for these traits to cull inferior lines in earlier generations (Battenfield et al., 2016; Belamkar et al., 2018; Gill et al., 2021; Ward et al., 2019). However, conventional univariate genomic prediction models have shown a weak PA for complex end-use quality traits owing to their complex genetic architecture and low heritabilities (Battenfield et al., 2016; Sandhu et al., 2021; Zhang-Biehn et al., 2021). In recent years, MTGP models have been proposed to improve the PA of complex traits when phenotypic data for secondary traits are available (Bhatta et al., 2020; Gaire et al., 2022; Jia & Jannink, 2012; Lado et al., 2018; Zhang et al., 2022). In this study, we used MTGP models and evaluated their PA for various end-use quality traits measured at different stages of the breeding program. Rather than limiting to just model comparison, we evaluated several combinations of traits that can be used as covariates in the MT models to predict complex traits like BAKEABS and LVOL. Furthermore, we restricted the choice of secondary traits for MT models to the traits which are

easy to score, inexpensive, and can be assessed on a large number of lines in early generations (Figure 5.1). This is the first study in HWW to evaluate MTGP models with these combinations of secondary traits for their use in forward prediction.

We observed significant variation in the phenotypic distribution of all 14 end-use quality traits. Broad-sense heritability estimates for most of the traits were moderate, with a few traits exhibiting high heritability (Table 2), which corroborated most findings from previous studies (Battenfield et al., 2016; Michel et al., 2018; Sandhu et al., 2021; Sandhu et al., 2022; Zhang-Biehn et al., 2021). Traits including MIXTM, MIXTOL, GI, and FLRSDS had moderate to high heritability, suggesting that most variation associated with these traits can be attributed to genetic factors. As expected, all three baking traits showed low to moderate heritability (Table 2). This suggests the possibility of leveraging highly heritable traits in the MT models to predict traits like BAKEABS and LVOL. Further, we observed significant phenotypic and genetic correlations among a few pairs of traits (Figures 5.4 and 5.6). However, varying degrees of correlations among different quality tests suggested that no single test can completely substitute for actual testing of end-use quality traits (Battenfield et al., 2016; Souza et al., 2002). Intriguingly, a rapid flour sedimentation test (FLRSDS) was found to be significantly correlated with various Glutomatic, Mixograph, and baking traits. FLRSDS also exhibited the highest positive correlation with LVOL (Figure 5.5) in corroboration with Seabourn et. al. (2012). In contrast to Seabourn et. al. (2012), the correlation between FLRSDS and FLRPROT was not significant in our study which might be due to the fact that hybrid SDS-SRC assay was performed in year 3 using residual grain stored in the lab. Since rapid tests for FLRSDS can be performed on a large scale along with NIR for FLRPROT, the genetic correlation of these traits with LVOL can be exploited in multi-trait GP models to select quality traits in early generations.

Three univariate genomic prediction models were evaluated for predicting various end-use quality traits and selecting the best model for comparison with MT models. We did not observe significant differences in the performance among the univariate models GBLUP, BA, and BB. Previous studies have also reported similar outcomes while predicting complex traits (Sandhu, et al., 2021; Zhang et al., 2022). Further, most end-use quality traits had low to moderate PA using ST models (Figure 5.8). Interestingly, the Mixograph traits, MIXTM and MIXTOL, showed better PA using the ST model when compared with other traits likely because of their high broad-sense heritabilities.

As discussed earlier, breeding programs rely on various types of quality assays during different stages of variety development to select lines with desirable end-use quality (Figure 5.1). Quality tests including Mixograph, Glutomatic analysis, and baking are resource- and cost-intensive and can only be conducted in later generations of cultivar development due to limited grain quantity available in early generations (Figure 5.1). We used the MT genomic prediction model to assess the PA of various traits from these analyses using different combinations of secondary traits (FLRPROT and FLRSDS) that can be analyzed rapidly. Overall, the MT model outperformed the ST-GBLUP model for all the traits evaluated in this study (Figure 5.10). However, the extent of improvement using the MT model varied for different sets of traits. In corroboration with previous studies (Arojju et al., 2020; Gill et al., 2021; Rutkoski et al., 2016; Sun et al., 2017; Zhang et al., 2022), the improvement in PA using MT model relied on various factors including heritability of primary and secondary traits, different combinations of secondary traits, and genetic correlation between primary and secondary traits.

The ST-GBLUP model yielded a PA of 0.26, 0.45, and 0.54 for MIXABS, MIXTM, and MIXTOL, respectively (Figure 5.8). We used the MT model using flour characteristics from NIR

analysis and FLRSDS from the hybrid SDS-SRC test as secondary traits to predict MIXABS, MIXTM, and MIXTOL. For MIXABS, the MT model yielded a PA of 0.64 when FLRPROT and FLRSDS were used as the covariates, resulting in more than a two-fold increase in PA over the ST-GBLUP model. We also observed an improvement in PA using MT model for MIXTIM and MIXTOL and the highest PA was observed while using FLRPROT and FLRSDS as secondary traits (Figure 5.9). The improvement in PA for MIXTOL was less than for MIXABS and MIXTIM when the MT model was used. This relates to the high heritability observed for MIXTOL, suggesting that the MT models are more useful for traits exhibiting low heritability and thus lower PA when using the ST models (Gill et al., 2021; Jia & Jannink, 2012; Rutkoski et al., 2016; Ward et al., 2019). Similar to Mixograph, the use of MTGP outperformed the ST-GBLUP model in predicting GI and WGC with both FLRPROT and FLRSDS being the most effective combination of covariates.

The primary end-use product made from HWW is yeast-leavened bread. BAKEABS and LVOL are important traits used by breeders to assess end-use quality potential. However, evaluation of these traits is restricted to later-generation elite materials only. Thus, GS can be a promising approach to inform selection based on these traits in early generations. The MT model outperformed ST-GBLUP for predicting all baking traits (Figure 5.10). For BAKEABS, we observed a PA of 0.75 when MIXABS was included in the MT model. Nevertheless, inclusion of both FLRPROT and FLRSDS yielded a PA of 0.61, which was higher than the ST-GBLUP model (0.36). The Mixograph analysis has been a regular practice for assessing dough properties of the mid to late-generation samples of the HWW breeding programs. The Mixograph procedure requires only a small quantity of flour (10g) and various mixograph parameters are used to infer the mixing properties of flour and estimate the breadmaking water absorption and

loaf volume without actual baking. However, recent discontinuation of the Mixograph instrument and unavailability of alternatives will limit the breeder's capacity to get this information in earlier stages. Our results suggest that MTGP can provide an opportunity to directly predict BAKEABS and other baking traits using information from simple assays like FLRPROT and FLRSDS, and potentially offset the gap created due to the discontinuation of the mixograph. Further, inclusion of both FLRPROT and FLRSDS resulted in higher PA for LVOL and SpLVOL compared to other combinations of secondary traits (Figure 5.10). Results from this study suggest that the MTGP models can effectively predict various end-use quality traits, including baking traits. Moreover, a combination of secondary traits (FLRPROT and FLRSDS) resulted in substantial improvement in PA for baking traits. A previous HWW study also reported that inclusion of GRPROT or FLRPROT improved the PA for Mixograph and baking traits while inclusion of other traits as covariates was not useful (Zhang-Biehn et al., 2021). It is also noteworthy that the MT genomic prediction will be useful if it performs better than the routine indirect selection of primary trait(s) using highly correlated secondary traits. Except for Mixograph traits, we observed that the predictive abilities using MT models were higher than the phenotypic correlations between primary and secondary trait(s), especially for important baking traits (LVOL and BAKEABS). Additionally, it is expected that further optimization of training populations and increased data points will further improve the PA of MT models. Overall, our results suggest that the MT genomic prediction for end-use quality traits such as LVOL can be a part of the genomic prediction pipeline of the HWW breeding program along with other agronomic traits.

In conclusion, wheat breeding programs can routinely perform rapid assays such as NIR-based characterization of flour or sedimentation tests in an earlier generation of a breeding

program. These assays are cost-effective and do not require intensive resources. For instance, the hybrid SDS-SRC for FLRSDS can be performed using 1g of whole wheat flour and 25-30 samples can be evaluated per hour (Seabourn et al., 2012). Similarly, flour characteristics such as FLRPROT can be quickly assayed using NIR-based spectroscopy and some newer NIR also provide an estimation of absorption and gluten index which can be further evaluated. Further, the availability of low-cost genotyping platforms has made it possible for breeding programs to sequence a large number of lines in earlier generations. Thus, the availability of these resources can help to implement MTGS for predicting traits like LVOL that are difficult to phenotype and exhibit low heritability. The MTGP models can be employed by combining FLRPROT and FLRSDS evaluated from earlier generations with a complete quality profile (including baking) from the advanced generations to predict baking traits in earlier generations (PYT or EOT). This will not only save considerable time and resources but will provide an opportunity for breeders to eliminate inferior material in earlier generations. Therefore, MTGS holds great promise in improving the selection efficiency of processing and end-use quality traits in hard winter wheat.

5.6 References

AACC International. (2000). *Approved methods of the American Association of Cereal Chemists, 10th Edition*. https://doi.org/https://www.cerealsgrains.org/resources/Methods/tools/Documents

AACC International. (2011). *Approved methods of analysis. American Association of Cereal Chemists,11th edition.* https://www.cerealsgrains.org/resources/Methods/Pages/default.aspx

Arojju, S. K., Cao, M., Trolove, M., Barrett, B. A., Inch, C., Eady, C., Stewart, A., & Faville, M. J. (2020). Multi-Trait Genomic Prediction Improves Predictive Ability for Dry Matter Yield and Water-Soluble Carbohydrates in Perennial Ryegrass. *Frontiers in Plant Science*, *11*, 1.

https://doi.org/10.3389/fpls.2020.01197

Bai, G., Kolb, F. L., Shaner, G., & Domier, L. L. (1999). Amplified fragment length polymorphism markers linked to a major quantitative trait locus controlling scab resistance in wheat. *Phytopathology*, *89*(4), 343–348. https://doi.org/10.1094/PHYTO.1999.89.4.343

Bassi, F. M., Bentley, A. R., Charmet, G., Ortiz, R., & Crossa, J. (2015). Breeding schemes for the implementation of genomic selection in wheat (Triticum spp.). *Plant Science*, *242*, 23–36. https://doi.org/10.1016/j.plantsci.2015.08.021

Battenfield, S. D., Guzmán, C., Gaynor, R. C., Singh, R. P., Peña, R. J., Dreisigacker, S., Fritz, A. K., & 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*(2). https://doi.org/10.3835/plantgenome2016.01.0005

Belamkar, V., Guttieri, M. J., Hussain, W., Jarquín, D., El-basyoni, I., Poland, J., Lorenz, A. J., & Baenziger, P. S. (2018). Genomic selection in preliminary yield trials in a winter wheat breeding program. *G3: Genes, Genomes, Genetics*, *8*(8), 2735–2747. https://doi.org/10.1534/g3.118.200415

Bhatta, M., Gutierrez, L., Cammarota, L., Cardozo, F., Germán, S., Gómez-Guerrero, B., Pardo, M. F., Lanaro, V., Sayas, M., & Castro, A. J. (2020). Multi-trait genomic prediction model increased the predictive ability for agronomic and malting quality traits in barley (Hordeum vulgare L.). *G3: Genes, Genomes, Genetics*, *10*(3), 1113–1124. https://doi.org/10.1534/g3.119.400968

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

Bioinformatics, *23*(19), 2633–2635. https://doi.org/10.1093/bioinformatics/btm308

Browning, S. R., & Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics*, *81*(5), 1084–1097. https://doi.org/10.1086/521987

Butler, D. G., Cullis, B. R., Gilmour, A. R., Gogel, B. J., & Thompson, R. (2018). *ASReml-R Reference Manual Version 4 ASReml estimates variance components under a general linear mixed model by residual maximum likelihood (REML)*. https://asreml.kb.vsni.co.uk/wp-content/uploads/sites/3/ASReml-R-Reference-Manual-4.pdf

Carter, A. H., Garland-Campbell, K., Morris, C. F., & Kidwell, K. K. (2012). Chromosomes 3B and 4D are associated with several milling and baking quality traits in a soft white spring wheat (Triticum aestivum L.) population. *Theoretical and Applied Genetics*, *124*(6), 1079–1096. https://doi.org/10.1007/s00122-011-1770-x

Cullis, B. R., Smith, A. B., & Coombes, N. E. (2006). On the design of early generation variety trials with correlated data. *Journal of Agricultural, Biological, and Environmental Statistics*, *11*(4), 381–393. https://doi.org/10.1198/108571106X154443

de los Campos, G., & Grüneberg, A. (2016). *MTM Package.* https://github.com/QuantGen/MTM/

Gaire, R., Arruda, M. P., Mohammadi, M., Brown-Guedira, G., Kolb, F. L., & Rutkoski, J. (2022). Multi-trait genomic selection can increase selection accuracy for deoxynivalenol accumulation resulting from fusarium head blight in wheat. *The Plant Genome*, *15*(1), e20188. https://doi.org/10.1002/tpg2.20188

Gill, H. S., Halder, J., Zhang, J., Brar, N. K., Rai, T. S., Hall, C., Bernardo, A., Amand, P. S., Bai, G., Olson, E., Ali, S., Turnipseed, B., & Sehgal, S. K. (2021). Multi-Trait Multi-Environment Genomic Prediction of Agronomic Traits in Advanced Breeding Lines of Winter Wheat. *Frontiers in Plant Science*, *12*. https://doi.org/10.3389/fpls.2021.709545

Gill, H. S., Halder, J., Zhang, J., Rana, A., Kleinjan, J., Amand, P. St., Bernardo, A., Bai, G., & Sehgal, S. K. (2022). Whole-genome analysis of hard winter wheat germplasm identifies genomic regions associated with spike and kernel traits. *Theoretical and Applied Genetics*, *1*, 3. https://doi.org/10.1007/s00122-022-04160-6

Hayes, B. J., Panozzo, J., Walker, C. K., Choy, A. L., Kant, S., Wong, D., Tibbits, J., Daetwyler, H. D., Rochfort, S., Hayden, M. J., & Spangenberg, G. C. (2017). Accelerating wheat breeding for end-use quality with multi-trait genomic predictions incorporating near infrared and nuclear magnetic resonance-derived phenotypes. *Theoretical and Applied Genetics, 130*(12), 2505–2519. https://doi.org/10.1007/s00122-017-2972-7

Heffner, E. L., Sorrells, M. E., & Jannink, J. L. (2009). Genomic selection for crop improvement. In *Crop Science* (Vol. 49, Issue 1, pp. 1–12). https://doi.org/10.2135/cropsci2008.08.0512

Ibba, M. I., Crossa, J., Montesinos-López, O. A., Montesinos-López, A., Juliana, P., Guzman, C., Delorean, E., Dreisigacker, S., & Poland, J. (2020). Genome-based prediction of multiple wheat quality traits in multiple years. *The Plant Genome*, *13*(3). https://doi.org/10.1002/tpg2.20034

IWGSC. (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science (New York, N.Y.)*, *361*(6403), eaar7191. https://doi.org/10.1126/science.aar7191

Jernigan, K. L., Godoy, J. V., Huang, M., Zhou, Y., Morris, C. F., Garland-Campbell, K. A., Zhang, Z., & Carter, A. H. (2018). Genetic Dissection of End-Use Quality Traits in Adapted Soft White Winter Wheat. *Frontiers in Plant Science*, *9*, 271. https://doi.org/10.3389/fpls.2018.00271

Jia, Y., & Jannink, J. L. (2012). Multiple-trait genomic selection methods increase genetic value prediction accuracy. *Genetics*, *192*(4), 1513–1522. https://doi.org/10.1534/genetics.112.144246

Juliana, P., Poland, J., Huerta-Espino, J., Shrestha, S., Crossa, J., Crespo-Herrera, L., Toledo, F. H., Govindan, V., Mondal, S., Kumar, U., Bhavani, S., Singh, P. K., Randhawa, M. S., He, X., Guzman, C., Dreisigacker, S., Rouse, M. N., Jin, Y., Pérez-Rodríguez, P., … Singh, R. P. (2019). Improving grain yield, stress resilience and quality of bread wheat using large-scale genomics. *Nature Genetics*, *51*(10), 1530–1539. https://doi.org/10.1038/s41588-019-0496-6

Juliana, P., Singh, R. P., Braun, H.-J., Huerta-Espino, J., Crespo-Herrera, L., Govindan, V., Mondal, S., Poland, J., & Shrestha, S. (2020). Genomic Selection for Grain Yield in the CIMMYT Wheat Breeding Program—Status and Perspectives. *Frontiers in Plant Science*, *11*, 1. https://doi.org/10.3389/fpls.2020.564183

Juliana, P., Singh, R. P., Singh, P. K., Crossa, J., Huerta-Espino, J., Lan, C., Bhavani, S., Rutkoski, J. E., Poland, J. A., Bergstrom, G. C., & Sorrells, M. E. (2017). Genomic and pedigree-based prediction for leaf, stem, and stripe rust resistance in wheat. *Theoretical and Applied Genetics*, *130*(7), 1415–1430. https://doi.org/10.1007/s00122-017-2897-1

Kiszonas, A. M., & Morris, C. F. (2017). Wheat Breeding for Quality: A Historical Review.

Cereal Chemistry Journal, *95*(1), CCHEM-05-17-0103-FI. https://doi.org/10.1094/CCHEM-05-17-0103-FI

Lado, B., Vázquez, D., Quincke, M., Silva, P., Aguilar, I., & Gutiérrez, L. (2018). Resource allocation optimization with multi-trait genomic prediction for bread wheat (Triticum aestivum L.) baking quality. *Theoretical and Applied Genetics*, *131*(12), 2719–2731. https://doi.org/10.1007/s00122-018-3186-3

Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. In *Genetics Soc America*. https://www.genetics.org/content/157/4/1819.short

Michel, S., Kummer, C., Gallee, M., Hellinger, J., Ametz, C., Akgöl, B., Epure, D., Löschenberger, F., & Buerstmayr, H. (2018). Improving the baking quality of bread wheat by genomic selection in early generations. *Theoretical and Applied Genetics*, *131*(2), 477–493. https://doi.org/10.1007/s00122-017-2998-x

Pérez, P., & De Los Campos, G. (2014). Genome-wide regression and prediction with the BGLR statistical package. *Genetics*, *198*(2), 483–495. https://doi.org/10.1534/genetics.114.164442

Poland, J., Endelman, J., Dawson, J., Rutkoski, J., Wu, S., Manes, Y., Dreisigacker, S., Crossa, J., Sánchez-Villeda, H., Sorrells, M., & Jannink, J. (2012). Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing. *The Plant Genome*, *5*(3), plantgenome2012.06.0006. https://doi.org/10.3835/plantgenome2012.06.0006

R Core Team. (2018). *R: A language and environment for statistical computing; 2015*. https://scholar.google.com/scholar?cluster=9441913529578809097&hl=en&as_sdt=5,42&sciodt=0,42

Roberts, S., Brooks, K., Nogueira, L., & Walters, C. G. (2022). The role of quality characteristics in pricing hard red winter wheat. *Food Policy, 108,* 102246. https://doi.org/10.1016/j.foodpol.2022.102246

Rutkoski, J., Benson, J., Jia, Y., Brown-Guedira, G., Jannink, J.-L., & Sorrells, M. (2012). Evaluation of Genomic Prediction Methods for Fusarium Head Blight Resistance in Wheat. *The Plant Genome, 5*(2), 51–61. https://doi.org/10.3835/plantgenome2012.02.0001

Rutkoski, J., Poland, J., Mondal, S., Autrique, E., Pérez, L. G., Crossa, J., Reynolds, M., & Singh, R. (2016). Canopy temperature and vegetation indices from high-throughput phenotyping improve accuracy of pedigree and genomic selection for grain yield in wheat. *G3: Genes, Genomes, Genetics*, *6*(9), 2799–2808. https://doi.org/10.1534/g3.116.032888

Sandhu, K., Aoun, M., Morris, C., & Carter, A. (2021). Genomic Selection for End-Use Quality and Processing Traits in Soft White Winter Wheat Breeding Program with Machine and Deep Learning Models. *Biology, 10*(7), 689. https://doi.org/10.3390/biology10070689

Sandhu, K., Patil, S. S., Pumphrey, M., & Carter, A. (2021). Multitrait machine- and deep-learning models for genomic selection using spectral information in a wheat breeding program. *The Plant Genome, 14*(3), e20119. https://doi.org/10.1002/tpg2.20119

Sandhu, K. S., Patil, S. S., Aoun, M., & Carter, A. H. (2022). Multi-Trait Multi-Environment Genomic Prediction for End-Use Quality Traits in Winter Wheat. *Frontiers in Genetics, 13*, 41. https://doi.org/10.3389/fgene.2022.831020

Seabourn, B. W., Xiao, Z. S., Tilley, M., Herald, T. J., & Park, S. H. (2012). A rapid, small-scale sedimentation method to predict breadmaking quality of hard winter wheat. *Crop Science*, *52*(3), 1306–1315. https://doi.org/10.2135/cropsci2011.04.0210

Souza, E. J., Guttieri, M. J., & Graybosch, R. A. (2002). Breeding wheat for improved milling and baking quality. In *Journal of Crop Production* (Vol. 5, Issues 1–2, pp. 39–74). Taylor & Francis Group . https://doi.org/10.1300/J144v05n01_03

Sun, J., Rutkoski, J. E., Poland, J. A., Crossa, J., Jannink, J., & 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*(2). https://doi.org/10.3835/plantgenome2016.11.0111

USDA ERS. (2022). *USDA Economic Reseasrch Service: Wheat Data*. https://www.ers.usda.gov/data-products/wheat-data/

USDA NASS. (2021). *Small Grains 2021 Summary*. https://www.nass.usda.gov/Publications/Todays_Reports/reports/smgr0921.pdf

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

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

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York. https://cran.r-project.org/web/packages/ggplot2/citation.html

William, R. (2013). *psych: Procedures for Personality and Psychological Research*. http://cran.r-project.org/package=psych

Zhang-Biehn, S., Fritz, A. K., Zhang, G., Evers, B., Regan, R., & Poland, J. (2021). Accelerating wheat breeding for end-use quality through association mapping and multivariate genomic prediction. *The Plant Genome*, *14*(3), e20164. https://doi.org/10.1002/tpg2.20164

Zhang, J., Gill, H. S., Brar, N. K., Halder, J., Ali., S., Liu, X., Bernardo, A., St Amand, P., Bai, G., Gill, U. S., Turnipseed, B., & Sehgal, S. K. (2022). Genomic prediction of Fusarium head blight resistance in early stages using advanced breeding lines in hard winter wheat. *The Crop Journal*. https://doi.org/10.1016/j.cj.2022.03.010

Zhu, T., Wang, L., Rimbert, H., Rodriguez, J. C., Deal, K. R., De Oliveira, R., Choulet, F., Keeble-Gagnère, G., Tibbits, J., Rogers, J., Eversole, K., Appels, R., Gu, Y. Q., Mascher, M., Dvorak, J., & Luo, M. (2021). Optical maps refine the bread wheat *Triticum aestivum* cv. Chinese Spring genome assembly. *The Plant Journal*, *107*(1), 303–314. https://doi.org/10.1111/tpj.15289

APPENDICES

Appendix 3.1 Details of all significant marker-trait associations (MTAs) identified by genome-wide association studies (GWAS) for spike and kernel traits in individual environments (E1, E2, E3, and E4) and combined analysis (CEnv).

Trait[a]	Env[b]	SNP	Chr	Position[c]	*P*.value	MAF	FDR-adj (*P*)	-$\log_{10}$(*P*)
KL	E3	S1A_299864277	1A	299,864,277	2.90E-06	0.07	0.012	5.53
	E4	S1A_299864277	1A	299,864,277	2.80E-05	0.07	0.112	4.56
	E1	S2A_17292018	2A	17,292,018	8.30E-07	0.38	0.003	6.08
	E3	S2B_660658322	2B	660,658,322	4.40E-05	0.18	0.089	4.35
	E1	S5A_413179144	5A	413,179,144	5.30E-05	0.48	0.107	4.27
	E1	S5A_413179162	5A	413,179,162	5.30E-05	0.48	0.107	4.27
	E3	S5A_455648025	5A	455,648,025	6.90E-07	0.15	0.006	6.16
	E3	S5A_476898590	5A	476,898,590	2.90E-05	0.41	0.079	4.53
	E1	S7A_717859384	7A	717,859,384	4.50E-08	0.13	4.00E-04	7.35
	E4	S7A_717859384	7A	717,859,384	4.10E-06	0.13	0.033	5.39
KW	E1	S1A_375151331	1A	375,151,331	3.10E-07	0.28	0.001	6.51
	E1	S2A_41083010	2A	41,083,010	6.20E-06	0.42	0.01	5.21
	E1	S2B_194454645	2B	194,454,645	6.50E-06	0.11	0.01	5.19
	E1	S2B_64463130	2B	64,463,130	1.20E-05	0.14	0.014	4.92
	E2	S2D_49291847	2D	49,291,847	1.90E-07	0.06	0.002	6.71
	E2	S3B_53438124	3B	53,438,124	4.00E-06	0.08	0.011	5.4
	E2	S4A_289871521	4A	289,871,521	2.40E-05	0.13	0.047	4.63
	E3	S4A_619197841	4A	619,197,841	3.20E-08	0.31	3.00E-04	7.5
	E1	S4A_625521699	4A	625,521,699	5.20E-08	0.49	4.00E-04	7.29
	E1	S5D_554548588	5D	554,548,588	9.00E-06	0.3	0.012	5.04
	E2	S7B_659759723	7B	659,759,723	1.40E-06	0.07	0.006	5.85
	E1	S7D_60662020	7D	60,662,020	1.20E-06	0.47	0.003	5.91
SD	E1	S1A_11441837	1A	11,441,837	1.50E-05	0.17	0.121	4.82
	E2	S1B_571072942	1B	571,072,942	5.20E-07	0.08	0.002	6.28
	E4	S1B_661583505	1B	661,583,505	1.90E-07	0.32	0.002	6.72

	E2	S3B_49841949	3B	49,841,949	7.20E-09	0.06	6.00E-05	8.14
	E3	S3D_71951480	3D	71,951,480	3.00E-07	0.09	0.002	6.52
	E4	S7D_51431227	7D	51,431,227	2.20E-06	0.07	0.009	5.66
SL	E2	S1A_13099591	1A	13,099,591	7.20E-06	0.08	0.029	5.14
	E4	S1A_13099591	1A	13,099,591	4.40E-05	0.08	0.071	4.36
	CEnv	S2B_16305395	2B	16,305,395	4.90E-06	0.07	0.023	5.31
	E4	S2B_16305395	2B	16,305,395	1.40E-05	0.08	0.038	4.84
	E4	S4B_593675948	4B	593,675,948	1.30E-06	0.11	0.005	5.88
	CEnv	S5B_432612793	5B	432,612,793	5.80E-06	0.14	0.023	5.24
	E2	S5B_432612793	5B	432,612,793	1.40E-07	0.14	0.001	6.86
	E4	S5B_432612793	5B	432,612,793	2.50E-07	0.14	0.002	6.6
	E1	S6B_619882604	6B	619,882,604	3.60E-06	0.08	0.014	5.45
	E4	S6B_622447936	6B	622,447,936	2.20E-05	0.25	0.044	4.66
	CEnv	S6B_667230947	6B	667,230,947	6.7504E-05	0.17	0.180	4.17
	E1	S7A_236941510	7A	236,941,510	1.10E-06	0.35	0.009	5.97
	E2	S7A_676732614	7A	676,732,614	1.20E-05	0.14	0.033	4.91
NSPS	E2	S2A_95491711	2A	95,491,711	5.70E-07	0.14	9.00E-04	6.25
	E2	S2B_16305395	2B	16,305,395	6.40E-06	0.07	0.007	5.19
	E3	S2B_16305395	2B	16,305,395	1.40E-05	0.07	0.029	4.84
	E4	S3A_647983369	3A	647,983,369	1.50E-05	0.09	0.03	4.83
	E2	S3A_651040581	3A	651,040,581	2.70E-07	0.06	5.00E-04	6.57
	E4	S5B_432612793	5B	432,612,793	4.00E-05	0.14	0.065	4.4
	E2	S5D_551528813	5D	551,528,813	6.10E-06	0.46	0.007	5.21
	E2	S7A_132414615	7A	132,414,615	6.50E-09	0.22	5.00E-05	8.18
	CEnv	S7A_132597623	7A	132,597,623	5.40E-09	0.26	4.00E-05	8.27
	E3	S7A_132597623	7A	132,597,623	7.90E-09	0.26	6.00E-05	8.1
	E4	S7A_132597623	7A	132,597,623	3.30E-08	0.26	3.00E-04	7.48
	E3	S7A_676621121	7A	676,621,121	1.00E-07	0.15	4.00E-04	6.99
	E2	S7A_676732614	7A	676,732,614	2.40E-08	0.15	1.00E-04	7.61
	E4	S7A_676732614	7A	676,732,614	9.60E-06	0.16	0.026	5.02
	CEnv	S7A_682556399	7A	682,556,399	9.20E-08	0.08	4.00E-04	7.04
	E3	S7B_634580423	7B	634,580,423	6.00E-06	0.23	0.016	5.22
	E4	S7B_86687472	7B	86,687,472	3.20E-07	0.08	0.001	6.49
	E2	S7D_532455282	7D	532,455,282	2.70E-07	0.25	5.00E-04	6.57

TKW	CEnv	S1A_58837340	1A	58,837,340	1.30E-05	0.15	0.026	4.89
	E1	S1D_203599718	1D	203,599,718	7.00E-06	0.08	0.028	5.16
	CEnv	S2B_66706142	2B	66,706,142	2.20E-06	0.14	0.009	5.66
	E2	S3A_60583645	3A	60,583,645	3.60E-06	0.42	0.015	5.44
	CEnv	S5A_476847493	5A	476,847,493	1.20E-05	0.35	0.026	4.93
	E2	S5A_5185086	5A	5,185,086	8.10E-06	0.16	0.022	5.09
	CEnv	S7D_60662020	7D	60,662,020	7.50E-11	0.48	6.00E-07	10.12
	E1	S7D_60662020	7D	60,662,020	5.10E-09	0.48	4.00E-05	8.3
	E2	S7D_60662020	7D	60,662,020	3.00E-09	0.48	2.00E-05	8.52
	E3	S7D_60662020	7D	60,662,020	3.80E-07	0.48	0.003	6.42

[a]SL, spike length; NSPS, spikelet number per spike; SD, spikelet density; TKW, thousand kernel weight; KL, kernel length; KW, kernel width; KA, kernel area

[b]Environments

[c]Physical position is based on IWGSC RefSeq v2.0 (IWGSC, 2018)

Appendix 3.2 Allelic distribution for selected MTAs in a set of selected elite winter wheat genotypes including released cultivars and breeding lines from the Great Plains region of the US. The alleles are coded as '0' for non-favorable allele, '1' for favorable allele, and '2' for missing values or heterozygotes.

Accession	Allelic constitution for selected MTAs						
	S1A_1 3099591	S2B_1 6305395	S5B_43 2612793	S6B_61 9882604	S7A_1 3241461	S7A_6 7673261	S7D_606 62020
1863	0	1	0	1	0	1	0
ALICE	2	1	0	1	0	1	0
ANTERO	0	1	2	1	0	1	1
ARAPAHOE	0	1	0	1	1	1	0
ART	0	1	0	1	1	1	0
BRYD	0	1	0	1	0	1	1
CEDER	0	0	0	1	0	1	1
CO09W040-F1	0	1	0	1	0	1	0
CO11D1316W	0	1	0	0	0	1	0
CO11D174	0	1	0	1	0	1	1
CO11D1767	0	1	0	0	0	0	1
CO11D346	0	1	0	0	0	1	1
CO11D446	0	1	2	1	0	1	1
DECADE	0	1	0	1	0	0	2
DENALI	0	1	0	0	0	1	1
EMERSON	0	1	1	2	0	2	1
EVEREST	0	1	2	1	0	1	1
EXPEDITION	0	1	0	1	0	1	1
FLATHEAD	0	1	0	1	0	1	0
FLOURISH	2	1	1	1	0	0	1
FREEMAN	0	1	0	1	0	1	1
GUARDIAN	0	1	0	1	0	1	1

HATCHER	0	0	0	1	0	1	1
IDEAL	0	1	0	0	0	1	1
JERRY	0	1	0	1	0	1	0
KELDIN	0	1	2	1	0	1	1
KS060084-M-4	0	1	0	1	0	1	0
KS060106-M-11	0	1	0	1	1	0	1
KS060476-M-6	0	1	0	1	2	1	0
KS11HW39-5-4	0	1	0	1	0	0	2
KS11HW39-6	0	1	0	1	0	1	0
LANGIN	0	1	2	1	0	1	1
LYMAN	0	1	2	1	0	1	1
MILLENIUM	0	1	0	1	0	1	1
MT1090	0	1	0	1	0	1	0
NE10507	0	1	0	1	0	0	0
NE10589	0	0	0	1	0	1	0
NE12444	0	1	0	1	0	2	0
NH11489	0	1	0	1	0	1	0
NI9710H	0	0	0	1	0	1	0
NW09627	0	0	0	1	0	1	0
OAHE	0	1	1	1	0	1	1
OK09125	0	1	0	1	0	0	1
OK10728W	0	1	0	1	0	0	0
OVERLAND	0	1	0	1	0	1	1
OVERLANDFH B1	0	1	0	1	0	1	0
REDFIELD	2	1	2	1	0	1	2
REDHAWK	0	1	0	1	0	1	0
ROBIDOUX	0	0	0	1	1	1	1
RUTH	0	0	0	1	0	1	0
SD08080	0	1	0	1	0	1	1
SD08200	0	1	0	1	0	1	1

SD09113	0	1	0	1	1	1	0
SD09138	0	1	0	1	1	1	0
SD09140	0	1	0	1	1	1	1
SD10026-3-1	0	1	0	1	0	1	1
SD10135	0	0	0	1	0	1	1
SD10257-2	0	1	0	1	0	1	2
SD10W089-3-5	1	1	0	1	0	1	0
SD10W153	0	1	0	1	1	1	0
SD11002-2	1	1	0	1	0	1	1
SD110036-2	0	1	0	1	2	1	0
SD110036-4	0	1	0	1	1	1	0
SD110038-3	1	1	0	1	0	0	0
SD110039-2	0	1	0	1	0	1	1
SD110040-5	0	1	0	1	0	1	1
SD110041-4	0	1	0	1	0	1	1
SD110044-6	0	1	0	1	2	1	2
SD110044-7	0	2	0	1	0	2	0
SD110049-7	0	1	0	1	1	1	1
SD110054-3	0	1	0	1	2	1	2
SD11005-5	0	1	0	0	0	1	0
SD110060-10	0	1	0	1	0	1	1
SD110060-7	0	1	0	1	0	1	1
SD110060-9	0	1	0	2	0	1	1
SD110085-1	0	1	0	1	0	1	0
SD110085-3	0	1	0	1	0	1	0
SD11009-5	1	1	0	1	1	1	0
SD11018-7	0	1	0	1	0	1	0
SD11023-8	1	1	0	1	2	1	0
SD12DHA00031	0	2	0	1	0	1	2
SD12DHA00969	0	1	1	1	0	2	0
SD12DHA01024	2	1	0	2	2	1	0

SD12DHA01038	0	1	1	1	0	1	1
SD12DHA01043	0	1	2	1	2	1	0
SD12DHA01131	2	1	2	1	2	1	0
SD12DHA01347	0	1	0	2	2	1	2
SD12DHA01353	0	2	0	1	2	1	0
SD12DHA01373	0	1	2	1	0	1	2
SD12DHA01556	0	1	0	0	2	1	0
SD12DHA01688	1	0	2	1	0	1	0
SD12DHA02135	0	1	0	1	0	1	0
SD12DHA03282	1	1	1	1	0	1	0
SD12DHA03290	0	1	0	2	0	0	2
SD12DHA03429	0	1	0	2	0	1	1
SD12DHA03614	0	1	0	1	0	1	1
SD13073-1	2	1	0	1	0	1	0
SD13153-3	0	0	0	1	0	1	1
SD13DHA02337	1	1	0	2	0	1	1
SD13DHA02346	0	1	0	2	2	1	0
SD13DHA02489	1	1	2	0	2	2	0
SD13DHA02497	0	1	0	2	0	1	0
SD13DHA02641	2	1	0	1	2	1	0
SD13W036-3	0	1	0	1	0	0	0
SD14059-4	0	1	0	1	2	1	0
SD14074-3	0	1	0	2	2	1	1
SD14076-1	0	1	2	1	0	1	2
SD14113-3	1	1	0	1	0	1	1
SD14115-5	1	1	0	1	0	1	0
SD14163-2	0	0	0	0	0	1	1
SD14182-1	0	1	0	1	0	2	0
SD14208-1	0	1	0	2	0	0	1
SD14239-2	2	1	2	1	2	1	2
SD14295-3	0	1	2	1	0	1	0

SD14303-3	1	2	0	1	0	1	0
SD14351-1	0	1	0	1	2	1	0
SD14355-2	0	1	2	2	1	1	0
SD14373-5	0	1	0	1	0	1	1
SD15002-8	0	1	0	0	2	2	2
SD15004-2	0	1	2	1	0	1	1
SD15007-11	0	1	0	2	0	1	2
SD15007-5	0	1	0	0	0	1	1
SD15009-1	0	2	2	1	1	2	0
SD15009-2	0	1	2	1	2	2	0
SD15025-1	0	1	0	1	0	1	0
SD15035-2	1	1	2	1	0	1	0
SD15050-2	0	1	2	1	0	1	0
SD15081-6	0	1	0	1	2	1	2
SD15083-2	0	1	2	1	0	1	1
SD15103-6	0	1	0	1	0	2	0
SD15108-1	0	2	2	2	0	1	1
SD15164-1	0	1	0	1	0	2	2
SD15205-1	0	1	0	0	0	0	2
SD15232-2	0	1	0	1	0	1	2
SD15240-2	2	1	0	1	0	2	2
SD16001	0	1	2	1	0	1	0
SD16006-3	0	2	2	1	0	1	2
SD16008-7	0	1	2	1	2	1	0
SD16010	0	1	0	1	1	1	2
SD17032	0	1	2	1	0	1	1
SD17078	0	1	0	1	1	1	1
SD17141	0	1	0	1	1	1	0
SD17181	0	1	0	1	0	2	0
SD17210	0	1	0	1	1	0	0
SD17246	1	1	0	1	0	1	0

SD17371	0	1	0	1	1	1	0
SD17420	0	1	2	1	1	1	1
SD18001-7	0	1	1	1	1	2	1
SD18003-11	1	1	1	1	0	1	0
SD18003-8	0	1	0	1	0	1	0
SD18005-1	0	1	0	1	0	1	2
SD18006-4	0	1	2	1	0	1	1
SD18007-1	1	1	0	1	0	1	0
SD18009-4	0	1	2	1	0	1	0
SD18012-5	0	1	2	1	0	1	0
SD18019-1	0	1	0	1	2	0	1
SD18020-2	1	1	2	1	1	2	0
SD18022-5	2	1	2	1	0	2	1
SD18023-3	0	1	0	1	0	1	0
SD18023-8	0	1	0	1	0	1	0
SD18025-8	0	1	1	1	0	2	1
SD18036-1	0	1	2	1	1	1	0
SD18037-7	0	1	2	1	2	1	1
SD18038-2	0	1	0	1	2	1	2
SD18039-2	0	1	1	1	1	1	1
SD18042-6	0	1	0	1	0	1	0
SD18067-9	1	1	0	1	1	1	1
SD18069-9	0	1	0	0	0	1	1
SD18072-2	2	1	2	1	0	1	1
SD18076-1	0	1	2	1	0	1	0
SD18076-2	0	1	0	1	1	1	2
SD18080-2	0	1	0	1	0	1	1
SD18083-8	0	1	0	1	1	1	1
SD18087-4	0	1	1	1	1	1	0
SD18113-1	0	1	2	1	0	1	1
SD18213-6	1	1	0	1	0	1	1

SD18231-8	0	1	0	1	1	0	0
SD18249-3	0	1	1	1	0	2	0
SD18272-3	0	1	0	1	1	1	1
SD19002-1	0	1	0	1	1	1	1
SD19002-2	0	1	0	1	2	1	1
SD19008-1	0	1	0	1	0	1	1
SD19011-2	0	1	0	1	2	1	1
SD19017-5	0	1	2	1	1	0	0
SD19019-2	0	1	0	1	1	1	0
SD19020-2	0	1	0	1	1	1	1
SD19033-2	0	1	0	1	1	1	0
SD19041-1	0	1	2	1	1	1	1
SD9140	0	1	0	1	1	1	0
SMOKY_HILL	0	1	0	1	1	1	0
T158	0	1	0	1	0	1	1
T163	0	1	0	1	0	2	1
THOMPSON	0	1	2	1	1	1	0
TX08A001249	0	1	0	1	0	1	0
TX08V7313	0	1	0	1	0	1	1
TX09A0091194	0	1	0	1	0	1	0
TX11A001295	0	1	0	1	0	1	1
TX12M4063	0	1	0	1	0	1	2
TX12M4065	0	1	0	1	0	2	1
WESLEY	0	1	0	1	0	1	1
WINNER	1	1	2	1	0	1	0

Appendix 3.3 Marker sequences for the SNPs associated with the stable MTAs identified in this study.

SNP	Sequence
S1A_13099591	CTCGTTCTCCTCCTCCTTCCTCCTGGAGGCACCTTTTTGACCGCGGTTTGGGATCCAGGCCCCTTG TTCAAAATTCAAGGGCTGAGATCGCTGGCGCGGCAGCCTTGGGTGGACACGCGTCGAGCCTGGA ATGGCCATCTCTGCAGGTCATGTCAGCGCAGCGCACTAGCCACTCCGCCCGCCACCCTGGGCGC GCTAC[T/G]CCACTCCGCCCGCGCATACCCGCCACCCTCCCCGCGCGTGGGCCACCCATGTAAGC GAGCGAGCGTAGGCAGAGTATGATCCATGGACCCACGACACTCGGCCGAAAACCCCGAGGCCA AAAACCCCGGCTGCCTCCGCCTCCGCCTCCGCCAGCGCAGGCGCTTCACTCCGCTCTAGCCTAGC CAGGGGAAGGAAGAGA
S1A_299864277	TTTCGAGGTGTTGGTGAACTTCTCAATTGGGGACGGCACGATCAGACAATTCTGGACCGACCCA TGGCTGCGGCGCCAAAGTCTATGCACAACTTACCCCGACCTTTTTGGCCAGTGCACCCTCTGACG CATAACCGTTGCTGCAGCTTTGCATAATGACAAGTGGATGAGACATTTCAAGGCCAACATGACT GCTGAC[A/G]CGCTTCTCCAGTTCACAAACCTATGGCACGACCTTCAAGCAGTGCACCTCAATCC GGATCAGCAGGACTCAATCTCATGGAGATGGACGGCCAATGGTGTTTACAACGTAGCCTCTGCG AACAGAATTCTCTTTGTTGCAACCATCAAGCAGGACTTTGCCAAAATGGCCTGGAACTCTGAGG CCCGGCCGAAGTGCCAG
S2B_16305395	CTCGAGGAGTCCCTGCAAGATGTGAGAGACTACATCGTTTTCATCCAGTGGCTTGCCGGTGGCT GCTAGCTCGTCGGCGATGCTCACCATCTTGGTGTAGTAGGCCGCAGCCGAGAGGTCTCCTTTGC GGGTATGCTCAATCACGGAGCGAAGCTGAATCACCCTCGCCCTTGACTGTGAGGCAAAGCTCTG CAGCAGG[G/A]CCTTCCAGAGCATGGTGGCGGTGGTGTGGGAGCTGACCTGTAGCGGAACCTCA CGAGAAAGAGATGAGATAAAAAACGTGAGGACTTGCTGATCTTGGGTTACCCACATAGCGTGTT CGGGATTGGGCCTCGATGTGATTTCCTTCTTGCCGGAGATATCCTTCTCCGAGAGGATCACGGCG GGAGGCTCCTGGATCGAA
S3A_647983369	TGTACACGCGCGGCGCAGCGAGCGCATGCCTGCCATCCTTCCTGCCTGCCTGCCTGCCTGCTGGC TGCGATGAGGAGGGTATAGATGGAGTCGTAAATGCATACCGGCCGGGAAGAATGCACCTCGGA GTCCACGTTTTGACACTACTTCTCGCTCCGCCTGCTGATGATGGATCGATCGATGCTCCTGCATG GCATGG[T/C]TGACCTGTCCTGTTGAACGGACGAGCAACGTCTACACTGGCCTGTACTGCAGTAG TACTCCCTCCGGTCCTTTTTACTTCGCACATTAGCTTTGTCTGAAGTCAAAGCTTGCTAAGTTTGA CCAAATTTGTATTAAAAAATATTAACATCTATAACATCTAATAAATATAATATGAAAATATATTC TAAGATGGATCTAA

S4A_619197841	CAAGGGGCTCGCCGGGTCAGTGCGAAGGCCGGCTGCAGGGAGAATTGGTAGGGGCGGCGAGGG GATCCGTTCCCGTGAGCGGAAGCCCGGGGAAGGCACCGGAGTTGGCCGGAATCGGGCGGTCAA TGGAAGGTTGCCGTCGCCGTTGAAGCGAGGGGATCGAAGGTCGGGGAGGCCCGACGTCTAGGA TCCAGGACCA[A/G]CGCGGACGCAGCGGTGGTGCCCCCAATCTCAGGCGCACCACTGCAGGGCG GAGGGCCAGGCGGCGGGAGGAATGGCGGGGAGGGTGTGCGTGAGTCATGCGCCGCCGCGGCTG GAGTTGAGCGTCTGGATAAGATAGAGGAAAGGAGCTCAGGTTGAAATCTAATAAATCCAGGGG TTAATTTGTAAAACCGAATCGTTT
S5A_476847493	GCTAGGAAGGAGCAGAGGAAACGATGCCTGAGGTGCGGCATCCTGTACCTGGACGAGGAGAAC TCCGCTGCTGCCTGCGCCTTCCATGGCCACATCACCGGTAGGTATCGTATCAGCACCACTGACTG CTGCATCACACTACAGTACACTAGTTGCAAAATATCTTCAGTTTGCGGTTGGGCTTGCGACTCTG TGCGTG[C/T]TGTGTACCACGAACGTCCTTGCCCATGCATACGACTGAAGCAGCCTATGCAAAAC TATAGGCCGGAGGAGCTTCCCGAGGCCCGTTTGAGAGTGGACATATGCGTGTTTGTGTTGCAGG TGAGAAGGGGCTGTTTTCGCTGTCGCCGCCGCACCAAGGGATCGACGGCGAGTGGAGCGACAA GAGCGGGGTCATCGTCTA
S5B_432612793	CTACGGAACATCGGTAAGCCTAGTGGATGCATGGATGGGCTATTTCAATTCGGAGAATCTCATT TTCATTCGGGTTGATCGTGAACGAACGTACTCTATCAATTATTTTAGTTAATCAACGCATCCTTT GCGTGTTGCAGGCGTACGTGACGCAGGACGACGTGCTGATGACGACGCTGACGGTGCGGGAGG CGGTGCG[C/G]TACTCGGCGTCGCTGCAGCTGCCGAGCGGCATGTCGTCGGCGGCGAAGCGGGA GCGGGCGGAGGAGACGCTGCGGGAGATGGGGCTGGAGGGTGCGGCGGACACGCGCATCGGCG GGTGGATGCACAAGGGGATCAGCGGCGGCCAGCGGCGCCGGGTCAGCATCTGCATGGAGATCC TCACCCGGCCGGCGCTGCTGTT
S6B_619882604	GTGTTTTTCATAACGATTCCAACCCTACTAATCAGACACACAATTAAATATATTTTTCCGGAAGG GTATAAATCTTCTAAATTTCCTCTGAAATTCCTTTGAATCAAAAGAGCTCTGAAGTTCCACTATG CCTTTTTTCTTCTCCACTGCAGATTGCTGAACCCTTATCAGCACAGGCCATCACAAGTGGGTAGT GCCT[C/G]ACGAGAAGTTTCCTCGCTCAGTTATACGGATGAGATCGAACATTTTGACTTCCCAAA ACCGGGGCCATCCGGCCATGCAAAGATGATACTAGACCCACACCAATCTTCAAATGCTCCATAC TGCCATCCGGTGAACCAGTGAAGGTTGCATCCGAGGGAGGGAGGAAGGGTGGCAGTGATCCGC TTCTCTGCACCACGAA
S7A_132414615	GGCCACAGCATTTGCTACATCAAACAGACCCGTATATGATGGTCATCGGCCACTTTAGCCACAC CCAAAGCTGCTTGAAAATGAGCAGGCTTGTTACATGCTCCGGGCCTATACTGGACACTTTATAT AACTAGATAAAACTTGTCTCCAGAAGGAATTTGATGTCCCTGGAAAGCTGACCCAGGGCTCGCA TTATCAT[C/A]CGCAGTGAAGGCGAGGGCCTGTTTGAGTACCAGGCCTCTCAATGCCGTGATGGT CTGCAGGCTTCACGGAGGGGCGCGGGCAATTGTTCCATCAGCGTGCAAGGCGTCCGCAGCTTCT

	AAAACTACGTCGCCATCGGAGGACATTTATTGTACGATGCAAGGAGGTGAAGGCAGTGTGGCG AGTCATGGGAACGGAAGCT
S7A_676732614	TTTGGGGATTTTAAGTAGGCCTTTGATGTCCGTCCTTTCGGCCCTTCCCGCGGCGGCGGGGGTGT ACAATAAAGTCATCTCCGGTGGCATGTCGTTGATGGTAGTTGATGGTGTCTGCTCGATGAATGAT TCTTCCTCAAGTCTTCTCATTTTGACAGTGATGTCTGTGGCAGTGTCGATGGTTCACAGTCGTAG ATCC[A/G]TACCATCCATTCTAGGGTTTGCAGGGTTACTCGGCATGGTGGCGGTGGTTGCTGCAG TTGGTTCGTCTAGCCTCATCGTTGGTGTTCATCGGCACGTTCAAAGGGGAAGAAGACAGTCAAC CCCATGGACTGTGGAGTGCTTCTTTGCTCTCGAACGTACACCGCCGCGCCTCCGGCGTGCTAAGC TGCTGACCCCATCA
S7A_717859384	TGTGTACTTGTGATCGAGGTATGCATGCACCAGACATCGTCTTGTAACACAGATACGTGCATTGT AGATTTTGTGGAATCCCTGATGACTCAATCCACACATACAAGTAGGACAAGTTGTCACCTTCTCC TTTGGAAATGAGGTCACTCCAAGTGCAGGTATTGCCATCCTGCAGTTGGTGAGGTGGGAATCGT GGGAA[A/G]CTAGGTATTTCTCTATTAGAGGAGAGTGGTTTATGCACTACTTTTGTGTATGTGGA TGTCCAGCCGTCATACAAGTGGATGAAAACGTTTCCGTGCCGGCAGTCTGTGATGTAGATATTCT TGCCCTTGTAGGCGTCAAAGCTGAAGTTTGCACGGTGGATGACTGTGCTAAGCTCTGAGGGCAG CAGCAGATTTGGCAT
S7D_60662020	GTGTAGGTGACCACGTTCCCCATGTACAGGTCCAGCGCGCCCTCCGACGGCTCCACGCACTTCC GGCCGTCGACGATCCACGCCATCCGGCAGTTTGGTAACTGCTGGGATATTTTTCTTCTAAAAAAT AATGATATAAACGCTCTTATATTTTTTTTATCGAGGGAATCGCCGACCGCACGGGCAGTTGGGGT CGGAG[T/G]CGCCGCCTGCAGTCAAAAGAAAGAAAAAAATAAATAAATAATTAATTTCTTTAAG AAAGAATGAACTAGTGACATTGGTATTACCTTTAGGAAGAAATTCAAAATAAAAAACCTGCACA TAACTTTGCACTGAAATGACACTTTATAAATTTGCATGAATAAACATATAATAGTGCAACTTTCA CACCCTAGTGTGCGAG

Appendix 4.1 Prediction accuracy for five traits recorded at five different environments in 2018-19 using different genomic prediction models (ST-CV1, single-trait model; MT-CV1, multi-trait model with CV1 scheme; MT-CV2, multi-trait model with CV2 scheme; MTME, multi-trait multi-environment model with CV1 scheme). The value in bold indicates the best performing model for given trait at respective location.

Trait	**Env**	**ST-CV1**		**MT-CV1**		**MT-CV2**		**MTME**	
		Mean PA	**S.E.**	**Mean PA**	**S.E.**	**Mean PA**	**S.E.**	**Mean PA**	**S.E.**
Yield	Brookings	0.28	0.005	0.29	0.02	**0.56**	0.02	0.26	0.03
	Dakota Lakes	0.32	0.004	0.28	0.03	**0.40**	0.02	0.36	0.02
	Hayes	0.38	0.004	0.35	0.02	**0.41**	0.02	0.25	0.03
	Onida	**0.43**	0.004	0.42	0.01	**0.43**	0.02	0.35	0.03
	Winner	0.13	0.005	0.03	0.02	0.15	0.03	**0.18**	0.03
	Average	0.31	-	0.27	-	0.**39**	-	0.28	-
Protein content	Brookings	0.40	0.004	0.41	0.02	**0.45**	0.02	0.33	0.03
	Dakota Lakes	0.50	0.004	**0.56**	0.02	0.51	0.02	0.45	0.03
	Hayes	0.32	0.004	0.34	0.02	**0.38**	0.02	0.26	0.03
	Onida	0.39	0.004	0.41	0.01	0.39	0.02	**0.46**	0.03
	Winner	0.15	0.004	0.20	0.02	**0.29**	0.02	0.13	0.04
	Average	0.35	-	0.38	-	0.40	-	0.32	-
Test weight	Brookings	0.31	0.004	0.31	0.02	**0.48**	0.02	0.35	0.03
	Dakota Lakes	0.23	0.005	0.23	0.01	**0.39**	0.02	0.32	0.02

	Hayes	0.50	0.004	0.49	0.02	**0.54**	0.02	0.52	0.02
	Onida	0.43	0.005	0.41	0.02	0.40	0.02	**0.47**	0.03
	Winner	0.35	0.005	0.36	0.02	0.32	0.02	**0.43**	0.02
	Average	0.36	-	0.36	-	0.42	-	0.42	-
Plant	Brookings	0.26	0.005	0.24	0.02	0.38	0.02	**0.44**	0.03
height	Dakota Lakes	0.16	0.005	0.16	0.02	0.38	0.02	**0.41**	0.04
	Hayes	0.33	0.004	0.21	0.02	0.31	0.02	**0.42**	0.03
	Onida	0.16	0.004	0.16	0.02	0.27	0.02	**0.45**	0.03
	Winner	0.32	0.004	0.30	0.02	0.34	0.02	**0.54**	0.03
	Average	0.25	-	0.21	-	0.34	-	0.42	-
Heading	Brookings	0.46	0.004	0.46	0.03	0.44	0.03	**0.49**	0.02
date	Dakota Lakes	0.33	0.005	0.32	0.02	**0.47**	0.02	0.40	0.04
	Hayes	0.23	0.005	0.24	0.02	**0.25**	0.02	0.16	0.04
	Onida	0.35	0.005	0.39	0.02	**0.40**	0.01	0.37	0.03
	Winner	0.35	0.005	0.35	0.02	**0.37**	0.02	0.37	0.03
	Average	0.34	-	0.35	-	0.38	-	0.36	-

Appendix 4.2 Prediction accuracy for five traits recorded at five different environments in 2019-20 using different genomic prediction models (ST-CV1, single-trait model; MT-CV1, multi-trait model with CV1 scheme; MT-CV2, multi-trait model with CV2 scheme; MTME, multi-trait multi-environment model with CV1 scheme). The value in bold indicates the best performing model for given trait at respective location.

Trait	**Env**	**ST-CV1**		**MT-CV1**		**MT-CV2**		**MTME**	
		Mean PA	**S.E.**	**Mean PA**	**S.E.**	**Mean PA**	**S.E.**	**Mean PA**	**S.E.**
Yield	Brookings	0.29	0.006	0.27	0.03	**0.52**	0.02	0.39	0.03
	Dakota Lakes	0.33	0.005	0.33	0.02	**0.57**	0.02	0.46	0.03
	Hayes	0.50	0.004	0.52	0.02	**0.71**	0.01	0.44	0.02
	Onida	0.44	0.003	0.41	0.02	**0.50**	0.02	0.46	0.03
	Winner	0.27	0.005	0.23	0.02	**0.67**	0.01	0.43	0.02
	Average	0.36	-	0.35	-	0.5**9**	-	0.43	-
Protein content	Brookings	0.41	0.005	0.41	0.02	0.58	0.02	**0.62**	0.02
	Dakota Lakes	0.34	0.004	0.36	0.02	0.59	0.01	**0.67**	0.02
	Hayes	0.40	0.004	0.40	0.02	0.56	0.02	**0.59**	0.02
	Onida	0.26	0.005	0.22	0.02	0.31	0.02	**0.52**	0.03
	Winner	0.34	0.004	0.35	0.02	**0.66**	0.01	0.58	0.02
	Average	0.35	-	0.35	-	0.54	-	0.60	-
	Brookings	0.56	0.003	0.57	0.01	0.59	0.01	**0.64**	0.02

Test	Dakota Lakes	0.58	0.004	0.58	0.02	0.61	0.01	**0.63**	0.02
weight	Hayes	0.58	0.004	0.57	0.02	0.64	0.01	**0.67**	0.02
	Onida	0.60	0.003	0.60	0.02	0.59	0.02	**0.66**	0.02
	Winner	0.37	0.005	0.33	0.02	0.50	0.02	**0.53**	0.02
	Average	0.54	-	0.53	-	0.59	-	0.63	-
Plant	Brookings	0.35	0.004	0.35	0.02	0.40	0.02	**0.51**	0.02
height	Dakota Lakes	0.31	0.005	0.26	0.02	0.43	0.02	**0.44**	0.02
	Hayes	0.43	0.004	0.41	0.02	0.53	0.02	**0.59**	0.02
	Onida	0.28	0.005	0.24	0.03	0.43	0.02	**0.48**	0.02
	Winner	0.28	0.005	0.27	0.02	0.36	0.02	**0.41**	0.03
	Average	0.33	-	0.31	-	0.43	-	0.49	-
Heading	Brookings	0.37	0.004	0.41	0.02	0.42	0.02	**0.58**	0.02
date	Dakota Lakes	0.29	0.004	0.27	0.02	0.30	0.02	**0.54**	0.02
	Hayes	0.29	0.005	0.36	0.02	0.32	0.03	**0.48**	0.04
	Onida	0.44	0.004	0.46	0.02	0.45	0.02	**0.56**	0.02
	Winner	0.35	0.004	0.31	0.02	0.39	0.02	**0.48**	0.02
	Average	0.35	-	0.36	-	0.38	-	0.53	-

Appendix 5.1 The mean prediction ability (PA) along with standard error (SE) for Mixograph traits using MTGP model. GRPROT, grain protein content; flour ash content; FLRPROT, flour protein content; FLRSDS, flour sedimentation weight value; MIXABS, Mixograph mixing absorption (%), MXTIM, Mixograph mix time (min); MIXTOL, Mixograph mix tolerance score.

Trait	Secondary trait(s)	PA	SE
MIXABS	GRPROT	0.55	0.012
MIXABS	FLRPROT	0.61	0.010
MIXABS	FLRSDS	0.32	0.014
MIXABS	FLRPROT + FLRASH	0.61	0.011
MIXABS	FLRPROT + FLRSDS	0.64	0.008
MIXTM	GRPROT	0.45	0.013
MIXTM	FLRPROT	0.45	0.012
MIXTM	FLRSDS	0.51	0.011
MIXTM	FLRPROT + FLRASH	0.43	0.014
MIXTM	FLRPROT + FLRSDS	0.50	0.011
MIXTOL	GRPROT	0.53	0.011
MIXTOL	FLRPROT	0.53	0.013
MIXTOL	FLRSDS	0.61	0.011
MIXTOL	FLRPROT + FLRASH	0.53	0.011
MIXTOL	FLRPROT + FLRSDS	0.58	0.011

Appendix 5.2 The mean prediction ability (PA) along with standard error (SE) for Glutomatic traits using MTGP model. FLRPROT, flour protein content; FLRSDS, flour sedimentation weight; WGC, wet gluten content; GI, gluten index.

Trait	Secondary trait(s)	PA	SE
WGC	FLRSDS	0.39	0.014
WGC	FLRPRO	0.36	0.016
WGC	FLRPRO + FLRSDS	0.43	0.015
GI	FLRSDS	0.62	0.010
GI	FLRPRO	0.52	0.013
GI	FLRPRO + FLRSDS	0.63	0.012

Appendix 5.3 The mean prediction ability (PA) along with standard error (SE) for Glutomatic traits using MTGP model. GRPROT, grain protein content; FLRASH, flour ash content; FLRPROT, flour protein content; FLRSDS, flour sedimentation weight value; WGC, wet gluten content; GI, gluten index; WGI, wet gluten index; MIXABS, Mixograph mixing absorption (%), MXTIM, Mixograph mix time (min); MIXTOL, Mixograph mix tolerance score; BAKEABS, bake absorption; LVOL, pup loaf volume; SpLVOL, loaf volume by weight.

Trait	Secondary trait(s)	PA	SE
BAKEABS	GRPROT	0.51	0.012
BAKEABS	FLRPROT	0.53	0.012
BAKEABS	FLRSDS	0.45	0.013
BAKEABS	GI	0.48	0.012
BAKEABS	WGC	0.39	0.013
BAKEABS	WGI	0.33	0.011
BAKEABS	MIXABS	0.75	0.007
BAKEABS	MIXTM	0.50	0.012
BAKEABS	MIXTOL	0.54	0.010
BAKEABS	FLRPROT + FLRASH	0.56	0.011
BAKEABS	FLRPROT + FLRSDS	0.61	0.010
BAKEABS	FLRPROT + GI	0.66	0.009
LVOL	GRPROT	0.34	0.016
LVOL	FLRPROT	0.46	0.010

LVOL	FLRSDS	0.44	0.011
LVOL	GI	0.34	0.017
LVOL	WGC	0.33	0.015
LVOL	WGI	0.31	0.013
LVOL	MIXABS	0.44	0.013
LVOL	MIXTM	0.36	0.015
LVOL	MIXTOL	0.37	0.016
LVOL	FLRPROT + FLRASH	0.47	0.010
LVOL	FLRPROT + FLRSDS	0.52	0.010
LVOL	FLRPROT + GI	0.47	0.013
SpLVOL	GRPROT	0.30	0.014
SpLVOL	FLRPROT	0.44	0.011
SpLVOL	FLRSDS	0.36	0.012
SpLVOL	GI	0.35	0.016
SpLVOL	WGC	0.32	0.012
SpLVOL	WGI	0.32	0.014
SpLVOL	MIXABS	0.37	0.013
SpLVOL	MIXTM	0.35	0.013
SpLVOL	MIXTOL	0.32	0.017
SpLVOL	FLRPROT + FLRASH	0.41	0.010
SpLVOL	FLRPROT + FLRSDS	0.45	0.013
SpLVOL	FLRPROT + GI	0.43	0.012

VITA

Harsimardeep Singh Gill was born in Sri Muktsar Sahib, Punjab, India to Mr. Rajbinder Singh and Mrs. Joginder Kaur. He received his B.S. (Agriculture) in 2015 from Punjabi University, India, and his M.S. (Plant Breeding and Genetics) in 2017 from Punjab Agricultural University, India. For his Ph.D., he joined South Dakota State University, Brookings, SD in 2018 and received the Doctorate in Plant Sciences specializing in Plant Breeding and Genetics in 2023 under the supervision of Dr. Sunish Sehgal.