DISSERTATION

GENOME-WIDE ASSOCIATION STUDY AND GENOMIC PREDICTION FOR END-USE QUALITIES IN HARD WINTER WHEAT

Submitted by

Meseret A. Wondifraw

Department of Soil and Crop Sciences

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Spring, 2024

Doctoral Committee:

Advisor: R. Esten Mason
Co-Advisor: Scott D. Haley

Davina Rhodes
Kevin Dorn

ABSTRACT

GENOME-WIDE ASSOCIATION STUDY AND GENOMIC PREDICTION FOR END-USE QUALITIES IN HRAD WINTER WHEAT

Wheat (*Triticum aestivum* L.) is a widely cultivated crop used primarily for human food, animal feed, and industrial products. Numerous wheat-based products have unique nutritional and functional requirements. In the global market, wheat quality is one of the determining factors of wheat's price and baked product characteristics. Thus, after grain yield, improving these qualities is one of the major breeding objectives in wheat.

Chapter One: This chapter outlines wheat's origins and global production. It explores major quality traits like water absorption and dough rheological properties, plus their measurement methods. Factors impacting wheat quality and pertinent genes are discussed. Finally, key challenges and opportunities around breeding for improved wheat quality are addressed.

Chapter Two: This chapter presents a genome-wide association study of water absorption capacity in hard winter wheat. Lines were phenotyped using the solvent retention capacity test and genotyped via genotyping-by-sequencing. Forty-three marker-trait associations were identified across 17 chromosomes, especially on chromosome 1B, indicating polygenic influence. Co-localization between identified marker-trait associations and the genes that have effects on water absorption was done, and some quantitative trait nucleotides (QTNs) were located near gluten glutenin, gliadin, and glycosyltransferase genes, confirming water absorption is a complex trait affected by different flour components.

Chapter Three: This chapter presents genome-wide prediction models to predict water absorption capacity using a training population of 497 hard winter wheat genotypes. Univariate models were compared to multivariate genomic prediction models using two validation approaches - cross-validation with 100 permutations and a 20-80 split and forward validation utilizing three years of data (2019-2021) from the CSU ELITE Trial. Multivariate genomic prediction models integrating highly correlated traits like break flour yield or all traits as covariates showed improved accuracy compared to univariate models in both validation approaches, demonstrating that incorporating related phenotypic traits as covariates in multivariate models can substantially improve the accuracy of predicting water absorption capacity.

Chapter Four: This chapter evaluates genomic prediction models for bread-baking quality traits in 790 wheat genotypes over the 2014-2022 growing seasons. Marker-trait associations identified via genome-wide association study (GWAS) were incorporated as fixed effects. Three models were compared using cross-validation and forward validation: a model without fixed effect, with *Glu-B1al* allele ($Bx7^{OE}$ + 8 subunit) kompetitive allele-specific PCR (KASP) marker data as a fixed effect, and with GWAS-identified markers as fixed effects. Overall, the model with GWAS-identified markers as fixed effects showed the highest prediction accuracy. However, prediction accuracy decreased for bake loaf volume prediction specifically, suggesting that trait-specific tuning is needed to optimize genomic prediction models for different baking quality traits.

These chapters reinforce the genetic complexity of water absorption capacity and baking quality traits in wheat. Polygenic inheritance was revealed for water absorption capacity. Genomic prediction that incorporates phenotypic covariates and GWAS-derived markers is the best approach to selecting water absorption and baking traits.

ACKNOWLEDGEMENTS

The journey of this dissertation was marked by numerous challenges, including the global pandemic, the loss of my beloved sisters, and multiple unprecedented events. Yet, I've made it through it all due to the excellent mentors, collaborators, friends, and family by my side.

I extend my deepest gratitude to Dr. Scott Haley for letting me join your lab, serving as my co-advisor and committee member, and being an incredible mentor and supervisor. While I have so many things to thank you for, for now, I will say thank you for sharing your encouragement and unwavering support whenever I needed it, and I am truly honored to have you by my side.

My immense appreciation goes to my Advisor, Dr. Esten Mason, for taking me on as your student, guiding me with patience and understanding, and providing the support necessary to sustain my pace. I'm grateful to have you in my corner.

My sincere thanks to committee members Dr. Davina Rhodes and Dr. Kevin Dorn for your guidance and insightful direction.

My heartfelt appreciation goes to Dr. Zachary Winn for your outstanding mentorship, support, and the profound impact you have made on my journey.

A special thank you to the wheat breeding team for fostering a supportive and scientifically enriching environment. I want to express my gratitude to John Stromberger for the lab training and guidance and for allowing me to work uninterrupted. Emily Hudson-Arns, I am thankful for your support, encouragement, amazing leadership during harvests, and for making hard work enjoyable, which have all been remarkable. My thanks also extend to Scott Seifert, Amanda Amsberry, Meena Meenakshi, Hong Wang, and Forest Wold-McGimsey for making my time at CSU so welcoming.

A big thank you to Mik Hammers and Selena Lopez for your encouragement and support throughout our shared journey in the lab. Thank you, Tori Anderson, for your encouragement and the smiles you've brought to my face whenever we meet.

Most importantly, my eternal gratitude to my family - my mother, Naniye (W/ro Alem Geze), and my siblings, Agerie, Abebech, and Aklog Wondifraw. You are the foundation and strength behind my every achievement. Finally, a heartfelt thank you to my friends, Dr. Demissew Desta, Dr. Workamaw Warsido, Heran Biza, and Banchi Tarekegn, for your unwavering support throughout this PhD journey.

DEDICATION

In loving memory of my beloved sisters, Beletu Wondifraw and Genet Wondifraw, who left this world too soon. Losing you has left an ache in my heart, a void that may never fully heal. Although you did not live to see this work completed, every page was written with you in my mind and heart and is dedicated to you.

TABLE OF CONTENTS

CHAPTER 1- INTRODUCTION

1.1 Wheat's Origin and Importance

Common wheat (*Triticum aestivum* L., 2n = 6x = 42, genomes AABBDD), a key member of the genus *Triticum* and the family *Poaceae,* has roots tracing back approximately 12,000 years to the Neolithic era, positioning it as a principal founder crop in early agriculture. Its domestication is believed to have started in the Fertile Crescent of Central Asia (Eckardt, 2010).

Wheat has been marked by its complex hexaploid genome due to hybridization events among three related *Triticum* species. Specifically, *Triticum monococcum* hybridized with *Triticum searsii,* forming the tetraploid *Triticum turgidum,* which was subsequently domesticated into Emmer wheat. A later hybridization between *Triticum turgidum* and *Aeglopes tauschii* led to the creation of the hexaploid *Triticum aestivum* (Dvorak et al., 2006; Dvorak et al., 1993; Dvorak et al., 1988).

Modern wheat cultivars primarily emerged from this hexaploid species, with durum wheat originating from tetraploid *Triticum turgidum* (Dubcovsky and Dvorak, 2007; McFadden and Sears, 1946). This series of ancestral hybridizations combined the genetic material of three grass genomes, giving birth to the wheat that presently serves as an essential dietary staple for over 2.5 billion individuals, contributing approximately 20% of the global caloric intake (Erenstein et al., 2022; Juliana et al., 2019).

1.2 Wheat Production

1.2.1 Global production

In 2022, global wheat production reached 781.31 million metric tons, an increase of 1.98 million MT or 0.25% compared to 2021 (World Economic Forum, 2022). China dominated as the top wheat producer with 138 million MT, followed closely by the European Union with 134.3 million MT, India with 103 million MT, and Russia with 91 million MT; the United States was not far behind. Positioned as the fifth-leading producer, the U.S. achieved a harvest of 44.9 million MT.

The world's average wheat production is approximately 3.5 t ha^{-1}. However, regional variances do exist. For instance, while Africa records an average yield of 2.6 t ha^{-1}, South Asia achieves around 3 t ha^{-1}, and peak yields are observed in East Asia and the European Union, ranging between 4.3 to 5.3 t ha^{-1}. Factors like crop management, technological adoption in agriculture, and regional agroecological characteristics play a significant role in these disparities (Grote et al., 2021).

1.2.2 U.S. production

According to data from the United States Department of Agriculture (USDA, 2023), wheat production in the United States for the 2022/23 season was 44.9 million MT. This is a slight increase of 100,000 metric tons compared to the 2021/22 season. However, the 2022/23 wheat production is 9 percent below the 5-year average. The average yield for all wheat varieties grown in the U.S. is 3.13 metric tons per hectare for 2022/23, which is 5 percent higher than the yield obtained in the 2021/22 season. The total harvested wheat area in the U.S. for 2022/23 is 14.4 million hectares, which is a record low (USDA Economic Research Service, 2023).

Wheat production varies widely among US states, with the Central Plains states of North Dakota (primarily spring wheat) and Kansas (primarily winter wheat) leading the country in production for 2022. Specifically, North Dakota produced the most wheat of any state at 7.8 million MT (USDA Economic Research Service, 2023). Kansas came in second, producing 6.3 million MT. Other top-producing states include Washington at 3.7 million MT, Montana at 3.6 million MT, and Idaho at 2.4 million MT. Additional wheat leaders were Minnesota (1.9 million MT), South Dakota (1.9 million MT), and Oklahoma (1.8 million MT).

In summary, while wheat is grown across much of the US, production is concentrated heavily in the Great Plains, led by North Dakota and Kansas. Together, they contributed 12.7 million MT, representing over a quarter of total US wheat production in 2022. Production drops substantially outside of the top two states, with the remaining top producers ranging between 1.8 to 3.7 million MT in 2022. However, Kansas has seen declining yields, with its 2022 output the lowest in almost 60 years. Meanwhile, states like Texas are expected to expand wheat production in 2023 based on the forecasts (USDA Economic Research Service, 2023).

1.2.3 Colorado wheat

Tracing back the history of wheat cultivation in Colorado, we find that as the earliest settlers ventured into the Central Plains, regions now known as Kansas, Colorado, and Nebraska, they prioritized wheat as a primary crop. By 1869, Colorado marked its inaugural documented wheat harvest, with initial yields being modest compared to the standards of today. The state's journey through the wheat production ranks has been quite dynamic, experiencing highs and lows. In 2010, Colorado claimed the fifth position in the US, only to experience fluctuations due to environmental challenges like droughts, freezes, and high-temperature events. However, by 2022, the state exhibited resilience by climbing to the thirteenth position nationally, producing an impressive 1.05 million MT, representing 2.33% of the U.S.'s total wheat production.

1.3 Overview of Different Wheat Market Classes

Wheat grown in the United States is classified into six market classes categorized by hardness, color, and growing season: hard red winter, hard red spring, hard white, soft white, soft red winter, and durum (US Wheat Associates, 2023). Winter wheat varieties are planted in the fall and emerge before going into dormancy when the winter arrives. They exit winter dormancy in spring and reach maturity in the summertime. Winter wheat has vernalization requirements, meaning a specified amount of time in the cold to bolt, flower, and produce seed. In contrast, spring wheat doesn't require vernalization to flower.

On average, winter wheat covers 70% of the total wheat production in the United States (U.S.), while spring wheat constitutes approximately 25% of the total U.S. wheat production (USDA Economic Research Service, 2022). Durum wheat accounts for the smallest share (2-5%) of U.S. wheat production. Hard red winter wheat is grown mainly in the Great Plains (Montana to northern Texas) and accounts for about 40% of U.S. wheat production. Hard red spring wheat covers about 25% of the total production and mainly grows in the Northern Great Plains, including Montana, Minnesota, and North and South Dakota. The third major wheat class in the United States is soft red winter, which accounts for about 15% of the total production. It is grown in states along the East Coast, the Midwest, and the South of the United States. Soft white wheat, which includes both spring and winter types, accounts for about 15% of U.S. production. They grow in different states, including Washington, Oregon, Idaho, Michigan, and New York (USDA Economic Service, 2022). Durum is the smallest wheat class and only accounts for 2-5% of the total wheat production, primarily grown in Montana and North Dakota.

The flour from hard red winter and hard red spring wheat is used to make pan bread and other leavened bread products. The flour from hard red spring wheat is known for its high protein concentration and is suitable for blending with low-protein flour. Soft red winter wheat is used to make cakes, cookies, and crackers. Both winter and spring white wheat are used to make noodle products, cookies, crackers, and crusted white breads, whereas durum wheat is used to produce pasta (US Wheat Associates, 2022).

1.4 Definition of Wheat Quality and Some Quality Traits

Wheat quality refers to the inherent characteristics of wheat grains that define their appropriateness for specific applications. End-use quality relates to how well wheat meets the requirements of its intended purpose, such as baking, milling, or other processing methods (Guzman et al., 2016).

Wheat quality is multilayered, determined by a wide array of traits that are essential for its processing and end-use. These traits are categorized into physical, chemical, nutritional, and functional properties. Physical traits include grain hardness, which influences milling properties and end-product texture; size and shape, affecting milling yield and efficiency; and color, impacting the aesthetic appeal of flour and its products (Hoseney, 1994;). Chemical traits encompass protein content, crucial for gluten development and baking quality, and starch composition, affecting product texture (Pomeranz,1968). Nutritional traits involve the content of fiber, vitamins, and minerals, contributing to the health benefits of wheat (Shwery et al., 2015; Šramková et al., 2009). Functional traits include water absorption, which dictates dough consistency; mixing time and tolerance, indicating dough's resilience and handling properties; and loaf volume, a critical quality indicator for bread, reflecting the dough's ability to rise and maintain structure (Khatkar et al., 2002).

1.4.1 Water absorption capacity

The water absorption capacity (WAC) of wheat flour is critically important for dough properties and baked product quality. Water absorption capacity refers to the amount of water flour can absorb during dough mixing (Kweon et al., 2011). For baked goods using hard winter wheat, like bread and pizza, higher WAC enables optimal dough strength and loaf volume (Buckley, 2013). Excessive WAC causes overly slack dough, while insufficient WAC results in a stiff, resilient dough that restricts fermentation rise. Both are problematic in commercial bakeries (Linlaud and Ferrero, 2009).

Water absorption capacity also has economic implications. Bakers require flour to meet optimum WAC to achieve the desired dough consistency and quality. Millers aim to meet these targets while optimizing flour extraction rates. If millers' wheat varieties yield flour below WAC specifications, adding vital gluten or higher protein wheat can raise WAC but adds costs. This demonstrates the importance of selecting optimal wheat varieties to balance dough quality and extraction efficiency for millers and bakers.

1.4.2 Bake Mixing time and Mixograph tolerance

Bake mixing time (BakeMT) and Mixograph tolerance (MixoT) are two other crucial quality parameters in wheat, influencing its dough properties and bread-making potential. BakeMT refers to the duration required to mix wheat dough to its optimal development, resulting in the best possible bread structure and texture. It's an indicator of the wheat flour's suitability for bread-making; some flours might require longer mixing times while others might form the desired dough consistency quicker (Huebner and Wall, 1976; Khalid et al., 2022). The optimal mixing time ensures that the gluten in the wheat flour is sufficiently developed to retain gases during fermentation, contributing to good loaf volume and texture (Decock and Cappelle, 2005;

Dobraszczyk and Morgenstern, 2003). Mixing time can be estimated using a Mixograph or Farinograph (AACC International, 2010), which plots the resistance of the dough to mixing over time and thus allows for the identification of the optimal mixing duration.

MixoT refers to how the dough responds to extended mixing beyond its optimal mixing time. If the dough starts to break down or weaken quickly after reaching its peak, it's said to have poor MixoT. On the other hand, a dough that maintains its consistency and doesn't deteriorate rapidly exhibits good MixoT. MixoT is vital to wheat flour quality because, in large-scale baking operations, slight overmixing can occur. Wheat flour with a good MixoT provides a safety net, ensuring that the dough remains usable even if mixed a bit longer than intended (Fowler and Kovacs, 2004; Gil-Humanes et al., 2012).

1.4.3 Bake loaf volume

Bake loaf volume refers to the volume occupied by a loaf of bread after baking, and it is a critical indicator of the bread's quality and the functional properties of the wheat flour used. A greater Bake loaf volume generally denotes a bread that has risen well (He and Hoseney, 1992) and possesses a desirable light and airy texture, often attributed to good quality wheat with suitable protein concentration and gluten properties (Kaukab et al., 2022; Whitley, 2009). The bake loaf volume holds significant importance in the wheat product market as consumers often prefer bread with good volume and texture (Nadimi et al., 2023; Osman et al., 2016).

1.5 Phenotyping Methods for Quality Traits

1.5.1 Water absorption capacity

Various methods are available for the determination of flour water absorption, including Mixograph, Farinograph, or doughLAB and the solvent retention capacity test (AACC International, 2010). However, evaluating the water absorption capacity using a Mixograph, Farinograph, or doughLAB may not be convenient since the latter is very slow and time-consuming, and the former is too subjective for reliable estimation of water absorption. Rapid, low-cost, and reliable tests are needed to select genotypes with essential end-use quality traits, especially in the late segregating and early advanced breeding process stages (Li et al., 2015). To fulfill this need, Yamazaki (1953) developed the alkaline water-retention capacity method (AWRC), which has been widely used to measure flour's water-absorption capacity, resulting from the cumulative contributions of all functional flour components. Later, Slade and Levine (1994) developed the solvent retention capacity (SRC) test (AACC International, 2010).

The SRC test is a test employed on hard and soft wheat flours to determine their end-use, baking quality, and hydration performance during mixing (Navrotskyi et al., 2020). The SRC uses different solvents to address the relative contributions to each flour component's water absorption (gluten, damaged starch, and pentosans). It enables prediction and a better understanding of each flour component's functional contribution to the overall flour functionality and the quality of the finished product (Guzman et al., 2015).

In the SRC test, the absorption and retention of different solvents are used to assess water absorption (water as solvent), gluten strength (lactic acid as solvent), damaged starch (sodium carbonate as solvent), and pentosans (sucrose as solvent) (Kweon et al., 2011; Navrotskyi et al., 2020). The water, sodium carbonate, and sucrose solvents, individually and collectively, can provide information on water absorption, while lactic acid provides information on dough rheological properties (Kweon et al., 2011). From the total water absorption of a standard wheat flour sample, gluten can hold 2.8 g of water per gram of dry gluten, starch 0.3 – 0.45 g of water per gram of dry starch, 1.5 g of water per gram of dry, damaged starch, 10 g of water per gram of pentosans (arabinoxylans) (Kweon et al., 2011).

1.5.2 Bake mixing time, Mixograph tolerance, and bake loaf volume

Both BakeMT and MixoT are evaluated through a graphical representation generated by a Mixograph, a farinograph, or a doughLAB during the mixing of the dough. The point at which the graph peaks denotes the mixing time in minutes, representing when the dough has reached its ideal consistency. MixoT, indicating the dough's resilience, is scored by observing the variations in the graph and assigning a value on a scale of 0-6. On the other hand, bake loaf volume is determined after the baking process using canola seed or rape seed displacement by placing the baked loaf in a vessel filled with the seeds; the volume of the displaced seeds is then measured, providing insight into the loaf's rise and texture, and thereby, its quality.

1.6 Genetic Factors Influencing Wheat Quality

1.6.1 Genes associated with quality traits

The genetics underlying wheat grain quality encompasses numerous known genes and quantitative trait loci (QTLs) that are distributed across all 21 chromosomes and impact diverse biochemical pathways and physiological processes determining end-use functionality (Li et al., 2023; Naraghi et al., 2019)

Grain color in wheat is determined primarily by anthocyanin content and other flavonoid pigments that are under genetic control. The red grain color is controlled by the *R-A1, R-B1,* and *R-D1* genes located on chromosomes 3A, 3B, and 3D, respectively, while the recessive allele (*r-1 null*) occurs at all three loci, resulting in white wheat (Flintham and Gale, 1982; Himi et al., 2011; Himi and Noda, 2005). Purple grain color is conferred by the *Pp* genes (*Pp1, Pp3b, and Pp3*), though their chromosomal locations are still unclear (Arbuzova and Maystrenko, 2000). Blue aleurone color is controlled by the *Ba* genes (Qualset et al., 2005), while yellow endosperm color (in durum wheat) is determined by *Psy1* and *Psy2* genes on chromosome groups five and seven

(Pozniak et al., 2007). Differences in grain color arise from the specific anthocyanin pigments produced through the flavonoid biosynthesis pathway, which is regulated by both structural genes (*CHS, F3H, DFR*) and transcription factors (*R2R3-MYB, bHLH*, *WD40*) (Khlestkina et al., 2015; Khlestkina, 2013).

Grain characteristics, including grain shape, size, hardness, and weight, significantly impact grain quality and marketability. Major genes influencing grain morphology include *TaGW8* controlling grain width on chromosome 7A (Chen et al., 2019), *TGW6* regulating grain weight on chromosome 6A (Hanif et al., 2016), and *GL7* determining grain length on chromosome 7A (Wang et al., 2012). In addition, numerous QTLs associated with grain shape and size have been identified across all chromosomes. The Hardness (Ha) locus on chromosome 5D, including the *Puroindoline a (Pina)* and *Puroindoline b (Pinb)* genes, determines wheat grain hardness. Mutations in these genes influence milling properties and end-use quality. This relationship was first described by Morris et al. (2002).

Pre-harvest sprouting tolerance is vital for maintaining functional grain quality in the field under variable environmental conditions prior to harvest. Major sprouting resistance genes include *Phs-A1* located on chromosome arm 4A (Torada et al., 2005), *Phs-B1* and *Phs-D1* on chromosome arms 4B and 4D (Kulwal et al., 2012), and additional dormancy regulators such as *TaMFT* on homoeologous chromosome arms 3A and 3B (Lei et al., 2013). Over 40 QTLs associated with pre-harvest sprouting tolerance have been identified, spanning all wheat chromosomes, with a high density on chromosome group four (Kulwal et al., 2012).

Starch composition, including the proportion of amylose and amylopectin fractions, significantly impacts flour functionality. Amylose content is primarily controlled by "*waxy" (Wx)* proteins encoded by the *Wx-A1, Wx-B1*, and *Wx-D1* genes located on chromosomes 7A, 4A, and

7D (Graybosch, 1998). Numerous starch synthase and starch branching enzyme isoforms on homoeologous chromosome groups 3, 4, and 7 influence amylopectin structure and starch physicochemical properties (Yamamori et al., 2000; Regina et al., 2006). In addition, starch-characteristic QTLs have been identified across all chromosomes, with some clustered in specific chromosomal regions.

Dietary fiber composition, particularly arabinoxylans, impacts digestion properties. Major arabinoxylan synthesis genes include *TaGT43* glycosyltransferase on chromosome arm 1A (Zeng et al., 2010) and the *TaXAT1* and *TaXAT2* acetyltransferase genes on chromosomes 4A and 7D (Anders et al., 2012). Numerous QTLs influencing fiber concentration and arabinoxylan levels have been identified on 10 chromosome arms, including 1A, 1B, 2B, 3A, 3B, 4A, 4B, 6A, 6B, 7A, and 7B (Gao et al., 2013; Tohver et al., 2015).

Gluten genes refer to the genes responsible for coding the proteins found in gluten, primarily gliadins and glutenins. These proteins are the main constituents of gluten, a protein complex found in wheat and several other cereals. Glutenin genes in bread wheat (Triticum aestivum) are divided into two types: high-molecular-weight (HMW) and low-molecular-weight (LMW) glutenins. HMW glutenin subunits are located on the long arms of group one chromosomes (1A, 1B, 1D) at the *Glu-A1*, *Glu-B1*, and *Glu-D1* loci. At the same time, LMW subunits are found on both the short arms of group one chromosome and chromosome six at loci such as *Glu-A3*, *Glu-B3*, and *Glu-D3* on the group one chromosome, and *Glu-A2*, *Glu-B2*, *Glu-D2* on chromosome six (Payne et al., 1983; Shewry et al., 2002). Gliadins are encoded by genes located on the short arms of the chromosome one group, specifically at the *Gli-A1*, *Gli-B1*, and *Gli-D1* loci on chromosomes 1AS, 1BS, and 1DS. Meanwhile, α-gliadins are located on chromosome six at the *Gli-A2*, *Gli-B2*, and *Gli-D2* loci on chromosomes 6AS, 6BS, and 6DS.

Additionally, the *Gli-2* locus, also encoding α-type gliadins, is situated on the short arm of the group six chromosomes (Shewry et al., 2003). The Grain Protein Content *(GPC)* genes, especially the *Gpc-B1* gene, play a critical role in determining grain protein concentration (Uauy et al., 2006).

Notably, the *Gpc-B1* gene is located on the short arm of chromosome 6B in bread wheat and was initially identified from a wild relative, *Triticum turgidum ssp. Dicoccoides* (Uauy et al., 2006). This gene has implications not just for grain protein concentration but also for increasing grain zinc and iron content (Uauy et al., 2006). Collectively, the understanding of these genes has significant implications for wheat breeding, focusing on both yield and quality improvement of wheat products.

Apart from known genes, numerous loci have been identified to be associated with wheat quality tarts. Quantitative trait locus mapping in bi-parental populations has been widely utilized for dissecting the genetic architecture of complex wheat grain and flour quality attributes. Recent studies employing high-resolution genetic maps have uncovered valuable QTLs associated with key end-use quality parameters. Li et al. (2020) constructed a high-density linkage map using a recombinant inbred line population and identified 30 total QTLs for grain hardness, protein concentration, and falling number across three diverse environments. A major QTL on chromosome 5D accounted for up to 48% of phenotypic variation in grain hardness, likely corresponding to the puroindoline genes regulating endosperm texture. However, the most consistent and largest effect QTL mapped to chromosome 3D, explaining up to 26% of the variation in falling number, which is a measure of grain sprouting tolerance.

Guo et al. (2020) performed QTL analysis in a bi-parental population across three years and uncovered 106 QTLs associated with an array of grain protein and starch quality traits. Twelve consistent QTL clusters were detected, including genomic regions harboring loci specifically affecting protein parameters, starch characteristics, or both. Numerous QTL markers showed homology to potential candidate genes, meriting further validation and characterization.

Additional QTL studies have dissected genetic architecture underlying grain protein concentration, dough rheology, pre-harvest sprouting tolerance, and diverse other quality attributes (Semagn et al., 2021; Yang et al., 2021). Meta-QTL analysis has also integrated results from multiple mapping studies to define consensus chromosome regions regulating wheat quality (Li et al., 2023).

Collectively, these efforts highlight the utility of QTL mapping for elucidating the polygenic control of complex wheat quality traits. Validated major-effect QTLs provide valuable targets for fine-mapping and marker-assisted breeding to improve grain composition and end-use functionality. Ongoing QTL discovery and characterization will enable more customized quality improvement and tailored development of elite varieties for diverse end-product applications.

1.6.2 Abiotic stress tolerance genes affecting wheat quality

In wheat, the interplay between abiotic stress resistance and grain quality is underscored by a series of genes whose roles in stress responses impact grain attributes. The *Dreb1/DREB2* genes, predominantly associated with drought response and located on group 1 and 5 chromosomes (Xiao et al., 2008), influence grain size and protein concentration through water availability modulation during grain development. *TaHKT,* which plays a pivotal role in salinity tolerance by maintaining Na+ and K+ balance (Munns et al., 2012), can indirectly affect grain development and protein deposition under saline conditions.

The photoperiod *(Ppd)* genes, which regulate flowering based on photoperiodic cues, are chiefly found on group two chromosomes (Beales et al., 2007). The *Ppd* genes play a key role in determining grain-filling duration, and this impacts grain size, weight, and yield. Distresses in the

TaSnRK2 signaling pathway, located on group 3 chromosomes (Zhang et al., 2016), due to drought or salinity, especially during grain filling, can induce changes in grain morphology and quality.

The vernalization (*Vrn*) genes, determining flowering in response to cold and positioned on groups 5 and 7 chromosomes (Yan et al., 2003), affect the grain-filling phase. Moreover, the *TaLEA* genes, recognized for desiccation tolerance and present on group 6 chromosomes (Gao et al., 2013), potentially affect grain moisture retention. These interconnected genetic roles accentuate the intricacies of breeding for stress resilience without unintentionally compromising grain quality in wheat.

1.7 Genome-Wide Association Studies for Quality Traits

Wheat grain and flour quality attributes like hardness, protein concentration, dough rheology, nutritional composition, and end-product functionality are major quality traits that are targeted in the breeding for quality. Over the two decades, genome-wide association studies (GWAS) have become a powerful approach for elucidating the intricate genetic architecture underlying these traits.

Genome-wide association studies exploit historical recombination and linkage disequilibrium in diverse germplasm panels or breeding populations to detect statistical associations between high-density DNA markers, typically single nucleotide polymorphisms (SNPs), and traits of interest. Genome-wide association study has been extensively applied in wheat for dissecting quality genetics and accelerating marker-assisted breeding. A growing body of GWAS research has uncovered numerous marker-trait associations providing insights into the genes, chromosomes, and genomic regions regulating vital wheat quality parameters.

Grain hardness significantly impacts wheat milling, processing, and end-use quality. Genome-wide association study has uncovered SNP markers associated with hardness on chromosomes 1B, 2B, 3A, 5A, 5B, 5D, 6B, and 7A, including multiple studies validating the role of *puroindolin*e and grain softness protein *(GSP)* genes known to control grain texture located on group 5 chromosomes (Aoun et al., 2021; Wurschum et al., 2017). A 1.3 centi Morgan (cM) haplotype block on 5DS harboring *Pinb* was found to be the major determinant of grain hardness through GWAS in U.S. winter wheat (Aoun et al., 2021). Additionally, GWASs have been utilized in the discovery of novel marker-trait associations (MTA) for grain hardness on chromosomes 2B, 3A, 3B, and 6B without known grain texture genes, highlighting new candidate regions for understanding endosperm hardness (Aoun et al., 2021).

Grain protein concentration, which strongly influences nutritional quality, dough rheology, and processing functionality, has been widely investigated through GWAS. A plethora of protein concentration-associated MTA has been revealed across all 21 wheat chromosomes. Key loci have been uncovered on chromosome arms 1A, 1B, 2B, 3B, 6A, 6B, 7A, and 7B, including several GWAS validating a highly consistent major QTL on 6B corresponding to the *NAM-B1* gene regulating grain nutrient remobilization and accumulation (Sukumaran et al. 2018; Wurschum et al. 2017). Genome-wide association studies using a North American breeding population found *NAM-B1* SNP markers on 6BS that explained up to 12.3% of the variation in grain protein concentration (Fiedler et al., 2017).

In addition to individual loci, a 2.2 Mega base pair (MBp) haplotype region on the short of chromosome 4B harboring the ubiquitin protease gene *TaUBP24* was found to have major effects on wheat grain protein concentration through GWAS integrated with transcriptome analysis (Liu et al., 2018). A genome-wide association study has also uncovered MTAs enabling simultaneous improvement of protein concentration and yield. Conditional GWAS uncovered grain protein QTL on chromosomes 3D and 6D independent of yield (Liu et al., 2018). Overall, GWAS has greatly expanded knowledge of genetic factors modulating wheat grain protein concentration.

Multiple studies have revealed major QTL clusters on group 1 and 7 chromosomes harboring *Glu-1* and *Glu-3* gluten protein loci known to regulate dough properties (Battenfield et al., 2018; Fiedler et al., 2017). The *Glu-D1* locus on 1DS significantly impacted dough resistance, extensibility, and mix time in U.S. winter wheat, with individual SNPs explaining up to 10.4% of the total phenotypic variation (Aoun et al., 2021).

A genome-wide association study in European elite germplasm revealed markers on 1DS corresponding to the *glutenin macropolymer* gene, which strengthens dough through polymerization (Wurschum et al. 2017). Thus, GWAS has validated the essential role of major gluten protein loci in conferring dough rheology while also uncovering novel QTL like chromosome 5B harboring a cell wall invertase gene influencing starch structure and dough behavior (Fiedler et al., 2017).

Grain appearance traits, including color, shape, size, density, and morphology, have been investigated through GWAS to understand regulators of wheat grain quality and market classes. Markers associated with grain pigmentation have been uncovered on chromosomes 1B, 2B, 3B, and 7B, corresponding to known genes controlling anthocyanins and carotenoids like *R-1, Ppr-1,* and *Zds* (Fiedler et al., 2017; Sukumaran et al., 2018; Talini et al., 2020). Loci affecting grain size and shape have been found across all chromosome groups. A major stable QTL for kernel weight mapped to chromosome 1B at 8.3 cM in U.S. winter wheat (Aoun et al., 2021). SNPs linked to grain dimensions have been uncovered on the chromosomes 2A, 3A, 4A, 5A, 6A, 7A, 1D, 2D, 3D, 5DL, 6D, and 7D through GWAS in European elite material (Wurschum et al. 2017). Markers provide tools to tailor wheat grain morphology for specific end-uses.

Milling properties, including flour yield, bran particle size, and coarseness, are critical quality parameters. GWAS uncovered flour yield QTL on 1B, 2A, 4A, 5A, and 5B in U.S. winter wheat (Aoun et al., 2021). The *Bx7*^{OE}+8 gene region on chromosome 1BL in wheat is linked to a reduced number of flour specks and its white color (Fiedler et al., 2017). Thus, GWAS enables the selection of alleles, improving flour purity and extraction. Quantitative trait loci for bran traits have been revealed on chromosomes 2B, 4D, and 6D through GWAS of milling performance (Fiedler et al., 2017).

Nutritional and biofortification traits have also been elucidated through GWAS. Markers linked to grain zinc, iron, magnesium, and calcium concentrations have been uncovered across all chromosome groups (Fiedler et al., 2017; Sukumaran et al., 2018). Major stable QTL for zinc and iron mapped to chromosomes 5A, 6B, and 7A, corresponding to known nutrient homeostasis genes like *NAM-A1* (Sukumaran et al. 2018; Fiedler et al. 2017). A genome-wide association study also identified single nucleotide polymorphisms (SNPs) associated with fiber components like arabinoxylan, carotenoid, flavonoid, and folate levels in grain (Fiedler et al., 2017; Talini et al., 2020). Thus, GWAS elucidates genetic factors modulating wheat nutrient bioavailability.

In summary, the multitude of GWAS studies over the past decade greatly accelerated the discovery of genetic factors controlling diverse wheat grain and flour quality traits, validating known genes and uncovering novel QTL. However, to complement GWAS, meta-analysis integrates results across studies to define consensus chromosome regions regulating complex quality parameters. Meta-GWAS of multi-year, unbalanced breeding trial data has proven effective in boosting the discovery power of modest individual mapping populations (Battenfield et al., 2018).

1.8 Genomic Selection for Quality Traits

Genomic selection leverages genome-wide marker data rather than individual genes to predict breeding values for quantitatively inherited traits through statistical models trained on phenotyped and genotyped populations (Heffner et al., 2009, 2011). Machine learning algorithms like random forest and neural networks implemented for genomic selection modeling have achieved promising prediction accuracies for diverse quality parameters (Jubair and Domaratzki, 2019). Continued optimization of training populations and model algorithms and incorporating multi-omics data into models will enable genomic selection to overcome prohibitive phenotyping costs and complexity barriers facing quality improvement (Li et al., 2019).

1.9 Overview of Genomic Selection Models and Their Assumptions

1.9.1 Ridge regression best linear unbiased prediction

Ridge regression best linear unbiased prediction *(rrBLUP)* was one of the earliest and is still one of the most widely utilized genomic selection models. It employs ridge regression to shrink the effects of all markers towards zero, which avoids extreme effect estimates (Endelman, 2011; Heffner et al., 2009). It then uses the best linear unbiased prediction (BLUP) methodology to derive genomic estimated breeding values (GEBVs).

The *rrBLUP* model assumes that all marker effects are normally distributed with an equal variance. While easy to implement, *rrBLUP* is less accurate than variable selection models for traits with major effect QTLs (Bernardo, 2021; Wang et al., 2015). For traits controlled by many small-effect QTLs (polygenic inheritance), linear models like *rrBLUP* tend to perform well and offer a good balance of speed and accuracy. The equal variance assumption approximates the underlying biology decently.

In summary, aligning genomic selection models with trait genetic architecture, data characteristics, and desired interpretability versus pure predictive power is key for optimizing accuracy and training efficiency. A suite of flexible modeling approaches is advantageous to cater to diverse datasets and breeding targets.

1.9.2 Genomic best linear unbiased prediction

Genomic best linear unbiased prediction *(GBLUP)* is another popular genomic selection method. While *rrBLUP* assumes all marker effects are independently and identically distributed, *GBLUP* models marker effects as correlated based on their genomic relationships (Bernardo, 1994). It uses a genomic relationship matrix (*GRM*) constructed from markers to capture relationships between individuals. The *GRM* contains pairwise relatedness measures for all samples.

In the mixed model, the *GRM* represents the variance-covariance structure for random genetic effects. Marker effects are not estimated directly. Instead, *GBLUP* fits individuals' total genomic breeding values as random effects, which are predicted using *BLUP* methodology, allowing borrowing information across the whole sample population based on genomic relationships (Bernardo, 1994; Gao et al., 2012).

Compared to *rrBLUP, GBLUP* better accounts for linkage disequilibrium between markers and underlying QTLs by modeling marker correlations (Merrick and Carter, 2021). It provides more accurate predictions, especially for traits affected by family relatedness and population structure. In general, *GBLUP* is computationally efficient, especially for large numbers of markers, as its complexity becomes independent of marker number after constructing the genomic relationship matrix (*GRM*), making it more scalable for high-density genotyping datasets compared to *rrBLUP.*

1.9.3 Bayesian ridge regression

Bayesian Ridge Regression (BRR) places a normal *(Gaussian)* prior marker effect to shrink the breeding values towards zero, similar to ordinary ridge regression but within a Bayesian framework (Gianola et al., 2006). Computationally, *BRR* is very efficient, even for models with hundreds of thousands of markers, as the marker effects can be integrated analytically. The key assumption of BRR is that all markers contribute some small non-zero effect to the overall genetic variance. By leveraging Bayesian priors, it estimates marker effects using both the data and this prior shrinkage and prevents overfitting, which is useful in high-dimensional genomic data where there are many more markers than samples (Gianola et al., 2006).

A major advantage of *BRR* is the ability to estimate many small effects across markers without overfitting. However, it lacks any variable selection, so all markers remain in the model. It also cannot capture outliers or major gene effects well, as the normal distribution prior limits large individual marker effects. Overall, *BRR* serves as a fast baseline or exploratory method when marker effects are suspected to be widely distributed with small individual effects. It is a simple yet efficient approach when genetic architecture is largely unknown. However, *BRR* makes strong assumptions on the homogeneity of marker variances (Gianola et al., 2006). It is outperformed by mixture models like *BayesB* or *BayesCπ* that can better handle traits shaped by both major genes and small effect loci.

1.9.4 Bayesian Lasso

The *Bayesian Lasso* is an adaptation of *Lasso* regression within a Bayesian framework, placing a Laplace prior to the regression coefficients to induce sparsity (Park and Casella, 2008). This double exponential prior shrinks coefficients strongly towards zero, setting many marker effects to exactly zero for variable selection. By zeroing out markers, *Bayesian Lasso* focuses on identifying the most influential markers for trait prediction, which works well in high-dimensional contexts where only a subset of markers affects the trait (Park and Casella, 2008). Computationally, *Bayesian Lasso* remains efficient and scalable.

The major downside is that the *Lasso* prior lacks the flexibility to model moderate to large effects, so it risks underestimating their sizes or excluding relevant markers. Additionally, tuning the prior is crucial but challenging to optimize sparsity and avoid overfitting effects (Park and Casella, 2008). Overall, *Bayesian Lasso* excels in very sparse architectures where a few markers dominate. But mixture models like *BayesB* may outperform in complex genetic landscapes shaped by both major genes with large effects and polygenes with small effects.

1.9.5 Bayes A

BayesA is a Bayesian linear regression model that, unlike Bayesian Ridge Regression (BRR), introduces individual variances for each marker effect. It uses a scaled inverse chi-square prior for these variances, allowing different markers to influence the trait to varying degrees (Meuwissen et al., 2001; Hayes et al., 2009). By accommodating heterogeneous effects across the genome, *BayesA* was developed to capture genetic architectures shaped by both major genes and polygenes. The potential of *BayesA* lies in modeling markers with disproportionately large effects, which *BRR* cannot handle well. However, the downside is significantly higher computational demand relative to BRR due to estimating marker-specific parameters. *BayesA* is best suited for situations where prior evidence suggests markers have highly variable effects on the trait (De Los Campos et al., 2013; Hayes et al., 2009). This heterogeneous variance assumption makes it advantageous over *BRR* for traits involving both major and minor effect QTLs, such as in whole genome selection.

1.9.6 BayesB

BayesB is a Bayesian sparse linear regression model that assumes that only a subset of markers affects the trait. It uses a mixture prior with a point mass at zero and a slab following a scaled inverse chi-square distribution (Habier et al., 2011; Meuwissen et al., 2001). The non-zero effects are assumed to arise from a relatively small number of influential markers/QTLs. By zeroing out the majority of marker effects, *BayesB* performs variable selection to focus model capacity on markers with the largest effects (De Roos et al., 2009). However, the marker-specific parameter estimation substantially increases computational demands compared to the simpler *BayesA* model. *BayesB* is best suited for traits shaped by a handful of major effect QTLs alongside numerous smaller background effects. However, it requires reasonable marker density around causative loci to zero out nonimportant markers successfully.

1.10 Comparison of Univariate and Multivariate models

1.10.1 Univariate and multivariate models

The univariate genomic prediction method uses genomic information to analyze a single trait at a time (Meuwissen et al., 2001). This approach assumes that the genomic data available directly influences the target trait. The main strength of the univariate method is its simplicity, which allows for straightforward interpretation of results and easy integration into breeding programs (Heffner et al., 2009; Meuwissen et al., 2001).

Additionally, it can often be more computationally efficient than multivariate methods. However, its downside is that it does not account for potential correlations between multiple traits, which could provide additional insights and improve prediction accuracy (Heslot et al., 2015). It is especially suitable when a particular trait is of prime importance or when computational resources are limited.

1.10.2 Multivariate genomic prediction models

Multivariate genomic prediction considers multiple traits simultaneously. This approach assumes that there might be genetic correlations between different traits, which can be harnessed to enhance predictive accuracy (Jia and Jannink, 2012). The primary advantage of multivariate methods is that they can provide more comprehensive insights into complex genetic architectures, where multiple traits influence one another. They can also be more efficient regarding data usage, as information from one trait can inform predictions about another (Montesinos-López et al., 2016).

However, they can be computationally intensive and require a more complex model setup. Additionally, they might not always result in better predictions for all traits in a given study. This method is best employed when traits are believed to be genetically correlated or when the goal is to gain a holistic understanding of an organism's genomic architecture (Guo et al., 2020).

1.10.3 Cross and forward validations

Cross-validation relies on sub-setting and iterative training and testing on data from the same population rather than temporal splitting (Meuwissen et al., 2001). In k-fold cross-validation, the available data are divided into k distinct subsets or folds. The model is then trained on k-1 folds and validated on the held-out fold in each iteration, which is repeated until all folds have served as the validation set. Cross-validation provides a more robust accuracy estimation compared to a single train-test split, as each observation is tested as part of an independent hold-out fold. The multiple test sets reduce overfitting risks and provide k-fold accuracy estimates that can be aggregated to reduce variance. However, cross-validation evaluations assume that samples are representative of the target population, which may not always be true (Wong et al., 2019).

Forward validation, on the other hand, is a valuable approach for evaluating the transportability and accuracy of a genomic selection model over time. In forward validation, a model is initially developed and trained using data from a set of individuals who were genotyped and phenotyped and then tested on data from subsequent years. The performance of this model is then tested by using it to predict the genetic merit or breeding values of new individuals that are genotyped and phenotyped in a subsequent year or season, which mimics the real-world application of genomic selection, where models developed on available data are deployed to predict the performance of future individuals not previously seen by the model (Hjorth and Hjort, 1982).

Forward validation provides a one-time estimate of predictive accuracy based on this temporal separation between training and validation data (Dawson et al., 2013). However, a limitation is that accuracy estimates can be inflated if the individuals genotyped in the later year are highly related to or descended from the original training set, reducing diversity. Therefore, forward validation is best suited for evaluating models that will be applied to related populations over time.

Cross-validation is well-suited for a more rigorous comparison of alternative models during development and parameter tuning based on their generalization ability before final assessment using forward validation. The complementary strengths of forward validation and cross-validation provide comprehensive safeguards against overfitting and overestimation that bolster confidence in deploying the best genomic selection model.

1.11 Studies Implementing Genomic Predictions for Additional Fixed Effects

Integrating known markers, including those identified through GWAS, as fixed effects in genomic selection models has emerged as an influential strategy to improve prediction accuracy.

Studies over the past several years have demonstrated that incorporating significant SNPs or genetic markers associated with major effect genes as fixed effects in genomic selection models can improve prediction accuracy for various traits in multiple crop species, including wheat, barley (*Hordeum vulgare*), and chili peppers (*Capsicum annuum*) (Sukumaran et al., 2017; Tsai et al., 2020; Kim et al., 2022). Wheat research has shown that integrating markers linked to genes controlling vernalization, photoperiod, disease resistance, yield, and other agronomic traits as fixed effects enhances prediction accuracy for those respective traits, compared to standard genomic selection models without fixed effects (Alemu et al., 2023; Odilbekov et al., 2019; Sarinelli et al., 2019; Sehgal et al., 2020; Sukumaran et al., 2017).

In a study of Nordic winter wheat landraces, Odilbekov et al. (2019) pinpointed ten markers associated with resistance to *Septoria tritici* blotch. Incorporating these markers into genomic selection models led to a significant increase in prediction accuracy. Similarly, Alemu et al. (2021) reported improved prediction accuracy for Septoria and powdery mildew resistance (*Blumeria graminis Tritici.)* by integrating GWAS-identified markers. For barley, Tsai et al. (2020) demonstrated the *Bayesian Lasso* model's superiority over the conventional *ridge regression BLUP*, particularly when evaluating traits linked to powdery mildew resistance. Moreover, Kim et al. (2022) applied this approach to chili peppers (*Capsicum annuum*), focusing on capsaicinoid concentration vital for its pungency. Their work illustrated the broader applicability of the strategy, which dramatically boosted prediction accuracies when integrated into genomic selection models.

Overall, the cumulative research on genomic selection over recent years demonstrates the potential for improved prediction accuracy by leveraging GWAS to identify major effect loci and incorporating these as fixed effects in genomic prediction models.

1.12 Challenges in Quality Improvement

1.12.1 Genetic complexity of quality traits

Wheat grain quality traits like flour water absorption capacity, dough rheology, nutritional composition, and end-product functionality are governed by highly complex genetic architectures involving many genes across the allohexaploid wheat genome. Most economically important quality parameters such as protein and mineral concentration, milling properties, and baking performance follow quantitative inheritance controlled by numerous small-effect loci interacting in complex molecular networks rather than single genes with large phenotypic effects (Battenfield et al., 2018).

This polygenic nature poses significant challenges for elucidating the specific contributions of individual genes to quality phenotypes. Quality traits are influenced by intricate molecular interactions between protein-coding genes and non-coding regulatory elements across the A, B, and D sub-genomes (Barnard, 2002). Epistasis, pleiotropy, and genetic linkages further confound, explaining the effects of specific loci, as a single gene may impact multiple traits or several genes may collaboratively influence one trait (Mackay et al., 2009). Environmental factors and management practices also modulate genetic expression, complicating genotype-phenotype associations (Blake et al., 2009).

1.12.2 Phenotyping limitation

The process of phenotyping quality traits in hard winter wheat presents several notable limitations. Firstly, the time-consuming nature of tests such as break flour and total flour, water absorption, BakeMT, and MixoT poses a significant challenge. From the initial sample preparation to the completion of these tests, the duration is extensive, often conflicting with the short interval between harvesting and the next planting season, typically not exceeding a month. This time constraint hampers the timely gathering and analysis of phenotypic data crucial for breeding decisions. Moreover, the expertise required for these phenotyping processes adds another layer of complexity. Certain tests demand experienced personnel for accurate execution and data interpretation, where subjective judgment can play a role, necessitating skilled evaluators.

Another critical bottleneck is the sample size requirement for these tests. While the actual tests like solvent retention capacity do not require large flour samples, the milling process to produce these samples demands significantly larger grain quantities. This is similarly true for tests such as the Mixograph and baking tests, which require a minimum of 10 grams of grain to produce sufficient flour. In early line development trials, such as head rows, obtaining adequate grain samples for these assessments is often unfeasible, leading to delays in phenotypic evaluation until later stages of line development. This constraint can impede the timely selection and advancement of superior wheat lines, affecting the overall efficiency and effectiveness of the breeding process.

1.12.3 Environmental variability and genotype by environment interactions

With rising temperatures, altered precipitation patterns, and increased frequency of extreme weather events, climate change profoundly affects wheat breeding to enhance quality traits (Fradgley et al., 2023; De Vita and Taranto, 2019). Elevated temperatures can lead to accelerated grain filling, potentially shriveled grain but increased grain protein concentration, thus impacting end-use qualities like baking and pasta-making. Changes in precipitation patterns can induce plant stress, affecting grain filling and influencing traits like grain size, hardness, and protein composition. Moreover, the unpredictability associated with climate change complicates field trials, making evaluations of potential varieties for quality traits inconsistent (Fradgley et al., 2023). Increased pests and diseases due to changing climate conditions further compound the challenges, as breeders may need to prioritize resistance traits over quality traits. In essence, the changing climate adds layers of complexity to wheat breeding, demanding more dynamic and adaptable strategies to ensure the grain's quality amidst evolving environmental challenges.

These challenges are compounded by the highly variable and changing environmental contexts in which new wheat varieties must perform. The diverse agro-ecological contexts spanning the global wheat canvas, from the heat of North Africa to the bitter cold of the Canadian prairies, means that a variety tailored for one set of environmental conditions may completely fail just a few hundred kilometers away under contrasting conditions (Bux et al., 2012; NAWG, 2023). Breeding programs thus face the monumental challenge of ensuring stable performance for key traits across an array of dynamic and unpredictable environmental conditions while also breeding specialized varieties tailored to excel in target environments.

1.13 Opportunities for Quality Improvement

1.13.1 Genomic tools enabling selection

Advancements in genotyping, omics technologies, and bioinformatics are significantly enhancing the precision and efficiency of breeding for wheat quality (Cobb et al., 2013; Taranto et al., 2018; Yang et al., 2021). High-density genotyping platforms, notably the 90K iSelect SNP arrays and genotype-by-sequencing (GBS) methods, have facilitated in-depth genomic profiling, uncovering population structure, diversity, and linkage disequilibrium patterns (Sukumaran et al., 2015; Crossa et al., 2016). Automated phenomics tools provide rapid and accurate assessments of quality parameters (Pratap et al., 2019).

Integrating genomics with transcriptomics, proteomics, and metabolomics has elucidated regulatory networks underpinning complex quality traits. The reduction in sequencing costs is indicative of a shift towards sequence-based genotyping enhancing genomic selection (GS) accuracy than what was possible based on pedigree-based *GRM* only. Genomic data promote the use of marker-assisted recurrent selection and GS, which presents advantages over phenotypic selection by enabling earlier generation screening and improved genotype heritability (Heffner et al., 2011; Battenfield et al., 2016). Marker-assisted stacking of quality genes can lead to additive allelic effects or favorable epistatic interactions (Gupta et al., 2022; Tyagi et al., 2014).

For the traits controlled by known major genes translating the genes and quantitative trait loci influencing wheat quality into functional markers for breeding. Platforms like kompetitive allele-specific PCR (KASP) based SNPs can facilitate high-throughput marker-assisted quality screening (Gupta et al., 2011; Pandurangan et al., 2021). Effective utilization of these genomic tools necessitates focused trait targeting, standardized phenotyping, and integration into breeding pipelines to optimize quality enhancement in wheat breeding programs.

1.13.2 Precision phenomics enabling genomic selection

Precision phenomics refers to the high-throughput, quantitative characterization of complex traits like grain quality using emerging digital and automation technologies combined with statistical analytics and machine learning (Araus and Kefauver, 2018). It involves precisely measuring phenotypic parameters like grain composition, morphology, and end-use properties in structured breeding populations using tools such as near-infrared spectroscopy, digital imaging, metabolite profiling, and automated rheology on thousands of samples (Mir et al., 2019). It generates extensive phenotypic datasets to uncover genotype-trait associations through genome-wide marker mapping and genomic selection models. By accelerating precise phenotyping, predictions of genotype performance and breeding values can be made early in the breeding process, enabling more rapid selection cycles. Additionally, the vast phenotypic data improve genomic prediction accuracy by providing larger training populations to optimize models.

The improved phenotypic and genomic predictions enabled by scaling up precision phenotyping can foster the acceleration of genetic gains in multi-year breeding programs by increasing selection intensity, reducing cycle times, and increasing selection accuracy. However, to achieve these potential improvements in breeding efficiency for complex quality traits, tight integration of phenomics data with genotyping, multi-environment field testing, and tailored data analytical methods is essential.

The reviewed literature underscores the intricate nature of wheat genetics, revealing the challenges inherent in their manipulation. Key genes and QTLs have been identified as both directly and indirectly influencing wheat quality traits. Despite the challenges, there are advanced genetic and statistical tools, such as GWAS, which help to dissect genetic complexity and serve as precise guides to pinpoint genetic markers tied to desired traits. Genome-wide association study and GS refine the selection processes, optimizing genetic gains and ensuring that resulting wheat genotypes align more closely with consumer and market demands.

Chapter 1 References

AACC Approved Methods of Analysis, 11th Ed. Physical dough tests and Methods. Approved November 3, 2010. *Cereals & Grains Association*, St. Paul, MN, U.S.A. https://www.cerealsgrains.org/resources/Methods/Pages/54PhysicalDoughTests.aspx

Alemu, A., Batista, L., Singh, P. K., Ceplitis, A., and Chawade, A. (2023). Haplotype-tagged SNPs improve genomic prediction accuracy for Fusarium head blight resistance and yield-related traits in wheat—*Theoretical and Applied Genetics*, *136*(4), 92. https://doi.org/10.1007/s00122-023-04352-8

Alemu, A., Brazauskas, G., Gaikpa, D. S., Henriksson, T., Islamov, B., Jørgensen, L. N., ... and Chawade, A. (2021). Genome-wide association analysis and genomic prediction for adult-plant resistance to *Septoria tritici blotch* and powdery mildew in winter wheat. *Frontiers in Genetics*, *12*, 661742. https:// doi.org/10.3389/fgene.2021.661742

Anders, N., Wilkinson, M. D., Lovegrove, A., Freeman, J., Tryfona, T., Pellny, T. K., ... and Mitchell, R. A. (2012). Glycosyl transferases in family 61 mediate arabinofuranosyl transfer onto xylan in grasses. *Proceedings of the National Academy of Sciences*, *109*(3), 989-993. https://doi.org/10.1073/pnas.1115858109

Araus, J. L., Kefauver, S. C., Zaman-Allah, M., Olsen, M. S., and Cairns, J. E. (2018). Translating high-throughput phenotyping into genetic gain. *Trends in Plant Science*, *23*(5), 451-466. https://doi.org/10.1016/j.tplants.2018.02.001

Arbuzova, V. S., and Maystrenko, O. I. (2000). The chromosomal location of genes for purple grain color introgressed in common wheat. *Cereal Research Communications*, 28, 235-237. https://doi.org/10.1007/BF03543599

Barnard, A. D., Labuschagne, M. T., and Van Niekerk, H. A. (2002). Heritability estimates of bread wheat quality traits in the Western Cape province of South Africa. *Euphytica*, *127*, 115-122. https://doi.org/10.1023/A:1019997427305

Battenfield, S. D., Sheridan, J. L., Silva, L. D., Miclaus, K. J., Dreisigacker, S., Wolfinger, R. D., ... and Poland, J. A. (2018). Breeding-assisted genomics: Applying meta-GWAS for milling and baking quality in CIMMYT wheat breeding program. *PLoS One*, *13*(11), e0204757. https://doi.org/10.1371/journal.pone.0204757

Battenfield, S. D., Guzmán, C., Gaynor, R. C., Singh, R. P., Peña, R. J., Dreisigacker, S., ... and Poland, J. A. (2016). Genomic selection for processing and end-use quality traits in the CIMMYT spring bread wheat breeding program. *The Plant Genome*, *9*(2), plantgenome2016-01. https://doi.org/10.3835/plantgenome2016.01.0005

Beales, J., Turner, A., Griffiths, S., Snape, J. W., and Laurie, D. A. (2007). A pseudo-response regulator is misexpressed in the photoperiod-insensitive Ppd-D1a mutant of wheat (Triticum aestivum L.). *Theoretical and Applied Genetics*, *115*, 721-733. https://doi.org/10.1007/s00122-007-0603-4

Bernardo, R. (1994). Prediction of maize single-cross performance using RFLPs and information from related hybrids. *Crop Science*, *34*(1), 20-25.

Bernardo, R. (2021). Predictive breeding in maize during the last 90 years. *Crop Science*, *61*(5), 2872-2881. https://doi.org/10.2135/cropsci1994.0011183X003400010003x

Blake, N. K., Lanning, S. P., Martin, J. M., Doyle, M., Sherman, J. D., Naruoka, Y., and Talbert, L. E. (2009). Effect of variation for major growth habit genes on maturity and yield in five spring wheat populations. *Crop Science*, *49*(4), 1211-1220. https://doi.org/10.2135/cropsci2008.08.0505

Buckley, E. (2013). Factors in hard winter wheat affecting water absorption tolerance (Doctoral dissertation, Kansas State University). http://hdl.handle.net/2097/16464

Bux, H., Ashraf, M., Hussain, F., Rattu, A. U. R., and Fayyaz, M. (2012). Characterization of wheat germplasm for stripe rust (*'Puccini striiformis'f.* sp.'*tritici'*) resistance. *Australian Journal of Crop Science*, *6*(1), 116-120.

Chen, S., Waghmode, T. R., Sun, R., Kuramae, E. E., Hu, C., and Liu, B. (2019). Root-associated microbiomes of wheat under the combined effect of plant development and nitrogen fertilization. *Microbiome*, *7*(1), 1-13. https://doi.org/10.1186/s40168-019-0750-2

Cobb, J. N., DeClerck, G., Greenberg, A., Clark, R., and McCouch, S. (2013). Next-generation phenotyping: requirements and strategies for enhancing our understanding of genotype–phenotype relationships and its relevance to crop improvement. *Theoretical and Applied Genetics*, *126*, 867-887. https://doi.org/10.1007/s00122-013-2066-0

Crossa, J., Jarquín, D., Franco, J., Pérez-Rodríguez, P., Burgueño, J., Saint-Pierre, C., ... and Singh, S. (2016). Genomic prediction of gene bank wheat landraces. *G3: Genes, Genomes, Genetics*, *6*(7), 1819-1834. https://doi.org/10.1534/g3.116.029637

Dawson, J. C., Endelman, J. B., Heslot, N., Crossa, J., Poland, J., Dreisigacker, S., ... and Jannink, J. L. (2013). The use of unbalanced historical data for genomic selection in an international wheat breeding program. *Field Crops Research*, *154*, 12-22. https://doi.org/10.1016/j.fcr.2013.07.020

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

De Roos, A. P. W., Hayes, B. J., and Goddard, M. (2009). Reliability of genomic predictions across multiple populations. *Genetics*, *183*(4), 1545-1553. https://doi.org/10.1534/genetics.109.104935

De Vita, P., and Taranto, F. (2019). Durum wheat (Triticum turgidum ssp. durum) breeding to meet the challenge of climate change. *Advances in Plant Breeding Strategies: Cereals: Volume 5*, 471-524. https://doi.org/10.1007/978-3-030-23108-8

Decock, P., and Cappelle, S. (2005). Bread technology and sourdough technology. *Trends in Food Science and Technology*, *16*(1-3), 113-120. https://doi.org/10.1016/j.tifs.2004.04.012

Dobraszczyk, B. J., and Morgenstern, M. P. (2003). Rheology and the breadmaking process. *Journal of Cereal Science*, *38*(3), 229-245. https://doi.org/10.1016/S0733-5210(03)00059-6

Dubcovsky, J., and Dvorak, J. (2007). Genome plasticity is a key factor in the success of polyploid wheat under domestication. *Science*, *316*(5833), 1862-1866. https://doi.org/10.1126/science.1143986

Dvorak, J., Akhunov, E. D., Akhunov, A. R., Deal, K. R., and Luo, M. C. (2006). Molecular characterization of a diagnostic DNA marker for domesticated tetraploid wheat provides evidence for gene flow from wild tetraploid wheat to hexaploid wheat. *Molecular Biology and Evolution*, 23(7), 1386-1396. https://doi.org/10.1093/molbev/msl004

Dvorak, J., Diterlizzi, P., Zhang, H. B., and Resta, P. (1993). The evolution of polyploid wheat - identification of the a-genome donor species. *Genome*, 36(1), 21-31. https://doi.org/10.1139/g93-004

Dvorak, J., McGuire, P. E., and Cassidy, B. (1988). Apparent sources of the genomes of wheat inferred from polymorphism in abundance and restriction fragment length of repeated nucleotide sequences. *Genome*, 30(5), 680-689. https://doi.org/10.1139/g88-115

Eckardt, N. A. (2010). *Evolution of domesticated bread wheat*. *American Society of Plant Biologists, Plant Cell*, 22(4), 993-993. 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). https://doi.org/10.3835/plantgenome2011.08.0024

Erenstein, O., Jaleta, M., Mottaleb, K. A., Sonder, K., Donovan, J., and Braun, H. J. (2022). Global trends in wheat production, consumption, and trade. In Wheat improvement: food security in a changing climate (pp. 47-66). Cham: *Springer International Publishing. https://doi.org/10.1007/978-3-030-90673-3*

*Fiedler, J. D., Salsman, E., Liu, Y., Michalak de J*iménez, M., Hegstad, J. B., Chen, B., ... and Li, X. (2017). Genome-wide association and prediction of grain and semolina quality traits in durum wheat breeding populations. *The PlantPgenome*, *10*(3), plantgenome2017-05. https://doi.org/10.3835/plantgenome2017.05.0038

Finney and Shogren, (1972). A ten-gram Mixograph for determining and predicting functional properties of wheat flours. *Bakers Digest*, *46*, 32-35. https://doi.org/10.1270/jsbbs.56.131

Flintham, J. E., and Gale, M. D. (1982). The Tom Thumb dwarfing gene, Rht3 in wheat, I. Reduced pre-harvest damage to breadmaking quality. *Theoretical and Applied Genetics*, *62*, 121-126. https://doi.org/10.1007/BF00293343

Fowler, D. B., and Kovacs, M. I. P. (2004). Influence of protein concentration on Farinograph absorption, mixing requirements, and Mixograph tolerance. *Canadian Journal of Plant Science*, *84*(3), 765-772. https://doi.org/10.4141/P03-148

Fradgley, N. S., Bacon, J., Bentley, A. R., Costa-Neto, G., Cottrell, A., Crossa, J., ... and Gardner, K. A. (2023). Prediction of near-term climate change impacts on UK wheat quality and the potential for adaptation through plant breeding. *Global Change Biology*, *29*(5), 1296-1313. https://doi.org/10.1111/gcb.16552

Gao, W., Bai, S., Li, Q., Gao, C., Liu, G., Li, G., and Tan, F. (2013). Overexpression of the TaLEA gene from Tamarix androssowii improves salt and drought tolerance in transgenic poplar (Populus simonii× P. nigra). *PLoS One*, *8*(6), e67462. https://doi.org/10.1371/journal.pone.0067462

Gao, H., Christensen, O. F., Madsen, P., Nielsen, U. S., Zhang, Y., Lund, M. S., and Su, G. (2012). Comparison of genomic predictions using three GBLUP methods and two single-step blending methods in the Nordic Holstein population. *Genetics Selection Evolution*, *44*(1), 1-8. https://doi.org/10.1186/1297-9686-44-8

Gianola, D., Fernando, R. L., and Stella, A. (2006). Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics*, *173*(3), 1761-1776. https://doi.org/10.1534/genetics.105.049510

Gil-Humanes, J., Pistón, F., Giménez, M. J., Martín, A., & Barro, F. (2012). The Introgression of RNAi Silencing of γ-Gliadins into Commercial Lines of Bread Wheat Changes the Mixing and Technological Properties of the Dough. *PLOS ONE*, *7*(9), e45937. https://doi.org/10.1371/journal.pone.0045937

Grote, U., Fasse, A., Nguyen, T. T., and Erenstein, O. (2021). Food security and the dynamics of wheat and maize value chains in Africa and Asia. *Frontiers in Sustainable Food Systems*, *4*, 617009. https://doi.org/10.3389/fsufs.2020.617009

Guo, J., Khan, J., Pradhan, S., Shahi, D., Khan, N., Avci, M., ... and 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

Gupta, P. K., Balyan, H. S., Chhuneja, P., Jaiswal, J. P., Tamhankar, S., Mishra, V. K., ... and Vishwakarma, M. K. (2022). Pyramiding of genes for grain protein content, grain quality, and rust resistance in eleven Indian bread wheat cultivars: A multi-institutional effort. *Molecular Breeding*, *42*(4), 21. https://doi.org/10.1007/s11032-022-01277-w

Gupta, P. K., Balyan, H. S., Mir, R. R., Kumar, J., Kumar, A., Kumar, S., ... and Kumari, S. (2011). QTL analysis, association mapping, and marker-assisted selection for some quality traits in bread wheat. An overview of the work done at CCS University, Meerut. *Journal of Wheat Research*, *3*(2), 1.

Guttieri, M. J., Bowen, D., Gannon, D., O'Brien, K., and Souza, E. (2001). Solvent retention capacities of irrigated soft white spring wheat flours. *Crop Science*, *41*(4), 1054-1061. https://doi.org/10.2135/cropsci2001.4141054x

Guzmán, C., Ibba, M. I., Álvarez, J. B., Sissons, M., and Morris, C. (2022). Wheat quality. In M. P. Reynolds and H. J. Braun (Eds.), *Wheat Improvement* (pp. Chapter 11). *Springer, Cham.* https://doi.org/10.1007/978-3-030-90673-3_11

Guzman, C., Peña, R. J., Singh, R., Autrique, E., Dreisigacker, S., Crossa, J., ... and Battenfield, S. (2016). Wheat quality improvement at CIMMYT and the use of genomic selection on it. *Applied and Translational Genomics*, *11*, 3-8. https://doi.org/10.1016/j.atg.2016.10.004

Guzmán, C., Posadas-Romano, G., Hernández-Espinosa, N., Morales-Dorantes, A., and Peña, R. J. (2015). A new standard water absorption criteria based on solvent retention capacity (SRC) to determine dough mixing properties, viscoelasticity, and bread-making quality. *Journal of Cereal Science*, *66*, 59-65. https://doi.org/10.1016/j.jcs.2015.10.009

Habier, D., Fernando, R. L., Kizilkaya, K., and Garrick, D. J. (2011). Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics*, *12*(1), 1-12. http://www.biomedcentral.com/1471-2105/12/186

Hanif, M., Gao, F., Liu, J., Wen, W., Zhang, Y., Rasheed, A., ... and Cao, S. (2016). TaTGW6-A1, an ortholog of rice TGW6, is associated with grain weight and yield in bread wheat. *Molecular Breeding*, 36, 1-8. https://doi.org/10.1007/s11032-015-0425-z

Hayes, B. J., Bowman, P. J., Chamberlain, A. C., Verbyla, K., and Goddard, M. E. (2009). Accuracy of genomic breeding values in multi-breed dairy cattle populations. *Genetics Selection Evolution*, *41*(1), 1-9.8. https://doi:10.1186/1297-9686-41-51

He, H., and Hoseney, R. C. (1992). Effect of the quantity of wheat flour protein on bread loaf volume. *Cereal Chem*, *69*(1), 17-19.

Heffner, E. L., Jannink, J. L., and Sorrells, M. E. (2011). Genomic selection accuracy using multifamily prediction models in a wheat breeding program. *The Plant Genome*, *4*(1). https://doi.org/10.3835/plantgenome2010.12.0029

Heffner, E. L., Sorrells, M. E., and Jannink, J. L. (2009). Genomic selection for crop improvement. *Crop Science*, *49*(1), 1-12. https://doi.org/10.2135/cropsci2008.08.0512

Heslot, N., Jannink, J. L., and Sorrells, M. E. (2015). Perspectives for genomic selection applications and research in plants. *Crop Science*, *55*(1), 1-12. https://doi.org/10.2135/cropsci2014.03.0249

Himi, E., Maekawa, M., Miura, H., and Noda, K. (2011). Development of PCR markers for Tamyb10 related to R-1, red grain color gene in wheat. *Theoretical and Applied Genetics*, *122*, 1561-1576. https://doi.org/10.1007/s00122-011-1555-2.

Himi, E., and Noda, K. (2005). The red grain color gene (R) of wheat is a Myb-type transcription factor. *Euphytica*, *143*, 239-242. https://doi.org/10.1007/s10681-005-7854-4.

Hjorth, U., and Hjort, U. (1982). Model selection and forward validation. *Scandinavian Journal of Statistics*, 95-105. https://www.jstor.org/stable/4615861

Hoseney, R. C. (1994). *Principles of cereal science and technology* (No. Ed. 2). *American Association of Cereal Chemists* (AACC).

Huebner, F. R., and Wall, J. S. (1976). Fractionation and quantitative differences of glutenin from wheat varieties varying in baking quality. *Cereal Chem* 53:258 - 269

Jia, Y., and 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

Jubair, S., and Domaratzki, M. (2019, November). Ensemble supervised learning for genomic selection. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 1993-2000). IEEE. https://doi.org/10.1109/BIBM47256.2019.8982998

Juliana, P., Poland, J., Huerta-Espino, J., Shrestha, S., Crossa, J., Crespo-Herrera, L., ... and 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

Kaukab, S., Mir, N. A., Ritika, and Yadav, D. N. (2022). Interventions in wheat processing quality of end products. In *New Horizons in Wheat and Barley Research: Global Trends, Breeding and Quality Enhancement* (pp. 789-808). Singapore: Springer Singapore. https://doi.org/10.1007/978-981-16-4449-8_30

Khatkar, B. S., Fido, R. J., Tatham, A. S., and Schofield, J. D. (2002). Functional properties of wheat gliadins. I. Effects on mixing characteristics and bread making quality. *Journal of Cereal Science*, *35*(3), 299-306.

Khalid, K. H., Ohm, J. B., and Simsek, S. (2022). Influence of bread-making method, genotype, and growing location on whole-wheat bread quality in hard red spring wheat. *Cereal Chemistry*, *99*(3), 467-481. https://doi.org/10.1002/cche.10509

Khlestkina, E. K. (2013). Genes determine the coloration of different organs in wheat. *Russian Journal of Genetics: Applied Research*, *3*, 54-65. https://doi.org/10.1134/S2079059713010085

Khlestkina, E. K., Shoeva, O. Y., and Gordeeva, E. I. (2015). Flavonoid biosynthesis genes in wheat. *Russian Journal of Genetics: Applied Research*, *5*, 268-278. https://doi.org/10.1134/S2079059715030077

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

Kulwal, P., Ishikawa, G., Benscher, D., Feng, Z., Yu, L. X., Jadhav, A., ... and Sorrells, M. E. (2012). Association mapping for pre-harvest sprouting resistance in white winter wheat. *Theoretical and Applied Genetics*, *125*,793-805. https://doi.org/10.1007/s00122-012-1872-0

Kumar, A., Mantovani, E. E., Simsek, S., Jain, S., Elias, E. M., and Mergoum, M. (2019). Genome-wide genetic dissection of wheat quality and yield related traits and their relationship with grain shape and size traits in an elite× non-adapted bread wheat cross. *PLoS One*, *14*(9), e0221826. https://doi.org/10.1371/journal.pone.0221826

Kweon, M., Slade, L., and Levine, H. (2011). Solvent retention capacity (SRC) testing of wheat flour: Principles and value in predicting flour functionality in different wheat-based food processes and in wheat breeding—A review. *Cereal Chemistry*, *88*(6), 537-552. https://doi.org/10.1094/CCHEM-07-11-0092

Lei, L., Zhu, X., Wang, S., Zhu, M., Carver, B. F., and Yan, L. (2013). TaMFT-A1 is associated with seed germination sensitive to temperature in winter wheat. *PloS one*, 12; *8*(9), e73330. https://doi.org/10.1371/journal.pone.0073330

Li, N., Miao, Y., Ma, J., Zhang, P., Chen, T., Liu, Y., ... and Yang, D. (2023). Consensus genomic regions for grain quality traits in wheat revealed by Meta-QTL analysis and in silico transcriptome integration. *The Plant Genome*, e20336. https://doi.org/10.1002/tpg2.20336

Li, Q., Pan, Z., Gao, Y., Li, T., Liang, J., Zhang, Z., ... and Yu, M. (2020). Quantitative trait locus (QTLs) mapping for quality traits of wheat based on high density genetic map combined with bulked segregant analysis RNA-seq (BSR-Seq) indicates that the basic 7S globulin gene is related to falling number. *Frontiers in Plant Science*, *11*, 600788. https://doi.org/10.3389/fpls.2020.600788

Li, D., Xu, Z., Gu, R., Wang, P., Lyle, D., Xu, J., ... and Wang, G. (2019). Enhancing genomic selection by fitting large-effect SNPs as fixed effects and a genotype-by-environment effect using a maize BC1F3: 4 population. *PLoS One*, *14*(10), e0223898. https://doi.org/10.1371/journal.pone.0223898

Li, A. L., Geng, S. F., Zhang, L. Q., Liu, D. C., and Mao, L. (2015). Making the bread: insights from newly synthesized allohexaploid wheat. *Molecular Plant*, *8*(6), 847-859. https://doi.org/10.1016/j.molp.2015.02.016

Linlaud, N. E., Puppo, M. C., and Ferrero, C. (2009). Effect of hydrocolloids on water absorption of wheat flour and Farinograph and textural characteristics of dough. *Cereal chemistry*, *86*(4), 376-382. https://doi.org/10.1094/CCHEM-86-4-0376

Liu, J., Feng, B., Xu, Z., Fan, X., Jiang, F., Jin, X., ... and Wang, T. (2018). A genome-wide association study of wheat yield and quality-related traits in southwest China. *Molecular Breeding*, *38*, 1-11. https://doi.org/10.1007/s11032-017-0759-9

Mackay, T. F., Stone, E. A., and Ayroles, J. F. (2009). The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics*, *10*(8), 565-577. https://doi.org/10.1038/nrg2612

McFadden, E. S., and Sears, E. R. (1946). The origin of Triticum spelta and its free-threshing hexaploid relatives. *Journal of Heredity*, 37(3), 81-89. https://doi.org/10.1093/oxfordjournals.jhered.a105590

Merrick, L. F., and Carter, A. H. (2021). Comparison of genomic selection models for exploring the predictive ability of complex traits in breeding programs. *The Plant Genome*, *14*(3), e20158. https://doi.org/10.1002/tpg2.20158

Meuwissen, T. H., Hayes, B. J., and Godard, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*(4), 1819-1829. https://doi.org/10.1093/genetics/157.4.1819

Mir, R. R., Reynolds, M., Pinto, F., Khan, M. A., and Bhat, M. A. (2019). High-throughput phenotyping for crop improvement in the genomics era. *Plant Science*, *282*, 60-72. https://doi.org/10.1016/j.plantsci.2019.01.007

Montesinos-López, O. A., Montesinos-López, A., Crossa, J., Toledo, F. H., Pérez-Hernández, O., Eskridge, K. M., and Rutkoski, J. (2016). A genomic Bayesian multi-trait and multi-environment model. *G3: Genes, Genomes, Genetics*, *6*(9), 2725-2744. https://doi.org/10.1534/g3.116.032359

Morris, C. F. (2002). Puroindolines: the molecular genetic basis of wheat grain hardness. *Plant Molecular Biology*, *48*, 633-647. https://doi.org/10.1023/A:1014837431178

Munns, R., James, R. A., Xu, B., Athman, A., Conn, S. J., Jordans, C., ... and Gilliham, M. (2012). Wheat grain yield on saline soils is improved by an ancestral Na+ transporter gene. *Nature Biotechnology*, *30*(4), 360-364. https://doi.org/10.1038/nbt.2120

Nadimi, M., Hawley, E., Liu, J., Hildebrand, K., Sopiwnyk, E., and Paliwal, J. (2023). Enhancing traceability of wheat quality through the supply chain. *Comprehensive Reviews in Food Science and Food Safety*. https://doi.org/10.1111/1541-4337.13150

Navrotskyi, S., Belamkar, V., Baenziger, P. S., and Rose, D. J. (2020). Insights into the genetic architecture of bran friability and water retention capacity are two important traits for whole grain end-use quality in winter wheat. *Genes*, *11*(8), 838. https://doi.org/10.3390/genes11080838

Naraghi, S. M., Simsek, S., Kumar, A., Al Rabbi, S. H., Alamri, M. S., Elias, E. M., and Mergoum, M. (2019). Deciphering the genetics of major end-use quality traits in wheat. *G3: Genes, Genomes, Genetics*, *9*(5), 1405-1427. https://doi.org/10.1534/g3.119.400050

NAWG (National Association of Wheat Growers) (2023). Retrieved from https://wheatworld.org/wheat-production-regions.

Odilbekov, F., Armoniené, R., Koc, A., Svensson, J., and Chawade, A. (2019). GWAS-assisted genomic prediction to predict resistance to Septoria tritici blotch in Nordic winter wheat at the seedling stage. *Frontiers in Genetics*, *10*, 1224. https://doi.org/10.3389/fgene.2019.01224

Osman, A. M., Almekinders, C. J. M., Struik, P. C., and van Bueren, E. L. (2016). Adapting spring wheat breeding to the needs of the organic sector. *NJAS-Wageningen Journal of Life Sciences*, *76*, 55-63. https://doi.org/10.1016/j.njas.2015.11.004

Pandurangan, S., Workman, C., Nilsen, K., and Kumar, S. (2021). Introduction to marker-assisted selection in wheat breeding. In *Accelerated Breeding of Cereal Crops* (pp. 77-117). New York, NY: Springer US. https://doi.org/10.1007/978-1-0716-1526-3_3

Park, T., and Casella, G. (2008). The *Bayesian Lasso*. *Journal of the American Statistical Association*, 103(482), 681-686. https://doi.org/10.1198/016214508000000337

Payne, P. I., and Lawrence, G. J. (1983). Catalog of alleles for the complex gene loci, Glu-A1, Glu-B1, and Glu-D1, which code for high-molecular-weight subunits of glutenin in hexaploid wheat. *Cereal Research Communications*, 29-35. https://www.jstor.org/stable/23781365

Pozniak, C. J., Knox, R. E., Clarke, F. R., and Clarke, J. M. (2007). Identification of QTL and association of a phytoene synthase gene with endosperm color in durum wheat. *Theoretical and Applied Genetics*, *114*, 525-537. https://doi.org/10.1007/s00122-006-0453-5

Pratap, A., Gupta, S., Nair, R. M., Gupta, S. K., Schafleitner, R., Basu, P. S., ... and Baek, K. H. (2019). Using plant phenomics to exploit the gains of genomics. *Agronomy*, *9*(3), 126. https://doi.org/10.3390/agronomy9030126

Pomeranz, Y. (1968). Relation between chemical composition and bread-making potentialities of wheat flour. *Advances in Food Research*, *16*, 335-455.

Qualset, C. O., Soliman, K. M., Jan, C. C., Dvorak, J., McGuire, P. E., and Vogt, H. E. (2005). Registration of UC66049 Triticum aestivum blue aleurone genetic stock. *Crop Science*, *45*(1), 432-433. https://www.crops.org/publications/cs

Regina, A., Bird, A., Topping, D., Bowden, S., Freeman, J., Barsby, T., ... and Morell, M. (2006). High-amylose wheat generated by RNA interference improves indices of large-bowel health in rats. *Proceedings of the National Academy of Sciences*, *103*(10), 3546-3551. https://doi.org/10.1073/pnas.0510737103

Sarinelli, J. M., Murphy, J. P., Tyagi, P., Holland, J. B., Johnson, J. W., Mergoum, M., ... and Brown-Guedira, G. (2019). Training population selection and use of fixed effects to optimize genomic predictions in a historical USA winter wheat panel. *Theoretical and Applied Genetics*, *132*, 1247-1261. https://doi.org/10.1007/s00122-019-03276-6

Sehgal, D., Rosyara, U., Mondal, S., Singh, R., Poland, J., and Dreisigacker, S. (2020). Incorporating genome-wide association mapping results into genomic prediction models for grain yield and yield stability in CIMMYT spring bread wheat. *Frontiers in Plant Science*, *11*, 197. https://doi.org/10.3389/fpls.2020.00197

Semagn, K., Iqbal, M., Chen, H., Perez-Lara, E., Bemister, D. H., Xiang, R., ... and Spaner, D. (2021). Physical mapping of QTL associated with agronomic and end-use quality traits in spring wheat under conventional and organic management systems. *Theoretical and Applied Genetics*, *134*, 3699-3719. https://doi.org/10.1007/s00122-021-03923-x

Shewry, P. R., and Hey, S. J. (2015). The contribution of wheat to human diet and health. *Food and Energy Security*, *4*(3), 178-202. https://doi.org/10.1002/fes3.64

Shewry, P. R., Halford, N. G., and Lafiandra, D. (2003). Genetics of wheat gluten proteins. *Advances in Genetics*, *49*, 111-184. https://doi.org/10.1016/S0065-2660(03)01003-4

Shewry, P. R., Halford, N. G., Belton, P. S., and Tatham, A. S. (2002). The structure and properties of gluten: an elastic protein from wheat grain. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *357*(1418), 133-142. https://doi.org/10.1098/rstb.2001.1024

Slade, L., and Levine, H. (1994). Water and the glass transition—dependence of the glass transition on composition and chemical structure: special implications for flour functionality in cookie baking. In *Water in Foods* (pp.143-188). Pergamon. https://doi.org/10.1016/0260-8774(95)90766-5

Šramková, Z., Gregová, E., and Šturdík, E. (2009). Chemical composition and nutritional quality of wheat grain. *Acta chimica slovaca*, *2*(1), 115-138.

Statista, (2023). Retrieved from https://www.statista.com/statistics/263977/world-grain-production-by-type.

Sukumaran, S., Reynolds, M. P., and Sansaloni, C. (2018). Genome-wide association analyses identify QTL hotspots for yield and component traits in durum wheat grown under yield potential, drought, and heat stress environments. *Frontiers in Plant Science*, *9*, 81. https://doi.org/10.3389/fpls.2018.00081

Sukumaran, S., Crossa, J., Jarquin, D., Lopes, M., and Reynolds, M. P. (2017). Genomic prediction with pedigree and genotype× environment interaction in spring wheat grown in South and West Asia, North Africa, and Mexico. *G3: Genes, Genomes, Genetics*, *7*(2), 481-495. https://doi.org/10.1534/g3.116.036251

Sukumaran, S., Dreisigacker, S., Lopes, M., Chavez, P., and Reynolds, M. P. (2015). 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*, 353-363. https://doi.org/10.1007/s00122-014-2435-3

Talini, R. F., Brandolini, A., Miculan, M., Brunazzi, A., Vaccino, P., Pè, M. E., and Dell'Acqua, M. (2020). Genome-wide association study of agronomic and quality traits in a world collection of the wild wheat relative Triticum urartu. *The Plant Journal*, *102*(3), 555-568. https://doi.org/10.1111/tpj.14650

Taranto, F., Nicolia, A., Pavan, S., De Vita, P., and D'Agostino, N. (2018). Biotechnological and digital revolution for climate-smart plant breeding. *Agronomy*, *8*(12), 277. https://doi.org/10.3390/agronomy8120277

Tohver, M., Kann, A., Täht, R., Mihhalevski, A., and Hakman, J. (2005). Quality of triticale cultivars suitable for growing and bread-making in northern conditions. *Food Chemistry*, *89*(1), 125-132. https://doi.org/10.1016/j.foodchem.2004.01.079

Tsai, H. Y., Janss, L. L., Andersen, J. R., Orabi, J., Jensen, J. D., Jahoor, A., and Jensen, J. (2020). Genomic prediction and GWAS of yield, quality and disease-related traits in spring barley and winter wheat. *Scientific Reports*, *10*(1), 3347. https://doi.org/10.1038/s41598-020-60203-2

Tyagi, S., Mir, R. R., Kaur, H., Chhuneja, P., Ramesh, B., Balyan, H. S., and Gupta, P. K. (2014). Marker-assisted pyramiding of eight QTLs/genes for seven different traits in common wheat (Triticum aestivum L.). *Molecular Breeding*, *34*, 167-175. https://doi.org/10.1007/s11032-014-0027-1

Uauy, C., Distelfeld, A., Fahima, T., Blechl, A., and Dubcovsky, J. (2006). A NAC gene regulating senescence improves grain protein, zinc, and iron content in wheat. *Science*, *314*(5803), 1298-1301. https://doi.org/10.1126/science.113364

Economic Research Service, U.S. Department of Agriculture. (2023). Wheat sector at a glance. Retrieved from https://www.ers.usda.gov/topics/crops/wheat/wheat-sector-at-a-glance

United states department of agriculture (USDA), (2023). Retrieved from https://www.ers.usda.gov/topics/crops/wheat

United states department of agriculture (USDA), Economic Research Service (2022). Retrieved from https://www.ers.usda.gov/topics/crops/wheat/wheat-sector-at-a-glance/#classes

US Wheat Associates, (2023). Retrieved from https://www.uswheat.org.

Wang, X., Yang, Z., and Xu, C. (2015). A comparison of genomic selection methods for breeding value prediction. *Science Bulletin*, *60*(10), 925-935. https://doi.org/10.1007/s11434-015-0791-2

Wang, S., Wu, K., Yuan, Q., Liu, X., Liu, Z., Lin, X., ... and Fu, X. (2012). Control of grain size, shape, and quality by OsSPL16 in rice. *Nature Genetics*, *44*(8), 950-954. https://doi.org/10.1038/ng.2327

Whitley, A. (2009). Bread matters. The state of modern bread and a definitive guide to baking your own. Andrews McMeel Publishing.

Wong, T. T., and Yeh, P. Y. (2019). Reliable accuracy estimates from k-fold cross-validation. *IEEE Transactions on Knowledge and Data Engineering*, *32*(8), 1586-1594. https://doi.org/10.1109/TKDE.2019.2912815

World Economic Forum (2022). Retrieved from https://www.weforum.org/agenda/2022/08/top-10-countries-produce-most-wheat.

Würschum, T., Langer, S. M., Longin, C. F. H., Tucker, M. R., and Leiser, W. L. (2017). A modern Green Revolution gene for reduced height in wheat. *The Plant Journal*, *92*(5), 892-903. https://doi.org/10.1111/tpj.13726

Xiao, H., Tattersall, E. A., Siddiqua, M. K., Cramer, G. R., and Nassuth, A. (2008). CBF4 is a unique member of the CBF transcription factor family of Vitis vinifera and Vitis riparia. *Plant, cell and environment*, *31*(1), 1-10. https://doi.org/10.1111/j.1365-3040.2007.01741.x

Yamamori, M., Fujita, S., Hayakawa, K., Matsuki, J., and Yasui, T. (2000). Genetic elimination of a starch granule protein, SGP-1, of wheat generates an altered starch with apparent high amylose. *Theoretical and Applied Genetics*, *101*, 21-29. https://doi.org/10.1007/s001220051444

Yamazaki, W. T. (1953). An alkaline water retention capacity test for the evaluation of cookie baking potentialities of soft winter wheat flours. *Cereal Chem*, *30*(3), 242-246.

Yan, L., Loukoianov, A., Tranquilli, G., Helguera, M., Fahima, T., and Dubcovsky, J. (2003). Positional cloning of the wheat vernalization gene VRN1. *Proceedings of the National Academy of Sciences*, *100*(10), 6263-6268. https://doi.org/10.1073/pnas.093739910

Yang, Y., Saand, M. A., Huang, L., Abdelaal, W. B., Zhang, J., Wu, Y., ... and Wang, F. (2021). Applications of multi-omics technologies for crop improvement. *Frontiers in Plant Science*, *12*, 563953. https://doi.org/10.3389/fpls.2021.563953

Zeng, W., Jiang, N., Nadella, R., Killen, T. L., Nadella, V., and Faik, A. (2010). A glucurono (arabino) xylan synthase complex from wheat contains members of the GT43, GT47, and GT75 families and functions cooperatively. *Plant Physiology*, *154*(1), 78-97. https://doi.org/10.1104/pp.110.159749

Zhang, H., Li, W., Mao, X., Jing, R., and Jia, H. (2016). Differential activation of the wheat SnRK2 family by abiotic stresses. *Frontiers in Plant Science*, *7*, 420. https://doi.org/10.3389/fpls.2016.00420

Chapter 2- Genome-Wide Association Study of Water Absorption Capacity in Hard Winter Wheat

2.1 Summary

Water absorption capacity (WAC) influences various aspects of bread making, such as loaf volume, bread yield, and shelf life. Despite its importance in the baking process and end-product quality, its genetic determinants are less explored. To address this limitation, a genome-wide association study was conducted on 337 hard wheat genotypes evaluated over five years in multi-environmental trials. Phenotyping was done using the solvent retention capacity (SRC) test with water (SRC-water), Sucrose (SRC-sucrose), lactic acid (SRC-lactic acid), and sodium carbonate (SRC-carbonate) as solvents. Individuals were genotyped using genotyping-by-sequencing to detect single nucleotide polymorphisms across the wheat genome. To detect the genomic regions that underline the SRCs and gluten performance index (GPI), a genome-wide association study was performed using six multi-locus models using the *mrMLM* package in R. Adjusted means for SRC-water ranged from 54.1% to 66.5%, while SRC-carbonate exhibited a narrow range from 84.9 % to 93.9 %. Moderate to high genomic heritability values were observed for SRCs and GPI, ranging from $h^2 = 0.61$ to 0.88. The GWAS identified a total of 42 quantitative trait nucleotides (QTNs), of which five explained over 10% of the phenotypic variation ($R^2 \geq 10\%$). Most of the QTNs were detected on chromosomes 1A, 1B, 3B, and 5B. Few QTNs, such as S1A_5190318, S1B_3282665, S4D_472908721, and S7A_37433960, were located near gliadin, glutenin starch synthesis, galactosyltransferase genes. Overall, these results show WAC to be under polygenic genetic control, with genes involved in the synthesis of key flour components influencing overall water absorption.

2.2 Introduction

The bread-baking quality of wheat (*Triticum aestivum* L.) is a primary determinant of the market price of the grain (Diriba et al., 2020). Water absorption capacity (WAC), which is a critical factor in the baking performance of wheat flour, refers to the amount of water a wheat flour can absorb to achieve optimum dough consistency for a desired baked product (Kweon et al., 2011). For hard winter wheat, a preferred wheat class to make bread and pizza dough, the WAC of flour influences bread production, both in terms of the bread's characteristics (Bushuk and Békés, 2002; Zghal et al., 2001) and the economy of its production (Bushuk and Békés, 2002; Pyler, 1979; Zghal et al., 2001). When flour has a high WAC, it can absorb more water, which provides two significant benefits for bakers. First, high WAC flour creates a larger volume of dough or bread per unit of flour, increasing bread yield. Bread yield can be maximized by increasing the amount of water in the formula without compromising bread quality and preventing water loss during baking (Puhr and D'Appolonia, 1992). Secondly, the use of high WAC flour may reduce the need for additional ingredients, which increases the costs of bread production.

The WAC of wheat flour is a complex trait that depends on various components present in the flour, including protein concentration, the degree of damaged starch, and pentosan/arabinoxylan concentration (Jelaca and Hlynka, 1971; Preston et al., 2001; Primo-Martin et al., 2003; Rakszegi et al., 2014; Tipples et al., 1978). Other factors, such as grain size (Morgan et al., 2000) and grain hardness, can also affect WAC since they are related to the degree of damaged starch that occurs during the milling process (Pasha et al., 2010; Tipples et al., 1978). Harder grains are more resistant to milling, which results in more damaged starch. High levels of damaged starch and pentosan concentration lead to higher flour WAC (Kweon et al., 2011; Zghal et al., 2001).

The WAC of flour can be measured using methods such as the Farinograph and Mixograph tests (AACC International, 2010). However, these procedures are time-intensive and require substantial sample sizes. Furthermore, the Mixograph offers only a subjective measure of WAC (Ram et al., 2005). An efficient alternative is the solvent retention capacity (SRC) test. The SRC test uses four solvents, namely, water, lactic acid, sodium carbonate, and sucrose, to assess the functional contributions of flour components to overall product quality. The SRC test is faster and more efficient for determining a flour's WAC and understanding the roles of its components.

Genome-wide association study (GWAS) is a genetic association mapping tool for identifying genomic regions related to complex traits in crops (Brachi et al., 2011; Zhao et al., 2011). Genome-wide association studies identify quantitative trait nucleotides, which can be used to elucidate the genetic architecture of a trait and to facilitate marker-assisted selection (MAS). Previously, a few quantitative trait loci (QTLs) mappings and genome-wide association studies have reported genomic regions that control wheat dough rheological and baking traits, including WAC. For instance, Ma et al. (2007) conducted a QTL mapping analysis on 92 doubled haploid populations of Australian wheat. The peak for water absorption capacity was located on chromosome 5A. Fox et al. (2013) conducted a QTL analysis for water absorption and flour yield using a doubled haploid population of 162 hard wheat individuals, and they reported QTLs on chromosome 4D. Similarly, Tsilo et al. (2013) performed composite interval mapping for farinograph water absorption and dough rheological properties using 139 hard red spring recombinant inbred lines (RILs). Six QTL on chromosomes 1A, 1B, 4B, 4D, and 5A were detected for farinograph water absorption. Campbell et al. (2001) reported several QTLs associated with damaged starch, alkaline water retention capacity (AWRC), and dough water absorption in a study using 78 RILs from a soft-hard wheat cross. For dough water absorption capacity, QTLs were identified on chromosomes 5A, 5B, and 5D for damaged starch and AWRC on chromosomes 4B, 4D, 5A, 5B, and 5D.

In addition, Xiao-ling et al. (2023), using linkage mapping and association mapping in a recombinant inbred line population consisting of 173 lines and an association panel of 205 common wheat varieties, identified 31 QTLs associated with SRC traits consisting of 57 QTNs, with two common chromosomal regions on chromosomes 1B and 4B found across both mapping populations. Similarly, Lou et al. (2021) conducted GWAS on 486 common wheat genotypes and reported 14 QTNs associated with water absorption capacity on chromosomes 2D, 3A, 3D, 5A, 5D, and 7A. Lastly, Navrotskyi et al. (2020) identified two single nucleotide polymorphisms (SNPs) on chromosome 4A associated with water absorption capacity in a study of 299 hard red winter wheat genotypes. Significant QTNs have been consistently reported on groups one, four, and five chromosomes across studies (Lou et al., 2021; Campbell et al., 2001; Navrotskyi et al., 2020), highlighting the crucial role of these chromosomes.

Despite WAC being a vital determinant of end-use quality in hard wheat, most of the previous GWAS studies were on soft wheat classes, and studies are still limited to hard winter wheat. Therefore, this study was conducted to: 1) elucidate the genetic architecture underlying WAC in hard winter wheat using multi-locus GWAS models, which offer higher power to detect marker-trait associations compared to previously reported single locus models; 2) to identify candidate genes involved in WAC and related pathways, thereby providing valuable insights into the genetic mechanisms governing this trait in hard wheat, and 3) Evaluate the impact of favorable allele combinations of significant marker-trait associations on each SRC trait which will help to determine the potential use in the breeding program.

2.3 Materials and Methods

2.3.1 Plant materials and experimental designs

Three hundred and thirty-seven genotypes generated by the Colorado State University (CSU) wheat breeding program were used in this study. Hereafter, the term 'genotype' will refer to both unreleased experimental lines and released cultivars. Genotypes were evaluated over a period of five years (2017-2021) in three independent breeding trials, namely the CSU Elite Trial (ELITE) and the Advanced Yield Nursery for conventionally derived genotypes (AYN) and doubled haploid-derived genotypes (AYND). Trials were carried out at CSU wheat breeding locations in the Great Plains wheat growing region of eastern Colorado, USA. All trials were grown in farmer's fields, except for trials conducted at the Agriculture Research, Development, and Educational Center (ARDEC) in Fort Collins, CO, and the United States Department of Agriculture Research Service Central Great Plains Center in Akron, CO. As such, the agronomic and crop management practices mirrored those used by the grower cooperators and varied according to the standard practices at each location.

Trials were arranged in resolvable, latinized row-column designs with partial replication following the methodology of John and Williams (1995) and Williams et al. (2011). In the CSU Elite trial, within a given location, half of the entries were replicated twice, and the other half were replicated once. A similar approach was used for the AYN and AYND, with approximately one-seventh of the entries having a second replicate at any given location. Each plot was six rows, 1.5 m wide and 3.7 m long, seeded at approximately 1.73 million seeds ha^{-1}. The grain was harvested from all six rows of each plot. A detailed description of the number of genotypes in each year-trial combination is given in Table 2.1.

2.3.2 Phenotyping

To obtain white flour samples for SRC tests, 50 g grain samples from each genotype were processed by tempering to 14% moisture and milled using a Quadrumat Junior or modified Quadrumat Senior experimental mill from Brabender (South Hackensack, NJ, USA). Flour moisture and protein concentration were determined using near-infrared reflectance spectroscopy (NIR) (Foss DS2500^{TM} Feed and Forage analyzer, Foss North America, Eden Prairie, MN).

Solvent retention capacity (SRC) tests were conducted on the milled flour samples. Two mL microcentrifuge tubes for each sample were labeled and pre-weighed. To conduct the SRC tests, 200 +/- 10 mg of flour from each sample was placed in each tube, and each of the four solvents (water, sucrose, lactic acid, sodium carbonate) was used, according to AACC International (2010). A 50% weight/volume (w/v) sucrose solution was prepared by dissolving one part of sucrose crystals in one part of double-distilled water, whereas the sodium carbonate and lactic acid solutions were prepared as 5% w/v. The flour and the solvent were vigorously mixed for 5 sec to suspend the flour. The mixtures were then shaken for 20 min on a rotator, allowing the flour to mix well. Samples were then immediately centrifuged for 15 min at room temperature. After removal from the centrifuge, the supernatant was decanted, and the remaining pellet was left to dry for 10 min before weighing. To obtain the pellet weight, the initial tube weight was subtracted from the subsequent weight of the tube with the pellet. The SRC value (%) for each solvent was determined using the following formula:

$$SRC(\%) = \left[\frac{(pellet\ weight(g)}{(flour\ weight(g)} - 1\right] x \left[\frac{(86)}{(100 - moisture(\%)}\right] x\ 100$$

where moisture (%) is the NIR-determined moisture of the flour. The SRC test produced four SRC values for each sample, corresponding to the four solvents used. Gluten performance index (GPI), a more comprehensive measurement of the overall performance of flour glutenin in the context of other modulating networks of flour polymers (Kweon et al., 2011), was calculated according to Zhang et al. (2007) using the following formula:

$$GPI\ = (SRC\text{–}\ lactic\ acid)/(SRC\text{–}\ sucrose\ +\ SRC\text{–}\ carbonate)$$

2.3.3 Phenotypic data analysis

All phenotypic data were analyzed using code and packages written in R statistical coding language (R Core Team, 2023). Across locations, best linear unbiased estimates (BLUEs) were calculated in a one-step analysis using *ASREML-R* (Butler et al., 2009). The following mixed linear model was used to estimate BLUEs for each genotype:

$$y_{ijk} = \mu + G_i + e{:}\,r_j + e{:}\,c_k + \varepsilon_{ijk}$$

Where Y_{ijk} is the response variable for i^{th} level of genotype, in the j^{th} level of the row, in the k^{th} level of the column; μ is the overall mean; G_i is the fixed genotype effect; $e{:}r_j$ is the random row within the trial ($r_j \sim N(0, \sigma_j^2)$, $e{:}c_k$ is the random column within the trial ($c_k \sim N(0, \sigma_k^2)$ and ε_{ijk} is the residual error term $\varepsilon_{ijk} \sim N(0, \sigma_\varepsilon^2)$. Summary statistics (minimum, maximum, and average values) for SRCs and GPI were also calculated. Pearson correlation among pairs of BLUEs was done and visualized using the *psych* package in R (Revelle and Revelle, 2017).

To estimate variance components and calculate genomic heritability, genotypes were assigned as a random effect and analyzed with a mixed linear model using *ASREML-R* (Butler et al., 2009). The following mixed linear model was used:

$$y = Xb + Zu + e$$

where X and Z are known design matrices; b is a vector of fixed effects; u is a vector of random genetic effects with distribution $u \sim N(0, \sigma_g^2 K)$ where K is the kinship matrix and σ_g^2 is the genetic variance; e is a vector of random residuals with distribution $e \sim N(0, \sigma_e^2 I)$ where σ_e^2 is the residual variance, and *I* is the identity matrix. Genomic heritability was then calculated using the following formula:

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

where σ_g^2 is the genetic variance and σ_ε^2 is the residual error associated with the trait.

2.3.4 Genotyping and quality control

Genomic DNA was extracted from one-week-old leaves in a 96-well format using the *MagMAX*™ Plant DNA Isolation kit (Thermo Fisher Scientific, MA, USA), optimized for use with the KingFisher Flex magnetic particle processor equipped with a 96DW (96-well-deep well setting) particle head. DNA concentration was determined using PicoGreen (Thermo Fisher Scientific, MA, USA), which allowed normalizing DNA concentration to 20 ng µL^{-1} for library construction. Libraries were created using the *PstI-MspI* restriction enzyme combination (Poland et al., 2012) and pooled together at a 384-plex level.

Sequencing was performed on an Illumina HiSeq 2500 system at a core lab at the University of Illinois. Single nucleotide polymorphism (SNP) calling was performed using *TASSEL GBSv2* pipeline (Glaubitz et al., 2014), with a 64-base *k*-mer length and a minimum *k*-mer count of five. The Burrows-Wheeler aligner version 0.7.10 (Li and Durbin, 2009) was used to align reads to the International Wheat Genome Sequencing Consortium (IWGSC) RefSeq v1.0' Chinese Spring' wheat reference genome (Appels et al., 2018).

The raw SNP data obtained from *TASSEL GBSv2* underwent a series of quality control steps. Initially, genotypes with over 50% missing calls and over 30% heterozygosity were filtered out. The filtration criteria were further tightened to retain only biallelic SNPs with a minor allelic frequency exceeding 5%, missing calls below 10%, and a maximum heterozygosity of 10%. Unaligned SNPs were discarded in the subsequent quality control phase. Imputation was then performed on the pruned dataset using the Beagle algorithm (Browning et al., 2018), resulting in a final set of 23,130 high-quality SNPs for downstream analyses. The marker data used in this study represents a marker density of approximately 1.36 SNPs per 1 megabase. On average, the SNP distribution over 21 chromosomes was 1,101 polymorphic SNPs per chromosome, but the distribution was non-uniform, with a minimum of 264 SNPs for chromosome 4D and a maximum of 1,955 SNPs for chromosome 2B (Figure 2.1).

2.3.5 Population Structure and Linkage Disequilibrium

For purposes of GWAS analysis, both the Q-matrix and Kinship matrix were considered in the GWAS analysis to control the confounding effects of population stratification and individual relatedness on the associations between genetic markers and traits. To ascertain population structure, the optimal number of subpopulations (denoted as "K") was determined using the entropy criterion. This method identifies the "K" value with the least cross-validation error, as detailed by Alexander and Lange (2011) and Frichot et al. (2014). By testing a range of 1 to 20 potential populations and iterating each ten times, the *sNMF* algorithm (implemented in the *LEA v 2.2.0* package in R) was employed. The outcome of this procedure was the ancestral coefficient (Q) for each genotype, based on the defined "K". Additionally, the genetic relationship between individuals (kinship) was calculated using *TASSEL* (Bradbury et al., 2007).

Additionally, as an alternative to the ancestral coefficient matrix (Q-matrix), principal component analysis (PCA) was performed using the raw marker data via the *prcomp* function in base R to analyze population structure (R Core Team, 2013). The optimal number of clusters (appropriate number of populations = K) was assigned based on the assumption of the majority rule from the *NbClust* package in R (Charrad et al., 2014). To support the population structure analyses, a neighbor-joining (NJ) phylogenetic tree was constructed using *TASSEL 5.2* (Bradbury et al., 2007) and visualized using *interactive Tree of Life (iTOL) v3*, an online tool for display annotation and management of phylogenetic tree (Letunic and Bork, 2016).

To estimate linkage disequilibrium (LD) among the markers, which further helps to delineate the size of the region influenced by GWAS-detected markers, pairwise LD values (r^2) were computed in *TASSEL* and plotted against physical distance (bp) in base R. The pattern of LD decay was determined as the distance where LD values are reduced to half of their maximum value (Remington et al., 2001).

2.3.6 Genome-Wide Association Study

To capture the genetic bases underlying WAC and contributing flour components, genome-wide association analysis was performed for each of the four SRC traits and GPI based on six different multi-locus GWAS models: Iterative Modified-Sure Independence Screening *EM-Bayesian Lasso (ISIS EM-BLASSO)* (Tamba et al., 2017); Multi-locus Random-SNP-Effect Mixed Linear Model *(mrMLM)* (Wang et al., 2016); Fast Multi-locus Random-SNP-Effect Efficient Mixed-Model Association *(FASTmrEMMA)* (Wen et al., 2018); Integration of Kruskal-Wallis Test with Empirical Bayes under polygenic background control *(pKWmEB)* (Ren et al., 2018); Integration of Least Angle Regression with Empirical Bayes *(pLARmEB)* (Zhang et al., 2017); and Fast Multi-locus Random-SNP-Effect Mixed Linear Model *(FASTmrMLM)* (Tamba and Zhang, 2018). All GWAS models were implemented using the *mrMLM* package in R (Wang et al., 2016). P-values were adjusted according to the Benjamini-Hochberg (1995) procedure, and those below 0.05 were considered significant.

Genotypes carrying one to six favorable alleles of significant MTAs were visualized using box plots to evaluate the effect of different combinations of favorable alleles. In addition, a stepwise linear regression analysis was conducted using the top six significant MTAs to understand the nature of their effects on the associated traits. Each quantitative trait nucleotide (QTN) was modeled independently to assess its additive effect. Subsequently, interaction terms were introduced to the model to identify potential non-additive effects, such as epistasis. The analysis utilized linear regression models, with the significance of QTNs and their interactions determined based on t-values and p-values. If individual QTNs show a significant p-value at a critical value of $\alpha < 0.05$, then the effect is considered additive. If the interaction terms are significant, it suggests the presence of non-additive effects.

2.3.7 Co-localization of QTNs and relevant genes for water absorption capacity

A co-localization analysis was done to reveal how closely significant QTNs were linked to genes previously reported to affect WAC either directly or indirectly through the synthesis of flour components such as gluten protein, starch, and arabinoxylans. The positions of the annotated genes involved in the synthesis of flour components and grain characteristics were retrieved from Ensemble Plants (*https://plants.ensembl.org*) with the International Wheat Genome Sequencing Consortium (IWGSC) RefSeq v1.0' Chinese Spring' (Appels et al., 2018). In addition, the KnetMiner (*https://knetminer.com/cereals*) wheat database was also used. Then, the positions of the genes were cross-referenced with the positions of the significant MTAs identified in this study.

Using the calculated LD decay values of $r^2 = 0.3$ as a threshold as a guideline (Figure 2.2), genes that are up to 4 Mb away from the significant QTNs were included as colocalized genes. To validate the associations between the identified genes and the QTNs within the specified search windows, an LD-block analysis was performed using the *gpart* package in R (Kim et al., 2019). This analysis aimed to confirm the robustness of the genetic associations uncovered in the GWAS results.

2.3 Results

2.3.1 Phenotypic evaluations and correlations

Summary statistics and heritability for SRCs and GPI are presented in Table 2.2. The SRC-water ranged from 54.1% to 66.5%, with a mean of 59.0%. SRC-sucrose ranged from 74.5 to 99.3%, with a mean of 87.6%. SRC-lactic acid ranged from 93.9% to 123.2%, with a mean of 111.0%. GPI had a mean of 0.70 and ranged from 0.60 to 0.90. All SRCs and GPI showed moderate to high heritability (Table 2.2), with the lowest heritability ($h^2 = 0.69 \pm 0.04$) recorded for SRC-water and the highest heritability ($h^2 = 0.88 \pm 0.2$) observed for SRC-sucrose.

Pearson's correlation coefficients ranged from r = -0.62 to r = 0.83; all identified correlations were highly significant ($p < 0.001$) (Figure 2.3). A strong positive correlation (r = 0.83; $p < 0.001$) was observed between SRC-water, which is the overall WAC, and SRC-carbonate, a proxy for the degree of damaged starch. Likewise, SRC-lactic acid, which is a measure of flour gluten strength, had a positive correlation (r = 0.38; $p < 0.001$) with SRC-sucrose, a measure of flour pentosan (arabinoxylan) concentration.

A negative correlation was observed between SRC-water and GPI (r = -0.62, $p < 0.001$). Strong positive correlations were observed among SRC-water, SRC-sucrose, and SRC-carbonate, with each correlation coefficient being above 0.70 ($p < 0.001$). Moderate correlations were observed between GPI and SRC-water (r = -0.45; $p < 0.001$), SRC-sucrose (r = 0.55; $p < 0.001$), and SRC-lactic acid, as well as between SRC-lactic acid and SRC-water (r = 0.29; $p < 0.001$) and SRC-sucrose (r = 0.38; $p < 0.001$). SRC-lactic acid and SRC-carbonate had a lower correlation (r = 0.22; $p < 0.001$).

2.3.2. Population structure

Principal component analysis (PCA) was conducted to discern and account for population stratification. The first three principal components explained a combined 22.94% of the genetic variance (Figure 2.4, A and B). In the PCA plot (A) PC1 vs. PC2, three distinct clusters were evident. The clustering might be attributed to the genetic backgrounds of the genotypes from the three main parent varieties in the CSU wheat breeding program. Cluster one, mainly encompassing 'Byrd' (PI 664257; Haley et al., 2012) and its derivatives, was positioned toward the higher end of PC2 and across PC1. Cluster two, consisting largely of 'Denali' (PI 664256; Haley et al., 2012) and its derivatives, fell at lower values on both PC1 and PC2. Cluster three, predominantly including 'Snowmass' (PI 658597; Haley et al., 2011) and 'Snowmass 2.0' (PI 691605) and their derivatives genotypes, occurred at the upper end of both PC1 and PC2. However, some genotypes did not align closely with any primary cluster, suggesting mixed genetic backgrounds. In PCA plot (B), PC1 vs. PC3, genotypes from clusters one and three maintained positions similar to plot (A). However, cluster two genotypes displayed a more dispersed pattern, merging somewhat with the other clusters.

To further elucidate the population structure, a Neighbor-Joining phylogenetic tree was constructed. Similar groupings were revealed among genotypes, with three major clades exhibiting substantial overlap and shared ancestral backgrounds (Figure 2.4 and Figure 2.5). The PCA and phylogenetic analyses provided clear visual representations of the genetic diversity and stratification within the studied genotypes. As such, a Q-matrix and a kinship matrix were included in the GWAS models to minimize spurious genotype-phenotype associations due to population substructure.

2.3.3 Genome-wide association study

Considering each of the four SRCs and GPI, a total of 42 QTNs across 17 chromosomes were identified using six GWAS models, some of which were associated with multiple traits (Table 2.3, Table 2.4; Figure 2.6). Chromosome 1B had the highest number of QTNs (11), with associations for each of the four SRCs and GPI. Chromosome 3B had five QTNs associated with SRC-sucrose and SRC-lactic acid. Chromosome 1A had four QTNs, three associated with SRC-sucrose and one associated with both SRC-lactic acid and GPI. Chromosomes 5B, 6B, and 7A each had three QTNs associated with specific SRC phenotypes. Chromosomes 2D and 4D each had two QTNs associated with different SRCs, and the remaining chromosomes had a single QTN, while chromosomes 1D, 3D, 5D, and 7D had no significant associations. The highest number of QTNs were associated with SRC-sucrose and SRC-lactic acid (14 for each). Nine QTNs were associated with SRC-water, seven were associated with SRC-carbonate, and four were associated with GPI.

The most significant QTN was S4D_39790394, identified by five different models with -$\log_{10}$(p) values ranging from 8.36 to 12.16, which were associated with SRC-water and SRC-carbonate. Another significant QTN was S1B_6922434, with -$\log_{10}$(p) values ranging from 6.93 to 11.90, which was associated with both SRC-water and SRC-sucrose. A third significant QTN was S1B_591353377, which was associated with both SRC-sucrose with -$\log_{10}$(p) of 10.02 and SRC-lactic acid, with -$\log_{10}$(p) value of 11.60. with MAF of 0.41

The proportion of phenotypic variance explained (R^2) by individual QTNs ranged from 1.58% to 13.15%, indicating small to moderate effects of individual QTNs (Table 2.3). The highest R^2 value was observed for QTN S1B_8678756, which was associated with SRC-carbonate (explaining 13.15% phenotypic variance), while the minimum R^2 value was for QTN

S1B_3282665 (1.58%), which was associated with SRC-water. Five QTNs (S1B_8678756, S4B_535089330, S1B_6922434, S4D_39790394 and S4D_472908721) exhibited ≥ 10% R^2.

Several QTNs were associated with more than one solvent retention capacity trait. Two QTNs, S1B_591353377 (R^2 = 3.24% - 5.32%) and S6B_147637875 (R^2 = 3.13% - 3.81%) were shared by SRC-sucrose and SRC-lactic acid. At the same time, S4B_535089330 (R^2 = 3.77% - 10.38%) and S4D_39790394 (R^2 = 6.56% - 7.73%) were shared by SRC-water and SRC-carbonate. Similarly, S1A_5190318 (R^2 = 3.85% - 6.41%) was shared by SRC-lactic acid and GPI, while S1B_6922434 (R^2 = 4.97% - 11.18%) was shared by SRC-water and SRC-sucrose.

Some of the QTNs identified in this study, such as S2A_787541163 and S1B_3767944, both associated with SRC-lactic acid, were rare in the population with minor allele frequencies (MAFs) of 0.06 and 0.07, respectively. Other QTNs, like S4B_535089330 with 0.08 MAF and S4D_39790394 with 0.20 MAF, both associated with SRC-water and SRC-carbonate, also had low frequencies. In contrast, QTNs such as S1B_595397803 (SRC-lactic acid), S1B_649243794 (SRC-carbonate), S5B_477143477, and S7A_3966595 (both SRC-water) were found at much higher frequencies, with MAFs of 0.49. The remaining significant QTNs were observed at moderate frequencies within the range between the highest and lowest MAFs observed.

In this study, several QTNs were captured by more than one model. Out of the 43 QTNs identified, 18 were captured by more than one model: eleven by two models, three by three models, three by four models, and one by five different models (Table 2.3, Table 2.4). The other 24 QTNs were identified by only one model. For instance, QTN S1B_649243794 was identified by the *FASTmrEMMA, FASTmrMLM*, *ISISEM-BLASSO*, *mrMLM,* and *pKWmEB* models, demonstrating associations with SRC-water and SRC-carbonate. Overall, both the *FASTmrMLM* and ISISEM-

Bayesian Lasso models detected QTNs associated with each trait, whereas the *pKWmEB* model did not detect any QTN for SRC-water or GPI.

Genotypes were grouped based on the favorable alleles of the significant marker-trait associations (MTAs) they carried to see the combined effects of the top six QTNs for each SRC trait. In general, with an increase in the number of favorable alleles present, the value of associated SRC traits also gradually increased (Figure 2.7). Genotypes carrying six favorable alleles for each SRC trait demonstrated a higher median value compared to genotypes carrying fewer than six QTNs. For all SRC traits, QTNs exhibited a more incremental additive effect with only small changes observed.

Stepwise linear regression analysis of all SRC traits showed that individual QTNs were significant ($\alpha < 0.05$), confirming that traits were affected by the associated QTNs in an additive manner. Analysis revealed no epistatic interactions among the markers, as evidenced by non-significant interaction terms (Table 2.6).

2.3.4 Co-localization of QTNs and relevant genes for water absorption capacity

To validate the relevance of newly identified QTNs and to understand their relationship with established genetic markers, co-localization with known genes was investigated. Genes affecting WAC both directly and indirectly through flour components such as gluten, arabinoxylan, and damaged starch, as well as through grain hardness, were considered. Forty-three annotated genes colocalized with 15 of the 42 significant QTNs identified in this study (Table 2.7), with key genes like *LMW-GS*, gliadins*, GT*, and *SS* located near the significant QTNs.

Some QTNs, such as S1B_3282665, S1B_3767944, S1B_6922434, S1B_7235112, and S1B_8678756, were located 1.72 Mb away from *TraesCS1B02G010600*, which encodes the Delta gliadin-B1 (*Gli-B1-1*) gene, and 1.68 Mb away from *TraesCS1B02G010400*, which encodes the

Gamma-gliadin-3 (*Gli-B1-3*) gene. These genes and the QTN 1B_3767944 are in strong LD, as evidenced in Figure 2.8 (d). Similarly, S1A_5190318 (associated with SRC-sucrose and SRC-lactic acid) was located near *TraesCS1A02G007405* (1.14 Mb), *TraesCS1A02G007400* (1.15 Mb), *TraesCS1A02G007300* (1.34 Mb), and *TraesCS1A02G007700* (1.13 Mb), all encoding Gamma gliadin-A3 (*Gli-A1-3*) genes. S1A_5190318 is again located near *TRITD1Av1G002790* (0.96 Mb), *TRITD1Av1G002360* (0.14 Mb), and *TRITD1Av1G002310* (0.24 Mb), known for encoding Low molecular weight glutenin genes (*Glu-B3*). LD block analysis in this region revealed that the QTN and the genes are in strong linkage disequilibrium (~ r2 = 0.9) (Figure 2.8 (a)). Additionally, on chromosome 1B, QTNs including S1B_3282665, S1B_3767944, S1B_6922434, S1B_7235112, and S1B_8678756, associated with various SRC traits, were found near *TraesCS1B02G010600*, *TraesCS1B02G010500*, *TraesCS1B02G010400*, and *TraesCS1B02G011000*, (1.56 Mb -1.89 Mb) which also encode gliadins.

Other QTNs associated with all SRC traits and GPI, found on chromosomes 1B, 3B, 4D, 6A, 6B, and 7A, were located near genes encoding the glycosyltransferase (*GT*) gene family. For example, S1B_66284969 (associated with GPI) on chromosome 1B is situated 0.73 Mb away from *TraesCS1B02G083100*, which encodes *GT* genes. Similarly, S1B_3282665 (associated with SRC-water) is 1.59 Mb from *TraesCS1B02G002100,* which encodes *GT* genes. Additionally, QTNs such as S3B_576577312 (associated with SRC-lactic acid) is located near *TraesCS3B02G313900* (2.23 Mb), and S4D_472908721 (associated with SRC-water) is near *TraesCS4D02G306800* (2.01 Mb), both encoding glycosyltransferase genes. S4D_472908721 is in strong LD with *TraesCS4D02G306800* (~r2 = 0.9) (Figure 2.8 (e)).

Likewise, S6A_6869286 (associated with SRC-lactic acid) is near *TraesCS6A02G018000* (1.92 Mb), S6B_147637875 (associated with SRC-sucrose and SRC-lactic acid) is near

TraesCS6B02G144400 (3.85 Mb), and S7A_31213757 (associated with SRC-sucrose) is near *TraesCS7A02G057400* (3.60 Mb), all encoding GT genes. The LD-block analysis on chromosomes 1B and 7A showed that S1A_5190318 and S7A_31213757 are in strong LD (approximately r2 = 0.9) with the *GT* genes (Figure 2.8 (a) and (f), respectively). Moreover, two QTNs, S1B_595397803 (associated with GPI) located near *TraesCS1B02G368500* (3.48 Mb) and S7A_37433960 (associated with SRC-sucrose) near *TraesCS7A02G070100* (1.66 Mb), encode for starch synthase (*SS*) and chloroplastic/amyloplastic genes.

2.4 Discussion

Hard winter wheat is a preferred class for bread making due to its higher protein and gluten concentration, as well as its higher water absorption capacity (Mallory et al., 2012; Sapirstein et al., 2018). Genotypes included in this study exhibited lower average SRC-water values (59%), which is low for bread-making. SRC-water measures the overall water absorption capacity of the flour and is influenced by various flour components, such as damaged starch, pentosans, and gluten proteins. Reduced levels of SRC-water could be attributed to the absence of favorable alleles controlling these traits or, if present, to their low expression levels. Environmental stressors, which are common in the U.S. Great Plains, including drought and extreme temperatures, can significantly impact the expression of genes associated with end-use quality (Alsamman et al., 2021; Pandey et al., 2022). These stressors also affect the synthesis of glutenin proteins (Ronga et al., 2020; Wan et al., 2022) and arabinoxylans (Henry et al., 1986; Tremmel-Bede et al., 2020).

Each of the evaluated SRC traits and GPI showed moderate to high heritability, in agreement with previous reports (Cabrera et al., 2015; Smith et al., 2011). High heritability suggests that the greater portion of the variation in these traits is due to genetic factors, and selection can result in genetic improvements (Bernardo, 2014). The strong positive correlation observed between SRC-sucrose and SRC-carbonate and the weak correlation observed between SRC-carbonate and SRC-lactic acid is consistent with other studies (Guttieri et al., 2001; Guzman et al., 2015). Furthermore, the strong positive correlation observed between SRC-water and both SRC-sucrose and SRC-carbonate indicates that the higher the pentosans and damaged starch in the flour, the higher the overall WAC, consistent with previous reports (Colombo et al., 2008; Duyvejonck et al., 2011; Gaines, 2000; Guttieri et al., 2001; Guzman et al., 2015).

In this study, the observed sharing of QTNs among SRC traits and GPI, particularly between SRC-lactic acid and SRC-sucrose, SRC-lactic acid and GPI, and SRC-water and SRC-carbonate, matches the observed phenotypic correlations between these traits. This alignment supports the previous findings of Guzman et al. (2015) and Jiang et al. (2017), reinforcing the genetic interconnectedness of these traits. Additionally, it suggests that the genes responsible for these traits could reside in the same region, or the same genes might affect both traits in a pleiotropic manner.

The genotypes, grouped based on the favorable alleles from the top six QTNs, showed an increase in the associated trait value as the number of alleles increased. In the stepwise regression analysis, the individual QTNs were significant, while the interaction terms were non-significant, suggesting that the marker effects predominantly reflect additive genetic influences and the absence of non-additive effects. This additive effect indicates that each favorable allele contributes directly and independently to the phenotypic expression of SRC traits. This finding suggests that selecting these specific alleles in breeding programs would lead to improvements in SRC traits.

Significant QTNs identified in this study were distributed across 17 chromosomes. Chromosomes 1A, 1B, and 3B are significant hubs harboring many QTNs (20 out of 42) that are associated with all SRCs and GPI, suggesting these chromosomes are the crucial regions for WAC and flour components associated with WAC. Previous studies reported that significant QTNs and QTLs on chromosomes 1A and 1B are associated with WAC, grain hardness, grain weight, and milling traits (Aoun et al., 2022; Gaire et al., 2019; Garcia-Santamaria et al., 2018; Ibba et al., 2021). At the same time, other studies have reported important QTLs in these regions for protein concentration, gluten concentration, starch concentration, and thousand kernel weight in both hard and soft wheat (Goel et al., 2019; Kerfal et al., 2010; McCartney et al., 2006; Pu et al., 2020). Current and previous studies suggest that, apart from the well-known gluten genes, there may be additional genes on chromosomes 1A and 1B that play a significant role in regulating the end-use quality of wheat.

The proximity of some QTNs on chromosomes 1A and 1B to the glutenin (*Glu-B3*) and gliadin genes (*Gli-A1-3, Gli-B1-1, Gli-B1-9*, and *Gli-B1-3*) suggests their influence on wheat's water absorption capacity. The HMW glutenin genes are located on chromosomes 1A, 1B, and 1D at the *Glu-A1*, *Glu-B1*, and *Glu-D1* loci, whereas LMW glutenin genes are located on the short arms of these chromosomes at the *Glu-A3*, *Glu-B3,* and *Glu-D3* loci (Payne et al., 1980; Shewry et al., 2003). Gliadin genes are primarily located on the short arms of group 1 chromosomes (1A, 1B, and 1D) at the *Gli-A1, Gli-B1*, and *Gli-D1* loci. These loci encode gamma, delta, and omega gliadins. Additionally, the *Gli-A2, Gli-B2, and Gli-D2* loci on the short arms of chromosomes 6A, 6B, and 6D encode for alpha and beta gliadins (Branlard et al., 2001; Payne et al., 1984; Shewry, and Halford, 2002).

The QTN S1A_5190318, which is associated with SRC-sucrose and SRC-lactic acid, was located near low molecular weight glutenin (*Glu-B3*) and gliadin (*Gli-A1-3*) genes. These genes were in strong LD (approximately $r^2 = 0.9$) with S1A_5190318, suggesting the association between them is not random, and they can be inherited together. SRC-sucrose is the proxy for pentosan (arabinoxylan) concentration and gliadin characteristics, while SRC-lactic acid is for glutenin characteristics, mainly gluten quality (Dubat et al., 2019; Labuschagne et al., 2021). The location of QTN S1A_5190318 near the glutenin and gliadin genes indicates its potential as a genetic marker. This marker could be instrumental in wheat breeding programs, especially for traits influenced by gluten proteins, such as water absorption capacity.

Similarly, QTNs such as S1B_3282665, S1B_3767944, S1B_6922434, S1B_7235112, and S1B_8678756 were 0.14 Mb to 2.09 Mb away from the gliadin genes. These genes and the QTNs are in a strong LD (approximately $r^2 = 0.7$) range, suggesting a non-random association between these QTNs and the gluten genes, indicating a strong likelihood of co-inheritance, implying that selecting gluten genes could enhance water absorption capacity in wheat.

Arabinoxylans/pentosans are cell wall polysaccharides that make up 65–70% of the polysaccharide content in wheat and affect the water absorption capacity of the flour (Izydorczyk, 2021; Kosik et al., 2017). Genes such as glycosyltransferase families, xyloglucan endotransglucosylase/ hydrolase (*XTH*) genes, and xylanases genes are reportedly involved in the regulation of pentosan synthesis and concentration (Lovegrove et al., 2013; Li et al., 2021). Glycosyltransferases are known for their involvement in the regulation of starchy and non-starch polysaccharides, including arabinoxylans (Li et al., 2021; Lovegrove et al., 2013). Some of the QTNs in this study, such as S1B_7235112 (SRC-carbonate), S1B_66284969 (GPI), S3B_576577312 (SRC-lactic acid), S4D_472908721(SRC-water) and S6B_147637875 (SRC-sucrose and SRC-lactic acid) were located near the glycosyltransferase genes (*GT*) indicates that the crucial role of the *GT* genes in water absorption capacity of a wheat flour and synthesis of flour components such as arabinoxylans. Downregulating the expression of glycosyltransferase gene families can reduce cell wall arabinoxylans in the endosperm by 40–50% (Izydorczyk, 2021; Kosik et al., 2017).

In addition, QTNs from the current study, including S1B_649243794 (SRC-carbonate), S1B_648546585 (SRC-sucrose), and S1B_6922434 (SRC-water and SRC-sucrose) were colocalized with QTNs from previous reports such as S1B_642650411, S1B_651557718, and S1B_14665450 that are associated with total and water extractable arabinoxylans (Ibba et al., 2021). Those QTNs from current and previous reports were located 2 Mb – 3 Mb away from each other. Given that the calculated LD decay ($R^2 = 0.3$) is very long (3.7 Mb), the current result and previously reported QTNs are still in strong LD, and both likely carry the same underlying variant.

Few QTNs, S1B_595397803 (associated with GPI) located near TraesCS1B02G368500 (3.48 Mb) and S7A_37433960 (associated with SRC-sucrose) near TraesCS7A02G070100 (1.66 Mb), encode for starch synthase (SS) and chloroplastic/amyloplastic genes. Starch synthase enzymes (*SS*) are responsible for the short chains of glucose polymers between branched clusters. These enzymes are important for the organization of the higher structure of starch granules (James et al., 2003). However, native starch (undamaged) absorbs relatively less water than damaged starch. In general, the proximal location of QTNs from the current study with annotated genes and previously reported QTNs and QTLs could be the conformation of the reliability of the marker-trait associations in this study.

Chapter 2 Tables

Table 2.1 Number of observations and unique genotypes used by trial and year.

Year	Trial	Number of Locations	Number of Observations	umber of Unique Genotypes
2017	AYN[a]	2	51	27
	AYND[b]	3	91	70
	ELITE[c]	2	70	35
2018	AYN	1	30	30
	AYND	1	18	18
	ELITE	4	155	50
2019	ELITE	6	785	100
2020	AYN	3	93	31
	AYND	3	116	53
	ELITE	4	600	100
2021	ELITE	4	590	100

Abbreviations are as follows: [a] Advanced Yield Nursery; [b] Advanced Yield Nursery of doubled haploid derived lines; [c] CSU Elite Trial.

Table 2.2 Summary of phenotypic variation and heritability of solvent retention capacity traits and gluten performance index.

	Descriptive Statistics			Heritability	
Trait	Mean	Minimum	Maximum	h^{2a}	Se[b]
SRC-water (%)	59.0	54.1	66.5	0.69	0.04
SRC-sucrose (%)	87.6	74.5	99.3	0.88	0.02
SRC-lactic acid (%)	111.0	93.9	123.2	0.82	0.05
SRC-carbonate (%)	84.9	74.1	93.9	0.80	0.04
Gluten performance index	0.70	0.60	0.90	0.61	0.05

Abbreviations: [a] broad sense heritability; [b] Standard error; SRC, solvent retention capacity.

Table 2.3 Significant quantitative trait nucleotides detected by each model based on the -log10(p) value.

QTN	Trait	Model	-log10(p)	R^2 (%)	MAF
S1A_353657879	SRC-S	3	6.04	2.75	0.42
S1A_373624388	SRC-S	5	8.61	7.46	0.25
S1A_377252109	SRC-S	6	6.09	2.65	0.21
S1A_5190318	SRC-L, GPI	5,6	6.02, 7.18	3.85, 6.41	0.21
S1B_3282665	SRC-W	3,4	6.28, 6.36	1.58, 4.32	0.22
S1B_3767944	SRC-L	3	7.49	3.86	0.07
S1B_432872392	SRC-S	4	6.79	5.28	0.20
S1B_591353377	SRC-S, SRC-L	2,4	10.02, 11.60	3.24, 5.32	0.41
S1B_595397803	SRC-L	6	7.47	4.01	0.49
S1B_648546585	SRC-S	5,6	6.77, 9.53	5.34, 6.11	0.36
S1B_649243794	SRC-C	2,3,4,5	6.65-8.23	3.17-5.25	0.49
S1B_66284969	GPI	2	6.45	2.85	0.20
S1B_6922434	SRC-W, SRC-S	1,2,5	6.93- 11.90	4.97-11.18	0.22
S1B_7235112	SRC-C	5	7.71	9.51	0.22
S1B_8678756	SRC-C	4,5,6	6.25-10.74	6.18-13.15	0.14
S2A_787541163	SRC-L	4	7.54	2.79	0.06
S2B_444557348	SRC-L	5	6.28	5.71	0.24
S2D_18933955	SRC-C	3,6	5.85, 6.92	6.37, 8.37	0.10
S2D_624600035	SRC-C	3	6.04	6.26	0.07
S3A_741636006	SRC-L	1,2,4,6	5.81-9.58	2.13-5.53	0.42
S3B_576577312	SRC-L	2,4	6.56, 8.82	4.56, 6.08	0.15
S3B_609002613	SRC-S	3	7.69	3.59	0.22
S3B_677739256	SRC-L	3,5	5.23, 7.53	5.62, 9.89	0.18
S3B_772483971	SRC-L	6	6.96	3.16	0.19
S3B_851549246	SRC-S	4,6	5.71, 6.53	2.62, 3.45	0.34
S4A_698874476	SRC-S	2,3,4,6	5.93- 7.85	3.89-5.64	0.23
S4B_535089330	SRC-W, SRC-C	2,4,6	5.87-8.63	3.77-10.38	0.08
S4D_39790394	SRC-W, SRC-C	1,2,3,4,6	8.36-12.16	6.56-11.73	0.10
S4D_472908721	SRC-W	3,5	8.67,11.41	6.53, 10.97	0.12
S5A_551174071	SRC-S	6	6.75	2.37	0.29
S5B_444994793	SRC-W	2	7.36	5.63	0.38
S5B_444994802	SRC-W	4	6.54	7.19	0.38
S5B_477143477	SRC-W	2	6.13	3.77	0.49
S6A_6869286	SRC-L	6	7.63	3.5	0.35
S6B_147637875	SRC-L, SRC-S	6	7.82	3.81	0.26
S6B_163949211	GPI	3	6.46	3.32	0.13
S6B_30806573	SRC-L	5	6.41	3.4	0.44
S6D_49209864	SRC-L	3	6.13	2.48	0.25
S7A_31213757	SRC-S	3,4	6.71, 7.77	2.70, 5.76	0.33
S7A_37433960	SRC-S	5,6	7.14, 7.27	4.47, 7.41	0.34
S7A_3966595	SRC-W	3	6.35	2.76	0.49

S7B_654488500	GPI	6	8.05	1.95	0.13

Note: Numbers before and after the underscore indicate the chromosome and position of the QTN on the chromosome, respectively.
Abbreviations: Model number 1, *FASTmrEMMA*; 2, *FASTmrMLM*; 3, *ISISEM-BLASSO*; 4, *mrMLM*[; 5, *pKWmEB*; 6, *pLARmEB*. Solvent retention capacity (SRC), SRC-W using water as a solvent, SRC-S using sucrose as a solvent, SRC-L using lactic acid as a solvent, SRC-C using sodium carbonate as a solvent; GPI, gluten performance index; QTN, quantitative trait nucleotide; LOD, logarithm of odds; R^2, coefficient of determination indicating percent phenotypic variance explained; −log10p, −log of p-value; MAF, Minor allele frequency.

Table 2.4 Significant marker-trait associations identified in each chromosome.

Chromosome	QTN detected	Traits	Models
1A	4	SRCS-S, SRC-L	3,5,6
1B	11	SRC-W, SRC-S, SRC-L, SRC-C, GPI	1,2,3,4,5,6
2A	1	SRC-L	5
2B	1	SRC-L	3,6
2D	2	SRC-C	3
3A	1	SRC-C	1,2,4,6
3B	5	SRC-S, SRC-L	3,5,6
4A	1	SRC-S	2,3,4,6
4B	1	SRC-C	2,3
4D	2	SRC-W, SRC-C	1,2,3,4,6
5A	1	SRC-S	6
5B	3	SEC-W	2,4
6A	1	SRC-L	6
6B	3	SRC-L, SRC-S, GPI	3,5,6
6D	1	SRC-L	3
7A	3	SRC-W, SRC-S	3,4,5,6
7B	1	GPI	6

Abbreviations: solvent retention capacity (SRC), SRC-water (SRC-W), SRC-sucrose (SRC-S), SRC-lactic acid (SRC-L), SRC-sodium carbonate (SRC-C), and gluten performance index (GPI).

Table 2.5 The number of Significant marker-trait associations captured by each model.

Model	Number of QTNs detected	Associated trait
FASTmrEMMA	3	SRC-W, SRC-S, SRC-L, SRC-C
FASTmrMLM	9	SRC-W, SRC-S, SRC-L, SRC-C, GPI
ISIS EM-BLASSO	14	SRC-W, SRC-S, SRC-L, SRC-C, GPI
mrMLM	12	SRC-W, SRC-S, SRC-L, SRC-C
pKWmEB	11	SRC-S, SRC-L, SRC-C, GPI
pLARmEB	16	SRC-W, SRC-S, SRC-L, SRC-C, GPI
Total QTNs	65	
Unique QTNs	42	
QTNs identified by different models	22	

Abbreviations: SRC, solvent retention capacity; SRC-W, using water as a solvent; SRC-S, using sucrose as a solvent; SRC-L, using lactic acid as a solvent; SRC-C, using sodium carbonate as a solvent; GPI, gluten performance index.

Table 2.6 Summary results from stepwise linear regression of the top six quantitative trait nucleotides for each SRC traits

Coefficients:(SRC-water)	Estimate	Std.Error	t value	Pr(>|t|)
Intercept	56.67	0.28	200.11	0.00E+00
S1B_3282665	0.79	0.22	3.59	3.82E-04
S4D_39790394	3.15	0.51	6.22	1.53E-09
S5B_477143477	0.56	0.19	2.97	3.16E-03
S7A_3966595	0.34	0.19	1.77	7.71E-02
S1B_662659763	1.38	0.23	5.97	6.01E-09
S4D_472908721	-0.25	0.48	-0.52	6.06E-01
Interaction of all QTNs	0.51	0.77	0.67	5.05E-01
Coefficients: (SRC-sucrose)	Estimate	Std.Error	t-value	Pr(>|t|)
Intercept	85.63	0.52	166.03	2.740681e-319
S1A_373624388	-0.97	0.37	-2.67	8.04E-03
S1B_6922434	0.74	0.84	0.87	3.83E-01
S1B_432872392	1.76	0.88	2.00	4.59E-02
S1B_648546585	-0.47	0.35	-1.37	1.71E-01
S3B_609002613	2.40	0.43	5.52	6.91E-08
S4A_698874476	2.31	0.44	5.26	2.55E-07
Interaction of all QTNs	2.76	1.03	2.67	8.03E-03
Coefficients:(SRC-lactic acid)	Estimate	Std.Error	t-value	Pr(>|t|)
Intercept	108.76	0.42	257.39	0.00E+00
S1B_3767944	3.84	0.95	4.04	6.56E-05
S2A_787541163	3.18	1.05	3.02	2.73E-03
S3A_741636006	1.84	0.47	3.88	1.25E-04
S3B_576577312	0.61	0.69	0.89	3.76E-01
S3B_677739256	3.75	0.61	6.09	3.10E-09
S6A_6869286	0.37	0.47	0.79	4.31E-01
Interaction of all QTNs	-0.04	2.53	-0.02	9.88E-01
Coefficients:(SRC-carbonate)	Estimate	Std.Error	t-value	Pr(>|t|)
Intercept	68.68	0.69	99.447	2e-16
S1B_7235112	1.63	0.38	4.31	2.13E-05
S1B_649243794	1.70	0.36	4.71	3.71E-06
S1B_654538334	0.23	0.41	0.56	0.017921
S2D_18933955	1.89	0.45	4.16	4.15E-05
S2D_624600035	1.47	0.49	2.98	0.00314
S4D_39790394	4.96	0.61	8.10	1.09E-14
Interaction of all QTNs	2.4108	1.6127	1.495	0.13589

Abbreviations: SRC solvent retention capacity, Std.Error, Standard error

Table 2.7 Summary of colocalized quantitative trait nucleotides identified in this study and annotated genes

Gene-ID	Chrom	Significant QTNs	Traits	Distance Mbp	Description
TraesCS1A02G007405	1A	S1A_5190318	SRC-L, GPI	1.14	Gamma gliadin-A3
TraesCS1A02G007400	1A	S1A_5190318	SRC-L, GPI	1.15	Gamma gliadin-A3
TraesCS1A02G007300	1A	S1A_5190318	SRC-L, GPI	1.34	Gliadin
TraesCS1A02G007700	1A	S1A_5190318	SRC-L, GPI	1.13	Gliadin
TraesCS1B02G010600	1B	S1B_3282665, S1B_3767944, S1B_6922434, S1B_7235112, S1B_8678756	All SRCs	1.72	Delta gliadin-B1
TraesCS1B02G010500	1B	S1B_3282665, S1B_3767944, S1B_6922434, S1B_7235112, S1B_8678756	All SRCs	1.7	Delta gliadin-B1
TraesCS1B02G010400	1B	S1B_3282665, S1B_3767944, S1B_6922434, S1B_7235112, S1B_8678756	All SRCs	1.68	Gamma-gliadin 3
TraesCS1B02G011000	1B	S1B_3282665, S1B_3767944, S1B_6922434, S1B_7235112, S1B_8678756	All SRCs	1.86	Gamma-gliadin B
TraesCS1B02G011471	1B	S1B_3282665, S1B_3767944, S1B_6922434, S1B_7235112, S1B_8678756	All SRCs	2.09	Gliadin
TraesCS1B02G011300	1B	S1B_3282665, S1B_3767944, S1B_6922434, 1B_7235112, S1B_8678756	All SRCs	1.89	Gliadin
TraesCS1B02G009877	1B	S1B_3282665, S1B_3767944, S1B_6922434, S1B_7235112, S1B_8678756	All SRCs	1.56	Gliadin
TRITD1Av1G002790	1A	S1A_5190318	SRC-L/GPI	0.96	LMW-GS
TRITD1Av1G002360	1A	S1A_5190318	SRC-L/GPI	0.14	LMW-GS
TRITD1Av1G002310	1A	S1A_5190318	SRC-L/GPI	0.24	LMW-GS
TraesCS1A02G196100	1A	S1A_353657879	SRC-S	0.64	Glycosyltransferase2
TraesCS1B02G023600	1B	S1B_7235112, 1B_8678756,	SRC-C	3.73, 2.29	Glycosyltransferase
TraesCS1B02G023300		S1B_6922434, S1B7235112, S1B_8678756	SRC-W, SRC-S, SRC-C	3.62,3.31,2.29	Glycosyltransferase
TraesCS1B02G083100	1B	S1B_66284969	GPI	0.73	Glycosyltransferase
TraesCS1B02G023700	1B	S1B_7235112, S1B_8678756	SRC-C	3.89, 2.45	Glycosyltransferase

TraesCS1B02G080400	1B	S1B_66284969	GPI	3.56	Glycosyltransferase
TraesCS1B02G002100	1B	S1B_3282665	SRC-W	1.59	Glycosyltransferase
TraesCS2B02G312100	2B	S2B_444557348	SRC-L	2.81	Glycosyltransferase
TraesCS3B02G313900	3B	S3B_576577312	SRC-L	2.23	Glycosyltransferase
TraesCS3B02G465600	3B	S3B_576577312	SRC-L	3.46	Glycosyltransferase
TraesCS3B02G144500	3B	S3B_576577312	SRC-L	2.73	Glycosyltransferase
TraesCS3B02G144900	3B	S3B_576577312	SRC-L	2.71	Glycosyltransferase
TraesCS3B02G035000	3B	S3B_576577312	SRC-L	1.89	Glycosyltransferase
TraesCS3B02G143100	3B	S3B_576577312	SRC-L	2.05	Glycosyltransferase
TraesCS3B02G143300	3B	S3B_576577312	SRC-L	2.19	Glycosyltransferase
TraesCS4D02G307700	4D	S4D_472908721	SRC-W	2.96	Glycosyltransferase
TraesCS4D02G306800	4D	S4D_472908721	SRC-W	2.01	Glycosyltransferase
TraesCS4D02G307000	4D	S4D_472908721	SRC-W	2.04	Glycosyltransferase
TraesCS4D02G306900	4D	S4D_472908721	SRC-W	2.03	Glycosyltransferase
TraesCS4D02G307300	4D	S4D_472908721	SRC-W	2.21	Xyloglucan
TraesCS6A02G018200	6A	S6A_6869286	SRC-L	1.99	Glycosyltransferase
TraesCS6A02G017900	6A	S6A_6869286	SRC-L	1.88	Glycosyltransferase
TraesCS6A02G018000	6A	S6A_6869286	SRC-L	1.92	Glycosyltransferase
TraesCS6B02G144400	6B	S6B_147637875	SRC-L, SRC-S	3.85	Glycosyltransferase
TraesCS7A02G057400	7A	S7A_31213757	SRC-S	3.6	Glycosyltransferase
TraesCS7A02G002600	7A	S7A_3966595	SRC-W	2.44	Glycosyltransferase
TraesCS7A02G015900	7A	S7A_3966595	SRC-W	2.83	Glycosyltransferase
TraesCS1B02G368500	1B	S1B_595397803	GPI	3.48	Starch synthase, amyloplastic
TraesCS7A02G070100	7A	S7A_37433960	SRC-S	1.66	Starch synthase, amyloplastic

Abbreviations: Mbp, Mega base pair; SRC, solvent retention capacity; SRC-W, using water as a solvent; SRC-S, using sucrose as a solvent; SRC-L, using lactic acid as a solvent; SRC-C, using sodium carbonate as a solvent; GPI, gluten performance index; chrom, chromosome, LMW-GS, low molecular weight glutenin subunit.

Chapter 2 Figures

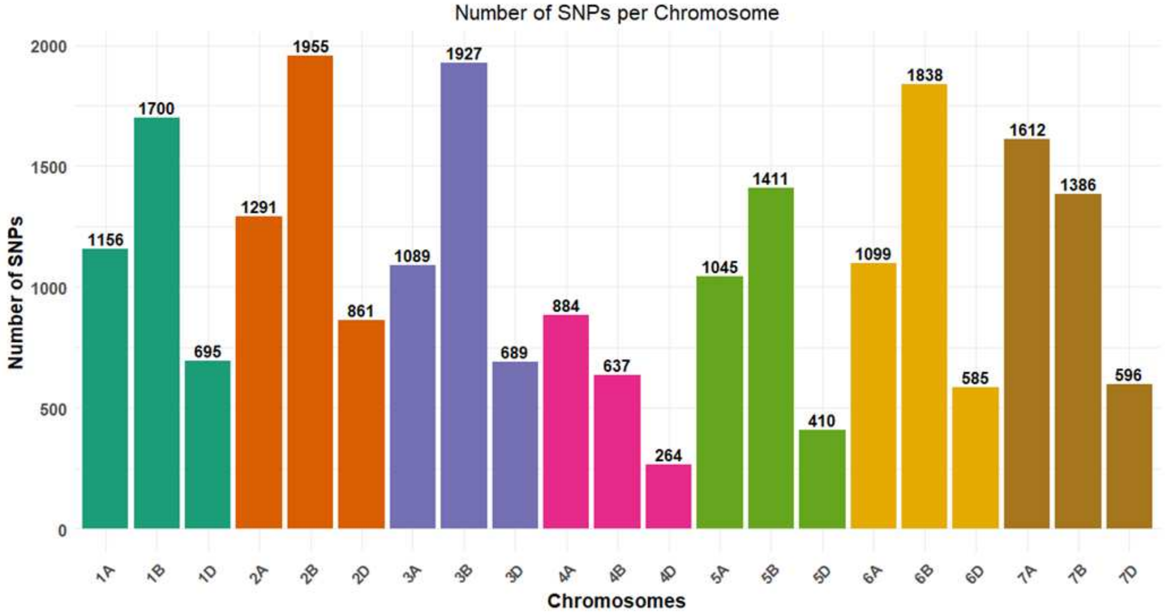


Figure 2.1 Distribution of Single Nucleotide Polymorphisms (SNPs) used in the current study across Chromosomes 1A to 7D.

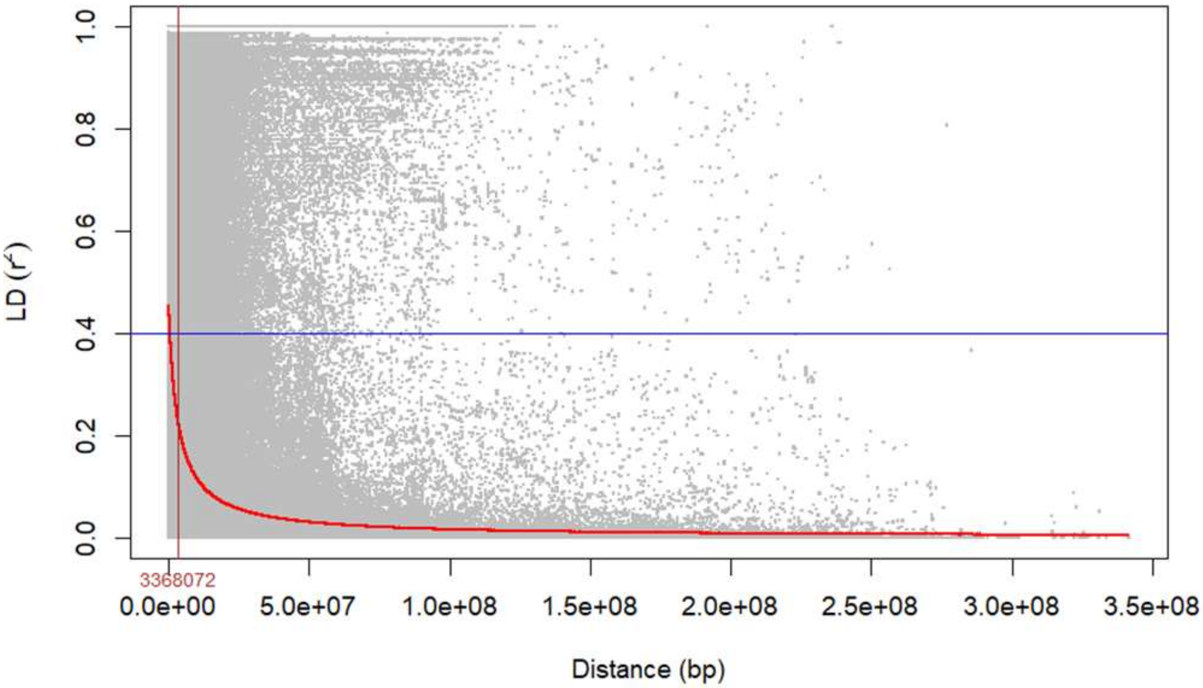


Figure 2.2 Scatterplot depicting Linkage Disequilibrium (LD) decay among 337 hard winter wheat genotypes. The plot shows the squared correlation coefficient (r^2) against the genetic distance (in base pairs). The green line marks the threshold where LD drops to 50% of its maximum value. The LD decay value at this cutoff point is highlighted on the x-axis in green. This plot provides insights into the rate of LD decay, a crucial factor for genome-wide association studies in this population.

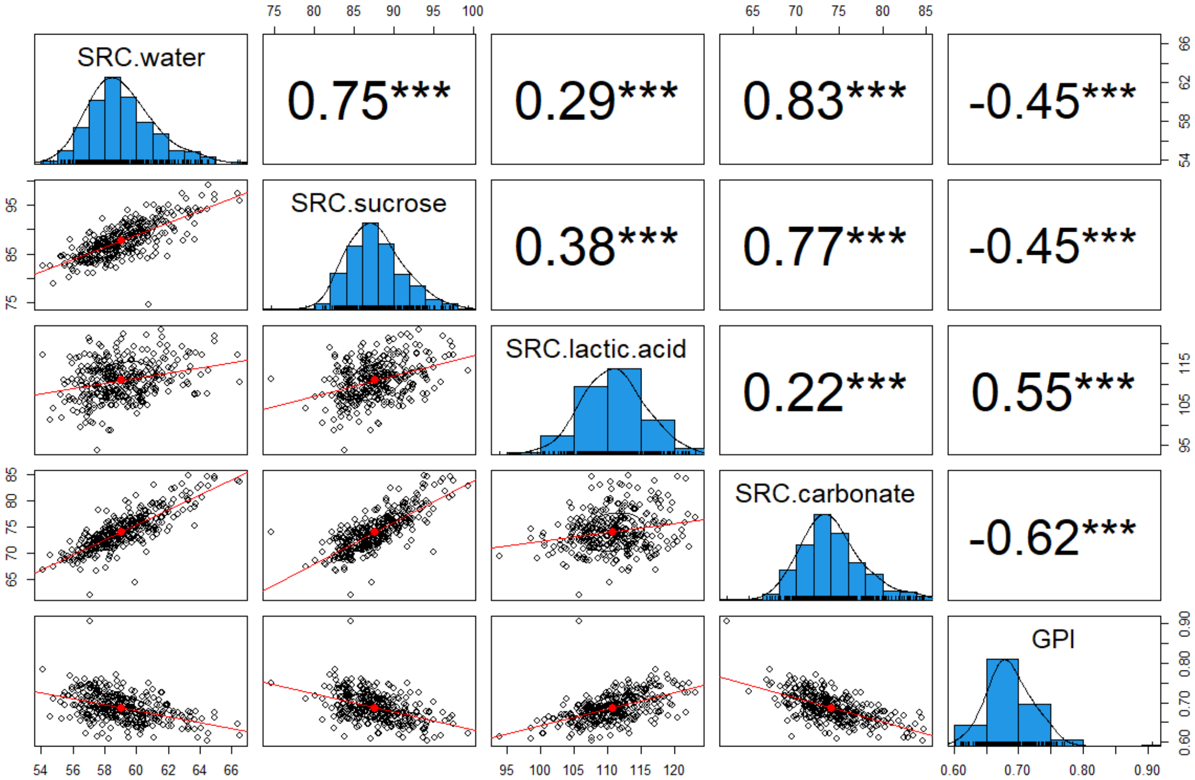


Figure 2.3 Pairwise correlation coefficients among traits studied in this study (SRCs and GPI) using best linear unbiased estimates (BLUEs) from across year-location analysis. The diagonal of the pair plot displays the frequency distribution for SRCs and GPI. The upper triangle shows Pearson's correlation coefficient (R) and the significance of the correlation for corresponding traits where *** Significant at the 0.001 probability level ($p \leq 0.001$). The lower triangle displays bivariate scatterplots with fitted lines.

Abbreviations: SRC, Solvent retention capacity, GPI, gluten performance index.

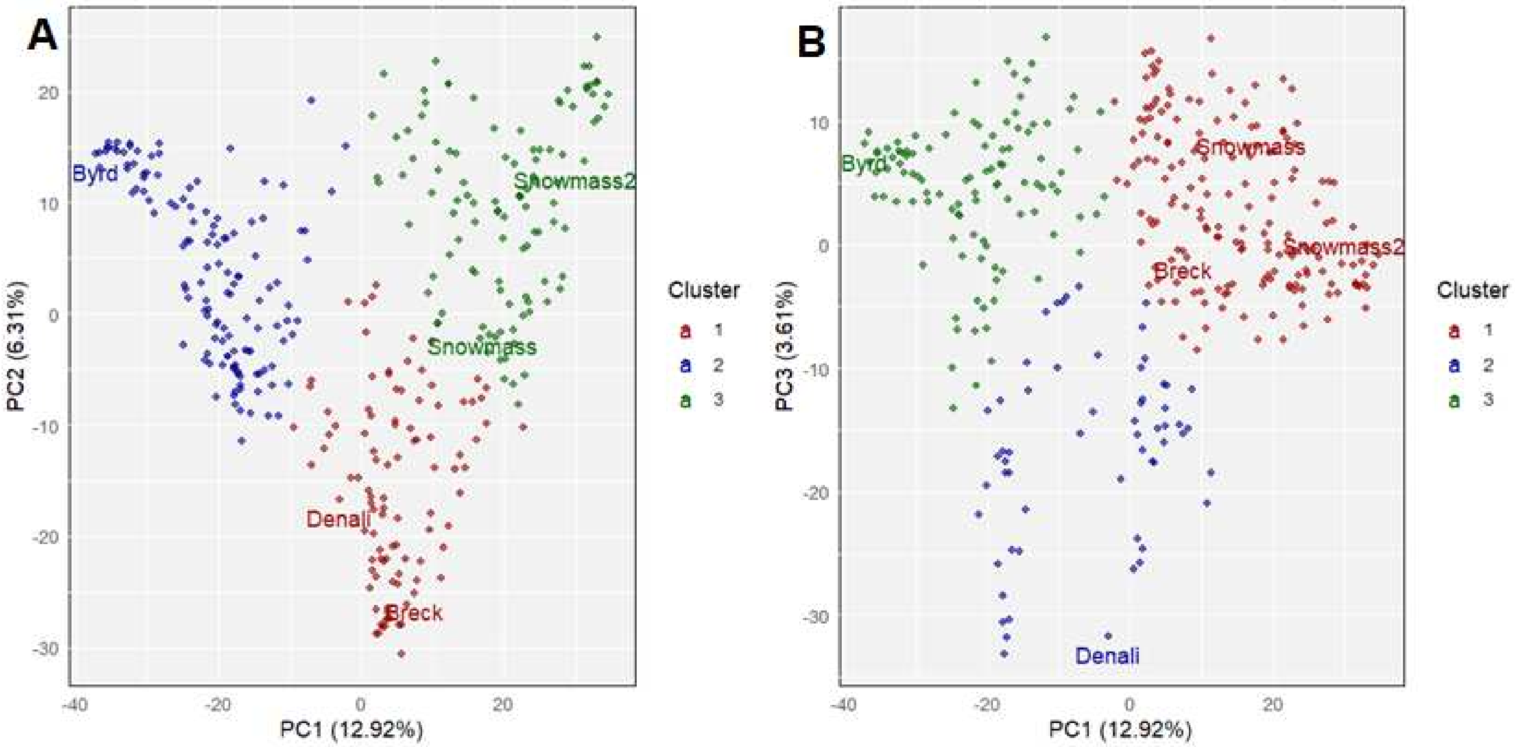


Figure 2.4 Scatter plot representing a) the first and second (PC1 and PC2) and b) the first and third (PC1 and PC3) principal components. The variance explained by each principal component is indicated in parentheses on each axis. Unique colors represent individual clusters.

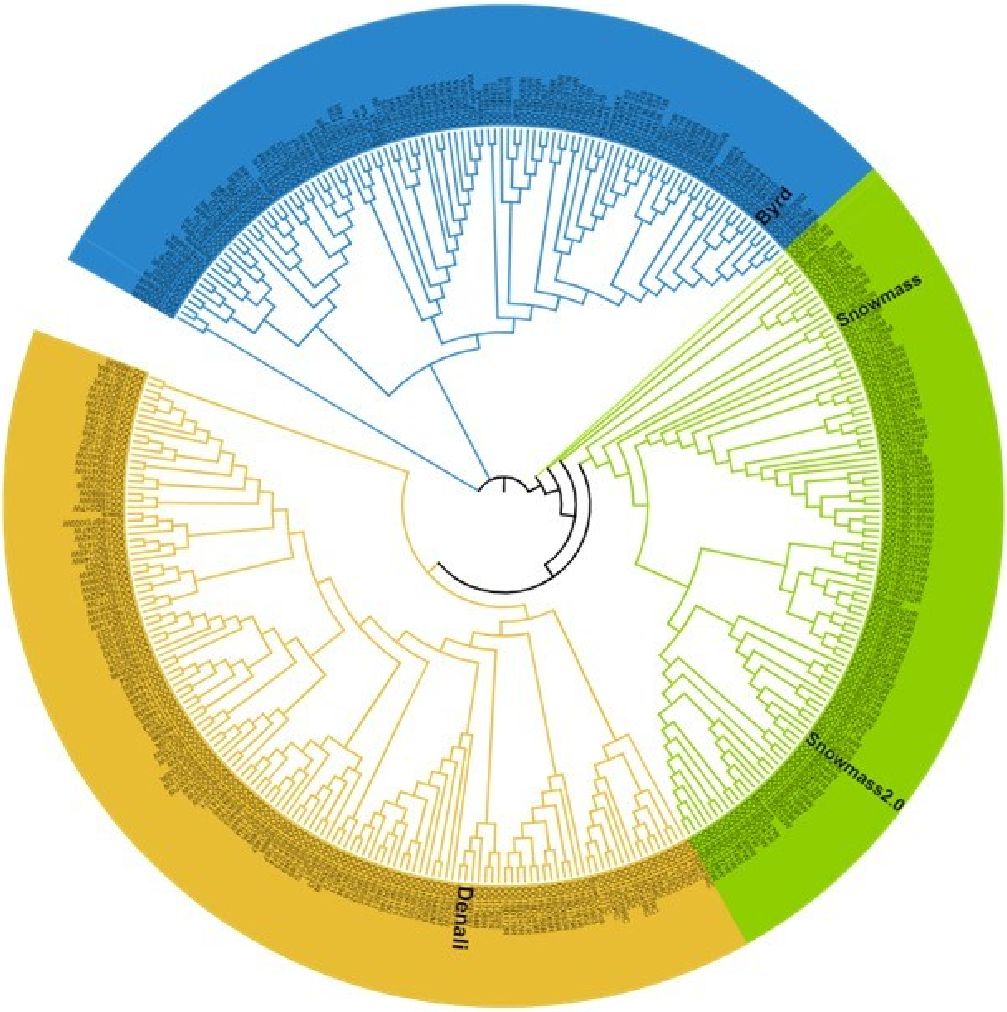


Figure 2.5 The Neighbor-Joining phylogenetic tree visualized using the Interactive *Tree of Life (iTOL)* software. The tree was derived from the analysis of dataset X and shows the genetic relationships among the studied taxa. Each branch represents a taxon, and the length of the branches indicates the genetic distance between them. Branch points, or nodes, represent hypothetical common ancestors.

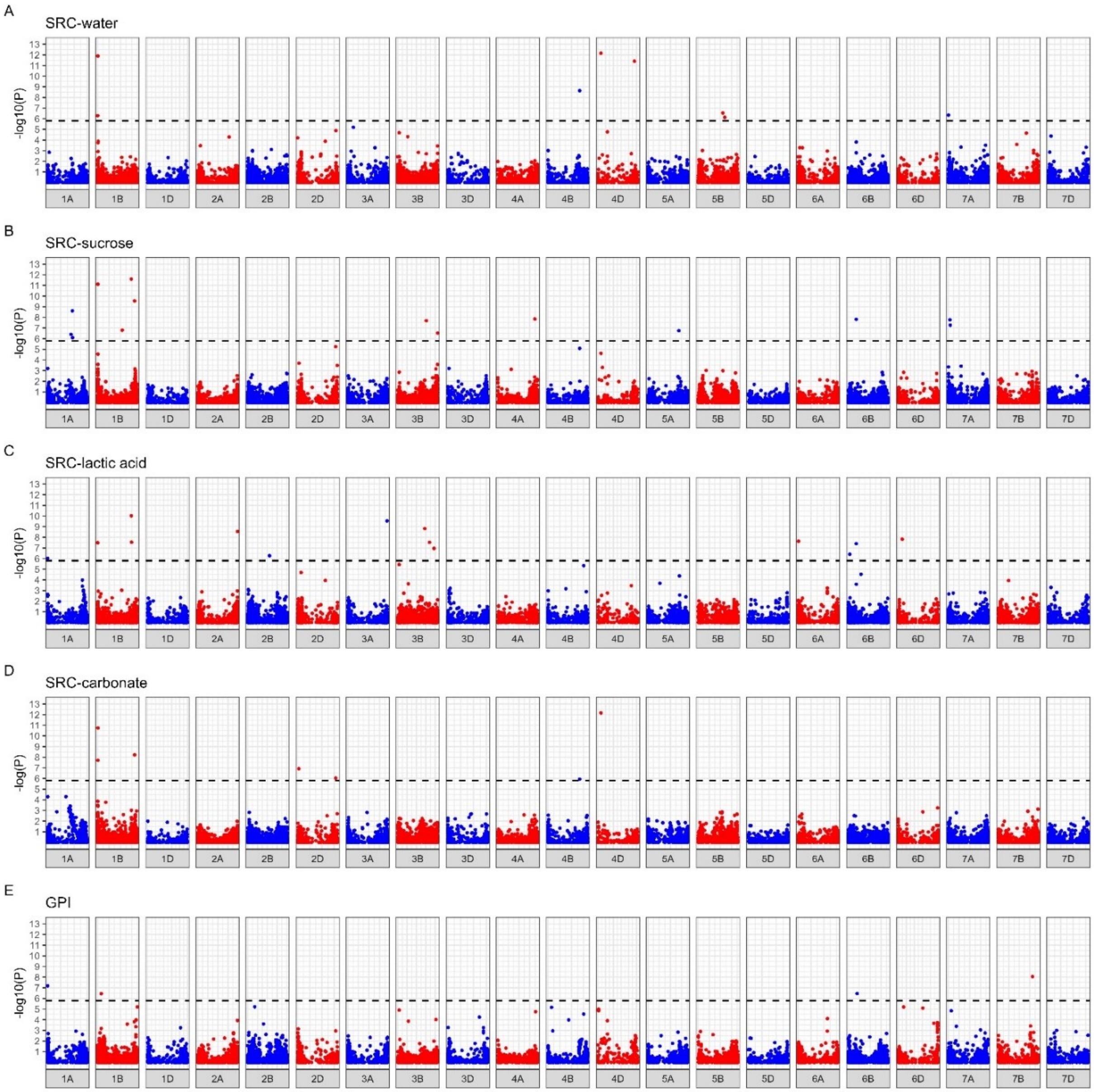


Figure 2.6 Manhattan plot illustrating -log$_{10}$(p) values from a genome-wide association study (GWAS) of solvent retention capacity (SRC) traits, including SRC-water, SRC-sucrose, SRC-lactic acid, SRC, carbonate, and gluten performance index (GPI) displayed with its -log$_{10}$(p) value on the y-axis and its chromosomal position on the x-axis. The dashed horizontal line marks the genome-wide significance threshold of 5.8 (-log10(p). Points above this line highlight SNPs showing significant association with the traits under study.

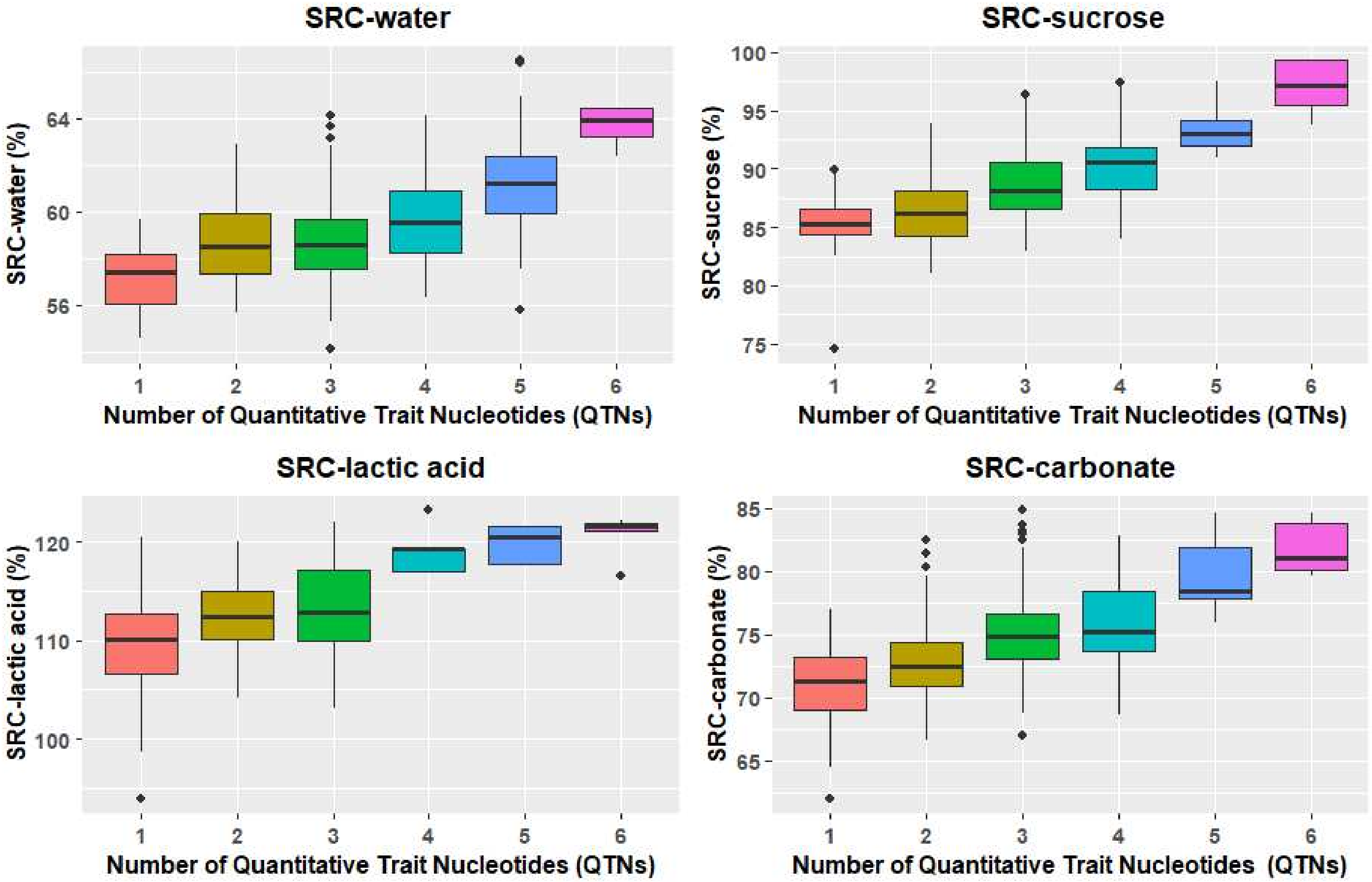
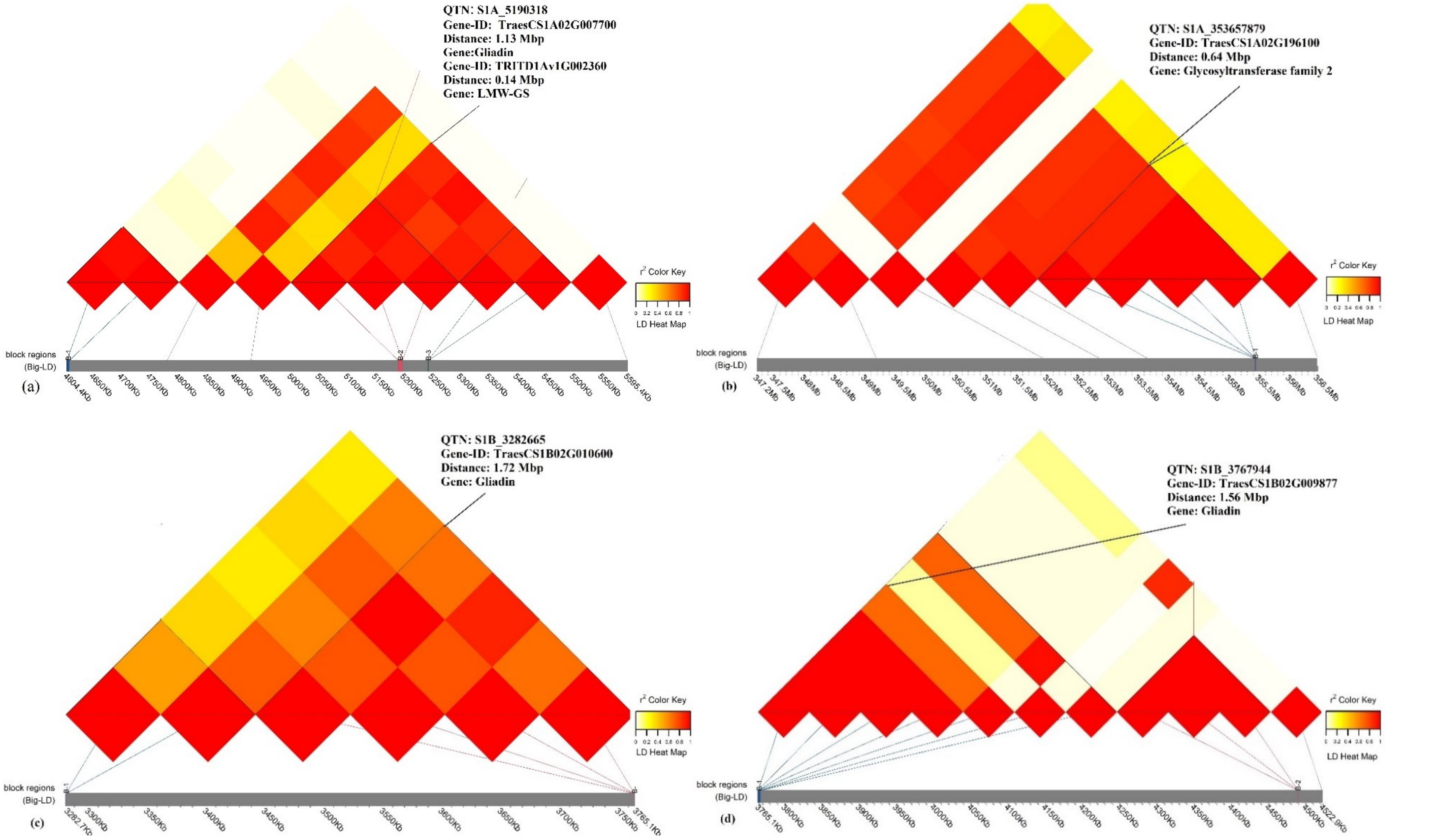


Figure 2.7 The cumulative effects of the top six significant markers associated with each SRC trait. Each box represents the interquartile range of the distribution across locations, with the best linear unbiased estimates, with the median indicated by a horizontal line within the box. Data points outside the box are considered outliers.
Abbreviations: SRC, solvent retention capacity.


QTN: S1A_5190318
Gene-ID: TraesCS1A02G007700
Distance: 1.13 Mbp
Gene:Gliadin
Gene-ID: TRITD1Av1G002360
Distance: 0.14 Mbp
Gene: LMW-GS
r2 Color Key
0 0.2 0.4 0.6 0.8 1
LD Heat Map
block regions (Big-LD)
B-1
B-2
B-3
(a)
4604.4Kb 4650Kb 4700Kb 4750Kb 4800Kb 4850Kb 4900Kb 4950Kb 5000Kb 5050Kb 5100Kb 5150Kb 5200Kb 5250Kb 5300Kb 5350Kb 5400Kb 5450Kb 5500Kb 5550Kb 5595.4Kb
QTN: S1A_353657879
Gene-ID: TraesCS1A02G196100
Distance: 0.64 Mbp
Gene: Glycosyltransferase family 2
r2 Color Key
0 0.2 0.4 0.6 0.8 1
LD Heat Map
block regions (Big-LD)
B-1
(b)
347.2Mb 347.5Mb 348Mb 348.5Mb 349Mb 349.5Mb 350Mb 350.5Mb 351Mb 351.5Mb 352Mb 352.5Mb 353Mb 353.5Mb 354Mb 354.5Mb 355Mb 355.5Mb 356Mb 356.5Mb
QTN: S1B_3282665
Gene-ID: TraesCS1B02G010600
Distance: 1.72 Mbp
Gene: Gliadin
r2 Color Key
0 0.2 0.4 0.6 0.8 1
LD Heat Map
block regions (Big-LD)
B-1
B
(c)
3282.7Kb 3300Kb 3350Kb 3400Kb 3450Kb 3500Kb 3550Kb 3600Kb 3650Kb 3700Kb 3750Kb 3765.1Kb
QTN: S1B_3767944
Gene-ID: TraesCS1B02G009877
Distance: 1.56 Mbp
Gene: Gliadin
r2 Color Key
0 0.2 0.4 0.6 0.8 1
LD Heat Map
block regions (Big-LD)
B-1
B-2
(d)
3765.1Kb 3800Kb 3850Kb 3900Kb 3950Kb 4000Kb 4050Kb 4100Kb 4150Kb 4200Kb 4250Kb 4300Kb 4350Kb 4400Kb 4450Kb 4500Kb 4522.9Kb

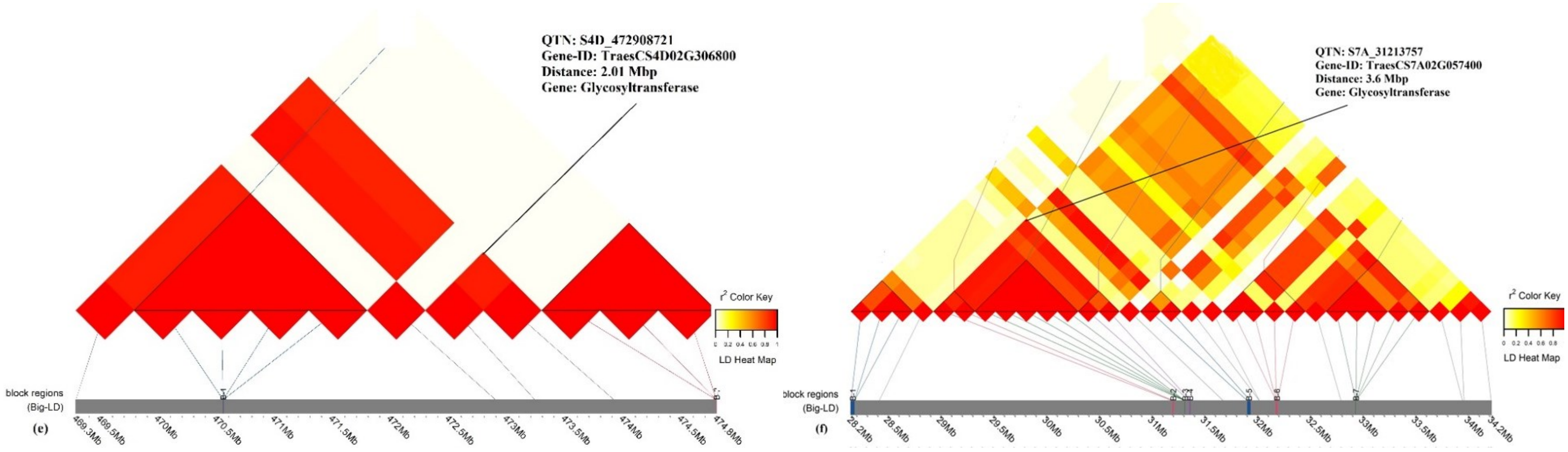


Figure 2.8 The LD (linkage disequilibrium) heatmap visualizes the strength of association between alleles at different loci on different wheat chromosomes. The color gradient, from red to yellow, indicates high to moderate linkage disequilibrium.

Chapter 2 References

AACC International. (2010). Approved Methods of Analysis, 11th Ed. Methods 10-53.01, 56-10.02, and 56-11.02. *Cereals & Grains Association*, St. Paul, MN, U.S.A. https://www.cerealsgrains.org/resources/Methods/Pages/54PhysicalDoughTests.aspx

Alexander, D. H., and Lange, K. (2011). Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics*, 12, 246. https://doi.org/10.1186/1471-2105-12-246

Alsamman, A. M., Bousba, R., Baum, M., Hamwieh, A., and Fouad, N. (2021). Comprehensive analysis of the gene expression profile of wheat at the crossroads of heat, drought, and combined stress. *Highlights Biosci*, *4*. https://dog.org/10.36462/H.BioSci.202104

Aoun, M., Carter, A. H., Morris, C. F., and Kiszonas, A. M. (2022). Genetic architecture of end-use quality traits in soft white winter wheat. *BMC Genomics*, *23*(1), 1-17. https://doi.org/10.1186/s12864-022-08676-5

Appels, R., Eversole, K., Feuillet, C., Keller, B., Rogers, J., Stein, N., . . . Manuscript Writing, T. (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science*, *361*(6403), 661-+, Article eaar7191. https://doi.org/10.1126/science.aar7191

Bernardo, R. (2014). Genome-wide selection occurs when major genes are known. *Crop Science*, *54*(1), 68-75. https://doi.org/10.2135/cropsci2013.05.0315

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society*: *Series B (Methodological),* 57(1), 289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Brachi, B., Morris, G.P., and Borevitz, J.O. (2011). Genome-wide association studies in plants: the missing heritability is in the field. *Genome Biology*, 12(10), 232. https://doi.org/10.1186/gb-2011-12-10-232

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

Branlard, G., Dardevet, M., Saccomano, R., Lagoutte, F., and Gourdon, J. (2001). Genetic diversity of wheat storage proteins and bread wheat quality. *Euphytica*, *119*(1-2), 59-67. https://doi.org/10.1023/A:1017586220359

Browning, B. L., Zhou, Y., and Browning, S. R. (2018). A one-penny imputed genome from next-generation reference panels. *The American Journal of Human Genetics*, *103*(3), 338–348.

Bushuk, W., and Bekes, F. (2002). Contribution of protein to flour quality. In A. Salgó, S. Tömösközi, and R.Lásztity (Eds.), Proceedings of the ICC: Novel Raw Materials, Technologies and Products – NewChallenge for Quality Control (pp. 14-19). Conference held in Budapest, Hungary, 26-29 May.https://doi.org/10.1016/j.foodres.2006.05.003

Butler, D., Cullis, B., Gilmour, A., and Thompson, R. (2007). Asreml-R: an R package for mixed models using residual maximum likelihood.

Cabrera, A., Guttieri, M., Smith, N., Souza, E., Sturbaum, A., Hua, D., O'Brien, K., Farmer, J., Sneller, C., Giroux, M., Griffey, C., Matnyazov, R., Koo, J.M., Taylor, J., Ohm, H., Patterson, F., and Udall, J. (2015). Identification of milling and baking quality QTL in multiple soft wheat mapping populations. *Theoretical and Applied Genetics*, 128, 2227-2242. https://doi:10.1007/s00122-015-2580-3

Campbell, K. G., Finney, P. L., Bergman, C. J., Gualberto, D. G., Anderson, J. A., Giroux, M. J., ... and Sorrells, M. E. (2001). Quantitative trait loci associated with milling and baking quality in a soft× hard wheat cross. *Crop Science*, *41*(4),1275-1285. https://doi.org/10.2135/cropsci2001.4141275x

Charrad, M., Ghazzali, N., Boiteau, V., and Niknafs, A. (2014). NbClust: an R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, *61*, 1-36. https://doi:10.18637/jss.v061.i06

Colombo, A., Pérez, G.T., Ribotta, P.D., and León, A.E. (2008). A comparative study of physicochemical tests for quality prediction of Argentine wheat flours used as correctors. *Journal of Cereal Science*, 48(3), 775-780. https://doi.org/10.1016/j.jcs.2008.05.003

Dubat, A., Berra, M., and Baik, B. K. (2019). Collaborative study report: automated measurement of wheat flour solvent retention capacity with the CHOPIN-SRC instrument (AACCI approved method 56-5.01). *Cereal Foods World*, 64(3). https://doi.org/10.1094/CFW-64-3-0033

Duyvejonck, A.E., Lagrain, B., Dornez, E., Courtin, C.M., and Delcour, J.A. (2011). Suitability of European wheat cultivars for dough liquor production. *Journal of Cereal Science*, 53(3), 345-348.

Fox, G. P., Martin, A., Kelly, A. M., Sutherland, M. W., Martin, D., Banks, P. M., and Sheppard, J. (2013). QTLs for water absorption and flour yield identified in the doubled haploid wheat population Lang/QT8766. *Euphytica*, *192*, 453-462. https://doi.org/10.1007/s10681-013-0885-3

Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G., and François, O. (2014). Fast and efficient estimation of individual ancestry coefficients. *Genetics*, 196(4), 973-983.

Gaines, C.S. (2000). Collaborative study of methods for solvent retention capacity profiles (AACC Method 56-11). *Cereal Foods World*, 45(1), 7-8.

Gaire, R., Huang, M., Sneller, C., Griffey, C., Brown-Guedira, G., and Mohammadi, M. (2019).Association analysis of baking and milling quality traits in an elite soft red winter wheat population. *Crop Science*, *59*(3), 1085-1094. https://doi.org/10.2135/cropsci2018.12.0751

Garcia-Santamaria, G., Hua, D., and Sneller, C. (2018). Quantitative trait loci associated with soft wheat quality in a cross of good by moderate quality parents. *PeerJ*, *6*, e4498. https://doi.org/10.7717/peerj.4498

Glaubitz, J. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q., and Buckler, E. S. (2014). *TASSEL-GBS*: a high-capacity genotyping by sequencing analysis pipeline. *PloS one*, *9*(2), e90346. https://doi.org/10.1371/journal.pone.0090346

Goel, S., Singh, K., Singh, B., Grewal, S., Dwivedi, N., Alqarawi, A. A., ... and Singh, N. K. (2019). Analysis of genetic control and QTL mapping of essential wheat grain quality traits in a recombinant inbred. *PLoS One*, *14*(3), e0200669. https://doi.org/10.1371/journal.pone.0200669

Guttieri, M.J., Bowen, D., Gannon, D., O'Brien, K., and Souza, E. (2001). Solvent retention capacities of irrigated soft white spring wheat flour. *Crop Science*, 41(4), 1054-1061. https://doi.org/10.2135/cropsci2001.4141054x

Guzmán, C., Posadas-Romano, G., Hernández-Espinosa, N., Morales-Dorantes, A., and Peña, R.J. (2015). A new standard water absorption criterion based on solvent retention capacity (SRC) to determine dough mixing properties, viscoelasticity, and bread-making quality. *Journal of Cereal Science*, 66, 59-65. https://doi.org/10.1016/j.jcs.2015.10.009

Haley, S.D., J.J. Johnson, F.B. Peairs, J.A. Stromberger, E.E. Heaton, S.A. Seifert, R.A. Kottke, J.B. Rudolph, G. Bai, R.L. Bowden, M.-S. Chen, X. Chen, Y. Jin, J.A. Kolmer, R. Chen, and B.W. Seabourn. 2011. Registration of 'Snowmass' wheat. *J. Plant* Reg. 5:1-4.

Haley, S.D., J.J. Johnson, F.B. Peairs, J.A. Stromberger, E.E. Hudson, S.A. Seifert, R.A. Kottke, V.A. Valdez, J.B. Rudolph, G. Bai, X. Chen, R.L. Bowden, Y. Jin, J.A. Kolmer, M.-S. Chen, and B.W. Seabourn. 2012. Registration of 'Byrd' wheat. *J. Plant* Reg. 6:302-305.

Haley, S.D., J.J. Johnson, F.B. Peairs, J.A. Stromberger, E.E. Hudson, S.A. Seifert, R.A. Kottke, V.A. Valdez, J.B. Rudolph, G. Bai, X. Chen, R.L. Bowden, Y. Jin, J.A. Kolmer, M.-S. Chen, and B.W. Seabourn. 2012. Registration of 'Denali' wheat. *J. Plant* Reg. 6:311-314.

Henry, R. J. (1986). Genetic and environmental variation in the pentosan and β-glucan contents of barley, and their relation to malting quality. *Journal of Cereal Science*, *4*(3), 269-277. https://doi.org/10.1016/S0733-210(86)80029-7.

Ibba, M. I., Juliana, P., Hernández-Espinosa, N., Posadas-Romano, G., Dreisigacker, S., Sehgal, D., ... and Guzmán, C. (2021). Genome-wide association analysis for arabinoxylan content in common wheat (T. Aestivum L.) flour. *Journal of Cereal Science*, *98*, 103166. https://doi.org/10.1016/j.jcs.2021.103166

Izydorczyk, M. S. (2021). Arabinoxylans. In *Handbook of hydrocolloids* (pp. 399-461). Woodhead Publishing. https://doi.org/10.1016/B978-0-12-820104-6.00016-4

James, M. G., Denyer, K., and Myers, A. M. (2003). Starch synthesis in the cereal endosperm. *Current opinion in plant biology*, *6*(3), 215-222. https://doi.org/10.1016/S1369-5266(03)00042-6

Jelaca, S., and Hlynka, I. (1971). Water-binding capacity of wheat flour crude pentosans and their relation to mixing characteristics of dough. *Cereal Chemistry*. 48, 211-222

Jiang, P., Zhang, P.P., Zhang, X., and Ma, H.X. (2017). Genetic diversity and association analysis for solvent retention capacity in the accessions derived from soft wheat Ningmai 9. *International Journal of Genomics*, 2017. https://doi.org/10.1155/2017/2413150

John, J. A., and Williams, E. R. (1995). Resolvable row-column designs. In *Cyclic and Computer-Generated Designs* (pp. 107-129). Springer US.

Kerfal, S., Giraldo, P., Rodríguez-Quijano, M., Vázquez, J. F., Adams, K., Lukow, O. M., ... and Carrillo, J. M. (2010). Mapping quantitative trait loci (QTLs) associated with dough quality in a soft× hard bread wheat progeny. *Journal of Cereal Science*, *52*(1), 46-52. https://doi.org/10.1016/j.jcs.2010.03.001

Kim, S. A., Brossard, M., Roshandel, D., Paterson, A. D., Bull, S. B., and Yoo, Y. J. (2019). gpart: human genome partitioning and visualization of high-density SNP data by identifying haplotype blocks. *Bioinformatics*, *35*(21), 4419-4421. https://doi.org/10.1093/bioinformatics/btz308

Kosik, O., Powers, S. J., Chatzifragkou, A., Prabhakumari, P. C., Charalampopoulos, D., Hess, L., ... and Lovegrove, A. (2017). Changes in the arabinoxylan fraction of wheat grain during alcohol production. *Food Chemistry*, *221*, 1754-1762. https://doi.org/10.1016/j.foodchem.2016.10.109

Kweon, M., Slade, L., and Levine, H. (2011). Solvent retention capacity (SRC) testing of wheat flour: Principles and value in predicting flour functionality in different wheat-based food processes and in wheat breeding A review. *Cereal Chemistry*, 88(6), 537-552. https://doi.org/10.1094/CCHEM-07-11-0092

Labuschagne, M.T., Botes, S., Bonthuys, B. (2021). Genome-wide association mapping of wheat solvent retention capacity and glutenin macropolymer traits in a South African spring wheat diversity panel. *Frontiers in Plant Science*, 12, 779406. https://doi.org/10.3390/plants10051000

Letunic, I., and Bork, P. (2016). Interactive Tree of Life (iTOL) v3: an online tool for the display and annotation of phylogenetics and other trees. *Nucleic Acids Research*, 44(W1), W242-W245. https://doi.10.1093/nar/gkw290

Li, J., Xie, L., Tian, X., Liu, S., Xu, D., Jin, H., Chen, L., Liang, Y., Xin, M., Wang, H., Li, X., Li, X., and Cao, S. (2021). TaNAC100 acts as an integrator of seed protein and starch synthesis, exerting pleiotropic effects on agronomic traits in wheat. *The Plant Journal*, 108(3), 829-840. https://doi.10.1111/tpj.15485

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*(14), 1754-1760. https://doi.org/10.1093/bioinformatics/btp324

Lou, H., Zhang, R., Liu, Y., Guo, D., Zhai, S., Chen, A., ... and Li, B. (2021). Genome-wide association study of six quality-related traits in common wheat (Triticum aestivum L.) under two sowing conditions. *Theoretical and Applied Genetics*, *134*, 399-418. https://doi.org/10.1007/s00122-020-03704-y

Lovegrove, A., Wilkinson, M. D., Freeman, J., Pellny, T. K., Tosi, P., Saulnier, L., ... and Mitchell, R. A. (2013). RNA interference suppression of genes in glycosyl transferase families 43 and 47 in wheat starchy endosperm causes large decreases in arabinoxylan content. *Plant Physiology*, *163*(1), 95-107. https://doi.org/10.1104/pp.113.222653

Ma, W., Sutherland, M. W., Kammholz, S., Banks, P., Brennan, P., Bovill, W., and Daggard, G. (2007). Wheat flour protein content and water absorption analysis in a doubled haploid population. *Journal of Cereal Science*, *45*(3), 302-308. https://doi.org/10.1016/j.jcs.2006.10.005

Mallory, E., Bramble, T., Williams, M., and Amaral, J. (2012). Understanding wheat quality- What bakers and millers need and what farmers can do. *University of Maine Cooperative Extension Publication. Available at the Web site.* http://umaine. edu/publications/1019e

McCartney, C. A., Somers, D. J., Lukow, O., Ames, N., Noll, J., Cloutier, S., ... and McCallum, B. D. (2006) QTL analysis of quality traits in the spring wheat cross RL4452×'AC Domain'. *Plant Breeding*, *125*(6), 565-575. https://doi.org/10.1111/j.1439-0523.2006.01256.x

Morgan, B., Dexter, J., and Preston, K. (2000). Relationship of kernel size to flour water absorption for Canada western red spring wheat. *Cereal Chemistry*, 77(3), 286-292. https://doi.org/10.1094/CCHEM.2000.77.3.286

Navrotskyi, S., Belamkar, V., Baenziger, P. S., and Rose, D. J. (2020). Insights into the genetic architecture of bran friability and water retention capacity are two important traits for whole grain end-use quality in winter wheat. *Genes*, *11*(8), 838. https://doi.org/10.3390/genes11080838

Pandey, A., Khobra, R., Mamrutha, H. M., Wadhwa, Z., Krishnappa, G., Singh, G., and Singh, G. P. (2022). Elucidating the drought responsiveness in wheat genotypes. *Sustainability*, *14*(7), 3957. https://doi.org/10.3390/su14073957

Pasha, I., Anjum, F., and Morris, C. (2010). Grain hardness: a major determinant of wheat quality. *Food Science and Technology International*, 16(6), 511-522. https://doi:10.1177/1082013210379691

Payne, P. I., Holt, L. M., Jackson, E. A., and Law, C. N. (1984). Wheat storage proteins: their genetics and their potential for manipulation by plant breeding. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, *304*(1120), 359-371.

Payne, P.I., Law, C.N., and Mudd, E.E. (1980). Control by homoeologous group 1 chromosomes of the high-molecular-weight subunits of glutenin, a major protein of wheat endosperm. *Theoretical and Applied Genetics,* 58, 113-120. https://doi:10.1007/BF00263101

Poland, J., Endelman, J., Dawson, J., Rutkoski, J., Wu, S., Manes, Y., Dreisigacker, S., Crossa, J., Sanchez Villeda, H., Sorrells, M., and Jannink, J.L. (2012). Genomic selection in wheat breeding using genotyping-by-sequencing. *The Plant Genome 5(3)*, https://doi.org/10.3835/plantgenome2012.06.0006

Preston, K., Hucl, P., Townley-Smith, T., Dexter, J., Williams, P., and Stevenson, S. (2001). Effects of cultivar and environment on Farinograph and Canadian short process mixing properties of Canada Western Red Spring wheat. *Canadian Journal of Plant Science*, 81(3), 391-398. https://doi.org/10.4141/P00-137

Primo-Martin, C., Valera, R., and Martinez-Anaya, M.A. (2003). Effect of pentosanase and oxidases on the characteristics of doughs and the glutenin macropolymer (GMP). *Journal of Agricultural and Food Chemistry*, 51(16), 4673-4679. https://doi.org/10.1021/jf0257695

Pu, Z., Ye, X., Li, Y., Liu, Z., Shi, B., Dai, S., ... and Zheng, Y. (2020). Genome-Wide association study identifies new elements on the genetic basis of quality-related traits in wheat across multiple environments. *Research Square*. https://doi.org/10.21203/rs.3.rs-99775/v1

Puhr, D., and D'Appolonia, B. (1992). Effect of baking absorption on bread yield, crumb moisture, and crumb water activity. *Cereal Chemistry*, 69, 582-582.

Pyler, E. (1979). Physical and chemical test methods. *Baking Science and Technology*, 2, 891-895.

R Core Team, R. (2013). R: A language and environment for statistical computing.

Rakszegi, M., Lovegrove, A., Balla, K., Láng, L., Bedő, Z., Veisz, O., and Shewry, P. R. (2014). Effect of heat and drought stress on the structure and composition of arabinoxylan and β-glucan in wheat grain. *Carbohydrate Polymers*, 102, 557-565. https://doi.org/10.1016/j.carbpol.2013.12.005

Ram, S., Dawar, V., Singh, R. P., and Shoran, J. (2005). Application of solvent retention capacity tests for the prediction of mixing properties of wheat flour. *Journal of Cereal Science*, 42(2), 261-266. https://doi.org/10.1016/j.jcs.2005.04.005

Remington, D. L., Thornsberry, J. M., Matsuoka, Y., Wilson, L. M., Whitt, S. R., Doebley, J., ... and Buckler IV, E. S. (2001). Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proceedings of the National Academy of Sciences*, *98*(20), 11479-11484.

Ren, W.-L., Wen, Y.-J., Dunwell, J. M., and Zhang, Y.-M. (2018). pKWmEB: integration of Kruskal–Wallis test with empirical Bayes under polygenic background control for multi-locus genome-wide association study. *Heredity*, 120(3), 208-218. https://doi.org/10.1038/s41437-017-0007-4

Revelle, W., and Revelle, M. W. (2015). Package 'psych.' *The comprehensive R archive network*, *337*(338). https://CRAN.R-project.org/package=psych

Ronga, D., Laviano, L., Catellani, M., Milc, J., Prandi, B., Boukid, F., ... and Francia, E. (2020). Influence of environmental and genetic factors on the content of toxic and immunogenic wheat gluten peptides. *European Journal of Agronomy*, *118*, 126091.https://doi.org/10.1016/j.eja.2020.126091

Sapirstein, H., Wu, Y., Koksel, F., and Graf, R. (2018). A study of factors influencing the water absorption capacity of Canadian hard red winter wheat flour. *Journal of Cereal Science*, *81*, 52-59. https://doi.org/10.1016/j.jcs.2018.01.012

Shewry, P. R., Halford, N. G., and Lafiandra, D. (2003). Genetics of wheat gluten proteins. *Advances in Genetics*, *49*, 111-184. https://doi.org/10.1016/S0065-2660(03)01003-4

Shewry, P. R., and Halford, N. G. (2002). Cereal seed storage proteins: structures, properties, and role in grain utilization. *Journal of Experimental Botany*, *53*(370), 947-958.

Smith, N., Guttieri, M., Souza, E., Shoots, J., Sorrells, M., and Sneller, C. (2011). Identification and validation of QTL for grain quality traits in a cross of soft wheat cultivars Pioneer Brand 25R26 and Foster. *Crop Science*, 51(4), 1424-1436. https://doi.org/10.2135/cropsci2010.04.0193

Tamba, C. L., and Zhang, Y.-M. (2018). A fast mrMLM algorithm for multi-locus genome-wide association studies. *Biorxiv,* 341784. https://doi.org/10.1101/341784

Tamba, C. L., Ni, Y.-L., and Zhang, Y.-M. (2017). Iterative sure independence screening EM-Bayesian LASSO algorithm for multi-locus genome-wide association studies. *PLoS Computational Biology*, 13(1), e1005357. https://doi.org/10.1371/journal.pcbi.1005357

Tipples, K.H., Meredith, J.O., and Holas, J. (1978). Factors affecting Farinograph and baking absorption. II.The relative influence of flour components. *Cereal Chemistry*, 55(5), 652-660.

Tremmel-Bede, K., Szentmiklóssy, M., Tömösközi, S., Török, K., Lovegrove, A., Shewry, P. R., and Rakszegi, M. (2020). Stability analysis of wheat lines with increased levels of arabinoxylan. *Plos one*, 15(5), e0232892. https://doi.org/10.1371/journal.pone.0232892

Tsilo, T.J., Nygard, G., Khan, K. et al. Molecular genetic mapping of QTL associated with flour water absorption and farinograph-related traits in bread wheat. *Euphytica* 194, 293–302 (2013). https://doi.org/10.1007/s10681-013-0906-2

Wan, L., Yuan, Z., Wu, B., Jia, H., Gao, Z., and Cao, F. (2022). Dissolution behavior of arabinoxylan from sugarcane bagasse in tetrabutylammonium hydroxide aqueous solution. *Carbohydrate Polymers*, *282*, 119037. https://doi.org/10.1016/j.carbpol.2021.119037

Wang, S.-B., Feng, J.-Y., Ren, W.-L., Huang, B., Zhou, L., Wen, Y.-J., . . . Zhang, Y.-M. (2016). Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. *Scientific Reports*, 6(1), 1-10. https://doi.org/10.1038/srep19444

Wen, Y. J., Zhang, H., Ni, Y. L., Huang, B., Zhang, J., Feng, J. Y., ... and Wu, R. (2018). Methodological implementation of mixed linear models in multi-locus genome-wide association studies. *Briefings in Bioinformatics*, *19*(4), 700-712. https://doi.org/10.1093/bib/bbw145

Williams, E., Piepho, H. P., and Whitaker, D. (2011). Augmented p-rep designs. *Biometrical Journal*, 53(1), 19-27. https://doi.org/10.1002/bimj.201000102

Xiao-ling, J., Hong-min, L., Yu-ling, A., Ji-shun, Z., Yan-yan, G., Xiao-jun, L., and Jian-sheng, C. (2023). Identification of solvent retention capacity quantitative trait loci by combined linkage and association mapping in wheat (Triticum aestivum L.). *Crop Science*, *63*(5), 2952-2962.

Zghal, M., Scanlon, M., and Sapirstein, H. (2001). Effects of flour strength, baking absorption, and processing conditions on the structure and mechanical properties of bread-crumbs. *Cereal Chemistry*, 78(1), 1-7. https://doi.org/10.1094/CCHEM.2001.78.1.1

Zhang, J., Feng, J.-Y., Ni, Y., Wen, Y., Niu, Y., Tamba, C., Zhang, Y. (2017). pLARmEB: integration of least angle regression with empirical Bayes for multi-locus genome-wide association studies. *Heredity*. https://doi.org/10.1038/hdy.2017.8

Zhang, Q., Zhang, Y., Zhang, Y., He, Z., and Peña, R. J. (2007). Effects of solvent retention capacities, pentosan content, and dough rheological properties on sugar snap cookie quality in Chinese soft wheat genotypes. *Crop Science*, 47(2), 656-662. https://doi.org/10.2135/cropsci2006.05.0357

Zhao, K., Tung, C. W., Eizenga, G. C., Wright, M. H., Ali, M. L., Price, A. H., ... and McCouch, S. R. (2011). Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa. *Nature Communications*, *2*(1), 467. https://doi.org/10.1038/ncomms1467

Chapter 3- Advancing Water Absorption Capacity in Hard Winter Wheat Using a Multivariate Genomic Prediction Approach

3.1 Summary

The water absorption capacity (WAC) of hard wheat flour affects end-use quality characteristics, including loaf volume, bread yield, and shelf life. Despite its importance, improving WAC through phenotypic selection is challenging. Phenotyping for WAC is time-consuming and, as such, is often limited to evaluation in the latter stages of the breeding process, resulting in the retention of suboptimal lines longer than desired. This study investigates the potential of univariate and multivariate genomic predictions as an alternative to phenotypic selection for improving WAC. A total of 497 hard winter wheat genotypes were evaluated in multi-environment advanced yield and elite trials over eight years (2014-2021). Phenotyping for WAC was done via the solvent retention capacity (SRC) using water as a solvent (SRC-W). Traits that exhibited a significant correlation ($r \geq 0.3$) with SRC-W and were evaluated earlier than SRC-W were included in the multivariate genomic prediction models. Kernel hardness and diameter were obtained using the single kernel characterization system (SKCS), and break flour yield (B-Flour) and total flour yield (T-Flour) were included. Cross-validation showed the mean univariate genomic prediction accuracy of SRC to be $r = 0.69 \pm 0.01$, while bivariate and multivariate models showed an improved prediction accuracy of $r = 0.82 \pm 0.00$. Forward validation showed a prediction accuracy up to $r = 0.81 \pm 0.00$ for a multivariate model that included SRC-W + All traits (SRC-W, Diameter, SKCS hardness and Diameter, F-Flour, and T-Flour). These results suggest that incorporating correlated traits into genomic prediction models can improve early-generation prediction accuracy.

3.2 Introduction

The primary objective of wheat breeding programs is to improve grain yield to meet the increasing demand for wheat *(Triticum aestivum L.)* due to rapid population growth (Alahmad et al., 2022; Xiong et al., 2008). Population growth not only affects wheat quantity but also changes food preferences and consumption trends (Schneider et al., 2011; Xiong et al., 2008), driving the need to improve the milling, dough mixing, and baking quality required for a variety of wheat-based foods like bread, noodles, crackers, cookies, pizza, and cereal bars.

The differences between hard and soft wheat classes are based on kernel texture, milling quality, protein strength, and water absorption (Bhave and Morris, 2008; Kiszonas et al., 2013; Souza et al., 2002). Hard wheat is favored over soft wheat for bread and pizza doughs due to its higher protein concentration, higher water absorption, and stronger gluten (Katyal et al., 2017; Wieser, 2007). Water absorption capacity (WAC) influences dough characteristics such as dough extensibility, crumb texture, loaf volume (Zghal et al., 2001), bread yield (Puhr and D'Appolonia, 1992), and shelf life (Pyler, 1979). To optimize flour's WAC, one can use flour from a variety with a higher absorption capacity or add extracted gluten, though the latter is not cost-efficient. A variety with a higher WAC is preferable for its sustainability and cost-effectiveness. Developing such a variety is challenging, as important quantitative traits like grain yield and quality are influenced by environmental factors, such as weather conditions, soil fertility, and disease and insect pests, as indicated by Bilsborrow et al. (2013) and Hua et al. (2022).

The short time between harvesting and planting in winter wheat breeding programs often limits opportunities to collect WAC and other end-use quality data in a timely manner (Nelson et al., 2016). Phenotyping quality traits is time- and labor-intensive and may require larger grain samples that are often unavailable during early-stage line development (Jernigan et al., 2018;

Kiszonas et al., 2013; Sandhu et al., 2021). Consequently, selection for quality traits is often delayed, leading to advanced breeding lines or released varieties with desirable yields but suboptimal quality (Bassi et al., 2016; Jernigan et al., 2018).

Genomic prediction is an indirect selection approach that enhances the accuracy of marker-assisted selection (MAS) by using genome-wide markers to capture quantitative trait loci (QTL) with small to large effects (Bernardo, 2002; Hayes et al., 2009; Larkin et al., 2019; Meuwissen et al., 2001). It offers several key advantages over phenotypic selection, including early selection, enhanced accuracy through genomic data integration, and selection for traits that are difficult to measure. One of the critical improvements genomic prediction brings is in reducing cycle time and cost, particularly relevant in quality traits where delayed parental selection until quality data are available has traditionally been a bottleneck. By enabling earlier selection decisions, genomic prediction overcomes this limitation and increases the rate of genetic gain. Additionally, it provides more accurate predictions of genetic potential by leveraging genomic information. Lastly, genomic prediction enhances selection efficiency by considering multiple traits and markers, which leads to a more efficient and precise selection of genotypes with desirable quality traits (Shahi et al., 2022).

The prediction accuracy of genomic prediction can be influenced by factors such as training population size, marker density, trait heritability, population structure, and relationship of the training population to selection candidates (Bassi et al., 2016; Larkin et al., 2019; Robertson et al., 2019). Various approaches have been proposed to improve prediction accuracies, including reducing confounding factors like population structure, increasing training population size, and selecting appropriate models (Larkin et al., 2019; Sallam et al., 2020). These models may incorporate marker-trait association scores from genome-wide association studies (GWAS)

(Medina et al., 2021; Zhang et al., 2019), machine learning models (Jubair and Domaratzki, 2023; Zhang et al., 2019), multi-environment models (Guo et al., 2018; Tomar et al., 2021), and multivariate genomic prediction models that include correlated traits (Atanda et al., 2022; Guo et al., 2020; Montesinos-López et al., 2021; Winn, et al., 2023). Several studies have applied multivariate genomic prediction methods to quality traits (Azizinia et al., 2022; Ibba et al., 2020; Montesinos-López et al., 2021; Sandhu et al., 2021). While several studies (Battenfield et al., 2016; Guo et al., 2020; Sandhu et al., 2021) have reported genomic predictions for water absorption capacity (WAC) using and solvent retention capacity of water (SRC-W) for soft white winter wheat, literature focusing on genomic prediction of WAC and SRC-W in hard winter wheat appears to be lacking.

Forward prediction, which predicts future generations using previous generations' data as the training population, is often employed in breeding programs (Belamkar et al., 2018; Calvert et al., 2020; Jarquín et al., 2017). Forward prediction assesses the performance of genomic predictions for lines and environments not yet phenotyped (Haikka et al., 2020). Although previous studies have aimed to improve prediction accuracy by adjusting (filtering, cleaning, and using more balanced data) and cross-validating training populations (Combs and Bernardo, 2013; Hayes et al., 2009; Larkin et al., 2019), the true effectiveness of GS lies in its practical application in breeding programs and validating the performance of genomic predictions.

Water absorption capacity in hard winter wheat is an important breeding target. Despite the complex and time-consuming nature of phenotyping WAC, the use of genomic tools can efficiently enhance selection for WAC in generations where phenotyping is impractical or infeasible. To assess the potential of genomic prediction of WAC, the following objectives were addressed in the current study: (1) explore the association between WAC, flour yield traits, and early generation-applicable traits from the single kernel characterization system (SKCS), to determine their potential for inclusion in multivariate genomic prediction models; 2) to assess the accuracy of univariate and multivariate genomic prediction for WAC through cross-validation, and (3) determine the accuracy of genomic prediction in a breeding scenario through forward validation.

3.3 Materials and Methods

3.3.1 Germplasm

A total of 497 hard winter wheat genotypes (experimental lines and check varieties) generated from the Colorado State University (CSU) wheat breeding program were utilized in this study. These genotypes were evaluated within their respective breeding cycles under different independent trials, including the CSU Elite Trial, the Advanced Yield Nursery (AYN), and the Advanced Yield Nursery comprised of doubled haploid lines (AYND). For a given nursery-year combination, genotypes phenotyped for quality traits may have been sampled from varying numbers of sites. A summary of the number of genotypes and observations in each nursery from 2014 - 2021 is provided (Table 3.1).

3.3.2 Experimental design and trait measurements

The CSU Elite Trials were organized using resolvable, latinized row-column designs with partial replication, as per the methodologies outlined by John and Williams (1995) and Williams et al. (2011). Within the CSU Elite Trials, at a given location, half of the entries were replicated twice, and half were replicated once. In the AYNs, at a given location, roughly one-seventh of the entries were replicated twice, and the remaining entries were replicated once. All genotypes included in this study were evaluated at CSU wheat breeding program sites located in the Great Plains wheat growing region of eastern Colorado, USA. All trials were conducted "on farm" in grower cooperators' fields, except for those at the Agricultural Research, Development, and

Educational Center (ARDEC) in Fort Collins, CO, Plainsman Research Center in Walsh, CO, and the United States Department of Agriculture – Agriculture Research Service Central Great Plains Center in Akron, CO. The agronomic and crop management practices mirrored those adopted by the grower cooperators and were adjusted based on the standard practices at each site. Each plot measured 1.5 m wide and 3.7 m long and was planted in six rows. The seeding rate was approximately 1.73 million seeds hectare^{-1}. All six rows of each plot were harvested, and a cleaned sample of the grain was used for subsequent quality analyses.

In this study, traits included as covariates were chosen based on their phenotypic correlation with values from the solvent retention capacity (SRC) test using water as a solvent (SRC-W) and their timing in the data collection process relative to SRC-W. The phenotypic correlations are presented in Figure 3.1. Traits that would normally be evaluated after SRC-W in the overall testing scheme or traits showing a correlation of $r \leq [0.3]$ were excluded: grain protein concentration (GPRO), bake water absorption (BakeAbs), bake mixing time (BakeMT), bake loaf volume (BakeV), Mixograph tolerance (MixoT), Mixograph mixing time (MPT), Mixograph midline left slope (MLS), Mixograph midline right slope (MRS), Mixograph midline peak value (MPV), Mixograph midline peak width (MPW), and Mixograph midline right width (MRW). Traits utilized for subsequent analyses were SKCS kernel hardness and kernel diameter, break flour yield (B-Flour), and total flour yield (T-Flour). The phenotyping details are presented only for the selected traits.

All traits included in this study were analyzed in the CSU Wheat Quality Lab located in Fort Collins, CO, following the procedures of the American Association of Cereal Chemists International (AACC International, 2010). SKCS kernel diameter and kernel hardness were evaluated using 10-15 g of grain using the Single Kernel Characterization System 4100

(PertenInstruments, Springfield, IL; AACC International, 2010). Grain samples (50 g) for each genotype were tempered to 15 g kg^{-1} (15 %) moisture and milled using a modified Brabender Quadrumat Senior Mill (Brabender Instruments, NJ, USA). The material passing the first break roll was separated by a two-stage sieve where only the first break flour passes through the sieve. After measuring the first break flour, the remaining material underwent further milling over the reduction rolls. The T-Flour was measured (in g kg^{-1}) as the combined total of white flour from the first and second mill streams. The B-Flour measured (in g kg^{-1}) of the white flour from the first pass through the mill rolls.

Water absorption capacity was determined from white flour samples using the SRC test using water as a solvent, according to AACC International (2010). Empty 2 mL microcentrifuge tubes were labeled and weighed for each sample. Each tube was then filled with 200 mg of flour for each sample. Tubes were mixed with 1 mL double distilled water and vortexed for 5 sec to suspend the flour. The mixture was then shaken for 20 min using a rotator, ensuring thorough mixing of the flour and water. Subsequently, samples were centrifuged for 15 min at 1000 g at room temperature. After centrifugation, the supernatant was removed, and the gel-containing tubes were dried for 10 min and then weighed. The gel (pellet) weight was calculated by subtracting the weight of the empty tube from the weight of the tube with the pellet. The SRC value for each sample was then calculated according to the following formula:

$$SRC(\%) = \left[\frac{pellet\ weight(g)}{(flour\ weight(g)} - 1\right] x \left[\frac{(86)}{(100 - moisture(\%)}\right] x\ 100$$

Where SRC (%) is the solvent retention capacity in percentage, and moisture (%) is the flour moisture determined using a Foss DA1650 near-infrared spectrometer (Foss North America, Eden Prairie, MN).

3.3.3 Phenotypic data analysis

Across trial-location-year, the best linear unbiased estimates (BLUEs) were calculated using the *ASREML-R* package in R statistical software (Butler et al., 2009; R Core Team, 2013). The following mixed linear model was used to estimate BLUEs across site-nursery-year combinations:

$$y_{ijk} = \mu + G_i + e{:}r_j + e{:}c_k + \varepsilon_{ijk}$$

Where y_{ijk} is the response variable for i^{th} level of genotype, in the j^{th} level of row, in the k^{th} level of column; μ is the overall mean; $G_i$is the fixed genotype effect; e:r_j is the random row within the trial-location effect ($r_j \sim N(0, \sigma_j^2)$), $e{:}c_k$ the random column within the trial-location effect ($ck \sim N(0, \sigma k^2)$) and ε_{ijk} is the residual error term ($\varepsilon_{ijk} \sim N(0, \sigma\varepsilon^2)$). Summary statistics (mean, minimum, maximum, and standard deviation) were calculated to assess the distribution of the data across the range of values in the population. Pearson correlation coefficients were calculated and visualized using the *psych* package in R (Revelle and Revelle, 2017).

A multivariate genomic mixed linear model was performed to calculate the genetic correlations and variance components to estimate heritability for all traits included in this study. The following model was used:

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

Where y is a matrix of observed phenotypic values (BLUEs) for the five traits, each column in y represents one of the traits (SRC-W, SKCS kernel hardness, SKCS kernel diameter, T-flour, B-flour), and each row corresponds to an individual; β is a vector of fixed effects; X is the design matrix for the fixed effects; u is a vector of random effects (BLUPs); Z is the design matrix relating genotypes to the BLUEs; and e is a vector of residual errors. Genomic heritability was calculated as follows:

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_\varepsilon^2}$$

Where σ_g^2 is the genetic variance and σ_ε^2 is the residual error.

3.3.4 Genotyping

Genotyping was done using genotyping-by-sequencing (GBS; Elshire et al., 2011). Genomic DNA was extracted from one-week-old wheat leaves in a 96-well format using the MagMax14 plant DNA kit and quantified using PicoGreen (Thermo Fisher Scientific, MA, USA) assays. For GBS library construction, a modified protocol from Poland et al. (2012) was used to create libraries (384-plex), which were sequenced on an Illumina HiSeq2500 platform at the University of Illinois. Single nucleotide polymorphism (SNP) calling was performed using the *TASSEL 5.0* GBSv2 pipeline (Glaubitz et al., 2014), with a 64-base *k*-mer length and a minimum *k*-mer count of five. Sequences were aligned to the International Wheat Genome Sequencing Consortium (IWGSC) RefSeq v2.0 'Chinese Spring' wheat reference sequence (Appels et al., 2018) using the Burrows-Wheeler aligner version 0.7.10 (Li and Durbin, 2009).

The raw SNP data generated from the *TASSEL* pipeline were filtered to remove individuals with more than 50% missing data and average heterozygosity greater than 30%. Markers were filtered to select biallelic SNPs with a minor allelic frequency greater than 5%, less than 10% missing data, and average heterozygosity less than or equal to 10%. Missing GBS data were imputed using the Beagle algorithm (Browning et al., 2018). All unaligned SNPs were dropped from the dataset prior to imputation. After filtering and imputation, 23,130 SNPs were used for subsequent analyses.

3.3.5 Univariate and Multivariate Genomic Prediction

Univariate genomic prediction for SRC-W was performed using a genomic best linear unbiased prediction *(GBLUP)* model using *ASREML* function with *ASREML-R* package as follows:

$$y = X\beta + Zu + \varepsilon i$$

Where u represents the vector of genotype effects, assumed to follow a normal distribution $u \sim N(0, G\sigma u^2)$, with G being the genomic relationship matrix which is calculated using the *GRM* function from the *gaston* package in R (Perdry, 2022) and σu^2 is the variance of individual genotype effects; β is the vector of fixed effects (the overall mean); and X and Z are the design matrices associated with fixed and random effects, respectively. The design matrix Z has m markers in its columns and no observed genotypes in its rows. The residual error at the ith locus, represented by εi, is assumed to have a normal distribution $\varepsilon i \sim N(0, I\sigma\varepsilon^2)$, where I is the identity matrix, and $\sigma^2\varepsilon$ is the residual error variance. The genomic estimated breeding value (GEBV) was calculated as the sum of the additive allele effects for a given genotype (Chen and Zhang, 2018; VanRaden, 2008).

Multivariate genomic prediction was carried out using *ASREML-R*. A general interpretation of the multivariate model is described as follows:

$$y = X\beta + Zu + \varepsilon i$$

Where y is an n x m matrix of observations, where n represents the number of observations and m indicates the different traits with each column in y corresponding to a different trait; X is the design matrix that corresponds to fixed effects (overall mean); β is a matrix of fixed effects, each column of which associated with a different trait; Z represents the design matrix of random effects and includes the genotypes of individuals; u is a vector representing the genotype effects, assumed to follow a multivariate normal distribution, denoted as $u \sim N(0, GA)$ with G representing the genomic relationship matrix and *A* being an unstructured covariance matrix that captures the variance and covariance of the genetic effects across the traits; εi is an *n x m* matrix of residual errors with *n* rows corresponding to each observation and *m* columns representing each trait. Each residual error indicated as εi at the i^{th} locus is assumed to follow a multivariate normal distribution, expressed as $\varepsilon i \sim N(0, IR)$ where *I* is the identity matrix and *R* is an unstructured covariance matrix capturing the variance and covariance of the residuals across traits. The genomic estimated breeding values (GEBVs) were calculated for each trait as the cumulative sum of the additive allele effects for a specific genotype.

The accuracy of genomic prediction was assessed through both cross-validation and forward validation. In cross-validation, the dataset was randomly split into training (80%), and testing (20%) sets, and the model was trained to predict GEBVs for the testing set. This process was repeated 100 times with different random subsets of the data for training and testing to evaluate the prediction accuracy of the model. Prediction accuracy was considered as the Pearson correlation coefficient (*r*) between the GEBVs and the observed data (BLUEs). The prediction accuracy of all models was visualized using a box plot constructed with the *ggplot2* package in R (Wickham, 2006).

The forward validation process involved the same 497 genotypes that were part of the cross-validation between 2014 and 2021. Seven different validation sets were formed using the CSU Elite Trial data from 2019-2021. Firstly, the BLUEs from the Elite trial were used as a validation set (FV_2019, FV_2020, and FV_2021). Subsequently, the data from each year with another year were combined to create two-year forward validation sets, designated as FV_2019and2020, FV_2019and2021, and FV_2020and2021. Finally, the data from all three years were assigned as a validation set (FV_2019-2021). The data remaining after the exclusion of the validation set were utilized as the training set for the corresponding evaluation. Unlike cross-validation, the entire prediction process in forward validation is non-iterated, making it a fixed, one-time assessment based on the initial data split. The number of observations for each genotype in each year is described (Table 3.1).

In both cross-validation and forward-validation multivariate genomic prediction, only the data for SRC-W in the validation set were masked for prediction in the training set. However, the data for the correlated traits of these genotypes remained in the model. For instance, during forward prediction model training, the SRC-W data for 58 genotypes from three years of CSU Elite Trials (2019-2021) were masked, while the covariate trait values associated with these genotypes were included in the models.

3.4 Results

3.4.1 Phenotypic variation

Summary statistics demonstrate that there is considerable genotypic variation for each trait (Table 3.2). SKCS kernel hardness displayed the greatest range between its minimum (49.2) and maximum values (86.6), indicating a wider variance in the total population for this trait. The mean SRC-W was 60.6 %, with a minimum of 54.4 % and a maximum of 71.1 %. Traits exhibited moderate heritability values, ranging from 0.50 ± 0.09 to 0.59 ± 0.08 (Table 3.2). The trait with the highest heritability was B-Flour ($h^2 = 0.59 \pm 0.08$), with a mean of 47.5 g kg^{-1}, followed by SRC-W with ($h^2 = 0.58 \pm 0.09$). The lowest heritability $h^2 = 0.50 \pm 0.08$ was observed for SKCS kernel diameter.

The Pearson correlation results showed both strong positive and negative correlations between the variables. The strongest positive correlation ($r = 0.82$, $p < 0.001$) was observed between B-Flour and T-Flour, whereas the strongest negative correlation ($r = -0.75$, $p < 0.001$) was observed between SRC-W and B-Flour (Figure 3.2 A). Similarly, T-Flour showed a significant negative correlation (r = -0.54, p < 0.001) with SRC-W. SKCS kernel hardness showed a significant positive correlation (r = 0.61, p < 0.001) with SRC-W and a significant negative correlation with B-Flour (r = -0.68, p < 0.001) and T-Flour (r = -0.46, p < 0.001). However, no correlation (r = 0.00, p > 0.05) was observed between SKCS kernel diameter and T-Flour, and a very weak correlation (r = 0.06, p > 0.05) was observed between SKCS kernel hardness and SKCS kernel diameter.

Genetic correlation values ranged from r = -0.76, p < 0.001, to r = 0.74, p < 0.001 (Figure 3.2B). There were observable differences compared to the phenotypic correlations, while some changed in magnitude. For instance, the phenotypic correlation between SKCS kernel hardness and T-Flour (r = -0.46, p < 0.001) was decreased in the genetic correlation (r = -0.21, p < 0.05). The same trend was shown for the genetic correlation between SRC-W and T-Flour, which reduced from (r = -0.54, p < 0.001) for the phenotypic correlation to r = -0.42, p < 0.01 for the genetic correlation. However, no change in the direction of the correlation was observed.

3.4.2 Univariate and multivariate models

The prediction accuracies of the univariate, bivariate, and full multivariate models for predicting SRC-W were validated using a cross-validation approach. These models exhibited varying degrees of accuracy, ranging from r = 0.69 ± 0.005 to r = 0.82 ± 0.003 (Figure 3.3). The analysis encompassed a progression from a univariate model to four bivariate models, concluding with a full multivariate model. The univariate model, utilizing only SRC-W as a predictor, showed a prediction accuracy of r = 0.69 ± 0.005. When SKCS kernel diameter was included in the first bivariate model, there was a marginal improvement in the prediction accuracy to r = 0.70 ± 0.004, representing only a one percent increase over the univariate model.

A higher prediction accuracy of r = 0.78 ± 0.004 was observed in the second bivariate model that included SKCS kernel hardness. This model demonstrated a notable 9% improvement over the univariate model. Regarding the two other bivariate models that integrated milling-related traits, prediction accuracy with the inclusion of T-Flour was r = 0.75 ± 0.004 and r = 0.82 ± 0.003 with the inclusion of B-Flour. The multivariate model, which combined all grain and flour-related traits, showed a prediction accuracy of r = 0.82 ± 0.003, identical to the highest accuracy observed with the bivariate model incorporating B-Flour. This represents a 13% prediction accuracy improvement over the univariate model, making it the highest accuracy observed among the six models.

3.4.3 Forward prediction accuracies in the CSU Elite trials

Forward validation was also done to assess the accuracy of the univariate, bivariate, and multivariate models. Data were partitioned into one-year, two-year, and three-year forward validation sets. In the three-year validation set (FV_2019-2021), the prediction accuracy ranged from r = 0.65 for the univariate model to r = 0.81 for the SRC-W + B-Flour bivariate model and a model including all measured traits (Figure 3.4). The bivariate model, including SRC-W and SKCS kernel hardness, had a prediction accuracy of r = 0.75, marking a 10% improvement compared to the univariate model from the FV_2019-2021 set. On the other hand, the inclusion of SKCS kernel diameter exhibited very little improvement in prediction accuracy (r = 0.66), only a one percent improvement from the univariate model.

Among all validation sets, the single-year set (FV_2019) exhibited the highest multivariate prediction accuracy (r = 0.93), as indicated by the multivariate model (Figure 3.5), which represents a 12% increase over the highest prediction accuracy observed from the three-year validation set (FV_2019-2021). Conversely, the lowest prediction accuracy range was found for the FV_2021 set, with values ranging from r = 0.44 (for the univariate SRC-W model) to r = 0.61 (for the multivariate SRC-W + All traits model) model (Figure 3.5). This marks a 20% decrease from the highest prediction accuracy observed for the FV_2019-2021 set and a substantial 32% decrease compared to the FV_2019 single-year validation set. Generally, all sets, including various combinations from the FV_2021, exhibited lower accuracies than those excluding FV_2021. The FV_2021 set comprised 33 genotypes, while the FV_2020 and FV_2019 sets had 29 and 37 genotypes, respectively. Notably, the FV_2019 set outperformed others in prediction accuracy, with sets integrating 2019 CSU Elite Trial data showing modest improvements.

3.5. Discussion

In the current study, a historical dataset of hard winter wheat quality data from multiple sites and years was analyzed together with their genotypic data. Univariate and multivariate models were compared in terms of their prediction accuracy for water absorption capacity as measured using the SRC test with water as a solvent (SRC-W). Both cross and forward-validation approaches were employed to validate the robustness and predictive accuracy of the models tested. Generally, for both cross-validation and forward prediction, the accuracy of predictions improved in multivariate models when the covariate demonstrated higher or nearly identical heritability and showed a significant correlation with the target trait. Traits used as covariates in this study include SKCS kernel hardness and diameter, break flour yield, and total flour yield.

The moderate heritability observed for all traits included in this study suggests a balance between genetic and environmental influences on their expression. This balance implies the potential for effective selection while also underscoring the role of environmental or management interventions in affecting their expression (Bernardo, 2002; Holland et al., 2003). Traits such as B-flour, T-flour, and SKCS kernel hardness showed almost the same heritability and a strong positive correlation with SRC-W, and when included in the models, they led to improved prediction accuracies. However, the SKCS kernel diameter showed a weak correlation with SRC-W, and its inclusion as a covariate did not result in an appreciable improvement in prediction accuracy. This indicates that strong correlations and heritability are the main factors in improving prediction accuracy in multivariate models, in agreement with previous studies (Crossa et al., 2017; Ibba et al., 2020).

The multivariate (SRC-W + All traits) and bivariate model (SRC-W + B-Flour) models showed a 13% increase in prediction accuracy compared to the univariate and the other bivariate models. These results underscore the significance of multivariate models over univariate models and indicate that these two models were the best fit for the data, successfully capturing the underlying genetic variation. Similarly, the bivariate model that included SRC-W and SKCS kernel hardness showed increased prediction accuracy. Here's how these three traits interplay: harder grains require greater milling force, which leads to increased starch damage, which in turn increases the water absorption capacity during dough mixing (Arya et al., 2015; Kweon et al., 2011; Sapirstein et al., 2018). On the other hand, B-Flour is negatively correlated with grain SKCS kernel hardness as harder grains produce less B-Flour, and thus more of the total product must pass over the reduction rolls, leading to greater starch damage and thus increased water absorption (Hogg et al., 2004; Symes, 1969). This suggests that both B-Flour and SKCS kernel hardness are critical traits that, in combination, influence the water absorption capacity of the milled flour and, by extension, the predictive accuracy of the models.

Previous studies have shown the effectiveness of multivariate models over univariate models. Gill et al. (2023) in winter wheat and Bhatta et al. (2020) in barley (*Hordeum vulgare L.*) found that a multivariate model outperformed a univariate model in predicting grain yield using agronomic traits as covariates. Similarly, Montesinos-López et al. (2021) reported improved prediction accuracy for grain yield with multivariate models using quality traits as covariates compared to univariate models with no covariates. Moreover, multivariate models outperformed univariate models for predicting grain yield and quality traits in wheat (Guo et al., 2020; Lozada and Carter, 2020; Sandhu et al., 2021). These findings support the advantages of utilizing multivariate models for predicting wheat quality traits, leading to improved prediction accuracy compared to univariate models. Incorporating a correlated trait in multivariate models may not always result in improved prediction accuracy, especially when the covariate has low heritability, and the correlation is not strong (Shahi et al., 2022; Sun et al., 2017). The correlation between SKCS kernel diameter and SRC-W was significant but relatively lower in magnitude, resulting in no improvement in prediction accuracy compared to the univariate model. This might be due to environmental variation or interactions that weren't captured by the respective models or due to the traits not being correlated (Sandhu et al., 2022). Other reports have also shown no improvement in certain multivariate models, including Lado et al. (2018) for wheat baking quality traits and Schulthess et al. (2016) for grain yield and protein concentration in rye (*Secale cereale* L.).

In this study, forward validation accuracy was similar to that observed for cross-validation (less than a five percent difference between the two validation approaches), indicating that the models would tend to be useful in a breeding program for predicting individuals that were not part of the training panel (unseen data) (Battenfield et al., 2016; Yadav et al., 2021). Such performance is a testament to a robust model that is applicable beyond its training data (Yadav et al., 2021). The similar prediction accuracy between the two validation methods further emphasizes the model's consistent performance, regardless of the validation approach used.

In most cases, achieving higher prediction accuracy for forward validation is challenging, as it involves predicting future outcomes where environmental conditions, management practices, and other factors may differ from the data used in the training set (Hoffstetter et al., 2016; Jarquín et al., 2017; Juliana et al., 2018; Lozada and Carter, 2020). Several possible reasons account for the higher prediction accuracy observed in this study using both validation approaches. One reason is that the genotypes used in both the validation and training sets were sourced from the same breeding program, suggesting a high probability of a shared genetic background. This similarity in genetic makeup can increase the representativeness of the training set, potentially leading to improved prediction accuracy (Battenfield et al., 2016). Another reason might be the quality of the phenotypic data used, which can significantly affect prediction accuracy. Accurate phenotypic data is a primary driving factor for increased prediction accuracy and genetic gain, and genomic prediction, in turn, relies on the quality of this phenotypic data (Bartholomé et al., 2022; Beyene et al., 2019; Sandhu et al., 2021). Another common concern is overfitting, where a model performs well on training data but underperforms with new data (Montesinos López et al., 2022). However, the similar accuracies observed in this study for both cross-validation and forward validation mitigate these concerns. In essence, the observed higher prediction accuracies suggest that the models are reliable, robust, and well-suited for generalizing to new individuals.

The comparable prediction accuracy results observed with both validation approaches contrast with previous reports of higher prediction accuracy from forward validation over cross-validation. For instance, Zhang et al. (2022) reported higher prediction accuracy with forward validation for Fusarium head blight resistance in hard winter wheat. Similarly, Haikka et al. (2020) observed a slight increase in prediction accuracy using forward validation in their genomic study of Finnish oats (*Avena sativa* L.) and barley. Azizinia et al. (2023) also reported improved forward prediction accuracy in multi-trait prediction for wheat quality. Furthermore, Fradgley et al. (2023) in wheat and Zystro et al. (2020) in sweet corn (*Zea mays* L.) also observed higher forward prediction accuracy than with a cross-validation approach.

However, several authors have reported lower prediction accuracies with forward validation. Jarquín et al. (2017) found lower prediction accuracy (by 37%) for grain yield in wheat using a forward validation approach. Similarly, Battenfield et al. (2016) reported a 19% decrease in forward-validation than cross-validation, and Sweeney et al. (2019) also reported reduced prediction accuracy (by 35%) from forward-validation for wheat end-use quality traits. Dawson et al. (2013) observed lower prediction accuracy (lower by 17%) from a forward validation approach for wheat grain yield.

Chapter 3 Tables

Table 3.1 Number of observations and genotypes used in this study per year per trial.

Year	Trial	No. of Locations	Number of Observations	Number of Genotypes
2014	AYND[a]	3	105	35
2015	AYN[b]	2	56	29
	ELITE[c]	2	68	34
2016	AYN	3	99	33
	ELITE	5	165	33
2017	ELITE	4	140	35
2018	AYN	3	90	30
	ELITE	5	95	19
2019	AYN	3	75	25
	AYND	3	78	35
	ELITE	6	185	37
2020	AYN	3	93	31
	AYND	3	96	33
	ELITE	4	116	29
2021	AYN	2	52	26
	ELITE	3	99	33

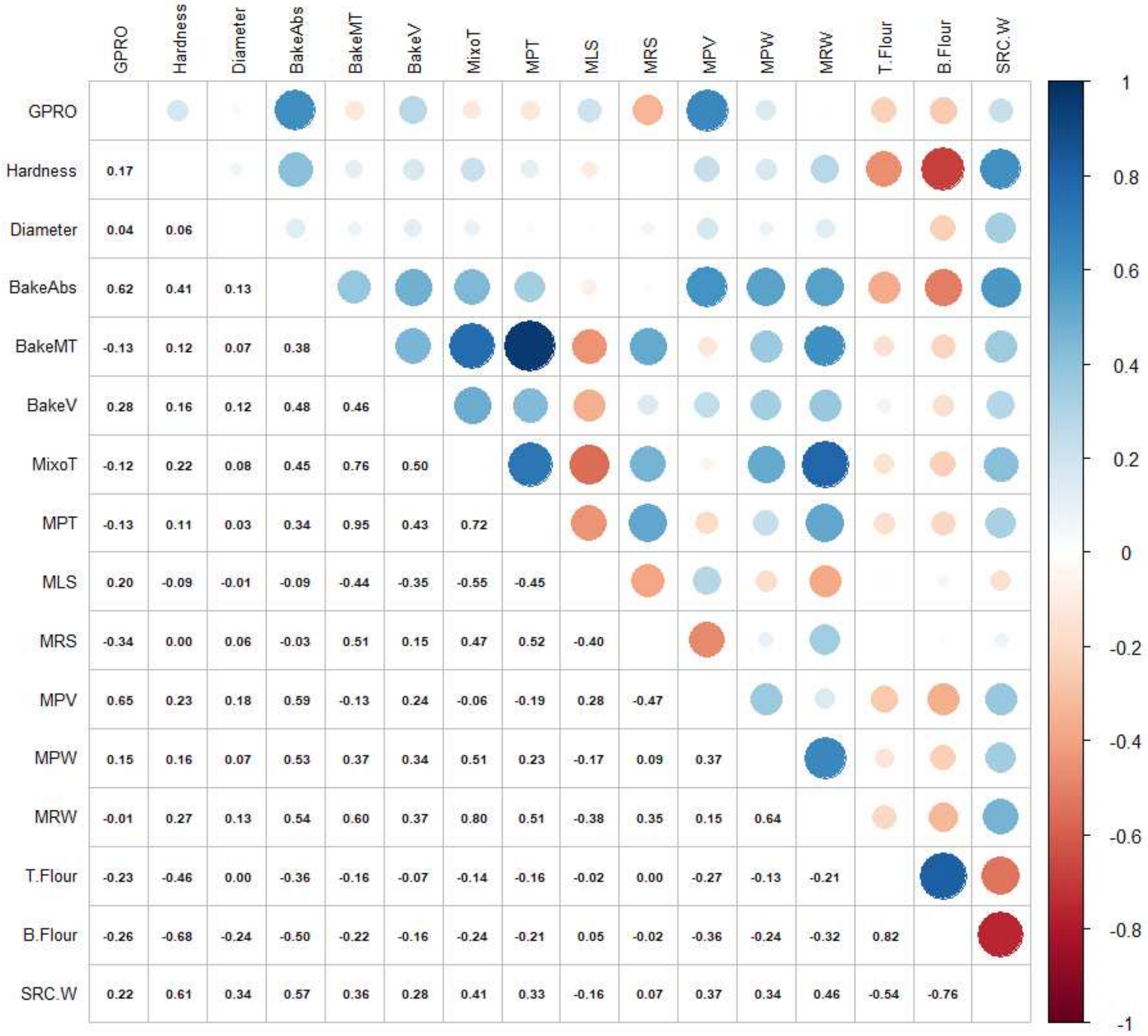
Abbreviations: [a] Advanced doubled haploid yield nursery. [b] Advanced yield nursery. [c] CSU Elite Trial

Table 3.2 Summary statistics for quality traits included in this study.

Traits	Mean	Min[a]	Max[b]	SD[c]	h^{2d}	SE h^{2e}
SRC-W	60.6	54.4	71.1	2.75	0.58	0.09
SKCS kernel diameter	2.6	2.3	2.9	0.09	0.50	0.09
SKCS kernel hardness	68.7	49.2	86.6	6.14	0.57	0.09
B-Flour	47.5	34.5	54.1	3.06	0.59	0.08
T-Flour	69.3	62.1	74.3	1.76	0.55	0.08

Abbreviations: [a] minimum; [b] maximum; [c] standard deviation; [d] heritability; [e] standard error of the heritability. SRC-W refers to the solvent retention capacity using water as a solvent, SKCS refers to a single kernel characterization system, B-Flour refers to break flour yield, and T-Flour refers to total flour yield.

Chapter 3 Figures



	GPRO	Hardness	Diameter	BakeAbs	BakeMT	BakeV	MixoT	MPT	MLS	MRS	MPV	MPW	MRW	T.Flour	B.Flour	SRC.W
GPRO																
Hardness	0.17															
Diameter	0.04	0.06														
BakeAbs	0.62	0.41	0.13													
BakeMT	-0.13	0.12	0.07	0.38												
BakeV	0.28	0.16	0.12	0.48	0.46											
MixoT	-0.12	0.22	0.08	0.45	0.76	0.50										
MPT	-0.13	0.11	0.03	0.34	0.95	0.43	0.72									
MLS	0.20	-0.09	-0.01	-0.09	-0.44	-0.35	-0.55	-0.45								
MRS	-0.34	0.00	0.06	-0.03	0.51	0.15	0.47	0.52	-0.40							
MPV	0.65	0.23	0.18	0.59	-0.13	0.24	-0.06	-0.19	0.28	-0.47						
MPW	0.15	0.16	0.07	0.53	0.37	0.34	0.51	0.23	-0.17	0.09	0.37					
MRW	-0.01	0.27	0.13	0.54	0.60	0.37	0.80	0.51	-0.38	0.35	0.15	0.64				
T.Flour	-0.23	-0.46	0.00	-0.36	-0.16	-0.07	-0.14	-0.16	-0.02	0.00	-0.27	-0.13	-0.21			
B.Flour	-0.26	-0.68	-0.24	-0.50	-0.22	-0.16	-0.24	-0.21	0.05	-0.02	-0.36	-0.24	-0.32	0.82		
SRC.W	0.22	0.61	0.34	0.57	0.36	0.28	0.41	0.33	-0.16	0.07	0.37	0.34	0.46	-0.54	-0.76	

Figure 3.1 Pearson correlation coefficients between end-use quality traits of hard winter wheat genotypes. The correlation coefficients are numerically displayed on the lower half of the plot and on the upper half by color-coded circles: blue for positive associations and red for negative ones. The intensity and size of the circles correlate with the strength of the relationship. The scale on the right indicates the correlation values, with 1 being a perfect positive correlation, -1 a perfect negative correlation, and 0 indicating no correlation.

Abbreviations: Grain protein (GPRO), SKCS kernel hardness (Hardness), SKCS kernel diameter (Diameter), bake absorption (BakeAbs), bake mixing time (BakeMT), bake volume (BakeV), Mixograph tolerance (MixoT), Mixograph mixing time (MPT), Mixograph midline left slope (MLS), Mixograph midline right slope (MRS), Mixograph midline peak value (MPV), Mixograph midline peak width (MPW), Mixograph midline right width (MRW), total flour yield (T-Flour), break flour yield (B-Flour) and solvent retention capacity using water as a solvent (SRC-W).

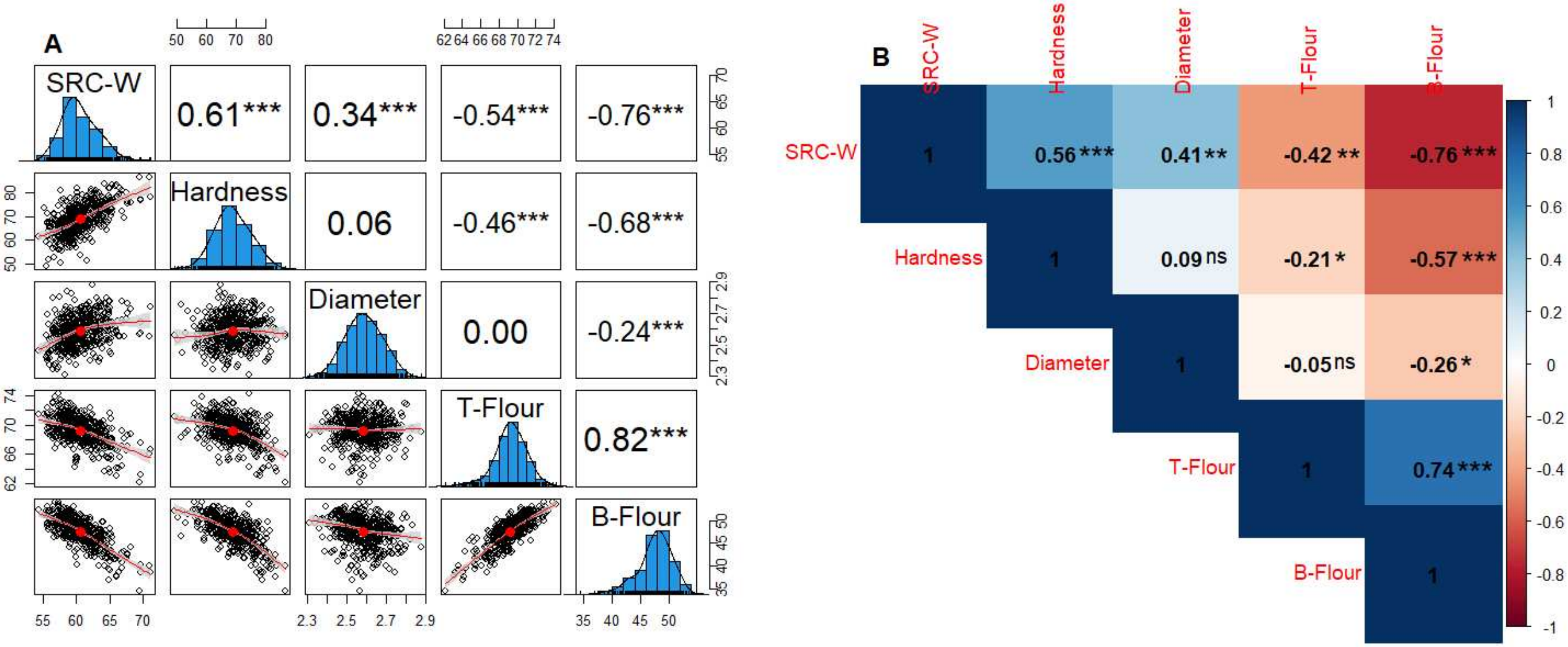


Figure 3.2 Phenotypic and genetic correlation among measured traits. [A] Pearson's correlation coefficient plot in which histograms and trait names are displayed on the central diagonal. The traits represented include solvent retention capacity using water as a solvent (SRC-W), single kernel characterization system (SKCS), SKCS kernel diameter (Diameter), SKCS kernel hardness (Hardness), break flour yield (B-flour), and total flour yield (T-Flour). The scatterplots of the traits and smoothed regression lines are shown in the lower half of the boxes. The values on the upper half of the diagonal represent the correlation coefficients between the traits. The significance of the correlation for corresponding traits where ns = non-significant, $* = 0.05$, $** = 0.01$, and $*** = 0.001$. If no stars are present, the correlation of the variables indicated by the boxes is not significant. [B] Heatmap correlation plot, which illustrates the pairwise genetic correlations between variables in the dataset. Each cell in the grid represents the genetic correlation coefficient between two variables, with color intensity indicating the strength and direction of the correlation. The darker blue colors indicate stronger positive correlations, the darker red colors indicate stronger negative correlations and lighter colors indicate weaker or no correlation.

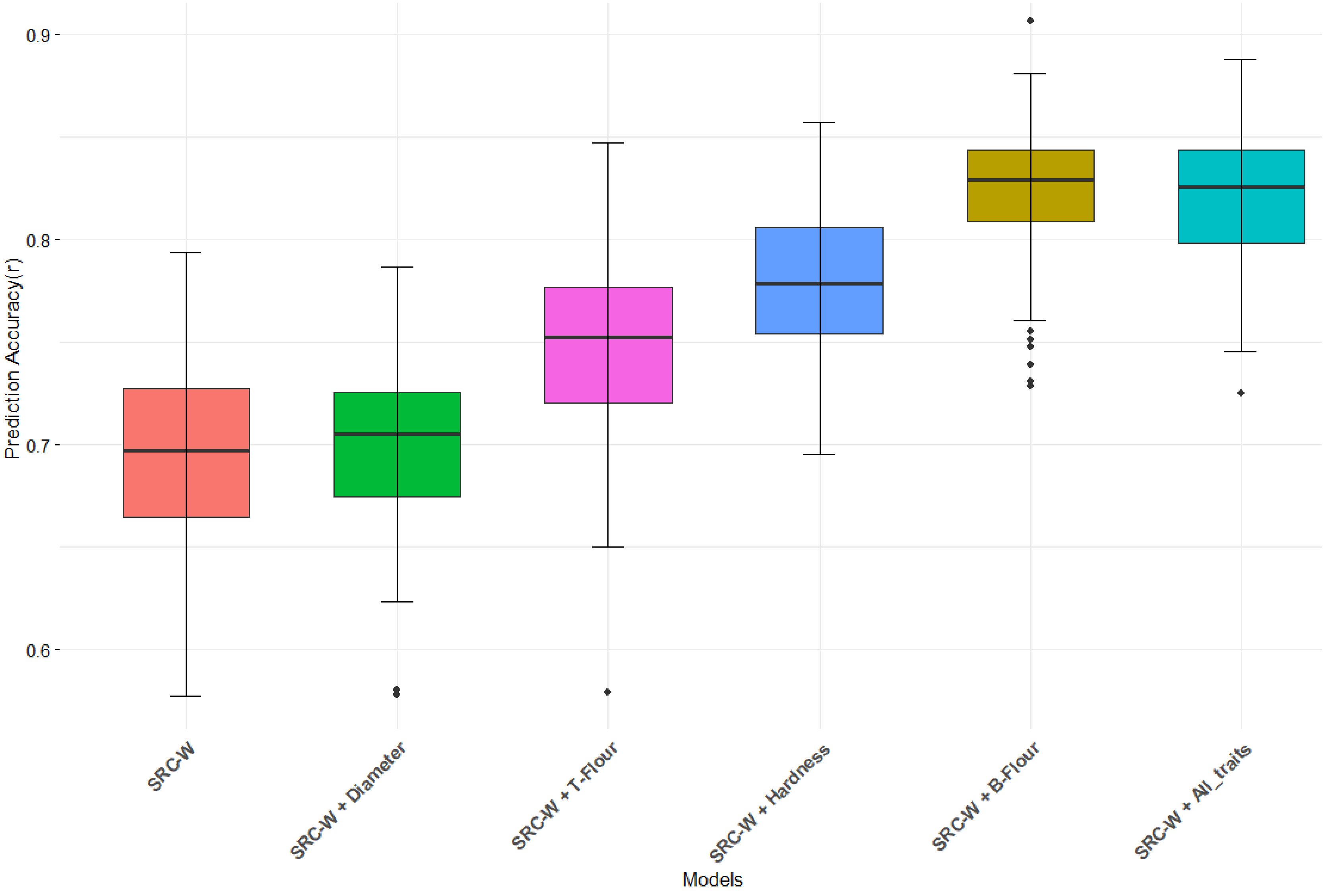

0.9
0.8
0.7
0.6
Prediction Accuracy(r)
SRC-W
SRC-W + Diameter
SRC-W +T-Flour
SRC-W + Hardness
SRC-W +B-Flour
SRC-W + All_traits
Models

Figure 3.3 Cross-validation prediction accuracy (*r*) of six different genomic selection models. The x-axis displays the model type, with the trait names such as solvent retention capacity using water as a solvent (SRC-W), single kernel characterization system (SKCS) kernel diameter (Diameter) and kernel hardness (Hardness), total flour yield (T-Flour), break flour yield (B-Flour), and All traits when all covariates were included in the model. The "+" symbol indicates that multiple traits are included in the model. The y-axis represents the correlation coefficient (r) between the genomic estimated breeding value and the best linear unbiased estimated value. Each box represents the interquartile range of the distribution of correlation coefficients, with the median indicated by a horizontal line within the box. Data points outside the box are considered outliers.

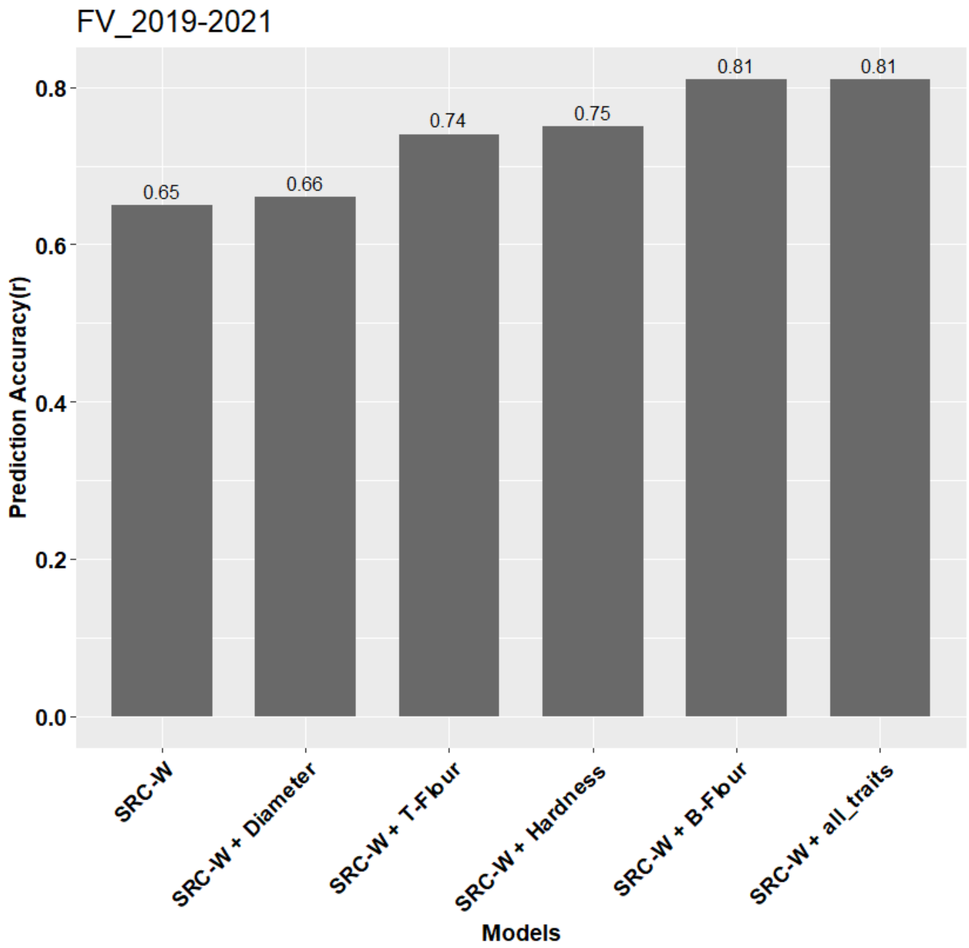


Figure 3.4 Bar charts of forward-validated prediction accuracies for solvent retention capacity using water as a solvent (SRC-W) derived from three years of Colorado State University (CSU) Elite Trials conducted by the Colorado State University Wheat Breeding Program. The prediction accuracy represented as Pearson's correlation coefficient (r) is shown on the y-axis and is also explicitly displayed on top of the corresponding bar. The x-axis shows the names of the respective models tested – one univariate, four bivariate, and one comprehensive multivariate model, in that order. Covariates are included in the models with Diameter (SKCS kernel diameter), Hardness (SKCS kernel hardness), B-flour (break flour yield), and T-Flour (total flour yield).

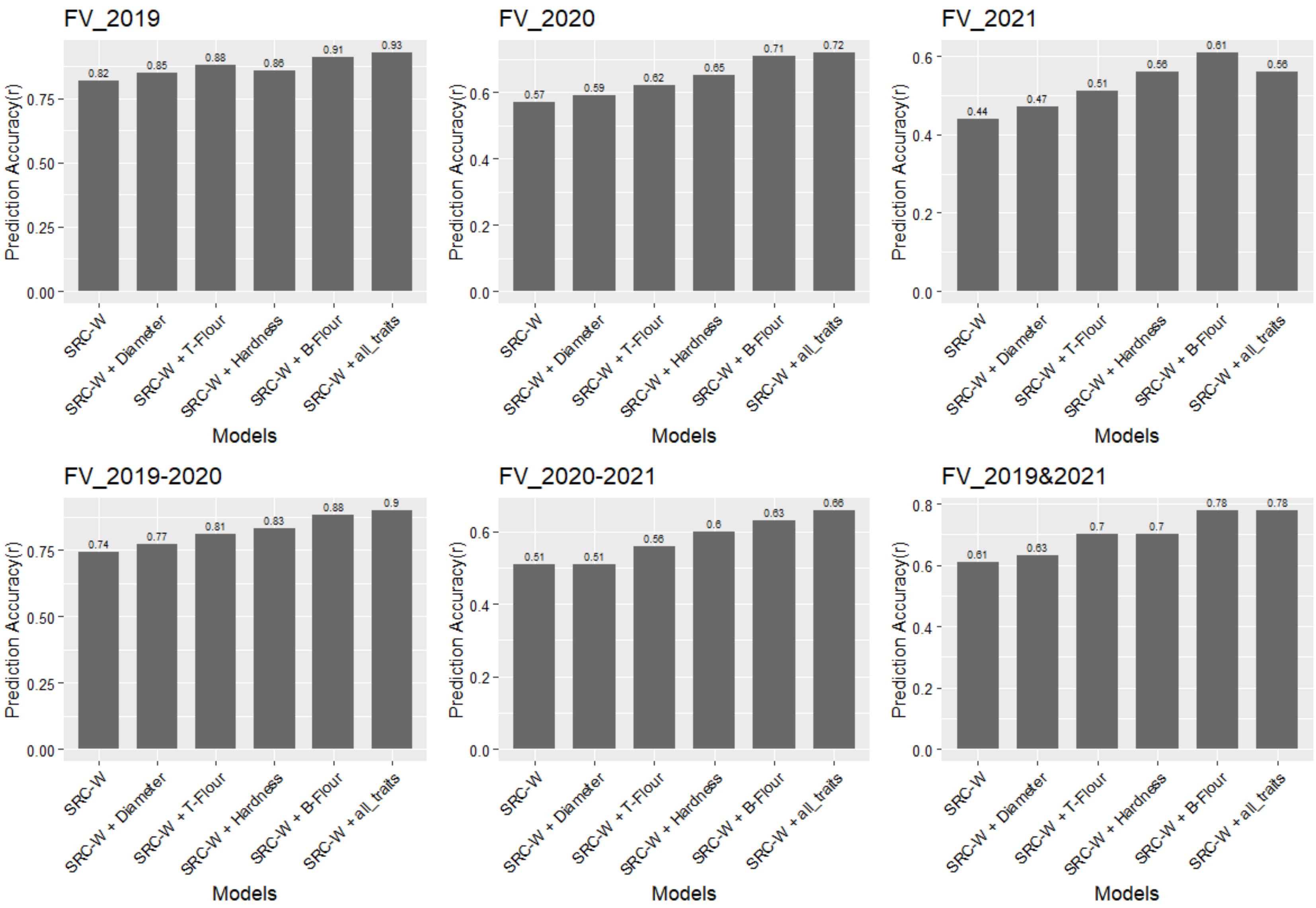

FV_2019
Prediction Accuracy(r)
0.00
0.25
0.50
0.75
0.82
0.85
0.88
0.86
0.91
0.93
FV_2020
Prediction Accuracy(r)
0.0
0.2
0.4
0.6
0.57
0.59
0.62
0.65
0.71
0.72
FV_2021
Prediction Accuracy(r)
0.0
0.2
0.4
0.6
0.44
0.47
0.51
0.56
0.61
0.56
FV_2019-2020
Prediction Accuracy(r)
0.00
0.25
0.50
0.75
0.74
0.77
0.81
0.83
0.88
0.9
FV_2020-2021
Prediction Accuracy(r)
0.0
0.2
0.4
0.6
0.51
0.51
0.56
0.6
0.63
0.66
FV_2019&2021
Prediction Accuracy(r)
0.0
0.2
0.4
0.6
0.8
0.61
0.63
0.7
0.7
0.78
0.78
SRC-W
SRC-W + Diameter
SRC-W + T-Flour
SRC-W + Hardness
SRC-W + B-Flour
SRC-W + all_traits
Models

Figure 3.5 Forward-validated prediction accuracies for solvent retention capacity using water as a solvent (SRC-W), derived from three years of CSU Elite Trials conducted by the Colorado State University Wheat Breeding Program. The data are organized as single-year and two-year combinations. The y-axis represents the prediction accuracy, measured as Pearson's correlation coefficient (r), and the corresponding value is explicitly displayed on top of each bar. The x-axis indicates the models used, including one univariate, four bivariate, and one multivariate model, listed in that order. The traits included in the models are solvent retention capacity using water as a solvent (SRC-W), single kernel characterization system (SKCS), kernel hardness (Hardness) and kernel diameter (Diameter), break flour yield (B-Flour), and total flour yield (T-Flour). Additionally, the plot titles at the top of each plot denote the validation data used to calculate the prediction accuracies represented in the graph. These labels, indicated as (FV_years), specify that the data utilized for validation corresponds to the specific year mentioned. For example, FV_2019 signifies that the forward validation set included only the 2019 CSU Elite Trial data.

Chapter 3 References

Alahmad, S., Rambla, C., Voss-Fels, K. P., & Hickey, L. T. (2022). Accelerating breeding cycles.In *Wheat improvement: Food security in a changing climate* (pp. 557–571). *Springer International Publishing Cham*. https://doi.org/10.1007/978-3-030-90673-3

Appels, R., Eversole, K., Feuillet, C., Keller, B., Rogers, J., Stein, N., . . . Manuscript Writing, T. (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome *Science*, *361*(6403), 661-+, Article eaar7191. https://doi.org/10.1126/science.aar7191

Arya, S., Sadawarte, P., & Ashish, W. (2015). Importance of damaged starch in bakery products-a review. *Starch*, *1*, 2019.

Atanda, S. A., Steffes, J., Lan, Y., Al Bari, M. A., Kim, J., Morales, M., Johnson, J. P., Saludares, R., Worral, H., & Piche, L. (2022). Multi-trait genomic prediction improves selection accuracy for enhancing seed mineral concentrations in pea. *The Plant Genome*, e20260. https://doi.org/10.1002/tpg2.20260

Atchison, J., Head, L., & Gates, A. (2010). Wheat as food, wheat as industrial substance; comparative geographies of transformation and mobility. *Geoforum*, *41*(2), 236–246. https://doi.org/10.1016/j.geoforum.2009.09.006

Azizinia, S., Mullan, D., Rattey, A., Godoy, J., Robinson, H., Moody, D., ... & Daetwyler, H. D. (2023). Improved multi-trait prediction of wheat end-product quality traits by integrating NIR-predicted phenotypes. *Frontiers in Plant Science*, *14*, 1167221. https://doi.org/10.3389/fpls.2023.1167221

Bartholomé, J., Prakash, P. T., & Cobb, J. N. (2022). Genomic Prediction: Progress and Perspectives for Rice Rice Improvement. *Genomic Prediction of Complex Traits: Methods and Protocols*, 569-617. https://doi.org/10.1007/978-1-0716-2205-6_21

Bassi, F. M., Bentley, A. R., Charmet, G., Ortiz, R., & Crossa, J. (2016). 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), *plantgenome2016*.01.0005. 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

Bernardo, R. (2002). Breeding for quantitative traits in plants (Vol. 1). Stemma press Woodbury.

Beyene, Y., Gowda, M., Olsen, M., Robbins, K. R., Pérez-Rodríguez, P., Alvarado, G., & Crossa, J. (2019). Empirical comparison of tropical maize hybrids selected through genomic and phenotypic selections. *Frontiers in plant science*, *10*, 1502. https://doi.org/10.3389/fpls.2019.01502

Bilsborrow, P., Cooper, J., Tétard-Jones, C., Średnicka-Tober, D., Barański, M., Eyre, M., Schmidt, C., Shotton, P., Volakakis, N., & Cakmak, I. (2013). The effect of organic and conventional management on the yield and quality of wheat grown in a long-term field trial. *European Journal of Agronomy*, *51*, 71–80. https://doi.org/10.1016/j.eja.2013.06.003

Bhatta, M., Gutierrez, L., Cammarota, L., Cardozo, F., Germán, S., Gómez-Guerrero, B., ... & 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

Browning, B. L., Zhou, Y., & Browning, S. R. (2018). A one-penny imputed genome from next-generation reference panels. *The American Journal of Human Genetics*, *103*(3), 338–348.

Calvert, M., Evers, B., Wang, X., Fritz, A., & Poland, J. (2020). Breeding Program Optimization for Genomic Selection in Winter Wheat. *BioRxiv*, 2020.10. 07.330415. https://doi.org/10.1101/2020.10.07.330415

Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). NbClust: An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, *61*, 1–36. https://doi.org/10.18637/jss.v061.i06

Chen, C. J., & Zhang, Z. (2018). iPat: Intelligent prediction and association tool for genomic research. *Bioinformatics*, *34*(11), 1925–1927. https://doi.org/10.1093/bioinformatics/bty015

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

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., De Los Campos, G., ... & Varshney, R. K. (2017). Genomic selection in plant breeding: methods, models, and perspectives. *Trends in plant science*, *22*(11), 961-975. https://doi.org/10.1016/j.tplants.2017.08.011

Dawson, J. C., Endelman, J. B., Heslot, N., Crossa, J., Poland, J., Dreisigacker, S., ... & Jannink, J. L. (2013). The use of unbalanced historical data for genomic selection in an international wheat breeding program. *Field Crops Research*, *154*, 12-22. https://doi.org/10.1016/j.fcr.2013.07.020

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. *The Plant Genome*, *4*(3). https://doi.org/10.3835/plantgenome2011.08.0024

Fradgley, N., Gardner, K.A., Bentley, A.R., Howell, P., Mackay, I.J., Scott, M.F., Mott, R., & Cockram, J. (2023). Multi-trait ensemble genomic prediction and simulations of recurrent selection highlight the importance of complex trait genetic architecture for long-term genetic gains in wheat. In Silico Plants, 5(1), diad002. https://doi.org/10.1093/insilicoplants/diad002

Gill, H., Brar, N., Halder, J., Hall, C., Seabourn, B., Chen, Y., Amand, P., Bernardo, A., Glover, K., Turnipseed, B., & Sehgal, S. (2023). Multi-trait genomic selection improves the prediction accuracy of end-use quality traits in hard winter wheat. *The Plant Genome*. https://doi.org/10.1002/tpg2.20331

Glaubitz, J. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q., & Buckler, E. S. (2014). TASSEL-GBS: a high-capacity genotyping by sequencing analysis pipeline. *PloS One*, *9*(2), e90346. https://doi.org/10.1371/journal.pone.0090346

Guo, X., Sarup, P., Jahoor, A., Jensen, J., & Christensen, O. F. (2022). Metabolomic-Genomic prediction drastically improves prediction accuracy of breeding values in crop breeding. *bioRxiv*, 2022-08. https://doi.org/10.1101/2022.08.03.502591

Guo, J., Pradhan, S., Shahi, D., Khan, J., Mcbreen, J., Bai, G., Murphy, J. P., & Babar, M. A. (2020). Increased prediction accuracy using combined genomic information and physiological traits in a soft wheat panel evaluated in multi-environments. *Scientific Reports*, *10*(1), 7023. https://doi.org/10.1038/s41598-020-63919-3

Guo, T., Yu, X., Li, X., Zhang, H., Zhu, C., Flint-Garcia, S., ... & Yu, J. (2019). Optimal designs for genomic selection in hybrid crops. *Molecular plant*, *12*(3), 390-401. https://doi.org/10.1016/j.molp.2018.12.022

Haikka, H., Knürr, T., Manninen, O., Pietilä, L., Isolahti, M., Teperi, E., Mäntysaari, E. A., & Strandén, I. (2020). Genomic prediction of grain yield in commercial Finnish oat (Avena sativa) and barley (Hordeum vulgare) breeding programmes. *Plant Breeding*, *139*(3), 550–561. https://doi.org/10.1111/pbr.12807

Hayes, B. J., Bowman, P. J., Chamberlain, A. J., & Goddard, M. E. (2009). Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of Dairy Science*, *92*(2), 433–443. https://doi.org/10.3168/jds.2008-1646

Hoffstetter, A., Cabrera, A., Huang, M., & Sneller, C. (2016). Optimizing Training Population Data and Validation of Genomic Selection for Economic Traits in Soft Winter Wheat. *G3 Genes|Genomes|Genetics*, *6*(9), 2919–2928. https://doi.org/10.1534/g3.116.032532

Hogg, A. C., Sripo, T., Beecher, B., Martin, J. M., & Giroux, M. J. (2004). Wheat puroindolines interact to form friabilin and control wheat grain hardness. *Theoretical and Applied Genetics*, *108*, 1089-1097. https://doi.org/10.1007/s00122-003-1518-3

Holland, J. B., Nyquist, W. E., Cervantes-Martínez, C. T., & Janick, J. (2003). Estimating and interpreting heritability for plant breeding: an update. *Plant breeding reviews*, *22*. https://doi.org/ 10.1002/9780470650202

Hua, S., Dal-Bianco, M., & Chen, Z.-H. (2022). Crop Yield and Quality Response to the Interaction Between Environment and Genetic Factors. *Frontiers in Genetics*, *13*, 823279. https://doi.org/10.3389/fgene.2022.823279

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), e20034. https://doi.org/10.1002/tpg2.20034

Jarquín, D., Lemes da Silva, C., Gaynor, R. C., Poland, J., Fritz, A., Howard, R., Battenfield, S., & Crossa, J. (2017). Increasing genomic-enabled prediction accuracy by modeling genotype× environment interactions in Kansas wheat. *The Plant Genome*, *10*(2), *plantgenome*2016.12.0130. https://doi.org/10.3835/plantgenome2016.12.0130

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

John, J. A., & Williams, E. R. (1995). Resolvable row-column designs. In *Cyclic and Computer-Generated Designs* (pp. 107-129). Springer US.

Jubair, S., & Domaratzki, M. (2023). Crop genomic selection with deep learning and environmental data: A survey. *Frontiers in Artificial Intelligence*, *5*. https://doi.org/10.3389/frai.2022.1040295

Juliana, P., Singh, R. P., Poland, J., Mondal, S., Crossa, J., Montesinos-López, O. A., Dreisigacker, S., Pérez-Rodríguez, P., Huerta-Espino, J., & Crespo-Herrera, L. (2018). Prospects and challenges of applied genomic selection. A new paradigm in breeding for grain yield in bread wheat. *The Plant Genome*, *11*(3), 180017. https://doi.org/10.3835/plantgenome2018.03.0017

Katyal, M., Singh, N., Virdi, A. S., Kaur, A., Chopra, N., Ahlawat, A. K., & Singh, A. M. (2017). Extraordinarily soft, medium-hard and hard Indian wheat varieties: composition, protein profile, dough and baking properties. *Food Research International*, *100*, 306-317.https://10.1016/j.foodres.2017.08.050

Kiszonas, A. M., Fuerst, E. P., & Morris, C. F. (2013). Wheat arabinoxylan structure provides insight into function. *Cereal Chemistry*, *90*(4), 387–395.https://doi.org/10.1094/CCHEM-02-13-0025-FI

Kweon, M., Slade, L., & Levine, H. (2011). Solvent retention capacity (SRC) testing of wheat flour: Principles and value in predicting flour functionality in different wheat-based food processes and in wheat breeding. A review. *Cereal Chemistry*, *88*(6), 537–552. https://doi.org/10.1094/CCHEM-07-11-0092

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*, 2719-2731. https://doi.org/10.1007/s00122-018-3186-3

Larkin, D. L., Lozada, D. N., & Mason, R. E. (2019). Genomic selection—Considerations for successful implementation in wheat breeding programs. *Agronomy*, *9*(9), 479. https://doi.org/10.3390/agronomy9090479

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform [Article]. *Bioinformatics*, *25*(14), 1754-1760. https://doi.org/10.1093/bioinformatics/btp324

Lozada, D. N., & Carter, A. H. (2020). Genomic Selection in Winter Wheat Breeding Using a Recommender Approach. *Genes (Basel)*, *11*(7). https://doi.org/10.3390/genes11070779

Medina, C. A., Kaur, H., Ray, I., & Yu, L.-X. (2021). Strategies to increase prediction accuracy in genomic selection of complex traits in alfalfa (Medicago sativa L.). *Cells*, *10*(12), 3372. https://doi.org/10.3390/cells10123372

Meuwissen, T. H., Hayes, B. J., & Goddard, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*(4), 1819–1829. https://doi.org/10.1093/genetics/157.4.1819

Montesinos López, O. A., Montesinos López, A., & Crossa, J. (2022). Overfitting, model tuning, and evaluation of prediction performance. In *Multivariate statistical machine learning methods for genomic prediction* (pp. 109-139). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-89010-0_4

Montesinos-López, A., Runcie, D. E., Ibba, M. I., Pérez-Rodríguez, P., Montesinos-López, O. A., Crespo, L. A., Bentley, A. R., & Crossa, J. (2021). Multi-trait genomic-enabled prediction enhances accuracy in multi-year wheat breeding trials. *G3*, *11*(10), *jkab*270. https://doi.org/10.1093/g3journal/jkab270

Nelson, M. C., Ingram, S. E., Dugmore, A. J., Streeter, R., Peeples, M. A., McGovern, T. H., Hegmon, M., Arneborg, J., Kintigh, K. W., & Brewington, S. (2016). Climate challenges, vulnerabilities, and food security. *Proceedings of the National Academy of Sciences*, *113*(2), 298–303. https://doi.org/10.1073/pnas.1506494113

Poland, J., Endelman, J., Dawson, J., Rutkoski, J., Wu, S., Manes, Y., Dreisigacker, S., Crossa, J., Sánchez-Villeda, H., & Sorrells, M. (2012). Genomic selection in wheat breeding using genotyping-by-sequencing. *The Plant Genome*, *5*(3). https://doi.org/10.3835/plantgenome2012.06.0006

Perdry, H., Dandine-Roulland, C., Bandyopadhyay, D., & Kettner, L. (2020). gaston: Genetic Data Handling (QC, GRM, LD, PCA) & Linear Mixed Models (1.5. 7).

Puhr, D. P., & D'Appolonia, B. L. (1992). Effect of baking absorption on bread yield, crumb moisture, and crumb water activity. *Cereal Chemistry*, *69*, 582–582.

Pyler, E. J. (1979). Physical and chemical test methods. *Baking Science and Technology*, *2*, 891–895.

Robertson, C. D., Hjortshøj, R. L., & Janss, L. L. (2019). Genomic selection in cereal breeding. *Agronomy*, *9*(2), 95. https://doi.org/10.3390/agronomy9020095

Sallam, A. H., Conley, E., Prakapenka, D., Da, Y., & Anderson, J. A. (2020). Improving prediction accuracy using multi-allelic haplotype prediction and training population optimization in wheat. *G3: Genes, Genomes, Genetics*, *10*(7), 2265–2273. https://doi.org/10.1534/g3.120.401165

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*, 831020. https://doi.org/10.3389/fgene.2022.831020

Sandhu, K. S., Aoun, M., Morris, C. F., & Carter, A. H. (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

Sapirstein, H., Wu, Y., Koksel, F., & Graf, R. (2018). A study of factors influencing the water absorption capacity of Canadian hard red winter wheat flour. *Journal of Cereal Science*, *81*, 52–59. https://doi.org/10.1016/j.jcs.2018.01.012

Schneider, U. A., Havlík, P., Schmid, E., Valin, H., Mosnier, A., Obersteiner, M., Böttcher, H., Skalskỳ, R., Balkovič, J., & Sauer, T. (2011). Impacts of population growth, economic development, and technical change on global food production and consumption. *Agricultural Systems*, *104*(2), 204–215. https://doi.org/10.1016/j.agsy.2010.11.003

Schulthess, A. W., Wang, Y., Miedaner, T., Wilde, P., Reif, J. C., & Zhao, Y. (2016). Multiple-trait-and selection indices-genomic predictions for grain yield and protein content in rye for feeding purposes. *Theoretical and Applied Genetics*, *129*, 273-287. https://doi.org/10.1007/s00122-015-2626-6

Shahi, D., Guo, J., Pradhan, S., Khan, J., Avci, M., Khan, N., McBreen, J., Bai, G., Reynolds, M., & Foulkes, J. (2022). Multi-trait genomic prediction using in-season physiological parameters increases prediction accuracy of complex traits in US wheat. *BMC Genomics*, *23*(1), 1–13. https://doi.org/10.1186/s12864-022-08487-8

Sun, J., Rutkoski, J. E., Poland, J. A., Crossa, J., Jannink, J. L., & 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), plantgenome2016-11. https://doi.org/10.3835/plantgenome2016.11.0111

Sweeney, D. W., Sun, J., Taagen, E., & Sorrells, M. E. (2019). Genomic selection in wheat. In Applications of genetic and genomic research in cereals (pp. 273-302). *Woodhead Publishing*. https://doi.org/10.1016/B978-0-08-102163-7.00013-2

Symes, K. J. (1969). Influence of a gene causing hardness on the milling and baking quality of two wheats. *Australian Journal of Agricultural Research*, *20*(6), 971-979. https://doi.org/10.1071/AR9690971

Timothy, J. T., Chicilo, F., Wiens, D. J., & Reaney, M. J. (2022). Beyond Bread and Beer: Value-Added Products from Wheat. In *Wheat*. IntechOpen.VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, *91*(11), 4414–4423. https://doi.org/10.5772/intechopen.102603

Tomar, V., Singh, D., Dhillon, G. S., Chung, Y. S., Poland, J., Singh, R. P., ... & Kumar, U. (2021). Increased predictive accuracy of multi-environment genomic prediction model for yield and related traits in spring wheat (Triticum aestivum L.). *Frontiers in Plant Science*, *12*, 720123. https://doi.org/10.3389/fpls.2021.720123

VanRaden, P. M. (2008). Efficient Methods to Compute Genomic Predictions [Article]. *Journal of Dairy Science*, *91*(11), 4414-4423. https://doi.org/10.3168/jds.2007-0980

Wickham, H. (2006). An introduction to ggplot: An implementation of the grammar of graphics in R. *Statistics*, 1–8.

Williams, E., Piepho, H., & Whitaker, D. (2011). Augmented p-rep designs. *Biometrical Journal*, *53*(1), 19–27. https://doi.org/10.1002/bimj.201000102

Winn, Z. J., Larkin, D. L., Lozada, D. N., DeWitt, N., Brown-Guedira, G., & Mason, R. E. (2023). Multivariate genomic selection models improve prediction accuracy of agronomic traits in soft red winter wheat. *Crop Science*, *63*(4), 2115-2130. https://doi.org/10.1002/csc2.20994

Xiong, Z. Q., Freney, J. R., Mosier, A. R., Zhu, Z. L., Lee, Y., & Yagi, K. (2008). Impacts of population growth, changing food preferences and agricultural practices on the nitrogen cycle in East Asia. *Nutrient Cycling in Agroecosystems*, *80*, 189–198. https://doi.org/10.1007/s10705-007-9132-4

Yadav, S., Wei, X., Joyce, P., Atkin, F., Deomano, E., Sun, Y., ... & Voss-Fels, K. P. (2021). Improved genomic prediction of clonal performance in sugarcane by exploiting non-additive genetic effects. *Theoretical and Applied Genetics*, *134*(7), 2235-2252. https://doi.org/10.1007/s00122-021-03822-1

Zghal, M. C., Scanlon, M. G., & Sapirstein, H. D. (2001). Effects of flour strength, baking absorption, and processing conditions on the structure and mechanical properties of bread crumb. *Cereal Chemistry*, *78*(1), 1–7. https://doi.org/10.1094/CCHEM.2001.78.1.1

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, 10(5), 1695-1704. https://doi.org/10.1016/j.cj.2022.03.010

Zhang, H., Yin, L., Wang, M., Yuan, X., & Liu, X. (2019). Factors affecting the accuracy of genomic selection for agricultural economic traits in maize, cattle, and pig populations. *Frontiers in Genetics*, *10*, 189. https://doi.org/10.3389/fgene.2019.00189

Zystro, J., Peters, T., Miller, K., & Tracy, W. F. (2021). Classical and genomic prediction of hybrid sweet corn performance in organic environments. *Crop Science*, *61*(3), 1698-1708. https://doi.org/10.1002/csc2.20400

Chapter 4-Improving Genomic Prediction Accuracy for Wheat End Use Quality Using Models with Informative Markers as Fixed Effects

4.1 Summary

The end-use quality of wheat is one determinant factor of grain price in the international market and affects the final characteristics of baked products. Improving end-use quality traits through phenotypic selection at the early stages of variety development is challenging because these traits are time-consuming to phenotype and require larger flour samples. Genomic prediction can be employed to select quality traits in earlier generations. In this study, 790 hard wheat genotypes from Colorado State University's (CSU) wheat breeding program were evaluated across multiple years (2014-2022) and locations. Phenotyping of quality traits was conducted at the CSU Wheat Quality Lab, while genotyping was performed using the GBS method and KASP markers specific to the *Bx7*$^{OE+8}$ *HMW-GS*. Three models, including one with genome-wide association study (GWAS), identified markers as fixed effects, *Bx7*$^{OE+8}$ allele as fixed effects, and another without fixed effects were employed. These models utilized *Bayesian Lasso* and *GBLUP* methods and were validated through the cross and forward-validation frameworks. The highest prediction accuracy (r = 0.82) was observed for BakeMT using both models with fixed effects, and r = 0.80 for MixoT from the first forward validation scenario. The lowest accuracy (r = 0.21) was noted in *GBLUP* models with *Bx7*$^{OE+8}$ *HMW-GS* as a fixed effect for loaf volume from the model without fixed effects in the second forward validation approach. Overall, models with fixed effects outperformed those without fixed effects, underscoring the importance of including known loci or GWAS-identified markers as fixed effects to enhance prediction accuracy.

4.2 Introduction

Wheat (*Triticum aestivum* L.) is a major source of cereal protein for both humans and livestock (Moore et al., 2016). The wheat endosperm consists of approximately 8% to 20% proteins (Žilić et al., 2011). Gluten proteins, the primary storage proteins in wheat grain, are mainly located in the endosperm, which is milled to produce white flour. These proteins form large polymers during grain development. When mixed with water, these polymers create a continuous network essential for the dough's structure (Tosi et al., 2011).

Gliadins and glutenins are the two main types of gluten proteins (Shewry et al., 2002; Wieser, 2007). Glutenins are larger water-insoluble proteins rich in cysteine (Shewry and Tatham, 1990; Payne et al., 1981). They are classified into two main groups: high molecular weight glutenins *(HMW-GS)* and low molecular weight glutenins (*LMW-GS*) (Payne et al., 1981). High molecular weight glutenin subunits are crucial for gluten structure, providing strength and elasticity to the dough (Shewry et al., 2000), whereas *LMW-GS* are believed to participate in the formation of disulfide bonds between *HMW-GS* and gliadins (Shewry et al., 2000). The relative proportions and specific types of gliadins and glutenins vary among wheat cultivars and significantly affect the wheat's baking quality. For example, wheat cultivars with a higher proportion of *HMW-GS* to *LMW-GS* and gliadins typically have stronger gluten and better bread-baking quality (De Santis et al., 2017; Singh et al., 1990).

The *HMW-GS* were located at *Glu-A1, Glu-B1*, and *Glu-D1* loci on the long arms of group 1 chromosomes (1A, 1B, and 1D) (Payne, 1987). There are two linked genes at these loci that encode for two distinct types of *HMW-GS*, the *x* and *y*-type subunits (Dai et al., 2018; Kumar et al., 2019; Payne et al., 1981; Shewry et al., 1992, 2003). These x and y-type subunits can be distinguished based on their electrophoretic mobility in sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE), whereas *x*-type subunits typically exhibit slower mobility and have higher molecular weight than *y*-type subunits leading to multiple alleles at the *Glu-B1* loci (Kumar et al., 2019; Mclntosh et al., 2008).

The *Glu-B1* gene, in particular, displays the most diversity among the three *Glu-1* loci and has a strong influence on the bread-making quality of wheat (Shewry et al., 1992; Anderson and Greene, 1997). This location produces several homoeologous x-type subunits: *Bx6, Bx7* (with variants like *Bx7*, $Bx7^{OE}$), Bx14, Bx17, Bx18, Bx20.* The $Bx7^{OE}$+8 *HMW-GS* subunit *Glu-B1al* allele (hereafter $Bx7^{OE}$) arises when *Bx7* is overexpressed due to a duplication in the gene sequence (Ragupathy et al., 2008; Gupta et al., 1991; Butow et al., 2003). Genotypes that lack the regular *Bx7* subunit show a lower quality of the gluten network's micro-structure, resulting in inferior baking quality (Gao et al., 2017; Wang et al., 2013). Conversely, wheat varieties with the $Bx7^{OE}$ subunit exhibit superior bread-baking quality compared with those containing the *Bx7* subunit (Chen et al., 2019; Gianibelli et al., 2001; Li et al., 2020). There is a reliable kompetitive allele-specific PCR (KASP) assay for the accurate genotyping of such alleles.

Phenotypic selection for baking traits such as BakeMT, MixoT, and BakeV has been challenging due to the time-intensive nature of phenotyping, the brief interval between winter wheat harvest and planting, and the larger flour sample size required, in particular for a baking test. To alleviate these challenges, genomic prediction can be used as a potential alternative for improving quality traits.

For many traits to which genomic prediction has been applied so far, improving prediction accuracy has been an important goal of many researchers (Crossa et al., 2017; Kumar et al., 2020). Several factors influence prediction accuracy, including population size, marker density, heritability, genetic architecture, genetic relatedness, and the choice of a genomic prediction model

(Crossa et al., 2017; Meuwissen et al., 2001; Wang et al., 2014). The choice of an appropriate genomic prediction model depends on the data type, model assumptions, and the objective of the analysis (De los Campos et al., 2013; Heslot et al., 2012; Bernardo and Yu, 2007). Genomic best linear unbiased prediction *(GBLUP)* is the most widely used genomic prediction model (Meuwissen et al., 2001; De los Campos et al., 2013). The *GBLUP* model assumes that markers follow a normal distribution with variance proportional to their relationship matrix (Habier et al., 2007). The *Bayesian Lasso* method includes a penalization step that serves to prevent all markers from having the same effect during the model training phase (Meuwissen et al., 2001; Habier et al., 2011). Adjusting factors that affect prediction accuracy can improve prediction accuracy (Desta and Ortiz, 2014; Isidro et al., 2015). For example, incorporating markers with major effects as fixed effects has increased genomic prediction accuracy in some cases (Spindel et al., 2016; Norman et al., 2018; Rice and Lipka, 2019; Rutkoski et al., 2014; Sehgal et al., 2020). However, fixed effects may not always improve accuracy depending on the genetic architecture of the target trait (Bernardo, 2014).

In this study, three end-use quality traits, including BakeV, BakeMT, and MixoT, were predicted using *Bayesian Lasso* and *gBLUP* methods adjusted in three forms: (1) with $Bx7^{OE}$ marker data from KASP assay as a fixed effect (2) with significant GWAS-identified markers as fixed effects, and (3) with no fixed effects beyond the default intercept. Prediction accuracy within each method and among methods was compared in cross- and forward-validations to assess the influence of model type and inclusion of fixed effects. Examining multiple model formulations provides insight into optimal prediction strategies to identify the ideal approach for prediction.

4.3 Materials and Methods

4.3.1 Germplasm

A total of 790 hard winter wheat genotypes (experimental lines and check varieties) generated from the Colorado State University (CSU) wheat breeding program were used in this study. The genotypes were evaluated in four different breeding trials, including the CSU Elite Trial (ELITE) and the Advanced Yield Nursery conventionally derived (AYN), and doubled haploid derived (AYND) genotypes, from 2014 to 2022 at CSU wheat breeding trial sites in Colorado, USA. A detailed description of the number of genotypes in each year-location-trial combination is given in (Table 4.1).

4.3.2 Experimental designs and trait measurements

The trials (CSU Elite, AYNs, and AYNDs) were arranged and planted in randomized, row-column designs with partial replication, following the methodologies outlined by Williams et al. (2011). In the CSU Elite Trials, half of the entries were replicated twice in each location, while the other half were replicated once at each location. In the AYNs, approximately one-seventh of the entries were replicated a second time at each location. Genotypes included in this study were evaluated at CSU wheat breeding sites in the Great Plains wheat-growing region of eastern Colorado, USA. All trials were grown "on farm," except for trials conducted at the Agricultural Research, Development, and Educational Center (ARDEC) in Fort Collins, CO, Plainsman Research Center in Walsh, CO, and the United States Department of Agriculture – Agriculture Research Service Central Great Plains Center in Akron, CO. The agronomic and crop management practices matched those used by the grower cooperators and varied according to the standard practices at each location. All genotypes were planted in a six-row plot, 1.5 m wide and 3.7 m long, at a seeding rate of approximately 1.73 million seeds ha^{-1}. All six rows of each plot were harvested, and a cleaned sample of the grain was used for subsequent quality analyses.

The end-use quality traits were phenotyped at the Colorado State University wheat quality lab located in Fort Collins, CO, following the procedures of the American Association of Cereal Chemists International (AACC International, 2010). For each genotype, a 300 g grain sample was assessed for protein and moisture with the Foss DA1650 NIRS (Eden Prairie, MN, USA) and then placed in 1 L Nalgene bottles (Thermo Fisher Scientific; Waltham, MA, USA). Samples were tempered to 15% moisture, and milling was done using a modified Brabender Quadrumat Senior mill Brabender Instruments, NJ, USA). This flour was then rescanned using NIRS to determine flour protein and moisture.

A 10 g Mixograph (National Manufacturing Co., Lincoln, NE, USA) was utilized for Mixograph tests on each sample that was baked (AACC, International, 2010). The initial water absorption of the flour was predicted based on the flour protein concentration determined through Near-Infrared Reflectance Spectroscopy (NIRS). The formula used for predicting water absorption is predicted as follows:

$$\text{Water absorption} = (42.7 + (1.69 \text{ x sample flour protein concentration}))$$

where the protein concentration was measured in the flour at a moisture level of 14% MB (Finney and Shogren, 1972). The predicted water absorption of the sample was then served as a starting point in the Mixograph. The Mixograph displayed a curve representing the dough's behavior during mixing. The width and slope of the curve after its peak were used to determine the MixoT, which was visually scored on a scale from 0 = poor to 6 = excellent.

BakeMT was first estimated as the Mixograph curve's highest point. As mixing time is not a perfect estimate, the actual BakeMT was visually assessed using the 100 g pup loaf bake test by observing when the dough reached its optimal form in the mixing bowl (National Manufacturing Co., Lincoln, NE, USA). Once an optimal dough was formed, the mixer was stopped, and the BakeMT was recorded. For bake loaf volume assessment, 100 g pup loaves were utilized (AACC, International, 2010). The composition of the baking recipe included 100 g white flour (14% MB), 6 g sugar (sucrose), 1.5 g salt (NaCl), 3.0 g Crisco®, and 2.12 g instant active yeast (Red Star®) and 5 mg DohTone® (fungal alpha amylase). The doughs were prepared based on a 6-minute bake schedule and a 90-minute fermentation. The fermentation cabinet was set at 86 °F with a relative humidity of 90-95 %. Following fermentation, samples were baked at 400 °F for 25 minutes. Upon completion of the baking process, the volume of the baked loaves was measured by rape seed displacement.

4.3.3 Phenotypic data analysis

For year-location-trial combinations, best linear unbiased estimates (BLUEs) were calculated using the *ASREML-R* package in R statistical software (Butler et al., 2009; R Core Team, 2013). The mixed linear model applied for estimating BLUEs across year-location-trial combinations was calculated as follows:

$$y_{ijk} = \mu + G_i + e{:}r_{jk} + e{:}c_{jl} + \varepsilon_{ijkl}$$

Where y_{ijk} is the response variable for i^{th} level of genotype, in the j^{th} level of the row, in the k^{th} level of the column; μ is the overall mean; G_i is the fixed genotype effect; $e{:}r_{jk}$ is the random row within the trial effect $(r_j \sim N(0, \sigma_j^2)$; $e{:}c_{jl}$ is the random column within the trial effect $(c_k \sim N(0, \sigma_k^2)$; and ε_{ijkl} is the residual error term $\varepsilon_{ijkl} \sim N(0, \sigma_\varepsilon^2)$.

Summary statistics were derived to evaluate the data distribution, and the Pearson correlation coefficient was estimated and displayed using the *psych* package in R (Revelle and Revelle, 2015). Variance components were estimated to derive heritability estimates for each trait using the mixed linear model by assigning genotypes as a random effect using *ASREML-R* (Butler et al., 2009). The mixed linear model applied for estimating variance components was formed as follows:

$$y = Xb + Zu + e$$

Where *X* and *Z* are design matrices; *b* is a vector of fixed effects; *u* is a vector of random genetic effects (BLUPs) with distribution $u \sim N(0, \sigma_g^2 K)$ where K is the kinship matrix and σ_g^2 is the genetic variance; *e* is a vector of random residuals with distribution $e \sim N(0, \sigma_e^2 I)$ where σ_e^2 is the residual variance, and *I* is the identity matrix. The genomic heritability of the traits was calculated using the following formula:

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

where σ_g^2 is the genetic variance and σ_e^2 is the residual error associated with the trait.

Pearson correlation analysis was calculated among BakeMT, MixoT, and BakeV and visualized using the *psych* package in R (Revelle and Revelle, 2017).

4.3.4 Genotyping

Genomic DNA was isolated from one-week-old wheat leaves using a 96-well plate and the MagMax-14™ DNA extraction kit (ThermoFisher Scientific, MA, USA). The DNA concentration was subsequently determined with PicoGreen (Thermo Fisher Scientific, MA, USA) in accordance with the manufacturer's instructions. Genotyping-by-sequencing (GBS) libraries were prepared using the *PstI-MspI* restriction enzyme combination (Poland et al., 2012) and pooled together at a 384 plex. Libraries were then sequenced on the Illumina HiSeq 2500 system at the University of Illinois. The identification of single nucleotide polymorphisms (SNPs) was done using *TASSEL GBSv2* pipeline (Glaubitz et al., 2014), applying a 64-base *k-mer* size and setting the minimum *k-mer* count to five. The SNPs were subsequently aligned with the International Wheat Genome

Sequencing Consortium's 'Chinese Spring' wheat reference genome v1.0 (Appels et al., 2018) using the Burrows-Wheeler Aligner version 0.7.10 (Li and Durbin, 2009).

Genotypic data quality control was done on the SNP data extracted through the *TASSEL* tool. Entries with more than 50% incomplete data or with an average heterozygosity surpassing 30% were discarded. The criteria for the retained SNP markers included being biallelic, having above 5% minor allele frequency, less than 10% incomplete data, and no more than 10% average heterozygosity. Unaligned SNPs, insertions, and deletions were removed prior to imputation. The *Beagle 5.4* algorithm (Browning et al., 2018) was used to impute missing marker data. After the filtration processes, a total of 23,130 SNPs were made available for analysis.

Genotyping of the $Bx7^{OE}$ marker was performed using the kompetitive allele-specific PCR (KASP) methodology. Initially, genomic DNA was extracted from selected wheat lines known to carry variations in the $Bx7^{OE}$ gene. Using the KASP assay, allele-specific primers were applied to discern the genotypic differences based on the presence or absence of a 43 bp insertion/deletion (Table 4.2). The PCR amplification was followed by a fluorescent-based detection to differentiate between the homozygous and heterozygous alleles. The fluorescent signals were read post-PCR, and the data obtained were analyzed to determine the genotype of each sample.

4.3.5 Marker data for inclusion as fixed effects

Marker data included as fixed effects in the genomic prediction models were prepared from two sources: GWAS-identified markers and $Bx7^{OE}$ KASP marker data for the glutenin allele. The $Bx7^{OE}$ KASP marker was developed from historical genotypic data derived by KASP assay by the Colorado State University wheat breeding program. The KASP assay provided the data in which individuals were denoted as 0 and 1; one refers to a genotype carrying $Bx7^{OE}$ (either homozygous or heterozygous state), and 0 signifies a genotype without *Bx7*^{OE}. Of the 790 individuals included in this study, 158 genotypes carried *Bx7*^{OE,} while 632 did not carry *Bx7*^{OE}.

A Genome-Wide Association Study (GWAS) was performed utilizing the *GWAS* function of the *rrBLUP* package in R (Endelman, 2011), aimed at identifying significant genetic markers for later inclusion as fixed effects in genomic prediction models. Prior to the GWAS, LD pruning was applied to the GBS marker data, comprising 23,120 markers, using an LD threshold of 0.9. Following LD pruning, a total of 9,884 SNP markers were retained as genotypic data for subsequent GWAS analysis. This preprocessing step aimed to reduce redundancies within the marker data and minimize the risk of multicollinearity, thereby ensuring the robustness of the GWAS findings and enhancing the reliability of the associations detected between markers and traits of interest (Vilhjálmsson et al., 2015).

The GWAS analysis was employed using a Q+K model (Yu et al., 2006) designed to control confounding effects attributed to population stratification. This was achieved by incorporating the first three principal components to delineate population structure alongside the kinship matrix for refined adjustment of individual relatedness. Significant marker-trait associations were declared based on the Bonferroni multiple-testing adjusted significance threshold. The Bonferroni significance threshold was calculated as:

$$\alpha = \frac{0.05}{n}$$

Where 0.05 is the alpha value for each test, and n is the number of markers used in the GWAS analysis. All marker-trait associations that showed a -log10(p) values greater than the Bonferroni threshold (5.67) were later included as fixed effects in the genomic prediction models.

For each permission of cross validation, the training population was subjected to a GWAS, and the significant MTAs identified in the training population were used as covariates in the testing population. This was done to avoid the "insider trading" phenomenon (Verges et al., 2021).

4.3.6 Genomic prediction

Three genomic prediction models were employed in *Bayesian Lasso* and *GBLUP* approaches, such as incorporating *Bx7^{OE}* KASP marker data as a fixed effect, with GWAS-identified markers as fixed effects, and without fixed effects. The *Bayesian Lasso* models were implemented with the *BGLR* package (version 1.1.1) in R (Pérez & de los Campos, 2014). This method choice was chosen for this analysis due to its ability to handle complex genetic relationships through Gaussian Random Field Models (*GRM*) for random genetic effects and potentially reduce model complexity by performing variable selection via the lasso penalty. The model can be conceptually represented by the following formula:

$$y = X\beta + Z\gamma + \varepsilon$$

Where y represents the response variable, X is the design matrix for the random genetic effects captured from the genotype data using the *GRM* within *BGLR*. β refers to the estimated coefficients for these effects, obtained through a Bayesian framework which accounts for uncertainty in their estimation. The lasso penalty can shrink some of these coefficients towards zero, performing variable selection among the genetic variants. Z is the design matrix for the fixed effects, captured from significant markers identified through GWAS and the *Bx7^{OE}* KASP marker data. γ represents the estimated coefficients for these fixed effects. ε denotes the error term, typically assumed to be normally distributed with mean 0 and variance $\sigma^2 e$ (represented as $\varepsilon \sim N(0, \sigma^2 e)$.

Bayesian Lasso models without additional fixed effects (only the intercept as a fixed effect) were formulated as follows:

$$y = \beta_0 + Zu + \varepsilon$$

Where y is the response variable. β_0 represents the intercept which is a constant term representing the average response of the phenotype in the absence of any genetic variation effects (the overall mean). Z (design matrix), this matrix captures information about the genotypes of individual subjects. u (random effect coefficients): captures random effects associated with unobserved genetic variations, assumed to be random and vary across individuals. The Laplace prior applied to the coefficients related to u (promoting sparsity by shrinking less influential markers towards zero). ε is the error term, accounts for all unexplained variation in the phenotype that is not captured by the model, assumed to be normally distributed with a mean of zero (Pérez & De los Campos, 2014). The model underwent 10,000 iterations of the Markov Chain Monte Carlo (MCMC) sampling process, with a burn-in period of 1,000 iterations.

The same models described in the *Bayesian Lasso* method were also fit under the *GBLUP* method using the *mmer* function from the *sommer* package in R (Covarrubias-Pazaran, 2016). The *GBLUP* method leverages marker information and a genomic relationship matrix (*GRM*) to estimate the genetic merit of individuals for the end-use quality traits. The *GBLUP* model with GWAS-derived markers or *Bx7*^{OE} as fixed effects were employed using the following formula.

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

Where y is the response variable; X is the design matrix for fixed effects. This matrix encodes the genotypes of significant markers identified through GWAS or KASP assays for each individual. β is the vector of fixed effects coefficients. Z is the design matrix for random, this is typically a vector of 1s, associating each individual with the random effect term (genetic merit); u represents the vector of random effects coefficients (genetic merit, each element represents the individual's deviation from the average phenotype due to its overall genetic makeup beyond the effects of the identified markers. The random effects (*u*) are assumed to follow a multivariate normal distribution with a mean of zero. The variance-covariance structure depends on the *GRM*, which captures the genetic similarity between individuals. *e* is the vector of residuals representing unexplained variation in the phenotype after accounting for fixed and random effects, are assumed to follow a normal distribution with a mean of zero. The variance structure of residuals might be specified based on "units" to account for potential correlations within experimental groups.

The model without additional fixed effects was fit using the following formula:

$$y = \beta_0 + Zu + e$$

Where y is the response variable, β_0 is the intercept term representing the average response of the phenotype when all other effects are zero. Z is the design matrix for random effects; u is the random effects coefficients - genetic merit, follows the multivariate normal distribution with a mean of zero, and e is the residuals, which are assumed to follow a normal distribution with a mean of zero (Covarrubias-Pazaran, 2016).

4.3.7 Cross validation

The accuracy of the predictive models was validated through the cross and forward-validation frameworks. Cross-validation was done by randomly dividing the full dataset into training (80%) and testing (20%) sets. The model was trained on each training set to predict genomic estimated breeding values (GEBVs) for the corresponding testing set. This cross-validation process was repeated 10 times with different random partitioning of the data into training and testing sets. Prediction accuracy was calculated as the Pearson correlation coefficient (*r*) between the predicted GEBVs and the BLUEs across the 10 permutations. Box plots visualizing the distribution of prediction accuracies across cross-validation iterations were generated for each model using the *ggplot2* graphics package in R (Wickham et al., 2016).

3.4.8 Forward validations

In the case of forward validation, the same 790 genotypes used in cross-validation were also used, but the training and testing set split was different. Two validation scenarios were used to see if the number of genotypes included in the validation set affects the prediction accuracy. The first validation scenario (n = 55) genotypes from ELITE trials of 2020, 2021, and 2022 were used as a testing set and the remaining genotypes (n = 735) were used as the training set. In the second validation scenario (n = 154), genotypes from three trials, including AYN, AYND, and ELITE trials from two years (2021 and 2022), were used as a testing set, and the remaining genotypes (n = 636) genotypes were used as a training set.

Models were trained using the training sets, and BLUEs of the validation set were masked. Genomic estimated breeding values (GEBVs) were then predicted for the validation set. Unlike cross-validation, no repeats or iterations were used. Prediction accuracy was calculated through correlation analysis between GEBVs and BLUEs.

4.4 Results

4.4.1 Phenotypic variations

Summary statistics were calculated for BakeV, BakeMT, and MixoT. These statistics revealed that genotypes exhibit variation for each trait (Table 4.3). The highest variation was observed in the BakeV, ranging from 689.96 cm^3 to 1184.94 cm^3 and a standard deviation of 85.81. The average BakeMT was 4.86 minutes with a standard deviation of 1.80. MixoT also presented a notable spread, with values ranging from a minimum of 0.82 to a maximum of 6.00. Among the traits analyzed, BakeMT demonstrated the highest heritability ($h^2 = 0.84 \pm 0.05$), followed by MixoT ($h^2 \pm 0.78 \pm 0.06$), and the lowest heritability was from BakeV ($h^2 = 0.64 \pm 0.06$).

Pearson's correlation analysis revealed significant correlations among all traits considered in this study (Figure 4.1). The highest correlation coefficient ($r = 0.78$, $p < 0.001$) was observed between BakeMT and MixoT, followed by the correlation between BakeV and MixoT ($r = 0.34$, $p < 0.001$). The lowest correlation ($r = 0.32$, $p < 0.001$) was observed between BakeV and BakeMT.

4.4.2 Genome-wide association study

A total of 28 significant marker-trait associations (MTAs) were identified across the three traits: 24 for BakeMT, 21 for MixoT, and 4 for BakeV (Table 4.4). There was overlap among the markers identified for these traits, resulting in 28 unique markers in total (Figure 4.2; Table 4.4). The MTAs for BakeMT and MixoT were found primarily on chromosomes 1B and 1D in regions (Figure 4.2). In addition to group one chromosomes, significant associations were also observed on chromosomes 2B, 5D, and 6D. For BakeV, only four marker-trait associations on chromosome 1D were identified.

4.4.3. Cross-validation

For MixoT, the *Bayesian Lasso* model without fixed effects showed a median prediction accuracy of $r = 0.55$. The *Bayesian Lasso* model, with *Bx7^{OE}* KASP marker data as a fixed effect, showed a mean prediction accuracy of $r = 0.62$, while the model with GWAS-identified markers as a fixed effect showed a mean prediction accuracy of $r = 0.74$. In contrast, the *GBLUP* model without fixed effects showed a mean prediction accuracy of $r = 0.49$, while the *GBLUP* model with *Bx7^{OE}* KASP marker data as a fixed effect exhibited a mean prediction accuracy of $r = 0.61$ (Figure 4.3). The model with GWAS-identified markers as a fixed effect showed a mean prediction accuracy of $r = 0.63$. Compared to the baseline model without fixed effects, the *Bayesian Lasso* model with GWAS-identified markers as fixed effect improved the mean prediction accuracy by 34%.

Comparing the model with GWAS-identified markers as a fixed effect under two methods, the *Bayesian Lasso* outperformed the *GBLUP* by a 15% margin. On the other hand, the *Bayesian Lasso* model without a fixed effect demonstrated a 12% improvement in prediction accuracy for MixoT over the *GBLUP* model without a fixed effect. In summary, coupling the *Bayesian Lasso* with the model with GWAS-identified markers as a fixed effect showed improved genomic prediction of BakeMT in cross-validation, surpassing all other modeling techniques, including *GBLUP* and *Bayesian Lasso* with *Bx7*^{OE} marker from KASP assay without a fixed effect.

For BakeMT, the *Bayesian Lasso* model with GWAS identified markers as fixed effect again demonstrated superior median prediction accuracy (r = 0.79) than the models with *Bx7*^{OE} KASP marker data as fixed effect and without fixed effect. The *Bayesian Lasso* model with *Bx7*^{OE} KASP marker as a fixed effect yielded a median prediction accuracy of r = 0.63, which is approximately 20% lower than the model with GWAS-identified markers fixed effect. However, the model with the *Bx7*^{OE} KASP marker as a fixed effect was still about 24% higher than the model without fixed effects (r = 0.51). In the Bayesian approach, the model without a fixed effect was the poorest predictor of BakeMT (Figure 4.3).

Similarly, the *GBLUP* model with *Bx7*^{OE} KASP marker data as a fixed effect showed higher median prediction accuracy (r = 0.64) when compared to the model without a fixed effect (r = 0.55). The median prediction accuracy of the *GBLUP* model with GWAS-identified markers was r = 0.67. Across all examined fixed effect models, the *Bayesian Lasso* models markedly and consistently outperformed the *GBLUP,* with a 22% improvement for predicting BakeMT. Additionally, the model with GWAS identified marker as a fixed effect proved to be the best predictor of BakeMT compared to the model with *Bx7*^{OE} KASP marker data as a fixed effect and the model without fixed effect models. For *GBLUP,* the two fixed effect models were pretty similar, whereas that is not shown in the *Bayesian* model.

For BakeV, the model with GWAS-identified markers as a fixed effect exhibited a prediction accuracy of r = 0.59 for both the *Bayesian Lasso* and *GBULP* approaches (Figure 4.3). The prediction accuracy of the *Bayesian Lasso* model with *Bx7*^{OE} KASP marker data as a fixed effect was r = 0.55. The *Bayesian Lasso* model without fixed effects yielded a prediction accuracy of r = 0.54, which was the lowest prediction accuracy for BakeV but was only nine percent lower than the top model in this framework (Figure 4.3). Bake volume prediction is not as good as the BakeMT and MixoT, regardless of the models used.

The *GBLUP* method with a model without fixed effects showed a median prediction accuracy of r = 0.57. The model with the *Bx7*^{OE} KASP marker data as a fixed effect appeared to be the poorest predictor for BakeV, with a median prediction accuracy of r = 0.53. In summary, the model with GWAS identified markers as fixed effects from both approaches emerged as the top model for predicting BakeV. Overall, the *GBLUP* model with *Bx7*^{OE} KASP marker data as a fixed effect was the poorest predictor of BakeV but not for *Bayesian Lasso*.

4.4.4 Forward-validation

Utilizing a training population of n = 55 genotypes, the *Bayesian Lasso* models with *Bx7*^{OE} KASP marker data and GWAS-identified markers as fixed effects both demonstrated superior prediction accuracy (r = 0.82) for BakeMT compared with the model without inclusion of fixed effects (Figure 4.4). The prediction accuracy represents a 10% increase over the model without a fixed effect. Similarly, for MixoT, the *Bayesian Lasso* model with GWAS identified markers as fixed effects exhibited the highest prediction accuracy (r = 0.80). The model with *Bx7*^{OE} KASP marker data as a fixed effect followed with an accuracy of r = 0.78, showing a 10% improvement over the model without fixed effects. For BakeV, only slight differences were observed among the three models tested using *Bayesian Lasso*. The model without a fixed effect had the lowest prediction accuracy (r = 0.41). In contrast, the model with *Bx7*^{OE} as a fixed effect had an accuracy of r = 0.44, and the model with GWAS-identified markers as a fixed effect had an accuracy of r = 0.46.

The *GBLUP* model generally performed similarly to the *Bayesian Lasso* method, as the inclusion of fixed effects in the models improved the prediction of BakeMT and MixoT, though not for BakeV. The *GBLUP* model with *Bx7*^{OE} KASP marker data as a fixed effect was the best predictor for BakeMT with a prediction accuracy of r = 0.75, representing a 25% improvement over the model without a fixed effect (r = 0.56) and only a three percent improvement over the model with GWAS identified markers as a fixed effect (r = 0.73). For BakeV, the *GBLUP* model without a fixed effect showed a prediction accuracy of r = 0.38, followed by the model with GWAS-identified markers as a fixed effect with r = 0.36. The least accurate predictor of BakeV was the model with *Bx7*^{OE} KASP marker data included as a fixed effect (r = 0.32).

For MixoT, the *GBLUP* model with the GWAS identified markers as a fixed effect was the top predictor, with a prediction accuracy of r = 0.80, followed by the model with *Bx7*^{OE} as a fixed effect with r = 0.77. The model without a fixed effect yielded an accuracy of r = 0.70. These results mirror those obtained under the *Bayesian Lasso* framework, with a minor difference in the prediction accuracy of the model with *Bx7*^{OE} KASP marker data, included as a fixed effect. In summary, incorporating *Bx7*^{OE} KASP marker data and GWAS-identified markers as fixed effects improved prediction accuracy for MixoT and BakeMT but did not yield a substantial improvement in BakeV prediction accuracy (Figure 4.4).

With the second forward validation approach, which had n = 154 individuals in the validation set and 636 in the training set, the *Bayesian Lasso* model with GWAS-identified markers as a fixed effect showed a higher prediction accuracy for BakeMT (r = 0.79) and MixoT (r = 0.71) (Figure 4.5). The *Bayesian Lasso* model without a fixed effect was the second-best predictor for BakeMT (r = 0.74) and MixoT (r = 0.60). The *Bayesian Lasso* model with $Bx7^{OE}$ as a fixed effect showed the lowest prediction accuracy for BakeMT (r = 0.71) and MixoT (r = 0.58). However, in the *GBLUP* method, the model with the $Bx7^{OE}$ as a fixed effect achieved the highest prediction accuracy for BakeMT (r = 0.69), marking a 12% improvement over a model with GWAS-identified markers and a 29% improvement over the model without the inclusion of fixed effects. When predicting MixoT via *GBLUP*, the model with GWAS identified markers as fixed effect ranked highest (r = 0.59), closely followed by the model without a fixed effect (r = 0.55).

For BakeV, all three models demonstrated lower predictive capability in both the *Bayesian Lasso* and *GBLUP* approaches, with accuracy ranging from r = 0.21 (for *GBLUP* with inclusion of $Bx7^{OE}$ KASP marker data as a fixed effect) to r = 0.29 in the (for *Bayesian Lasso* with inclusion of GWAS identified markers as a fixed effect). Overall, when considering the two-year validation approaches, the inclusion of GWAS-identified markers and $Bx7^{OE}$ KASP marker information as fixed effects consistently improved predictions for MixoT and BakeMT.

4.5 Discussion

Improving genomic prediction accuracy has been a major goal since genomic selection was introduced in plant breeding (Meuwissen et al., 2001; Desta and Ortiz, 2014). Prediction accuracy depends on several important factors, including marker density, trait heritability, training population size and composition, population structure, and model selection (Crossa et al., 2017; Heslot et al., 2012). Various statistical models that leverage these parameters have been developed and compared to improve accuracy (Jarquín et al., 2017; Heslot et al., 2012). This comparison is crucial for determining the most effective models for specific traits and breeding scenarios. Modifying the model components, model structure, and the weighting of markers by significance or effect size can also increase accuracy, suggesting the importance of model customization for different traits and conditions (Bernardo, 2014; Bian and Holland, 2017).

In this study, *Bayesian Lasso* models showed a notable advantage over *GBLUP* in predicting BakeMT and MixoT traits, evident in both cross- and forward-validation frameworks. *Bayesian Lasso* is generally better suited for scenarios where a small number of significant genetic markers are influential, as it efficiently performs shrinkage and variable selection (Park and Casella, 2008; Li and Sillanpää, 2012). However, *GBLUP* models are often preferred for highly polygenic traits due to their assumption of multiple small effect loci underlying trait expression (Meuwissen et al., 2001; Habier et al., 2011; Heslot et al., 2012; VanRaden, 2008). The *Bayesian Lasso's* approach of emphasizing major loci provided a more accurate prediction for these specific traits. This finding suggests that while *GBLUP* has broad applicability, the precision of *Bayesian Lasso* in identifying and weighting key markers makes it especially effective for traits like BakeMT and MixoT.

Previous studies have reported improved prediction accuracy with the *Bayesian Lasso* model over *GBLUP*. For instance, Lozada and Carter (2020) found that *Bayesian Lasso* showed superior prediction accuracy over *GBLUP* for grain yield and agronomic traits in winter wheat. Similarly, Ogutu et al. (2012) compared different models and found that *Bayesian Lasso* outperformed the others. Additionally, Gianola and Fernando (2020) observed a clear advantage of *Bayesian Lasso* over *GBLUP* when evaluating pine tree (*Pinus taeda*) traits. Sahebalam et al. (2022) compared the *Bayesian Lasso* and *GBLUP* models using simulated livestock data and reported that, although *GBLUP* showed improved overall accuracy, *Bayesian Lasso* was more effective in predicting oligogenic traits with high heritability. Zhang et al. (2019) also found that *GBLUP* improved prediction accuracy for low heritable, polygenic maize (*Zea mays*) traits, while *Bayesian* models were more suited to traits affected by a few major genes. A similar trend is observed in the current study, where the *Bayesian Lasso* model demonstrates superior prediction accuracy for traits with higher heritability, such as BakeMT and MixoT, underscoring its effectiveness at leveraging the genetic architecture of traits with significant heritable variation. In contrast, traits like BakeV, characterized by lower heritability, also exhibit reduced prediction accuracy under the *Bayesian Lasso* model, highlighting its sensitivity to the heritability of the traits it aims to predict.

Even though the models we tested showed improvements under *Bayesian Lasso* compared to *GBLUP*, the observed improvement was not the same for all traits. For example, in the first forward validation scenario, the three *Bayesian Lasso* models showed less than a 10% difference in prediction accuracy compared to *GBLUP* for BakeV and MixoT. Similarly, for BakeMT, the two models with fixed effects showed similar prediction accuracy, while the model without fixed effects exhibited considerable improvement for the *Bayesian Lasso* method compared to *GBLUP*. The comparable prediction accuracy of these models in both approaches suggests that they can be used interchangeably for predicting these traits. This observation aligns with the findings of Li et al. (2022), who reported closer performance between *Bayesian Lasso* and *GBLUP* models for cotton (*Gossypium hirsutum)* fiber and yield traits. Despite its potential accuracy benefits, *Bayesian Lasso* requires more computational time than *GBLUP* primarily because it involves complex *Bayesian* computations, such as Markov Chain Monte Carlo (MCMC) methods, to estimate a larger number of parameters and require numerous iterations to achieve accurate estimates. In contrast, *GBLUP* is a more direct method that often achieves similar accuracy but with reduced computational time requirements (Thijssen et al., 2016). Such considerations are crucial when selecting an appropriate model for practical applications (De los Campos et al., 2013; Heslot et al., 2015; Larkin et al., 2019).

The observed improvement in prediction accuracy in our study, particularly when integrating GWAS-identified markers as fixed effects, highlights the essential role of marker-trait associations underlying the trait's genetic architecture. These markers seemingly capture a significant portion of the genetic variation underlying the trait phenotype. Models with GWAS-identified markers as fixed effects enable a focus on relevant loci, reduce dimensionality, minimize noise from non-significant markers, and better capture phenotypic variance (Arruda et al., 2015; Bernardo, 2014). This approach potentially improves estimator efficiency and the signal-to-noise ratio (Arruda et al., 2015).

Previous studies support these findings. For instance, Li et al. (2019) found improved genomic prediction accuracy in maize by integrating large-effect GWAS-identified markers. Similarly, Spindel et al. (2016) noted the superior performance of a genomic prediction model with GWAS-identified markers over six other models in tropical rice (*Oryza sativa*), underscoring the value of incorporating GWAS-identified markers in genomic prediction models. In wheat, Rutkoski et al. (2014) showed improved prediction accuracy against stem rust upon factoring in the *Sr2* locus as a fixed effect. Kim et al. (2022) observed heightened accuracy for predicting capsaicinoid concentrations in chili pepper (*Capsicum annuum*) with GWAS-identified markers as fixed effects. Odilbekov et al. (2019) reported similar findings for resistance to *Septoria tritici* blotch (*Zymoseptoria tritici*) in Nordic winter wheat when integrating GWAS-identified markers as a fixed effect. Furthermore, Chen et al. (2023) affirmed that including GWAS-identified markers in genomic prediction models elevates the predictive capability in Norway spruce (*Picea abies*) clones.

From models with fixed effects, the model with *Bx7*^{OE} KASP marker data as a fixed effect outperformed the model without a fixed effect, yet it didn't surpass the model that included GWAS-identified markers as fixed effects. This outcome likely arises because the *Bx7*^{OE} allele targets a specific genomic region related to glutenin, thus addressing a limited portion of the genetic variation. In contrast, GWAS-identified markers span the entire genome, capturing a broader share of genetic variation that influences the trait. However, its inclusion as a fixed effect improved prediction accuracy for BakeMT and MixoT, most likely due to its association with glutenin composition and dough properties (Butow et al., 2003; Li et al., 2020). However, *Bx7*^{OE} was not the best-fixed effect to include for predicting BakeV, possibly because *Bx7*^{OE} is associated with mainly dough strength characteristics such as BakeMT and MixoT not directly influencing the BakeV (Butow et al., 2003; Radovanovic et al., 2002) and the fact that BakeV is highly affected by alleles at the *Glu-D1* locus, as shown in the Manhattan plot (Figure 4.2).

Forward validation approaches mimic real breeding scenarios, which is crucial to estimating the accuracy of genomic prediction, especially when considering their implementation in practical breeding programs (Crossa et al., 2017; Habier et al., 2013; He et al., 2016; Rutkoski et al., 2015; Sandhu et al., 2021). In this study, using a validation set of n = 55 individuals and n = 735 individuals for training, predictions were higher than those in cross-validation. While prediction accuracy for BakeMT and MixoT traits in this scenario surpassed accuracy observed in cross-validation, it declined for BakeV. This observation seems to contradict previous reports that have suggested that forward validation usually provides a more conservative estimate of prediction accuracy than cross-validation (Battenfield et al., 2016; Habier et al., 2013; He et al., 2016; Sandhu et al., 2021).

As highlighted by Crossa et al. (2014), cross-validation might tend to overestimate prediction accuracy, as may occur with a close degree of relationship between training and testing sets. However, with well-designed models and correctly assigned variables, forward validation can be as predictive as cross-validation, as evident in the current study and previous studies (Azizinia et al., 2023; Covarrubias-Pazaran et al., 2018; Okeke et al., 2017; Winn et al., 2023).

Chapter 4 Tables

Table 4.1 The number of individual genotypes per trial per location per year.

Year	Trial	Number of Locations	Number of observations	Number of unique Genotypes
2014	AYN[a]	3	105	35
	AYN	2	58	29
2015	AYND[b]	2	80	41
	ELITE[c]	2	68	34
	AYN	3	99	33
2016	AYND	3	69	49
	ELITE	5	165	33
	AYN	2	52	27
2017	AYND	2	52	31
	ELITE	4	140	35
	AYN	3	93	31
2018	AYND	2	36	18
	ELITE	3	75	25
	AYN	5	95	19
2019	AYND	2	78	45
	ELITE	5	185	37
	AYN	3	93	31
2020	AYND	2	108	54
	ELITE	4	116	29
2021	AYN	2	52	26
	ELITE	3	117	39
	AYN	3	75	25
2022	AYND	2	72	36
	ELITE	3	84	28

Abbreviations: [a] Advanced Yield Nursery; [b] Advanced Yield Nursery of Doubled Haploids; [c] CSU Elite Trial; CSU, Colorado State University.

Table 4.2 The primer sequence for KASP maker for *Bx7*^{OE}

Primer	Target	Sequence
A1	Insertion-*Bx7*^{OE}	CGGCAACAACTTGTGGGGGC
F1	Insertion-*Bx7*^{OE}	GTTGTTGCCGGAATATTTTACAATATATTTAAG
A2	Deletion-*Bx7*	TATTCCGGCAACAACTTGTGGGGTA
F2	Deletion-*Bx7*	CACTTCTTCTCTCGTTGGCCTTATCT

Table 4.3 Summary statistics and heritability of bake mixing time (BakeMT), Mixograph tolerance (MixoT), and bake loaf volume (BakeV).

Traits	Mean	Minimum	Maximum	SD[a]	h^{2b}	SE h^2
BakeMT	4.86	1.04	14.94	1.80	0.84	0.05
BakeV	922.62	689.96	1184.94	85.81	0.64	0.06
MixoT	3.70	0.82	6.00	1.22	0.78	0.06

Abbreviations: [a] standard deviation; [b] heritability; BakeMT, bake mixing time; BakeV, bake loaf volume; and MixoT, Mixograph tolerance.

Table 4.4 Significant marker-trait associations identified in a genome-wide association study (GWAS) for bake mixing time (BakeMT), Mixograph tolerance (MixoT), and bake loaf volume (BakeV) traits.

Trait	Marker	Chromosome	-log$_{10}$(p)	MAF
BakeMT	S1B_3282665	1B	6.02	0.31
BakeMT	S1B_3282705	1B	10.70	0.23
BakeMT	S1B_3765044	1B	17.31	0.30
BakeMT	S1B_3765060	1B	11.82	0.22
BakeMT	S1B_4528524	1B	7.81	0.19
BakeMT	S1B_4935662	1B	42.64	0.33
BakeMT	S1B_6437407	1B	23.31	0.31
BakeMT	S1B_6561394	1B	19.42	0.35
BakeMT	S1B_6923012	1B	6.94	0.32
BakeMT	S1B_7692865	1B	5.93	0.41
BakeMT	S1B_7939480	1B	5.91	0.18
BakeMT	S1D_885391	1D	5.52	0.28
BakeMT	S1D_414593582	1D	8.52	0.36
BakeMT	S1D_414927498	1D	7.89	0.27
BakeMT	S1D_415205039	1D	7.96	0.20
BakeMT	S1D_415641567	1D	7.99	0.21
BakeMT	S1D_415821975	1D	7.76	0.29
BakeMT	S1D_416711152	1D	7.08	0.16
BakeMT	S1D_416855528	1D	7.45	0.23
BakeMT	S1D_417603470	1D	7.80	0.42
BakeMT	S1D_417945958	1D	7.42	0.25
BakeMT	S2B_788719721	2B	17.61	0.31
BakeMT	S5D_565697447	5D	17.81	0.19
BakeMT	S6D_475591931	6D	18.51	0.17
MixoT	S1B_3765044	1B	6.41	0.33
MixoT	S1B_4935662	1B	14.39	0.43
MixoT	S1B_6437407	1B	7.87	0.29
MixoT	S1B_6561394	1B	6.67	0.12
MixoT	S1D_885391	1D	6.52	0.22
MixoT	S1D_414593582	1D	10.73	0.18
MixoT	S1D_414927498	1D	9.17	0.31
MixoT	S1D_415205039	1D	10.95	0.42
MixoT	S1D_415641567	1D	10.77	0.33
MixoT	S1D_415821975	1D	10.88	0.26
MixoT	S1D_416711152	1D	11.43	0.38
MixoT	S1D_416711191	1D	7.788	0.19
MixoT	S1D_416855528	1D	11.91	0.23
MixoT	S1D_417603455	1D	7.23	0.32
MixoT	S1D_417603470	1D	11.04	0.19
MixoT	S1D_417945958	1D	11.76	0.28

MixoT	S1D_418175745	1D	7.25	0.31
MixoT	S1D_418754660	1D	7.01	0.33
MixoT	S2B_788719721	2B	9.56	0.24
MixoT	S5D_565697447	5D	8.86	0.29
MixoT	S6D_475591931	6D	7.07	0.21
BakeV	S1D_414593582	1D	6.24	0.34
BakeV	S1D_416711152	1D	5.83	0.22
BakeV	S1D_417603470	1D	6.08	0.28
BakeV	S1D_417945958	1D	6.53	0.29

Abbreviation: MAF, marker allele frequency

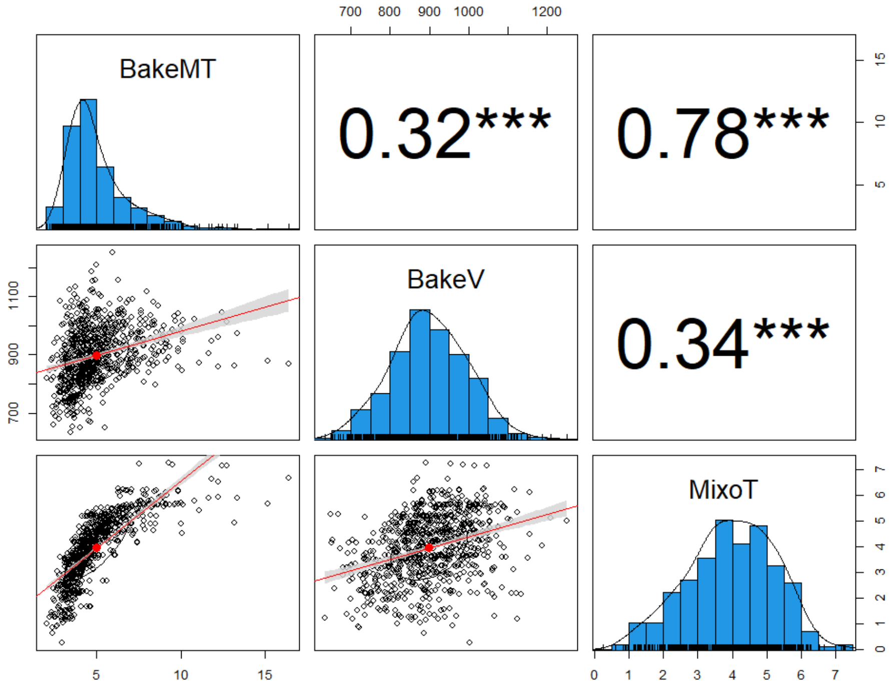


Figure 4.1 Pair plots showing the best linear unbiased estimation of end-use quality traits across year-location-trials. Histograms and trait names are displayed on the diagonal. The scatterplots of the traits with linear regression line fit are represented in the lower half of the figure. The numbers on the upper half of the diagonal represent Pearson's correlation coefficients between the traits. The three asterisks (***) denote a p-value of 0.001, indicating a high level of statistical significance for the correlation among corresponding traits.

Abbreviations: BakeMT, bake mixing time; BakeV, bake loaf volume; MixoT, Mixograph tolerance.

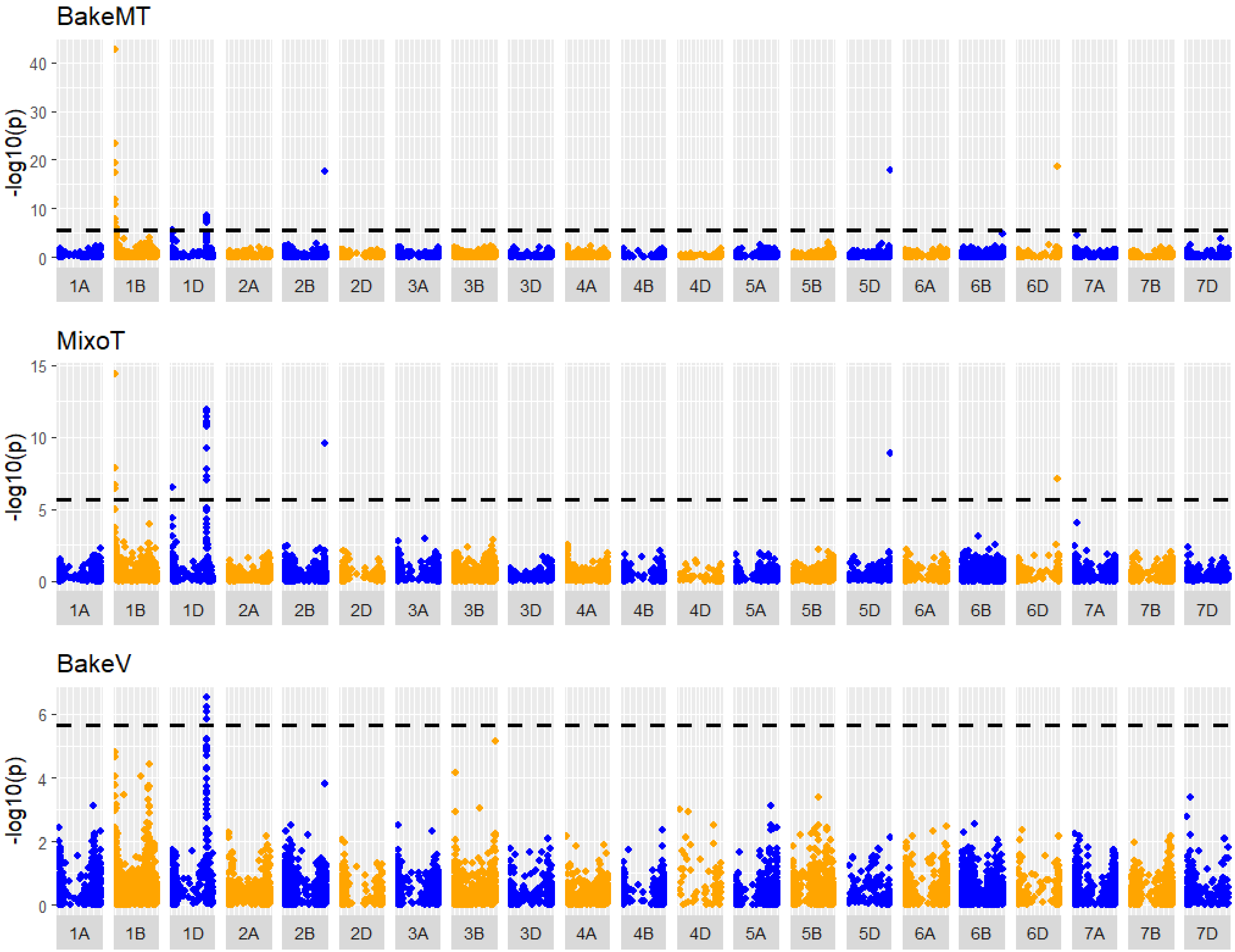

BakeMT
-log10(p)
40
30
20
10
0
1A
1B
1D
2A
2B
2D
3A
3B
3D
4A
4B
4D
5A
5B
5D
6A
6B
6D
7A
7B
7D
MixoT
-log10(p)
15
10
5
0
1A
1B
1D
2A
2B
2D
3A
3B
3D
4A
4B
4D
5A
5B
5D
6A
6B
6D
7A
7B
7D
BakeV
-log10(p)
6
4
2
0
1A
1B
1D
2A
2B
2D
3A
3B
3D
4A
4B
4D
5A
5B
5D
6A
6B
6D
7A
7B
7D

Figure 4.2 Manhattan plot illustrating -$\log_{10}$(p) values from a genome-wide association study (GWAS) of A, bake mixing time (BakeMT); B, Mixograph tolerance (MixoT); C, bake loaf volume (BakeV) of 790 CSU hard winter wheat genotypes. Each point in the plot indicates a single nucleotide polymorphism (SNP), displayed with its -$\log_{10}$(p) value on the vertical axis and its chromosomal position on the horizontal axis. The dashed horizontal line marks the genome-wide significance threshold using the Bonferroni test (-$\log_{10}$(p)= 5.67). Peaks above this line highlight SNPs showing significant association with the respective trait under study.

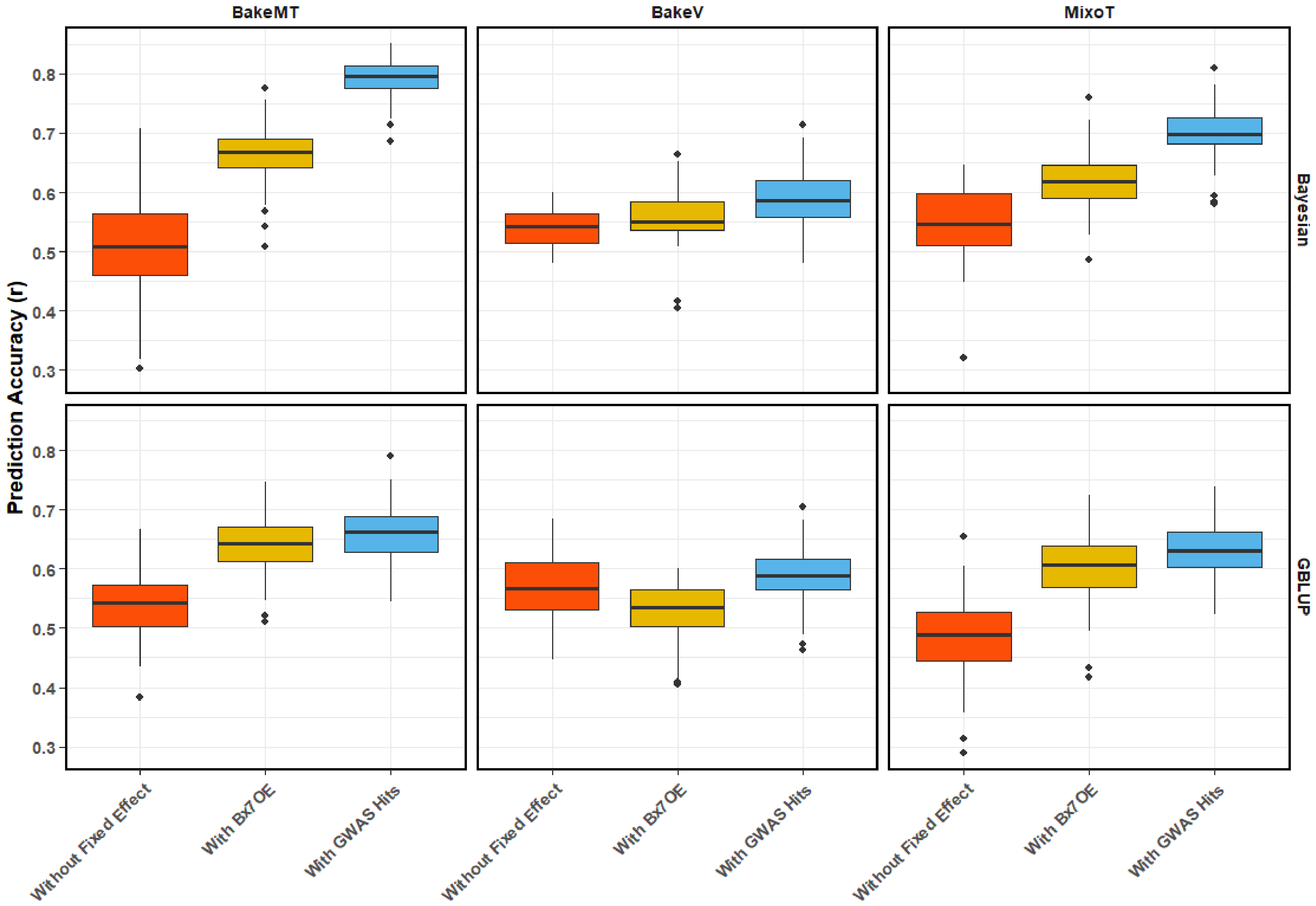

BakeMT
BakeV
MixoT
Bayesian
GBLUP
Prediction Accuracy (r)
0.8
0.7
0.6
0.5
0.4
0.3
Without Fixed Effect
With Bx7OE
With GWAS Hits
Models

Figure 4.3 Box plots depicting the distribution of prediction accuracies for bake mixing time (BakeMT), bake loaf volume (BakeV), and Mixograph tolerance (MixoT), measured as the correlation (r) between the genomic estimated breeding values (GEBVs) of the validation population and the best linear unbiased estimates (BLUEs) of the lines within that population. For each approach, three models were evaluated: no fixed effect, *Bx7^{OE}* KASP marker data as a fixed effect, and genome-wide association study (GWAS) hits (GWAS-identified markers) as fixed effects. Each box plot shows the distribution of correlation coefficients over 100 cross-validation iterations, with the median marked as a horizontal line within each box. The box depicts the interquartile range, while outliers are data points outside the box.

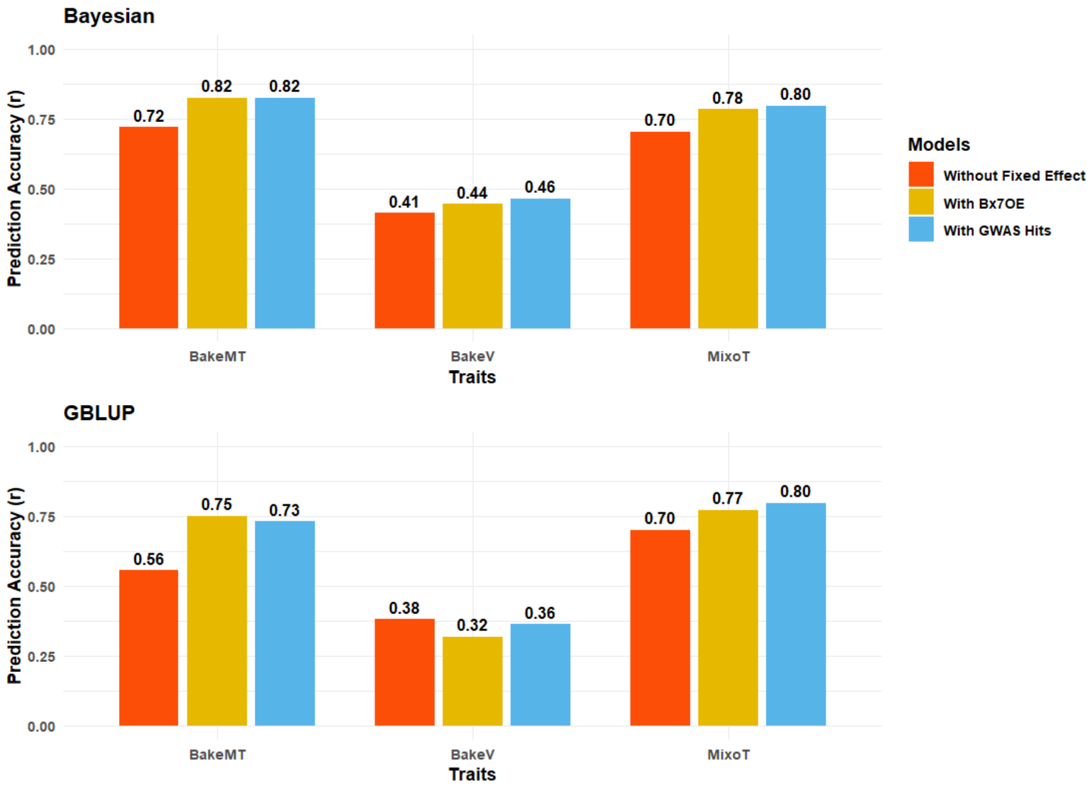


Figure 4.4 Forward validation accuracy of models for end-use quality traits using *Bayesian Lasso* and *GBLUP* for the first forward validation scenario where the elite trials from 2020, 2021, and 2022 were used as a testing set. The bar plot displays prediction accuracy values for models incorporating different fixed effects across three quality traits: bake mixing time (BakeMT), Mixograph tolerance (MixoT), and bake loaf volume (BakeV). Models utilizing genome-wide association study (GWAS) markers as a fixed effect, the $Bx7^{OE}$ KASP marker data as a fixed effect, and without fixed effects are represented by colored bars. Numeric values above each bar indicate prediction accuracy for each model-trait combination under the two modeling approaches.

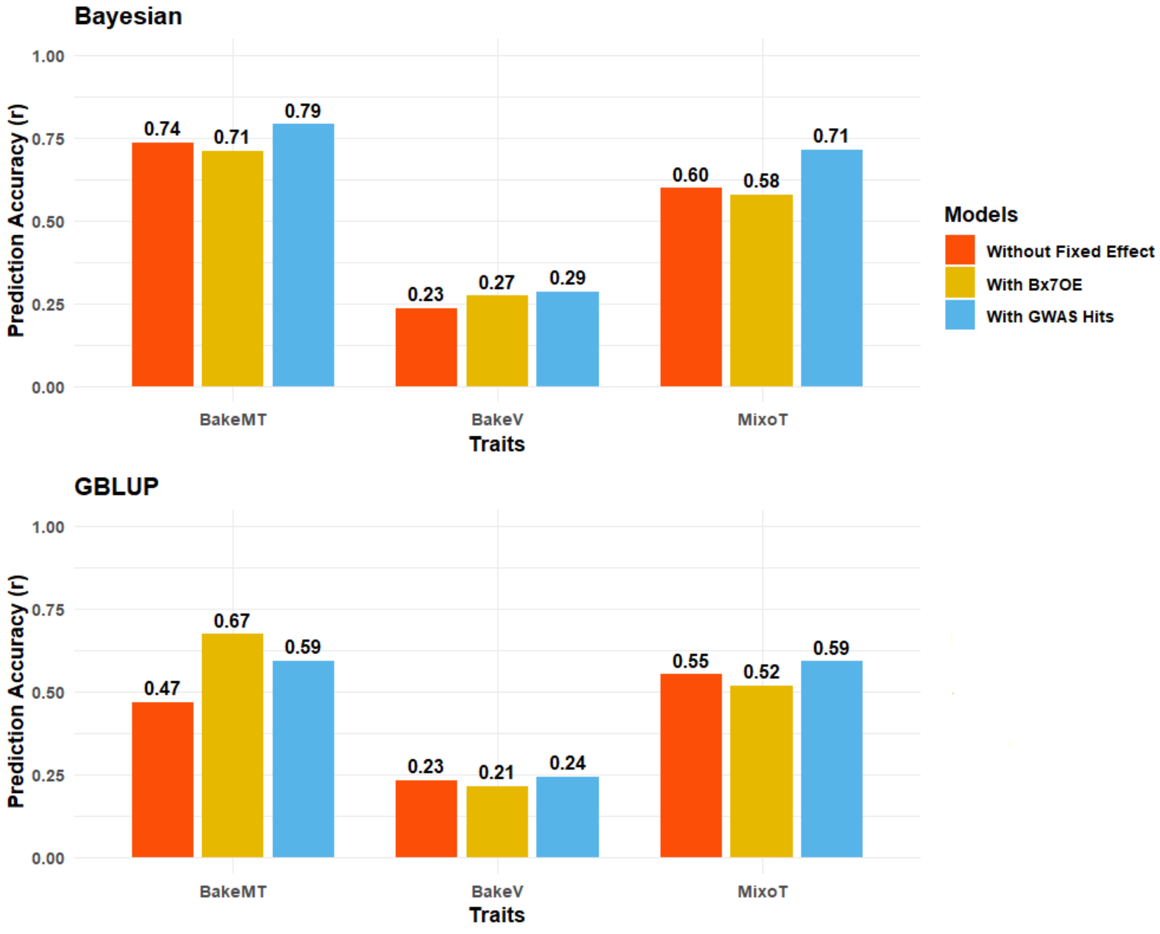


Figure 4.5 Forward validation accuracy of models for end-use quality traits using *Bayesian Lasso and GBLUP* using two years of data from all trials as a validation set. The bar plot displays prediction accuracy for models incorporating different fixed effects across three baking traits: bake mixing time (BakeMT), Mixograph tolerance (MixoT), and bake loaf volume (BakeV). Models utilizing genome-wide association study (GWAS) markers as a fixed effect, the *Bx7*^{OE} KASP marker data as a fixed effect, and no fixed effects are represented by colored bars. Numeric values above each bar indicate prediction accuracy for each model-trait combination under the two *GBLUP* and *Bayesian Lasso* methods.

Chapter 4 References

AACC Approved Methods of Analysis, 11th Ed. Methods 10-05.01 Guidelines for Measurement of Volume by Rapeseed Displacement, 2010. Cereals and Grains Association, St. Paul, MN, U.S.A. https://doi.org/10.1094/AACCIntMethod-02-03.02

AACC Approved Methods of Analysis, 11th Ed. Methods 54-40.02 Guidelines for 10g Mixograph test, 2010. Cereals and Grains Association, St. Paul, MN, U.S.A.

Anderson, O. D., and Greene, F. C. (1997). The α-gliadin gene family. II. DNA and protein sequence variation, subfamily structure, and origins of pseudogenes. *Theoretical and Applied Genetics,* 95, 59-65. https://doi.org/10.1007/s001220050532

Appels, R., Eversole, K., Feuillet, C., Keller, B., Rogers, J., Stein, N., . . . Manuscript Writing, T. (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome [Article]. *Science*, *361*(6403), 661-+, Article eaar7191. https://doi.org/10.1126/science.aar7191

Arruda, M. P., Brown, P. J., Lipka, A. E., Krill, A. M., Thurber, C., and Kolb, F. L. (2015). Genomic selection for predicting Fusarium head blight resistance in a wheat breeding program. *The Plant Genome*, 8(3), plantgenome2015-01. https://doi.org/10.3835/plantgenome2015.01.0003

Azizinia, S., Mullan, D., Rattey, A., Godoy, J., Robinson, H., Moody, D., ... and Daetwyler, H. D. (2023). Improved multi-trait prediction of wheat end-product quality traits by integrating NIR-predicted phenotypes. *Frontiers in Plant Science*, *14*, 1167221. https://doi.org/10.3389/fpls.2023.1167221

Barbano, R., Arridge, S., Jin, B., and Tanno, R. (2022). Uncertainty quantification in medical image Synthesis. In *Biomedical Image Synthesis and Simulation* (pp. 601-641). Academic Press. https://doi.org/10.1016/B978-0-12-824349-7.00033-5

Battenfield, S. D., Guzmán, C., Gaynor, R. C., Singh, R. P., Peña, R. J., Dreisigacker, S., ... and Poland, J. A. (2016). Genomic selection for processing and end-use quality traits in the CIMMYT spring bread wheat breeding program. *The plant genome*, 9(2), plantgenome2016-01. https://doi.org/10.3835/plantgenome2016.01.0005

Bernardo, R. (2014). Genome-wide selection is when major genes are known. *Crop Science*, 54(1), 68-75. https://doi.org/10.2135/cropsci2013.05.0315

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

Bian, Y., and Holland, J. B. (2017). Enhancing genomic prediction with genome-wide association studies in multi-parental maize populations. *Heredity*, 118(6), 585-593. https://doi.org/10.1038/hdy.2017.4

Browning, B. L., Zhou, Y., and Browning, S. R. (2018). A one-penny imputed genome from next-generation reference panels. *The American Journal of Human Genetics*, *103*(3), 338–348.

Butler, D., Cullis, B., Gilmour, A., and Thompson, R. (2007). Asreml-R: an R package for mixed models using residual maximum likelihood.

Butow, B. J., Ma, W., Gale, K. R., Cornish, G. B., Rampling, L., Larroque, O., and Békés, F. (2003). Molecular discrimination of Bx7 alleles demonstrates that a highly expressed high-molecular-weight glutenin allele has a major impact on wheat flour dough strength. *Theoretical and Applied Genetics,* 107, 1524-1532. https://doi.org/10.1007/s00122-003-1396-8

Chen, Z. Q., Klingberg, A., Hallingbäck, H. R., and Wu, H. X. (2023). The preselection of QTL markers enhances the accuracy of genomic selection in Norway spruce. *BMC Genomics*, 24(1), 1-16.

Chen, Q., Zhang, W., Gao, Y., Yang, C., Gao, X., Peng, H., ... and Yao, Y. (2019). High molecular weight glutenin subunits 1Bx7 and 1By9 encoded by *the Glu-B1* locus affect wheat dough properties and the sponge cake quality. *Journal of Agricultural and Food Chemistry*, 67(42), 11796-11804. https:// dio.org/10.1021/acs.jafc.9b05030.

Covarrubias-Pazaran, G., Schlautman, B., Diaz-Garcia, L., Grygleski, E., Polashock, J., Johnson-Cicalese, J., ... and Zalapa, J. (2018). Multivariate GBLUP improves the accuracy of genomic selection for yield and fruit weight in biparental populations of Vaccinium macrocarpon Ait. *Frontiers in Plant Science*, *9*, 1310. https://doi.org/10.3389/fpls.2018.01310

Covarrubias-Pazaran, G. (2016). Genome-assisted prediction of quantitative traits using the R package *sommer*. *PloS one*, *11*(6), e0156744. https://doi.org/10.1371/journal.pone.0156744

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., De los Campos, G., Burgueño, J., González-Camacho, J. M., Pérez-Elizalde, S., and Beyene, Y. (2017). Genomic selection in plant breeding: Methods, models, and perspectives. *Trends in Plant Science*, 22(11), 961–975. https://doi.org/10.1016/j.tplants.2017.08.011

Crossa, J., Perez, P., Hickey, J., Burgueno, J., Ornella, L., Cerón-Rojas, J., ... and Mathews, K. (2014). Genomic prediction in CIMMYT maize and wheat breeding programs. *Heredity*, 112(1), 48-60. https://doi.org/10.1038/hdy.2013.16

Dai, S. F., Xu, D. Y., Wen, Z. J., Song, Z. P., Chen, H. X., Li, H. Y., ... and Yan, Z. H. (2018). Characterization of a Novel 4.0-kb Y-type HMW-GS from Eremopyrum distance. *Cereal Research Communications,* 46(3), 499-509. https://doi.org/10.1556/0806.46.2018.018

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

De Santis, M. A., Giuliani, M. M., Giuzio, L., De Vita, P., Lovegrove, A., Shewry, P. R., and Flagella, Z. (2017). Differences in gluten protein composition between old and modern durum wheat genotypes in relation to 20th-century breeding in Italy. *European Journal of Agronomy*, 87, 19–29. https://doi.org/10.1016/j.eja.2017.04.003

Desta, Z.A., and Ortiz, R. (2014). Genomic selection: genome-wide prediction in plant improvement. *Trends in Plant Science*, 19(9), 592-601. https://doi.org/10.1016/j.tplants.2014.05.006

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. *The Plant Genome*, 4(3). https://doi.org/10.3835/plantgenome2011.08.0024

Finney, K. F. (1972). A ten-gram mixograph for determining and predicting functional properties of wheat flours. *Bakers Digest*, *46*, 32-35. Bakers Digest 46 32-35,38-42,77, 1972

Gao, J., Hu, X., Gao, C., Chen, G., Feng, H., Jia, Z., Zhao, P., Yu, H., Li, H., Geng, Z., Fu, J., Zhang, J., Cheng, Y., Yang, B., Pang, Z., Xiang, D., Jia, J., Su, H., Mao, H., … Li, Q. (2023). Deciphering genetic basis of developmental and agronomic traits by integrating high-throughput optical phenotyping and genome-wide association studies in wheat. *Plant Biotechnology Journal*. https://doi.org/10.1111/pbi.14104

Gianibelli, M. C., Larroque, O. R., MacRitchie, F., and Wrigley, C. W. (2001). Biochemical, genetic, and molecular characterization of wheat endosperm proteins. *Cereal Chem*, 78(6), 635-646.

Gianola, D. (2013). Priors in whole-genome regression: the Bayesian alphabet returns. *Genetics*, 194(3), 573-596. https://doi.org/10.1534/genetics.113.151753

Gianola, D., and Fernando, R. L. (2020). A multiple-trait *Bayesian Lasso* for genome-enabled analysis and prediction of complex traits. *Genetics,* 214(2), 305-331. https://doi.org/10.1534/genetics.119.302934

Glaubitz, J. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q., and Buckler, E. S. (2014). TASSEL-GBS: A High-Capacity Genotyping by Sequencing Analysis Pipeline. *PLOS ONE,* 9(2), e90346. https://doi.org/10.1371/journal.pone.0090346

Gupta, R. B., Shepherd, K. W., and MacRitchie, F. (1991). Genetic control and biochemical properties of some high molecular weight albumins in bread wheat. *Journal of Cereal Science,* 13(3), 221-235. https://doi.org/10.1016/S0733-5210(09)80002-7

Habier, D., Fernando, R. L., and Garrick, D. J. (2013). Genomic BLUP decoded: a look into the black box of genomic prediction. *Genetics*, *194*(3), 597-607. https://doi.org/10.1534/genetics.113.152207

Habier, D., Fernando, R. L., Kizilkaya, K., and Garrick, D. J. (2011). Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics*, 12(1), 1-12. https://doi.org/10.1186/1471-2105-12-186

Habier, D., Fernando, R. L., and Dekkers, J. (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics,* 177(4), 2389-2397. https://doi.org/10.1534/genetics.107.081190

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

Heslot, N., Jannink, J. L., and Sorrells, M. E. (2015). Perspectives for genomic selection applications and research in plants. *Crop Science*, *55*(1), 1-12. https://doi.org/10.2135/cropsci2014.03.0249

Heslot, N., Yang, H. P., Sorrells, M. E., and Jannink, J. L. (2012). Genomic selection in plant breeding: a comparison of models. *Crop Science*, 52(1), 146-160.

Isidro, J., Jannink, J. L., Akdemir, D., Poland, J., Heslot, N., and Sorrells, M. E. (2015). Training set optimization under population structure in genomic selection. *Theoretical and Applied Genetics*, 128, 145-158. https://doi.org/10.2135/cropsci2011.06.0297

Jarquin, D., De Leon, N., Romay, C., Bohn, M., Buckler, E. S., Ciampitti, I., ... and Lorenz, A. (2021). Utility of climatic information via combining ability models to improve genomic prediction for yield within the genomes to fields maize project. *Frontiers in Genetics*, 11, 592769. https://doi.org/10.3389/fgene.2020.592769

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

Kumar, A., Agarwal, D. K., Kumar, S., Reddy, Y. M., Chintagunta, A. D., Saritha, K. V., and Kumar, S. J. (2019). Nutraceuticals derived from seed storage proteins: implications for health wellness. *Biocatalysis and Agricultural Biotechnology*, 17, 710-719. https://doi.org/10.1016/j.bcab.2019.01.044

Kumar, S., Kumari, J., Bhusal, N., Pradhan, A. K., Budhlakoti, N., Mishra, D. C., Chauhan, D., Kumar, S., Singh, A. K., and Reynolds, M. (2020). A genome-wide association study reveals genomic regions associated with ten agronomical traits in wheat under late-sown conditions. *Frontiers in Plant Science*, 11, 1420. https://doi.org/10.3389/fpls.2020.549743

Larkin, D. L., Lozada, D. N., and Mason, R. E. (2019). Genomic selection—considerations for successful implementation in wheat breeding programs. *Agronomy*, *9*(9), 479. https://doi.org/10.3390/agronomy9090479

Li, Z., Liu, S., Conaty, W., Zhu, Q. H., Moncuquet, P., Stiller, W., and Wilson, I. (2022). Genomic prediction of cotton fiber quality and yield traits using Bayesian regression methods. *Heredity*, 129(2), 103-112. https://doi.org/10.1038/s41437-022-00537-x

Li, S., Liu, Y., Tong, J., Yu, L., Ding, M., Zhang, Z., and Gao, X. (2020). The overexpression of high-molecular-weight glutenin subunit Bx7 improves the dough rheological properties by altering secondary and micro-structures of wheat gluten. *Food Research International*, 130, 108914. https://doi.org/10.1016/j.foodres.2019.108914

Li, D., Xu, Z., Gu, R., Wang, P., Lyle, D., Xu, J., Zhang, H., and Wang, G. (2019). Enhancing genomic selection by fitting large-effect SNPs as fixed effects and a genotype-by-environment effect using a maize BC1F3:4 population. PLoS One, 14 (10), e0223898. https://doi.org/10.1371/journal.pone.0223898

Li, Z., and Sillanpää, M. J. (2012). Overview of *Lasso*-related penalized regression methods for quantitative trait mapping and genomic selection. *Theoretical and Applied Genetics*, 125, 419-435.https://doi.org/10.1007/s00122-012-1892-9

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*(14), 1754-1760. https://doi.org/10.1093/bioinformatics/btp324

Lozada, D. N., and Carter, A. H. (2020). Genomic selection in winter wheat breeding using a recommender approach. *Genes*, *11*(7), 779. https://doi.org/10.3390/genes11070779

McIntosh, S. R., Brushett, D., and Henry, R. J. (2008). GTP cyclohydrolase one expression and folate accumulation in the developing wheat seed. *Journal of Cereal Science*, 48(2), 503-512. https://doi.org/10.1016/j.jcs.2007.11.008

Meuwissen, T.H., Hayes, B.J., and Goddard, M.E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4), 1819-1829. https://doi.org/10.1093/genetics/157.4.1819

Moore, K. L., Tosi, P., Palmer, R., Hawkesford, M. J., Grovenor, C. R., and Shewry, P. R. (2016). The dynamics of protein body formation in developing wheat grain. *Plant Biotechnology Journal*, 14(9), 1876-1882. https://doi.org/10.1111/pbi.12549

Norman, A., Taylor, J., Edwards, J., and Kuchel, H. (2018). Optimizing genomic selection in wheat: effect of marker density, population size, and population structure on prediction accuracy. G3: *Genes, Genomes, Genetics*, 8(9), 2889-2899. https://doi.org/10.1534/g3.118.200311

Odilbekov, F., Armonienė, R., Koc, A., Svensson, J., and Chawade, A. (2019). GWAS-Assisted Genomic Prediction to Predict Resistance to Septoria Tritici Blotch in Nordic Winter Wheat at Seedling Stage. *Frontiers in Genetics*, 10. https://doi.org/10.3389/fgene.2019.01224

Ogutu, J. O., Schulz-Streeck, T., and Piepho, H. P. (2012, December). Genomic selection using regularized linear regression models: ridge regression, *Lasso*, elastic net, and their extensions. In *BMC proceedings* (Vol. 6, pp. 1-6). *BioMed Central*. https://doi.org/10.1186/1753-6561-6-S2-S10

Okeke, U. G., Akdemir, D., Rabbi, I., Kulakow, P., and Jannink, J. L. (2017). Accuracies of univariate and multivariate genomic prediction models in African cassava. *Genetics Selection Evolution*, 49(1), 1-10. https://doi.org/10.1186/s12711-017-0361-y

Park, T., and Casella, G. (2008). The *Bayesian Lasso*. *Journal of the American Statistical Association*, 103(482), 681-686. https://doi.org/10.1198/016214508000000337

Payne, P. I. (1987). Genetics of wheat storage proteins and the effect of allelic variation on bread-making quality. *Annual Review of Plant Physiology*, 38(1), 141–153. doi.org/10.1146/annurev.pp.38.060187.001041

Payne, P. I., Holt, L. M., and Law, C. N. (1981). Structural and genetic studies on the high-molecular-weight subunits of wheat glutenin: Part 1: Allelic variation in subunits amongst varieties of wheat (Triticum aestivum). *Theoretical and Applied Genetics,* 60, 229-236. https://doi.org/10.1007/BF02342544

Pérez, P., and 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., and Rife, T. W. (2012). Genotyping-by-sequencing for plant breeding and genetics. *The Plant Genome*, 5(3). https://doi.org/10.3835/plantgenome2012.05.0005

R Core Team, R. (2013). R: A language and environment for statistical computing.

Radovanovic, N., Cloutier, S., Brown, D., Humphreys, D. G., and Lukow, O. M. (2002). Genetic variance for gluten strength contributed by high molecular weight glutenin proteins. *Cereal Chemistry*, *79*(6), 843-49.https://doi.org/10.1094/CCHEM.2002.79.6.843

Ragupathy, R., Naeem, H. A., Reimer, E., Lukow, O. M., Sapirstein, H. D., and Cloutier, S. (2008). Evolutionary origin of the segmental duplication encompassing the wheat GLU-B1 locus encoding the overexpressed Bx7 (Bx7 OE) high molecular weight glutenin subunit. *Theoretical and Applied Genetics,* 116, 283-296. https://doi.org/10.1007/s00122-007-0666-2

Revelle, W., and Revelle, M. W. (2015). Package *'psych'. The Comprehensive R Archive Network*, *337*(338). https://CRAN.R-project.org/package=psych

Rice, B., and Lipka, A. E. (2019). Evaluation of RR-BLUP genomic selection models that incorporate peak genome-wide association study signals in maize and sorghum. *The Plant Genome,* 12(1), 180052. https://doi.org/10.3835/plantgenome2018.07.0052

Rutkoski, J., Singh, R. P., Huerta-Espino, J., Bhavani, S., Poland, J., Jannink, J. L., and Sorrells, M. E. (2015). Genetic gain from phenotypic and genomic selection for quantitative resistance to stem rust of wheat. *The Plant Genome*, *8*(2), plantgenome2014-10. https://doi.org/10.3835/plantgenome2014.10.0074

Rutkoski, J. E., Poland, J. A., Singh, R. P., Huerta-Espino, J., Bhavani, S., Barbier, H., ... and Sorrells, M. E. (2014). Genomic selection for quantitative adult plant stem rust resistance in wheat. *The Plant Genome*, 7(3), plantgenome2014-02. https://doi.org/10.3835/plantgenome2014.02.0006

Sahebalam, H., Gholizadeh, M., Hafezian, H., and Ebrahimi, F. (2022). Evaluation of Bagging approach versus *GBLUP* and *Bayesian Lasso* in genomic prediction. *Journal of Genetics*, 101(1), 19. https://doi.org/10.1007/s12041-022-01358-x

Sandhu, K. S., Aoun, M., Morris, C. F., and Carter, A. H. (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

Sehgal, D., Rosyara, U., Mondal, S., Singh, R., Poland, J., and Dreisigacker, S. (2020). Incorporating genome-wide association mapping results into genomic prediction models for grain yield and yield stability in CIMMYT spring bread wheat. *Frontiers in Plant Science*, 11, 197. https://doi.org/10.3389/fpls.2020.00197

Shewry, P. R., Halford, N. G., Tatham, A. S., Popineau, Y., Lafiandra, D., and Belton, P. S. (2003). The high molecular weight subunits of wheat glutenin and their role in determining wheat processing properties. *Elsevier*. https://doi.org/10.1016/S1043-4526(03)45006-7

Shewry, P. R., Halford, N. G., Belton, P. S., and Tatham, A. S. (2002). The structure and properties of gluten: an elastic protein from wheat grain. Philosophical Transactions of the Royal Society of London. Series B: *Biological Sciences*, 357(1418), 133–142. https://doi.org/10.1098/rstb.2001.1024

Shewry, P. R., Popineau, Y., Lafiandra, D., and Belton, P. (2000). Wheat glutenin subunits and dough elasticity: findings of the EUROWHEAT project. *Trends in Food Science and Technology,* 11(12), 433-441. https://doi.org/10.1016/S0924-2244(01)00035-8

Shewry, P. R., Halford, N. G., and Tatham, A. S. (1992). High molecular weight subunits of wheat glutenin. *Journal of Cereal Science*, 15(2), 105-120. https://doi.org/10.1016/S0733-5210(09)80062-3

Shewry, P. R., and Tatham, A. S. (1990). The prolamin storage proteins of cereal seeds: structure and evolution. *Biochemical Journal*, *267*(1),1. https://doi/10.1042/bj2670001

Spindel, J., Begum, H., Akdemir, D. et al. (2016). Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. *Heredity*, 116(5), 395–408. https://doi.org/10.1038/hdy.2015.113

Singh, N. K., and Shepherd, K. W. (1988). Linkage mapping of genes controlling endosperm storage proteins in wheat: 1. Genes on the short arms of group 1 chromosomes. *Theoretical and Applied Genetics,* 75, 628-641. https://doi.org/10.1007/BF00289132

Singh, N. K., Donovan, R., and MacRitchie, F. (1990). As a Measure of Bread-making Quality. *Cereal Chem*, 67(2), 161-170.

Thijssen, B., Dijkstra, T. M., Heskes, T., and Wessels, L. F. (2016). BCM: toolkit for Bayesian analysis of computational models using samplers. *BMC Systems Biology*, *10*(1), 1-8. https://doi.org/10.1186/s12918-016-0339-3.

Tosi, P., Gritsch, C. S., He, J., and Shewry, P. R. (2011). Distribution of gluten proteins in bread wheat (Triticum aestivum) grain. *Annals of Botany*, 108(1), 23-35. https://doi.org/10.1093/aob/mcr098

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

Verges, V. L., Brown-Guedira, G. L., & Van Sanford, D. A. (2021). Genome-wide association studies combined with genomic selection as a tool to increase Fusarium head blight resistance in wheat. *Crop Breeding, Genetics and Genomics*, *3*(4).

Vilhjálmsson, B. J., Yang, J., Finucane, H. K., Gusev, A., Lindström, S., Ripke, S., ... & Marsal, S. (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *The American Journal of Human Genetics*, *97*(4), 576-592.

Wang, Y., Mette, M. F., Miedaner, T., Gottwald, M., Wilde, P., Reif, J. C., and Zhao, Y. (2014). The accuracy of prediction of genomic selection in elite hybrid rye populations surpasses the accuracy of marker-assisted selection and is equally augmented by multiple field evaluation locations and test years. *BMC Genomics*, 15(1), 1-12. https://doi.org/10.1186/1471-2164-15-556

Wang, S., Yu, Z., Cao, M., Shen, X., Li, N., Li, X., ... and Yan, Y. (2013). Molecular mechanisms of HMW glutenin subunits from 1Sl genome of Aegilops longissima positively affecting wheat bread-making quality. *PLoS One*, 8(4), e58947. https://doi.org/10.1371/journal.pone.0058947

Wickham, H., Chang, W., and Wickham, M. H. (2016). Package 'ggplot2'. Create elegant data visualizations using the grammar of graphics. *Version*, 2(1), 1-189.

Wieser, H. (2007). Chemistry of gluten proteins. *Food Microbiology*, 24(2), 115-119. https://doi.org/10.1016/j.fm.2006.07.004

Williams, E., Piepho, H. P., and Whitaker, D. (2011). Augmented p-rep designs. *Biometrical Journal,* 53(1), 19-27. https://doi.org/10.1002/bimj.201000102

Winn, Z. J., Larkin, D. L., Lozada, D. N., DeWitt, N., Brown-Guedira, G., and Mason, R. E. (2023). Multivariate genomic selection models improve the prediction accuracy of agronomic traits in soft red winter wheat. *Crop Science*, 63(4), 2115-2130. https://doi.org/10.1002/csc2.20994

Zhang, H., Yin, L., Wang, M., Yuan, X., and Liu, X. (2019). Factors affecting the accuracy of genomic selection for agricultural economic traits in maize, cattle, and pig populations. *Frontiers in Genetics*, 10, 189. https://doi.org/10.3389/fgene.2019.00189

Žilić, S., Barać, M., Pešić, M., Dodig, D., and Ignjatović-Micić, D. (2011). Characterization of proteins from grains of different bread and durum wheat genotypes. *International Journal of Molecular Sciences*, *12*(9), 5878-5894. https://doi.org/10.3390/ijms12095878

Chapter 5-Conclusions

Breeding for quality traits in hard winter wheat has been challenging since the nature of the breeding cycle is very brief between harvesting and the next planting, and the phenotyping of the quality traits is time intensive. This dissertation highlights the need for integrating genomic approaches such as genome-wide association study and genomic prediction to enhance the breeding effort toward quality traits.

Identifying the genetic architecture of the traits provides information to choose the best approach to improve the traits. The multiple loci identified for SRCs and GPI traits across 17 chromosomes contribute smaller effects to the phenotype, suggesting that water absorption capacity (WAC) is a polygenic, quantitatively inherited trait. The significant QTNs identified in this study, which appear to exhibit only additive effects in the studied population, provide an opportunity for utilization in a combined haplotype approach for marker-assisted selection. This strategy offers a more pronounced effect on the phenotype than using individual QTNs in wheat breeding programs. The notable concentration of significant QTNs on chromosomes 1A, 1B, 3B, and 5B near genes involved in gliadin, glutenin, and starch synthesis underscores the potential of these genomic regions in improving WAC. Low molecular weight glutenin (*Glu-B3*) gamma gliadins (*Gli-A1-3, Gli-B1-3*), and delta gliadins (*Gli-B1-1*) have been located near the significant QTNs, as have starch synthase genes (SS) on chromosomes 1A and 6B. These genes are considered potential candidates for affecting WAC in wheat. Breeders could leverage these QTNs for marker-assisted selection to improve multiple traits simultaneously.

The current work has demonstrated the strong potential of multivariate genomic prediction for improving water absorption capacity in hard winter wheat. By including easily obtainable covariates in genomic prediction, the accuracy of SRC-W prediction was significantly improved, highlighting the importance of carefully selecting traits when constructing prediction models. Although incorporating flour yield values as covariates may pose challenges due to time-consuming and costly phenotyping processes, the SRC-W + SKCS model presented in this study offers a practical and cost-effective solution.

The integration of fixed effects such as *Bx7*^{OE} and GWAS-identified markers significantly enhances prediction accuracy for end-use quality traits such as bake mixing time and mixing tolerance. The effectiveness of including *Bx7*^{OE} KASP marker data in predicting BakeMT and MixoT highlights the role of glutenin genes on these traits. Since gluten genes have multiple effects on end-use quality traits, including water absorption capacity, bake mixing time, and Mixograph tolerance, selection for higher gluten (selection for gluten genes) genotypes could potentially improve the overall end-use quality of the wheat. The *Bayesian Lasso* method is a more effective approach than *GBLUP* for predicting baking traits in wheat, as evidenced by its superior performance in both cross and forward-validation. While the incorporation of specific genetic markers like GWAS-identified markers and *Bx7*^{OE} is key to elevating prediction accuracy in wheat breeding programs, the choice of predictive model should be tailored to the trait's genetic heritability.