Genetic and genomic tools for improving end-use quality in wheat

by

Emily Elizabeth Delorean

B.S., Colorado State University, 2014
M.S., Colorado State University, 2016

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Interdepartmental Genetics
College of Agriculture

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2021

Abstract

Wheat accounts for 20% of daily caloric intake of the world population and has one of the widest cultivation distributions of any crop. With increasing demand for both quantity and quality, wheat yields must increase while also maintaining acceptable end-use quality. However, measuring end-use quality is complex, requires large volumes grain and significant effort. The overarching goal of this dissertation research was to develop genetic and genomic tools to facilitate breeding for end-use quality in wheat.

Building on initial work with genomic prediction of wheat quality, we continued application of genomic prediction models to the International Maize and Wheat Improvement Center (CIMMYT) wheat breeding program. For practical application in the breeding program to advance selection, we focused on forward prediction in each cycle of the bread wheat program. Models were built on 12 years of past data including over 18,000 entries with quality data. Predictions for 10,000 yield trial lines were generated each year for selection, with forward prediction accuracies of 0.40 to 0.73, and approached heritability. This is one of the largest scale applications of genomic selection.

We also studied the interaction of climate change and the important quality genes, high-molecular weight glutenins (HMW-GS) and low-molecular weight glutenins (HMW-GS). A diverse panel of 54 CIMMYT wheat varieties were grown in 2 levels of drought stress, heat stress and optimal growth conditions. Quality traits, HMW-GS and LMW-GS alleles were measured. We fit a mixed linear model for each quality trait with HMW-GS, LMW-GS, environment, and the interactions of those as predictors. Overall, the superior glutenin alleles either maintained or increased quality in stressful environments. This work confirmed that superior alleles should always be selected for, regardless of target environment.

To increase the genetic diversity for wheat quality, we analyzed *Glu-D1* gene diversity on the wheat D genome donor, *Aegilops tauschii*. We constructed *Glu-D1* molecular haplotypes from sequence data of 234 *Ae. tauschii* accessions and found 15 subclades and over 45 haplotypes, representing immense gene diversity. We found evidence that the *5+10* allele originated from a newly described Lineage 3 of *Ae. tauschii*, further supporting that this unique lineage contributed to modern bread wheat. We also observed rare recombinant haplotypes between the x and y subunits of any HMW-GS locus. This work will facilitate incorporation of *Ae. tauschii Glu-D1* alleles into modern wheat.

Given that certain HMW-GS alleles are highly desirable, we set out to develop a high-throughput, high resolution genotyping method for HMW-GS alleles that would fit within genotyping already done for genomic prediction models. This ‘sequence based genotyping’ approach uses diagnostic k-mers developed to predict alleles in skim-sequenced breeding material. Prediction accuracies for *Glu-D1* and *Glu-A1* were very good, but lower for the *Glu-B1* alleles where many alleles are highly related. Overall, SBG offers a high throughput method to call alleles from existing data.

These genetic and genomic tools developed and implemented for end-use quality selection in wheat offer promising resources for continued improvement of both yield and quality in wheat breeding.

Genetic and genomic tools for improving end-use quality in wheat

by

Emily Elizabeth Delorean

B.S., Colorado State University, 2014
M.S., Colorado State University, 2016

A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Interdepartmental Genetics
College of Agriculture

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2021

Approved by:

Major Professor
Jesse Poland

Copyright

Abstract

Wheat accounts for 20% of daily caloric intake of the world population and has one of the widest cultivation distributions of any crop. With increasing demand for both quantity and quality, wheat yields must increase while also maintaining acceptable end-use quality. However, measuring end-use quality is complex, requires large volumes grain and significant effort. The overarching goal of this dissertation research was to develop genetic and genomic tools to facilitate breeding for end-use quality in wheat.

Building on initial work with genomic prediction of wheat quality, we continued application of genomic prediction models to the International Maize and Wheat Improvement Center (CIMMYT) wheat breeding program. For practical application in the breeding program to advance selection, we focused on forward prediction in each cycle of the bread wheat program. Models were built on 12 years of past data including over 18,000 entries with quality data. Predictions for 10,000 yield trial lines were generated each year for selection, with forward prediction accuracies of 0.40 to 0.73, and approached heritability. This is one of the largest scale applications of genomic selection.

We also studied the interaction of climate change and the important quality genes, high-molecular weight glutenins (HMW-GS) and low-molecular weight glutenins (HMW-GS). A diverse panel of 54 CIMMYT wheat varieties were grown in 2 levels of drought stress, heat stress and optimal growth conditions. Quality traits, HMW-GS and LMW-GS alleles were measured. We fit a mixed linear model for each quality trait with HMW-GS, LMW-GS, environment, and the interactions of those as predictors. Overall, the superior glutenin alleles either maintained or increased quality in stressful environments. This work confirmed that superior alleles should always be selected for, regardless of target environment.

To increase the genetic diversity for wheat quality, we analyzed *Glu-D1* gene diversity on the wheat D genome donor, *Aegilops tauschii*. We constructed *Glu-D1* molecular haplotypes from sequence data of 234 Ae. tauschii accessions and found 15 subclades and over 45 haplotypes, representing immense gene diversity. We found evidence that the *5+10* allele originated from a newly described Lineage 3 of *Ae. tauschii*, further supporting that this unique lineage contributed to modern bread wheat. We also observed rare recombinant haplotypes between the x and y subunits of any HMW-GS locus. This work will facilitate incorporation of *Ae. tauschii Glu-D1* alleles into modern wheat.

Given that certain HMW-GS alleles are highly desirable, we set out to develop a high-throughput, high resolution genotyping method for HMW-GS alleles that would fit within genotyping already done for genomic prediction models. This 'sequence based genotyping' approach uses diagnostic k-mers developed to predict alleles in skim-sequenced breeding material. Prediction accuracies for *Glu-D1* and *Glu-A1* were very good, but lower for the *Glu-B1* alleles where many alleles are highly related. Overall, SBG offers a high throughput method to call alleles from existing data.

These genetic and genomic tools developed and implemented for end-use quality selection in wheat offer promising resources for continued improvement of both yield and quality in wheat breeding.

Table of Contents

List of Figures

List of Tables

Acknowledgements

This work would not have been possible without the support and guidance from mentors past and present, and family and friends. Thank you to the fellowships that supported my dissertation, the Kansas State University Genetics Doctoral Program Fellowship and the Bayer's Beachell-Borlaug International Scholar Program Fellowship. It absolutely would not have been possible without the network of brilliant scientists, technicians and support staff at Kansas State University and our collaborating partners. I would especially like to thank everyone at the International Maize and Wheat Improvement Center for their tireless and enthusiastic work everyday to address global hunger. Thank you to everyone involved in the Kansas State University Wheat Genetic Resource Center. Your work made the wheat wild relative work presented here possible through thorough record keeping and patient cultivation of often uncooperative wild plants. Thank you to Dr. Allan Fritz and Dr. Mary Guttieri who put in the long hours to harness wild wheat genetic diversity. Thank you to everyone in the Wheat Genetics lab, the lab to which I have proudly and lovingly belonged, for all the things big and small that supported this work and shaped me into the scientist I am today. It has been a great, joyous adventure to work alongside all of you. Thank you most of all to Jesse Poland, my advisor and mentor. Advising is difficult, mentoring even more so, but still you always found the time and the energy. You made the Wheat Genetics lab a wonderland of scientific questions and the data we needed to answer them. Your guidance made me a sharper scientist and your mentorship made me a better person. Thank you. Thank you to my husband, Travis, who has always whole heartedly supported my decisions and ambitions. You bring light and happiness to my life. Thank you to my family, Travis’s family, our friends, and our friends who have become family. I am lucky to know each and every one of you, and even luckier still to have you in my life. I hope

you, the reader, can glean some new knowledge or inspiration from this dissertation though it is not perfect. Thank you for reading it.

Chapter 1 - Introduction to the Genetics of Wheat Quality

Wheat accounts for 20% of daily caloric intake of the world population and has one of the widest cultivation distributions of any crop (P. R. Shewry, Halford, & Lafiandra, 2003). Unlike other major cereals, wheat is almost exclusively processed into baked goods before consumption. Wheat production and nutrition must increase to meet projected demands of the growing human population. Yet, yields are constrained by abiotic and biotic stresses, and further threatened by a changing climate (Reynolds, 2010). Furthermore, wheat could be more nutritious and more efficient with resources in order to better meet the needs of global food security. Both advancements in wheat management techniques and improvement of wheat through new varieties are required to make wheat more resilient to stressors and sustainable to produce.

One might argue that the unique properties of wheat flour that allow it to be baked into goods such as bread are in fact the driving force for the widespread cultivation and consumption of wheat. Which in turn, would mean that the gluten, which is major underlying factor of wheat end-use quality, is the reason wheat is so popular. Gluten is a large heterogenous protein matrix found in Triticeae tribe. The proteins of gluten are divided into three classes, the high molecular weight glutenin grain storage proteins (HMW-GS), the low molecular weight glutenin grain storage proteins (LMW-GS) and the mostly monomeric gliadins. The HMW-GS proteins form the backbone of the matrix through disulphide bonds between themselves. LMW-GS proteins act as chain branchers and terminators within the matrix, also through disulphide bonds. Gliadins primarily interact through hydrostatic forces with the matrix.

HMW glutenins are the most widely studied of the gluten proteins and allelic differences have historically have been shown to have the largest impact on quality. Each genome in wheat contains one HMW glutenin locus on the long arm of the group 1 chromosomes at which two genes are located, an x subunit and a y subunit. The SDS-PAGE mobility of these subunits together gives rise the HMW glutenin allele names, e.g. *Glu-D1 5x+10y.* In domesticated wheat, the *Glu-1 Ay* subunit is always silent due to a transposon insertion, except in rare instances where an active y subunit has likely entered the gene pool from a wheat wild relative (Margiotta et al., 1996). HMW-GS genes consist of highly conserved N and C terminal regions, and a highly variable central repetitive domain. The central repetitive domain of the x subunits is comprised of pentapeptide and nonapeptide repeats whereas the y subunit also contains a hexapeptide motif. Size differences between allelic proteins are most likely due to losses or gains of repeat units due to unequal crossing over.

LMW glutenins are the next most widely studied of the gluten proteins. These genes are located on the short arms on the group 1 chromosomes and share the locus with gliadin genes. There are typically about 4 LMW genes in each locus. Due to the increased complexity, these proteins are named by the fingerprint of their SDS-PAGE bands, e.g. *Glu-A3 a*.

Gliadins are the most complex and least studied of the gluten proteins. The exact number of gliadin genes, not alleles, in wheat has not been elucidated, but estimates come to between 60 and 200. The genes for the omega, gamma and delta gliadins share the ~ 13 mega-base LMW locus on the short arms of the group 1 chromosomes, but when speaking about the gliadins, the locus is called *Gli-1*. The alpha/beta gliadins are located on the short arms of the group 6 chromosomes, at the *Gli-2* locus. Gliadin alleles are also named for their electrophoretic fingerprints.

Gliadins are regarded as monomeric proteins and glutenins as polymeric. Gliadins are monomeric because their cysteine residues, except for omega gliadins which lack cysteines altogether, form intramolecular (within molecule) disulphide bonds to stabilize the protein structure. Glutenins also contain intramolecular disulphide bonds between cysteines, but also some free cysteines that form intermolecular (with other molecules) bonds. LMW glutenins are more similar to gliadins in amino acid sequence than to HMW glutenins. The key difference being that LMW glutenins have cysteine residues that interact with HMW glutenins.

HMW glutenins are thought to form the backbone of the gluten protein network to which LMW glutenins act as chain terminators. The number of free cysteines in HMW glutenins is correlated with quality, e.g. *1Dx5* that has an extra free cysteine compared to 1Dx*2* and *1Bx20* is missing two conserved cysteines. The density of disulphide bonds and non-covalent interactions contribute to dough elasticity. Increasing the number of intermolecular bonds, through genetic factors or protein concentration, increases elasticity. Gliadins interact with the protein matrix through non-covalent bonds to contribute to dough viscosity. Together, the balance of gluten protein to other dough constituents and the balance of monomeric/polymeric gluten influence dough viscoelastic properties.

Both dough strength and elasticity are determined by the structure of the gluten matrix. The backbone of the gluten matrix is formed by the covalent disulphide bonds between cysteine residues in HMW glutenin proteins. HMW glutenins typically possess between 4 and 6 cysteine residues, with the number and positions likely being determinant for the ability of that protein to covalently bond with another and thus to continue chains within in the matrix. LMW glutenins are structurally very similar to third class of gluten proteins, the gliadins, but were originally classified as glutenins because they covalently participated in the gluten matrix. Although some gliadins possess cysteine residues, most form intrachain disulphide bonds and therefore are not covalently bound in the gluten matrix. LMW glutenins typically possess an odd number of cysteine residues which allow them participate covalently in the gluten matrix (P. R. Shewry et al., 2003). Both LMW glutenins and especially gliadins increase the extensibility of dough interrupting the HMW glutenin backbone and providing an avenue for molecular rearrangement of the gluten matrix during applied mechanical stress.

HMW glutenins are encoded by a relatively simple locus on the long arm of the group one chromosomes in wheat. Hexaploid wheat, comprised of the A, B and D genomes, contains three HMW glutenin loci, *Glu-A1, Glu-B1* and *Glu-D1*. Each locus harbors two HMW glutenin genes known as the x and y subunit, that are separated by 57 kilobases pairs (kb) in *Glu-D1* to a couple hundred kilobase pairs as seen in *Glu-B1* and *Glu-A1* (Walkowiak et al., 2020). The x and y subunit genes arose from a duplication event that occurred approximately 7.2-10 million years ago, before the origin of the ancestral wheat genomes (ABD) at approximately 5.0 – 6.9 MYA (Allaby, Banerjee, & Brown, 1999). Although each locus contains two genes for a cumulative total of 6 in common wheat and 4 in durum wheat, some genes are inactive, either through a transposon insertion (as is the case with all *Glu-A1y* genes in common and durum wheat) or an early stop codon (*Glu-A1x* in the null allele or *Glu-B1y* in the *7x* allele). There also a notable instance of a complete duplication *Glu-B1x* gene that occurs in the *Glu-B1 7OE+8* allele.

Allelic differences in all three gluten proteins contribute to the conformation of the gluten matrix. The position and number of cysteine residues being an obvious and exciting example. But multiple studies have also shown that differences in promoter elements between alleles lead to earlier and greater accumulation of protein in the grain (for example *Glu-D1 5+10* as opposed to *2+12*). Additionally, the duplicated x subunit in the *Glu-B1 7OE+8* is associated with higher accumulation of the 7x protein and therefore is associated with better end use quality.

Of the HMW glutenin alleles, the greatest impact on end-use quality is commonly attributed to the *5+10* allele at the *Glu-D1* locus. Given the highly sought after quality characteristics imparted by *5+10*, numerous studies have set out to understand what characteristics of this allele are responsible for such superior quality characteristics. The 5x protein has a unique cysteine residue just within the central repeat domain. The accumulation timing and amount for this allele is greater than that of the other *Glu-D1* alleles, in particular *2+12*. The central repeat domains appear to not possess any strange or unique characteristics compared to HMW glutenin alleles.

Which HMW glutenin locus has the greatest effect on quality has been a point of interest for numerous studies (Lawrence, MacRitchie, & Wrigley, 1988). One study in particular showed rather elegantly the importance of each locus through EMS knockouts of Glu-A1, Glu-B1 and Glu-D1 in the Chinese wheat cultivar Xiaoyan 81 (Wang et al., 2017). The researchers measured the effect of the loss of an individual locus through not only gluten strength parameters, but also through sodium dodecyl sulfate (SDS) unextractable polymeric protein and insoluble glutenin. The latter two are measurements of the glutenin macropolymer, which consists of HMW and LMW glutenins. Knock out of Glu-D1 resulted in the greatest negative effect on the GMP size. It was interesting that Glu-D1 knockout still had the greatest effect on the GMP because Xiaoyan 81 carries is the inferior the *2+12* allele associated with inferior quality. One would have expected such results with 5+10, but seeing a similar trend for even the inferior *2+12* indicates that Glu-D1 overall does indeed have a greater effect on quality than the other loci and that the effects attributed to *Glu-D1* are not solely due to any individual allele.

Given that *Glu-D1* has two genes, the x and y subunit, the same researchers then wondered which of the two has the greater effect on the glutenin macropolymer. With EMS knockout mutants of Xiaoyan 54, a parent of Xiaoyan 81 with the same HMW glutenin alleles, they found that the loss of *2x* of *Glu-D1* decreased the relative amount of the insoluble glutenin to a greater extent than the loss of *12y*. Maybe the most exciting result was that the other HMW glutenin proteins (*1Ax1*, *1Bx14*, *1By15*) and the LMW-GS in the IG fraction were also reduced when *1Dx2* or *1Dy12* were knocked out. This indicates that the *Glu-D1* proteins promote incorporation of other glutenin proteins into the glutenin macropolymer. This was further supported by comparing the same effect for *Glu-A1* and *Glu-B1* knockouts, where they found that the effect was much less pronounced. The glutenin proteins missing in the insoluble fraction were found in the soluble fraction (indicating that they were not covalently bound in the glutenin macropolymer).

The conclusion that the 2x protein has a greater impact on the continuity of the GMP is supported by other studies that have found x subunit proteins form dimers with other x subunits and with y subunits, but y-y dimers have not been detected (Lindsay & Skerritt, 1998). Although Werner et al. (1992) were only able to identify homodimers of x-type HMW-GS. Additionally, the studies concluded that the *Glu-D1* subunits were more important to the glutenin backbone than *Glu-B1* and *Glu-A1* because the *Glu-D1* subunits were almost always present in the most difficult to reduce protein fractions.

Some of the earliest work indicating that certain HMW glutenin proteins form oligomers bound so strongly to each other that they are resistant to reducing agents was by (Lawrence & Payne, 1983). They showed that the usually ignored bands at the top of an SDS-PAGE gel are actually oligomers, perhaps dimers, of HMW glutenin subunits that survived reduction with SDS. By applying a stronger reducing agent and separating those bands on a 2D SDS-PAGE gel they were able to deduce the identities of some of the HMW glutenin proteins that participated in these oligomers. Interestingly, they found that some bands consisted of x-x oligomers while others were x-y, but they never observed a y-y combination. They found that the oligomer bands were consistent across varieties and were HMW glutenin allele specific. Their findings showed that particular HMW glutenin alleles combine with each other in predictable ways and we can deduce that these affinities underlie the gluten strength attributed to specific alleles.

The molecular mechanisms imparting the extensibility and elasticity of gluten matrix have also been something of curiosity. After learning from scanning electron microscopy of *Glu-B1 20x+20y* (durum wheat) that the HMW glutenins have a rod like appearance and deducing from analysis of the central repeat domain that the repeats appear have a beta turn structure, it was thought that the HMW glutenins might act like springs. Deforming under pressure and then retaking the spring shape under relaxation. However, further analysis of the disulphide linkages showed that the structure of glutenins is more of globular nature, especially in the C and N terminal regions where intermolecular bonds take place.

Although gluten genes and proteins have been widely studied and are considered the major determinants of quality, other factors also a play a role. Quality traits are influenced by both genetics and environment, as we will explore in Chapter 2 of this dissertation. Higher protein concentration results typically in greater gluten strength. Though there exist some genetic control of protein concentration, such as the GPC genes, this trait is primarily controlled by environment. Drought conditions typically result in high grain protein, but this is because starch accumulation has been limited and therefore yield decreased. Grain protein is also impacted by nitrogen and sulfur availability. If these nutrients are restricted, then both protein and yield are limited. However, increasing these nutrients above a threshold will not increase protein concentration as yield (starch) is also increased.

End-use quality research and breeding are important in achieving global food security. Not only do we need to develop higher yielding crops, more resilient and sustainable crops, but those crops must also be nutritious with end-use quality sufficient to make the products they are intended for. The issue that wheat breeders face is that screening for end-use quality is labor intensive, expensive and requires a large amount of grain. Therefore, measuring end-use quality only occurs very late in the variety development and on relatively few candidates. The real world impact of this is that the newest, best and highest yielding varieties often make poor bread. Genetic studies have shown that end-use quality in wheat depends largely on which gluten proteins a variety possesses, among other smaller effect genes. Profiling gluten genes is currently done with a protein gel. While faster than quality testing, these gels are still too slow and too low resolution to broadly screen breeding candidates. With this information, my research goals were to develop molecular biology and bioinformatic tools that allow breeders to select for improved end-use quality earlier in the program (and therefore on far more candidates) and to tailor their varieties for the target growers based on the gluten genes.

My specific questions were (1) are gluten genes resilient to increasing heat and drought in the changing climate, (2) can we reliably predict the quality of candidate lines from their genetics, (3) can we determine the gluten genes of a variety from only its DNA sequences, (4) do wheat wild ancestors possess a greater diversity of gluten genes?

Chapter 2 - Glutenin Gene by Environment Interaction for Quality in Bread Wheat

Introduction

Wheat is major staple crop on which people rely for 20% of their carbohydrates and 20% of the protein. Wheat is unique among food crops because it is not consumed in its raw or whole cooked form, but is first milled into flour then the flour is used to make a diversity of foods. This ability is primarily attributed to a gluten storage protein family found only in the Triticeae tribe which interact to form a matrix when hydrated and mixed during dough development. The matrix is strong, extensible and elastic, allowing the baker to form the dough or batter. Once baked, and water removed, the gluten matrix holds its form as bread, cake, crackers or noodles.

Gluten is made up of two classes of proteins, called glutenins and gliadins. The viscoelastic properties of dough are primarily determined by the ratio of glutenin to gliadin (Sissons, Ames, Hare, & Clarke, 2005; Southan & MacRitchie, 1999), available cysteine residues for intramolecular disulphide bonds, primary structure of individual proteins (S. Li et al., 2020), and overall protein proportion in grain. Gliadins confer mainly extensibility and flowability whereas glutenins confer strength and elasticity Shifting of the glutenin/gliadin ratio and the overall amount of protein changes the extensibility and strength which determines end-use profile. This ratio is controlled by relative gene expression which is inversely related (De Santis et al., 2017) and influenced by environmental conditions (Altenbach, Tanaka, & Seabourn, 2014; De Santis et al., 2017; Giuliani et al., 2015; Hurkman, Tanaka, Vensel, Thilmony, & Altenbach, 2013; Yongfang Wan, Gritsch, Hawkesford, & Shewry, 2014).

Glutenins are divided into two subclasses, high molecular weight glutenins and low molecular weight glutenins. HMW glutenins are present on the long arms of 1A, 1B and 1D. At each locus are two HMW glutenin genes, an x subunit and a y subunit. LMW glutenins are more diverse and the loci composition more complicated. Between 1 and 4 functional LMW glutenin genes are present at each of the loci on the short arms of 1A, 1B, and 1D. These loci are shared with the γ-gliadins, δ-gliadins, and ω-gliadins. LMW glutenin gene sequences and protein structures more closely resemble γ-gliadins, but possess cysteine residues available for disulphide bonding and therefore are polymeric proteins.

The gluten matrix is formed through intermolecular disulphide bonds primarily between HMW glutenins and LMW glutenins. HMW glutenins have between 5 and 7 cysteine residues available to form disulphide bonds with neighboring HMW and LMW glutenins to form the polymeric gluten matrix. Monomeric gliadins contain an even number of cysteine residues that form only intramolecular bonds, however, some gliadin proteins have an uneven number of cysteines and therefore act as chain terminators.

Gliadins are equally important determinants of wheat quality and also under well-defined genetic control forming primarily monomeric proteins without available cysteine residues. There have been 5 families of gliadins described which include the γ-, δ-, ω-, α-, and β- gliadins. While the α /β gliadins are on the group 6 homologous chromosomes, γ-, δ-, ω-gliadins share loci with the LMW glutenins. The interaction of these proteins forms the large heterogenous protein matrix called gluten.

Beyond genetic factors on the amount and type of glutenin, quality is highly influenced by environmental growing conditions with stress conditions both increasing and decreasing quality. Drought and heat stress generally decrease yield by limiting starch accumulation in the grain which results in a higher protein concentration and therefore higher processing and end-use quality. This is because protein accumulates in the wheat kernel within 15 - 20 days post anthesis, but there after increases in grain size are primarily due to starch accumulation. Beyond protein concentration in the grain, the quality and structure of starch and protein are highly impacted by environment. In particular, heat stress decreases milling quality by causing shriveling and increased hardness in the grain which results in higher damaged starch during milling.

A strong genotype by environment interaction for quality traits is well documented in literature, however little work has explored the impact of the gene gluten x environment interaction on quality. In order to assemble the best possible varieties for the target environments, the genes underlying the genotypes must be elucidated. We set out to (i) determine if gene x environment exists for stress environments, (ii) if all three HMW glutenin and all three LMW glutenin (and also gliadin) loci show gene x environment interaction, and (iii) estimate which alleles at these loci are favorable and most stable across environments or best suited for specific environments.

Methods

Plant Material

54 spring bread wheat varieties representing the past 50 years of wheat breeding program at CIMMYT were grown in Obregon, Mexico during cropping seasons 2012-2013 and 2013-2014. To test environmental effects on quality parameters, the lines were grown in 6 environments, representing two types of stress: heat and drought. The environments were selected to simulate the mega-environments (ME) targeted by the CIMMYT wheat breeding

program (Hernandez 2017) including ME1, ME2, ME4, and ME5. The types of stress environments are summarized in Table 1, which included optimal irrigated conditions (ME1), basin irrigation (ME2), mild heat stress (ME5), severe heat stress (ME5), mild drought stress (ME4), severe drought stress (ME4). Each variety was grown in three replicates in a randomized complete block design. Weather data from the experimental station recorded a negligible amount of precipitation during the growing period.

Heat and drought stress environments were compared to the control environments for optimal yield. The control environments were both flood irrigated with more than 500 mm water applied over the growing season. These two environments differed in planting bed design, with optimal irrigation being planted in raised beds and basin irrigation being flat planted. Optimal irrigation experiment also received more nitrogen than all other experimental environments. Plants were sown in November for control and drought stress environments to achieve maximum temperatures of 31-32 ºC during grain filling in March and April. The heat stress environments were sown later in January for mild heat stress and in February for severe heat stress, to achieve higher maximum temperatures during grain filling in May of 35-39 C. Reduced irrigation was applied to drought stress environmental fields, with mild drought stress receiving 300 mm and severe drought receiving 180 mm over the growing season. Details of field experiment can be found in (Guzmán et al., 2016; Hernández-Espinosa et al., 2018).

Gluten Allele Determination

Genotyping for the glutenin alleles was conducted using sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) as described by (Pena, Amaya, Rajaram, & Mujeeb-Kazi, 1990). The glutenin and gliadin protein fractions were separated as described by

(N. K. Singh, Shepherd, & Cornish, 1991) with a modification. The concentration of the separation gel was 12.5% with a 0.97% cross linker and the gel was run with a current of 12.5 mA. Alleles were identified using the nomenclatures proposed by (Payne & Lawrence, 1983) for high molecular weight glutenins and by (Branlard, Dardevet, Amiour, & Igrejas, 2003) for low molecular glutenins.

High Molecular Weight Glutenin Molecular Markers

Genotyping-by-sequencing was coincidentally conducted on a portion of the 54 wheat lines in this study during the course of genotyping for the Chapter 3 study of genomic selection models. See the materials and methods section of Chapter 3 for details of how variants were called. Briefly, genotyping by sequencing reads were aligned to the Chinese Spring reference genome (RefSeq v1.0) and single nucleotide polymorphisms were called with TASSEL v5. Variants within the *Glu-A1, Glu-B1* or *Glu-D1* locus were examined to determine if molecular haplotypes could be elucidated. The locus regions were defined as the x and y subunit genes, the space between the subunits and 2 kb flanking up or down stream. The regions were chr1D: 412158785-412221631 for *Glu-D1,* chr1B:555763126-555937716 for *Glu-B1,* and for *Glu-A1.*

Statistical Analysis

Statistical analysis was performed in R version 3.5.2. Correlation between quality traits were found with the cor (R Core Team, 2018) package and plotted with the corrplot package (Wei and Simko, 2017). Collinearities between glutenin loci, that is how often certain alleles cooccurred in wheat varieties, were determined using chi square tests with xtabs (R Core Team, 2018). Nearly all glutenin loci exhibited pairwise collinearity, indicating that glutenin alleles at different loci were not random within varieties and allelic effect could be confounded. Fitting a model with all loci at once resulted in an overfit model and inflated effect estimates. Therefore, all loci were fit separately. Each quality trait was fit as separate response variable in the following model using lme4 package (Bates, Mächler, Bolker, & Walker, 2015).

$$y_{ijk} = \mu + e_i + g_j + \gamma_k + e\gamma_{ik} + \varepsilon_{ijk}$$

Where y_{ijk} is the adjusted value of a given quality trait, μ is the overall mean, e_i is the fixed effect of the i^{th} environment, g_j is the fixed effect of the jth allele for the given glutenin gene, γ_k is the fixed effect of the k^{th} year, $e\gamma_{ik}$ is the fixed effect of the interaction between the i^{th} environment and and k^{th} year, and ε_{ijkl} is the residual variance, where $\varepsilon_{ijkl} \sim\sim N(0, \sigma_e^2)$. The marginal means and contrasts were calculated using emmeans (Russel Length, 2018). Differences in quality traits between alleles within an environment were considered significant at $\alpha < 0.05$. Variance attributable to each glutenin locus was found by fitting the above model as a random effect only model and all glutenin loci simultaneously.

$$y_{ijk} = \mu + e_i + g_j + \gamma_k + (eg\gamma)_{ijk} + v_l + (ve\gamma)_{l(i(k))} + \varepsilon_{ijkl}$$

Graphical displays were created in R with ggplot2 (Wickham, 2016) and viridis (Garnier, 2018).

Results

High Molecular Weight Glutenin Markers

Given the results found in Chapter 5 of this dissertation that multiple *Glu-B1* haplotypes exist within several SDS-PAGE alleles, we attempted to resolve haplotypes of the HMW glutenin loci. We found that a portion of these wheat accessions were genotyped in the course of genotyping-by-sequencing of the CIMMYT wheat breeding program for making genomic predictions. However, few to no variants were found within the HMW glutenin loci. No SNPs were found in the Glu-A1 locus.

Within the 63 kb Glu-D1 locus region (chr1D: 412158785-412221631, which is 2 kb flanking the Glu-D1 locus) , 2 SNPs were detected (at 412169433 and 412169869, only 436 bp between the two). At both sites, the homozygous reference alleles were perfectly correlated with 5+10 and the homozygous alternative alleles were perfectly correlated with *2+12.* This was unexpected given that the reference genome used for alignments is Chinese Spring, which has *2+12.*

We found only 1 SNP within the 175kb Glu-B1 locus (chr1B:555763126-555937716) and it was not reliably predictive of any allele. The SNP (at 555765141) occurred in a homozygous state in GID5794687 (*13+16*) and heterozygous in GID13396 (*13+16*). These two accessions were the only two with the *13+16* SDS-PAGE allele, indicating that the variant site could be predictive for the allele, but not in this dataset.

Environmental Effect

Environment had a substantial effect on yield and quality parameters. Grain protein, flour protein and grain yield were highly correlated with each other (Figure 2.1), and were particularly affected by environment (Figure 2.2), with environmental variance accounting for up to 75% of the total variance (Table 2.2). The greatest negative effect on grain yield was severe heat stress (Figure 2.3). Wheat variety typically explained between 15 – 30% of phenotypic variance and the variety by environment interaction explained another 15-30%. The main effects of gluten alleles explained upwards of 25% of variation for quality traits (Tables 2.3-2.5). While the *Glu-A1* and *Glu-D1* allele by environment interaction explained a small, but measurable amount of variation for some quality traits, between 0.5 – 3.5%. Other gluten genes appeared to have a negligible interaction with environment, such as Glu-B1 which explained no variation for most traits.

Glu-D1 By Environmental Interaction

The swelling index of gluten was significantly different between the alleles in severe drought stress and the difference in SDS sedimentation was nearly significant (p-value = 0.058) in severe heat (Figure 2.4). GMP, UPP %, SIG, SDS sedimentation all attempt to measure the amount of polymeric protein in flour. Polymeric protein, the protein aggregates of HMW and LMW glutenins joined by disulphide bonds, are the intermediate building blocks of the gluten network that will be formed in the dough. Studies have shown that the amount of polymeric protein in flour is correlated with gluten strength in dough. The primary difference between the swelling index of gluten and the SDS sedimentation test is that the SDS soluble (i.e. gluten monomers) are removed in SIG before measuring gluten polymer swelling.

The impact of environment x *Glu-D1* on gluten strength measurements (Mixograph mix time and torque at peak, alveograph W and P/L, swelling index of gluten and loaf volume) was seen in the magnitude of the difference between alleles. *Glu-D1 5+10* always conferred stronger gluten than *2+12,* but the effect was most pronounced in severe drought stress for all gluten strength measurements. Severe heat stress also increased the effect size between alleles, but to a lesser degree and less consistently. Therefore, although we saw increased grain and flour protein associated with *5+10* under severe heat and drought, a clear impact on gluten strength wasn’t found. We hypothesized this may be due to controlling for flour protein content when modelling gluten strength parameters, but the same trends were found when dropping flour protein content from the linear model.

The environment x *Glu-D1* interaction accounted for the greatest variation in gluten strength measurements among all the environment x glutenin interactions. However, the attributable variance was still small, between 1.8 and 5.4% of the total variation. The main effect of *Glu-D1* often dwarfed the environment x *Glu-D1* variance, reaching upwards of 26% of the total variation. The variety x environment variance was also always lower than the *Glu-D1* variance of gluten strength traits, except for alveograph P/L. These results indicate that the genotype x environment variance of the quality parameters is partially, but not wholly attributable to the *Glu-D1* x environment interaction. From the results, we can clearly conclude that *Glu-D1 5+10* conferred superior gluten strength than *2+12* in all environments with some GxE interaction observed for Glu-D1 alleles. The beneficial effects of *5+10* were increased under stress environments, especially severe drought, but mitigated in the most favorable, high yielding environments.

Glu-A1 By Environmental Interaction

Flour protein concentration exhibited environment x *Glu-A1* interaction, with *Ax1* having a higher concentration than *Ax2** in severe heat and drought stress. The same trend was not seen in grain protein concentration. This may be due to differences in flour yield. The environment x *Glu-A1* interaction was significant for severe heat and severe drought, with allele *Ax2** having a higher flour yield.

Overall, gluten strength measurements were not significantly different between alleles in any environments for *Glu-A1*. The only exception was for alveograph W, where allele *Ax1* was higher in basin irrigation, severe heat and severe drought (Figure 2.5). As seen in *Glu-D1,* the measurement of glutenin polymer size (swelling index of gluten) was significantly different between alleles in severe drought and nearly significant in severe heat, with allele *Ax1* being higher than *Ax2*.* Although some gene x environment interactions were seen for protein concentration and gluten strength, no difference between alleles was detected for loaf volume. These results indicate that although the *Glu-A1* x environment interaction impacts glutenin polymer and flour protein, but does have a measurable impact on final bread making quality.

Discussion

Many of the differences in allele effects under heat stress are likely due to the shortened grain filling period. Previous studies have shown that glutenins begin accumulation in the grain earlier than gliadins resulting in a higher glutenin/gliadin ratio (Koga et al., 2016). Given that glutenins confer greater gluten strength, it is possible that shortening the grain filling period would lead to greater mixograph parameters and loaf volume. However, reports conflict as to the effect of heat stress on quality traits. (Irmak, Naeem, Lookhart, & MacRitchie, 2008; Naeem & MacRitchie, 2005) found that heat stress decreased end-use quality, while (Blumenthal et al., 1991; Y. Li, Wu, Hernandez-Espinosa, & Peña, 2013; Maphosa, Langridge, Taylor, Emebiri, & Mather, 2015) found increased quality under heat stress. These differences could be due to the stage when the heat stress was applied, but also due to the glutenin allele composition of the varieties.

In a study of NILs for 5 *Glu-B1* alleles (*6+8, 7+8, 7+9, 15+15, 17+18*) in a Chinese cultivar found that 6+8 had the significantly slowest rate of accumulation that ultimately led to a lower unextractable polymeric protein (UPP) content (T. Liu et al., 2016). Additionally, other studies with NILs showed similar results for *Glu-D1 2+12 vs 5+10, Glu-B1 20+20 vs 7+9* and *Glu-B1 7+8* vs *7OE+8* (Naeem & MacRitchie, 2005) (S. Li et al., 2020). All four alleles with delayed UPP in grain filling are associated with poor quality characteristics. It therefore could be hypothesized that these alleles interact more negatively with heat stress than their counterparts.

Previous studies with near isogenic lines differing between *5+10* and *2+12* have shown that *5+10* has an earlier and more rapid unextractable polymeric protein accumulation (Irmak et al., 2008). It was hypothesized that because of this, *5+10* lines would have better quality characteristics under environmental stresses that shorten grain filling duration, such as severe heat and drought. To test whether the earlier and more rapid accumulation of *Glu-D1* glutenins provided better quality characteristics, we predicted that the superior gluten strength effects of *5+10* would be greater under severe heat and drought stress. This was consistent with the previous physiology model as we observed that *Glu-D1 5+10* was associated with significantly higher grain and flour protein content than *Glu-D1 2+12* under severe heat stress and severe drought stress.

To evaluate confounding quality factors, the two *Glu-D1* alleles did not differ in test weight, thousand kernel weight or grain yield. NIL studies of *Glu-D1* alleles have shown that *5+10* is associated with both a greater gluten macropolymer and a greater flour protein concentration (Don, Lookhart, Naeem, MacRitchie, & Hamer, 2005).

The swelling index of gluten was significantly different between the alleles in severe drought stress and the difference in SDS sedimentation was nearly significant (p-value = 0.058) in severe heat (Figure 2.4). GMP, UPP %, SIG, SDS sedimentation all attempt to measure the amount of polymeric protein in flour. Polymeric protein, the protein aggregates of HMW and

LMW glutenins joined by disulphide bonds, are the intermediate building blocks of the gluten network that will be formed in the dough. Studies have shown that the amount of polymeric protein in flour is correlated with gluten strength in dough. The primary difference between the swelling index of gluten and the SDS sedimentation test is that the SDS soluble (i.e. gluten monomers) are removed in SIG before measuring gluten polymer swelling.

Conclusion

In this study we observed the large and consistent main effect of *Glu-D1* on the quality parameters for bread wheat. We also observed that the *5+10* allele at *Glu-D1* was superior in all environments. Most importantly, we found that a detectable gene x environment interaction exists for *Glu-D1* exists. The gene x environment interaction was due to an increased positive effect of the *5+10* allele under environmental stress. We also observed an environmental interaction between *Glu-A1* alleles, with *Ax1* being superior to *Ax2** for gluten strength. These results support the approach for breeders uniformly selecting *5+10* allele regardless of the target environment and suggest that *Ax1* may be desirable over *Ax2**.

Table 2.1 - Environmental treatment summary.

Heat and drought stress environments were compared to the control environments for optimal yield. The two control environments were referred to as ‘optimal’, which was fully irrigated with a drip system, and as ‘basin irrigation’, which was flood irrigated. The heat stress environments, ‘mild heat stress’ and ‘severe heat stress’ were subjected to maximum daily temperatures between 35 – 39 °C during grain filling. The drought stress environments, ‘mild heat stress’ and ‘severe heat stress’, were planted in November to avoid heat stress, and received reduced irrigation levels of 300 mm for mild drought and 180 mm for severe drought. Details of field experiment can be found in Hernández-Espinosa et al. (2017) and Guzman et al. (2016).

Environment	Planting date	Irrigation (mm)	Nitrogen (kg/ha)	Max temp during grain filling (°C)
Optimal irrigation	November	> 500 (raised bed - flood)	300	31-32
Basin irrigation	November	> 500 (flat - flood)	200	31-32
Mild heat stress	January	500 (drip)	200	35-39
Severe heat stress	February	500 (drip)	200	35-39
Mild drought stress	November	300 (drip)	200	31-32
Severe drought stress	November	180 (drip)	200	31-32

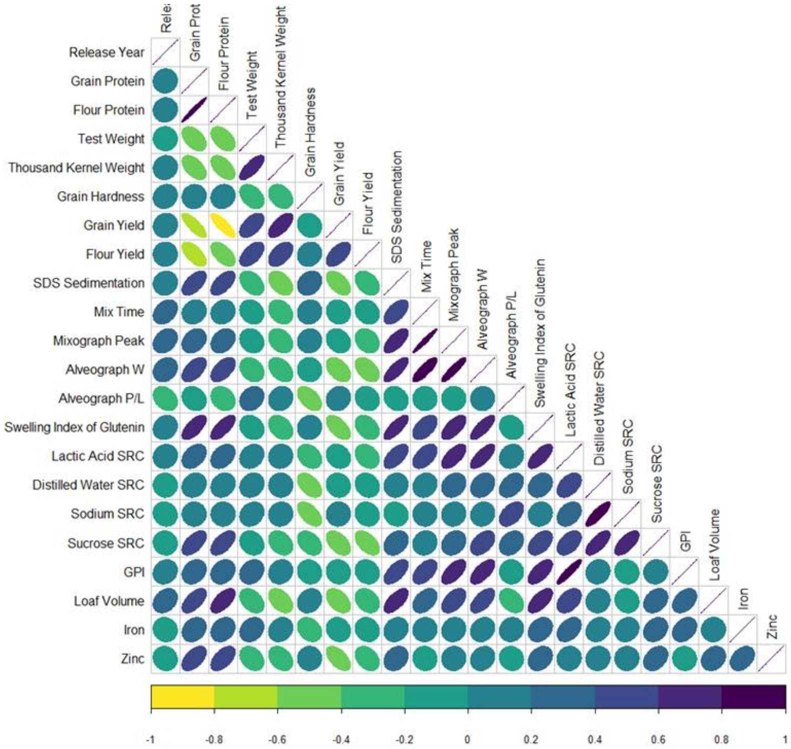


Figure 2.1 - Correlation among grain traits.

Phenotypic correlations over all 6 environments and both years showed that in general the quality traits were positively correlated with each other and either not correlated or slightly inversely correlated with yield traits. The exception being that grain and flour protein content were strongly negatively correlated with grain yield. Grain yield had a strong negative correlation with flour protein, but very little to no correlation with gluten strength indicating that high yield does not necessarily equal poor quality.

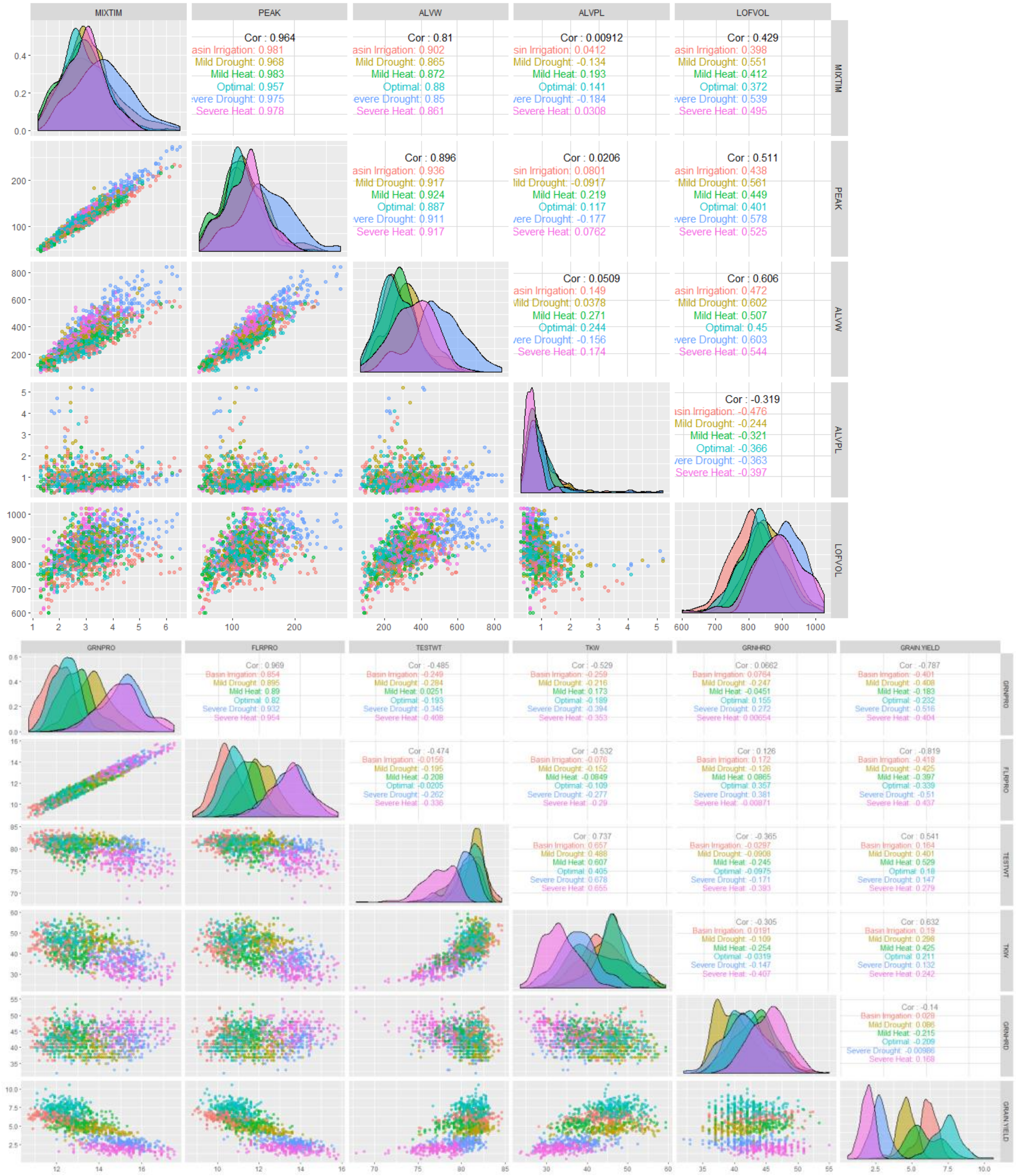


Figure 2.2 - Trait distributions and correlations dissected by environment.

Diagonal elements show the trait distributions dissected by the 6 environments. Optimal in teal, basin irrigation in salmon, mild drought in yellow, mild heat in green, severe drought in blue, severe heat in violet. The lower off diagonal elements show scatter plots between the trait listed on the x axis and the trait on the y axis, also dissected by environment. The upper off diagonal elements show the corresponding Pearson's correlation coefficients.

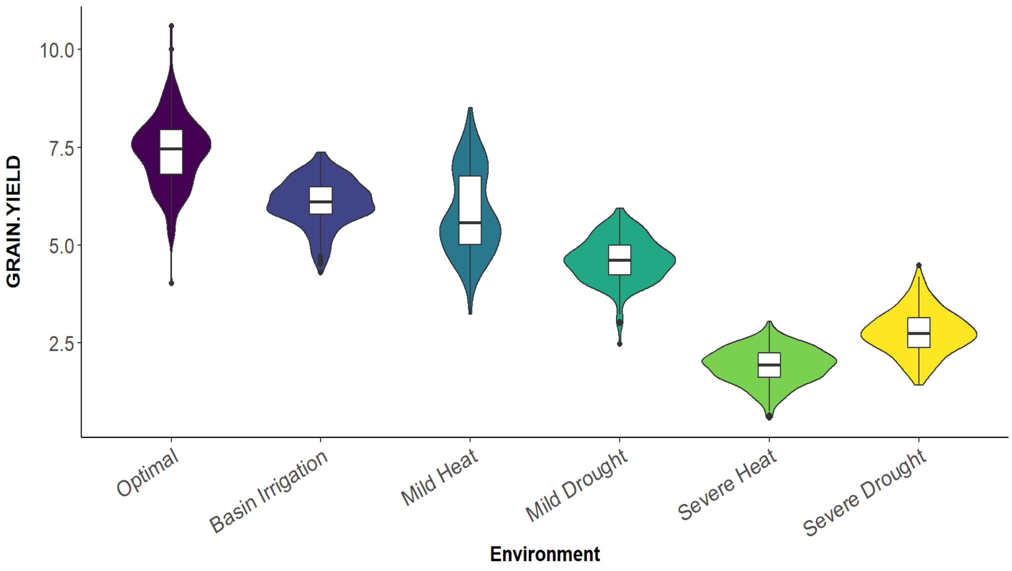


Figure 2.3 - Grain yield distributions over environments. Grain yield (tons/ha) for all 54 CIMMYT varieties over 2 years.

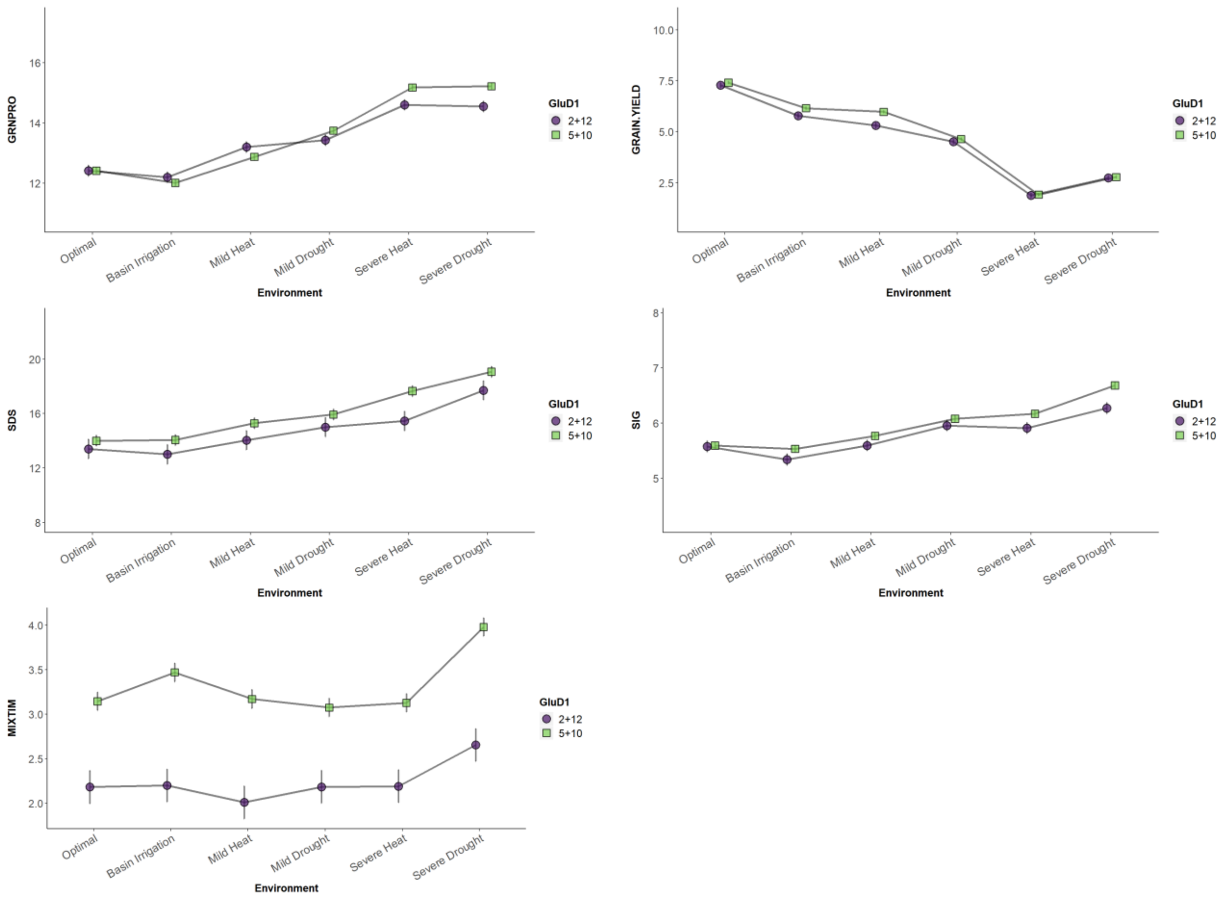


Figure 2.4 – Environment interaction of *Glu-D1* alleles.

Estimated values for alleles *5+10* (green squares) and *2+12* (purple circles) and standard error bars are shown for each trait. GRNPRO: grain protein content (% at 12.5% moisture); GRAIN.YIELD: grain yield (tons/ha) SDS: flour sodium dodecyl sulfate sedimentation volume (mL); SIG: swelling index of gluten; MIXTIM: optimal mixing time from Mixograph (min).

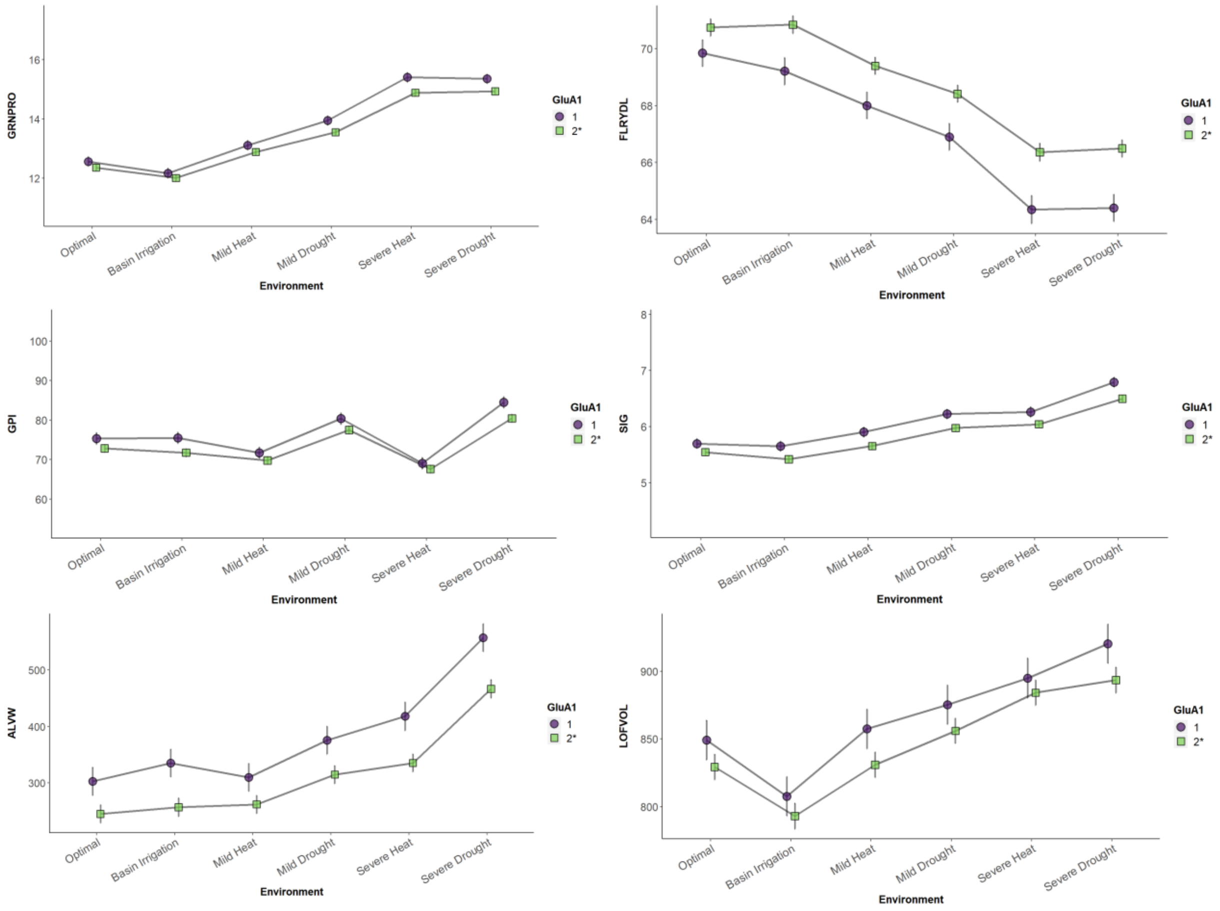


Figure 2.5 - Environmental interaction of *Glu-A1* alleles.

Estimated values for alleles *2** (green squares) and *1* (purple circles) and standard error bars are shown for each trait. GRNPRO: grain protein content (% at 12.5% moisture); FLRYLD: flour yield from milling (% recovered); GPI: gluten protein index; SIG: swelling index of gluten; ALVW: Alveograph W (work value from alveograph curve); LOFVOL, pup loaf volume (cm^2).

Table 2.2 - Variance explained by model components in yield traits.

Percent of total variance (% var) shows the attributable variance to each effect divided by the total variance. TESTWT: test weight (kg/hL); TKW: thousand kernel weight (grams); GRNHRD: Grain hardness (PSI, %); GRNYLD: grain yield (tons/ha); FLRYLD: flour yield from milling (% recovered).

	TESTWT		TKW		GRNHRD		GRNYLD		FLRYLD	
	vcov	% var	vcov	% var	vcov	% var	vcov	% var	vcov	% var
ENVIRONMENT	1.6	36.8	18.0	35.5	2.3	15.2	2.7	75.8	3.1	24.4
YEAR	0.0	0.0	0.0	0.0	1.6	10.5	0.0	1.4	1.3	10.6
ENVIRONMENT:YEAR	0.6	14.5	9.2	18.1	0.7	4.9	0.1	2.8	1.1	8.6
FLRPRO	0.0	0.0	0.0	0.0	0.0	0.1	0.2	6.6	0.0	0.0
RELEASE.YEAR	0.0	0.0	0.0	0.0	0.9	6.0	0.0	0.9	0.0	0.0
GID	0.6	13.8	8.5	16.7	2.3	15.6	0.0	0.8	1.7	13.6
GID:ENVIRONMENT:YEAR	0.6	12.5	2.5	4.9	1.6	10.3	0.1	2.4	1.5	12.4
GluA1	0.0	0.0	0.0	0.0	0.3	2.3	0.0	0.1	0.4	3.1
GluA1:ENVIRONMENT:YEAR	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.0	0.0
GluA3	0.0	0.0	6.5	12.8	0.0	0.0	0.0	0.0	0.0	0.0
GluA3:ENVIRONMENT:YEAR	0.0	0.2	0.2	0.4	0.0	0.2	0.0	0.0	0.0	0.0
GluB1	0.1	2.2	0.0	0.0	0.5	3.5	0.0	0.0	0.6	4.8
GluB1:ENVIRONMENT:YEAR	0.0	0.9	0.7	1.4	0.0	0.1	0.0	0.1	0.0	0.0
GluB3	0.0	0.0	1.6	3.2	2.6	17.2	0.0	0.0	0.0	0.0
GluB3:ENVIRONMENT:YEAR	0.1	1.4	0.1	0.2	0.2	1.6	0.0	0.6	0.0	0.0
GluD1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GluD1:ENVIRONMENT:YEAR	0.2	4.0	1.3	2.6	0.2	1.3	0.0	0.5	0.4	2.8
GluD3	0.0	0.0	0.1	0.2	0.0	0.3	0.1	2.8	0.0	0.0
GluD3:ENVIRONMENT:YEAR	0.0	0.3	0.5	0.9	0.3	1.7	0.0	0.0	0.0	0.0
X1B.1R	0.4	8.1	0.0	0.1	0.0	0.0	0.0	0.0	1.9	14.8
Phenotypic variance	4.5		47.7		12.8		4.2		10.1	
Model total variance	4.4		50.8		15.1		3.5		12.5	
Residual	0.2	5.1	1.6	3.2	1.4	9.0	0.2	5.0	0.6	4.9

Table 2.3 - Variance explained by model components in dough mixing traits.

Percent of total variance (% var) shows the attributable variance to each effect divided by the total variance. MIXTIM: optimal mixing time from Mixograph (min); PEAK: Mixograph mixing midline peak (torque); ALVW: Alveograph W (work value from alveograph curve); ALVPL: Alveograph P (tenacity) divided by L (extensibility) (mm/mm).

	MIXTIM		PEAK		ALVW		ALVPL	
	vcov	% var	vcov	% var	vcov	% var	vcov	% var
ENVIRONMENT	0.0	3.7	201.3	8.6	6626.9	22.8	0.0	3.9
YEAR	0.0	0.0	0.0	0.0	0.0	0.0	0.0	6.0
ENVIRONMENT:YEAR	0.1	3.8	48.9	2.1	531.7	1.8	0.0	4.2
FLRPRO	0.0	0.1	0.2	0.0	0.0	0.0	0.0	0.0
RELEASE.YEAR	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GID	0.2	14.6	303.3	12.9	2767.8	9.5	0.1	22.1
GID:ENVIRONMENT:YEAR	0.1	8.2	143.5	6.1	1636.7	5.6	0.1	23.7
GluA1	0.0	0.0	0.0	0.0	345.9	1.2	0.0	0.0
GluA1:ENVIRONMENT:YEAR	0.0	0.1	5.3	0.2	127.4	0.4	0.0	0.0
GluA3	0.1	6.6	165.0	7.0	2489.4	8.6	0.0	1.0
GluA3:ENVIRONMENT:YEAR	0.0	0.2	3.1	0.1	85.7	0.3	0.0	0.0
GluB1	0.0	1.5	67.2	2.9	781.9	2.7	0.0	4.4
GluB1:ENVIRONMENT:YEAR	0.0	0.5	9.1	0.4	16.0	0.1	0.0	0.8
GluB3	0.0	0.0	0.0	0.0	0.0	0.0	0.1	17.0
GluB3:ENVIRONMENT:YEAR	0.0	0.8	26.6	1.1	202.4	0.7	0.0	7.7
GluD1	0.3	25.7	518.6	22.1	3430.8	11.8	0.0	0.0
GluD1:ENVIRONMENT:YEAR	0.0	2.0	31.0	1.3	406.7	1.4	0.0	5.0
GluD3	0.1	7.6	156.3	6.7	1924.0	6.6	0.0	0.0
GluD3:ENVIRONMENT:YEAR	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.8
X1B.1R	0.3	22.3	632.9	26.9	7292.4	25.1	0.0	0.0
Phenotypic variance	0.9		1425.2		18715.8		0.3	
Model total variance	1.3		2349.6		29091.6		0.3	
Residual	0.0	2.1	37.4	1.6	425.7	1.5	0.0	3.5

Table 2.4- Variance explained by model components in gluten strength traits.

Percent of total variance (% var) shows the attributable variance to each effect divided by the total variance. SDS: flour sodium dodecyl sulfate sedimentation volume (mL); SIG: swelling index of gluten; GPI: gluten protein index, LOFVOL, pup loaf volume (cm^2).

	SDS		SIG		GPI		LOFVOL	
	vcov	% var	vcov	% var	vcov	% var	vcov	% var
ENVIRONMENT	2.7	18.2	0.1	28.5	22.8	20.1	1110.5	20.6
YEAR	0.0	0.0	0.0	7.3	0.0	0.0	19.5	0.4
ENVIRONMENT:YEAR	0.5	3.2	0.0	3.4	4.6	4.0	23.7	0.4
FLRPRO	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0
RELEASE.YEAR	0.1	0.8	0.0	1.7	8.1	7.1	125.4	2.3
GID	1.9	13.1	0.0	7.6	4.3	3.8	1357.5	25.1
GID:ENVIRONMENT:YEAR	1.1	7.4	0.0	7.5	5.6	4.9	819.2	15.2
GluA1	0.1	0.5	0.0	2.3	5.2	4.6	0.1	0.0
GluA1:ENVIRONMENT:YEAR	0.0	0.0	0.0	0.0	0.2	0.2	0.0	0.0
GluA3	2.1	14.5	0.0	6.6	7.0	6.2	44.5	0.8
GluA3:ENVIRONMENT:YEAR	0.1	0.5	0.0	0.5	0.5	0.5	6.1	0.1
GluB1	0.4	2.7	0.0	0.7	0.0	0.0	0.0	0.0
GluB1:ENVIRONMENT:YEAR	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GluB3	1.2	8.0	0.0	0.0	0.0	0.0	976.1	18.1
GluB3:ENVIRONMENT:YEAR	0.2	1.2	0.0	0.9	0.9	0.8	101.3	1.9
GluD1	0.0	0.0	0.0	1.7	0.0	0.0	367.6	6.8
GluD1:ENVIRONMENT:YEAR	0.2	1.6	0.0	1.6	1.0	0.8	112.5	2.1
GluD3	0.0	0.0	0.0	0.0	0.0	0.0	80.4	1.5
GluD3:ENVIRONMENT:YEAR	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
X1B.1R	3.8	26.3	0.1	24.9	50.3	44.3	15.3	0.3
Phenotypic variance	11.2		0.3		64.2		5279.9	
Model total variance	14.6		0.5		113.4		5403.6	
Residual	0.3	2.0	0.0	4.8	2.9	2.5	244.0	4.5

Table 2.5 - Variance explained by model components in solvent retention capacity traits.

Percent of total variance (% var) shows the attributable variance to each effect divided by the total variance. SRC LA: solvent retention capacity of lactic acid solution; SRC H_2O: solvent retention capacity of distilled water; SRC SOD: solvent retention capacity of sodium carbonate solution; SRC SUC: solvent retention capacity of sucrose solution.

	SRC LA		SRC H2O		SRC SOD		SRC SUC	
	vcov	% var	vcov	% var	vcov	% var	vcov	% var
ENVIRONMENT	127.8	38.3	1.7	12.1	2.5	11.1	8.8	26.6
YEAR	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ENVIRONMENT:YEAR	5.4	1.6	1.7	12.4	1.9	8.7	1.4	4.2
FLRPRO	0.0	0.0	0.0	0.0	0.1	0.3	0.0	0.1
RELEASE.YEAR	7.5	2.2	0.0	0.0	0.0	0.0	0.0	0.0
GID	37.9	11.4	5.3	37.5	8.4	37.7	8.0	24.2
GID:ENVIRONMENT:YEAR	15.5	4.6	1.8	13.1	2.4	10.6	3.2	9.8
GluA1	16.2	4.8	0.0	0.0	0.0	0.0	0.0	0.0
GluA1:ENVIRONMENT:YEAR	0.4	0.1	0.0	0.0	0.2	0.8	0.1	0.3
GluA3	22.5	6.7	0.2	1.2	0.0	0.0	0.4	1.3
GluA3:ENVIRONMENT:YEAR	1.2	0.4	0.1	0.9	0.0	0.0	0.8	2.4
GluB1	1.5	0.5	0.0	0.0	0.0	0.0	0.5	1.5
GluB1:ENVIRONMENT:YEAR	0.0	0.0	0.1	0.7	0.1	0.3	0.0	0.0
GluB3	0.0	0.0	0.3	2.5	1.1	4.7	0.0	0.0
GluB3:ENVIRONMENT:YEAR	3.1	0.9	0.1	0.9	0.7	3.0	0.4	1.3
GluD1	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0
GluD1:ENVIRONMENT:YEAR	2.3	0.7	0.1	0.9	0.3	1.3	0.4	1.1
GluD3	0.0	0.0	0.4	2.9	0.0	0.1	0.6	1.7
GluD3:ENVIRONMENT:YEAR	0.0	0.0	0.0	0.0	0.1	0.3	0.0	0.0
X1B.1R	85.7	25.7	0.7	5.1	2.9	13.1	5.9	17.8
Phenotypic variance	242.9		13.7		20.1		28.5	
Model total variance	333.7		14.1		22.2		33.0	
Residual	6.7	2.0	1.4	9.8	1.7	7.8	2.6	7.8

Chapter 3 - Genomic Selection Models for End Use Quality Traits

Introduction

The top priority of the CIMMYT wheat breeding program is to release stably higher yielding varieties for the many mega-environments. The yield increases should be as stable as possible across the environments and conditions that the variety may encounter.

Yield is limited by abiotic and biotic stress; therefore, environment relevant tolerance and resistance are also targeted. Heat and drought stress are major yield limiting stresses that are projected to become larger issues with climate change. The CIMMYT wheat breeding program tests lines under these stresses in Obregon and sites in around the world. Disease resistance, especially to leaf and stem rust, are a top priority and lines are screened throughout inbreeding generations. The Global Wheat Program also conducts pre-breeding specifically for these traits by crossing with wild relatives to increase genetic diversity under the SeeD program and by making strategic crosses under the Trigo and IWYP projects.

Quality and nutrition are secondary, but important, priorities in the CIMMYT wheat breeding program. The variety should have acceptable quality characteristics for the products in the target region. The demand for higher quality, higher protein wheat is increasing in places where mechanized production of bread products is becoming common. The nutritional quality of the grain is also under selection in the Harvest Plus program with the aim of releasing varieties with high zinc and/or high iron content. Although yield is the number one priority of the breeding program, releasing high end-use and nutritional quality varieties is important for consumer health and acceptance.

Although end-use quality is largely determined by high molecular weight glutenin alleles and these can be selected for through SDS-PAGE and molecular marker analysis, other traits also play a role and only selecting for superior HMW-GS alleles does not paint a complete picture of the end-use quality characteristics of a breeding line. Grain hardness, overall protein content, starch quality and lipid quality all also play a role. Each of these are also under some degree of genetic control. Therefore, implementing genomic selection, which takes into account genetic markers across the genome and estimates the quality value of a given breeding line, is most informative for end-use quality selections.

In 2011, PhD student Sarah McNeil (previously Sarah Battenfield) and Dr. Jesse Poland of Kansas State University collaborated with Dr. Carlos Guzman of the International Center for Maize and Wheat Improvement (CIMMYT) to implement genomic prediction models for end-use quality traits on the CIMMYT spring wheat breeding program. Starting with 1,245 entries with both genetic markers and end-use quality data in 2012 they were able to make genomic predictions with reasonable forward prediction accuracies (Battenfield et al., 2016). Adding approximately 900 entries every year to the training set showed promising gains in forward prediction accuracy which were approaching the heritability of each end-use quality trait. The culmination of this work was published in 2016 (Battenfield et al., 2016) and will be referenced often in this chapter.

In 2018, the focus of this work shifted from showing proof of concept to making routine end-use quality predictions for the approximately 10,000 yield trial entries in the CIMMYT wheat breeding program. This chapter will summarize the results of this shift.

Materials and Methods

Germplasm and Genotyping

Germplasm used as the training set were from the elite yield trails, international bread wheat spring nurseries and crossing blocks of the CIMMYT hard spring wheat breeding program between years 2010 and 2019. All of these populations have both genotyping and phenotyping information. The germplasm population for which genomic predictions were made was from the yield trial population for the given year. This population has only genotyping information at the time of genomic predictions, but a subset of the population is advanced and therefore phenotyped for end-use quality in late summer to early fall.

DNA was collected from a single seeding per yield trial entry in early March of each year. DNA was extracted at the CIMMYT Genotyping lab and sent to Kansas State University to prepare genotyping-by-sequencing (GBS) libraries according to the protocol described in (Poland, Brown, Sorrells, & Jannink, 2012), but with the modifications of 100 ng of DNA and 10 ul of restriction digest mix instead of 200 ng and 20 ul, respectively. GBS libraries were sequenced on the Illumina HiSeqX10 platform. Raw sequencing data was processed into single nucleotide polymorphisms with Tassel 5 GBS v2 (Bradbury et al., 2007) with Chinese Spring (IWGSC RefSeq v1.0) as the reference genome. The resulting vcf files of the given year were merged with vcf files from previous years and filtered for missing calls < 30%. The filtered file was then imputed using Beagle 4.1 (Browning & Browning, 2016). DNA was collected from a single seeding per yield trial entry in early March of each year. DNA was extracted at the CIMMYT Genotyping lab and sent to Kansas State University to prepare genotyping-by-sequencing (GBS) libraries according to the protocol described in (XX). GBS libraries were sequenced on the Illumina HiSeqX10 platform. Raw sequencing data was processed into single nucleotide polymorphisms with Tassel 5 GBS v2 with Chinese Spring (IWGSC RefSeq v1.0) as the reference genome. The resulting vcf files of the given year were merged with vcf files from previous years and filtered for missing calls < 30%. The filtered file was then imputed using Beagle 4.1.

Phenotyping

End-use quality phenotypes were collected at the CIMMYT wheat quality lab and are described in (Battenfield et al., 2016). The method for measuring grain hardness changed in 2015 from particle size index estimated with near infrared spectroscopy (AACC Method 39-70.02) to force required to crush kernels measured by the Single Kernel Classification System (AACC Method 55-31.01). Grain hardness measurements prior to the change are excluded from the training set. Polyphenol oxidase (PPO) activity assay was also added in 2017 (AACC Method 22-85.01). PPO influences discoloration of grain and products over time. The best linear unbiased estimates (BLUEs) for quality traits were found with lm1 package of R with only the breeding line as a fixed effect.

Genomic Selection Models

Genomic predictions were conducted in R (R Core Team, 2017) with the rrBLUP package v. 4 (Endelman, 2011). rrBLUP v4 requires the vector of training set phenotypes, the vector of prediction set breeding line identification numbers and the matrix of genotypes that contains both training set and prediction set breeding lines. From the matrices of genotypes, the kinship matrix was calculated using the gaussian kernel, which estimates Euclidean distances between genotypes. Gaussian kernel was chosen because Battenfield (2016) and Endelmen

(2011) both found that it outperformed A matrix in cross validation and performed equally well in forward prediction. Genomic estimated breeding values for each quality trait were predicted separately for simplicity and speed.

Only forward prediction accuracies were measured to gauge the performance of the models and cross validations were not considered. Forward prediction accuracy was calculated as the Pearson's correlation coefficient between predicted value and unadjusted observed value. It was calculated with default settings in the cor() function of base R.

To estimate heritability, we fit a random effects model using all observed phenotypic data from 2010 to 2019. Each trait was the response variable and the random effects predictor was breeding line. The mixed linear model was fit using the using lmer() and variance components estimated with VarCorr(), both of the lme4 v1.1-26 package of R-4.0.4 (Bates et al., 2015). Heritability was calculated as the breeding line variance divided by phenotypic variance (breeding line variance plus error variance).

Results and Discussion

Each year, approximately 400 elite yield trial, 1200 international bread wheat spring nursery, and 200 crossing block entries were genotyped using GBS, phenotyped for end use quality traits at the Wheat Chemistry and Quality Laboratory at the International Maize and Wheat Improvement Center (CIMMYT) in Texcoco, Mexico and then added to the training set for the GS models (Table 3.1). The size of the training set grew from 5,520 entries in 2015 to over 19,000 in 2020. By 2020, just over 4,300 of the lines were phenotyped in more than 1 year, allowing us to make across year heritability estimates for traits. Across year heritability ranged between 0.49 and 0.84 (Table 3.2) and were higher than those reported by Battenfield in 2016.

Distributions of the end-use quality phenotypes varied over the years (Figure 3.1). Overall, most traits had an approximately normal distribution. Those traits that had both approximately normal distributions and which vary the least year to year tended to have the highest prediction accuracies, i.e., Mixograph Mixing Peak (MP) which is measure of gluten strength. Other traits which varied more over the years i.e., loaf volume (LOFVOL) or had distributions with long tails i.e., Alveograph P/L (ALVPL) had the lowest prediction accuracies (Figure 3.2). Trait distributions from the last ten years showed that overall, quality trait values maintained similar values. This reflects that selection at CIMMYT is to maintain the already acceptable quality values and discard candidate wheat lines with unacceptably poor quality. Interestingly, grain protein content did appear to increase over time.

For each year between 2018 and 2020, the forward predictions were made on approximately 10,000 first year yield trails. The genomic prediction models ran on the high performance computing cluster with 90 Gb of memory and two tasks per node. Each year there were approximately 90,000 SNPs passing filtering criteria. Predictions completed in 7-10 days and were output as a comma separated file of two columns, breeding line identification number and the corresponding GEBV. The GEBV were combined into one excel file before sending the CIMMYT teams. Predictions were provided before breeders made selections in all years between 2018 and 2020.

From 2018 to 2020, the forward prediction accuracies were close to the theoretical threshold of the heritability estimates of each trait. On the whole, the forward prediction accuracies hovered just at or below the heritability (Figure 3.1). Year to year environmental variation influenced the predictability. Some years were better for predictability yield related traits such as thousand kernel weight (TKW) and correspondingly worse for the predictability of gluten strength traits such as mixing time (MIXTIM).

Conclusion

In this work we showed that genomic prediction models for end-use quality in the large CIMMYT spring bread wheat breeding program has become routine. For most traits, the forward prediction accuracies were near the theoretical maximum of the heritability for that trait. For traits were improvements in forward prediction accuracy could be made, covariates such as grain protein content and weather data could be added to the genomic prediction models. Pressingly, a concise estimate of overall end-use quality class as described by Guzman et al. (2016) could also facilitate interpretation of the 13 quality traits during selection. With this work, we showed that genomic predictions provide reliable end-use quality information to breeders before selections are made on otherwise unobserved material.

Table 3.1 - Summary of training set entries by year.

Training set entries are from the crossing block, elite yield trial and international bread wheat spring nursery populations at CIMMYT. Each entry has quality phenotypic data and genotyping by sequencing markers.

Year	Number of training set entries
2010	1258
2011	1000
2012	1580
2013	1844
2014	2010
2015	2118
2016	2118
2017	2096
2018	2136
2019	2006
2020	1500
TOTAL	19,666

Table 3.2 - Heritability of quality traits.

A subset of approximately 4,000 training set entries were phenotyped in more than one year. Each trait heritability was was calculated as the genetic variance over phenotypic variance (genetic variance and residual variance).

Trait	Heritability (h^2)
Thousand Kernel Weight (TKW)	0.76
Test weight (TESTWT)	0.65
Polyphenol Oxidase activity (PPO)	0.77
Grain hardness (GRNHRD)	0.75
Grain protein (GRNPRO)	0.69
Flour protein (FLRPRO)	0.76
Flour SDS sedimentation (FLRSDS)	0.71
Flour yield (FLRYLD)	0.49
Mixograph mix peak (MP)	0.84
Alveograph P/L (ALVPL)	0.61
Alveograph W (ALVW)	0.80
Loaf volume (LOFVOL)	0.70

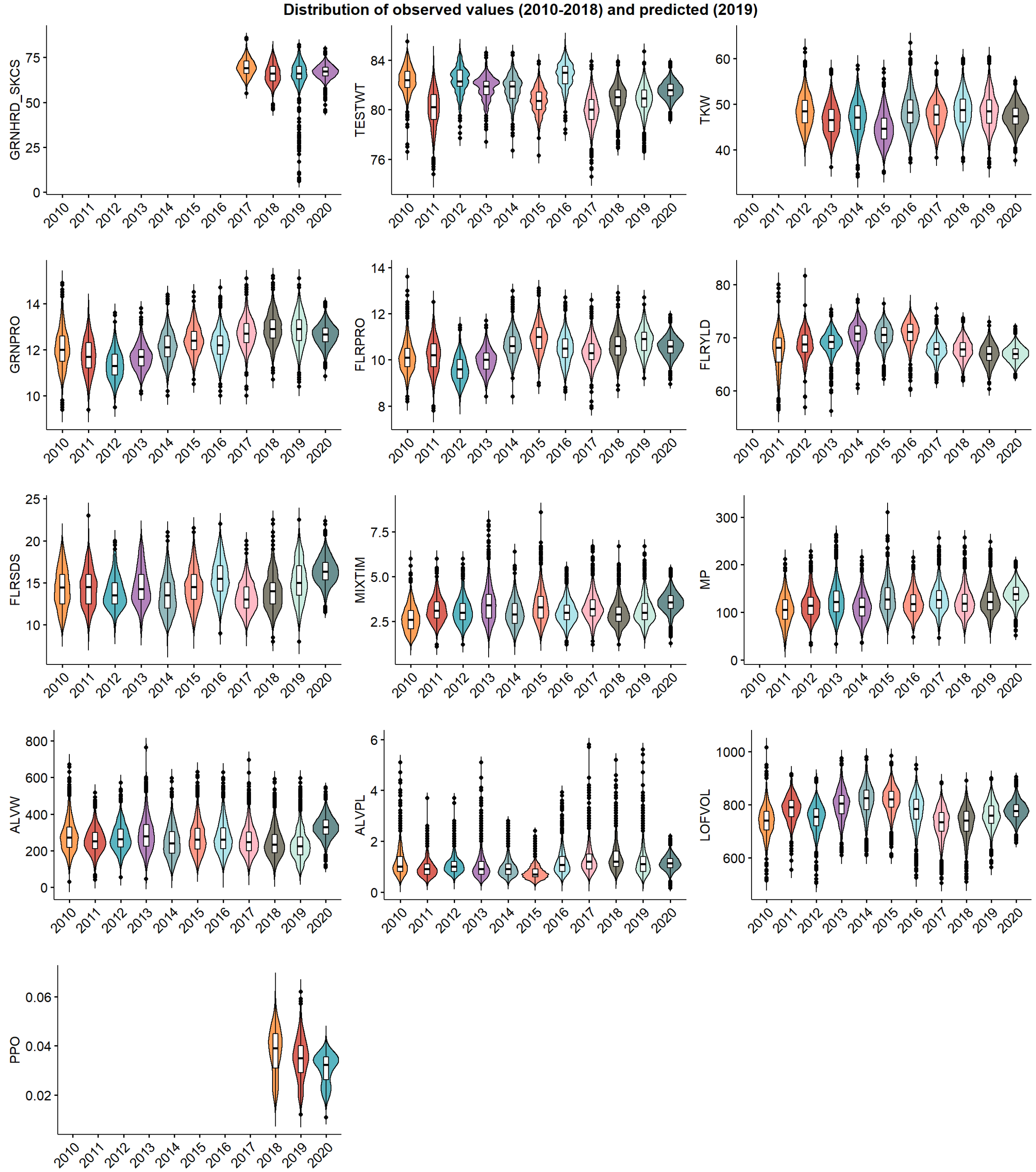


Figure 3.1 - Distribution of end-use quality phenotypes.

Years 2010-2019 were the observed phenotypic values for the training set entries. Whereas Year 2020 showed the distribution of the 2020 genomic predictions for the first year yield trial entries. GRNHRD_SKCS: Grain hardness (single kernel characterization system hardness index); TESTWT: test weight (kg/hL); TKW: thousand kernel weight (grams); GRNPRO: grain protein content (% at 12.5% moisture); FLRPRO: flour protein content (% at 14% moisture); FLRYLD: flour yield from milling (% recovered); FLRSDS: flour sodium dodecyl sulfate sedimentation volume (mL); MIXTIM: optimal mixing time from Mixograph (min); MP: Mixograph mixing midline peak (torque); ALVW: Alveograph W (work value from alveograph curve); ALVPL: Alveograph P (tenacity) divided by L (extensibility) (mm/mm); LOFVOL, pup loaf volume (cm^2).

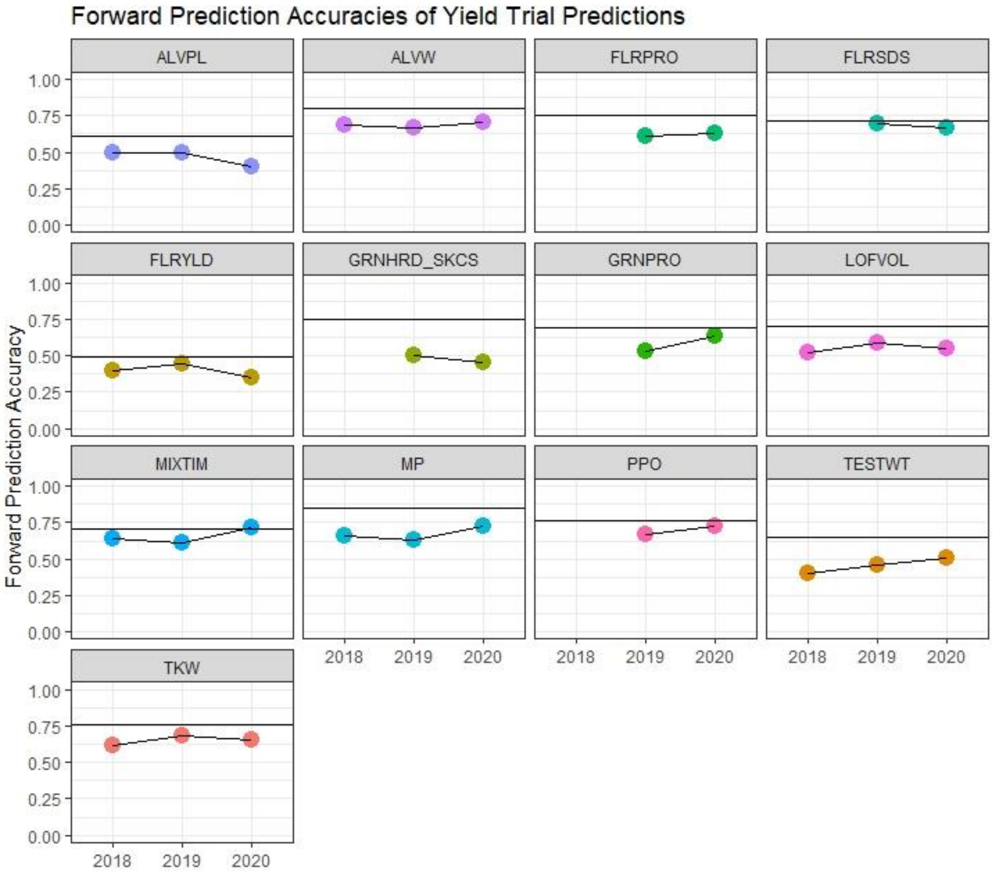


Figure 3.2 - Forward prediction accuracies of genomic selection models.

Prediction accuracies shown over time as colored dots. Although all 10,000 yield trial lines were predicted, only a subset of about 2,000 were advanced in the breeding cycle and then phenotyped. The prediction accuracies were therefore for the subset of those breeder selected lines. Black horizontal lines denote the heritability that was calculated with 2010-2019 data. Forward prediction accuracies were calculated as the Pearson's correlation coefficient between predicted values and observed values. GRNHRD_SKCS: Grain hardness (single kernel characterization system hardness index); TESTWT: test weight (kg/hL); TKW: thousand kernel weight (grams); GRNPRO: grain protein content (% at 12.5% moisture); FLRPRO: flour protein content (% at 14% moisture); FLRYLD: flour yield from milling (% recovered); FLRSDS: flour sodium dodecyl sulfate sedimentation volume (mL); MIXTIM: optimal mixing time from Mixograph (min); MP: Mixograph mixing midline peak (torque); ALVW: Alveograph W (work value from alveograph curve); ALVPL: Alveograph P (tenacity) divided by L (extensibility) (mm/mm); LOFVOL, pup loaf volume (cm^2).

Chapter 4 - High Molecular Weight Glutenin Gene Diversity in *Aegilops tauschii* demonstrates Unique Origin of Superior Wheat Quality

Abstract

Central to the diversity of wheat products was the origin of hexaploid bread wheat, which added the D-genome of *Aegilops tauschii* to tetraploid wheat giving rise to superior dough properties in leavened breads. The polyploidization, however, imposed a genetic bottleneck, with only limited diversity introduced in the wheat D-subgenome. To understand genetic variants for quality, we sequenced 273 accessions spanning the known diversity of *Ae. tauschii*. We discovered 45 haplotypes in *Glu-D1*, a major determinant of quality, relative to the two predominant haplotypes in wheat. The wheat allele *2+12* was found in *Ae. tauschii* Lineage 2, the donor of the wheat D-subgenome. Conversely, the superior quality wheat allele *5+10* allele originated in Lineage 3, a recently characterized lineage of *Ae. tauschii*, showing a unique origin of this important allele. These two wheat alleles were also quite similar relative to the total observed molecular diversity in *Ae. tauschii* at *Glu-D1*. *Ae. tauschii* is thus a reservoir for unique *Glu-D1* alleles and provides the genomic resource to begin utilizing new alleles for end-use quality improvement in wheat breeding programs.

Introduction

Originating in the Fertile Crescent some 10,000 years ago, hexaploid wheat (*Triticum aestivum*) is now grown and consumed around the world (Salamini, Özkan, Brandolini, Schäfer-

Pregl, & Martin, 2002). The global consumption of wheat as a staple crop is owed principally to the unique viscoelastic properties of wheat dough that lend it the capacity to make diverse baked products such as leavened bread, tortillas, chapati, pastries, and noodles. The uniqueness of wheat dough can also be described as the strength to resist deformation and elasticity to recover the original shape as well as the viscosity to permanently deform under persistent stress. Elasticity is important for the product to hold shape, while viscosity allows the dough to be worked and formed. The balance of the competing properties determines what baked goods a dough is suitable for, such as a dough with greater strength for leavened pan bread compared to the more extensible dough that is desired for a chapati or tortilla.

Bread wheat is an allohexaploid with the A-, B- and D-subgenomes contributed by different, but related, species. The closest relative to the wheat A-subgenome is diploid *Triticum urartu*, with other diploid A-genome species including the wild and domesticated Einkorn wheat (*Triticum monococcum*). While the exact ancestor of the B-genome is unknown and presumed extinct, it is believed that *Ae. speltoides* (S-genome) is the closest living relative. These two species were brought together to form a tetraploid wheat species with AABB genome composition, which is known as durum or pasta wheat (*Triticum durum*). The D genome from *Aegilops tauschii* was the most recent addition forming the hexaploid genome. This addition of the D-subgenome, to form hexaploid wheat, led to a much broader adaptation and superior bread making quality compared to the tetraploid and diploid ancestors (Dubcovsky & Dvorak, 2007). However, the original hexaploid species originated from very few *Ae. tauschii* accessions and limited subsequent cross-hybridization likely caused by ploidy barriers with the diploid *Ae. tauschii* (J. Wang et al., 2013). This genetic bottleneck resulted in limited genetic diversity in the wheat D-subgenome (Zhou et al., 2020).

The utility of wheat and the variation of wheat products and consumption is driven by the strength and elasticity of the dough which is determined by the structure of the gluten matrix. This matrix is formed from a combination of high-molecular weight (HMW) and low-molecular weight (LMW) glutenin proteins and gliadins (P. Shewry, 2019). The backbone of the gluten matrix is developed under mixing by the covalent disulphide bonds between cysteine residues in HMW glutenins (Lutz, Wieser, & Koehler, 2012). These glutenins, therefore, are some of the most important genes giving wheat its unique dough properties. They are encoded by a relatively simple locus on the long arm of the group one chromosomes of the Triticeae. Hexaploid wheat, comprised of the A-, B- and D-genomes, thus contains three HMW glutenin loci; *Glu-A1*, *Glu-B1* and *Glu-D1*. Each locus harbors two HMW glutenin genes known as the x and y subunit, that are tightly linked but separated by tens to hundreds of kilobase pairs (kb) (Anderson, Rausch, Moullet, & Lagudah, 2003; Gu et al., 2006; Kong, Gu, You, Dubcovsky, & Anderson, 2004). Each subunit consists of short, unique N and C terminal domains which flank a central highly repetitive region that accounts for 74-84% of the total protein length (P. R. Shewry, Halford, Belton, & Tatham, 2002).

Allelic differences in all three gluten proteins contribute to the conformation of the gluten matrix and variable end-use quality. The D-subgenome locus, however, is the major driver of bread quality and absence of the D-genome leads to substantially different dough qualities found in tetraploid pasta wheats (Wang et al., 2017). The two common alleles at *Glu-D1* found in bread wheat are *Glu-D1a* (SDS-PAGE allele designation *2+12)* and *Glu-D1d* (*5+10*), with the latter associated with superior breadmaking quality . Following the domestication and breeding of wheat, there is limited variation at the *Glu-D1* locus in the D-genome with only these two alleles found in the vast majority of bread wheat throughout the world (Payne, Holt, & Law, 1981; Payne & Lawrence, 1983). Of the HMW glutenin alleles on the three sub-genomes, the greatest impact on end-use quality is imparted by the *Glu-D1* locus. Thus, the addition of the wheat D-subgenome and specifically variation at *Glu-D1* has substantial impact on wheat quality globally. This is arguably the single greatest defining feature of bread wheat.

Reflecting the importance of *Glu-D1* in determining the end-use quality of wheat, focus has been given to understanding the variation present in *Ae. tauschii* for this locus. Much of the work has utilized sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) protein analysis of *Ae. tauschii* collections (Gianibelli, Gupta, Lafiandra, Margiotta, & MacRitchie, 2001; Mackie, Lagudah, Sharp, & Lafiandra, 1996; William, Peña, & Mujeeb-Kazi, 1993; Xu, Khan, Klindworth, & Nygard, 2010). From this work, over 37 SDS-PAGE *Glu-D1* alleles have been named in *Ae. tauschii.* However, due to the limited resolution of SDS-PAGE, many of the alleles have indistinguishable SDS-PAGE mobilities from the common *Glu-D1* hexaploid alleles, *2+12* and *5+10*, or are difficult to reliably distinguish. By changing the polyacrylamide percentage or acidity in the SDS-PAGE, it was shown that the *Ae. tauschii 2+12* and *5+10* alleles were slightly different than the common wheat alleles (Lagudah & Halloran, 1988). These *Ae. tauschii* alleles are therefore given the designations *2t+12t* and *5t+10t*. In addition to the *2t+12t* and *5t+10t* alleles, a large number of SDS-PAGE alleles have been described, supporting the hypothesis that *Ae. tauschii* could be a vast resource for untapped diversity at *Glu-D1* and that this diversity could be utilized for wheat quality improvement.

Here we characterized the *Glu-D1* allelic diversity in a panel of 273 sequenced *Ae. tauschii* accessions. The panel spans the known genetic diversity of *Ae. tauschii* and is a powerful resource for association mapping and gene identification (Gaurav et al., 2021). From the sequenced *Ae. tauschii* panel, we discovered hundreds of genetic variants which defined dozens of unique haplotypes. This gives the needed molecular information to track these alleles in breeding germplasm, which will in turn enable targeted assessment of the novel *Ae. tauschii* HMW glutenin alleles in hexaploid backgrounds leading to utilization of favorable alleles for wheat quality improvement.

Materials and Methods

Plant Material

This study included 273 *Aegilops tauschii* accessions, of which 241 were from the Wheat Genetics Resource Center (WGRC) collection at Kansas State University in Manhattan, KS, USA. Another 28 were from the National Institute for Agricultural Botany (NIAB) in Cambridge, United Kingdom. An additional 2 were from the Commonwealth Scientific and Industrial Research Organisation (CSIRO) in Canberra, Australia. The final accession, AL8/78, was obtained from the John Innes Center (JIC) in Norwich, Norfolk, England. Data regarding original collection sites for the WGRC accessions is detailed in Supplementary Data 1(N. Singh et al., 2019). *Aegilops tauschii* is divided into two subspecies, spp. *tauschii* (Lineage 1) and the wheat D-genome donor spp. *strangulata* (Lineage 2). In this data set, 117 accessions were Lineage 1 and 143 Lineage 2. An additional eight accessions (five non-redundant) belonged to the newly described Lineage 3(Gaurav et al., 2021).

SDS-PAGE Analysis

The sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) analysis of 72 of the *Ae. tauschii* accessions, was conducted at the Wheat Chemistry and Quality Laboratory at the International Maize and Wheat Improvement Center (CIMMYT) in Texcoco, Mexico according to Singh *et al.* (1991)(N. K. Singh et al., 1991) with the following modifications. Specifically, 20 mg of whole meal flour were mixed at 1,400 rpm with 0.75 ml of 50% propanol (v/v) for 30 min at 65°C in a Thermomixer Comfort (Eppendorf). The tubes were then centrifuged for 2 min at 10,000 rpm, and the supernatant containing the gliadins was discarded. The pellet was then mixed with 0.1 ml of a 1.5% (w/v) DTT solution in a Thermomixer for 30 min at 65°C, 1,400 rpm, and centrifuged for 2 min at 10,000 rpm. A 0.1 ml volume of a 1.4% (v/v) vinylpyridine solution was then added to the tube which was subsequently placed again in a Thermomixer for 15 min at 65°C, 1,400 rpm, and centrifuged for 5 min at 13,000 rpm. The supernatant was mixed with the same volume of sample buffer (2% SDS (w/v), 40% glycerol (w/v), and 0.02% (w/v) bromophenol blue, pH 6.8) and incubated in the Thermomixer for 5 min at 90°C and 1,400 rpm. Tubes were centrifuged for 5 min at 10,000 rpm, and 8 ml of the supernatant were used for the glutenin gel. Glutenins were separated in polyacrylamide gels (15% or 13% T) prepared using 1 M Tris buffer, pH of 8.5. Gels were run at 12.5 mA for ~19 h. Alleles were identified using the nomenclatures proposed by Payne and Lawrence (1983)(Payne & Lawrence, 1983) for bread wheat high molecular weight glutenins and Lagudah and Halloran (1988)(Lagudah & Halloran, 1988) for previously described *Ae. tauschii* high molecular weight glutenins.

DNA Sequencing

Whole genome Illumina paired-end sequencing to 10x coverage for most accessions, and 30x coverage for select accessions, was obtained from TruSeq PCR-free libraries with 350 bp insert with Illumina paired end sequencing of 150 bp according to manufacturer recommendations. Sequence datasets are detailed in Gaurav *et al.* (2021).

Variant Calling and Duplicated Accession Analysis

Paired-end reads of the *Ae. tauschii* samples were aligned to the *Ae. tauschii* AL8/78 genome assembly (Aet v4.0; NCBI BioProject PRJNA341983, accession AL8/78) and hexaploid wheat samples aligned to an *in silico* reference assembly including the hexaploid wheat A and B genomes from 'Jagger' (Walkowiak et al., 2020) (Aet v4.0; NCBI BioProject PRJNA341983, accession AL8/78) combined with the Aetv4.0 D genome using HISAT2 version 2.1.0 with default parameters (Kim, Paggi, Park, Bennett, & Salzberg, 2019). Alignments were sorted and indexed using samtools v1.9 (H. Li et al., 2009). Variants for coding regions of the x and y subunits of *Glu-D1* were called using bcftools version 1.9 (Heng Li, 2011) 'mpileup' and 'call' commands with a minimum alignment quality of 20 (--q 20) (L. Gao, 2020). Duplicated accessions were identified as sharing greater than 99.8% variant calls.

Molecular Haplotype Analysis

Ae. tauschii and hexaploid wheat variant call format (vcf) files were merged in R and variant calls were recoded to reference (-1) and alternate (1) alleles in R and heterozygous calls were set to missing. Variants were filtered on the following criteria: a variant must be present in either hexaploid wheat or *Ae. tauschii*, must have a quality score greater than 30 and be present in greater than 50% of samples. Given that we expected novel alleles present in single accessions, no minimum minor allele frequency was set. Samples sharing the same variants were considered to share the same molecular haplotype.

Genetic distances were calculated as the Euclidean distance on the A matrix of the variants in R. The A matrix was calculated with 'A.mat()' from the rrBLUP package (Endelman, 2011) and Euclidean distances with 'dist()'. Hierarchical clustering of the genetic distances were found using hclust() and converted to a dendrogram object before plotting with the dendextend package (Galili, 2015).

Molecular haplotypes were designated by the subclade number of the x and y subunits together and then by the letter corresponding to the individual gene level haplotype within. For example, molecular haplotype *x1a + y1b* represents the *a*th x haplotype and *b*th y haplotype within the subclade 1. It should be noted that letter designations across subclades have no correspondence. The *ath* x haplotype of subclade 1 is different than that of subclade 2.

Results and Discussion

Molecular Diversity of Glu-D1 in Ae. tauschii

Through the Open Wild Wheat Consortium, we obtained Illumina 150 bp paired-end short reads from 234 unique *Ae. tauschii* accessions each sequenced to greater than 7-fold coverage(Gaurav et al., 2021). These were aligned to the *Ae. tauschii* AL8/78 reference genome and sequence variants at the annotated *Glu-D1* locus were extracted. We also included three wheat cultivars in this analysis to compare *Ae. tauschii* variants to the common *5+10* (variety 'CDC Stanley') and *2+12* (varieties 'Chinese Spring' and 'LongReach Lancer') alleles. From this panel, we identified a total of 310 variants at *Glu-D1,* which were used to generate haplotypes and evaluate molecular diversity at this locus.

From the *Ae. tauschii* germplasm collection we identified 32 and 33 haplotypes within the coding sequence for the x and y subunits of the *Glu-D1* locus, respectively (Figure 4.1). When considering the complete *Glu-D1* locus with combination of the x and y subunit, a total of 45 haplotypes were identified (Tables 4.2 and 4.3). The various x and y subunit haplotypes were almost exclusively associated with each other, demonstrating the close physical association and limited recombination between the two genes. We included the 2500 bp up- and downstream sequences in our analysis to see if this resulted in further differentiation of alleles as short-read sequences often do not align uniquely to the central, highly repetitive region of the HMW glutenin genes. Including the flanking regions did not result in additional haplotypes. Thus, it appears that the identified variants are sufficient for faithfully differentiating alleles at *Glu-D1*.

We then calculated genetic distances and determined a gene-level phylogeny at *Glu-D1* for all of the *Ae. tauschii* accessions (Figure 4.1). Haplotypes clustered into three major clades, two of which were associated predominantly with Lineage 2 and one with Lineage 1. A unique group of *Glu-D1* alleles from the newly characterized Lineage 3 accessions were found within a narrow clade with Lineage 2. Among the three major clades, we designated 16 subclades that were clearly distinguished by variants and coincided with a Euclidean distance of 4. Of the 16 subclades, eight were associated exclusively with Lineage 2, five with Lineage 1, and one with Lineage 3. The Lineage 3 accessions all fell within the Lineage 2 major clades, but occupied a unique subclade therein. Thus, the gene-level phylogeny at this locus agrees very closely with the overall previously described population structure of the *Ae. tauschii* lineages(Gaurav et al., 2021; N. Singh et al., 2019). We also observed one clade (9) that had representative accessions from both Lineage 1 and 2. This could represent an ancestral haplotype found in both lineages which underwent incomplete lineage sorting, or a case of recent interlineage haplotype exchange. Cases of haplotypes shared across Lineages 1 and 2 were also observed for pest (*Cmc4*) and disease resistance (*Sr46*) genes (Gaurav et al., 2021).

Lineage 2, the recognized ancestral diploid donor of the D-subgenome of hexaploid wheat (J. Wang et al., 2013), had greater *Glu-D1* molecular haplotype diversity than Lineage 1. Not only were there more subclades associated with Lineage 2, there were also more haplotypes

(Figure 4.1). As expected, the haplotypes of wheat clustered within Lineage 2 subclades (Figure 4.1). Within Lineage 2, we observed *Ae. tauschii* accessions with a matching sequence haplotype to the wheat *2+12* allele consistent with the D-subgenome origin from Lineage 2. Interestingly, the wheat *5+10* allele clustered within the unique Lineage 3 sub-clade. Supporting the inheritance of the *5+10* allele from Lineage 3, (Gaurav et al., 2021) observed genome-wide contribution of Lineage 3 to wheat ancestry. These findings reveal that the Lineage 3 contribution to the wheat D-subgenome included the very valuable *Glu-D1 5+10* allele, arguably one of the most important genes defining the quality of bread wheat.

Given the large difference in quality between wheat cultivars carrying *2+12* and *5+10* alleles, we hypothesized that these two haplotypes would not be similar at a molecular level. However, we found that *2+12* and *5+10* clustered relatively closely within major-clade III, with much greater overall diversity detected across *Ae. tauschii* particularly when including the Lineage 1 accessions which had very different haplotypes. When comparing the *2+12* and *5+10* haplotypes to those found in Lineage 1, it becomes apparent that *Ae. tauschii* carries alleles that are very unlike anything seen in bread wheat and may offer unique functional characteristics when introgressed into hexaploid backgrounds.

Geographic Diversity

Given the known geographic structure and distribution of *Ae. tauschii* which is associated with various levels of population structure (N. Singh et al., 2019), we evaluated the *Glu-D1* diversity relative to the geographic origin of the *Ae. tauschii* accessions. Molecular haplotypes were strongly associated with geographic origin, consistent with the overall genome-wide picture (N. Singh et al., 2019), and genetic distances between alleles increased with the geographic distance between collection sites of the *Ae. tauschii* accessions (Figure 4.2). The greatest concentration of haplotype diversity was located along the shores of the Caspian Sea in Iran (Figure 3.2). Consistent with a hypothesis of admixture between Lineage 1 and Lineage 2 leading to shared gene-level haplotypes across the lineages, the accessions from Lineage 1 and 2 with the same *Glu-D1* haplotype (within subclade 9) were collected very near one another.

Molecular Haplotypes Identify Novel Glu-D1 Alleles

We employed SDS-PAGE analysis, the traditional standard for differentiating HMW glutenin loci, to determine if the haplotype molecular sequence diversity would also reflect differences in protein mobility. We evaluated at total of 72 unique accessions with SDS-PAGE and differentiated 9 alleles for the x subunit and 8 alleles at the y subunit from this protein mobility assay. Analysis of the Lineage 1 and Lineage 2 variants revealed that molecular haplotypes were consistent with the proteins differentiated by SDS-PAGE (Tables 4.1 and 4.2). For the majority of the alleles that were differentiated by SDS-PAGE, we were able to unambiguously correlate the observed SDS-PAGE alleles with the molecular variants. Although specific molecular haplotypes were associated with specific SDS-PAGE mobilities, there was little concordance between gene level variation and SDS-PAGE mobility as similar alleles at the molecular level were observed with very different SDS-PAGE mobilities. Alternatively, very different molecular haplotypes were observed with the same SDS-PAGE. This supports our hypothesis that the observed sequence variants are effectively in complete linkage disequilibrium and tagging the size variants from the central repeat region. Similarly, the SDS-PAGE diversity was lower having less differentiating power than the molecular haplotypes. As noted, the same SDS-PAGE mobilities were observed in both Lineage 1 and Lineage 2 haplotypes, but the molecular haplotypes were clearly differentiated (Figure 4.1). The protein mobility differences are considered to be primarily due to variation in the central repetitive region and therefore are not directly detectable with short-read sequencing, though the variable central repeats are completely phased with diagnostic haplotype variants within the terminal coding regions. Thus, we conclude that a sequence-based resource such as this *Ae. tauschii* panel provides a superior tool for identification and tracking of unique *Glu-D1* alleles in molecular breeding.

We also examined the connection between the glutenin protein mobility in *Ae. tauschii* compared to hexaploid wheat. *Ae. tauschii* haplotype *Dx1a+Dy1a* matched with the wheat *2+12* haplotype and exhibited the same SDS-PAGE mobility. Although we found an *Ae. tauschii* haplotype identical to the wheat *2+12* allele haplotype, the exact wheat *5+10* haplotype was not detected in this panel, although a very closely related Lineage 3 haplotype was found. Additionally, no *5+10* SDS-PAGE mobilities were observed. This was a surprising observation given that previous studies reported *Ae. tauschii* alleles with a *5+10* SDS-PAGE mobility (William et al., 1993). However, Williams *et al.* (1993)(William et al., 1993) did not reveal the identities of the *Ae. tauschii* accessions with 5+10 SDS-PAGE mobility. Interestingly, the haplotype *Dx7a+Dy7a* in the newly characterized Lineage 3(Gaurav et al., 2021) was most similar to *5+10* on the molecular level, however it carried eight variant differences. This current panel, however, only has five unique accessions representing Lineage 3. It is possible therefore that exploration of additional Lineage 3 accessions would reveal a haplotype exactly matching the wheat *5+10* with the same mobility.

Cryptic Haplotypes

One of the most valuable findings of this study was the high prevalence of cryptic molecular haplotypes hidden within SDS-PAGE mobilities. Within every SDS-PAGE mobility pattern there were multiple molecular haplotypes, often from very different subclades and occasionally from entirely different clades (Figure 4.1). The cryptic SDS-PAGE haplotypes, accordingly, were geographically disperse (Figure 4.3). For example, within SDS-PAGE *2+12* were four haplotypes; one which was the same as wheat *2+12* (*Dx1a + Dy1a*), another which was within the same subclade (*Dx1c+Dy1d*), and two from entirely different major-clades (*Dx9a+Dy9b* and *Dx13b+Dy13a*). Also, within subclade 9 were the SDS-PAGE mobilities *Dx2+Dy10* and *Dx2+Dy11*, and within subclade 13 were the SDS-PAGE mobilities *1t+12*, *2.1*+12.1**, and *4+10* further supporting that these haplotypes are not all similar to the wheat *2+12* haplotype at the molecular level. However, the proteins still migrate similarly on an SDS-PAGE. These results suggest that SDS-PAGE alone is insufficient when characterizing HMW glutenin diversity in wild relatives and will not be a suitable tool for tracking novel alleles in the hexaploid wheat germplasm.

While most molecular haplotypes delineated along the three *Ae. tauschii* lineages (Figure 1), a notable exception was within the predominantly Lineage 1 major-clade, subclade 9, where the same three haplotypes (*Dx9a+Dy9a*, *Dx9a+Dy9b*, and *Dx9a+Dy9c*) were observed in both Lineage 1 and Lineage 2 accessions. Interestingly, while there were three haplotypes at the y subunit, there was only a single x haplotype associated with all three of these. The x subunit mobility was the same for all three haplotypes, indicating that the x allele is in fact the same. However, the y subunit was differentiated with the mobility *Dy9b* was faster than that of *Dy9a* and *Dy10c*.

Recombinant Haplotypes Identified

The close proximity of the glutenin genes results in such tight linkage that recombination is extremely rare. To date, a recombination between the x and y subunit of any HMW-GS locus has yet to be verified. Among the 242 *Ae. tauschii* accessions studied here, we found a clear example of a historical recombination at *Glu-D1* in the accession TA1668 (Lineage 2). SDS-PAGE mobility of TA1668 matches that of TA10081 (*Dx2+Dy10.2*), and though the y haplotype of TA1688 is the same as the y haplotype of TA10081, the x subunit is very different and matching the Lineage 1 clade (Figure 4.4). Within this clade, the subclade 9 contains both Lineage 1 and Lineage 2 accessions, indicating that there was incomplete lineage sorting or admixture between the two lineages that lead to the introgression of a lineage *Glu-D1* haplotype into the Lineage 2 population. In the presence of both haplotypes, it appears there was a rare recombination between the Lineage 1 and Lineage 2 *Glu-D1* haplotypes, leading to the recombinant haplotype *Dx9a+Dy5a* found in TA1688.

The Lineage 3 accession TA2576 also appears to carry a recombinant haplotype (*Dx7b + Dy15b*) (Figure 4.4). However, our dataset did not contain the exact haplotypes involved in the recombination that led to *Dx7b + Dy15b.* The closest x subunit haplotype is *Dx7a*, the only other Lineage 3 haplotype, from major-clade III and the closest y subunit is the Lineage 2 haplotype *Dy15a* from major-clade II (Lineage 2). We therefore designated the x and y subunit haplotypes of TA2576 haplotypes within subclades 7 and 15. Geographical analysis reveals that TA2576 was collected from a region shared with other Lineage 3 accessions. However, the accessions containing *Dy15a* haplotype were not collected from a shared region with the L3 accessions. Although not conclusive, the most parsimonious explanation is therefore that *Dx7b + Dy15b* represents a recombinant haplotype between the x and y subunits from two different alleles.

Within our current panel, however, we are unable to differentiate exactly which original haplotypes gave rise to this recombinant haplotype.

Conclusion

Importance of Glu-D1 Diversity

The *Glu-D1* locus of wheat provides the greatest contribution to gluten strength, regardless of the allele present (Wang et al., 2017). The allelic diversity of *Glu-D1* in wheat is limited to two predominant alleles *2+12* and *5+10*, and a few rare alleles (*3+12*, *4+10*) which are associated with similar end-use quality as *2+12* (Dong et al., 2013; Y. Wan et al., 2005) . The unique *2.2+12* SDS-PAGE allele, which is found at high frequency in Japanese wheat, was shown to be identical to the *2+12* haplotype with the exception of additional repeats in the internal repeat domain of the x subunit(Payne, Holt, & Lawrence, 1983; Shimizu et al., 2020; Y. Wan et al., 2005). The x subunit protein from *5+10* has a unique cysteine residue just within the central repeat domain which is suspected to increase disulfide bonds in the forming dough. The early expression and greater transcription of this allele is also greater than that of the other *Glu-D1* alleles, in particular *2+12(Don et al., 2005)*. It is unclear which of these characteristics, or the combination of the two, lend *5+10* the superior quality characteristics. The unique origin of *5+10* from Lineage 3, however, further supports the important contributions of this lineage to the wheat D genome consistent with the findings by Gaurav *et al*. (2021).

Unique and Valuable Sources Of Diversity

Our haplotype analysis revealed that the x and y subunits are strongly associated even in diverse germplasm and that the *Glu-D1* haplotypes were clustered to specific geographic origins.

Consistent with the findings of Gaurav *et al.* (2021), we found evidence of two lineages (Lineage 2 and Lineage 3) contributing to the D genome of wheat with the superior *5+10* allele found associated with Lineage 3 accessions. Given the excellent end-use quality imparted by *5+10*, understanding this unique origin of the wheat allele support the further exploration and evaluation of novel *Glu-D1* alleles to further improve wheat quality. This also greatly supports the potential of novel alleles and unique haplotypes from the breadth of *Ae. tauschii* diversity.

Wheat grain quality remains one of the most important targets for breeders to develop superior wheat cultivars. Wild wheat relatives have been shown as a valuable resource for accessing novel genetic diversity to improve a range of wheat breeding targets including yield and disease resistance. For quality evaluation, however, the large quantity of grain needed for milling and baking and the confounding morphological characteristic needed for quality evaluation, such as suitable seed size for milling, make direct evaluation of various end-use quality traits intractable to phenotype directly these wild relatives, including *Ae. tauschii*. In this work, we therefore took the first step in a reverse genetics approach in *Ae. tauschii* by identifying and characterizing variants at the important *Glu-D1* locus. This demonstrated the unique origin of the *Glu-D1* allele in wheat as well as uncovering novel allele variants and haplotypes that can now be targeted for breeding. We also established the relation of wheat alleles to those of *Ae. tauschii* and have shown that *Ae. tauschii* contains a trove of unique *Glu-D1* alleles very unlike the alleles in current wheat germplasm. With accessible germplasm resources such as synthetic hexaploids (Gaurav et al., 2021), the diagnostic variants will enable marked-assisted selection of novel *Ae. tauschii* introgressions into wheat, characterization of their end-use quality, and utilization in wheat improvement.

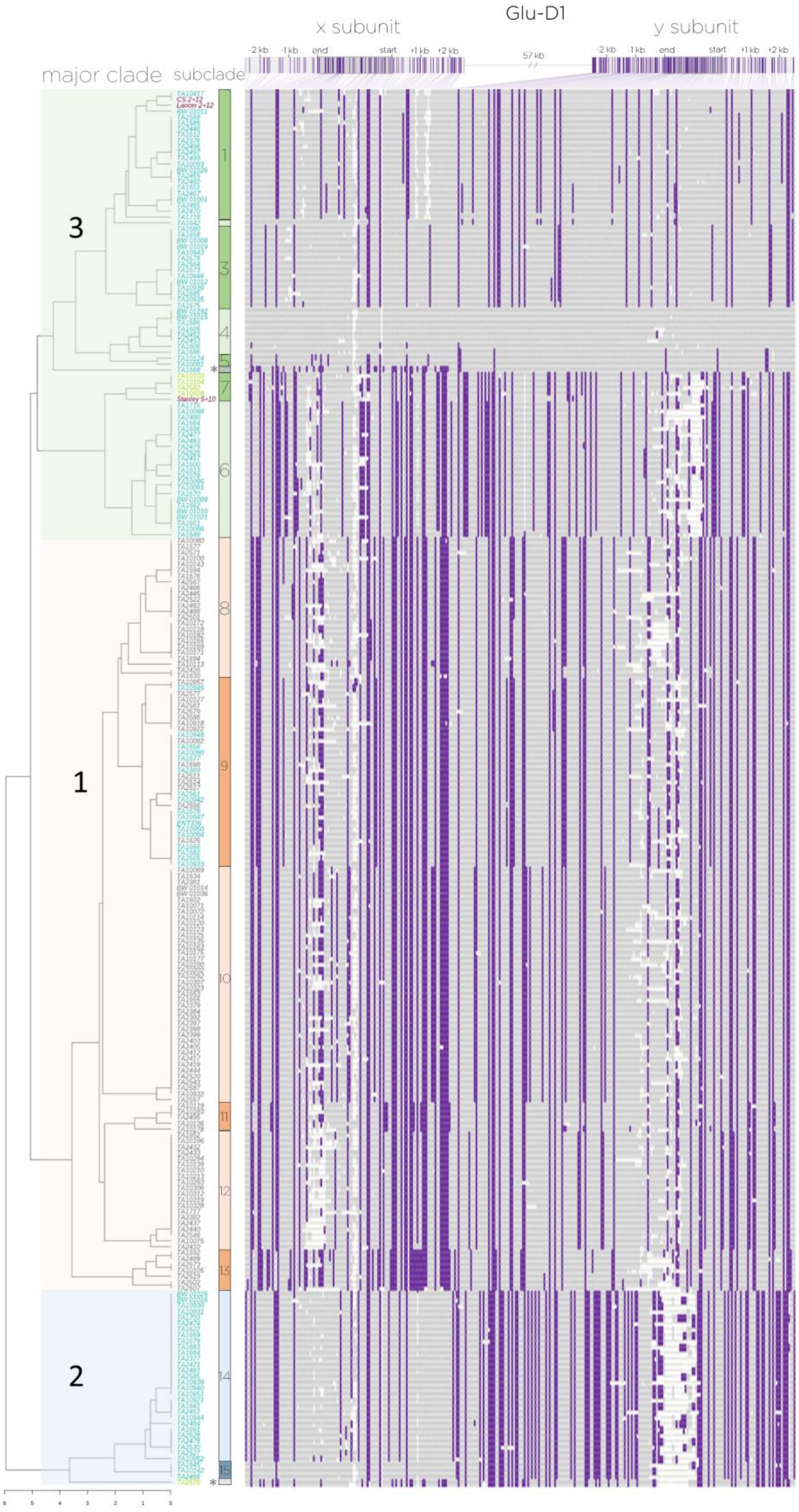

Glu-D1
x subunit
y subunit
-2 kb
-1 kb
end
start
+1 kb
+2 kb
57 kb
major clade
subclade
3
1
2
CS 2+12
Lancer 2+12
Stanley 5+10
1
3
4
5
*
7
6
8
9
10
11
12
13
14
15
*
6
5
4
3
2
1
0

Figure 4.1 - Molecular haplotypes of *Glu-D1 Ae. tauschii.*

Variant positions within the x and y subunit coding sequences and their 2.5 kb flanking sequences are marked with purple bars on schematic of the genes. The molecular haplotypes with position as column and accession as row are shown to the right of the dendrogram, reference allele is in gray and alternate allele is in purple. Dendrogram of combined x and y subunit haplotypes is shown to the left. The corresponding lineage of each accession is colored in gray (Lineage 1), aqua (Lineage 2) and lime green (Lineage 3) and burgundy (hexaploid wheat). Haplotypes with major clades and subclades are designated by numbers. The wheat alleles *Glu-D1* alleles *2+12* (varieties Chinese Spring and Long Reach Lancer) and *5+10* (variety CDC Stanley) are shown in purple and recombinant haplotypes are shown in grey with asterisk. Major clades are designated with numerical values 1, 2 and 3. Subclades are designated with numerical values 1-15 in the colored rectangles between the dendrogram and molecular haplotypes.

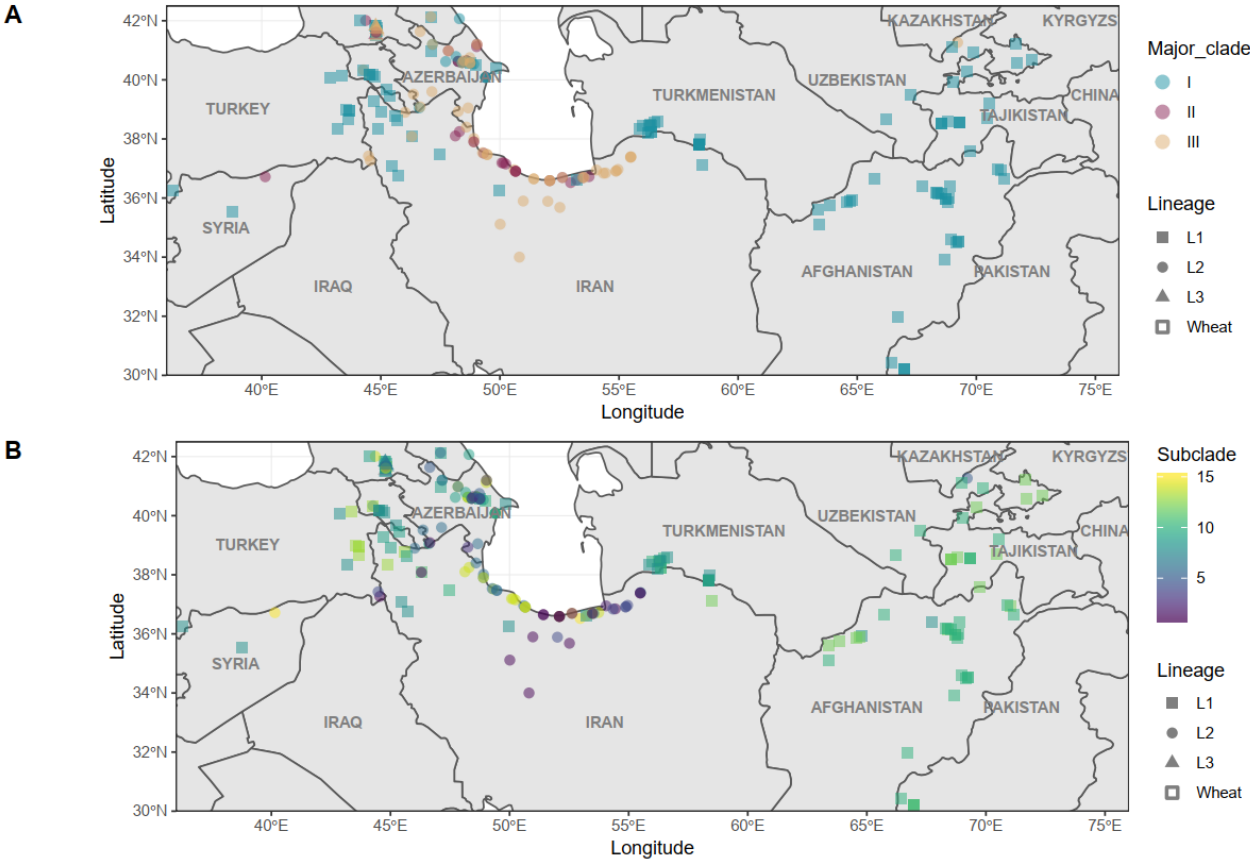


Figure 4.2 - Geographic distribution of *Glu-D1* haplotypes in *Ae. tauschii.*

Molecular haplotypes for Glu-D1 shown at the collection site for the given *Ae. tauschii* accession. (**a**) Distribution of accessions according to *Glu-D1* major clades with Clade I in blue, Clade II in red and Clade III in orange. (**b**) Distribution of accessions according to *Glu-D1* haplotype subclades. Haplotypes are shown on a scale from purple to yellow according to dendrogram order (Figure 4.1). Lineages are shown as circles for Lineage 1, triangles for Lineage 2 and squares for Lineage 3.

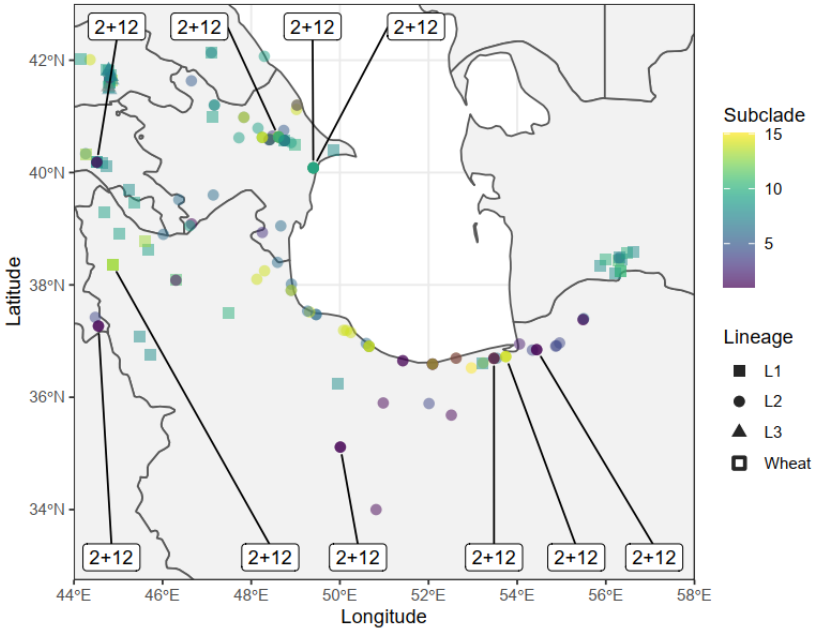


Figure 4.3 - Cryptic haplotypes within SDS-PAGE alleles *Ae. tauschii*.

A map of the region around the Caspian Sea where the *Aegilops tauschii* accessions were collected. Each point represents the collection site of a single accession. Points are colored by the high molecular weight glutenin haplotype subclade. Subclade numbers correspond to hierarchical clustering order by Euclidean distances based on genetic variants, therefore, subclade 15 (yellow) would be most distantly related to subclade 1 (purple). *Ae. tauschii* lineages are designated with the shape of the points, square (Lineage 1), circle (Lineage 2), triangle (Lineage 3) and wheat (unfilled square).

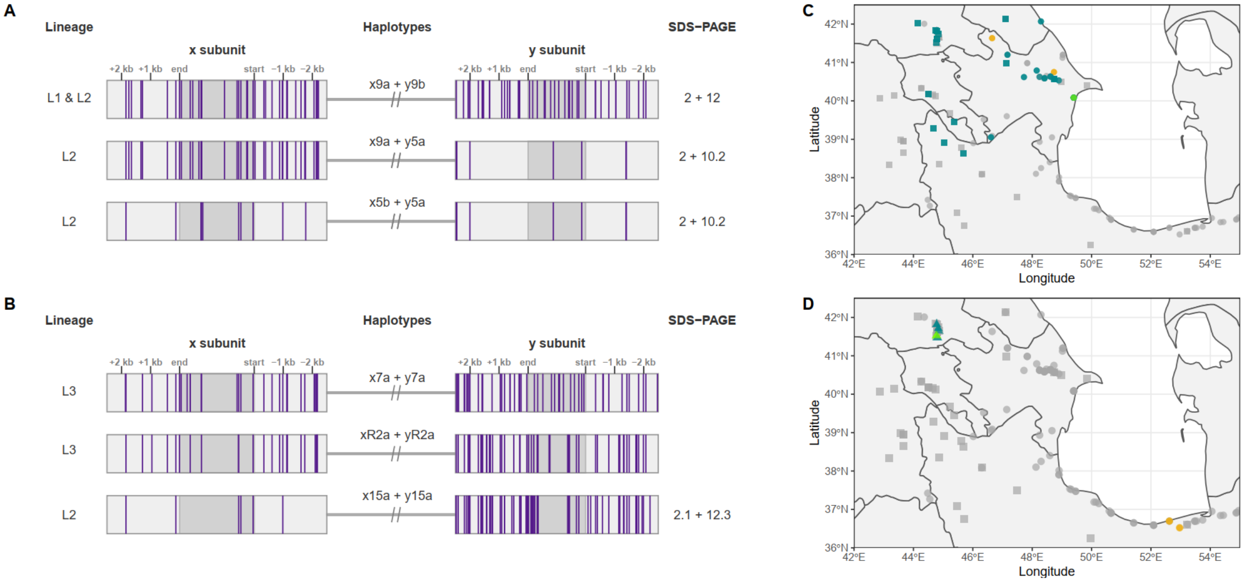


Figure 4.4 - *Glu-D1* x and y subunit recombinants in *Ae. tauschii.*

Molecular haplotype representation of recombinant *Glu-D1* haplotypes for accessions (**a**) TA1668 (Lineage 2) and (**b**) TA2576 (Lineage 3). Vertical purple bars represent the alternate allele variants as called against the AL8 7/8 genome assembly. SDS-PAGE allele for the given haplotypes are show to right of each gene model. Haplotype *Dx9a* is present in both Lineage 1 and Lineage 2 accessions. The closest potential x and y subunit haplotypes involved in the recombinant haplotype *xR2+yR2a* of TA2576 are *x7a* (Lineage 3) and *y15a* (Lineage 2). SDS-PAGE protein mobilities for *x7a + y7a* and *xR2+yR2a* were not analyzed. Geographical distribution of recombinant Glu-D1 haplotypes for (**c**) TA1668 (Lineage 2) and (**d**) TA2576 (Lineage 3). Collection site of recombinant accessions is marked in lime green, whereas turquoise and orange designate the collection sites of accessions carrying x subunit and y subunit haplotypes, respectively. Accessions with unrelated haplotypes are in light gray. Lineages are shown in squares (Lineage 1), circles (Lineage 2) or triangles (Lineage 3).

Table 4.1 - *Glu-D1* gene positions in *Ae. tauschii* reference genome assembly (Aet v4).

	Chr	Start position	End position	Length
Glu-D1x	1D	419306988	419309556	2568 bp
Glu-D1y	1D	419364015	419365995	1980 bp

Table 4.2 - *Glu-D1x* molecular haplotypes in *Ae. tauschii.*

The combination of x and y coding sequence haplotypes resulted in a cumulative total of forty-five haplotypes detected among 242 Ae. tauschii accessions and 33 unique x subunit haplotypes. Haplotypes are listed by major clade, followed by alphanumerical haplotype for the x subunit within the clade. The number of accessions possessing each haplotype is indicated along with the associated SDS-PAGE mobilities.

Major Clade	Lineages	Number of accessions	Haplotype	SDS-PAGE
III	L2	21	x1a	2x
III	L2	2	x1b	2.1x
III	L2	8	x1c	2x, 1.5x
III	L2	1	x2a	2.1x
III	L2	14	x3a	null, 2x, 2.1x
III	L2	2	x3b	
III	L2	11	x4a	2.1x
III	L2	1	x4b	1.5x
III	L2	3	x5a	4x
III	L2	1	x5b	2x
III	L2	19	x6a	4x
III	L2	8	x6b	4x
III	L2	2	x6c	4x
III	L3	7	x7a	
I	L1	9	x8a	2x
I	L1	6	x8b	2x
I	L1	7	x8c	2x
I	L1	1	x8d	
I	L1	2	x8e	2x
I	L1, L2	36	x9a	2x
I	L1	1	x10a	
I	L1	1	x10b	
I	L1	39	x10c	3x
I	L1	5	x11a	2x, 3x
I	L1	20	x12a	3x
I	L1	3	x13a	1tx
I	L1	2	x13b	2x, 4x
I	L1	4	x13c	2.1*x
II	L2	29	x14a	1.5x, 2x
II	L2	3	x14b	2x, 3x
II	L2	1	x14c	
II	L2	3	x15a	2.1x
II	L3	1	x16a	

Table 4.3 - *Glu-D1y* molecular haplotypes in *Ae. tauschii.*

The combination of x and y coding sequence haplotypes resulted in a cumulative total of forty-five haplotypes detected among 242 Ae. tauschii accessions and 32 unique y subunit haplotypes. Haplotypes are designated as subclade number, followed by alphabetical haplotype for the y subunit within the subclade. The number of accessions possessing each haplotype is given in ‘n accession’ column followed by the associated SDS-PAGE mobilities.

Major Clade	Lineages	Number of accessions	Haplotype	SDS-PAGE
III	L2	16	y1a	12y
III	L2	1	y1b	
III	L2	3	y1c	
III	L2	1	y1e	
III	L2	10	y1d	12y, 12.1*y
III	L2	1	y2a	12y
III	L2	14	y3a	10.1y, 10.2y
III	L2	2	y3b	
III	L2	4	y4?	10y
III	L2	4	y4a	10.1y
III	L2	3	y4b	10y
III	L2	1	y4c	10.1y
III	L2	4	y5a	10.2y
III	L2	21	y6a	10.2y, 12y, 12.1*y
III	L2	8	y6b	10y
III	L3	7	y7a	
I	L1	25	y8a	10y, 10.3y
I	L1	13	y9a	10y
I	L1, L2	14	y9b	11y, 12y
I	L1, L2	9	y9c	10y
I	L1	41	y10a	10y, 10.1y
I	L1	3	y11a	10y
I	L1	1	y11b	10y
I	L1	1	y11c	10.1y
I	L1	19	y12a	10.1y
I	L1	1	y12b	
I	L1	5	y13a	12y, 10y
I	L1	4	y13b	12.1*y
II	L2	30	y14a	12.2y, 12y
II	L2	3	y14b	12.2y
II	L2	3	y15a	12.3y
II	L3	1	y16a	

Chapter 5 - Haplotype analysis enables sequence-based-genotyping of high molecular weight glutenin alleles in wheat

Abstract

The quality of products produced from wheat flour depends largely on gluten forming proteins. These proteins are comprised of the high-molecular weight (HMW) glutenins, the low molecular weight (LMW) glutenins, and the gliadins. HMW and LMW glutenin proteins form inter- and intra-molecular bonds to create a large matrix that gives the viscoelastic properties which defines wheat dough. Depending on the glutenin alleles constituting the gluten matrix, the rheological properties of the wheat dough change, influencing the end-use application of the flour. Given that HMW glutenins are major determinants of the end-use quality, they are an important target of selection in wheat breeding programs. However, for molecular breeding of HWM glutenins, genotyping is often based on protein gels or PCR markers, both of which have limitations in low-throughput or limited differentiation of alleles. To understand the molecular structure of HWM loci and address these breeding challenges, we used the *de novo* assemblies of 11 wheat genomes to identify extensive sequence and structural variation across the HWM loci which were diagnostic for known SDS-PAGE alleles. Diagnostic 50-mers were developed from whole genome sequencing of 96 CIMMYT hard spring wheat founder varieties and 41 western plains hard winter wheat varieties and then leveraged in a genotyping pipeline that utilizes high-throughput skim-sequencing of breeding lines. We obtained 89% and 98% cross validation prediction accuracy for genotyping Glu-A1 and Glu-D1 respectively. This approach has the potential to offer a low-cost, high-throughput and diagnostic alternative to gel methods for gluten genotyping in breeding programs.

Introduction

End-use quality in wheat is due to the viscoelastic properties of the dough, which is conferred by the strong and elastic gluten protein network. The gluten network traps gas particles enabling the rising of bread and gives the porous texture of wheat baked goods. Gluten is a heterogeneous network of proteins comprised of the high molecular weight (HMW) glutenins, the low molecular weight (LMW) glutenins and the gliadins. HMW glutenins form the protein backbone of the gluten network through intramolecular disulphide bonds between cysteine residues. It is these disulphide bonds which provide strength and cohesivity of the gluten network. The number and position of cysteine residues available for intramolecular bonding that result in a strong or weak network. Typically, more cysteine residues result in a stronger dough, but position of the residues within the protein is also an important factor (Buonocore, Caporale, & Lafiandra, 1996; X. Gao, Zhang, Newberry, Chalmers, & Mather, 2012). The number and position of cysteine residues determines if glutenin will extend or end a chain in the gluten network. Provided that HMW glutenin alleles differ in the number and position of cysteine residues, and therefore also in the gluten strength in the dough, breeders can select for superior end-use quality by targeting specific alleles.

The unique viscoelastic properties of wheat are attributed primarily to the formation of the gluten network during dough development (P. R. Shewry & Halford, 2002). Of the gluten proteins, high molecular weight (HMW) glutenins are particularly important for dough strength and have the largest overall effect on end-use quality (Payne, Corfield, Holt, & Blackman, 1981). HMW glutenins are located on the long arms of the group 1 chromosomes with two paralogous genes at each locus, resulting in six HMW glutenin genes in each hexaploid genome. The paralogous genes are designated as x and y subunits which are separated by 50 to 200kb. Given the vast importance of these genes, we leveraged the wheat pan-genome assemblies which enabled characterization of these complex loci in detail.

Norin 61 carries the *Glu-D1f* allele that produces the *2.2 + 12* subunits of the high-molecular-weight (HMW) glutenin, first described by Payne et al. (1983) in Japanese hexaploid wheat varieties. HMW glutenin proteins are integral to end-use quality characteristics through the formation of the gluten matrix. The *Glu-D1* locus contains two genes, each encodes for an x and y subunits, that together constitute the HMW glutenin types (i.e. *2.2x + 12y* or *2x + 12y*). The *2.2 + 12* allele is also referred to as the f allele, and *2 + 12* as the *a* allele (Payne and Lawrence 1983). The very low electrophoretic mobility of the *2.2x* protein on sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) suggested that the protein was much larger in size than any other hexaploid HMW glutenin subunits known at the time. It was then hypothesized that *2.2x* arose from an unequal crossing over in the central repeat domain of *2x* that increased the size of the gene and that this novel allele is very recent, within the modern breeding history. This is supported by failure to identify *2.2+12* in surveys of *Glu-D1* SDS-PAGE mobilities in *Ae. tauschii* (Lagudah and Halloran 1988, William et al. 1993, Mackie et al. 1996), the D genome donor to hexaploid wheat, and that *2.2+12* is present primarily only in Japanese germplasm (Nakamura et al. 1999, Yanaka et al. 2016) indicated that 2.2x arose from recent mutation within hexaploid wheat. Further evidence for this hypothesis was provided by the full gene CDSs comparing 2.2x to 2x that showed 2.2x has a perfect 396 nucleotide duplication in the central repeat region (Wan et al. 2005).

Current genotyping methods for glutenins can differentiate a large number of alleles but are relatively low-throughput or contrastingly, high-throughput but restricted on the number of alleles that can be differentiated. The available genotyping methods include sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE) or PCR based methods such as KASP markers. SDS-PAGE for typing glutenin alleles is a methodology first described by (Moonen, Scheepstra, & Graveland, 1982) and modified by Singh *et al.* (N. K. Singh et al., 1991). It involves isolating the gluten network from ground kernels, treating the network with sodium dodecyl sulfate (SDS) to break the disulphide bonds and then separating the proteins on a polyacrylamide gel (PAGE). SDS-PAGE provides information about all three HMW-glutenin loci and the three LMW glutenin loci simultaneously, making it the gold standard for glutenin genotyping. However, it often cannot separate alleles with similar mobilities and can be difficult for many labs to perform and analyze. Additionally, the throughput is too low to for large-scale screening in a breeding program and typically only possible for genotyping advanced breeding lines and breeding program parents.

Recognizing the limitations of SDS-PAGE for genotyping, significant effort has been made to develop high-throughput PCR and KASP markers for HMW-glutenin alleles in the search for a more rapid throughput and sensitive system than SDS-PAGE. Some of these efforts have been successful with development of PCR marker assays for *Ax2** at *Glu-A1, Dx2, Dx5, Dy10, Dy12* at *Glu-D1* (S. Liu, Chao, & Anderson, 2008). While others have developed KASP markers from publicly published HMW-GS sequences (Ravel et al., 2020). However, the results are often limited to differentiating only a few alleles and ambiguous, and therefore these have not been widely adopted yet. Due to these various constraints, genetic profiling for glutenins is often limited to charactering breeding program parents rather than screening large, early-generation populations.

Here we sought to first understand the haplotype variation in the wheat HWM glutenin loci leveraging newly assembled wheat genomes for important varieties representing global diversity. We observe large structural variation across the known alleles and identified diagnostic polymorphisms between HMW glutenin alleles which correspond to the known alleles differentiated by SDS-PAGE. We also developed a methodology using sequence-based genotyping (SBG) that uses diagnostic k-mers to genotype HMW glutenin alleles from very low coverage skim-sequencing data. SBG is as accurate as SDS-PAGE while also being high-throughput and scalable to whole breeding programs and could be implemented in parallel with skim-sequencing for whole-genome profiling for genomic selection.

Materials and Methods

Populations

For this study, we utilized a set of 11 wheat varieties with completed *de novo* assemblies from the 10+ Wheat Genomes project (Walkowiak et al., 2020) to develop a set of diagnostic k-mers for tagging known glutenin alleles (Supplemental Table 1). We also included 93 CIMMYT hard spring wheat founder lines and 45 western plains hard winter wheat founder varieties.

To validate the sequence tags we utilized a set of 93 CIMMYT wheat parental lines that have been developed in the CIMMYT bread wheat breeding program. The 10+ Genomes and the CIMMYT parental lines were characterized for known HWM glutenin alleles with SDS-PAGE (Supplemental Table 1). The CIMMYT parental lines were sequenced to ~ 10-fold coverage with Illumina pair-end sequencing. PCR-free genomic libraries were constructed with targeted 350bp inserts and sequenced with 2 x 150bp reads.

For validation of genotyping pipeline, we utilized a large set of breeding lines in the CIMMYT bread wheat program that were typed with SDS-PAGE as they entered the crossing block. This set consisted of 805 advanced lines (F5 or F6-dervied). These lines were skim-sequenced on Illumina NovoSeq to approximately 0.05x coverage using the methods of Adhikari, Shrestha et al., (PENDING).

SDS-PAGE Analysis

SDS-PAGE analysis was conducted for the 10+ Genomes lines and 60 of the CIMMYT founder lines as described by Pena (2004) and allele nomenclature in accordance with Payne and Lawrence (Payne & Lawrence, 1983). SDS-PAGE alleles for the western plains hard winter wheat varieties were referenced from literature reports (Table 5.1).

Haplotype Analysis of 10+ Genomes High Molecular Weight Glutenin Genes

A completely assembled reference allele for each of the HMW glutenins was created by manually combining the respective HMW glutenin coding sequences of Chinese Spring RefSeq v1.0 (Appels et al., 2018) and Chinese Spring Triticum 3.1 (A. V. Zimin et al., 2017). The positions of the HMW glutenins of each assembly were identified using blastn of NCBI BLAST v2.6.0 (Altschul, Gish, Miller, Myers, & Lipman, 1990) with E value < 0.0005, output format 6 and query sequences from Triticum 3.1 The incompletely assembled nature of the HMW glutenins made sequence retrieval difficult, therefore a custom R script was developed that scanned reads to identify the start and stop of each gene and is available in the Github repository for this study. The script output a bed file of gene positions and the getfasta command of BEDtools v2.19.1 (Quinlan & Hall, 2010) was used to retrieve sequences as a fasta file. The Chinese Spring sequences were aligned manually with Jalview v2 (Waterhouse, Procter, Martin, Clamp, & Barton, 2009) and the final reference alleles output as a fasta file. Of particular note is the *Glu-A1y* (null) allele in Chinese Spring that is inactivated by a 10.8 kb wis-2 insertion. This large insertion was removed from the reference allele for better comparison to the other alleles.

The HMW glutenin positions and sequences were retrieved from the other 10 genomes as described above, but with the addition of the Chinese Spring reference alleles to the query fasta file. Alleles for each HMW glutenin gene were aligned with Clustal Omega Multiple Sequence Alignment tool with default parameters (Madeira et al., 2019). Alignments were manually corrected with Jalview v2. Sequence variants were identified and positions calculated from the start of the Chinese Spring reference allele. Sequence haplotypes were determined by exact sharing of variants.

Sequence variants were confirmed by variant calling using the raw sequencing data for each of 11 genomes using the paired end 2x250 bp Illumina sequencing data of 470bp insert libraries. Adapters and low quality bases were trimmed from raw sequencing data using the Bbduk tool of the Bbtools package (http://jgi.doe.gov/data-and-tools/bbtools/). The cleaned forward and reverse reads were aligned together to the Chinese Spring RefSeq v1.1 genome assembly with HISAT2 default parameters and preventing spliced alignments with '--no-spliced-alignment' option (Kim et al., 2019) with an average depth of 10x over the HMW glutenins. Variants for HMW glutenins were called with bcftools v1.6 (H. Li, 2011) for each genome independently. First the genotype likelihoods were found with mpileup with options -f and -R to generate a faidx indexed fasta file for specified HMW glutenin regions. The file was piped into the call tool with options -m and -Ou for multiallelic calling and to generate an uncompressed vcf file. The vcf file was filtered for genotype quality (GQ) scores greater than 20 and read depth greater than 1 with the filter tool.

Nucleotide variation in the Glu-D1 locus of Norin 61 was determined by aligning 470-bp PE Illumina reads of all 10 + Genome varieties to the CS reference genome v1 with HISAT2 v2.1.0 using default parameters (Kim et al., 2019) and alignment sorting and indexing done with samtools v1.6 (H. Li et al., 2009). Variants within the Glu-D1 subunits, the 57 kb between subunits and 100 kb flanking were called with bcftools v1.11 ‘mpileup’ and ‘call’ with minimum alignment quality score of 20 and ‘–group-samples–’ option (Li 2011). Heterozygous calls were set to missing and missing variants exceeding a proportion of 0.1 filtered out. Structural variation between the *2.2+12* locus of Norin 61 genome assembly v1.1 and *2+12* locus of CS genome assembly v1 was assessed with MUMmer v3.23 using default parameters (Kurtz et al., 2004).

Chinese Spring v1.0 Augmented Genome Assembly

To recover reads unique to all of the HMW glutenin alleles while minimizing alignment computing requirements, we developed an augmented reference genome assembly for the Chinese Spring RefSeq v1.0. For this augmented assembly the HMW glutenin loci from each of the 10+ Genomes assemblies (Walkowiak et al., 2020) and the *Aegilops tauschii* v4 assembly (Luo et al., 2017) were concatenated to the Chinese Spring v1.0 reference genome (Appels et al., 2018) and referred to hereafter as CSv1_HMW-GS assembly. This augmented assembly enabled alignment and extraction of unique sequence reads that were not present in the HWM *Glu* alleles of Chinese Spring.

Haplotype Analysis For Expanded Founders Populations

Molecular haplotypes of high molecular weight glutenin loci were determined using methodology similar to that described in Chapter 2 of this dissertation. Whole genome sequencing data was trimmed using FASTp and aligned to the Chinese Spring RefSeq v1.0 assembly using HISAT2 default parameters. Alignments were filtered for mapping quality greater than 20 (-q 20), converted to bam format, sorted and indexed using samtools. All variants from reads with mapping quality greater than 20 (-q 20) were called within the HMW-GS loci with bcftools. Each sample was treated as its own group with '–groupsamples –' to reduce heterozygous calls. Variants were filtered twice, first for read quality depths for reference or the alternate allele greater than or equal to 5 and for estimated variant quality greater than 100 (FMT/DP>=5 & FMT/AD>=5 & QUAL>100) with bcftools. The variants were filtered again for minor allele frequency > 0.05, heterozygosity < 10%, and missing calls < 95%.

Identification of Diagnostic K-mers

To minimize impact of assembly and sequencing errors for variant calling and k-mer identification, we employed the raw-sequencing data from the 10+ Wheat Genome assemblies. The data for these genomes utilized Illumina sequenced to 30x coverage with 250 bp paired-end reads of 470 bp insert libraries on Illumina HiSeq2500. Adapters and low quality bases were trimmed from raw sequencing data using the Bbduk tool of the Bbtools package (http://jgi.doe.gov/data-and-tools/bbtools/). The cleaned forward and reverse reads were aligned together to the respective genome assembly with HISAT2 default parameters and preventing spliced alignments (Kim et al., 2019). Alignments were sorted and indexed using samtools v1.6 and reads aligning to the HMW glutenin loci were extracted using samtools view command (H. Li et al., 2009). For each of the genomes the HMW glutenin regions were defined as the x and y subunit, the sequence space between and additional 2000bp proximal and terminal (Supplemental Table 2). The read file was converted from a SAM to a fasta using awk. 50-mers were generated using kmerextract of Bbtools with minimum count of 10 to exclude those due to sequencing errors. This set of k-mers was defined as the starting set for identification of diagnostic k-mers.

Testing and Validation Of Genotyping Pipeline

The validation set of approximately 800 CIMMYT crossing block lines were skim-sequenced from ~600bp insert libraries using the modified Nextera libraries. Adapters and low quality based were trimmed as before and aligned to the CSv1_HMW-GS assembly as described above using HISAT2. Reads aligning to the HMW glutenin loci, including the augmented set, were extracted using samtools view command as described previously and K-mers were generated using kmerextract of BBtools with minimum count of 1.

Results and Discussion

Haplotype Analysis of 10+ Genomes High Molecular Weight Glutenin Genes

The HMW glutenins possess short and highly conserved N (243 – 312 bp) and C terminal domains (42 bp) separated by a complex repeat domain ranging from approximately 1.4 to 2kb (P. R. Shewry, Halford, & Tatham, 1992). Even with very high-quality assemblies, we found these complex repeat domains were not fully assembled in the genomes, though completed to a single scaffold and correctly ordered. This incomplete assembly with many gaps was likely due to the short-read technology utilized for the 10+ Genomes assemblies as we observed the Chinese Spring Triticum 3.1 assemblies (Aleksey V. Zimin et al., 2017), which utilized long read sequencing, had complete assembly of the complex repeats. To overcome this limitation in the assemblies we combined and manually curated the HMW glutenin loci for two Chinese Spring genome assemblies, RefSeq v1.0 (Appels et al., 2018) and Triticum 3.1 assemblies (Aleksey V. Zimin et al., 2017). The resulting manually curated and sized genes were used as templates for the multiple sequence alignments of the other 10+ Genomes HMW glutenin alleles.

The different HMW glutenin alleles are classically determined by sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE) profiling of the proteins (Payne, Corfield, & Blackman, 1979; Peter I. Payne et al., 1981). To examine the molecular structure of the different known alleles, we compared SDS-PAGE alleles with newly characterized molecular variants at each HMW glutenin locus in the wheat genomes (Figure 5.1). We found that molecular haplotypes directly corresponded with each of the known HWM alleles, in all but two cases. In the first, we were able to differentiate two different sequence haplotypes that were sized as the same *7+8* allele of at *Glu-B1*. These 'crypic' haplotypes are differentiated at the molecular level, but not at the protein level in the SDS-PAGE. In the second, *Glu-D1 2.2+12* could not be differentiated from *2+12*. The *2.2+12* allele was first described in Japanese hexaploid wheat as having a unique x subunit with lower SDS-PAGE mobility than any other x subunits known at that time (Payne et al., 1983). It was previously hypothesized that unequal crossing over in the central repetitive domain of 2x in the 2+12 allele led to the larger *2.2x* protein. By comparing the coding sequences of 2*.2+12* and 2*+12,* as well as the SNPs within the Glu-D1 locus, we found evidence to support the hypothesis that 2.2+12 originated in hexaploid wheat from the 2+12 allele due to uneven crossing over resulting in an expansion of the repeat region.

Within the 257-kb region around the *Glu-D1* locus, we surveyed SNPs. There exists only 10 SNPs differentiating *2.2 + 12* from *2 + 12* of CS and other wheat varieties (Long Reach Lancer, ArinaLrFor, Mace, SY Mattis). The number was much less than the 72 SNPs differentiating the *2 + 12* sequences among the wheat varieties CS, Long Reach Lancer, ArinaLrFor, Mace and SY Mattis in the same region. Analysis with MUMmer showed no large-scale structural differences between *2.2 + 12* and *2 + 12* of CS, aside from 11 gaps ranging between 33 and 5,381 bp. Gaps were likely assembly artifacts, evidenced by Norin 61 reads aligning to the CS ‘gap’ regions. Due to the highly repetitive nature of the central repeat domain and the limitations of short-read sequencing, we were unable to directly detect the 396 duplication in *2.2x* relative to *2x*. However, with the genome assembly of Norin 61, we support here that the unique *2.2 + 12* in Japanese germplasm arose from a recent mutation of *2 + 12*, as well as identify several molecular variants that could be used as a high-throughput molecular marker in breeding to differentiated *2.2 + 12* from the common *2 + 12.*

Norin 61 carries the *2.2+12* allele at the *Glu-D1* locus, first described by (Payne et al., 1983) in Japanese hexaploid wheat varieties. The very low electrophoretic mobility of *2.2x* on sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE) suggested that the protein was much larger in size than any other hexaploid HMW glutenin subunits known at the time. It was hypothesized then that *2.2x* arose from an unequal crossing over in the central repeat domain of *2x* that increased the size of the gene. Failure to identify *2.2+12* in surveys of *Glu-D1* SDS-PAGE mobilities in *Aegilops tauschii* (Lagudah & Halloran, 1988; Mackie et al., 1996; William et al., 1993), the D genome donor to hexaploid wheat, and that *2.2+12* is present primarily only in Japanese germplasm indicated that *2.2x* arose from recent mutation within hexaploid wheat. Further evidence for this hypothesis was provided by the full gene coding sequences comparing *2.2x* to *2x* that showed *2.2x* has a perfect 396 nucleotide duplication in the central repeat region (Y. Wan et al., 2005).

Within the 257 kb region around the *Glu-D1* locus, there exists only 1 single nucleotide polymorphism differentiating *2.2+12* from *2+12* of Chinese Spring. An additional SNP differentiates Norin 61 and Chinese Spring *Glu-D1* locus from the *2+12* present in other wheat varieties (Long Reach Lancer, ArinaLrFor, Mace, SY Mattis). Analysis with MUMMER showed no large scale structural differences between *2.2+12* and *2+12* of Chinse Spring. With the genome assembly of Norin 61 we confirm here that *2.2+12* arose from a recent mutation of *2+12*. (Payne & Lawrence, 1983)

The overall conserved sequence haplotypes within HMW glutenin alleles of globally representative wheat genomes indicates that the functional alleles at the important glutenin genes are each of single origin and shared among global breeding programs. The completed assemblies and detailed sequence characterization of the alleles also enables development of effective high-throughput markers for molecular breeding as well as differentiating and predicting alleles using high-throughput skim sequencing.

Haplotype Analysis Of Founder Lines

Over 131 wheat founder varieties in the CIMMYT and western plains, we discovered few variants within the coding regions of the x and y subunits. However, when we expanded the haplotype space to include the region between each subunit and 2 kb flanking the edge of each subunit, we were able to distinguish haplotypes more reliably through 2429 variants at *Glu-A1,* 523 at *Glu-B1* and 550 at *Glu-D1*. From this haplotype analysis we were able to differentiate three haplotypes *Glu-A1,* two for *Glu-D1,* and ten for *Glu-B1*.

We found that the SDS-PAGE alleles and molecular haplotypes perfectly corresponded for *Glu-A1* and *Glu-D1* (Figures 5.2 and 5.4), owing most likely to the fact that there were few alleles at these two loci which were highly differentiated. Likewise, we were able to find correspondence between molecular haplotypes at SDS-PAGE alleles at *Glu-B1*. Notably, however, we discovered cryptic molecular haplotypes within *Glu-B1* SDS-PAGE alleles (Figure 5.3). For some *Glu-B1* SDS-PAGE alleles, such as *7+8*, there existed only one molecular haplotype. While for others, in particular *7+9*, there were three different molecular haplotypes. Two of these haplotypes were relatively closely related while others were very different (Figure 5.3). These results show that *7+9* mobility is shared among many different alleles, including highly divergent alleles.

There exists confusion in the wheat breeding community surrounding the 7+8 allele. It has been noted that there exists more than one *7+8* allele as indicated by slight differences of the 7x subunit in SDS-PAGE. The reference *7+8* allele (*Glu-B1b)* is found in Chinese Spring, and the slightly different allele was given the designation *7*+8 (Glu-B1u).* The *7+8* allele was thought to be very rare in North American germplasm and the commonly seen allele was in fact *7*+8* (Wrigley et al., 2009). In our population of western hard spring wheat and CIMMYT hard spring wheats, the Chinese Spring *7+8* haplotype was unique and not found in any other accessions. However, there were many haplotypes associated with *7+8* SDS-PAGE mobilities. Although Duster was reported to possess *7*+8* (Edwards et al., 2012), the molecular haplotype was most closely related to *7OE+8.* KS Hatchett (KS090049K-8), a progeny of Duster, also shared this haplotype.

From the CIMMYT program, breeding lines, GID7634384 and GID8059716, also possessed the *7OE*+8 haplotype, though GID7634384 appeared to have *7+8* on SDS-PAGE. There are up to three *Bx7OE* alleles. One confers good quality (*Glu-B1al*) and one confers poor quality (*Glu-B1br*). *Glu-B1br* is present in an Australian cultivar H45, also known as Galaxy (X. Gao, Appelbee, Mekuria, Chalmers, & Mather, 2012). The poor quality of the allele is thought to be due to a SNP in the repetitive region which results in a cysteine residue there (X. Gao, Zhang, Newberry, Chalmers, & Mather, 2013). The researchers postulated that the cysteine residue causes interference in gluten polymerization, however, experimental evidence of this hypothesis has not yet been provided. The conclusion from their work is that not all *Bx7OE* alleles are good for quality. Additionally, and what we are probably seeing, is that it appears Kanred has a related (possibly ancestral) allele to *Bx7OE* (Butow et al., 2004). Kanred is present seven times in the pedigree of Duster, but none of the known Bx*7OE* carriers are.

Another interesting, and unexpected, finding was that West Bred Cedar possesses a highly unusual Glu-B1 allele which is most similar to *6.1+22.1* in spelt wheat. WB Cedar resulted from a cross between TAM302 and experimental germplasm with the Pioneer variety 2180. Seeing that 2180 is in many other varieties, such as of Endurance (http://ofss.okstate.edu/seed/wheat03) and Gallagher (Marburger et al., 2021), we do not believe the allele originated from 2180. Another Pioneer variety, 2145, also appeared to possess a spelt type *Glu-B1* allele. However, it is difficult to confirm given that the 2145 sample was heterogeneous and thus all of its spelt type variants were also heterozygous.

Development of SBG Pipeline

We hypothesized that the conserved and highly divergent haplotypes between alleles at the *Glu-D1* or *Glu-A1* locus could be used to predict allele identities from sequencing data. As a proof of concept, we first cross validated prediction ability on the whole genome sequenced founder lines (10+ Genomes, western hard winter founders and CIMMYT hard spring founders). All founder lines were aligned to the augmented Chinese Spring assembly, filtered for alignment quality and reads aligning to the locus of interest were converted to 50-mers (Figure 5.7).

Twenty percent of samples within each haplotype were held out of the diagnostic k-mer training and these became the prediction set. Diagnostic k-mers were trained on the remaining 80% of samples. Diagnostic k-mer training was relatively naïve. A k-mer was considered diagnostic for a haplotype as long as it was unique to only one haplotype. Predictions were on a per training set sample basis, that is, the predicted allele was the one that belonged to training set sample with the highest correlation to the prediction set sample. The SBG pipeline was repeated 5 times with the assignment of lines to either the training or prediction set being random.

The resulting prediction accuracies were 89% and 98% for *Glu-A1* and *Glu-D1* respectively. Glu-D1 alleles *2+12* and *5+10* both had good prediction accuracies, with *2+12* successfully predicted 93% of the time and *5+10* at 99% (Figure 5.6). In *Glu-A1*, the most common allele, *2**, was the most successfully predicted while the less common alleles were often incorrectly predicted as *2** (Figure 5.6). Suspecting that the sheer representational imbalance of alleles in *Glu-A1* was responsible for the disparity, we attempted to balance the training set data by drawing fewer lines of the *2** haplotype. This did improve the prediction accuracy for the minor allele to a small degree, but at the expense of predictive ability of the major allele and therefore the overall predictive ability of the pipeline. We therefore concluded that additional data improves predictive ability and that less than 5 representative accessions of the minor allele restricts predictive power for that allele.

Encouraged by the results of the founder cross validation, we tested SBG pipeline on a skim sequenced set of 965 CIMMYT accessions. The accessions were chosen because they participated in the crossing block of the hard spring wheat breeding nursery in the last 10 years. They therefore capture the HMW-GS diversity in the CIMMYT wheat breeding program. Additionally, the accessions all have SDS-PAGE mobilities determined for both HMW-GS and LMW-GS. Although, we did not apply SBG to LMW-GS in this study, we were aware that if SBG were successful for HMW-GS then the next logical step would be to expand it to predicting LMW-GS alleles. Most importantly, these accessions were also characterized for the entire battalion of end-use quality characteristics. We expected to detect cryptic alleles, particularly in *Glu-B1*. If SBG were to successfully distinguish these alleles then the effect on end-use quality could be determined as well.

One hundred of the accessions were measured in more than 1 year for both SDS-PAGE and end-use quality traits. This allowed us to calculate the SDS-PAGE accuracy. Discrepancies between SDS-PAGE alleles could be due to heterogeneity, heterozygosity, or sample errors. SDS-PAGE accuracy was measured as number of total number of accessions minus number of discrepancies at any HMW-GS locus, over total number of accessions. The measured accuracy was 94% and the theoretical maximum prediction accuracy of SBG on this sample set.

The alignment and k-mer making pipeline portion of SBG was applied the same as before in the cross validation approach (Figure 5.6). The overall number and depth of the 50-mers were much less for the skim sequencing data. The approximate depth of skim sequencing on the crossing block set was 0.05x whereas the whole genome sequencing depth of the founder lines was closer to 10x. Forty one samples failed to yield enough sequencing data for any reads to align to any of the HMW-GS loci and were discarded.

Diagnostic 50-mers for the crossing block samples were found by considering data from only the 93 CIMMYT founders and excluding any founders that appeared to have predominately heterozygous calls for haplotype variants. Again, highest correlation between the kmers of a given sample and the prediction set was used to determine allele in the sample. Given that skim sequencing data is too low coverage to determine the actual haplotypes, the SDS-PAGE alleles were used as a proxy for haplotype. This assumed that there were not additional HMW-GS haplotypes in the crossing block samples set that were not present in the founder training set, although we are fairly certain that this assumption is false given the results of our *Ae. tauschii*

HMW-GS diversity analysis and that the CIMMYT breeding program utilizes material with wild relative crosses.

The overall prediction accuracies of SBG on skim sequencing data were 61% and 89% for *Glu-A1* and *Glu-D1*, respectively (Figure 5.6). These prediction accuracies were theoretically constrained by the cross validation prediction accuracies. We saw in both *Glu-A1* and *Glu-D1* that prediction ability was greatest for major alleles from the training set and that attempting to balance the training set by withholding samples of the major allele did not improve prediction accuracy. Notably, *Glu-D1 2+12* prediction accuracy in the entire WGS cross validation was 93%, however it felt to just 11% when the CIMMYT breeding lines were used as the training set and the crossing block as the prediction set. There were nearly 10 fold more CIMMYT founder lines with *5+10* than *2+12*, and therefore also 10 fold more diagnostic kmers for *5+10* than for *2+12*. Extreme imbalances between alleles such as this in the training set should therefore be avoided.

Glu-B1 predictions for SBG are not shown here because most alleles were too highly related and complex to be predicted with any accuracy. Further work is required to determine if better training of the model through training set selection, identification of diagnostic k-mers or other statistical relatedness measurements could improve prediction accuracies. At this time, SBG is not recommended for *Glu-B1* allele genotyping. Instead, KASP markers for any of the 523 variant sites identified in this work would likely provide a much more reliable genotyping method for *Glu-B1* haplotypes.

Conclusion

In this work, we characterized the molecular haplotypes the high molecular weight glutenin loci of 145 bread wheat varieties. The varieties represented global breeding programs with wider consideration given to the western plain hard winter wheats and CIMMYT hard spring wheats. We used SNPs from high coverage whole genome sequencing to develop high resolution molecular haplotypes of each HMW-GS locus. The haplotypes revealed strongly conserved loci at *Glu-A1* and *Glu-B1.* Given the results from Chapter 2 of this dissertation, we expected the intergenic regions between the x and y subunits of *2+12* and *5+10* alleles of *Glu-D1* to be structurally diverged due to their proposed origins in two separate lineages of *Ae. tauschii.*

The haplotype analysis revealed that while the three *Glu-A1* and two *Glu-D1* common alleles were shared among the relevant breeding programs, that the *Glu-B1* alleles were sometimes unique within programs. We discovered cryptic molecular haplotypes within two *Glu-B1* SDS-PAGE alleles. One of the most interesting cases being that the *Glu-B1 17+18* SDS-PAGE allele in the CIMMYT program is an entirely different allele than the *17+18* in western plains breeding programs. We also found a Glu-B1 haplotype in the CIMMYT program that is very unlike any of the others.

Future work is required to determine the end-use quality characteristics of the cryptic and novel *Glu-B1* haplotypes. Given that the varieties sequenced in this study are awaiting SDS-PAGE analysis and already have been measured for end-use quality traits, the prelimary analysis should be straightforward and forthcoming. With the SNPs identified in this study, KASP markers can be designed to tag the haplotypes for future studies on additional germplasm.

KASP markers are the most practical genotyping approach for *Glu-B1* given the highly conserved sequences between alleles. However, the sequence and structural divergence between *Glu-A1* and *Glu-D1* alleles are amenable to the novel genotyping approach described in this chapter, sequence-based-genotyping (SBG). We showed that SBG can reliably predict *Glu-A1* and *Glu-D1* haplotypes in high coverage sequencing data, and that it has acceptable prediction accuracy for skim sequenced samples. SBG could be improved by optimizing the training set and methodology for detecting diagnostic k-mers.

Table 5.1: Positions of HMW-GS genes in 10+ Genomes wheat assemblies.

Each locus contains the x subunit and y subunit. Locus start is -2 kb from start position of the x subunit and +2 kb from the end of the y subunit. The augmented assembly, CS HMW-GS, is the full Chinese Spring RefSeq v1.0 assembly with the HMW-GS loci of each of the following assemblies added. Each locus has its own fasta header in order to behave like a chromosome during the sequence-based-genotyping pipeline.

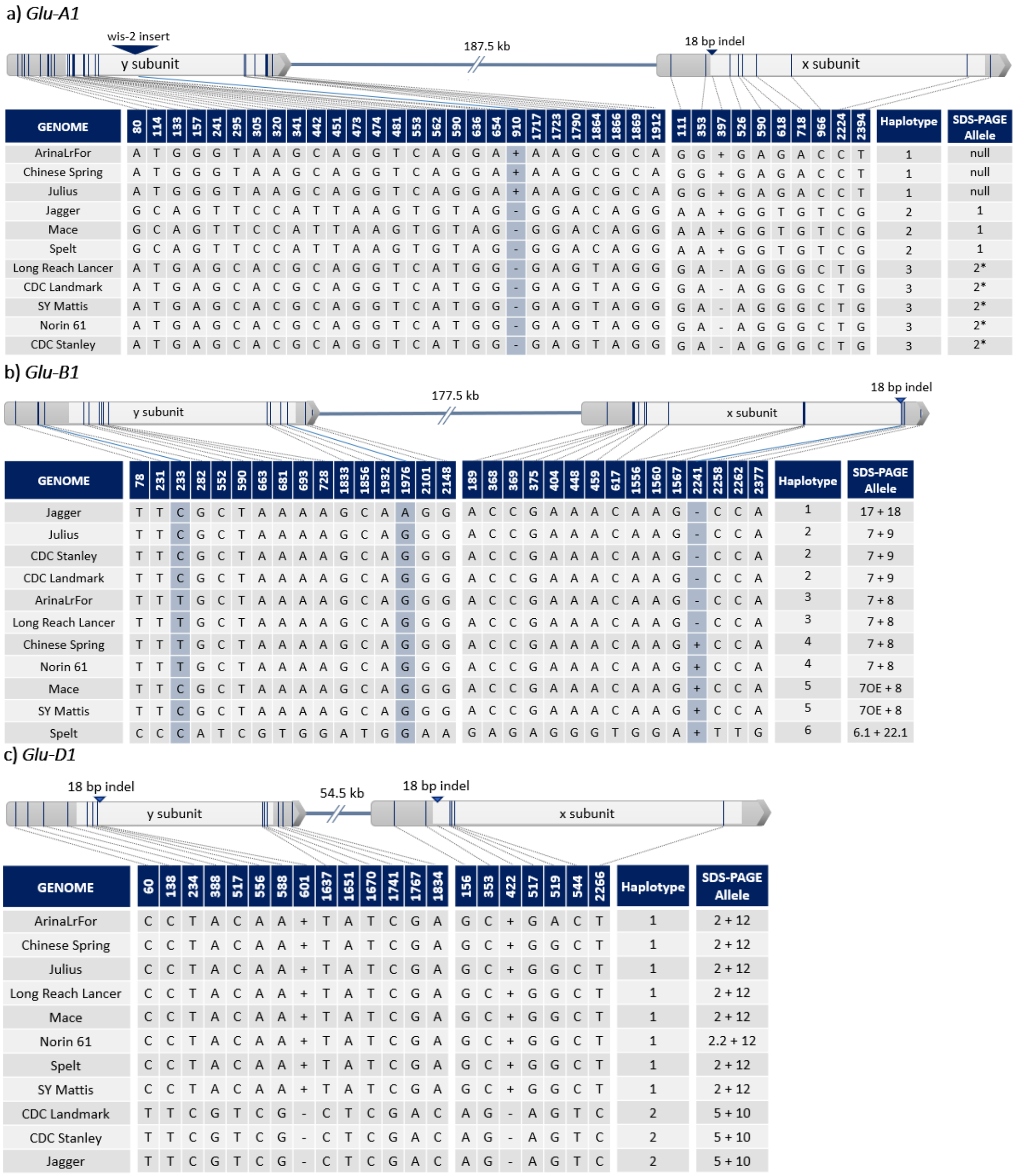
Genome	Locus	Subunit	SDS-PAGE Allele	Chr	Start position	End position
Chinese Spring	Glu-A1	x	null	chr1A	508726319	508723998
Chinese Spring	Glu-A1	y	null	chr1A	508916062	508914850
Chinese Spring	Glu-A1	y	null	chr1A	508925603	508924816
Chinese Spring	Glu-B1	x	7+8	chr1B	555766152	555765126
Chinese Spring	Glu-B1	y	7+8	chr1B	555935716	555933488
Chinese Spring	Glu-D1	x	2+12	chr1D	412163311	412160785
Chinese Spring	Glu-D1	y	2+12	chr1D	412219631	412217775
Arina	Glu-A1	x	null	chr1A	515185153	515183154
Arina	Glu-A1	y	null	chr1A	515373772	515372680
Arina	Glu-A1	y	null	chr1A	515390113	515389203
Arina	Glu-B1	x	7+8	chr1B	561737430	561735889
Arina	Glu-B1	y	7+8	chr1B	561910781	561909141
Arina	Glu-D1	x	2+12	chr1D	407196086	407194829
Arina	Glu-D1	y	2+12	chr1D	407252053	407250608
CDC Landmark	Glu-A1	x	2*	chr1A	507679179	507676625
CDC Landmark	Glu-A1	y	2*	chr1A	507866629	507864735
CDC Landmark	Glu-B1	x	7+9	chr1B	564481946	564479636
CDC Landmark	Glu-B1	y	7+9	chr1B	564659128	564656812
CDC Landmark	Glu-D1	x	5+10	chr1D	415924677	415921653
CDC Landmark	Glu-D1	y	5+10	chr1D	415980417	415978297
Jagger	Glu-A1	x	1	chr1A	509714763	509712422
Jagger	Glu-A1	y	1	chr1A	509913341	509911544
Jagger	Glu-B1	x	17+18	chr1B	565994772	565992542
Jagger	Glu-B1	y	17+18	chr1B	566179164	566176837
Jagger	Glu-D1	x	5+10	chr1D	409324900	409322316
Jagger	Glu-D1	y	5+10	chr1D	409379863	409378215
Julius	Glu-A1	x	null	chr1A	504708294	504706362
Julius	Glu-A1	y	null	chr1A	504897408	504895526
Julius	Glu-A1	y	null	chr1A	504907052	504906036
Julius	Glu-B1	x	7+9	chr1B	569169531	569167842
Julius	Glu-B1	y	7+9	chr1B	569340366	569338961
Julius	Glu-D1	x	2+12	chr1D	409051667	409050013
Julius	Glu-D1	y	2+12	chr1D	409107438	409106167
Long Reach Lancer	Glu-A1	x	2*	chr1A	509405648	509403898
Long Reach Lancer	Glu-A1	y	2*	chr1A	509592869	509591373
Long Reach Lancer	Glu-B1	x	7+8	chr1B	556729416	556727640
Long Reach Lancer	Glu-B1	y	7+8	chr1B	556897611	556896136
Long Reach Lancer	Glu-D1	x	2+12	chr1D	408915948	408914277
Long Reach Lancer	Glu-D1	y	2+12	chr1D	408972046	408970432

Mace	Glu-A1	x	1	chr1A	506595893	506593879
Mace	Glu-A1	y	1	chr1A	506801909	506800293
Mace	Glu-B1	x	7OE+8	chr1B	557705753	557705087
Mace	Glu-B1	x	7OE+8	chrUn	260082208	260081542
Mace	Glu-B1	x	7OE+8	chrUn	368348997	368348608
Mace	Glu-B1	y	7OE+8	chr1B	557886456	557884850
Mace	Glu-D1	x	2+12	chr1D	408824036	408822140
Mace	Glu-D1	y	2+12	chr1D	408880150	408878489
Norin 61	Glu-A1	x	2*	chr1A	508841175	508839370
Norin 61	Glu-A1	y	2*	chr1A	509027573	509025997
Norin 61	Glu-B1	x	7+8	chr1B	558708800	558707057
Norin 61	Glu-B1	y	7+8	chr1B	558889881	558888539
Norin 61	Glu-D1	x	2.2+12	chr1D	410249827	410248302
Norin 61	Glu-D1	y	2.2+12	chr1D	410305402	410304148
Spelt	Glu-A1	x	1	chr1A	512591475	512589368
Spelt	Glu-A1	y	1	chr1A	512797732	512796049
Spelt	Glu-B1	x	6.1+22.1	chr1B	552718526	552717008
Spelt	Glu-B1	y	6.1+22.1	chr1B	552882289	552880515
Spelt	Glu-D1	x	2+12	chr1D	410126384	410124867
Spelt	Glu-D1	y	2+12	chr1D	410182430	410180707
Stanley	Glu-A1	x	2*	chr1A	509949749	509947467
Stanley	Glu-A1	y	2*	chr1A	510137232	510135477
Stanley	Glu-B1	x	7+9	chr1B	571416847	571414586
Stanley	Glu-B1	y	7+9	chr1B	571593965	571591754
Stanley	Glu-D1	x	5+10	chr1D	411371897	411369226
Stanley	Glu-D1	y	5+10	chr1D	411426824	411424924
SY Mattis	Glu-A1	x	2*	chr1A	513772223	513770326
SY Mattis	Glu-A1	y	2*	chr1A	513958964	513957324
SY Mattis	Glu-B1	x	7OE+8	chr1B	556900092	556897933
SY Mattis	Glu-B1	x	7OE+8	chrUn	332545430	332544739
SY Mattis	Glu-B1	y	7OE+8	chr1B	557069193	557067525
SY Mattis	Glu-D1	x	2+12	chr1D	403928738	403927042
SY Mattis	Glu-D1	y	2+12	chr1D	403984712	403983201

Table 5.2 – High molecular weight glutenin SDS-PAGE alleles in hexaploid wheat founders.

All of the 10+ Genomes and CIMMYT crossing block populations were typed for SDS-PAGE alleles. Sixty of the CIMMYT founders were typed. None of the WP founders were typed and reported alleles are from variety release publications. Rare alleles and heterozygous calls are not reported.

HMW-GS LOCUS SDS-PAGE ALLELE	10+ Genomes (n)	Western plains hard winter wheat founders (n)	CIMMYT hard spring wheat founders (n)	CIMMYT hard spring wheat crossing block (n)
Glu-A1				
0 (null)	3	0	0	14
1	3	3	14	148
2*	5	6	45	637
Glu-B1				
7	0	0	2	104
7+8	4	3	2	50
7*+8	0	1	0	0
7+9	3	5	33	370
7OE+8	2	0	0	12
17+18	1	3	20	224
13+16	0	0	2	12
Glu-D1				
2+12	7	2	3	45
5+10	3	10	56	760



a) *Glu-A1*

GENOME	80	114	133	157	241	295	305	320	341	442	451	473	474	481	553	562	590	636	654	910	1717	1723	1790	1864	1866	1869	1912	111	353	397	526	590	618	718	966	2224	2394	Haplotype	SDS-PAGE Allele
ArinaLrFor	A	T	G	G	G	T	A	A	G	C	A	G	G	T	C	A	G	G	A	+	A	A	G	C	G	C	A	G	G	+	G	A	G	A	C	C	T	1	null
Chinese Spring	A	T	G	G	G	T	A	A	G	C	A	G	G	T	C	A	G	G	A	+	A	A	G	C	G	C	A	G	G	+	G	A	G	A	C	C	T	1	null
Julius	A	T	G	G	G	T	A	A	G	C	A	G	G	T	C	A	G	G	A	+	A	A	G	C	G	C	A	G	G	+	G	A	G	A	C	C	T	1	null
Jagger	G	C	A	G	T	T	C	C	A	T	T	A	A	G	T	G	T	A	G	-	G	G	A	C	A	G	G	A	A	+	G	G	T	G	T	C	G	2	1
Mace	G	C	A	G	T	T	C	C	A	T	T	A	A	G	T	G	T	A	G	-	G	G	A	C	A	G	G	A	A	+	G	G	T	G	T	C	G	2	1
Spelt	G	C	A	G	T	T	C	C	A	T	T	A	A	G	T	G	T	A	G	-	G	G	A	C	A	G	G	A	A	+	G	G	T	G	T	C	G	2	1
Long Reach Lancer	A	T	G	A	G	C	A	C	G	C	A	G	G	T	C	A	T	G	G	-	G	A	G	T	A	G	G	G	A	-	A	G	G	G	C	T	G	3	2*
CDC Landmark	A	T	G	A	G	C	A	C	G	C	A	G	G	T	C	A	T	G	G	-	G	A	G	T	A	G	G	G	A	-	A	G	G	G	C	T	G	3	2*
SY Mattis	A	T	G	A	G	C	A	C	G	C	A	G	G	T	C	A	T	G	G	-	G	A	G	T	A	G	G	G	A	-	A	G	G	G	C	T	G	3	2*
Norin 61	A	T	G	A	G	C	A	C	G	C	A	G	G	T	C	A	T	G	G	-	G	A	G	T	A	G	G	G	A	-	A	G	G	G	C	T	G	3	2*
CDC Stanley	A	T	G	A	G	C	A	C	G	C	A	G	G	T	C	A	T	G	G	-	G	A	G	T	A	G	G	G	A	-	A	G	G	G	C	T	G	3	2*

b) *Glu-B1*

GENOME	78	231	233	282	552	590	663	681	693	728	1833	1856	1932	1976	2101	2148	189	368	369	375	404	448	459	617	1556	1560	1567	2241	2258	2262	2377	Haplotype	SDS-PAGE Allele
Jagger	T	T	C	G	C	T	A	A	A	A	G	C	A	A	G	G	A	C	C	G	A	A	A	C	A	A	G	-	C	C	A	1	17 + 18
Julius	T	T	C	G	C	T	A	A	A	A	G	C	A	G	G	G	A	C	C	G	A	A	A	C	A	A	G	-	C	C	A	2	7 + 9
CDC Stanley	T	T	C	G	C	T	A	A	A	A	G	C	A	G	G	G	A	C	C	G	A	A	A	C	A	A	G	-	C	C	A	2	7 + 9
CDC Landmark	T	T	C	G	C	T	A	A	A	A	G	C	A	G	G	G	A	C	C	G	A	A	A	C	A	A	G	-	C	C	A	2	7 + 9
ArinaLrFor	T	T	T	G	C	T	A	A	A	A	G	C	A	G	G	G	A	C	C	G	A	A	A	C	A	A	G	-	C	C	A	3	7 + 8
Long Reach Lancer	T	T	T	G	C	T	A	A	A	A	G	C	A	G	G	G	A	C	C	G	A	A	A	C	A	A	G	-	C	C	A	3	7 + 8
Chinese Spring	T	T	T	G	C	T	A	A	A	A	G	C	A	G	G	G	A	C	C	G	A	A	A	C	A	A	G	+	C	C	A	4	7 + 8
Norin 61	T	T	T	G	C	T	A	A	A	A	G	C	A	G	G	G	A	C	C	G	A	A	A	C	A	A	G	+	C	C	A	4	7 + 8
Mace	T	T	C	G	C	T	A	A	A	A	G	C	A	G	G	G	A	C	C	G	A	A	A	C	A	A	G	+	C	C	A	5	7OE + 8
SY Mattis	T	T	C	G	C	T	A	A	A	A	G	C	A	G	G	G	A	C	C	G	A	A	A	C	A	A	G	+	C	C	A	5	7OE + 8
Spelt	C	C	C	A	T	C	G	T	G	G	A	T	G	G	A	A	G	A	G	A	G	G	G	T	G	G	A	+	T	T	G	6	6.1 + 22.1

c) *Glu-D1*

GENOME	60	138	234	388	517	556	588	601	1637	1651	1670	1741	1767	1834	156	353	422	517	519	544	2266	Haplotype	SDS-PAGE Allele
ArinaLrFor	C	C	T	A	C	A	A	+	T	A	T	C	G	A	G	C	+	G	A	C	T	1	2 + 12
Chinese Spring	C	C	T	A	C	A	A	+	T	A	T	C	G	A	G	C	+	G	G	C	T	1	2 + 12
Julius	C	C	T	A	C	A	A	+	T	A	T	C	G	A	G	C	+	G	G	C	T	1	2 + 12
Long Reach Lancer	C	C	T	A	C	A	A	+	T	A	T	C	G	A	G	C	+	G	G	C	T	1	2 + 12
Mace	C	C	T	A	C	A	A	+	T	A	T	C	G	A	G	C	+	G	G	C	T	1	2 + 12
Norin 61	C	C	T	A	C	A	A	+	T	A	T	C	G	A	G	C	+	G	G	C	T	1	2.2 + 12
Spelt	C	C	T	A	C	A	A	+	T	A	T	C	G	A	G	C	+	G	G	C	T	1	2 + 12
SY Mattis	C	C	T	A	C	A	A	+	T	A	T	C	G	A	G	C	+	G	G	C	T	1	2 + 12
CDC Landmark	T	T	C	G	T	C	G	-	C	T	C	G	A	C	A	G	-	A	G	T	C	2	5 + 10
CDC Stanley	T	T	C	G	T	C	G	-	C	T	C	G	A	C	A	G	-	A	G	T	C	2	5 + 10
Jagger	T	T	C	G	T	C	G	-	C	T	C	G	A	C	A	G	-	A	G	T	C	2	5 + 10

Figure 5.1 - Coding sequences haplotypes of high molecular weight glutenins in the 10+ Genomes wheat varieties.

Positions of SNPs (horizontal blue lines) and InDels (downward triangles) detected in the coding sequences of the x and y subunits are presented for each locus. The gray arrows indicate the coding direction of the genes. Nucleotide positions are relative to the coding orientation of each subunit in the curated assembly combining Chinese Spring v1.0 and Triticum 3.1. Figure is to scale.

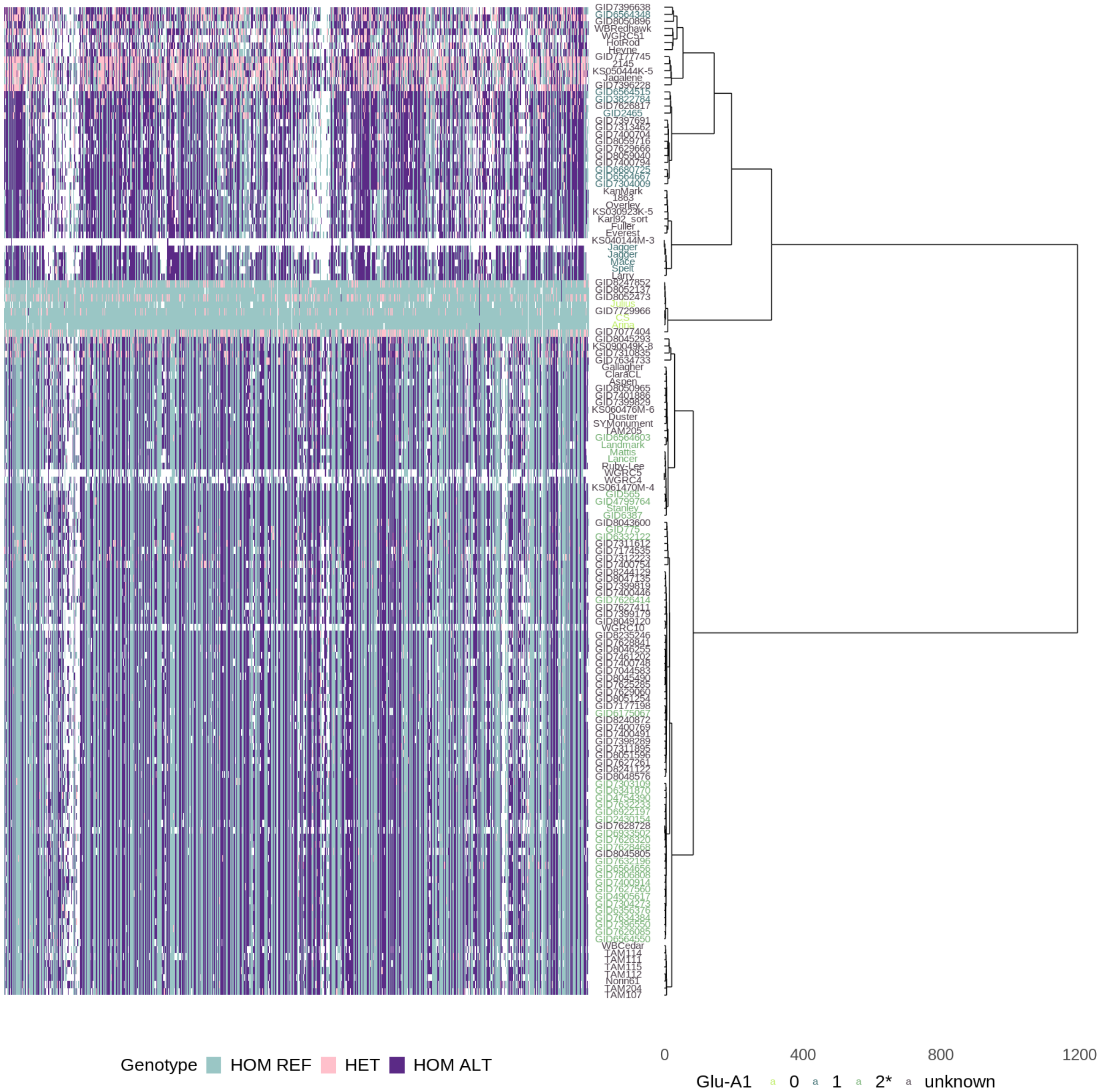


Figure 5.2 - Haplotypes of *Glu-A1* locus in hexaploid wheat founders.

Dendrogram of hexaploid wheat accessions is shown to the right of the molecular haplotypes. Each row is an accession and each column a variant site. The variant states for the of homozygous reference allele are shown in aqua, heterogyzgous in pink and homozygous alternate allele in purple. Wheat accession identifiers are shown as the leaves on the dendrogram and colored by the SDS-PAGE allele is known. 10+ Genomes varieties are Chinese Spring, ArinaLrFor, Jagger, Julius, Long Reach Lancer, CDC Landmark, AUS Mace, SY Mattis, CDC Stanley and Spelt. The CIMMYT varieties are prefaced with 'GID' and all others belong to the USA western plains population.

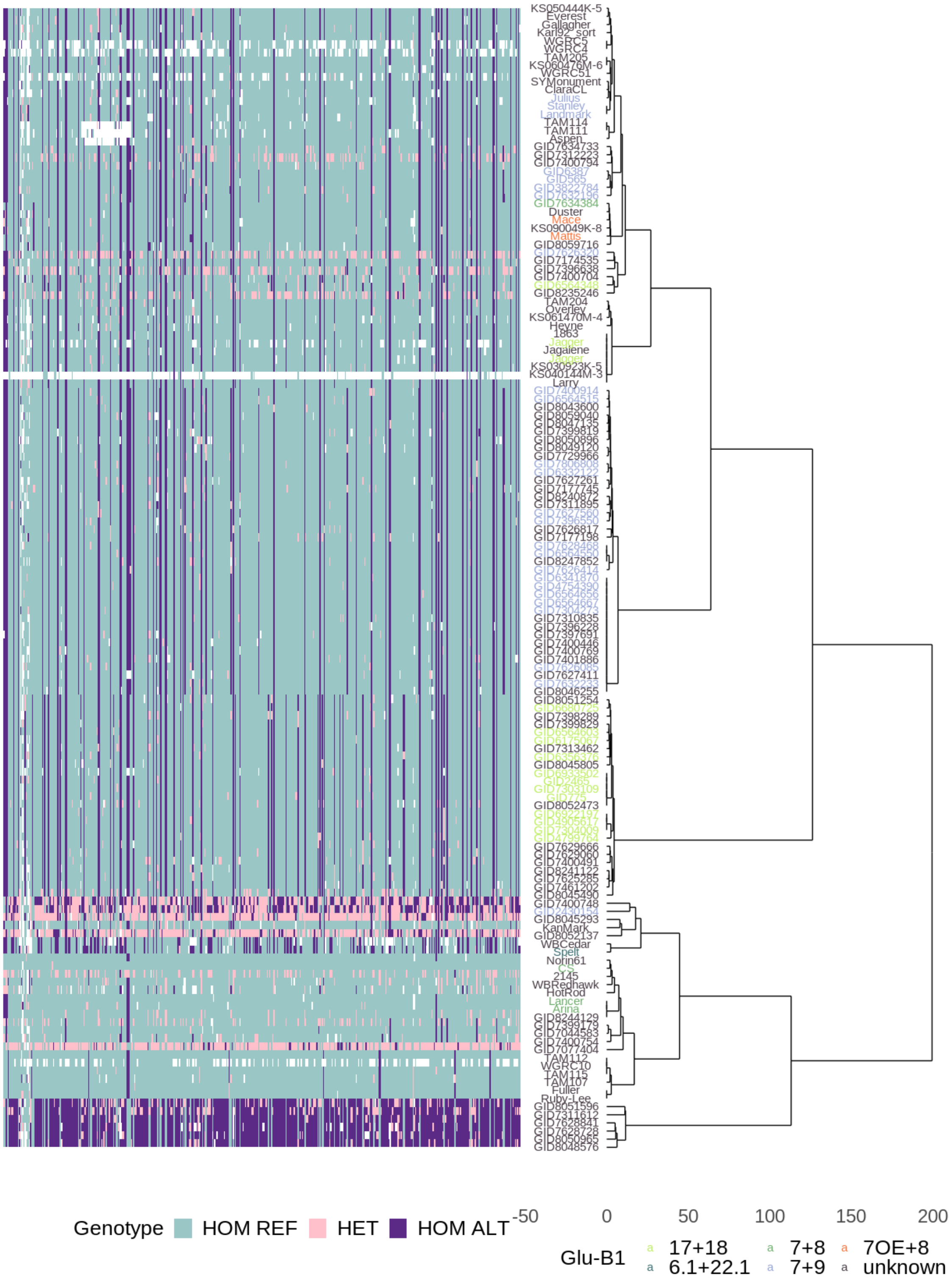


Figure 5.3 – Haplotypes of *Glu-B1* locus in hexaploid wheat founders.

Dendrogram of hexaploid wheat accessions is shown to the right of the molecular haplotypes. Each row is an accession and each column a variant site. The variant states for the of homozygous reference allele are shown in aqua, heterogyzgous in pink and homozygous alternate allele in purple. Wheat accession identifiers are shown as the leaves on the dendrogram and colored by the SDS-PAGE allele is known. 10+ Genomes varieties are Chinese Spring, ArinaLrFor, Jagger, Julius, Long Reach Lancer, CDC Landmark, AUS Mace, SY Mattis, CDC Stanley and Spelt. The CIMMYT varieties are prefaced with 'GID' and all others belong to the USA western plains population.

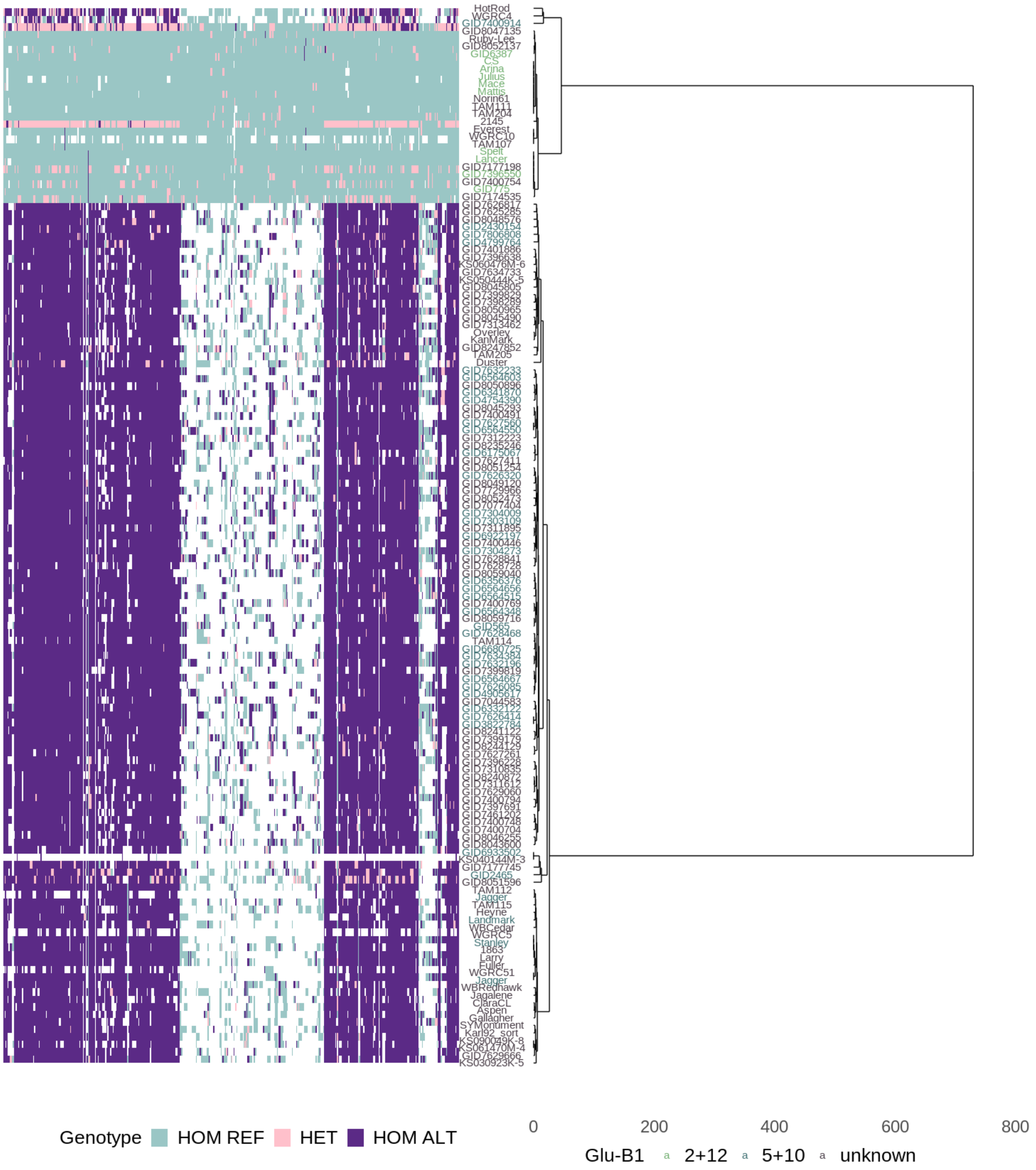
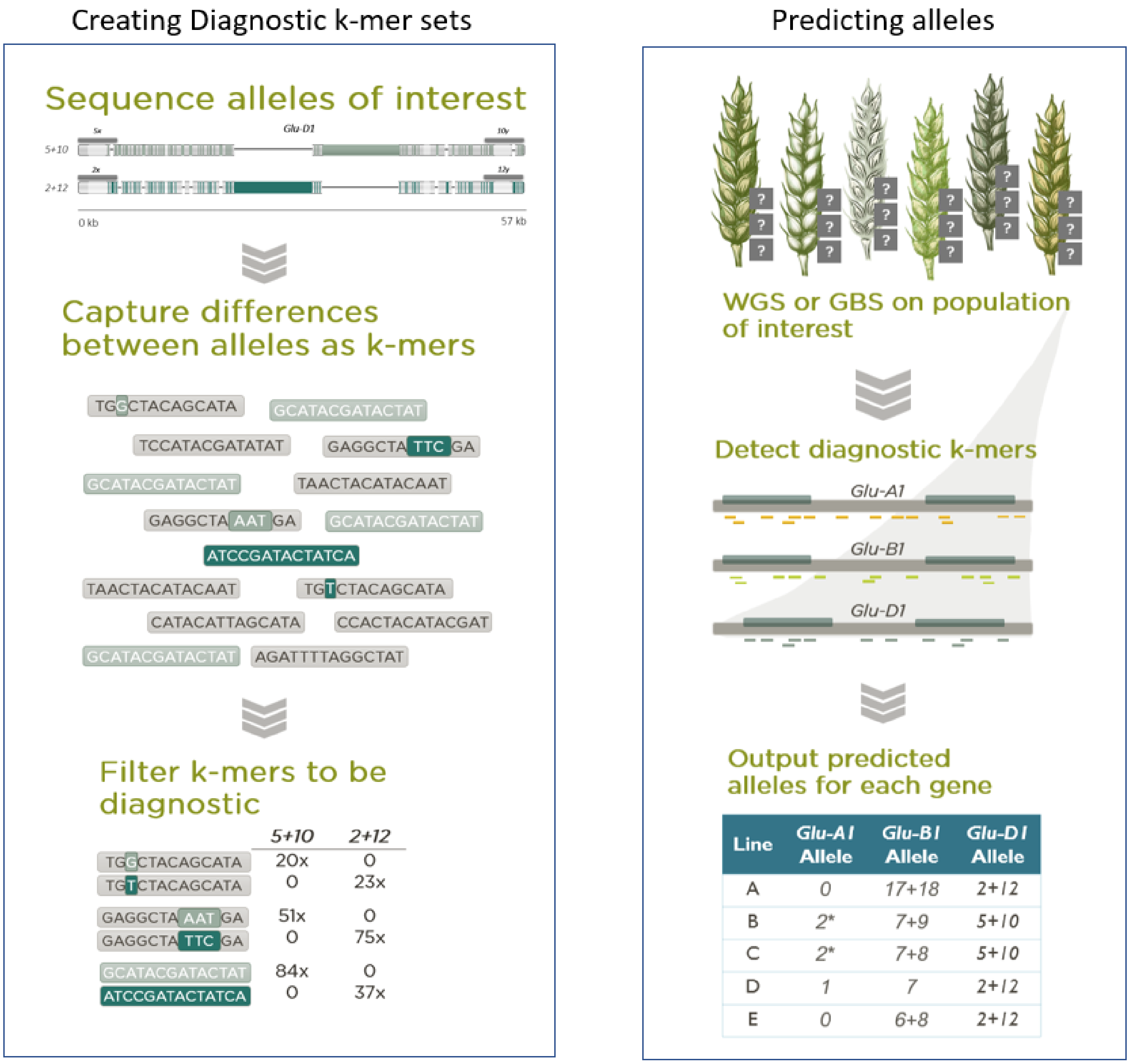


Figure 5.4 - Haplotypes of *Glu-D1* locus in wheat founders.

Dendrogram of hexaploid wheat accessions is shown to the right of the molecular haplotypes. Each row is an accession and each column a variant site. The variant states for the of homozygous reference allele are shown in aqua, heterogyzgous in pink and homozygous alternate allele in purple. Wheat accession identifiers are shown as the leaves on the dendrogram and colored by the SDS-PAGE allele is known. 10+ Genomes varieties are Chinese Spring, ArinaLrFor, Jagger, Julius, Long Reach Lancer, CDC Landmark, AUS Mace, SY Mattis, CDC

Stanley and Spelt. The CIMMYT varieties are prefaced with ‘GID’ and all others belong to the USA western plains population.



	5+10	2+12
TGGCTACAGCATA	20x	0
TGTCTACAGCATA	0	23x
GAGGCTAAATGA	51x	0
GAGGCTATTCGA	0	75x
GCATACGATACTAT	84x	0
ATCCGATACTATCA	0	37x

Line	Glu-A1 Allele	Glu-B1 Allele	Glu-D1 Allele
A	0	17+18	2+12
B	2*	7+9	5+10
C	2*	7+8	5+10
D	1	7	2+12
E	0	6+8	2+12

Figure 5.5 - Generalized sequence-based-genotyping workflow.

In the training segment of SBG, germplasm carrying the alleles are interested are sequenced. The sequences aligning to the genes of interest are extracted and broken into k-mers. K-mers are sorted by allele and filtered to be diagnostic for a given allele. During the prediction segment, skim sequencing is done on the population of interest. Diagnostic k-mers from the training set are detected and the allelic state of each line within the population is predicted based on which k-mers that line carries.

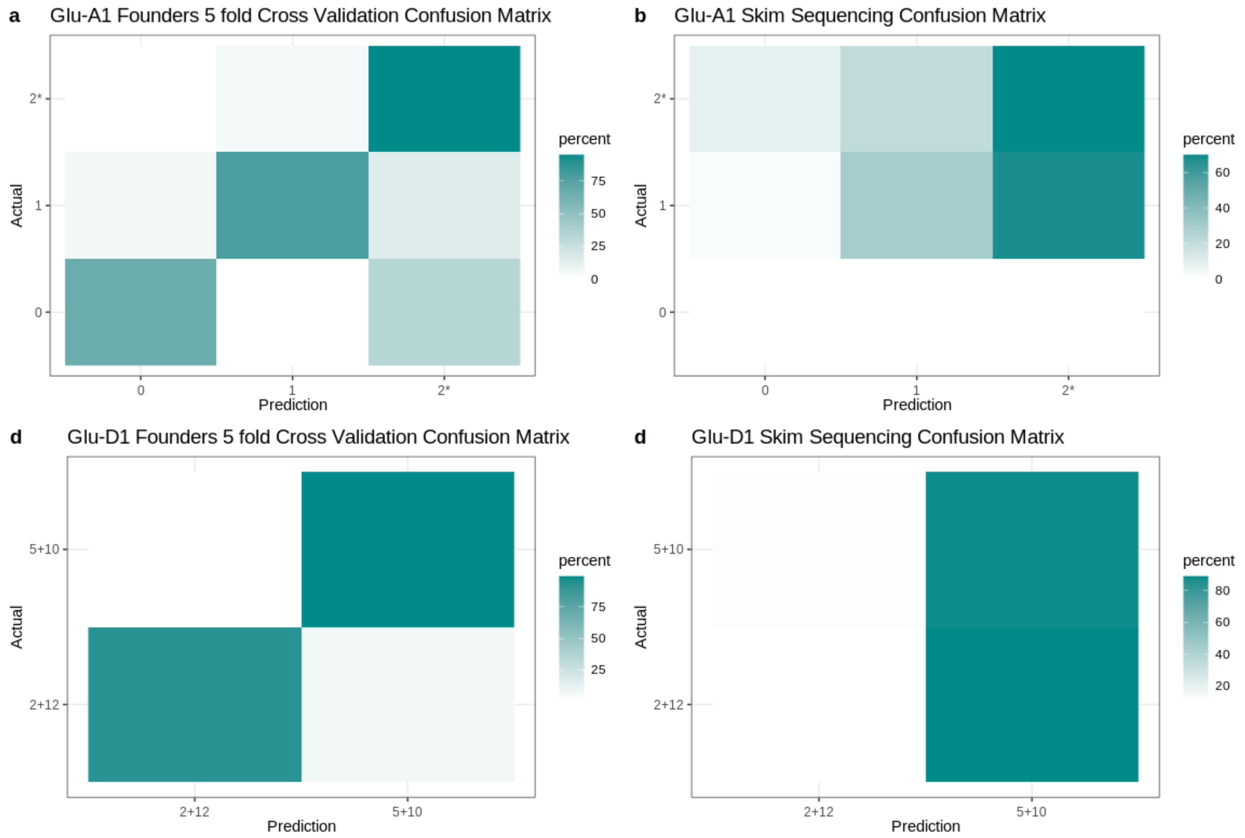


Figure 5.6 - Confusion matrices for sequence based genotyping allele predictions on *Glu-A1* and *Glu-D1*.

Confusion matrices show the prediction accuracy of the sequence-based-genotyping as percentage of wheat accessions correctly or incorrectly predicted. The x-axis shows the SBG prediction of the allele and the y-axis shows the true allele. Darker squares along the diagonal and white squares on the off diagnonal indicate higher prediction accuracy. Dark squares on the off diagonal indicate incorrect predictions. Initial testing was on done 5 fold cross validation within the training set for *Glu-A1* (**a**) and *Glu-D1* (**c**). Further testing was done on the skim sequenced (0.05x coverage) crossing block population for *Glu-A1* (**b**) and *Glu-D1* (**d**).

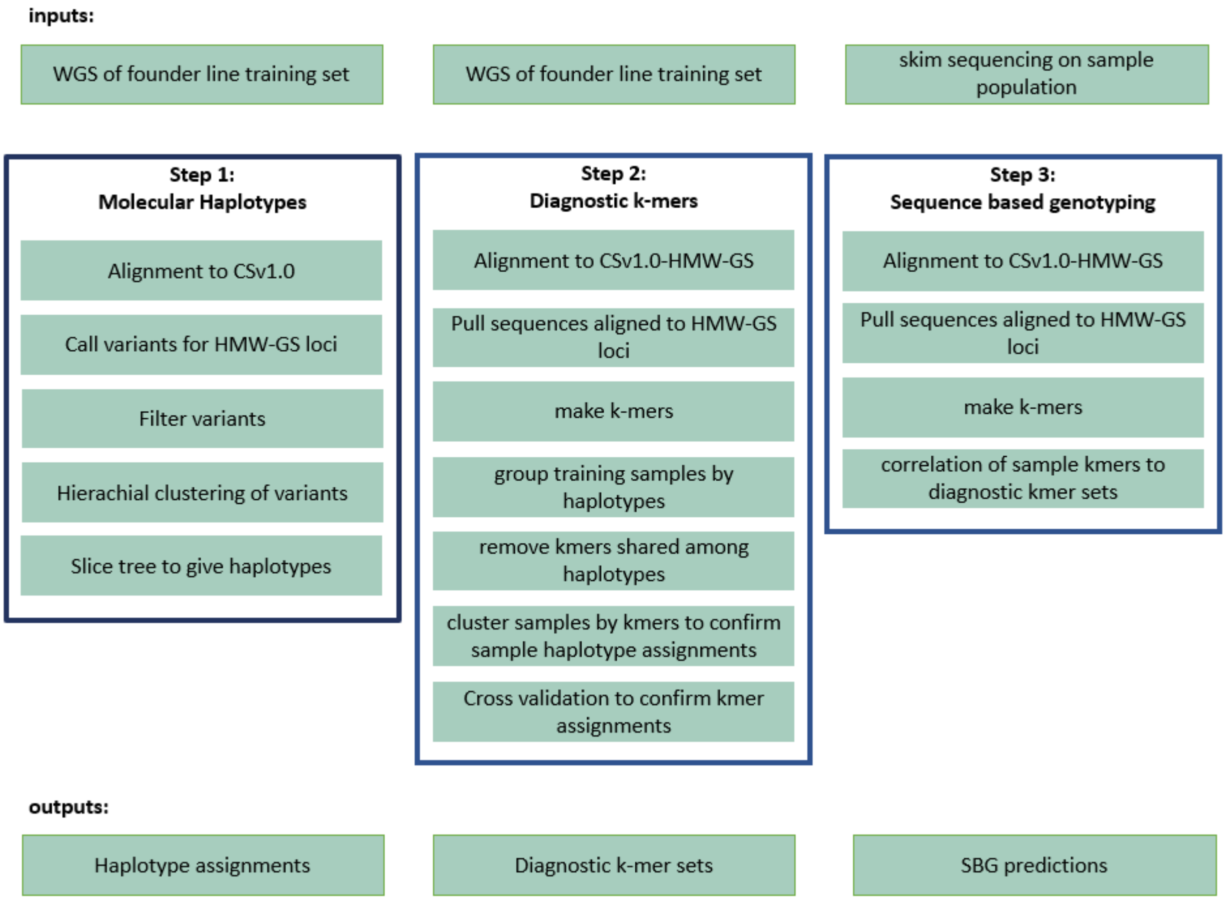

inputs:
WGS of founder line training set
WGS of founder line training set
skim sequencing on sample population
Step 1:
Molecular Haplotypes
Alignment to CSv1.0
Call variants for HMW-GS loci
Filter variants
Hierachial clustering of variants
Slice tree to give haplotypes
Step 2:
Diagnostic k-mers
Alignment to CSv1.0-HMW-GS
Pull sequences aligned to HMW-GS loci
make k-mers
group training samples by haplotypes
remove kmers shared among haplotypes
cluster samples by kmers to confirm sample haplotype assignments
Cross validation to confirm kmer assignments
Step 3:
Sequence based genotyping
Alignment to CSv1.0-HMW-GS
Pull sequences aligned to HMW-GS loci
make k-mers
correlation of sample kmers to diagnostic kmer sets
outputs:
Haplotype assignments
Diagnostic k-mer sets
SBG predictions

Figure 5.7 - Technical sequence-based-genotyping workflow.

Chapter 6 - Discussion of future work

Genomic Selection

Quality is a critical component of wheat variety development. In many breeding programs, acceptable processing and end-use quality parameters have been achieved. As shown in the CIMMYT breeding program, that acceptable quality characteristics are wide and diverse depending on the target region and the wheat products made there (Guzman et al., 2016). The potential of wheat flour for many end-use products, such as soft and weak gluten being ideal for cake flour and hard grain with high gluten strength being ideal for bread flour, also means that acceptable quality traits are relatively easy to attain and a particular variety can be targeted for a particular end-use. For an excellent overview of the quality traits suited for different products, see Guzman et al. (2016).

The goal of the wheat breeding program at CIMMYT is maintenance of acceptable quality within the higher yielding and more resilient wheat varieties. We see very clearly from wealth of quality data collected over past 10 years that quality traits have not moved substantially in any direction (Figure XX). The values of quality traits over the populations change year to year, but these values follow closely the yearly weather trends, especially the average global surface temperature. This is to be expected, given that environment is known to play a role in quality characteristics (XX) and we additionally showed in Chapter 2 of this work that some high molecular weight glutenins interact with environment to influence final quality traits.

In Chapter 3 of this dissertation, we showed that genomic predictions for 13 quality traits have become routine for the phenotypically unobserved 10,000 first year trial wheat lines in the

CIMMYT wheat breeding program. While the predictions are straight forward to attain, how to incorporate 130,000 data points into a coherent index for selection recommendation is complicated. Quality is complex and cannot be distilled into a single value nor do we necessarily aim to drive any of the traits in one direction. In initial years of this work, we naively attempted to predict the quality classes as defined by Guzman et al. (2016) from the predicted alveograph W, alveograph P/L and grain protein values in an attempt to distill quality predictions into a more palatable value for the breeders. It failed spectacularly and was quickly abandoned, no further attempts of creating a selection criterion were made.

What then, would be the best use of the genomic predictions? Given that only of one the quality classes is unacceptable, that the goal of the Wheat Chemistry and Quality Laboratory at the International Maize and Wheat Improvement Center (CIMMYT) is to maintain acceptable quality for a diverse range of products and that replicated yield trial plots are costly in labor and resources, then it appears that the most efficient use of genomic predictions in the first year yield trials would be to use them identify the poorest quality lines to not be selected. One must be careful to remember though that poorest quality does not necessarily mean lowest trait values. Instead, a range, perhaps defined as within two standard deviations from the mean in either direction should be used to flag candidate lines for potential removal. A weighted flag could be applied when any of the more critical traits, such as grain protein, grain hardness, alveograph W, and alveograph P/L have strayed outside the acceptable range. These flags could help breeders quickly decide which of the 10,000 first year yield trial wheat lines to not select for advancement.

Future work would involve developing this flagging system and testing its success. Success could be measured by measuring a correlation between if a line was recommended to be discarded and if that line was discarded. This measure will be complicated by the fact that quality is not the main source for selection, but instead a plethora of other traits such as disease resistance for which genomic predictions are made within CIMMYT itself. Additional work could be done to test if including covariates such as grain protein content or weather data could improve prediction accuracies.

Additionally, moderate and large effect loci such as the high molecular weight glutenin haplotypes, low molecular weight glutenin haplotypes and gliadin haplotypes could be included as covariates in the models. The eventual objective of implementing sequence based genotyping would provide haplotype predictions on the first year yield trials to use as both quality traits themselves for selection and as covariates in the genomic predictions models. An important consideration is that genomic prediction and SBG are very unlikely to identify rare alleles that contribute to unique or highly desirable quality characteristics. Therefore, those alleles are likely to be lost in population unless molecular markers tagging those alleles are used or SBG is specifically trained for the allele. Detecting the existence of these alleles, their effects and their markers will require continued research alongside breeding efforts.

In conclusion, simple genomic prediction models have been shown in this work to provide predictions at breeding program scale. Continued research to refine the models could potentially improve prediction accuracy. In the near future, an index to concisely summarize the quality predictions would likely be very useful for breeders during selection.

High molecular weight glutenin haplotypes

We used dense single nucleotide polymorphism (SNP) markers to determine molecular haplotypes of the high molecular weight glutenin loci in the 10+ Wheat Genome assemblies, CIMMYT hard spring wheat founders and USA western plains hard winter wheat representative varieties. The *Glu-A1* and *Glu-D1* haplotypes corresponded to the SDS-PAGE alleles as expected, but several interesting observations were made regarding *Glu-B1.* The first was that most of the *Glu-B1* haplotypes were much more similar to each other than was seen within the *Glu-A1* or *Glu-D1* loci, indicating that many of the modern alleles are recent mutations after the hybridization events that led to the hexaploid wheat genome. The second was that within SDS-PAGE alleles were more than one cryptic haplotype, and in the case of *Glu-B1 17+18* the haplotype present in the CIMMYT spring material is different than the one present winter wheat material. The final observation was that there were two very different haplotypes, one corresponding to a spelt *Glu-B1* allele and another of yet unknown origin, perhaps another spelt allele, a durum allele or a recently introgressed wild allele.

Further work is required to elucidate the significance of these findings. One potential project is to compare the quality effects of the newly discovered *Glu-B1* molecular haplotypes. Past analyses have used SDS-PAGE to determine quality effects of *Glu-B1* alleles, as seen for example in Chapter 2 of this dissertation, but we now know that multiple alleles exist within the SDS-PAGE profiles. Quality data has already been gathered for the CIMMYT spring material, but further analysis will be required to include the 10+ Genomes and USA western plains winter wheat varieties. Ideally, the wheat lines would be analyzed by the CIMMYT quality lab to ensure standard protocols across all samples.

Another potential project would be to determine the factors underlying quality effect differences between the *Glu-B1* alleles. There are several hypotheses for which factors lead to quality differences. The first is that the number and position of cysteine residues determines how well a protein can interact in the gluten matrix. The second relating to the length and structure of the repetitive domain. A third is that the transcription timing and rate determines the accumulation of aggregated gluten proteins in the kernel. The first hypothesis could begin to be tested by estimating the amino acid composition of the N and C terminal regions of the *Glu-B1* haplotypes from the short read sequencing data available through this study. Substitutions leading to loss or gains of cysteine residues would be relatively straightforward to determine. However, changes in amino acid composition within the central repetitive domains of the glutenin genes would not be possible with the present data and long read sequencing would be required to transverse the region. With long read sequencing, the size and structure of the repetitive domain would also be determined to test hypothesis two. Testing the third hypothesis would require expression studies and possibly accumulation pattern analysis during grain development (S. Wang et al., 2013). Any of these studies combined with the wealth of quality phenotypic data available from Chapter 4 of this dissertation would provide a powerful dataset for quantifying the quality effects attributable to the different alleles and molecular characteristics they possess. A unique factor in using *Glu-B1* alleles for this study is that the alleles appear to be much less diverged than those at either *Glu-A1* or *Glu-D1,* which could allow a more precise determination of the individual molecular characteristics associated with the quality differences.

The *Ae. tauschii Glu-D1* work provided dozens of markers within the coding sequences or very close to the high molecular weight glutenins. These markers can be used to track introgression of novel *Ae. tauschii Glu-D1* alleles in the wheat breeding programs. In particular, in Dr. Allan Fritz's wide cross populations at Kansas State University. The markers will help to prioritize candidates with the novel alleles for advancement and then quality trait characterization. This same haplotype analysis framework can also be applied to other wheat wild relatives with potential for domestic wheat improvement. For example, in the wild emmer (*Triticum turgidum* ssp. *dicoccoides*) by domesticated wheat populations being made by Dr. Mary Guttieri of the USDA in Manhattan, KS. Wild emmer crosses have the potential of introducing novel *Glu-A1* and *Glu-B1* alleles in the wheat genetic pool.

This research provides the framework for molecular haplotype diversity analysis of genes. The framework can be applied to other genes for diversity analysis and also to identify useful variant sites for molecular marker development. The markers identified provide immediate support for tracking novel alleles in introgression populations. Future work will characterize the quality traits imparted by the novel wild wheat alleles.

Low molecular weight glutenin and gliadin haplotypes

Although the focus of much of this dissertation was devoted to high molecular weight glutenins, the importance of low molecular weight glutenins and gliadins shouldn't be overlooked. We chose to apply these initial studies to high molecular weight glutenins because the loci are relatively simple compared to the other gluten loci, the high molecular weight glutenins are the most widely and deeply studied at this time, and that the SDS-PAGE mobilities are simpler to interpret. The next phases for both the hexaploid wheat and *Ae. tauschii* DNA sequence data sets is to apply the haplotype analysis framework to the low molecular weight and gliadin loci. The SDS-PAGE analysis for low molecular weight glutenins is currently underway at the CIMMYT quality lab. The next stage is simply to call and filter variants for the low the molecular weight glutenin and gliadin loci, and to conduct the phylogenetic analysis for molecular haplotypes. Harnessing the quality data available for the CIMMYT founder lines combined with the molecular haplotypes for the high molecular weight glutenin, low molecular weight glutenin and gliadin loci will further refine the quality effects attributable to the alleles at each locus and compare the effects between the loci.

An additional study, which is currently in progress in collaboration with Dr. Brett Carver of Oklahoma State University, is to determine the source of superior quality traits in the OSU released winter wheat variety, Ruby Lee. The high molecular weight glutenin composition (*2**, *7+8* and *2+12*) and pedigree would indicate that it should not have quality traits it does indeed possess. We are particularly interested in this variety because it may harbor unique gluten genes from *Ae. tauschii*. Ruby Lee has two unique *Ae. tauschii* accessions in its pedigree from the Kansas State University Wheat Genetic Resource Center (WGRC), TA2460 and TA2470, from which novel D genome alleles could have been inherited. Through the course of the high molecular weight glutenin haplotype analysis on hexaploid wheat and *Ae. tauschii* in Chapters 4 and 5, we've determined that Ruby Lee does not have a unique *Glu-D1* haplotype from *Ae. tauschii*. Therefore, the next leading hypothesis is that Ruby Lee has a novel *Ae. tauschii* allele at a low molecular weight glutenin or gliadin locus. A haplotype and SDS-PAGE analysis comparing these loci in Ruby Lee, other hexaploid wheat varieties and the tauschii accessions would test this hypothesis. The required whole genome sequencing data and alignments required are available through the Chapter 5 project of this dissertation. The samples are currently being analyzed together for SDS-PAGE mobilities at the CIMMYT quality lab. If this hypothesis turns out to also be false, then a haplotype binning analysis could be done to determine if Ruby Lee has any *Ae. tauschii* introgressions. Additionally, the in depth quality analysis of Ruby Lee and other wheat varieties being conducted in Dr. Brett Carver's group could be leveraged against the whole genome sequencing data currently available to carry out a genome wide association study to identify genomic regions associated with the superior quality of Ruby Lee. If, however, we are able to show that Ruby Lee possesses a novel *Ae. tauschii* low molecular weight glutenin or gliadin allele then this would become a powerful story showcasing the utilization of wild relatives for genetic diversity in modern wheat.

Conclusion

In this dissertation research, we've shown the utility of applying genomic tools to quality improvement in wheat. We showed that superior high molecular weight glutenin alleles often confer greater effects under environmental stress. We applied genomic selection annually to the CIMMYT spring wheat first year yield trials. Predictions for those 10,000 otherwise unobserved entries were made available ahead of breeder selections. We characterized high molecular weight glutenin gene diversity through the wheat pan genome projects, 10+ Wheat Genomes and Open Wild Wheat Consortium. Hundreds of variant sites were identified for molecular marker development. We showed initial results of the k-mer based allele prediction methodology, sequence-based-genotyping. Future work will entail quality effect characterization of the cryptic domestic alleles and the novel wild alleles identified in Chapters 4 and 5, refinement of the genomic prediction and sequence-based-genotyping models, and the application of sequence-based-genotyping to wheat breeding programs. Through this work, we've shown the power of applying genomic tools to breeding programs.

References

Allaby, R. G., Banerjee, M., & Brown, T. A. (1999). Evolution of the high molecular weight glutenin loci of the A, B, D, and G genomes of wheat. *Genome =, 42*(2), 296-307. doi:http://dx.doi.org/10.1139/gen-42-2-296

Altenbach, S. B., Tanaka, C. K., & Seabourn, B. W. (2014). Silencing of omega-5 gliadins in transgenic wheat eliminates a major source of environmental variability and improves dough mixing properties of flour. *BMC Plant Biology, 14*(1), 393. doi:10.1186/s12870-014-0393-1

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol, 215*(3), 403-410. doi:10.1016/s0022-2836(05)80360-2

Anderson, O., Rausch, C., Moullet, O., & Lagudah, E. (2003). The wheat D-genome HMW-glutenin locus: BAC sequencing, gene distribution, and retrotransposon clusters. *Functional & Integrative Genomics, 3*(1), 56-68. doi:10.1007/s10142-002-0069-z

Appels, R., Eversole, K., Stein, N., Feuillet, C., Keller, B., Rogers, J., . . . Wang, L. (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science, 361*(6403), eaar7191. doi:10.1126/science.aar7191

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *2015, 67*(1), 48. doi:10.18637/jss.v067.i01

Battenfield, S. D., Guzmán, C., Gaynor, R. C., Singh, R. P., Peña, R. J., Dreisigacker, S., . . . Poland, J. A. (2016). Genomic Selection for Processing and End-Use Quality Traits in the CIMMYT Spring Bread Wheat Breeding Program. *Plant Genome, 9*(2). doi:10.3835/plantgenome2016.01.0005

Blumenthal, C. S., Bekes, F., Batey, I. L., Wrigley, C. W., Moss, H. J., Mares, D. J., & Barlow, E. W. R. (1991). Interpretation of grain quality results from wheat variety trials with reference to high temperature stress. *Australian Journal of Agricultural Research, 42*(3), 325-334.

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

Branlard, G., Dardevet, M., Amiour, N., & Igrejas, G. (2003). Allelic diversity of HMW and LMW glutenin subunits and omega-gliadins in French bread wheat (Triticum aestivum L.). *Genetic Resources and Crop Evolution, 50*(7), 669-679. doi:10.1023/A:1025077005401

Browning, Brian L., & Browning, Sharon R. (2016). Genotype Imputation with Millions of Reference Samples. *The American Journal of Human Genetics, 98*(1), 116-126. doi:10.1016/j.ajhg.2015.11.020

Buonocore, F., Caporale, C., & Lafiandra, D. (1996). Purification and Characterisation of HighMrGlutenin Subunit 20 and its Linked y-type Subunit from Durum Wheat. *Journal of Cereal Science, 23*(3), 195-201. doi:https://doi.org/10.1006/jcrs.1996.0020

Butow, B. J., Gale, K. R., Ikea, J., Juhász, A., Bedö, Z., Tamás, L., & Gianibelli, M. C. (2004). Dissemination of the highly expressed Bx7 glutenin subunit (Glu-B1al allele) in wheat as revealed by novel PCR markers and RP-HPLC. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik, 109*(7), 1525-1535. doi:10.1007/s00122-004-1776-8

De Santis, M. A., Giuliani, M. M., Giuzio, L., De Vita, P., Lovegrove, A., Shewry, P. R., & 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. doi:https://doi.org/10.1016/j.eja.2017.04.003

Don, C., Lookhart, G., Naeem, H., MacRitchie, F., & Hamer, R. J. (2005). Heat stress and genotype affect the glutenin particles of the glutenin macropolymer-gel fraction. *Journal of Cereal Science, 42*(1), 69-80. doi:https://doi.org/10.1016/j.jcs.2005.01.005

Dong, Z., Yang, Y., Li, Y., Zhang, K., Lou, H., An, X., . . . Wang, D. (2013). Haplotype Variation of Glu-D1 Locus and the Origin of Glu-D1d Allele Conferring Superior End-Use Qualities in Common Wheat. *PLoS ONE, 8*(9), e74859. doi:10.1371/journal.pone.0074859

Dubcovsky, J., & Dvorak, J. (2007). Genome plasticity a key factor in the success of polyploid wheat under domestication. *Science (New York, N.Y.), 316*(5833), 1862-1866. doi:10.1126/science.1143986

Edwards, J., Hunger, R., Smith, E., Horn, G., Chen, M.-S., Yan, L., . . . Carver, B. (2012). ‘Duster’ Wheat: A Durable, Dual-Purpose Cultivar Adapted to the Southern Great Plains of the USA. *Journal of Plant Registrations, 6,* 37. doi:10.3198/jpr2011.04.0195crc

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

Galili, T. (2015). dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics, 31*(22), 3718-3720. doi:10.1093/bioinformatics/btv428

Gao, L. (2020). Open Wild Wheat Consortium Variant (SNP) dataset [Data set]. . Zenodo.

Gao, X., Appelbee, M. J., Mekuria, G. T., Chalmers, K. J., & Mather, D. E. (2012). A second 'overexpression' allele at the Glu-B1 high-molecular-weight glutenin locus of wheat: sequence characterisation and functional effects. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik, 124*(2), 333-343. doi:10.1007/s00122-011-1708-3

Gao, X., Zhang, Q., Newberry, M., Chalmers, K., & Mather, D. (2012). A cysteine in the repetitive domain of a high-molecular-weight glutenin subunit interferes with the mixing properties of wheat dough. *Amino Acids*. doi:10.1007/s00726-012-1441-5

Gao, X., Zhang, Q., Newberry, M. P., Chalmers, K. J., & Mather, D. E. (2013). A cysteine in the repetitive domain of a high-molecular-weight glutenin subunit interferes with the mixing properties of wheat dough. *Amino Acids, 44*(3), 1061-1071. doi:10.1007/s00726-012-1441-5

Gaurav, K., Arora, S., Silva, P., Sánchez-Martín, J., Horsnell, R., Gao, L., . . . Wulff, B. B. H. (2021). Evolution of the bread wheat D-subgenome and enriching it with diversity from Aegilops tauschii. *bioRxiv,* 2021.2001.2031.428788. doi:10.1101/2021.01.31.428788

Gianibelli, M. C., Gupta, R. B., Lafiandra, D., Margiotta, B., & MacRitchie, F. (2001). Polymorphism of High MrGlutenin Subunits in Triticum tauschii: Characterisation by Chromatography and Electrophoretic Methods. *Journal of Cereal Science, 33*(1), 39-52. doi:https://doi.org/10.1006/jcrs.2000.0328

Giuliani, M. M., Palermo, C., De Santis, M. A., Mentana, A., Pompa, M., Giuzio, L., . . . Flagella, Z. (2015). Differential Expression of Durum Wheat Gluten Proteome under Water Stress during Grain Filling. *Journal of Agricultural and Food Chemistry, 63*(29), 6501-6512. doi:10.1021/acs.jafc.5b01635

Gu, Y. Q., Salse, J., Coleman-Derr, D., Dupin, A., Crossman, C., Lazo, G. R., . . . Chalhoub, B. (2006). Types and rates of sequence evolution at the high-molecular-weight glutenin locus in hexaploid wheat and its ancestral genomes. *Genetics, 174*(3), 1493-1504. doi:10.1534/genetics.106.060756

Guzmán, C., Autrique, J. E., Mondal, S., Singh, R. P., Govindan, V., Morales-Dorantes, A., . . . Peña, R. J. (2016). Response to drought and heat stress on wheat quality, with special emphasis on bread-making quality, in durum wheat. *Field Crops Research, 186*, 157-165. doi:https://doi.org/10.1016/j.fcr.2015.12.002

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

Hernández-Espinosa, N., Mondal, S., Autrique, E., Gonzalez-Santoyo, H., Crossa, J., Huerta-Espino, J., . . . Guzmán, C. (2018). Milling, processing and end-use quality traits of CIMMYT spring bread wheat germplasm under drought and heat stress. *Field Crops Research, 215*, 104-112. doi:https://doi.org/10.1016/j.fcr.2017.10.003

Hurkman, W. J., Tanaka, C. K., Vensel, W. H., Thilmony, R., & Altenbach, S. B. (2013). Comparative proteomic analysis of the effect of temperature and fertilizer on gliadin and glutenin accumulation in the developing endosperm and flour from Triticum aestivum L. cv. Butte 86. *Proteome Science, 11*(1), 8. doi:10.1186/1477-5956-11-8

Irmak, S., Naeem, H. A., Lookhart, G. L., & MacRitchie, F. (2008). Effect of heat stress on wheat proteins during kernel development in wheat near-isogenic lines differing at Glu-D1. *Journal of Cereal Science, 48*(2), 513-516. doi:https://doi.org/10.1016/j.jcs.2007.12.002

Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology, 37*(8), 907-915. doi:10.1038/s41587-019-0201-4

Koga, S., Böcker, U., Wieser, H., Koehler, P., Uhlen, A., & Moldestad, A. (2016). Polymerisation of gluten proteins in developing wheat grain as affected by desiccation. *Journal of Cereal Science, 73*. doi:10.1016/j.jcs.2016.12.003

Kong, X. Y., Gu, Y. Q., You, F. M., Dubcovsky, J., & Anderson, O. D. (2004). Dynamics of the evolution of orthologous and paralogous portions of a complex locus region in two genomes of allopolyploid wheat. *Plant Mol Biol, 54*(1), 55-69. doi:10.1023/B:PLAN.0000028768.21587.dc

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., & Salzberg, S. L. (2004). Versatile and open software for comparing large genomes. *Genome Biology, 5*(2), R12. doi:10.1186/gb-2004-5-2-r12

Lagudah, E. S., & Halloran, G. M. (1988). Phylogenetic relationships of Triticum tauschii the D genome donor to hexaploid wheat. *Theoretical and Applied Genetics, 75*(4), 592-598. doi:10.1007/BF00289125

Lawrence, G. J., MacRitchie, F., & Wrigley, C. W. (1988). Dough and baking quality of wheat lines deficient in glutenin subunits controlled by the Glu-A1, Glu-B1 and Glu-D1 loci. *Journal of Cereal Science, 7*(2), 109-112. doi:https://doi.org/10.1016/S0733-5210(88)80012-2

Lawrence, G. J., & Payne, P. I. (1983). Detection by Gel Electrophoresis of Oligomers Formed by the Association of High-Molecular-Weight Glutenin Protein Subunits of Wheat Endosperm. *Journal of Experimental Botany, 34*(3), 254-267. doi:10.1093/jxb/34.3.254

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics, 27*(21), 2987-2993. doi:10.1093/bioinformatics/btr509

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics (Oxford, England), 27*(21), 2987-2993. doi:10.1093/bioinformatics/btr509

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England), 25*(16), 2078-2079. doi:10.1093/bioinformatics/btp352

Li, S., Liu, Y., Tong, J., Yu, L., Ding, M., Zhang, Z., . . . 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. doi:https://doi.org/10.1016/j.foodres.2019.108914

Li, Y., Wu, Y., Hernandez-Espinosa, N., & Peña, R. J. (2013). The influence of drought and heat stress on the expression of end-use quality parameters of common wheat. *Journal of Cereal Science, 57*(1), 73-78. doi:https://doi.org/10.1016/j.jcs.2012.09.014

Lindsay, M. P., & Skerritt, J. H. (1998). Examination of the Structure of the Glutenin Macropolymer in Wheat Flour and Doughs by Stepwise Reduction. *Journal of Agricultural and Food Chemistry, 46*(9), 3447-3457. doi:10.1021/jf980315m

Liu, S., Chao, S., & Anderson, J. A. (2008). New DNA markers for high molecular weight glutenin subunits in wheat. *Theor Appl Genet, 118*(1), 177-183. doi:10.1007/s00122-008-0886-0

Liu, T., Gao, X., Li, L., Du, D., Cheng, X., Zhao, Y., . . . Li, X. (2016). Effects of HMW-GS at Glu-B1 locus on the polymerization of glutenin during grain development and on the secondary and micro-structures of gluten in wheat (Triticum aestivum L.). *Journal of Cereal Science, 72*, 101-107. doi:https://doi.org/10.1016/j.jcs.2016.10.007

Luo, M.-C., Gu, Y. Q., Puiu, D., Wang, H., Twardziok, S. O., Deal, K. R., . . . Dvořák, J. (2017). Genome sequence of the progenitor of the wheat D genome Aegilops tauschii. *Nature, 551*(7681), 498-502. doi:10.1038/nature24486

Lutz, E., Wieser, H., & Koehler, P. (2012). Identification of disulfide bonds in wheat gluten proteins by means of mass spectrometry/electron transfer dissociation. *Journal of Agricultural and Food Chemistry, 60*(14), 3708-3716. doi:10.1021/jf204973u

Mackie, A. M., Lagudah, E. S., Sharp, P. J., & Lafiandra, D. (1996). Molecular and Biochemical Characterisation of HMW Glutenin Subunits fromT. tauschiiand the D Genome of Hexaploid Wheat. *Journal of Cereal Science, 23*(3), 213-225. doi:https://doi.org/10.1006/jcrs.1996.0022

Madeira, F., Park, Y. M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., . . . Lopez, R. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic acids research, 47*(W1), W636-W641. doi:10.1093/nar/gkz268

Maphosa, L., Langridge, P., Taylor, H., Emebiri, L. C., & Mather, D. E. (2015). Genetic control of grain protein, dough rheology traits and loaf traits in a bread wheat population grown in three environments. *Journal of Cereal Science, 64*, 147-152. doi:https://doi.org/10.1016/j.jcs.2015.05.010

Marburger, D. A., Silva, A. d. O., Hunger, R. M., Edwards, J. T., Van der Laan, L., Blakey, A. M., . . . Carver, B. F. (2021). 'Gallagher' and 'Iba' hard red winter wheat: Half-sibs inseparable by yield gain, separable by producer preference. *Journal of Plant Registrations, 15*(1), 177-195. doi:https://doi.org/10.1002/plr2.20116

Moonen, J. H. E., Scheepstra, A., & Graveland, A. (1982). Use of the SDS-sedimentation test and SDS-polyacrylamidegel electrophoresis for screening breeder's samples of wheat for bread-making quality. *Euphytica, 31*(3), 677-690. doi:10.1007/BF00039206

Naeem, H. A., & MacRitchie, F. (2005). Polymerization of glutenin during grain development in near-isogenic wheat lines differing at Glu-D1 and Glu-B1 in greenhouse and field. *Journal of Cereal Science, 41*(1), 7-12. doi:https://doi.org/10.1016/j.jcs.2004.04.009

Payne, P. I., Corfield, K. G., & Blackman, J. A. (1979). Identification of a high-molecular-weight subunit of glutenin whose presence correlates with bread-making quality in wheats of related pedigree. *Theor Appl Genet*, *55*(3-4), 153-159. doi:10.1007/bf00295442

Payne, P. I., Corfield, K. G., Holt, L. M., & Blackman, J. A. (1981). Correlations between the inheritance of certain high-molecular weight subunits of glutenin and bread-making quality in progenies of six crosses of bread wheat. *Journal of the Science of Food and Agriculture, 32*(1), 51-60. doi:10.1002/jsfa.2740320109

Payne, P. I., Holt, L. M., & Law, C. N. (1981). Structural and genetical 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*(4), 229-236. doi:10.1007/bf02342544

Payne, P. I., Holt, L. M., & Lawrence, G. J. (1983). Detection of a novel high molecular weight subunit of glutenin in some Japanese hexaploid wheats. *Journal of Cereal Science*, *1*(1), 3-8. doi:https://doi.org/10.1016/S0733-5210(83)80003-4

Payne, P. I., & Lawrence, G. J. (1983). Catalogue 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, 11*(1), 29-35.

Pena, R. J., Amaya, A., Rajaram, S., & Mujeeb-Kazi, A. (1990). Variation in quality characteristics associated with some spring 1B/1R translocation wheats. *Journal of Cereal Science, 12*(2), 105-112. doi:https://doi.org/10.1016/S0733-5210(09)80092-1

Poland, J. A., Brown, P. J., Sorrells, M. E., & Jannink, J.-L. (2012). Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach. *PLoS ONE, 7*(2), e32253. doi:10.1371/journal.pone.0032253

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics, 26*(6), 841-842. doi:10.1093/bioinformatics/btq033

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Ravel, C., Faye, A., Ben-Sadoun, S., Ranoux, M., Dardevet, M., Dupuits, C., . . . Branlard, G. (2020). SNP markers for early identification of high molecular weight glutenin subunits (HMW-GSs) in bread wheat. *Theoretical and Applied Genetics, 133*(3), 751-770. doi:10.1007/s00122-019-03505-y

Reynolds, M. P. (2010). *Climate Change and Crop Production*: CABI.

Salamini, F., Özkan, H., Brandolini, A., Schäfer-Pregl, R., & Martin, W. (2002). Genetics and geography of wild cereal domestication in the near east. *Nature Reviews Genetics, 3*(6), 429-441. doi:10.1038/nrg817

Shewry, P. (2019). What Is Gluten—Why Is It Special? *Frontiers in Nutrition, 6*(101). doi:10.3389/fnut.2019.00101

Shewry, P. R., & Halford, N. G. (2002). Cereal seed storage proteins: structures, properties and role in grain utilization. *Journal of Experimental Botany, 53*(370), 947-958. doi:10.1093/jexbot/53.370.947

Shewry, P. R., Halford, N. G., Belton, P. S., & 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. doi:10.1098/rstb.2001.1024

Shewry, P. R., Halford, N. G., & Lafiandra, D. (2003). Genetics of wheat gluten proteins. *Adv Genet, 49*, 111-184. doi:10.1016/s0065-2660(03)01003-4

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

Shimizu, K. K., Copetti, D., Okada, M., Wicker, T., Tameshige, T., Hatakeyama, M., . . . Handa, H. (2020). De Novo Genome Assembly of the Japanese Wheat Cultivar Norin 61 Highlights Functional Variation in Flowering Time and Fusarium Resistance Genes in East Asian Genotypes. *Plant and Cell Physiology*. doi:10.1093/pcp/pcaa152

Singh, N., Wu, S., Tiwari, V., Sehgal, S., Raupp, J., Wilson, D., . . . Poland, J. (2019). Genomic Analysis Confirms Population Structure and Identifies Inter-Lineage Hybrids in Aegilops tauschii. *Frontiers in Plant Science, 10*(9). doi:10.3389/fpls.2019.00009

Singh, N. K., Shepherd, K. W., & Cornish, G. B. (1991). A simplified SDS-PAGE procedure for separating LMW subunits of glutenin. *Journal of Cereal Science, 14*(3), 203-208. doi:https://doi.org/10.1016/S0733-5210(09)80039-8

Sissons, M., J., Ames, N., P., Hare, R., A., & Clarke, J., M. (2005). Relationship between glutenin subunit composition and gluten strength measurements in durum wheat. *Journal of the Science of Food and Agriculture, 85*(14), 2445-2452. doi:10.1002/jsfa.2272

Southan, M., & MacRitchie, F. (1999). Molecular Weight Distribution of Wheat Proteins. *Cereal Chemistry, 76*(6), 827-836. doi:10.1094/CCHEM.1999.76.6.827

Walkowiak, S., Gao, L., Monat, C., Haberer, G., Kassa, M. T., Brinton, J., . . . Pozniak, C. J. (2020). Multiple wheat genomes reveal global variation in modern breeding. *Nature, 588*(7837), 277-283. doi:10.1038/s41586-020-2961-x

Wan, Y., Gritsch, C. S., Hawkesford, M. J., & Shewry, P. R. (2014). Effects of nitrogen nutrition on the synthesis and deposition of the ω-gliadins of wheat. *Annals of Botany, 113*(4), 607-615. doi:10.1093/aob/mct291

Wan, Y., Yan, Z., Liu, K., Zheng, Y., D'Ovidio, R., Shewry, P. R., . . . Wang, D. (2005). Comparative analysis of the D genome-encoded high-molecular weight subunits of glutenin. *Theoretical and Applied Genetics, 111*(6), 1183-1190. doi:10.1007/s00122-005-0051-y

Wang, J., Luo, M. C., Chen, Z., You, F. M., Wei, Y., Zheng, Y., & Dvorak, J. (2013). Aegilops tauschii single nucleotide polymorphisms shed light on the origins of wheat D-genome genetic diversity and pinpoint the geographic origin of hexaploid wheat. *New Phytol, 198*(3), 925-937. doi:10.1111/nph.12164

Wang, S., Yu, Z., Cao, M., Shen, X., Li, N., Li, X., . . . Yan, Y. (2013). Molecular Mechanisms of HMW Glutenin Subunits from 1Sl Genome of Aegilops longissima Positively Affecting Wheat Breadmaking Quality. *PLoS ONE, 8*(4), e58947. doi:10.1371/journal.pone.0058947

Wang, Z., Li, Y., Yang, Y., Liu, X., Qin, H., Dong, Z., . . . Wang, D. (2017). New insight into the function of wheat glutenin proteins as investigated with two series of genetic mutants. *Scientific Reports*, *7*(1), 3428. doi:10.1038/s41598-017-03393-6

Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M., & Barton, G. J. (2009). Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics, 25*(9), 1189-1191. doi:10.1093/bioinformatics/btp033

Wickham, H. (2009) *ggplot2: elegant graphics for data analysis*. Springer New York.

William, M. D. H. M., Peña, R. J., & Mujeeb-Kazi, A. (1993). Seed protein and isozyme variations in Triticum tauschii (Aegilops squarrosa). *Theoretical and Applied Genetics, 87*(1), 257-263. doi:10.1007/BF00223774

Wrigley, C., Asenstorfer, R., Batey, I., Cornish, G., Day, L., Mares, D., & Mrva, K. (2009). The Biochemical and Molecular Basis of Wheat Quality *Wheat Science and Trade* (pp. 495-520).

Xu, S., Khan, K., Klindworth, D., & Nygard, G. (2010). Evaluation and characterization of high-molecular weight 1D glutenin subunits from Aegilops tauschii in synthetic hexaploid wheats. *Journal of Cereal Science, 52*. doi:10.1016/j.jcs.2010.05.004

Zhou, Y., Zhao, X., Li, Y., Xu, J., Bi, A., Kang, L., . . . Lu, F. (2020). Triticum population sequencing provides insights into wheat adaptation. *Nature Genetics, 52*(12), 1412-1422. doi:10.1038/s41588-020-00722-w

Zimin, A. V., Puiu, D., Hall, R., Kingan, S., Clavijo, B. J., & Salzberg, S. L. (2017). The first near-complete assembly of the hexaploid bread wheat genome, Triticum aestivum. *GigaScience, 6*(11). doi:10.1093/gigascience/gix097

Zimin, A. V., Puiu, D., Hall, R., Kingan, S., Clavijo, B. J., & Salzberg, S. L. (2017). The first near-complete assembly of the hexaploid bread wheat genome, Triticum aestivum. *GigaScience, 6*(11), 1-7. doi:10.1093/gigascience/gix097

1.