GENOMIC SELECTION AND ASSOCIATION MAPPING FOR WHEAT PROCESSING AND END-USE QUALITY

by

SARAH DENISE BATTENFIELD

M. S., Oklahoma State University, 2011

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Genetics
College of Agriculture

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2015

Abstract

Globally, wheat (*Triticum aestivum* L.) is the second most widely grown cereal grain and is primarily used as a food crop. To meet the demands for human consumption, cultivars must possess suitable end-use quality for release and acceptability. However, breeding for quality traits is often considered a secondary goal, largely due to amount of seed needed and overall expense of such testing. Without testing and selection, many undesirable materials tend to be advanced.

Here we demonstrate two methods, mega-genome-wide association mapping and genomic selection, to enhance selection accuracy for quality traits in the CIMMYT bread wheat breeding program. The methods were developed using high-density SNPs detected from genotyping-by-sequencing and processing and end-use quality evaluations from unbalanced yield trial entries (n = 4,095) during 2009 to 2014, at Ciudad Obregon, Sonora, Mexico.

Genome-wide association mapping, with covariates for population structure and kinship, was applied for each trait to each site-year individually and results were combined across years in a mega-analysis using an inverse variance, fixed effect model in JMP-Genomics. This method presents a new way to detect genes of interest within a breeding program and develop markers for selection of these traits, which can then be used in earlier generations.

Genomic selection prediction models were developed using ridge regression, Gaussian kernel, partial least squares, elastic net, and random forest models in R. With these predictions genomic selection (GS) can be applied at earlier stages and undesirable materials culled before implementing expensive yield and quality screenings. In general, prediction accuracy increased over time as more data was available to train the model. Based on these prediction accuracies, we conclude that genomic selection can be a useful tool to facilitate earlier generation selection for end-use quality in CIMMYT bread wheat breeding.

Genomic selection was conducted for processing and end-use quality traits in the Kansas hard red winter wheat breeding unit. Genomic predictions demonstrate increases in accuracy with added data over time. These data demonstrate that current genomic selection models will need more data to continue improvement in prediction accuracy.

GENOMIC SELECTION AND ASSOCIATION MAPPING FOR WHEAT PROCESSING AND END-USE QUALITY

by

SARAH DENISE BATTENFIELD

M. S., Oklahoma State University, 2011

A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Genetics
College of Agriculture

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2015

Approved by:

Major Professor
Dr. Allan K. Fritz

Copyright

Abstract

Globally, wheat (*Triticum aestivum* L.) is the second most widely grown cereal grain and is primarily used as a food crop. To meet the demands for human consumption, cultivars must possess suitable end-use quality for release and acceptability. However, breeding for quality traits is often considered a secondary goal, largely due to amount of seed needed and overall expense of such testing. Without testing and selection, many undesirable materials tend to be advanced.

Here we demonstrate two methods, mega-genome-wide association mapping and genomic selection, to enhance selection accuracy for quality traits in the CIMMYT bread wheat breeding program. The methods were developed using high-density SNPs detected from genotyping-by-sequencing and processing and end-use quality evaluations from unbalanced yield trial entries (n = 4,095) during 2009 to 2014, at Ciudad Obregon, Sonora, Mexico.

Genome-wide association mapping, with covariates for population structure and kinship, was applied for each trait to each site-year individually and results were combined across years in a mega-analysis using an inverse variance, fixed effect model in JMP-Genomics. This method presents a new way to detect genes of interest within a breeding program and develop markers for selection of these traits, which can then be used in earlier generations.

Genomic selection prediction models were developed using ridge regression, Gaussian kernel, partial least squares, elastic net, and random forest models in R. With these predictions genomic selection (GS) can be applied at earlier stages and undesirable materials culled before implementing expensive yield and quality screenings. In general, prediction accuracy increased over time as more data was available to train the model. Based on these prediction accuracies, we conclude that genomic selection can be a useful tool to facilitate earlier generation selection for end-use quality in CIMMYT bread wheat breeding.

Genomic selection was conducted for processing and end-use quality traits in the Kansas hard red winter wheat breeding unit. Genomic predictions demonstrate increases in accuracy with added data over time. These data demonstrate that current genomic selection models will need more data to continue improvement in prediction accuracy.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank Monsanto Beachell-Borlaug International Scholars Program for funding my PhD. This generous fellowship included research, tuition, and travel funds. Funding for the CIMMYT projects was also provided by US Agency for International Development Feed the Future Initiative (USAID Cooperative Agreement No. AID-OAA-A-13-0005) and the Bill & Melinda Gates Foundation through a grant to Cornell University for "Genomic Selection: The next frontier for rapid gains in maize and wheat improvement." Support for phenotyping of quality traits was provided by CGIAR CRP WHEAT, Durable Rust Resistance Project, and Fondo Sectorial SAGARPA-CONACYT (No. 146788 – "Sistema de mejoramiento genético para generar variedades resistentes a royas, de alto rendimiento y alta calidad para una producción sustentable en México de trigo") of the Mexican government. The computing for this project was performed on the Beocat Research Cluster at Kansas State University, which is funded in part by NSF grants CNS-1006860, EPS-1006860, and EPS-0919443. This work represents contribution number 16-037-J from the Kansas Agricultural Experiment Station. Sponsorship for the Kansas State University wheat breeding program, quality testing and genotyping was provided by the Kansas Wheat Alliance.

I would like to thank my PhD advisors Drs. Allan Fritz, Jesse Poland, Rebecca Miller, and Vara Prasad for their guidance and assistance throughout the program. Also, I would like to thank my CIMMYT collaborating advisors Drs. Ravi Singh, Carlos Guzman, and Javier Peña for their help with research, field experience, and international travel.

I would like to thank the colleagues who provided assistance and commentary on this project. CIMMYT and Kansas State University wheat quality labs provided phenotype data for these analyses, specifically C. Guzman and J. Peña, CIMMYT, and R. Miller, KSU and all of their staff. S. Wu, KSU, and S. Driesigacker, CIMMYT, and C. Gaynor, KSU, collected and prepared DNA for these studies. J. Crossa, CIMMYT, L. Silva, SAS, K. Miclaus, SAS, and R. Wolfinger, SAS, consulted and assisted with statistical analysis. E. Jackson and J. Sheridan were instrumental in the association mapping project and making connections to SAS. Additionally, I would like to thank my lab mates in the Fritz breeding program for assistance, support, critique and discussion of these projects.

Finally, I would like to thank my family for their support through this PhD project. My husband, Kyle, especially was very supportive through all of my projects, deadlines, and international travel. I would also like to thank my father, Wes, and gradmother, Arena, for introducing me to wheat farming.

Chapter 1 - Literature review

The human population is growing exponentially with current projections predicting a population of greater than 9 billion by the year 2050 (Gerland, *et al.*, 2014). Currently bread wheat (*Triticum aestivum* L.) per capita consumption is 65 kg per year, supplying nearly 16 g of protein daily on average for each person in the world (Faostat, 2013). This consumption pattern is increasing fastest in the world's least developed countries, which were predicted to have the largest population increase over the next century (Faostat, 2013, Gerland, *et al.*, 2014). An intersection of improved agronomic practices and improved crop varieties will be imperative to meet the amount of food available for this population. While overall production must increase, there is also growing demand to produce higher-quality and more nutritive food products. Here, we review testing and genetic control of wheat processing and end-use quality for a wheat breeding program.

Bread wheat (*Triticum aestivum* L.) flour is traditionally used for a variety of products including leavened, unleavened and steamed breads, as well as cookies, cakes, and pastries. Each of these products demands flours with specific best-fit quality profile (Morris, 2002, Peña, 2002). Additionally, wheat is used for malt in brewing, ready-to-eat breakfast cereals, and there are growing markets for healthier whole-wheat alternatives and convenience foods such as frozen or refrigerated dough products which are purchased ready-to-cook. These food products represent even wider specific end-use quality requirements for optimal production in an industrial process.

End-use and processing quality for wheat requires a multi-faceted description. Several phenotypes using wheat as grain, flour, dough, and final products must be assessed to determine an overall best end-use product, for a given wheat cultivar or breeding line (Peña, 2002). In general, hard grain with high protein content and strong, extensible gluten is marketed for making industrial pan bread, whereas soft grain with low protein content and weak, extensible gluten is marketed for use in making cookies, cakes, and pastries (Peña, 2002). Many tests must be considered to ensure that wheat lines fall into basic marketing classes, or if there are alternate end-use products for which a wheat line is well suited to produce.

The tests of grain can be done on a small scale, quickly, and cheaply. Several of these tests can be conducted on small samples, many using near infrared spectroscopy (NIRS) or

single kernel characterization (AACC, 2000), making them possible to implement in high throughput programs. However, dough rheology and end-use tests require larger quantities of grain for milling into flour, which implies these tests cannot be conducted until later in the breeding program. In addition, tests conducted post-milling are, in general, more costly and time consuming, which indicates they will likely be performed on fewer samples. Independently these tests can be useful for material culling thresholds, but are often interpreted collectively or used to inform for further stages of testing. Overall, selecting for wheat processing and end-use quality is an art and science, as are the other parts of wheat breeding.

Grain testing

Wheat grain is assessed for pre-milling characteristics, which impact marketing. These tests include kernel weight, test weight per volume, color, hardness, vitreousness of the kernel, and total protein content. In industrial markets, kernel size, volume, and protein tests are often used for bulk purchasing and allow the wheat to be sorted in large marketing classes (e.g. hard white, hard red winter, soft white, etc.), which will later impact milling and flour mixing by millers to ensure consistent end-use products over time. In local markets in the developing world, the visual characteristics of a cultivar are extremely important, as much of this wheat will be milled and used in the home.

Grain color is an indication of amount of phlobaphene, a polyphenol compound, in the aluerone layer of wheat grain (Miyamoto and Everson, 1958). Grain color can be determined visually or through digital imaging, and is typically classified as either white or red, although blue grain is also found (Abdel-Aal and Hucl, 2003). White grain is favored in local, soft, and whole-wheat end-use product markets. Both red and white grains are used in bread-making. Though grain color does not impact endosperm quality or color, regional preference is common for red or white.

Grain color is genetically controlled by the three *R* genes on the long arms of the series 3 chromosomes (McIntosh, *et al.*, 2000), which are transcription factors for the flavonoid biosynthesis pathway (Himi and Noda, 2005). Since these transcription factors inhibit the pathway, red (*a* allele) is dominant to white (*b* allele) at each locus. Additional genetic and environmental variation has been found for grain color beyond the three major alleles (Matus-Cadiz, *et al.*, 2003).

Grain weight per kernel and volume are important characteristics to milling efficiency of wheat. Grain weight is often indicated as thousand kernel weight (TKW). TKW can be measured on using machines which analyze individual seeds, or using a grain counter followed by weighing a specific number of kernels (AACC, 2000). Test weight measures grain weight per volume and is represented in lb bu^{-1} or kg hl^{-1}. There are several instruments available to measure weight per unit volume at varying scales (AACC, 2000).

Grain weight is a parameter of total grain yield along with grains per spike and spikes per unit area. Thus, grain weight is typically highly correlated with grain yield, which is highly impacted by environment and genotype by environment interactions. Grain weight per kernel and per volume were found to have increased in CIMMYT and Great Plains bread wheats over time and was significantly correlated with yield (Cox, *et al.*, 1988, Aisawi, *et al.*, 2015). The grain protein content homeologue on chromosome 6A was recently found to also impact grain yield and protein content as it was associated with senescence (Cormier, *et al.*, 2015). Several other unannotated genetic regions for TKW and test weight have been found across the genome using association mapping strategies (Breseghello and Sorrells, 2006, Liu, *et al.*, 2010, Neumann, *et al.*, 2010, Reif, *et al.*, 2011, Mir, *et al.*, 2012, Edae, *et al.*, 2014).

Wheat endosperm texture also plays an important role in milling and end-use targets. Hardness refers to the strength required to crush wheat grain. Hardness can be measured by particle size index, force required to crush grains, or NIR calibration (AACC, 2000). Hard and soft wheat differ in the strength which starch granules are attached to the protein matrix (Barlow, *et al.*, 1973). Hard wheat has much stronger attachment, thus requiring more force in milling and damaging more starch than in soft endosperm wheat (Giroux and Morris, 1997). Higher damaged starch increases the amount of water which may be absorbed by the dough. Higher amounts of water absorbed are favored in bread baking, as opposed to making cookies and pastries.

Genetically, hardness is qualitatively controlled by *Ha* hardness genes located on the short arm of chromosome 5D, which control hardness class (Morris, 2002, McIntosh, *et al.*, 2013). Wheat lines with wild-type alleles have soft grain, while wheat with null alleles at these loci have hard grain, as is the case in durum wheat which is missing these two loci. Previous results have shown very high proportions of haplotype of hardness genes *Pina-D1-b* with *Pinb-D1-a* in CIMMYT breeding material (Lillemo, *et al.*, 2006). While the hardness loci control much genetic variation for grain hardness, other genetic modifiers are present which help

account for the variation found within class (Morris, 2002, Pasha, *et al.*, 2010). The largest portion of variance in an association study has been attributed to the *Pin* alleles, however other regions were found which also demonstrated genetic control of grain hardness (Bordes, *et al.*, 2011).

Grain protein content is a pre-milling test which is both correlated to yield and quality performance metrics. Protein can be measured through a Kjeldahl combustion method or estimated by an NIR calibration to this method (AACC, 2000). Grain protein content is highly correlated with grain hardness, dough strength (Borghi, *et al.*, 1995, Blandino, *et al.*, 2015), loaf volume in pan breads (Bushuk, 1997), and overall baking of hard wheats quality (Garg, *et al.*, 2006). However, grain protein content is often negatively correlated with yield, as increased grain fill is attributed to increase in starch deposition, and is highly impacted by environment and agronomic management (Terman, *et al.*, 1969, Borghi, *et al.*, 1995, Blandino, *et al.*, 2015). Still, increases have been made in protein content, as well as TKW, over time due to breeding in the Great Plains and CIMMYT (Cox, *et al.*, 1989, Aisawi, *et al.*, 2015).

One gene from durum wheat which was found to increase grain protein content is *Gpc-B1* on the short arm of chromosome 6B (Uauy, *et al.*, 2006). This gene has homeologues on the other 6 series chromosomes as well as on 2 series chromosomes (Cormier, *et al.*, 2015). Though *Gpc-B1* has been known for a longer period of time, the homeologue on chromosome 6A is the one found to show most genetic variation in a diverse panel of bread wheats (Cormier, *et al.*, 2015). Alleles of this gene were found to either promote or delay senescence, which would change the amount of grain filling duration, thus impacting final yield and grain protein concentration. Additionally, studies have shown multiple loci in locations across the genome demonstrating genetic control for grain protein concentration (Neumann, *et al.*, 2010, Bordes, *et al.*, 2011, Reif, *et al.*, 2011).

Flour yield is impacted by TKW and test weight, as discussed earlier, as well as genetic and environmental factor. Increases in flour yield are beneficial to millers, but it is important to note that optimal flour yield is attained when mill rollers and sieves are set appropriately for the common shape and size of a specific wheat line. As such, experimental test mills cannot be reset for each genotype and commercial mills are milling a mixture of many different varieties unless a cultivar is specifically sourced. Significant associations marker-trait associations for flour yield have been found on wheat chromosomes 2D and 5B (Breseghello and Sorrells, 2006).

Endosperm proteins

Wheat is special among cereals for its viscoelastic ability to rise and extend, while still retaining shape and connectivity. The viscoelastic properties of wheat dough originate from the storage proteins; glutenins and gliadins (Payne, *et al.*, 1987, Garg, *et al.*, 2006, Zheng, *et al.*, 2009, Delcour and Hoseney, 2010). Glutenins are responsible for the elasticity and resistance to extension properties of wheat dough (Delcour and Hoseney, 2010). The multiallelic glutenin profile refers to alleles present in the high molecular weight glutenins *Glu-A1*, *Glu-B1*, and *Glu-D1* on the long arms of 1A, 1B, and 1D, respectively, and low molecular weight glutenin alleles *Glu-A3*, *Glu-B3*, and *Glu-D3* on the short arms of 1A, 1B, and 1D, respectively (Payne and Lawrence, 1983, Payne, *et al.*, 1987, Branlard, *et al.*, 1992). The low molecular weight glutenins are in tight linkage with the γ and ω Gli-1 gliadins *Gli-A1*, *Gli-B1*, *Gli-D1* on chromosomes 1A, 1B, 1D, respectively (Payne and Lawrence, 1983, Payne, *et al.*, 1987). Additionally there are α and β Gli-2 gliadins *Gli-A2*, *Gli-B2*, and *Gli-D2*, on chromosomes 6A, 6B, and 6D, respectively (Payne and Lawrence, 1983, Payne, *et al.*, 1987). The gliadins are responsible for the cohesive, visousproperties of wheat dough, which allow it to rise and retain gas in the leavening process (Delcour and Hoseney, 2010). Additionally, other constituents of the wheat kernel, such as non-starch polysaccharides, enzymes, oligosaccharides, phytic acid, lipids, vitamins and minerals, and damaged starch may also have impacts on dough rheology and end-use quality (Delcour and Hoseney, 2010).

Flour sodium dodecyl sulfate (SDS) sedimentation is correlated with both protein concentration and protein quality. In this test flour is mixed with water and rested, then lactic acid is added and solution is again rested. The volume of sediments precipitated is measured. High sedimentation volumes are associated with greater protein and higher gluten strength (AACC, 2000). Micro SDS-sedimentation can be used as a high-throughput test of wheat flour to screen many lines quickly for gluten strength, which indicates it will be impacted by the type and quantity of glutenins present. Additionally, significant marker-trait associations for small quantitative traits controlling SDS-sedimentation were found on 2B, 3A, and 7A (Neumann, *et al.*, 2010), as well as 1B, 2A, 2D, and 5B in a diversity panel (Reif, *et al.*, 2011).

Dough rheology

Flour is mixed with water to create viscoelastic dough. In this process gluten is formed as a complex protein aggregate of the glutenins and gliadins. The gluten-containing dough then

has the ability to maintain a cohesive unit which is both elastic and extensible. The dough can be used in leavened breads which can rise by trapping carbon dioxide within its structure. The exact properties of wheat dough highly impact the optimal end-use product (Peña, 2002).

Dough rheology traits measure attributes related to performance of wheat dough during and following mixing and resting. Several tests are available which may measure gluten strength, extensibility, elasticity, optimum dough development times, tolerance to over mixing, optimum water absorption, starch gelatinization, and starch pasting. Mixograph, farinograph, alveograph, and other equipment are also widely used in dough rheology testing.

These tests are all impacted by amount of water added to the dough. All methods have first estimates of water absorption criteria in the AACC (2000) method, but an experienced researcher may need to optimize water absorption from the estimates made in the method. Alternatively, optimal water absorption may be estimated in using solvent retention capacity tests (Guzmán, *et al.*, 2015).

The mixograph (National Manufacturing, Lincoln, NE) records the resistance of dough mixing on the pins in the mixer (AACC, 2000). The height and width of the resistance curve changes as the flour is initially absorbing water, developing to the optimal mixing time (peak), through time beyond optimal mix (breakdown). The optimal mixing time is important as commercial bakeries demand a specific interval of optimal mixing times and for use in empirical bake tests. Wheat gluten strength can be measured digitally and through assessment of the overall mixing curve. Weak gluten wheats have short mixing time and low tolerance to overmixing. Overmixing tolerance is measured by the width of the curve past peak mix time. Digitally, using the MixSmart software (National Manufacturing, Lincoln, NE), percent torque at the optimal mixing time is an indicator of gluten strength, mix time, and over mixing tolerance can all be measured.

Alveograph is a test of dough functionality which measures the force required to make and break a bubble blown in rested wheat dough. Gluten strength and extensibility can be measured through the curve recorded over time. Dough strength is highly controlled by glutenins (Payne, *et al.*, 1987, Zheng, *et al.*, 2009).

End-use testing

Dough rheology tests are useful for determining various attributes of the wheat line and its protein quality. Dough rheology tests and grain protein tests are moderate to highly correlated

with bread baking volume (Miller, *et al.*, 1956, Kaur, *et al.*, 2004). However, they do not explain full phenotypic variation in wheat bread loaf volume.

Bread making is a very complex system. Baking will be impacted by water and additives applied, mixing speed and time, resting periods, and sheeting, as well as the baker. Controlled wheat bread making processes control for these variations on various scale sizes of the testing (AACC, 2000).

In addition to the flour and dough properties, genetic control for wheat bread loaf volume have been recently identified in the wheat bread making, *wbm*, gene found using RNA-seq (Furtado, *et al.*, 2015). This gene is expressed 14 to 30 days post anthesis. This may control more of the previously unknown variance from protein content and protein quality to final end-use product.

In summary, wheat processing and end-use quality testing is a multi-faceted process involving screening of several traits to determine overall acceptability of a wheat line for an end-use product. Many tests are available which can demonstrate a characteristic of wheat grain or dough properties, but these individual properties do not necessarily indicate that an overall good end-use product can always been made. Thus, selecting for wheat quality is an art and a science. However, recent efforts to digitize and increase throughput of wheat quality testing may help in the breeding process as more modern wheat breeding programs increase their throughput.

References

AACC. 2000. Approved Methods of the American Association of Cereal ChemistsAmer Assn of Cereal Chemists.

Abdel-Aal, E.-S.M. and P. Hucl. 2003. Composition and stability of anthocyanins in blue-grained wheat. Journal of Agricultural and Food Chemistry 51: 2174-2180.

Aisawi, K.A.B., M.P. Reynolds, R.P. Singh and M.J. Foulkes. 2015. The Physiological Basis of the Genetic Progress in Yield Potential of CIMMYT Spring Wheat Cultivars from 1966 to 2009. Crop Science 55: 1749. doi:10.2135/cropsci2014.09.0601.

Barlow, K., M. Buttrose, H. Simmonds and M. Vesk. 1973. The nature of the starch-protein interface in wheat endosperm.

Blandino, M., F. Marinaccio, P. Vaccino and A. Reyneri. 2015. Nitrogen Fertilization Strategies Suitable to Achieve the Quality Requirements of Wheat for Biscuit Production. Agronomy Journal 107: 1584-1594. doi:10.2134/agronj14.0627.

Bordes, J., C. Ravel, J. Le Gouis, A. Lapierre, G. Charmet and F. Balfourier. 2011. Use of a global wheat core collection for association analysis of flour and dough quality traits. Journal of Cereal Science 54: 137-147. doi:10.1016/j.jcs.2011.03.004.

Borghi, B., G. Giordani, M. Corbellini, P. Vaccino, M. Guermandi and G. Toderi. 1995. Influence of crop rotation, manure and fertilizers on bread making quality of wheat (Triticum aestivum L.). European journal of agronomy 4: 37-45.

Branlard, G., J. Pierre and M. Rousset. 1992. Selection indices for quality evaluation in wheat breeding. Theoretical and Applied Genetics 84: 57-64.

Breseghello, F. and M.E. Sorrells. 2006. Association mapping of kernel size and milling quality in wheat (Triticum aestivum L.) cultivars. Genetics 172: 1165-1177. doi:10.1534/genetics.105.044586.

Bushuk, W. 1997. Wheat breeding for end-product use. Wheat: Prospects for global improvement. Springer. p. 203-211.

Cormier, F., M. Throude, C. Ravel, J. Gouis, M. Leveugle, S. Lafarge, et al. 2015. Detection of NAM-A1 Natural Variants in Bread Wheat Reveals Differences in Haplotype Distribution between a Worldwide Core Collection and European Elite Germplasm. Agronomy 5: 143-151. doi:10.3390/agronomy5020143.

Cox, T., M. Shogren, R. Sears, T. Martin and L. Bolte. 1989. Genetic improvement in milling and baking quality of hard red winter wheat cultivars, 1919 to 1988. Crop Science 29: 626-631.

Cox, T., J. Shroyer, L. Ben-Hui, R. Sears and T. Martin. 1988. Genetic improvement in agronomic traits of hard red winter wheat cultivars 1919 to 1987. Crop Science 28: 756-760.

Delcour, J. and R.C. Hoseney. 2010. Principles of cereal science and technology. status: published.

Edae, E.A., P.F. Byrne, S.D. Haley, M.S. Lopes and M.P. Reynolds. 2014. Genome-wide association mapping of yield and yield components of spring wheat under contrasting moisture regimes. Theor Appl Genet 127: 791-807. doi:10.1007/s00122-013-2257-8.

Faostat, F. 2013. Statistical Databases. Food and Agriculture Organization of the United Nations.

Furtado, A., P. Bundock, P. Banks, G. Fox, X. Yin and R. Henry. 2015. A novel highly differentially expressed gene in wheat endosperm associated with bread quality. Scientific reports 5.

Garg, M., H. Singh, H. Kaur and H.S. Dhaliwal. 2006. Genetic Control of High Protein Content and Its Association with Bread-Making Quality in Wheat. Journal of Plant Nutrition 29: 1357-1369. doi:10.1080/01904160600830134.

Gerland, P., A.E. Raftery, H. Ševčíková, N. Li, D. Gu, T. Spoorenberg, et al. 2014. World population stabilization unlikely this century. Science 346: 234-237.

Giroux, M. and C. Morris. 1997. A glycine to serine change in puroindoline b is associated with wheat grain hardness and low levels of starch-surface friabilin. Theoretical and Applied Genetics 95: 857-864.

Guzmán, C., G. Posadas-Romano, N. Hernandez-Espinosa, A. Morales-Dorantes and R.J. Pena. 2015. A new standard water absorption criteria based on solvent retention capacity (SRC) to determine dough mixing properties, viscoelasticity, and bread-making quality. Journal of Cereal Science.

Himi, E. and K. Noda. 2005. Red grain colour gene (R) of wheat is a Myb-type transcription factor. Euphytica 143: 239-242.

Kaur, K., O. Lukow, K. Preston and L. Malcolmson. 2004. How well do early-generation quality tests predict flour performance? Canadian journal of plant science 84: 71-78.

Lillemo, M., F. Chen, X. Xia, M. William, R.J. Peña, R. Trethowan, et al. 2006. Puroindoline grain hardness alleles in CIMMYT bread wheat germplasm. Journal of Cereal Science 44: 86-92.

Liu, L., L. Wang, J. Yao, Y. Zheng and C. Zhao. 2010. Association mapping of six agronomic traits on chromosome 4A of wheat (Triticum aestivum L.). Molecular Plant Breeding 1.

Matus-Cadiz, M., P. Hucl, C. Perron and R. Tyler. 2003. Genotype× environment interaction for grain color in hard white spring wheat. Crop Science 43: 219-226.

McIntosh, R., K. Devos, J. Dubcovsky and W. Rogers. 2000. Catalogue of gene symbols for wheat: 2000 supplement. Wheat Information Service: 33-70.

McIntosh, R., Y. Yamazaki, J. Dubcovsky, W. Rogers, C. Morris, D. Somers, et al. 2013. MacGene 2012: catalogue of gene symbols for wheat.

Miller, B., B. Hays and J. Johnson. 1956. Correlation of farinograph, mixograph, sedimentation, and baking data for hard red winter wheat flour samples varying widely in quality. Cereal Chemistry 33: 277-290.

Mir, R.R., N. Kumar, V. Jaiswal, N. Girdharwal, M. Prasad, H.S. Balyan, et al. 2012. Genetic dissection of grain weight in bread wheat through quantitative trait locus interval and association mapping. Molecular Breeding 29: 963-972. doi:10.1007/s11032-011-9693-4.

Miyamoto, T. and E. Everson. 1958. Biochemical and physiological studies of wheat seed pigmentation. Agronomy Journal 50: 733-734.

Morris, C.F. 2002. Puroindolines: the molecular genetic basis of wheat grain hardness. Plant molecular biology 48: 633-647.

Neumann, K., B. Kobiljski, S. Denčić, R.K. Varshney and A. Börner. 2010. Genome-wide association mapping: a case study in bread wheat (Triticum aestivum L.). Molecular Breeding 27: 37-58. doi:10.1007/s11032-010-9411-7.

Pasha, I., F. Anjum and C. Morris. 2010. Grain hardness: a major determinant of wheat quality. Food Science and Technology International: 1082013210379691.

Payne, P.I. and G.J. Lawrence. 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: 29-35.

Payne, P.I., M.A. Nightingale, A.F. Krattiger and L.M. Holt. 1987. The relationship between HMW glutenin subunit composition and the bread-making quality of British-grown wheat varieties. Journal of the Science of Food and Agriculture 40: 51-65.

Peña, R. 2002. Wheat for bread and other foods. Bread wheat improvement and production. Food and Agriculture Organization of the United Nations. Rome: 483-542.

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

Terman, G., R. Ramig, A. Dreier and R. Olson. 1969. Yield-protein relationships in wheat grain, as affected by nitrogen and water. Agronomy Journal 61: 755-759.

Uauy, C., A. Distelfeld, T. Fahima, A. Blechl and J. Dubcovsky. 2006. A NAC gene regulating senescence improves grain protein, zinc, and iron content in wheat. Science 314: 1298-1301.

Zheng, S., P.F. Byrne, G. Bai, X. Shan, S.D. Reid, S.D. Haley, et al. 2009. Association analysis reveals effects of wheat glutenin alleles and rye translocations on dough-mixing properties. Journal of cereal science 50: 283-290.

Chapter 2 - Applying genomic selection for prediction of processing and end-use quality traits in CIMMYT spring bread wheat breeding program

Sarah D. Battenfield, Jesse A. Poland, R. Chris Gaynor, Ravi P. Singh, Carlos Guzman, R. Javier Peña, and Allan K. Fritz

S. D. Battenfield and A. K. Fritz, Dep. of Agron., Kansas State Univ., 2004 Throckmorton Plant Sci. Ctr., Manhattan, KS, 66506; J.A. Poland, Dep. of Plant Pathology, Kansas State Univ., 4011 Throckmorton Plant Sci. Ctr., Manhattan, KS, 66506; R. C. Gaynor, The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, Midlothian, UK; R. P. Singh, C. Guzman, and R. J. Peña, Global Wheat Program, International Maize and Wheat Improvement Center, Mexico, D.F., Mexico.

Abbreviations

ALVPL, Alveograph P/L value; ALVW, Alveograph W value; AVE, Average of all genomic selection models on line-mean basis; BLUP, Best linear unbiased predictor; CIMMYT, International Center of Maize and Wheat Improvement (Spanish acronym); EN, Elastic net; FLRPRO, Flour protein; FLRSDS, Flour sodium dodecyl sulfate sedimentation; FLRYLD, Flour yield; GAUSS, Gaussian kernel; GBS, Genotyping-by-sequencing; GRNHRD, Grain hardness; GRNPRO, Grain protein; LOFVOL, pup loaf volume; MIXTIM, Mixograph mix time; MP, Mixograph % torque at the integral of the midline peak; NIRS, Near infrared spectroscopy; PCA, Principal component analysis; PLSR, Partial least squares regression; TKW, Thousand kernel weight; TESTWT, Test weight; RF, Random forest; RRBLUP, Ridge regression best linear unbiased prediction; SDS, Sodium dodecyl sulfate;

Abstract

Wheat (*Triticum aestivum* L.) is the second most widely grown cereal grain, primarily used as a food crop. To meet the demands for human consumption, cultivars must possess suitable end-use quality for release and acceptability. However, breeding for quality traits is often considered a secondary goal, largely due to amount of seed needed and overall expense of such testing. Without testing and selection, many undesirable materials tend to be advanced. Here we develop and validate whole genome prediction models for end-use quality phenotypes routinely generated by the CIMMYT bread wheat breeding program. With these predictions genomic selection (GS) can be applied at earlier stages and undesirable materials culled before implementing expensive yield and quality screenings. Prediction accuracy was tested using quality data from unbalanced yield trials from 2009 to 2014 (n = 4,095) at Ciudad Obregon, Mexico evaluated for quality parameters: test weight, thousand kernel weight, grain hardness, grain and flour protein, flour yield, SDS-sedimentation, Mixograph and Alveograph performance, and bread loaf volume. High-density markers were generated with genotyping-by-sequencing and SNPs were imputed. Prediction models were developed using ridge regression, Gaussian kernel, partial least squares, elastic net, and random forest models in R. In general, prediction accuracy increased over time as more data was available to train the model. Mean forward prediction accuracies (r) for quality parameters in 2014 ranged from 0.262 (grain hardness) to 0.593 (mix-time). Based on these prediction accuracies, we conclude that GS can be a useful tool to facilitate early generation selection for end-use quality in wheat.

Introduction

The human population is growing exponentially with current projections predicting a population of greater than 9 billion by the year 2050 (Gerland, *et al.*, 2014). An intersection of improved agronomic practices and improved crop varieties will be imperative to meet the food required for this population. While overall production must increase, there is also growing demand to produce higher-quality food products. Bread wheat (*Triticum aestivum* L.) flour is used for a variety of products including leavened breads, unleavened breads, noodles, cookies, cakes, and pastries. Each of these products demands flours with specific best-fit quality profile (Peña, 2002).

End-use and processing quality for wheat is difficult to define by any one given parameter. Several phenotypes using wheat as grain, flour, dough, and final products must be assessed to determine an overall best end-use product, for a given wheat cultivar or breeding line. Typically, hard grain with high protein and strong and extensible gluten is acceptable for making industrial pan bread, whereas soft grain with low protein and weak and extensible gluten is more acceptable for cookies, cakes, and pastries (Peña, 2002). However, many tests must be considered to ensure that wheat lines fall into these basic marketing classes. Some of these tests are useful on their own with general selection thresholds, whereas others should be interpreted collectively or are used to inform for further stages of testing. Overall, selecting for wheat processing and end-use quality is an art and science, as are the other parts of wheat breeding.

Wheat grain is assessed for pre-milling characteristics, which impact marketing. These tests include kernel weight, weight per volume, color, hardness, vitreousness of the kernel, and total protein content. Many of these characteristics are strongly correlated with grain yield with varying levels of heritability. Grain weight was found to have increased in CIMMYT bread wheats over time and was significantly correlated with yield (Aisawi, *et al.*, 2015). In contrast, grain protein content is often negatively correlated with yield and is highly impacted by environment and agronomic management (Terman, *et al.*, 1969).

Wheat endosperm texture also plays an important role in milling and end-use targets. Hard and soft wheat differ in the strength of which starch granules are attached to the protein matrix. Hard wheat has much stronger attachment, thus requiring more force in milling and damaging more starch than in soft endosperm wheat (Giroux and Morris, 1997). Higher damaged starch increases the amount of water which may be absorbed by the dough, and higher amounts

of water are favored in bread baking as compared to making cookies and pastries. Genetically, hardness is qualitatively controlled by *Ha* hardness genes located on the short arm of chromosome 5D. In industrial markets, kernel size, volume, and protein tests are often used for bulk purchasing and allow the wheat to be sorted in large marketing classes (e.g. hard white, hard red winter, soft white, etc.), which will later impact milling and flour mixing by millers to ensure consistent end-use products over time. In local markets in the developing world, the visual characteristics of a cultivar are extremely important as much of this wheat will be milled and used in the home.

The next stage of testing produces information regarding milling value and protein concentration of the flour. Increases in flour yield can be beneficial to millers, but it is important to note that optimal flour yield is attained when mill rollers and sieves are set appropriately for the common shape and size of a specific wheat line. As such, experimental test mills cannot be reset for each genotype and commercial mills are running a mixture of all different varieties. The other two tests measure the amount of protein in the flour sample, as this is very important to be correct for further testing.

Wheat is special among cereals for its viscoelastic ability to rise and extend, while still retaining shape and connectivity. Dough rheology and end-use tests involve mixing flour into dough to determine the viscoelastic properties of strength, elasticity, and tolerance. These tests are time consuming, costly, and require large quantities of flour. However, each of these tests are collectively necessary to predict an ideal end-use product for a specific wheat line (Peña, 2002).

The viscoelastic properties of wheat mostly originate from the storage proteins; glutenins and gliadins (Payne, *et al.*, 1987, Garg, *et al.*, 2006, Zheng, *et al.*, 2009, Delcour and Hoseney, 2010). Glutenins are responsible for the elasticity and resistance to extension properties of wheat flour dough. The multiallelic glutenin profile (Payne and Lawrence, 1983, Payne, *et al.*, 1987) in the high molecular weight glutenin alleles *Glu-A1*, *Glu-B1*, and *Glu-D1* are on the long arms of 1A, 1B, and 1D, and low molecular weight glutenin alleles *Glu-A3*, *Glu-B3*, and *Glu-D3* are on the short arms of 1A, 1B, and 1D (Branlard, *et al.*, 1992). Gliadins, *Gli-A1*, *Gli-B1*, *Gli-D1 Gli-A2*, *Gli-B2*, and *Gli-D2*, on chromosomes 1A, 1B, 1D, 6A, 6B, and 6D, respectively (Payne and Lawrence, 1983, Payne, *et al.*, 1987), are responsible for the cohesive properties of wheat dough, which allow it to rise and retain gas. Additionally, other constituents of the wheat kernel, such as

non-starch polysaccharides, enzymes, oligosaccharides, phytic acid, lipids, vitamins and minerals, and damaged starch may also have impacts on dough rheology and end-use quality.

Historically, wheat breeding has focused on using yield and visual selection for lines with improved agronomic performance, and disease resistance. Quality traits are generally evaluated as a final performance test because the tests are intensive, expensive, and usually cannot occur until later in the breeding program due to the large amount of grain necessary. This often results in advancement of promising wheat cultivars that should not be released due to poor quality. In addition, there is limitation for developing wheat cultivars with good and specialized end-use traits. Accurate processing and end-use quality prediction models would allow breeding programs to cull unacceptable lines or target specific lines earlier in the pipeline, before money and time was invested in lines which would not pass the final test.

Due to the polygenic nature of these traits, marker assisted selection with previously identified significant markers is not a fully applicable solution to the problem (Heffner, *et al.*, 2011). Genomic selection (GS) models, however, utilize high-density genotype data sets simultaneously model all additive genetic variance. These models use entries with known phenotype and genotype to train an algorithm, cross-validate the prediction, and then predict quantitative traits in materials with only genotype information available. This approach was first introduced into animal breeding by Meuwissen, *et al.* (2001)advocating for ridge regression and Bayesian approaches to solve this problem. Their claim that attaining large amount of markers would become cheaper than phenotyping each individual is coming to fruition (Poland and Rife, 2012). Many methods have been proposed that handle the problem of multicollinearity from massively more predictors (markers) than observations available (Lorenz, *et al.*, 2011). Taking all this into consideration, GS could serve as a way to predict processing and end-use quality phenotypes earlier in the pipeline before breeders have enough seed for testing and allow predictions of more individuals than would be possible to phenotype.

Genomic selection has been tested many times for wheat yield and disease resistance (Heffner, *et al.*, 2009, Crossa, *et al.*, 2010, Rutkoski, *et al.*, 2010, Rutkoski, *et al.*, 2012, Dawson, *et al.*, 2013, Crossa, *et al.*, 2014, Rutkoski, *et al.*, 2014), but not thoroughly for wheat processing and end-use quality. GS was tested in wheat end-use quality in a biparental population and a small breeding population (Heffner, *et al.*, 2009, Heffner, *et al.*, 2011). These studies utilized cross-validation, rather than forward prediction approaches and used soft wheat quality traits

from a small market. They did find processing and end-use quality traits to be more highly predictive than grain yield. Here we conducted forward prediction in the breeding program with GS models on all end-use and processing quality traits regularly assessed by the CIMMYT bread wheat breading program. The objective of this study was to determine prediction accuracy of several GS models in several wheat processing and end-use traits over time, with the intention of introducing this method to the CIMMYT bread wheat breeding program.

Materials and Methods

Germplasm

Wheat lines used in training and testing the GS models were from first year yield trial materials advanced to quality testing in CIMMYT bread wheat breeding program. All wheat lines were grown in Ciudad Obregon, Sonora, Mexico, over subsequent years. However, a given line was only evaluated for quality in one year, to allow for the largest, least selected training set possible. Materials were planted in a lattice design with 28 entries to every 2 checks, in 2 replications. Only those selected for superior yield or other agronomic performance were advanced to processing and end-use quality testing. A single sample from one replication was used to measure grain, flour, dough, and end-use quality phenotypes for each selected wheat line.

Phenotypes

Grain morphological characteristics were evaluated with digital image system SeedCount SC5000 (Next Instruments, Australia) to obtain thousand kernel weight (TKW, g) and test weight (TESTWT, kg hl^{-1}). Grain protein (GRNPRO), hardness (GRNHRD), and moisture content were determined by near-infrared spectroscopy (NIRS), using NIR Systems 6500 (Foss, Denmark) by the official methods of the American Association of Cereal Chemists (AACC) 39-10, 39-70A, and 39-00, respectively (AACC, 2000). GRNPRO was reported at 12.5% moisture basis. Grain samples were tempered and milled using a Brabender Quadrumat Jr. experimental mill (C. W. Brabender OHG, Germany). Flour protein (FLRPRO) and moisture content were estimated by NIRS using an Antaris II FT-NIR Analyzer (Thermo, USA). Both NIRS instruments were calibrated for particle size index (AACC Method 55-30), moisture (AACC Method 44-15A), and protein (AACC Method 46-11A). Sodium dodecyl sulfate (SDS)-sedimentation (FLRSDS) was conducted as in Peña, *et al.* (1990).

Dough rheology was assessed using the Mixograph (National Mfg. Co., USA) according to AACC Method 54-40A (AACC, 2000), and the Chopin Alveograph (Tripette & Renaud, France), AACC Method 54-30A (AACC, 2000). These methods were adjusted to allow for variable water content based on Solvent Retention Capacity, as in Guzmán, *et al.* (2015). Optimal mix time (MIXTIM) and torque (MP) were measured by Mixograph. Dough strength, work value (ALVW), and tenacity vs. extensibility, the ratio of height to length (P/L, ALVPL), were measured using Alveograph. Alveograph P/L values were log transformed prior to analysis for normalization. Bread was baked to test end-use productivity as pan bread with AACC Method 10-09 (AACC, 2000). Pup loaf baking also utilized the Guzmán, *et al.* (2015) adjustment for optimal water absorption. Bread loaf volume (LOFVOL) was measured by rapeseed displacement in accordance with AACC Method 10-05.01 (AACC, 2000).

Genotypes

Leaf tissue was collected and bulked from five plants per line and DNA was extracted using a modified CTAB protocol (Saghai-Maroof, *et al.*, 1984). DNA was quantified and normalized, then digested with a two-enzyme approach, barcoded, amplified, then sequenced (Poland, *et al.*, 2012). Sequences were trimmed to 64 base pairs, unique sequence tags were aligned, and single nucleotide polymorphisms (SNPs) were recoded numerically as (-1, 0, 1) using a modification of TASSEL 5v2 (Bradbury, *et al.*, 2007) in Java script. The SNPs were aligned with the IWGSC draft reference map (International Wheat Genome Sequencing, 2014) using Bowtie 2 (Langmead and Salzberg, 2012). SNPs were investigated for percent missing and heterozygosity. Markers with greater than 20% missing data or greater than 20% percent heterozygous and individuals with greater than 80% missing data were removed from further analysis. Remaining missing SNPs were imputed using mean imputation based on marker frequency using R (R Development Core Team, 2014) package 'rrBLUP' (Endelman, 2011).

Analyses

Genomic selection models were constructed using packages in R (R Development Core Team, 2014). Ridge regression best linear unbiased predictor (RR-BLUP) and reproducing kernel Hilbert space, here referred to as Gaussian kernel (GAUSS), models were conducted using the package 'rrBLUP', as described in Endelman (2011). Partial least squares regression (PLSR), elastic net (EN), and random forest (RF) were tested using R packages 'pls' (Mevik and

Wehrens, 2007), 'glmnet' (Friedman, *et al.*, 2009), and 'randomForest' (Liaw and Wiener, 2002), respectively. These models were combined by Gaynor (2015) into R package 'GSwGBS'.

The average prediction across prediction models was calculated using standardized values to avoid overly weighting the average towards any single prediction model. This was accomplished by first calculating standardized values, *z*, using the equation: $z_{ij} = x_{ij} - \bar{x}_j s_j$, where x_{ij} is the predicted value of the *i*-th individual from the *j*-th model and $s_j = \sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j) \,/\, (n-1)}$ is the sample standard deviation for predictions from the *j*-th model. The standardized predictions for each individual were then averaged across prediction methods. This average, $\bar{z_i}$, was then returned to its original units, $\bar{x_i}^*$, using a back-transformation: $\bar{x_i}^* = \bar{z_i} s_p + \bar{\bar{x}}$, where $s_p = \sqrt{\sum_{j=1}^{m} {s_j}^2 / m}$ is the square root of the pooled variance.

RRBLUP uses a mixed model to solve for individual random marker effects (Endelman, 2011). These effects are then multiplied by the marker matrix of the line to be predicted. GAUSS is similar to ridge regression, except it utilizes a kernel effect based on the Euclidean distance between lines to determine genetic covariance instead of marker matrix (Endelman, 2011). PLSR is similar to regression based on principal components (Mevik and Wehrens, 2007). We used a 10-fold cross validation to train the PLSR model for optimal number of components to be used in the prediction algorithm. EN fits a generalized linear model with penalized maximum likelihood (Friedman, *et al.*, 2009). EN was also trained using a 10-fold cross validation in training data to tune the alpha, mixing, and lambda, regularization, parameters. RF regression is based on a decision tree method (Breiman, 2001). The RF predictions were made using 1000 trees.

Models were tested using temporal forward and cross validation predictions. Forward predictions were conducted using data as it would have historically become available to predict the following year (i.e. 2009 predicts 2010, 2009 and 2010 predict 2011, etc.). Cross validation predictions were conducted on all data across all years with 20% random masked, which was replicated 10 times. Predictions were examined using linear models comparing the predicted and actual values in (R Development Core Team, 2014) and correlations between predicted phenotypes and empirical phenotypes are presented.

Results and Discussion

Materials and genotypes

Phenotypic assessment was conducted on 6,398 lines in first year yield trials between 2009 and 2014 for processing and end-use quality. In the first year of the project, 2009, only individuals promoted to advanced testing were genotyped, whereas other years all individuals in the first year yield test were genotyped. This resulted in many fewer individuals present in the first year. Individuals were filtered for large amounts of missing data per individual. Overall this resulted in 4,095 individuals with high quality genotype and phenotype for GS (Table 1). Originally, 20,833 SNPs were found using the TASSEL pipeline which also aligned to the IWGSC reference sequence (International Wheat Genome Sequencing, 2014). These SNPs were then restricted no more than 20% missing to ensure higher accuracy through reduced reliance on imputation, resulting in 3,075 SNPs that were used in the GS models.

Phenotype distributions of all traits within all years followed an approximately normal distribution (Fig. 1 & 2), except Alveograph P/L (Fig. 2), which was log transformed for subsequent analysis (Box, 1964). Phenotype mean and standard deviations are presented by year (Table 2). Since materials were not replicated across years in this model, heritability was not calculated, however, it is generally assumed that the heritability of most processing and end-use quality traits is intermediate to high (Breseghello and Sorrells, 2006, Kuchel, *et al.*, 2006).

Protein assessments were highly correlated to each other (Fig. 3). This is expected since the majority of the protein in the wheat kernel is stored in the endosperm (Delcour and Hoseney, 2010). Most dough rheology traits evaluated here were highly correlated among themselves, with the exception of Alveograph P/L (Fig. 3). Phenotypic correlations in this study again demonstrate that no single quality test is a substitute for end-use testing, as the correlations from all other parameters are present, but not strongly correlated to final pup loaf volume (Fig. 3). This further supports classification systems for end-use as a function of several quality phenotypes (Peña, 2002, Guzmán, *et al.*, 2014).

Genomic selection

GS models used in this study all tended to produce highly correlated results, with the exception of random forest (Fig. 4-6). For traits of varying genetic architecture, however, models may have differing accuracy. Model averaging has been shown as a valid option when the ideal prediction model is unknown, as is the case in forward prediction (Raftery, *et al.*, 1997, Raftery,

et al., 2010), thus, model averaging with prediction models normalized for mean and standard deviation was conducted for each entry. In general, GAUSS was the best prediction model in cross validation (Fig. 7). However, the AVE model produced consistently high correlations in forward predictions between predicted and empirical phenotypes for all traits and were not heavily impacted by fluctuations in accuracy by trait as individual prediction models were (Fig. 4-6). This indicates that either the models with lower prediction accuracy are adding information to the overall mean or the other models are overfitting to the training data, thus favoring them in cross validation.

TKW and TESTWT were evaluated here since they impact milling and were assessed in the wheat quality laboratory. Data for TKW was only available starting in 2012 (Table 2). Predictions were better with an increase from 0.40 to 0.44 in the subsequent year (Table 3). Random forest performed worst for this trait (Fig. 4). It is unclear whether the accuracy for this trait will continue to increase with time, but the cross validation accuracies for TKW seem to indicate there is room for continued improvement (Fig. 7). Test weight predictions increased from ~ 0.1 correlation in the first year to 0.36 in 2014 (Table 3 and Fig. 4). This trait is highly impacted by environment and relatively low in heritability, like yield. However, it is promising to see prediction accuracy tripling over time.

There was no predictive ability for GRNHRD in 2011. With a larger training set over time, this increased to 0.26 correlation between the observed and predicted (Table 3 and Fig. 3). Additionally, GRNHRD has one of the lowest predictive accuracies in cross validation (Fig. 7). These results corroborate with Heffner, *et al.* (2011) who found that softness had lower prediction accuracy than other quality traits. While there is a normally distributed phenotypic range for GRNHRD (Fig. 1), most material in this data set was classified as hard or semi-hard, with few soft lines present. A high proportion of the CIMMYT historical and breeding lines previously tested had the haplotype *Pina-D1b* and *Pinb-D1a* alleles for the hardness, *Ha,* genes on the short arm of chromosome 5D (Lillemo*, et al.*, 2006). Protein concentration, where more protein leads to harder grain, may be one of the factors responsible for some of the smaller differences found within hardness class.

FLRYLD data was not available until 2011 for prediction in 2012 (Tables 2 and 3). The predictions for FLRYLD were highest in the first year of testing and dropped slightly in the two following years (Table 3 and Fig. 4). Grain and flour protein are very highly correlated

phenotypes (Fig. 3), and follow very closely to one another in predictive ability using GS (Table 3 and Fig. 5). In these traits we saw a general increase over time. Protein traits could continue to increase in accuracy as they are still not nearing cross validation accuracy (Fig. 7). FLRSDS, which is correlated to both protein and dough rheology traits (Fig. 3), is fairly highly predictive (Figures 5 and 7), but may have come to a forward accuracy plateau of between 0.5 and 0.6 (Fig. 5 and Table 3).

Dough rheology traits MIXTIM, MP, and ALVW are all highly correlated (Fig. 3). MP and ALVW are measures of gluten strength, while ALVPL is a better indication of the balance of viscoelasticity. These traits are the foundation of determining gluten strength classification (Peña, unpublished), which informs end-use quality type. For example, strong gluten is typically favored in pan breads, medium strength gluten is better for flat breads and noodles, weak gluten is best for cakes, cookies, and pastries, and tenacious gluten is only acceptable as wheat for animal feed (Peña, 2002). MIXTIM, MP, and ALVW are all highly predictive with forward (Fig. 5, 6, and Table 3) and cross validation (Fig. 7 and Table 3) GS models. However, ALVPL has a lower forward and cross validation prediction accuracy (Fig. 6, 7, and Table 3). Accuracies have increased approximately 5% with log transformation of ALVPL (data not shown). These varying dough rheology prediction accuracies could be due to dough strength having high genetic control from the high and low molecular weight glutenins, whereas ALVPL values less than 0.8 or greater than 1.2 would be influenced by more factors (with more environment dependence) apart from specific glutenin profile.

Baking a pup loaf is the final end-use quality test to determine appropriateness of a wheat line for industrial pan bread. This test gives quantitative and qualitative results not only of how big the resultant loaf is, but also the appearance of loaf and crumb structure. Here we demonstrate that forward prediction accuracy of LOFVOL is approximately 0.45 for the last three years (Fig. 7 & Table 3), but could reach as high as cross validation accuracy 0.67 (Fig.6 & Table 3).

Prediction accuracy of whole-genome models was lower in forward prediction (Fig. 4-6) than cross validation (Fig. 7) for all traits (Table 3). This is likely due to cross validation models using training and testing data containing all years, thus better accounting for environmental variation prior to training the prediction models. Another reason could be due to the possibility of full-siblings being randomly assigned to training and testing sets. The selection procedure in

the breeding program keeps all good material, regardless of their relationship, and sometimes favors advancement of large groups of full siblings. In the full yield trial of 2014 (n=7,672), there was an average of 5.3 entries per cross, with a maximum of 51 full siblings for one specific cross (data not shown). Thus, we assume cross validation represents an over inflation compared to forward predictions, and could possibly represent a ceiling of highest attainable prediction accuracy for a given trait and model.

Conclusions

Wheat quality is typically not the primary breeding objective, often secondary to yield, agronomic performance, and disease resistance. However, with the implementation of GS for wheat quality, predictions that are available in earlier generations of selection will enable better selection for quality and even targeting of wheat lines to potential areas of specific end use. The models here demonstrate that GS for processing and end-use quality has sufficient accuracy for implementation in the breeding program. In addition, the accuracy increased over time, likely due to increasing training population size. Finally, we corroborate previous research (Gaynor, 2015) showing that model averaging gives stable high forward prediction accuracy among all methods.

Phenotyping for wheat processing and end-use quality for the traits included in this study can take 1 kg of seed and cost approximately $60 US dollars at the internal rate, at high throughput, without assessment of indirect costs. Genotyping a wheat line with GBS currently costs ~$10 USD per line, which can also be utilized for multiple trait GS and other analyses. Wheat breeding programs may screen lines for traits which can be assessed in small samples, such as protein, SDS-sedimentation, or mixograph, as early as head or line row stage, but typically do not have enough seed for all tests until after preliminary yield tests. This makes GS much less expensive and has potential for predictions years earlier than phenotyping, especially when considering that the cost is applicable to many traits at one time. Still, the authors note that currently GS is not considered to be a complete replacement for phenotypic selection, but that it can be used to make more informed selection decisions for material advancement between harvest and planting of the next cycle and prioritizing what is evaluated in the quality labs.

Computation time and resource availability is often a consideration in GS model choice as training and testing populations grow, along with the number of traits and possibly models

used in GS. GS models should run quickly in order to produce phenotypic prediction in a timely manner for a breeding program. These models were computed on the Beocat high performance computing cluster at Kansas State University. While larger resources were available, the models for forward prediction all ran on a modest amount of resources in a relatively short amount of time; 8 cores running with 4 Gb memory per core for 11 traits all ran within 24 hours on these data sets with maximum 20% missing markers. However, models with large amounts of computations for relational structure or bootstrapping are more time consuming. As we moved forward into the 2015 breeding program predictions for ~9,000 lines, computation time and resources were increased with GS models running within 1 week.

GS for processing and end-use quality at CIMMYT has now been in development since 2012. In 2014, predictions for end-use and processing quality were available in the fall before phenotype assessments were completed. In 2015, quality phenotypes were predicted in the spring around the time of harvest of 9,000 lines in preliminary yield trails. These predicted phenotypes, with the assumed accuracy from the 2014 cycle, were available to breeders as selections for advancement were made. It is expected that predictive information regarding end-use quality earlier in the breeding program will enable selections to be made for specific end-use quality products in the near future of wheat breeding.

Tables

Table 2-1: Materials available for genomic selection modeling

Trial Harvest Year	Total in Yield Trial	Screened for quality	Phenotype and genotype available
2010	4,956	1,258	250
2011	6,685	1,000	995
2012	10,196	1,580	850
2013	9,436	1,215	886
2014	7,672	1,345	1,114
Total	38,945	6,398	4,095

Table 2-2: Phenotype means and standard deviations by year.

Year	2010		2011		2012		2013		2014	
Entries	250		995		850		886		1114	
	MEAN	SE	MEAN	SE	MEAN	SE	MEAN	SE	MEAN	SE
TKW					48.30	0.13	46.57	0.11	47.79	0.11
TESTWT	82.43	0.06	80.15	0.05	82.37	0.03	81.83	0.03	81.74	0.03
GRNHRD	40.75	0.36	45.77	0.15	40.31	0.16	42.95	0.11	43.56	0.09
GRNPRO	12.07	0.05	11.73	0.02	11.31	0.02	11.70	0.02	12.23	0.02
FLRYLD			67.55	0.11	68.83	0.08	69.35	0.06	70.57	0.06
FLRPRO	10.22	0.05	10.20	0.02	9.57	0.02	9.99	0.02	10.71	0.02
FLRSDS	14.86	0.15	14.35	0.07	13.83	0.08	14.05	0.26	13.68	0.06
MIXTIM	2.75	0.04	3.15	0.02	3.11	0.02	3.35	0.03	2.97	0.02
MP			106.12	1.00	116.41	0.92	123.02	1.10	113.16	0.83
ALVW	285.88	5.70	256.68	2.17	271.58	2.33	291.74	3.10	253.06	2.49
ALVPL	1.04	0.02	0.93	0.01	1.03	0.01	0.99	0.01	0.96	0.01
LOFVOL	746.12	4.22	785.25	1.49	752.46	2.48	807.83	1.85	822.59	1.72

TKW- thousand kernel weight (g), TESTWT- test weight (kg hL^{-1}), GRNHRD- grain hardness (PSI), GRNPRO- grain protein (at 12.5% moisture basis), FLRYLD- flour yield from milling (% recovered), FLRPRO- flour protein (at 14% moisture basis), FLRSDS- SDS-sedimentation volume (mL), MIXTIM- optimum mix time (min), MP- torque at the integral of the midline

peak, ALVW- work value from alveograph curve (J), ALVPL- Alveograph P, strength, divided by L, extensibility, (mm mm^{-1}), LOFVOL pup loaf volume (cc).

Table 2-3: Average GS prediction accuracies of forward and cross-validation. Forward predictive models trained on all prior data, whereas cross-validation trained on a random 80% of the data to predict the remaining masked 20%. Cross validation was replicated 10 times. Average was conducted after variance within each model was standardized.

Validation population	2011	2012	2013	2014	Cross Validation
Training size	250	995	2095	2981	3276
Testing size	995	850	886	1,114	819
TKW			0.400	0.443	0.620
TESTWT	0.102	0.276	0.263	0.362	0.643
GRNHRD	-0.014	0.114	0.217	0.256	0.497
FLRYLD		0.410	0.357	0.364	0.517
GRNPRO	0.284	0.452	0.433	0.477	0.646
FLRPRO	0.259	0.410	0.369	0.435	0.647
FLRSDS	0.378	0.542	0.578	0.521	0.668
MIXTIM	0.404	0.532	0.657	0.595	0.696
MP		0.453	0.592	0.559	0.683
ALVW	0.338	0.523	0.615	0. 551	0.674
ALVPL	0.252	0.288	0.354	0.470	0.514
LOFVOL	0.309	0.453	0.460	0.448	0.667

TKW- thousand kernel weight (g), TESTWT- test weight (kg hL^{-1}), GRNHRD- grain hardness (PSI), GRNPRO- grain protein (at 12.5% moisture basis), FLRYLD- flour yield from milling (% recovered), FLRPRO- flour protein (at 14% moisture basis), FLRSDS- SDS-sedimentation volume (mL), MIXTIM- optimum mix time (min), MP- torque at the integral of the midline peak, ALVW- work value from alveograph curve (J), ALVPL- Alveograph P, strength, divided by L, extensibility, (mm mm^{-1}), LOFVOL pup loaf volume (cc).

Figures

Figure 2-1: Distribution of thousand kernel weight, test weight, grain hardness, flour yield, grain protein, and flour protein across all years.

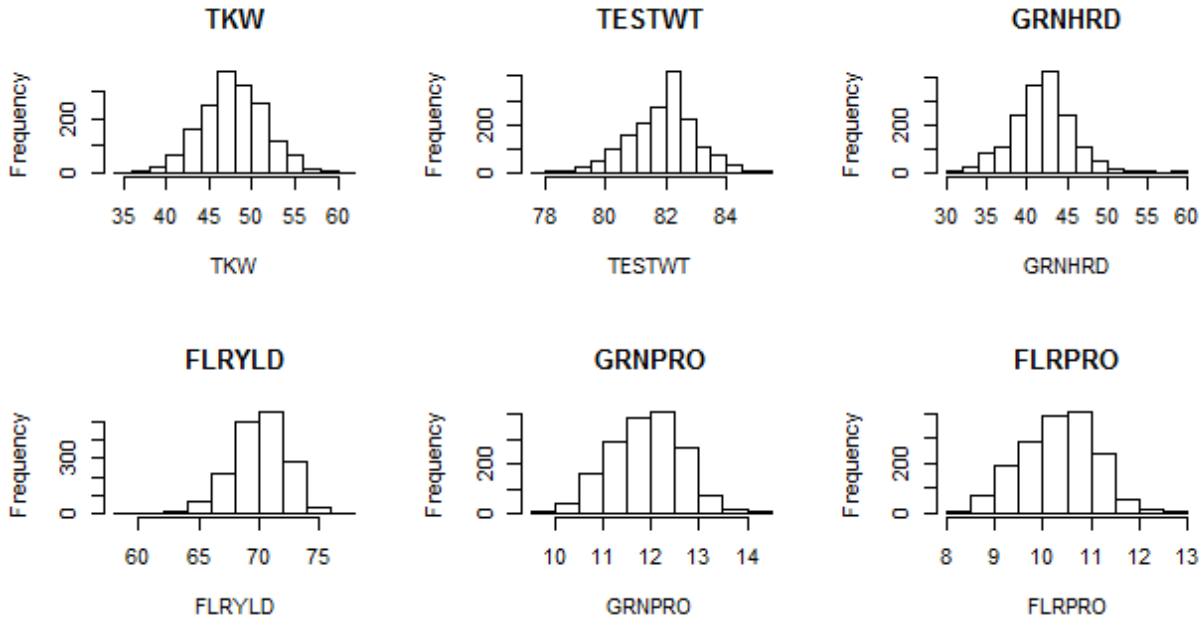


Figure 2-2: Distribution of flour SDS-sedimentation, Mixograph mix time and midline peak, Alveograph W and P/L, and loaf volume across all years.

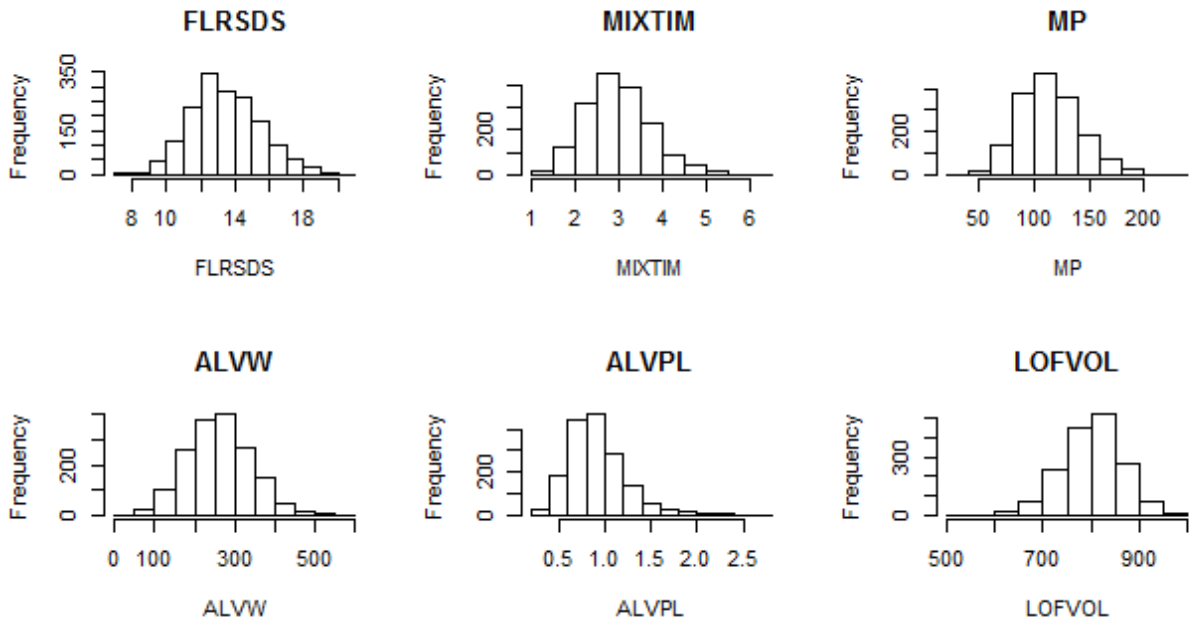


Figure 2-3: Correlation scatterplot of all processing and end-use quality phenotypes from 2014.

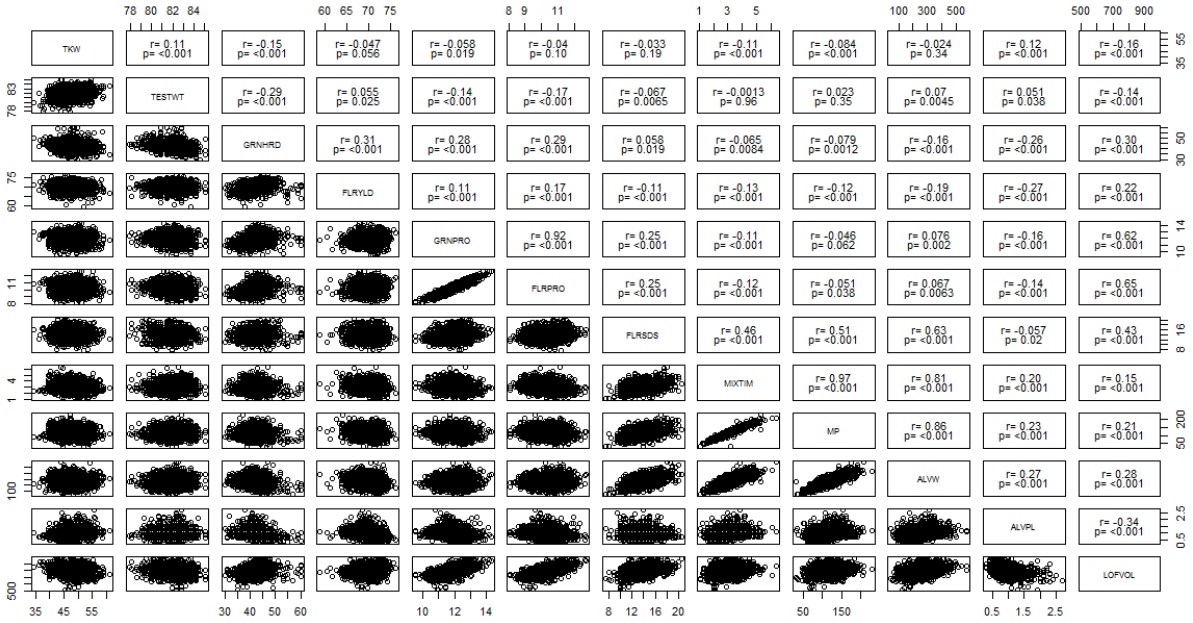


Figure 2-4: GS prediction accuracies for thousand kernel weight, test weight, grain hardness, and flour yield over time.

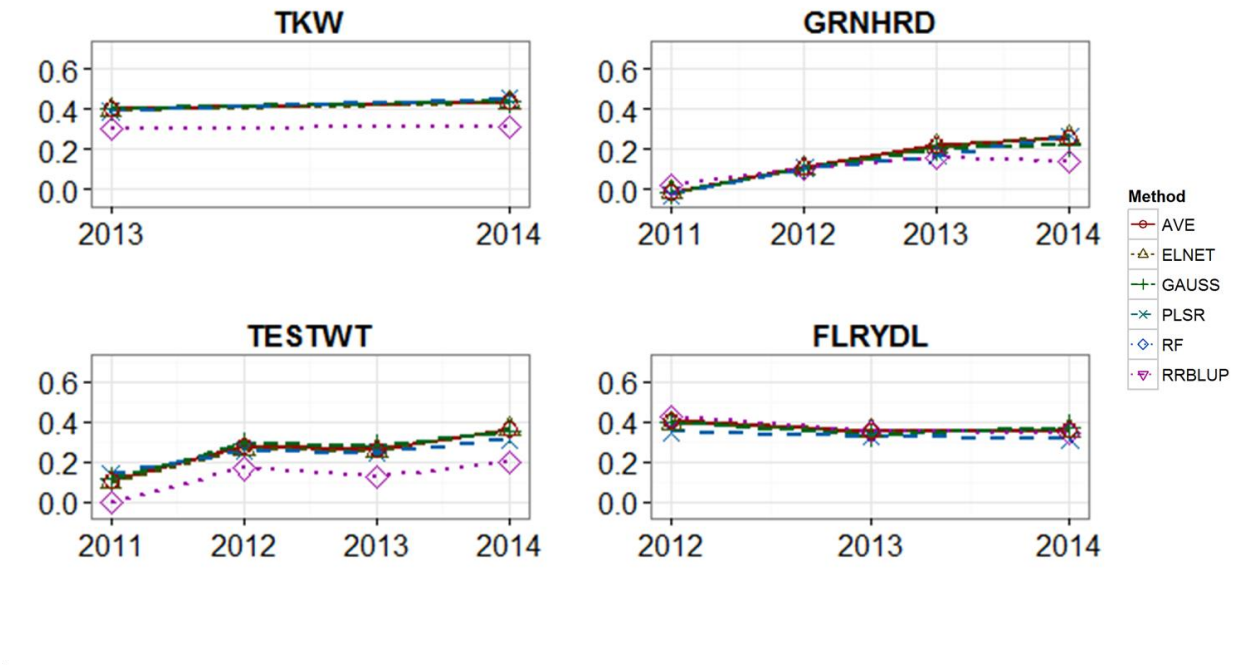


Figure 2-5: GS prediction accuracies for grain protein, flour protein, flour SDS-sedimentation, and Mixograph mix time over time.

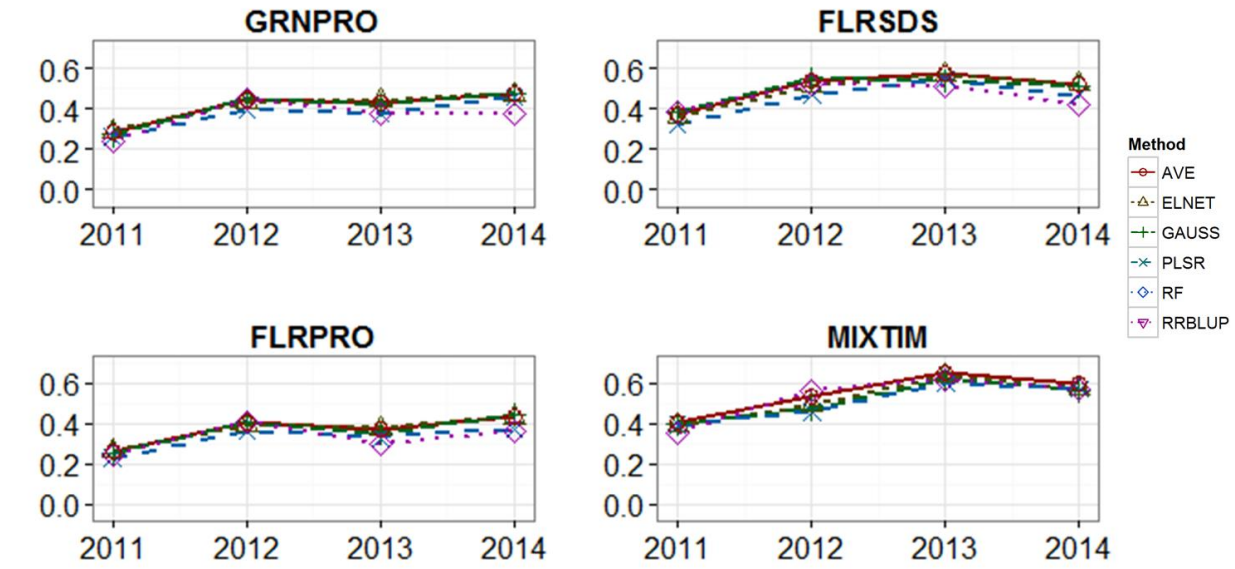


Figure 2-6: GS prediction accuracies for Mixograph torque, Alveograph W and P/L, and loaf volume over time.

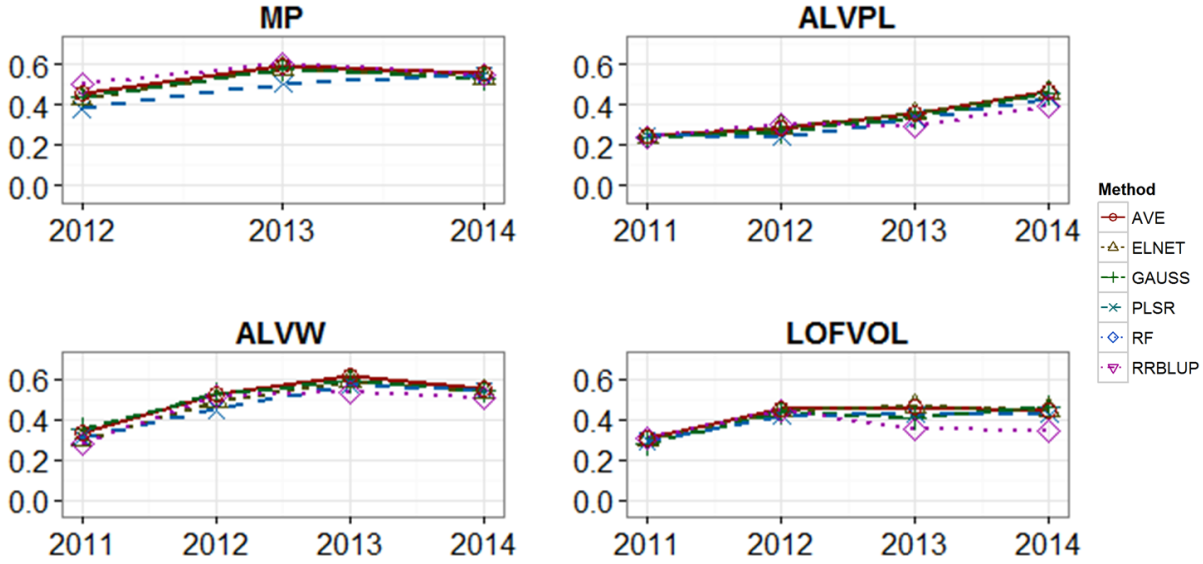
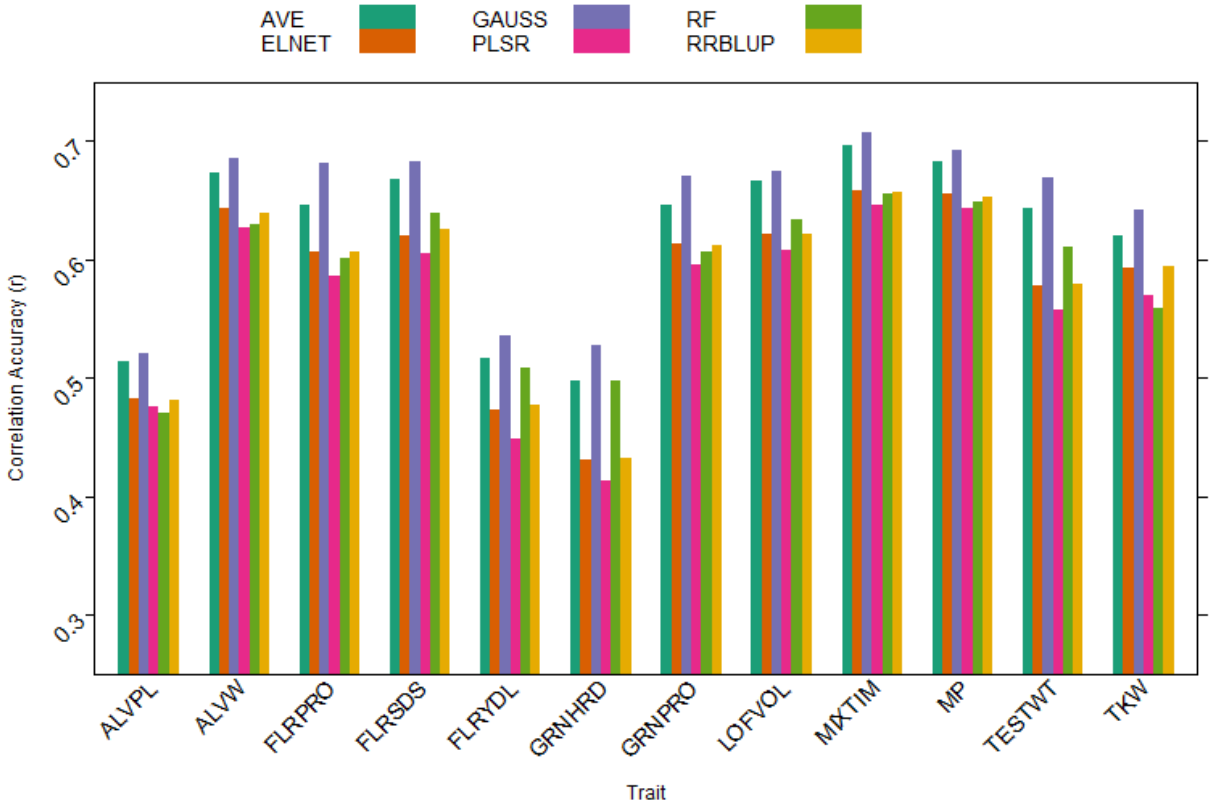


Figure 2-7: Genomic selection cross validation accuracies



Acknowledgements

Battenfield's research was supported through a Monsanto Beachell-Borlaug International Scholars Program fellowship. Funding for this project was provided by US Agency for International Development Feed the Future Initiative (USAID Cooperative Agreement No. AID-OAA-A-13-0005) and the Bill & Melinda Gates Foundation through a grant to Cornell University for "Genomic Selection: The next frontier for rapid gains in maize and wheat improvement." Support for phenotyping of quality traits was provided by CGIAR CRP WHEAT, Durable Rust Resistance Project, and Fondo Sectorial SAGARPA-CONACYT (No. 146788 – "Sistema de mejoramiento genético para generar variedades resistentes a royas, de alto rendimiento y alta calidad para una producción sustentable en México de trigo") of the Mexican government. The computing for this project was performed on the Beocat Research Cluster at Kansas State University, which is funded in part by NSF grants CNS-1006860, EPS-1006860, and EPS-0919443. This work represents contribution number 16-037-J from the Kansas Agricultural Experiment Station.

References

AACC. 2000. Approved Methods of the American Association of Cereal ChemistsAmer Assn of Cereal Chemists.

Aisawi, K.A.B., M.P. Reynolds, R.P. Singh and M.J. Foulkes. 2015. The Physiological Basis of the Genetic Progress in Yield Potential of CIMMYT Spring Wheat Cultivars from 1966 to 2009. Crop Science 55: 1749. doi:10.2135/cropsci2014.09.0601.

Box, G.E.P.C.D.R. 1964. An Analysis of Transformation. Journal of the Royal Statistical Society 26: 211-252.

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

Branlard, G., J. Pierre and M. Rousset. 1992. Selection indices for quality evaluation in wheat breeding. Theoretical and Applied Genetics 84: 57-64.

Breiman, L. 2001. Random forests. Machine learning 45: 5-32.

Breseghello, F. and M.E. Sorrells. 2006. Association mapping of kernel size and milling quality in wheat (Triticum aestivum L.) cultivars. Genetics 172: 1165-1177. doi:10.1534/genetics.105.044586.

Crossa, J., L. Campos Gde, P. Perez, D. Gianola, J. Burgueno, J.L. Araus, et al. 2010. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186: 713-724. doi:10.1534/genetics.110.118521.

Crossa, J., P. Perez, J. Hickey, J. Burgueno, L. Ornella, J. Ceron-Rojas, et al. 2014. Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity (Edinb) 112: 48-60. doi:10.1038/hdy.2013.16.

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

Delcour, J. and R.C. Hoseney. 2010. Principles of cereal science and technology. status: published.

Endelman, J.B. 2011. Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. The Plant Genome Journal 4: 250. doi:10.3835/plantgenome2011.08.0024.

Friedman, J., T. Hastie and R. Tibshirani. 2009. glmnet: Lasso and elastic-net regularized generalized linear models. R package version 1.

Garg, M., H. Singh, H. Kaur and H.S. Dhaliwal. 2006. Genetic Control of High Protein Content and Its Association with Bread-Making Quality in Wheat. Journal of Plant Nutrition 29: 1357-1369. doi:10.1080/01904160600830134.

Gaynor, R.C. 2015. GSwGBS: an R package Genomic Selection with Genotyping-by-Sequencing. Genomic Selection for Kansas Wheat. K-State Research Exchange.

Giroux, M. and C. Morris. 1997. A glycine to serine change in puroindoline b is associated with wheat grain hardness and low levels of starch-surface friabilin. Theoretical and Applied Genetics 95: 857-864.

Guzmán, C., A.S. Medina-Larqué, G. Velu, H. González-Santoyo, R.P. Singh, J. Huerta-Espino, et al. 2014. Use of wheat genetic resources to develop biofortified wheat with enhanced grain zinc and iron concentrations and desirable processing quality. Journal of Cereal Science 60: 617-622. doi:10.1016/j.jcs.2014.07.006.

Guzmán, C., G. Posadas-Romano, N. Hernandez-Espinosa, A. Morales-Dorantes and R.J. Pena. 2015. A new standard water absorption criteria based on solvent retention capacity (SRC) to determine dough mixing properties, viscoelasticity, and bread-making quality. Journal of Cereal Science.

Heffner, E.L., J.-L. Jannink, H. Iwata, E. Souza and M.E. Sorrells. 2011. Genomic Selection Accuracy for Grain Quality Traits in Biparental Wheat Populations. Crop Science 51: 2597. doi:10.2135/cropsci2011.05.0253.

Heffner, E.L., J.-L. Jannink and M.E. Sorrells. 2011. Genomic selection accuracy using multifamily prediction models in a wheat breeding program. The Plant Genome 4: 65-75.

Heffner, E.L., M.E. Sorrells and J.-L. Jannink. 2009. Genomic Selection for Crop Improvement. Crop Science 49: 1. doi:10.2135/cropsci2008.08.0512.

International Wheat Genome Sequencing, C. 2014. A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. Science 345: 1251788. doi:10.1126/science.1251788.

Kuchel, H., P. Langridge, L. Mosionek, K. Williams and S. Jefferies. 2006. The genetic control of milling yield, dough rheology and baking quality of wheat. Theoretical and Applied Genetics 112: 1487-1495.

Liaw, A. and M. Wiener. 2002. Classification and regression by randomForest. R news 2: 18-22.

Lillemo, M., F. Chen, X. Xia, M. William, R.J. Peña, R. Trethowan, et al. 2006. Puroindoline grain hardness alleles in CIMMYT bread wheat germplasm. Journal of Cereal Science 44: 86-92.

Lorenz, A.J., S. Chao, F.G. Asoro, E.L. Heffner, T. Hayashi, H. Iwata, et al. 2011. Genomic Selection in Plant Breeding. 110: 77-123. doi:10.1016/b978-0-12-385531-2.00002-5.

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

Mevik, B.-H. and R. Wehrens. 2007. The pls package: principal component and partial least squares regression in R. Journal of Statistical Software 18: 1-24.

Payne, P.I. and G.J. Lawrence. 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: 29-35.

Payne, P.I., M.A. Nightingale, A.F. Krattiger and L.M. Holt. 1987. The relationship between HMW glutenin subunit composition and the bread-making quality of British-grown wheat varieties. Journal of the Science of Food and Agriculture 40: 51-65.

Peña, R. 2002. Wheat for bread and other foods. Bread wheat improvement and production. Food and Agriculture Organization of the United Nations. Rome: 483-542.

Peña, R.J., A. Amaya, S. Rajaram and A. Mujeeb-Kazi. 1990. Variation in quality characteristics associated with some spring 1B/1R translocation wheats. Journal of Cereal Science 12: 105-112.

Poland, J.A., P.J. Brown, M.E. Sorrells and J.L. Jannink. 2012. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS One 7: e32253. doi:10.1371/journal.pone.0032253.

Poland, J.A. and T.W. Rife. 2012. Genotyping-by-Sequencing for Plant Breeding and Genetics. The Plant Genome Journal 5: 92. doi:10.3835/plantgenome2012.05.0005.

R Development Core Team. 2014. R: A lanugauge and environment for statistical computing. R Foundation for Statistical Computing.

Raftery, A.E., M. Kárný and P. Ettler. 2010. Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill. Technometrics 52: 52-66.

Raftery, A.E., D. Madigan and J.A. Hoeting. 1997. Bayesian model averaging for linear regression models. Journal of the American Statistical Association 92: 179-191.

Rutkoski, J., J. Benson, Y. Jia, G. Brown-Guedira, J.-L. Jannink and M. Sorrells. 2012. Evaluation of Genomic Prediction Methods for Fusarium Head Blight Resistance in Wheat. The Plant Genome Journal 5: 51. doi:10.3835/plantgenome2012.02.0001.

Rutkoski, J.E., E.L. Heffner and M.E. Sorrells. 2010. Genomic selection for durable stem rust resistance in wheat. Euphytica 179: 161-173. doi:10.1007/s10681-010-0301-1.

Rutkoski, J.E., J.A. Poland, R.P. Singh, J. Huerta-Espino, S. Bhavani, H. Barbier, et al. 2014. Genomic Selection for Quantitative Adult Plant Stem Rust Resistance in Wheat. The Plant Genome 7: 0. doi:10.3835/plantgenome2014.02.0006.

Saghai-Maroof, M.A., K.M. Soliman, R.A. Jorgensen and R. Allard. 1984. Ribosomal DNA spacer-length polymorphisms in barley: Mendelian inheritance, chromosomal location, and population dynamics. Proceedings of the National Academy of Sciences 81: 8014-8018.

Terman, G., R. Ramig, A. Dreier and R. Olson. 1969. Yield-protein relationships in wheat grain, as affected by nitrogen and water. Agronomy Journal 61: 755-759.

Zheng, S., P.F. Byrne, G. Bai, X. Shan, S.D. Reid, S.D. Haley, et al. 2009. Association analysis reveals effects of wheat glutenin alleles and rye translocations on dough-mixing properties. Journal of cereal science 50: 283-290.

Chapter 3 - Mega-GWAS: Method for applying GWAS to an unbalanced breeding population.

Sarah D. Battenfield[1], J. Sheridan[2], L. Silva[3], C. Guzmán[4], K. Miclaus[3], R. Wolfinger[3], J. Peña[4], R. Singh[4], J. Poland[5], A. Fritz[1], & E. Jackson[2]

[1]Kansas State University, Department of Agronomy, Manhattan, KS, USA

[2]General Mills, Manhattan, KS, USA

[3]SAS, JMP-Genomics Division, Cary, NC, USA

[4]International Maize and Wheat Improvement Center, Mexico, D.F., Mexico

[5]Kansas State University, Department of Plant Pathology, Manhattan, KS, USA

Acronyms:

ALVPL- Alveograph P/L value; ALVW- Alveograph W value; CIMMYT- International Center of Maize and Wheat Improvement (Spanish acronym); FDR- False discovery rat; FLRPRO- Flour protein; FLRSDS- Flour sodium dodecyl sulfate sedimentation (test); FLRYLD- Flour yield; GBS- Genotyping-by-sequencing; GRNHRD- Grain hardness; GRNPRO- Grain protein; GWAS- Genome wide association study; IBD- Inbreeding by descent; K- Kinship [matrix]; LD- linkage disequilibrium; LOFVOL- Pup loaf volume; MIXTIM- Mixograph mix time; MP- Mixograph % torque at the integral of the midline peak; NIRS- Near infrared spectroscopy; PCA- Principal component analysis; Q- Population structure [matrix]; QTL- Quantitative trait locus/loci; SDS- Sodium dodecyl sulfate; TKW- Thousand kernel weight; TESTWT- Test weight

Abstract

Bread wheat (*Triticum aestivum* L.) is a staple cereal grain which must be processed into food products for human consumption. Many breeding programs have focused on increasing wheat yield, but processing and end-use quality must also be considered for feeding the rising population of the next century. End-use quality traits are expensive to test, and many require large amounts of seed, thus they cannot be tested until late in breeding programs. Thus, our goal is to have more reliable markers for utilization within breeding programs for quality traits. Here we describe a new method to identify marker-trait associations within a breeding program using a mega-genome wide association study. This method allowed for mapping in 4,095 individuals for all of the quantitative processing and end-use quality phenotypes and high- and low-molecular weight glutenins from advanced breeding lines of the CIMMYT bread wheat breeding program from 2009 – 2014. Using the mega- genome-wide analysis we have identified new marker-trait associations for grain protein and loaf volume, as well as known marker-trait associations for high- and low-molecular weight glutenins, which impact dough rheology. Many detected associations indicate the major allele in the breeding program as detrimental for the trait of interest, which indicates there is continued room for improvement. These results are promising for increasing processing and end-use quality as the alleles are already found within breeding populations, and can be altered through marker-assisted selection.

Globally, bread wheat (*Triticum aestivum* L.) per capita consumption is 65 kg per year, supplying nearly 16 g of protein daily on average for each person in the world (Faostat, 2013). This consumption pattern is increasing fastest in the world's least developed countries (Faostat, 2013), which also are predicted to have the largest increases in population over the next century (Gerland, *et al.*, 2014). Continuing to increase wheat yield potential through breeding remains vital, and in addition, will be important to tailor these wheat breeding lines to specific, local processing and end-use specifications.

Processing and end-use quality are paramount characteristics in a wheat cultivar for ensuring market acceptance. Wheat is commonly used in making leavened breads, flat breads, cookies, and crackers. There are also many wheat end-use products which have historically been region specific, such as a variety of dense breads, flat breads, steamed breads, and bread wheat noodles. Now, there are also growing trends for end-use products that fit well within the convenience and health foods models, such as refrigerated and frozen dough products, and more nutritious products made from whole wheat flour. All of these products are best made with wheat flours with specific attributes regarding grain color and hardness, protein content, and dough rheology which can be targeted through breeding (Peña, *et al.*, 2002).

Wheat breeding programs typically breed for regionally specific end-use product or products, and release cultivars with not only results of yield and disease resistance, but also end-use quality performance. In wheat breeding programs, wheat is assessed as grain, flour, dough, and end-use products to determine the genetic aptitude of wheat lines for processing and end-use quality (Peña, 2002). The tests of grain can be done on a small scale, quickly, and cheaply. Several of these tests can be conducted using near infrared spectroscopy (NIRS) (AACC, 2000), making them possible to implement in high throughput programs. However, dough rheology and end-use tests require large quantities of grain for milling into flour, are more costly, and more time consuming, which implies these tests cannot be conducted until later in the breeding program.

Wheat grain is assessed for basic characteristics which impact grain marketing. Wheat is separated into marketing groups based on growth habit of the plant, grain hardness and color (hard white, hard red winter, soft white, etc.). In general, hard wheat is preferred for pan bread, whereas soft wheat is preferred for cookie and cake. However, several cultivars, possibly from differing regions, will be mixed in industrial mills to ensure consistency with end-use quality

specifications over time. In local markets in the developing world, the visual characteristics of a cultivar, such as kernel size, vitreousness, which often is an indicator of hardness (Delcour and Hoseney, 2010), and color, are extremely important as much of this wheat will be milled and used in the home.

The viscoelastic properties of wheat allow for rise and retention of gas, while still retaining shape and connectivity. Dough rheology tests involve mixing flour with water to make dough, and then measuring the viscoelastic properties of strength, elasticity, and tolerance. End-use testing requires making the target product to assess final internal and external appearance, as well as size, of the end-use product when the dough is optimally mixed. Both dough rheology and end-use tests are time consuming, costly, and require large amounts of grain to conduct. However, combinations of tests of grain hardness, protein content, and dough rheology are necessary to predict best suited end-use products for a specific wheat line (Peña, 2002).

Since these traits cannot be assessed until late in the breeding program there is interest to use marker assisted selection in aiding breeding for end-use quality. Typical quality markers used in the CIMMYT bread wheat breeding program are for grain protein content on 6B (Uauy, *et al.*, 2006), grain hardness on 5D (Gautier, *et al.*, 1994), and high and low molecular weight glutenins on 1A, 1B, and 1D (Liu, *et al.*, 2008, Ragupathy, *et al.*, 2008, Wang, *et al.*, 2009, Wang, *et al.*, 2010). However, these markers are mostly assessed on parents of crosses, and rarely within segregating or testing material.

Gene mapping in plants has historically been conducted using biparental mapping to detect genetic makers in linkage disequilibrium (LD) with quantitative trait loci (QTL). Recently, mapping with other structured populations has become more common, such as nested association mapping (Yu, *et al.*, 2008) or multi-parent advanced generation inter-cross (Cavanagh, *et al.*, 2008) populations. These approaches attempt to decrease LD distance and genetic background effect. However, the structured populations are still limited in size, have ascertainment bias due to population founders, and the genetic background effect may not be thoroughly detected.

GWAS is a tool commonly used in human genetics where making structured populations cannot ethically be conducted to map traits of interest. In this method a population of individuals is used to associate markers to phenotypes. The use a large population reduces the ascertainment bias and genetic background effect issues, and historical recombination through many generations decreases LD between the detected marker and the causative genomic region.

Spurious associations may be found due to relatedness of individuals. However population structure and relatedness, Q and K, respectively, can be accounted for GWAS (Yu, *et al.*, 2006) to reduce these limitations. GWAS results can further be strengthened by results confirmed over several years or in several studies through mega-analysis or meta-analysis, respectively (Begum, *et al.*, 2012). These powerful meta- and mega-analyses have not previously been shown in detecting QTL in wheat breeding programs.

GWAS has previously been utilized several times to detect yield and disease resistance in wheat (Crossa, *et al.*, 2007, Liu, *et al.*, 2010, Neumann, *et al.*, 2010, Edae, *et al.*, 2014, Juliana, *et al.*, 2015), but has not been utilized thoroughly for wheat processing and end-use quality. Three of these studies included thousand kernel weight (TKW), test weight, or protein content in their analyses (Liu, *et al.*, 2010, Mir, *et al.*, 2012, Edae, *et al.*, 2014). However, few studies have focused solely on processing and end-use quality traits.

Two previous studies have examined association mapping for quality in soft wheat lines (Breseghello and Sorrells, 2006, Reif, *et al.*, 2011), and one has investigated a core collection of hexaploid wheat, likely containing both hard and soft types (Bordes, *et al.*, 2011). Breseghello and Sorrells (2006) focused solely on chromosomes 2D, 5A, and 5B based on prior information to detect associations for kernel size and milling quality not related to glutenins. Reif, et al., (2011) investigated TKW, test weight, protein content, SDS-sedimentation volume, and starch content in a genome-wide scan of 207 soft wheat breeding lines. Both of these studies utilized low-density simple sequence repeat (SSR) markers, whereas now, higher density marker platforms are available which could differentiate more quantitative traits. The core collection study (Bordes, *et al.*, 2011) used a high-density genome wide scan of 372 core diversity lines in their study. They found marker-trait associations for grain protein content, grain hardness, viscosity, flour color, and Mixograph parameters for dough consistency, strength, elasticity, and optimal mix time.

None of the previous studies have shown associations for Alveograph dough rheology or end-use pup loaf volume. Additionally, GWAS has not previously been conducted in tandem with high and low molecular weight glutenins, which have long been deemed biologically relevant to the final outcomes in wheat quality (Payne, *et al.*, 1987). Finally, these studies used panels of germplasm designed to maximize the potential of the association mapping, whereas here we present GWAS using data from a breeding program, which can immediately continue to

be applied for further improvement without new linkage drag. The objectives of this study were to conduct mega genome-wide association mapping for all quantitative processing and end-use quality traits in the CIMMYT bread wheat breeding program to identify SNPs associated with these traits which could be used in breeding programs.

Materials and Methods

Wheat lines used in association mapping for wheat quality were materials in the preliminary and advanced yield trials of the CIMMYT bread wheat breeding program between 2009 and 2014. All wheat lines were grown in Ciudad Obregon, Sonora, Mexico, in at least one year, under full irrigation. Site-years were treated individually for the QK-Mixed model GWAS and were considered eligible for analysis if there were greater than 200 entries tested. Best materials for agronomic and quality traits were advanced in the breeding program and grown and tested a second year under full irrigation. The full set for association mapping, n=4,095, included both replicated and non-replicated entries to increase the size of the association mapping panel and show validity of the mega-GWAS method.

Phenotype Assessment

Continuous quality phenotypes for thousand kernel weight, test weight, grain hardness, flour yield, grain protein, flour protein, SDS-sedimentation, Mixograph mix time and torque, Alveograph W and PL^{-1}, and pup loaf volume, were measured according to AACC (2000) with water absorption modifications (Guzmán, *et al.*, 2015) as in Battenfield et al. (2015, chapter 2). High and low molecular weight glutenins were assessed in a subset 952 lines using SDS-PAGE (Gupta and Shepherd, 1990, Singh, *et al.*, 1991). Glutenins were recorded as binary for the presence or absence of each allele. Only glutenin classes with greater than 5% frequency were analyzed, removing rare alleles, which could be better screened in a more targeted panel.

Genotype assessment

Tissue collection for DNA (Saghai-Maroof, *et al.*, 1984) and GBS protocol (Poland, *et al.*, 2012) was conducted using TASSEL 5 v2 pipeline. GBS tags were aligned to the soft-masked *Triticum aestivum* IWGSC genome assembly version 2.25 (IWGSC, 2014) and indexed using Bowtie 2 (Langmead and Salzberg, 2012). SNPs were identified by chromosome and position number and numerically coded for major, minor, heterozygous, or missing classes.

SNPs were then curated in JMP-Genomics 7.1 (SAS, Cary, NC) to maintain maximum data accuracy with the large amounts of missing data found using genotyping-by-sequencing. Individuals with greater than 35% missing data were removed from further analysis. Markers with greater than 25% missing data, greater than 20% percent heterozygous, or less than 5% minor allele frequency were also removed. Polymorphism information content was calculated for each marker. Linkage disequilibrium (LD) was plotted and markers were removed which showed excessive LD over long genomic distances.

The final annotated and curated set of SNPs was aligned with POPSEQ (Mascher, *et al.*, 2013). POPSEQ allowed determination of approximate cM position of the markers in reference to the Synthetic W7984 x Opata M85 recombinant inbred line mapping population (Sorrells, *et al.*, 2011, Chapman, *et al.*, 2015). These reference positions were used to display significant marker-trait associations in mapped positions more consistent with previously detected loci.

Data analysis

Population and relationship structure were investigated and added to the association mapping analysis as covariates to help prevent spurious associations (Yu, *et al.*, 2006). Principal component analysis was conducted using "PCA for population stratification" in JMP-Genomics 7.1 (SAS, Cary, NC) to create the population structure matrix, Q. Relative kinship between individuals was also analyzed using JMP-Genomics. This was conducted using identity by descent method in the "Relationship matrix" program with false discovery rate (FDR) multiple testing correction, resulting in the K matrix (Benjamini and Hochberg, 1995).

Association mapping for continuous phenotypes was conducted using a "Q-K mixed model" in JMP-Genomics 7.1 (SAS, Cary, NC) for each site-year with false discovery rate (FDR) multiple testing correction applied (Benjamini and Hochberg, 1995, Yu, *et al.*, 2006). The site-year marker-trait associations were combined using "GWAS meta-analysis" using an inverse-variance, fixed effects model where each site-year was treated as a fixed effect (Begum, *et al.*, 2012). Multi-year marker-trait associations were corrected again for multiple testing using FDR (Benjamini and Hochberg, 1995). Probabilities were transformed using –log10(*p*), and shown at the $\alpha < 0.05$, 0.01, and 0.001 levels, but reported as significant at $\alpha < 0.001$.

Association mapping for high and low molecular weight glutenins was conducted with all 952 samples pooled across all years since these were measured simply on presence or absence of the allele in question. PCA and IBD were determined as in the continuous traits. In order to

efficiently map these binary glutenins, a K-matrix compression was conducted for each trait separately in JMP-Genomic 7. The GWAS was then fit using 3 PCAs and the compressed K-matrix for the trait of interest. Less power was available in the binary screening, thus α was set more stringently to 5×10^{-8}.

Results

Genotypes

A total of 1,625 SNP markers were identified to have high quality and be acceptable for use in GWAS. Mean SNP counts were 93, 116, and 23, for the A-, B-, and D-series, respectively with Bowtie alignment (Figure 1). 906 of the 1,625 SNPs aligned with cM positions from PopSeq, with mean 47, 68, and 14 SNPs per A-, B-, and D-series chromosomes represented. Still, all 1,625 SNPs were used in mapping marker-trait associations. SNP markers were well distributed across the chromosomes, except in the weakly represented 3D and 4D chromosomes.

Population and kinship structure

Population structure and relationship structure were present in these data (Figure 2). PCA of the genotype matrix demonstrated that 4 principal components explained the largest portions of the variance before approaching a plateau (Figure 2). There was also significant structure in the genetic relationship based on probability of IBD. There was no correlation between the PCA and IBD principal components (Table 1). Therefore, 4 principal components and the IBD matrix were used as Q and K matrix covariates, respectively, in GWAS.

Significant associations by year, significant associations across site-years

Significant marker-trait associations were found across years on all chromosomes except 3D and 4D. 127 significant marker-trait associations were found over the all traits. These collectively represent 77 unique SNPs (Figs. 3-14; Table 2). The loci in the same region of a chromosome with more than one collective effect will be referred to as hotspots. Position of glutenins found by binary Q-K mixed model GWAS were only used in reference to the other marker-trait associations.

Significant marker-trait association hotspots were found in the regions of high molecular weight glutenins for dough rheology traits and grain hardness. Markers associated with the differentiation of *Glu-D1*- 2+12 or 5+10 had the largest effect on dough rheology traits. The responses at this locus for ALVW, ALVPL, MIXTIM, and MP, demonstrated that the major

allele in the CIMMYT program, *Glu-D1* 5+10, was highly beneficial for these traits. Associations in other high and low molecular weight glutenin regions demonstrated that most prominent alleles at these loci had mix impacts on the protein, dough rheology, and loaf volume traits (Table 2; Figs 3-14).

Largest marker-trait associations for grain and flour protein were found on chromosome 6A at 62-65 and 70 cM, which were also hotspots with loaf volume (Figs 4, 5, 6, 9, and 10). The major alleles found in the 62-65 cM region had a negative impact on protein content, thus also negative associations with dough strength and loaf volume. The major allele found around 70 cM had a positive impact on protein concentration and loaf volume (Table 2). Additionally, smaller impact QTL for protein concentration were found on 6B.

Wheat bread loaf volume was most significantly impacted by QTL on chromosome 7A (Figure 10). This significant QTL is located near 83 cM (Table 2). This QTL is expressed larger in final loaf volume, but appears to be first impacting extensibility through ALVPL (Table 2). This hotspot is found as the minor allele in the CIMMYT breeding program, thus selection for this QTL could improve bread making performance.

Conclusions

Here we present a new application of a mega-study of genome wide association mapping directly to unreplicated data from a breeding program. The results agree with empirical expectations, and thus seems this method is a good fit for finding alleles present in a breeding program that impact trait outcomes and can be immediately selected in the breeding program. We also identified two new regions of interest on 6A and 7A for grain protein concentration, dough extensibility, and loaf volume, which have some recent empirical evidence and will be more thoroughly investigated.

The viscoelastic properties of wheat mostly originate from the storage proteins glutenins and gliadins (Delcour and Hoseney, 2010). Wheat high and low molecular weight glutenins have been highly studied and found responsible for the elasticity and resistance to extension properties of wheat dough (Payne, *et al.*, 1987, Zheng, *et al.*, 2009). The multiallelic glutenin profile (Payne and Lawrence, 1983, Payne, *et al.*, 1987) in the high molecular weight glutenin alleles *GluA-1*, *GluB-1*, and *GluD-1* on the long arms of 1A, 1B, and 1D, and low molecular weight glutenin alleles *GluA-3*, *GluB-3*, and *GluD-3* on the short arms of 1A, 1B, and 1D has been found to be

related to end-use properties (Branlard, *et al.*, 1992). Gliadins, *GliA-1*, *GliB-1*, *GliD-1 GliA-2*, *GliB-2*, and *GliD-2*, on chromosomes 1A, 1B, 1D, 6A, 6B, and 6D, respectively (Payne and Lawrence, 1983, Payne, *et al.*, 1987), are responsible for the cohesive properties of wheat dough, which allow it to rise and retain gas (Delcour and Hoseney, 2010). We report significant associations for Alveograph, Mixograph, flour SDS, grain hardness, flour yield, and loaf volume relating to glutenins. These associations collectively make over half of the significant hits in this data set, and have the most significant results found in dough rheology and SDS traits.

Grain protein content is highly correlated with dough strength (Borghi, *et al.*, 1995, Blandino, *et al.*, 2015), loaf volume in pan breads (Bushuk, 1997), and overall baking of hard wheats quality (Garg, *et al.*, 2006). However, grain protein content is often negatively correlated with yield and is highly impacted by environment and agronomic management (Terman, *et al.*, 1969, Borghi, *et al.*, 1995, Blandino, *et al.*, 2015). This inverse relationship between grain protein and yield can be explained through TKW as grain fill over time is mostly starch deposition. While yield is most often the metric considered for cost of wheat, not quality, thus yield is economically favored over quality in many wheat production markets. However, both TKW and grain protein concentration been shown to be increasing over time (Cox, *et al.*, 1989, Aisawi, *et al.*, 2015).

One gene found to increase grain protein content is *Gpc-B1* on the short arm of chromosome 6B (Uauy, *et al.*, 2006). This gene was first characterized in *Triticum turgidum* L. *ssp durum*, but must be specifically incorporated into bread wheat. Additionally there are homeologues for this gene on 6A and 6D, and paralogues on 2A, 2B, and 2D (Cormier, *et al.*, 2015). Here, in agreement with Cormier, *et al.* (2015), we find that the 6A homeologue of *Gpc-B1*, *Gpc-A1*, or *NAM*, as it is named in other species, controls the largest portion of the variance for grain protein content. *Gpc-A1* is particularly interesting for breeding impact as the two primary alleles control the tradeoff between senescence timing, TKW, and grain protein concentration (Cormier, *et al.*, 2015). We find that both the *a* and *d* alleles discussed in Cormier, *et al.* (2015) appear to be present in this population, leading one to believe that there is reason to select both haplotypes in different breeding situations.

Hardness, or wheat endosperm texture, appears to be mostly controlled by *Ha* hardness genes the short arm of chromosome 5D. Hard wheat contains starch granules which are more tightly attached to the protein matrix, thus requiring more force in milling and damaging more

starch than in soft endosperm wheat (Giroux and Morris, 1997). Previous results have shown very high proportions of haplotype pin with pin in CIMMYT breeding material (Lillemo, *et al.*, 2006). We believe the overwhelming majority for one haplotype of *pin* alleles leads to the lack of their detection in this study. Beyond that, the remainder of genetic variation seems to be attributable to protein concentration and glutenin alleles.

Final loaf volume is complex trait as it is impacted by both type and amount of storage proteins present in the flour. Here, our results agree that loaf volume is impacted by the glutenin profile on the 1 series chromosomes, as well as the protein concentration QTL on 6A. In addition, there is a hotspot for dough extensibility and loaf volume found on chromosome 7A. We believe this could be an impact of the recently discovered wheat bread making, *wbm*, gene identified through RNA-seq post anthesis (Furtado, *et al.*, 2015). The favorable allele at this locus is currently minor within the breeding program, so selection for this allele could further increase dough extensibility within this breeding program.

Acknowledgements

Battenfield's research was supported through a Monsanto Beachell-Borlaug International Scholars Program fellowship. Funding for this project was provided by US Agency for International Development Feed the Future Initiative (USAID Cooperative Agreement No. AID-OAA-A-13-0005) and the Bill & Melinda Gates Foundation through a grant to Cornell University for "Genomic Selection: The next frontier for rapid gains in maize and wheat improvement." Support for phenotyping of quality traits was provided by CGIAR CRP WHEAT, Durable Rust Resistance Project, and Fondo Sectorial SAGARPA-CONACYT (No. 146788 – "Sistema de mejoramiento genético para generar variedades resistentes a royas, de alto rendimiento y alta calidad para una producción sustentable en México de trigo") of the Mexican government. The computing for this project was performed on the Beocat Research Cluster at Kansas State University, which is funded in part by NSF grants CNS-1006860, EPS-1006860, and EPS-0919443. This work represents contribution number 16-037-J from the Kansas Agricultural Experiment Station.

References

AACC. 2000. Approved Methods of the American Association of Cereal ChemistsAmer Assn of Cereal Chemists.

Aisawi, K.A.B., M.P. Reynolds, R.P. Singh and M.J. Foulkes. 2015. The Physiological Basis of the Genetic Progress in Yield Potential of CIMMYT Spring Wheat Cultivars from 1966 to 2009. Crop Science 55: 1749. doi:10.2135/cropsci2014.09.0601.

Begum, F., D. Ghosh, G.C. Tseng and E. Feingold. 2012. Comprehensive literature review and statistical considerations for GWAS meta-analysis. Nucleic acids research: gkr1255.

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

Blandino, M., F. Marinaccio, P. Vaccino and A. Reyneri. 2015. Nitrogen Fertilization Strategies Suitable to Achieve the Quality Requirements of Wheat for Biscuit Production. Agronomy Journal 107: 1584-1594. doi:10.2134/agronj14.0627.

Bordes, J., C. Ravel, J. Le Gouis, A. Lapierre, G. Charmet and F. Balfourier. 2011. Use of a global wheat core collection for association analysis of flour and dough quality traits. Journal of Cereal Science 54: 137-147. doi:10.1016/j.jcs.2011.03.004.

Borghi, B., G. Giordani, M. Corbellini, P. Vaccino, M. Guermandi and G. Toderi. 1995. Influence of crop rotation, manure and fertilizers on bread making quality of wheat (Triticum aestivum L.). European journal of agronomy 4: 37-45.

Branlard, G., J. Pierre and M. Rousset. 1992. Selection indices for quality evaluation in wheat breeding. Theoretical and Applied Genetics 84: 57-64.

Breseghello, F. and M.E. Sorrells. 2006. Association mapping of kernel size and milling quality in wheat (Triticum aestivum L.) cultivars. Genetics 172: 1165-1177. doi:10.1534/genetics.105.044586.

Bushuk, W. 1997. Wheat breeding for end-product use. Wheat: Prospects for global improvement. Springer. p. 203-211.

Cavanagh, C., M. Morell, I. Mackay and W. Powell. 2008. From mutations to MAGIC: resources for gene discovery, validation and delivery in crop plants. Curr Opin Plant Biol 11: 215-221. doi:10.1016/j.pbi.2008.01.002.

Cormier, F., M. Throude, C. Ravel, J. Gouis, M. Leveugle, S. Lafarge, et al. 2015. Detection of NAM-A1 Natural Variants in Bread Wheat Reveals Differences in Haplotype Distribution between a Worldwide Core Collection and European Elite Germplasm. Agronomy 5: 143-151. doi:10.3390/agronomy5020143.

Cox, T., M. Shogren, R. Sears, T. Martin and L. Bolte. 1989. Genetic improvement in milling and baking quality of hard red winter wheat cultivars, 1919 to 1988. Crop Science 29: 626-631.

Cox, T., J. Shroyer, L. Ben-Hui, R. Sears and T. Martin. 1988. Genetic improvement in agronomic traits of hard red winter wheat cultivars 1919 to 1987. Crop Science 28: 756-760.

Crossa, J., J. Burgueno, S. Dreisigacker, M. Vargas, S.A. Herrera-Foessel, M. Lillemo, et al. 2007. Association analysis of historical bread wheat germplasm using additive genetic covariance of relatives and population structure. Genetics 177: 1889-1913.

Delcour, J. and R.C. Hoseney. 2010. Principles of cereal science and technology. status: published.

Edae, E.A., P.F. Byrne, S.D. Haley, M.S. Lopes and M.P. Reynolds. 2014. Genome-wide association mapping of yield and yield components of spring wheat under contrasting moisture regimes. Theor Appl Genet 127: 791-807. doi:10.1007/s00122-013-2257-8.

Faostat, F. 2013. Statistical Databases. Food and Agriculture Organization of the United Nations.

Furtado, A., P. Bundock, P. Banks, G. Fox, X. Yin and R. Henry. 2015. A novel highly differentially expressed gene in wheat endosperm associated with bread quality. Scientific reports 5.

Garg, M., H. Singh, H. Kaur and H.S. Dhaliwal. 2006. Genetic Control of High Protein Content and Its Association with Bread-Making Quality in Wheat. Journal of Plant Nutrition 29: 1357-1369. doi:10.1080/01904160600830134.

Gautier, M.-F., M.-E. Aleman, A. Guirao, D. Marion and P. Joudrier. 1994. Triticum aestivum puroindolines, two basic cystine-rich seed proteins: cDNA sequence analysis and developmental gene expression. Plant molecular biology 25: 43-57.

Gerland, P., A.E. Raftery, H. Ševčíková, N. Li, D. Gu, T. Spoorenberg, et al. 2014. World population stabilization unlikely this century. Science 346: 234-237.

Giroux, M. and C. Morris. 1997. A glycine to serine change in puroindoline b is associated with wheat grain hardness and low levels of starch-surface friabilin. Theoretical and Applied Genetics 95: 857-864.

Gupta, R. and K. Shepherd. 1990. Two-step one-dimensional SDS-PAGE analysis of LMW subunits of glutelin. Theoretical and Applied Genetics 80: 65-74.

Guzmán, C., G. Posadas-Romano, N. Hernandez-Espinosa, A. Morales-Dorantes and R.J. Pena. 2015. A new standard water absorption criteria based on solvent retention capacity (SRC) to determine dough mixing properties, viscoelasticity, and bread-making quality. Journal of Cereal Science.

Himi, E. and K. Noda. 2005. Red grain colour gene (R) of wheat is a Myb-type transcription factor. Euphytica 143: 239-242.

IWGSC. 2014. A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. Science 345: 1251788. doi:10.1126/science.1251788.

Juliana, P., J.E. Rutkoski, J.A. Poland, R.P. Singh, S. Murugasamy, S. Natesan, et al. 2015. Genome-Wide Association Mapping for Leaf Tip Necrosis and Pseudo-black Chaff in Relation to Durable Rust Resistance in Wheat. The Plant Genome. doi:10.3835/plantgenome2015.01.0002.

Langmead, B. and S.L. Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. Nature methods 9: 357-359.

Lillemo, M., F. Chen, X. Xia, M. William, R.J. Peña, R. Trethowan, et al. 2006. Puroindoline grain hardness alleles in CIMMYT bread wheat germplasm. Journal of Cereal Science 44: 86-92.

Liu, L., L. Wang, J. Yao, Y. Zheng and C. Zhao. 2010. Association mapping of six agronomic traits on chromosome 4A of wheat (Triticum aestivum L.). Molecular Plant Breeding 1.

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

Mascher, M., G.J. Muehlbauer, D.S. Rokhsar, J. Chapman, J. Schmutz, K. Barry, et al. 2013. Anchoring and ordering NGS contig assemblies by population sequencing (POPSEQ). The Plant Journal 76: 718-727.

Mir, R.R., N. Kumar, V. Jaiswal, N. Girdharwal, M. Prasad, H.S. Balyan, et al. 2012. Genetic dissection of grain weight in bread wheat through quantitative trait locus interval and association mapping. Molecular Breeding 29: 963-972. doi:10.1007/s11032-011-9693-4.

Neumann, K., B. Kobiljski, S. Denčić, R.K. Varshney and A. Börner. 2010. Genome-wide association mapping: a case study in bread wheat (Triticum aestivum L.). Molecular Breeding 27: 37-58. doi:10.1007/s11032-010-9411-7.

Payne, P.I. and G.J. Lawrence. 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: 29-35.

Payne, P.I., M.A. Nightingale, A.F. Krattiger and L.M. Holt. 1987. The relationship between HMW glutenin subunit composition and the bread-making quality of British-grown wheat varieties. Journal of the Science of Food and Agriculture 40: 51-65.

Peña, R. 2002. Wheat for bread and other foods. Bread wheat improvement and production. Food and Agriculture Organization of the United Nations. Rome: 483-542.

Peña, R.J., R. Trethowan, W.H. Pfeiffer and M.V. Ginkel. 2002. Quality (End-Use) Improvement in Wheat. Journal of Crop Production 5: 1-37. doi:10.1300/J144v05n01_02.

Poland, J.A., P.J. Brown, M.E. Sorrells and J.L. Jannink. 2012. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS One 7: e32253. doi:10.1371/journal.pone.0032253.

Ragupathy, R., H.A. Naeem, E. Reimer, O.M. Lukow, H.D. Sapirstein and S. Cloutier. 2008. Evolutionary origin of the segmental duplication encompassing the wheat GLU-B1 locus encoding the overexpressed Bx7 (Bx7OE) high molecular weight glutenin subunit. Theor Appl Genet 116: 283-296. doi:10.1007/s00122-007-0666-2.

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

Saghai-Maroof, M.A., K.M. Soliman, R.A. Jorgensen and R. Allard. 1984. Ribosomal DNA spacer-length polymorphisms in barley: Mendelian inheritance, chromosomal location, and population dynamics. Proceedings of the National Academy of Sciences 81: 8014-8018.

Singh, N., K. Shepherd and G. Cornish. 1991. A simplified SDS—PAGE procedure for separating LMW subunits of glutenin. Journal of Cereal Science 14: 203-208.

Sorrells, M.E., J.P. Gustafson, D. Somers, S. Chao, D. Benscher, G. Guedira-Brown, et al. 2011. Reconstruction of the synthetic W7984× Opata M85 wheat reference population. Genome 54: 875-882.

Terman, G., R. Ramig, A. Dreier and R. Olson. 1969. Yield-protein relationships in wheat grain, as affected by nitrogen and water. Agronomy Journal 61: 755-759.

Uauy, C., A. Distelfeld, T. Fahima, A. Blechl and J. Dubcovsky. 2006. A NAC gene regulating senescence improves grain protein, zinc, and iron content in wheat. Science 314: 1298-1301.

Wang, L., G. Li, R.J. Peña, X. Xia and Z. He. 2010. Development of STS markers and establishment of multiplex PCR for Glu-A3 alleles in common wheat (Triticum aestivum L.). Journal of Cereal Science 51: 305-312. doi:10.1016/j.jcs.2010.01.005.

Wang, L., X. Zhao, Z. He, W. Ma, R. Appels, R. Peña, et al. 2009. Characterization of low-molecular-weight glutenin subunit Glu-B3 genes and development of STS markers in common wheat (Triticum aestivum L.). Theoretical and Applied Genetics 118: 525-539.

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

Yu, J., G. Pressoir, W.H. Briggs, I. Vroh Bi, M. Yamasaki, J.F. Doebley, et al. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38: 203-208. doi:10.1038/ng1702.

Zheng, S., P.F. Byrne, G. Bai, X. Shan, S.D. Reid, S.D. Haley, et al. 2009. Association analysis reveals effects of wheat glutenin alleles and rye translocations on dough-mixing properties. Journal of cereal science 50: 283-290.

Figures

Figure 3-1: Marker distribution by counts for all chromosomes

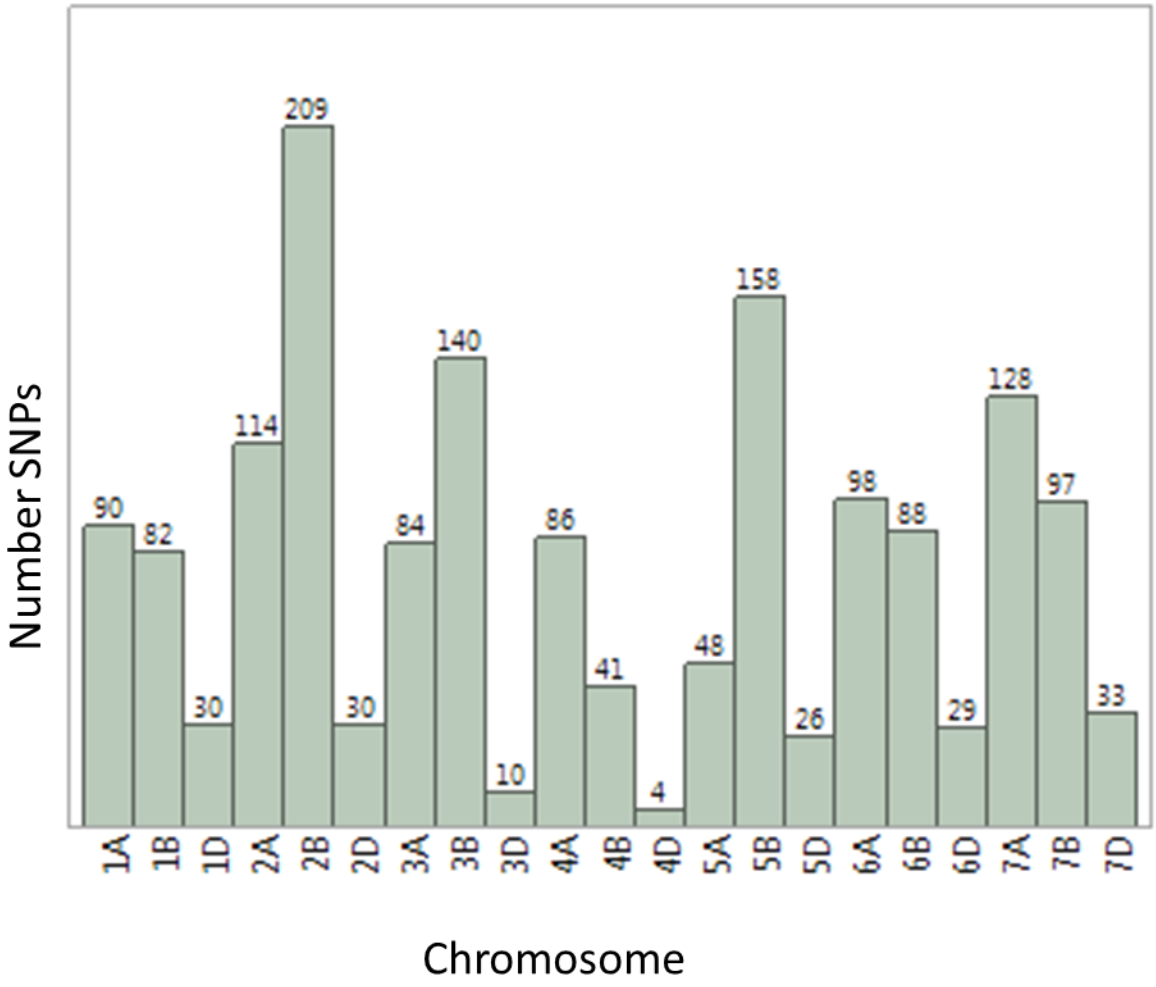


Figure 3-2: Popularion structure demonstrated by Principal Coordinate Analysis (PCA) and Inbreeding by Descent (IBD). Where PCA is on the left and IBD is on the right. a) and b) show the three-dimensional representation while c) and d)show the two-dimension representation of each component of the population structure explained by PCA and IBD, respectively. Plots e) and f) show the scree plots of the explained variance by each component for PCA and IBD, respectively.

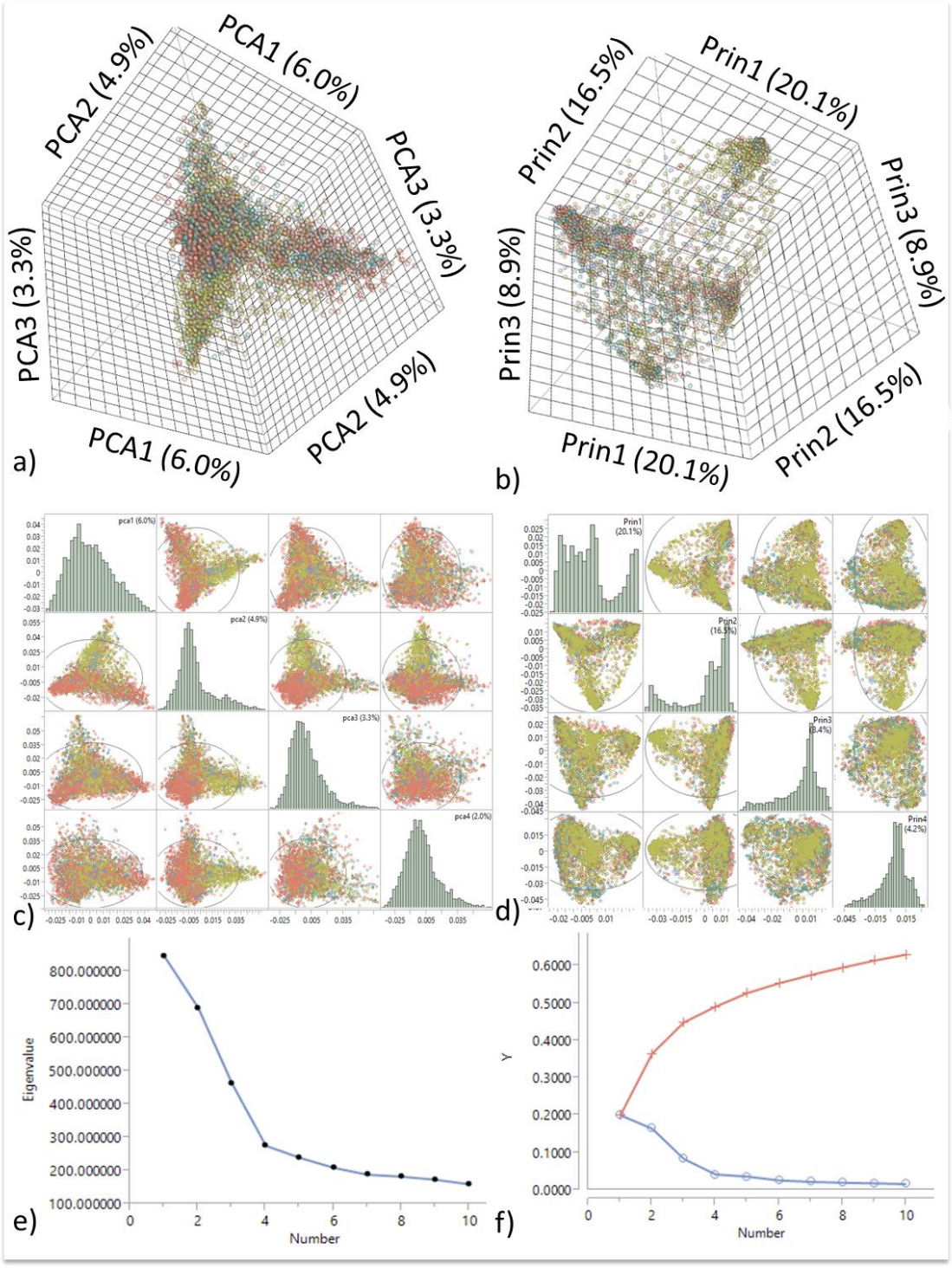


Figure 3-3: Manhattan plot of ALVPL

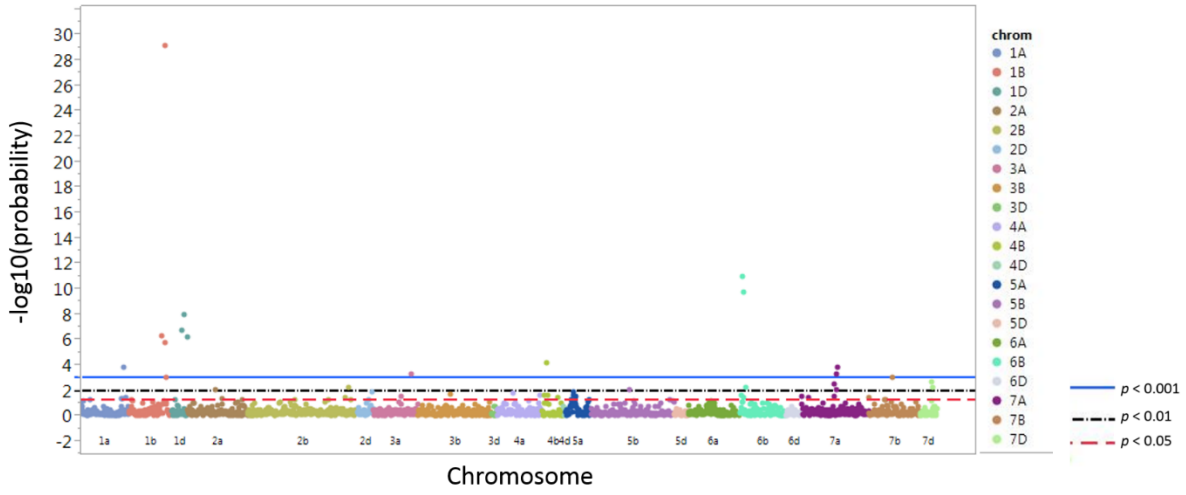


Figure 3-4: Manhattan plot of ALVW

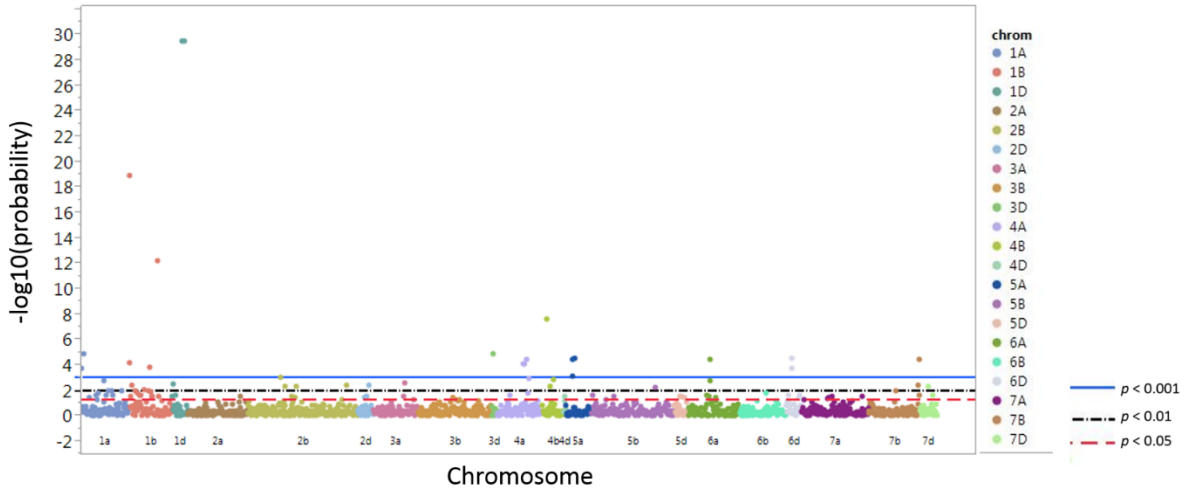


Figure 3-5: Manhattan plot of FLRPRO

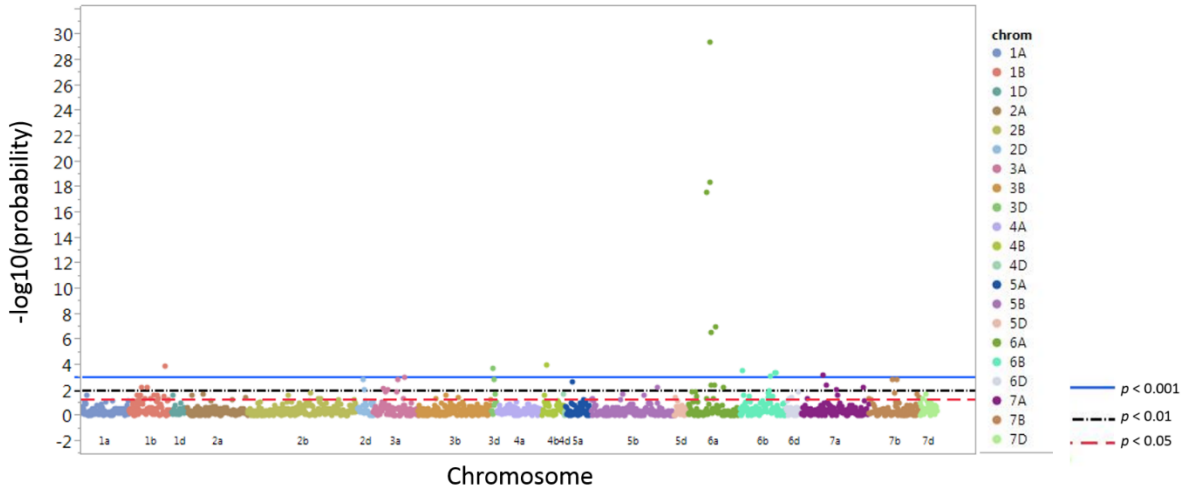


Figure 3-6: Manhattan plot of FLRSDS

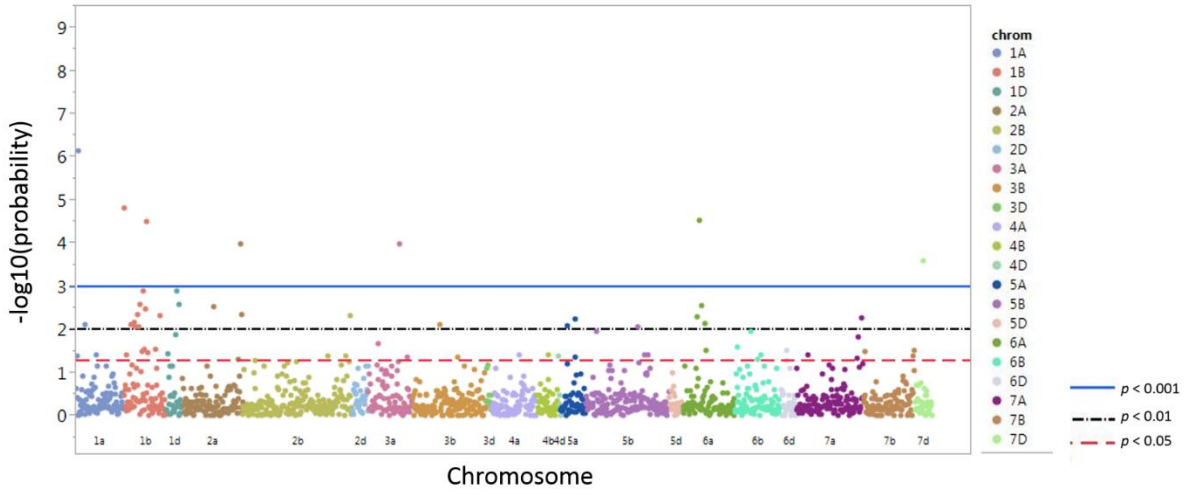


Figure 3-7: Manhattan plot of FLRYLD

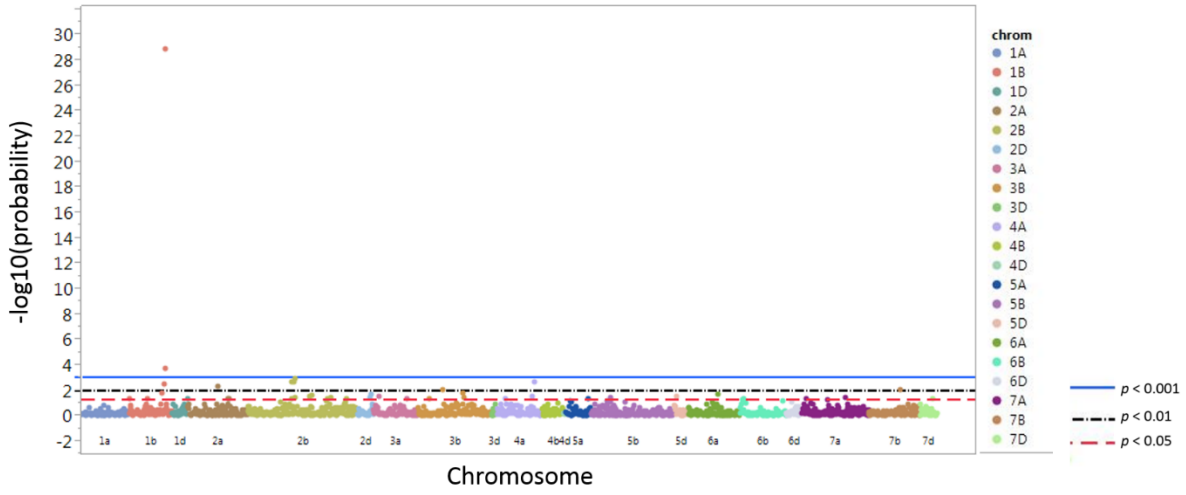


Figure 3-8: Manhattan plot of GRNHRD

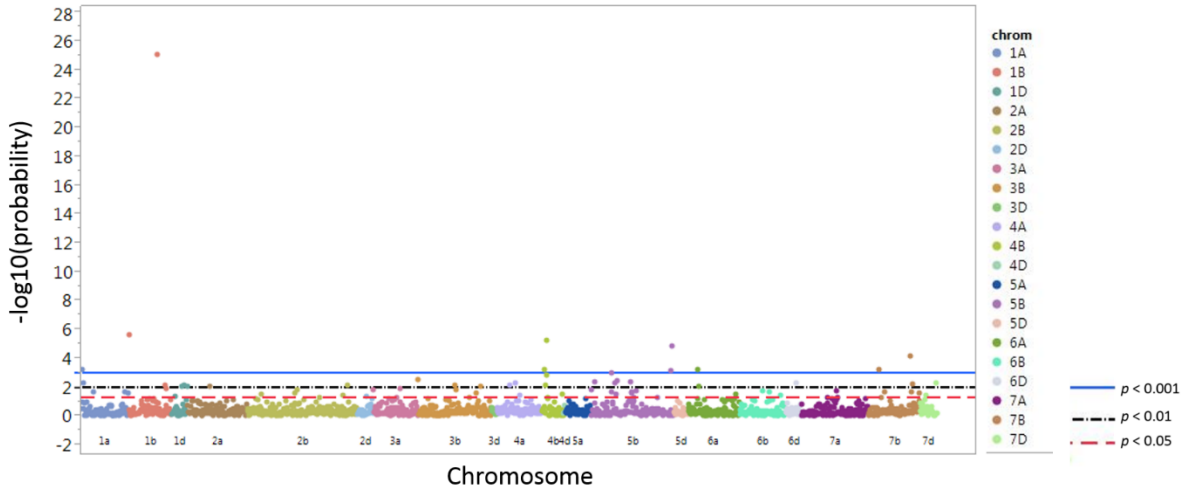


Figure 3-9: Manhattan plot of GRNPRO

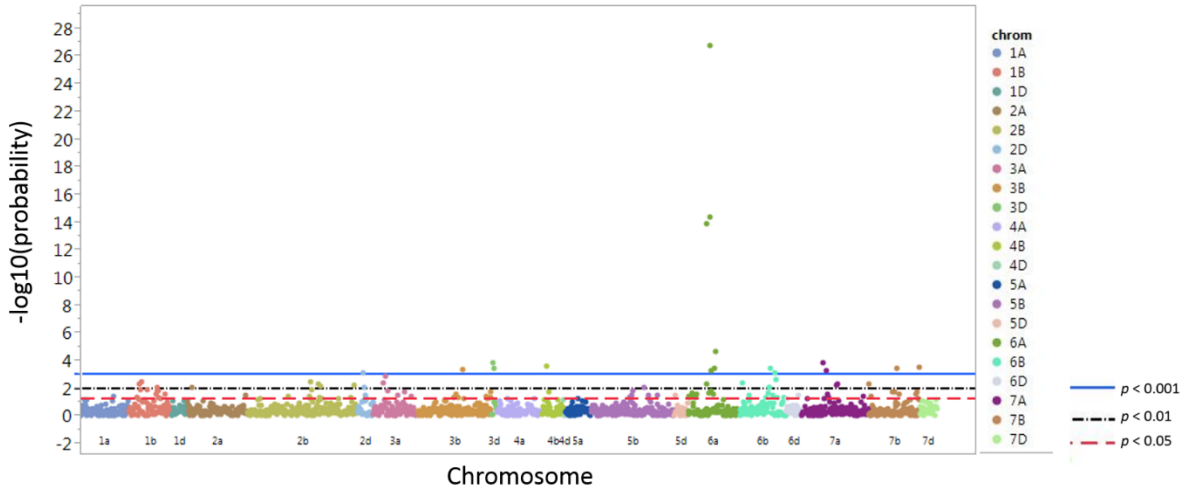


Figure 3-10: Manhattan plot of LOFVOL

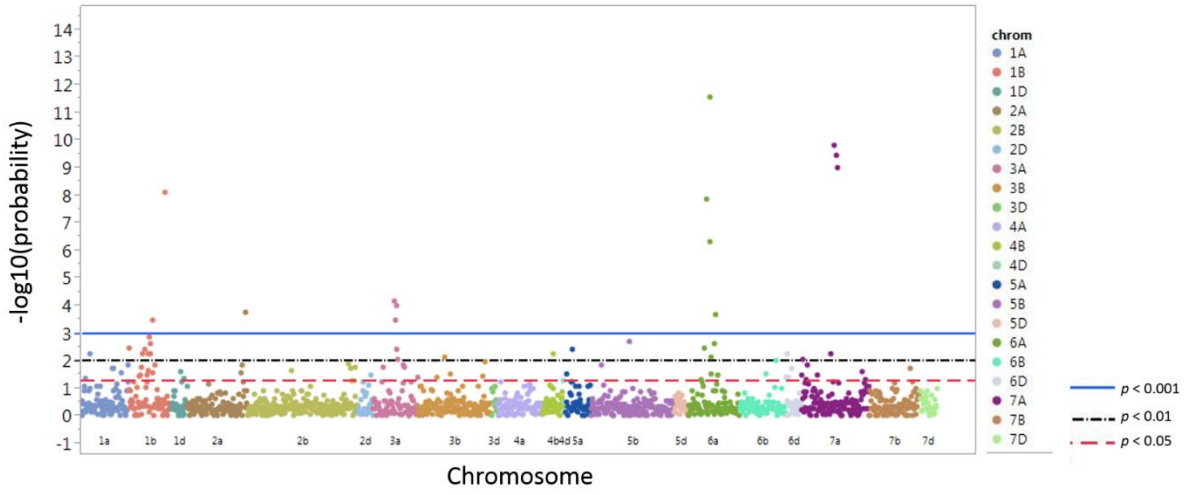


Figure 3-11: Manhattan plot of MIXTIM

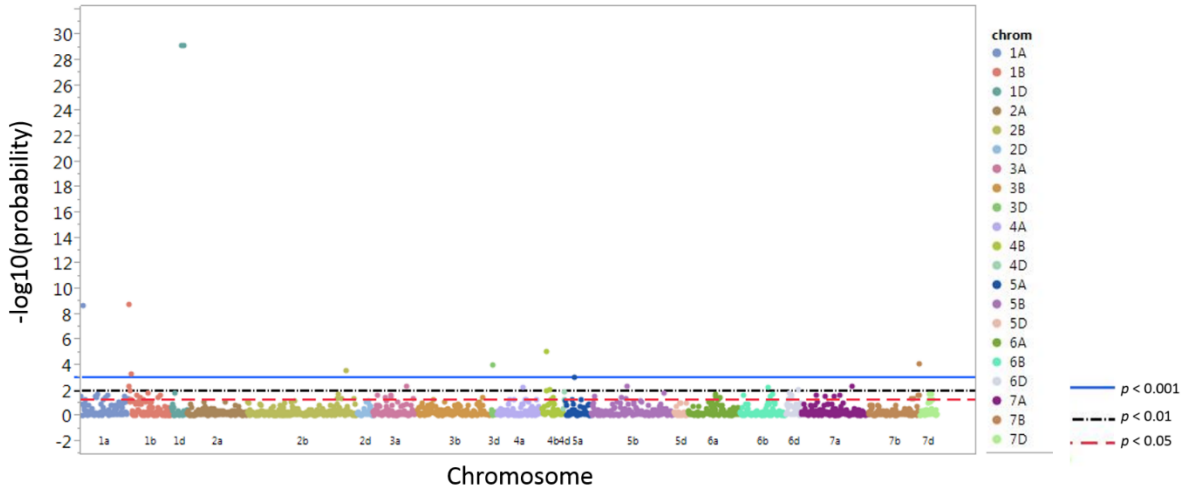


Figure 3-12: Manhattan plot of MP

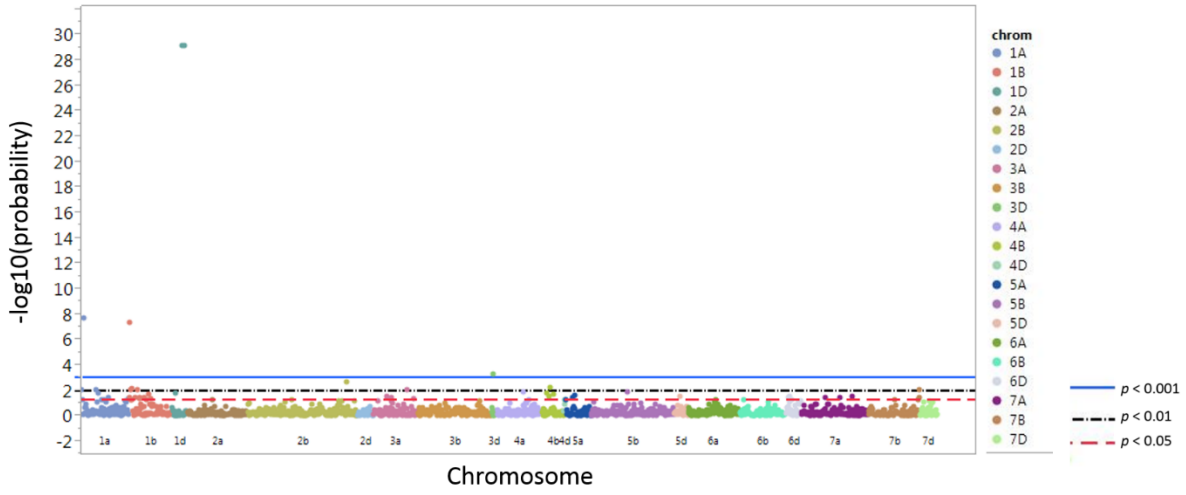


Figure 3-13: Manhattan plot of TESTWT

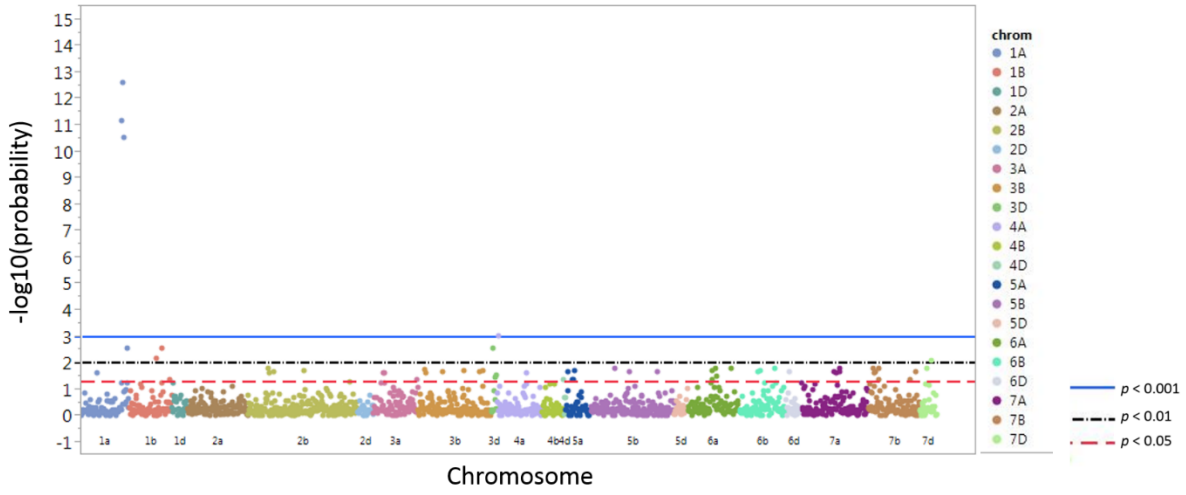
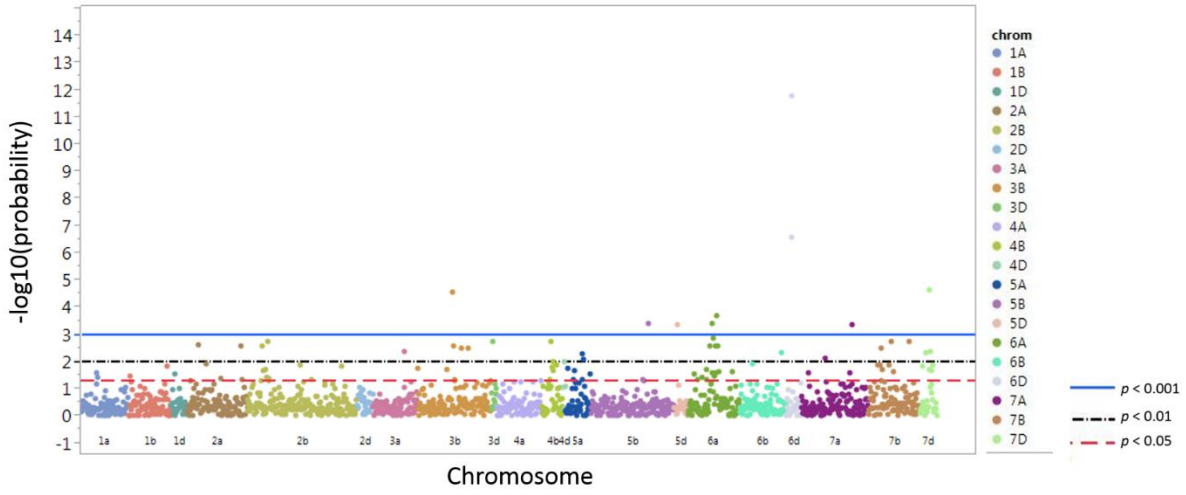


Figure 3-14: Manhattan plot of TKW



Tables

Table 3-1: Correlation between principal components from structure analysis (PCA 1-4) and principal components from kinship structure (IBD 1-4).

	PCA 1 (6.0%)	PCA 2 (4.9%)	PCA 3 (3.3%)	PCA 4 (2.0%)	IBD 1 (20.1%)	IBD 2 (16.5%)	IBD 3 (8.4%)	IBD 4 (4.2%)
PCA 1 (6.0%)	1.0000	0.0000	0.0000	0.0000	0.0083	-0.0060	-0.0200	0.0054
PCA 2 (4.9%)	0.0000	1.0000	0.0000	0.0000	0.0111	-0.0382	-0.0184	-0.0119
PCA 3 (3.3%)	0.0000	0.0000	1.0000	0.0000	0.0276	0.0635	-0.0201	-0.0208
PCA 4 (2.0%)	0.0000	0.0000	0.0000	1.0000	0.0176	-0.0210	0.0063	-0.0093
IBD 1 (20.1%)	0.0083	0.0111	0.0276	0.0176	1.0000	-0.0004	-0.0038	-0.0020
IBD 2 (16.5%)	-0.0060	-0.0382	0.0635	-0.0210	-0.0040	1.0000	-0.0034	-0.0018
IBD 3 (8.4%)	-0.0200	-0.0184	-0.0201	0.0063	-0.0037	-0.0034	1.0000	-0.0163
IBD 4 (4.2%)	0.0054	-0.0119	-0.0208	-0.0093	-0.0020	-0.0018	-0.0163	1.0000

Table 3-2: Significant marker trait associations with Bowtie and POPSEQ alignment, and overall effect, standard error, and False Discovery Rate adjusted –Log10(p-value).

Marker	Chr.	Position	PS Chr.	PS cM	Trait	Overall Effect	Overall SE	FDR Neg LOG10(p)
Loci_S1_460472	1A	460472	1A	0	ALVW	-8.02	1.72	5.51
Loci_S1_460474	1A	460474	1A	0	GRNHRD	-0.61	0.13	3.25
Loci_S1_1246860	1A	1246860	1A	4.562	ALVW	-8.62	1.60	7.15
Loci_S1_1246860	1A	1246860	1A	4.562	FLRSDS	-0.31	0.05	6.14
Loci_S1_1246860	1A	1246860	1A	4.562	MIXTIM	-0.11	0.02	8.70
Loci_S1_1246860	1A	1246860	1A	4.562	MP	-4.19	0.63	7.74
Loci_S1_243975850	1A	243975850			TESTWT	-0.19	0.02	11.15
Loci_S1_244605162	1A	244605162			TESTWT	0.20	0.02	12.63
Loci_S1_245341120	1A	245341120	1A	139.761	ALVPL	-0.03	0.01	3.84
Loci_S1_245341120	1A	245341120	1A	139.761	TESTWT	0.18	0.02	10.54
Loci_S2_2163545	1B	2163545			ALVW	-9.96	2.01	6.15
Loci_S2_2163545	1B	2163545			GRNHRD	-0.61	0.10	5.63
Loci_S2_2532201	1B	2532201			ALVW	-15.87	1.64	21.36
Loci_S2_2532201	1B	2532201			FLRSDS	-0.29	0.05	4.82
Loci_S2_2532201	1B	2532201			MIXTIM	-0.11	0.02	8.79

Loci_S2_2532201	1B	2532201			MP	-4.27	0.67	7.36
Loci_S2_7828321	1B	7828321	1B	26.891	MIXTIM	-0.10	0.02	3.28
Loci_S2_247110912	1B	247110912			ALVW	-10.89	2.30	5.66
Loci_S2_252844209	1B	252844209			FLRSDS	0.33	0.06	4.51
Loci_S2_258464672	1B	258464672	1B	74.787	LOFVOL	-5.39	1.16	3.48
Loci_S2_266250969	1B	266250969	1B	79.233	ALVW	9.72	1.23	14.66
Loci_S2_266250969	1B	266250969	1B	79.233	GRNHRD	0.76	0.07	25.06
Loci_S2_276759258	1B	276759258	1B	94.045	ALVPL	-0.04	0.01	6.26
Loci_S2_282749164	1B	282749164	1B	106.272	ALVPL	0.11	0.01	29.17
Loci_S2_282749164	1B	282749164	1B	106.272	FLRPRO	-0.05	0.01	3.93
Loci_S2_282749164	1B	282749164	1B	106.272	FLRYLD	-0.52	0.04	28.88
Loci_S2_282749164	1B	282749164	1B	106.272	LOFVOL	-6.66	1.00	8.11
Loci_S2_282801547	1B	282801547			ALVPL	0.05	0.01	5.78
Loci_S2_282801547	1B	282801547			FLRYLD	-0.27	0.05	3.72
Loci_S2_284583852	1B	284583852			ALVPL	0.03	0.01	3.04
Loci_S3_108397610	1D	108397610	1D	78.754	ALVPL	0.07	0.01	6.76
Loci_S3_108397610	1D	108397610	1D	78.754	ALVW	45.26	2.41	32.00
Loci_S3_108397610	1D	108397610	1D	78.754	MIXTIM	0.52	0.02	29.16
Loci_S3_108397610	1D	108397610	1D	78.754	MP	19.21	0.88	29.18
Loci_S3_113356875	1D	113356875	1D	78.474	ALVPL	0.08	0.01	7.98
Loci_S3_113356875	1D	113356875	1D	78.474	ALVW	44.42	2.36	32.00
Loci_S3_113356875	1D	113356875	1D	78.474	MIXTIM	0.53	0.02	29.16
Loci_S3_113356875	1D	113356875	1D	78.474	MP	19.79	0.88	29.18
Loci_S3_134641057	1D	134641057			ALVPL	-0.04	0.01	6.19
Loci_S4_253161415	2A	253161415	4A	0	FLRSDS	-0.31	0.06	3.98
Loci_S4_253161415	2A	253161415	4A	0	LOFVOL	-6.76	1.41	3.76
Loci_S5_41827195	2B	41827195			ALVW	-11.21	2.60	4.80
Loci_S5_332355376	2B	332355376	2B	116.178	MIXTIM	0.16	0.03	3.61
Loci_S6_16493617	2D	16493617	2D	77.464	GRNPRO	0.06	0.01	3.10
Loci_S7_14150522	3A	14150522	3A	55.919	LOFVOL	5.90	1.18	4.14
Loci_S7_24772605	3A	24772605	3A	63.04	LOFVOL	5.28	1.14	3.48
Loci_S7_70434692	3A	70434692			LOFVOL	-6.91	1.40	3.99
Loci_S7_164428314	3A	164428314	3A	102.094	FLRPRO	-0.06	0.01	3.04
Loci_S7_164428314	3A	164428314	3A	102.094	FLRSDS	-0.29	0.06	3.98
Loci_S7_178923859	3A	178923859	3A	159.687	ALVPL	0.03	0.01	3.33
Loci_S8_511291591	3B	511291591			TKW	0.70	0.13	4.54
Loci_S8_652811064	3B	652811064	3B	94.67	GRNPRO	0.06	0.01	3.34
Loci_S9_50810552	3D	50810552	3D	67.442	ALVW	-7.43	1.37	7.21
Loci_S9_50810552	3D	50810552	3D	67.442	FLRPRO	-0.05	0.01	3.76
Loci_S9_50810552	3D	50810552	3D	67.442	GRNPRO	-0.06	0.01	3.87
Loci_S9_50810552	3D	50810552	3D	67.442	MIXTIM	-0.07	0.01	4.04
Loci_S9_50810552	3D	50810552	3D	67.442	MP	-2.58	0.54	3.27
Loci_S9_96868734	3D	96868734	3D	82.278	GRNPRO	0.05	0.01	3.44

Loci_S10_24957419	4A	24957419	4A	58.701	TESTWT	-0.11	0.02	3.02
Loci_S10_201716835	4A	201716835			ALVW	10.61	2.17	5.98
Loci_S10_202066152	4A	202066152			ALVW	11.52	2.36	5.99
Loci_S10_203231427	4A	203231427			ALVW	-11.72	2.32	6.39
Loci_S11_4861934	4B	4861934	4B	34.63	GRNHRD	0.35	0.08	3.25
Loci_S11_9212182	4B	9212182	4B	46.92	ALVPL	-0.04	0.01	4.18
Loci_S11_9212182	4B	9212182	4B	46.92	FLRPRO	-0.06	0.01	3.98
Loci_S11_9212182	4B	9212182	4B	46.92	GRNHRD	0.44	0.08	5.22
Loci_S11_9212182	4B	9212182	4B	46.92	GRNPRO	-0.06	0.01	3.58
Loci_S11_9212182	4B	9212182	4B	46.92	MIXTIM	0.09	0.02	5.08
Loci_S11_9791147	4B	9791147			ALVW	-13.94	2.15	10.02
Loci_S13_77029577	5A	77029577	5A	10.487	ALVW	7.88	1.55	6.46
Loci_S13_77296972	5A	77296972	5A	10.487	ALVW	6.64	1.52	4.92
Loci_S13_82914716	5A	82914716			ALVW	-7.78	1.49	6.73
Loci_S13_82914716	5A	82914716			MIXTIM	-0.07	0.01	3.06
Loci_S14_236729118	5B	236729118			TKW	0.45	0.10	3.38
Loci_S14_270906880	5B	270906880	5B	165.726	GRNHRD	-0.61	0.14	3.11
Loci_S14_271965160	5B	271965160			GRNHRD	-0.82	0.15	4.81
Loci_S15_143849002	5D	143849002			TKW	0.51	0.11	3.35
Loci_S16_5034804	6A	5034804			GRNHRD	-0.50	0.11	3.25
Loci_S16_8917119	6A	8917119			FLRSDS	-0.31	0.06	4.54
Loci_S16_19072856	6A	19072856	6A	62.364	FLRPRO	-0.12	0.01	17.63
Loci_S16_19072856	6A	19072856	6A	62.364	GRNPRO	-0.12	0.01	13.86
Loci_S16_19072856	6A	19072856	6A	62.364	LOFVOL	-7.89	1.21	7.84
Loci_S16_24998762	6A	24998762	6A	63.546	FLRPRO	-0.13	0.01	18.40
Loci_S16_24998762	6A	24998762	6A	63.546	GRNPRO	-0.13	0.01	14.38
Loci_S16_24998762	6A	24998762	6A	63.546	LOFVOL	-7.77	1.31	6.32
Loci_S16_33226117	6A	33226117	6A	65.769	ALVW	-9.47	1.85	6.51
Loci_S16_33226117	6A	33226117	6A	65.769	FLRPRO	-0.18	0.01	29.43
Loci_S16_33226117	6A	33226117	6A	65.769	GRNPRO	-0.16	0.01	26.77
Loci_S16_33226117	6A	33226117	6A	65.769	LOFVOL	-10.15	1.28	11.54
Loci_S16_50275005	6A	50275005			FLRPRO	0.13	0.02	6.60
Loci_S16_50275005	6A	50275005			GRNPRO	0.10	0.02	3.24
Loci_S16_94542206	6A	94542206			TKW	0.66	0.14	3.38
Loci_S16_143466155	6A	143466155			GRNPRO	0.07	0.01	3.44
Loci_S16_150663555	6A	150663555	6A	70.269	FLRPRO	0.14	0.02	6.99
Loci_S16_150663555	6A	150663555	6A	70.269	GRNPRO	0.13	0.02	4.62
Loci_S16_150663555	6A	150663555	6A	70.269	LOFVOL	10.30	2.17	3.67
Loci_S16_154516738	6A	154516738			TKW	-0.51	0.10	3.70
Loci_S17_5974923	6B	5974923	6B	25.394	ALVPL	-0.07	0.01	10.96
Loci_S17_5974923	6B	5974923	6B	25.394	FLRPRO	0.07	0.02	3.61
Loci_S17_6513799	6B	6513799			ALVPL	-0.07	0.01	9.78
Loci_S17_164481606	6B	164481606	6B	61.6255	FLRPRO	0.07	0.01	3.16

Loci_S17_164481606	6B	164481606	6B	61.6255	GRNPRO	0.07	0.02	3.44
Loci_S17_176721264	6B	176721264			FLRPRO	-0.06	0.01	3.40
Loci_S17_176721264	6B	176721264			GRNPRO	-0.06	0.01	3.10
Loci_S17_180461356	6B	180461356	6B	67.906	FLRPRO	-0.07	0.01	3.38
Loci_S18_20387611	6D	20387611	6D	59.749	ALVW	6.52	1.39	5.58
Loci_S18_20387611	6D	20387611	6D	59.749	TKW	0.58	0.07	11.76
Loci_S18_111388404	6D	111388404	6D	66.9795	ALVW	7.26	1.40	6.69
Loci_S18_111388404	6D	111388404	6D	66.9795	TKW	0.46	0.07	6.57
Loci_S19_15975054	7A	15975054	7A	41.985	FLRPRO	-0.07	0.02	3.25
Loci_S19_15975054	7A	15975054	7A	41.985	GRNPRO	-0.08	0.02	3.82
Loci_S19_18380157	7A	18380157			GRNPRO	-0.07	0.02	3.30
Loci_S19_78415889	7A	78415889			LOFVOL	-8.63	1.18	9.79
Loci_S19_101352522	7A	101352522			ALVPL	0.04	0.01	3.33
Loci_S19_101352522	7A	101352522			LOFVOL	-8.33	1.17	9.43
Loci_S19_112027332	7A	112027332	7A	82.926	ALVPL	0.04	0.01	3.84
Loci_S19_112027332	7A	112027332	7A	82.926	LOFVOL	-8.49	1.22	8.98
Loci_S19_162307306	7A	162307306	7A	110.34	TKW	-0.44	0.09	3.35
Loci_S20_17646955	7B	17646955	7B	61.943	GRNHRD	-0.64	0.14	3.19
Loci_S20_173057509	7B	173057509			ALVPL	-0.04	0.01	3.04
Loci_S20_211863525	7B	211863525			GRNPRO	-0.07	0.01	3.44
Loci_S20_235418274	7B	235418274	7B	125.056	GRNHRD	-0.44	0.09	4.18
Loci_S20_251235369	7B	251235369	7B	151.479	ALVW	7.66	1.51	6.37
Loci_S20_251235369	7B	251235369	7B	151.479	GRNPRO	-0.06	0.01	3.51
Loci_S20_251235369	7B	251235369	7B	151.479	MIXTIM	0.08	0.02	4.12
Loci_S21_51496571	7D	51496571	7D	117.486	FLRSDS	0.27	0.06	3.59
Loci_S21_55673430	7D	55673430	7D	115.977	TKW	-0.51	0.09	4.65

Table 3-3: Tag sequence with polymorphic index content, heterozygous frequency, and minor allele frequency for significant marker-trait associations.

Marker	Tag Sequence	PIC	Het Freq.	MAF
Loci_S1_460472	TTCAGGCCGAGTCACTGCACCGACCCGTCCATGC GTGCTCGACGGTGGGATTGGACGAGCTGCA	0.30	0.04	0.24
Loci_S1_460474	TTCAGGCCGAGTCACTGCACCGACCCGTCCATGC GTGCTCGACGGTGGGATTGGACGAGCTGCA	0.10	0.01	0.06
Loci_S1_1246860	TGCAGTCAATGATCCAGTTCCTCCGACCAAAGAC CTCGCAACAGAACAACTGCCAGTTGAAGCG	0.32	0.03	0.28
Loci_S1_1246860	TGCAGTCAATGATCCAGTTCCTCCGACCAAAGAC CTCGCAACAGAACAACTGCCAGTTGAAGCG	0.32	0.03	0.28

Loci_S1_1246860	TGCAGTCAATGATCCAGTTCCTCCGACCAAAGAC CTCGCAACAGAACAACTGCCAGTTGAAGCG	0.32	0.03	0.28
Loci_S1_1246860	TGCAGTCAATGATCCAGTTCCTCCGACCAAAGAC CTCGCAACAGAACAACTGCCAGTTGAAGCG	0.32	0.03	0.28
Loci_S1_243975850	TGCAGTCAAGGGCCTCGTCAGCTCCTCCACCATC TATCTTTTGTCATGCAGAGTTTCACTACAT	0.38	0.04	0.50
Loci_S1_244605162	TGCAGCTCGGCCGCCATGGCGAGATCCATCCACT GGAACGCAACCCTGGTTTTGCAGCGCCAGC	0.38	0.04	0.50
Loci_S1_245341120	CCCCAAAACATCACTGGCTGCTGGAACACAGGT TATCTCCGAAAAGGGGCAGATGAACACTGCA	0.37	0.03	0.49
Loci_S1_245341120	CCCCAAAACATCACTGGCTGCTGGAACACAGGT TATCTCCGAAAAGGGGCAGATGAACACTGCA	0.37	0.03	0.49
Loci_S2_2163545	TGCAGTAGAGAGCCCCAATGCCTGATGGACTCA GTACCACTCAAGACAAACATTGGTAGATGAT	0.18	0.02	0.11
Loci_S2_2163545	TGCAGTAGAGAGCCCCAATGCCTGATGGACTCA GTACCACTCAAGACAAACATTGGTAGATGAT	0.18	0.02	0.11
Loci_S2_2532201	TGCAGCGTTTCTTCTTCTTCTTTGCCTTGATGATC GTTTGCCTTGCGTTTTTGCAGCGAGAATA	0.28	0.02	0.21
Loci_S2_2532201	TGCAGCGTTTCTTCTTCTTCTTTGCCTTGATGATC GTTTGCCTTGCGTTTTTGCAGCGAGAATA	0.28	0.02	0.21
Loci_S2_2532201	TGCAGCGTTTCTTCTTCTTCTTTGCCTTGATGATC GTTTGCCTTGCGTTTTTGCAGCGAGAATA	0.28	0.02	0.21
Loci_S2_2532201	TGCAGCGTTTCTTCTTCTTCTTTGCCTTGATGATC GTTTGCCTTGCGTTTTTGCAGCGAGAATA	0.28	0.02	0.21
Loci_S2_7828321	TGCAGAAACTAATGTATACTTCCTACTCCCTTCA GCCTGCTTGCTGATGGTTCTGTGGTCTCGT	0.21	0.02	0.14
Loci_S2_247110912	TGCAGCTAAACTTTACTTGTACGGTCGTACGTGC CGTACTGTCCG	0.19	0.02	0.12
Loci_S2_252844209	TGCAGGTGTCCGACATGGACATGTATTACCATTT ACGTGACCCTGTTTTTGTTGATGTCATGAT	0.26	0.03	0.19
Loci_S2_258464672	TGCAGAGGAATGGAGGAGGGAACCTGCTGAGAC GGGAGGTGGCGGCGTGGGGAAGAAGGTCTCT	0.37	0.04	0.49
Loci_S2_266250969	TGCAGCGTCACCCCCTGCACGCTCACCCCCCTGC ACTCCTCGAAATGCAACGCCTACAACATCC	0.38	0.04	0.50
Loci_S2_266250969	TGCAGCGTCACCCCCTGCACGCTCACCCCCCTGC ACTCCTCGAAATGCAACGCCTACAACATCC	0.38	0.04	0.50

Loci_S2_276759258	AGATAAAACCTCTCGAAGTCTTCTCGATAACCTC GCTGTCGTGGATCCGTAGTAAGTCGCTGCA	0.35	0.03	0.33
Loci_S2_282749164	GGGTCCATCCAACAAATCTGTGACCCTAAGTTGC TTGCATGTTTCGCACACAAGTGAATCTGCA	0.37	0.03	0.45
Loci_S2_282749164	GGGTCCATCCAACAAATCTGTGACCCTAAGTTGC TTGCATGTTTCGCACACAAGTGAATCTGCA	0.37	0.03	0.45
Loci_S2_282749164	GGGTCCATCCAACAAATCTGTGACCCTAAGTTGC TTGCATGTTTCGCACACAAGTGAATCTGCA	0.37	0.03	0.45
Loci_S2_282749164	GGGTCCATCCAACAAATCTGTGACCCTAAGTTGC TTGCATGTTTCGCACACAAGTGAATCTGCA	0.37	0.03	0.45
Loci_S2_282801547	TGCAGCAGCGTGAACCGTGAAGCAAGGAACCAC CAACGGAGAGATCGGAAGAGCGGTTCAGCAG	0.31	0.03	0.27
Loci_S2_282801547	TGCAGCAGCGTGAACCGTGAAGCAAGGAACCAC CAACGGAGAGATCGGAAGAGCGGTTCAGCAG	0.31	0.03	0.27
Loci_S2_284583852	TGCAGTTGAGAGATATGTATGTATCAGCGCCAC AAGCAGAGGTCAAGCATCAACAAGGTAACCG	0.31	0.03	0.26
Loci_S3_108397610	TGCAGAGGAGGTCAGAGTTCCTCATCTCTGAGGT GGGGCTGGAACCGACATACATTGCTCATCG	0.15	0.01	0.09
Loci_S3_108397610	TGCAGAGGAGGTCAGAGTTCCTCATCTCTGAGGT GGGGCTGGAACCGACATACATTGCTCATCG	0.15	0.01	0.09
Loci_S3_108397610	TGCAGAGGAGGTCAGAGTTCCTCATCTCTGAGGT GGGGCTGGAACCGACATACATTGCTCATCG	0.15	0.01	0.09
Loci_S3_108397610	TGCAGAGGAGGTCAGAGTTCCTCATCTCTGAGGT GGGGCTGGAACCGACATACATTGCTCATCG	0.15	0.01	0.09
Loci_S3_113356875	CTAGTAATAACTAGGCTGATGTGATGTAGCGCAT GTGTGCCTCGCCGCTGCCTGGCTGCCTGCA	0.16	0.01	0.10
Loci_S3_113356875	CTAGTAATAACTAGGCTGATGTGATGTAGCGCAT GTGTGCCTCGCCGCTGCCTGGCTGCCTGCA	0.16	0.01	0.10
Loci_S3_113356875	CTAGTAATAACTAGGCTGATGTGATGTAGCGCAT GTGTGCCTCGCCGCTGCCTGGCTGCCTGCA	0.16	0.01	0.10
Loci_S3_113356875	CTAGTAATAACTAGGCTGATGTGATGTAGCGCAT GTGTGCCTCGCCGCTGCCTGGCTGCCTGCA	0.16	0.01	0.10
Loci_S3_134641057	TGCAGATGTCATCGTCAGTATTTCACTCACTAGA ACTAGCCGCAACATCAACATGTCAGCAGCA	0.38	0.06	0.50
Loci_S4_253161415	GCTCCGTGCGGAGCTGTCGGAGCTGCGGGCTAA AACAGTAGAGTTGAAAAATAGGCACTCTGCA	0.35	0.04	0.34

Loci_S4_253161415	GCTCCGTGCGGAGCTGTCGGAGCTGCGGGCTAA AACAGTAGAGTTGAAAAATAGGCACTCTGCA	0.35	0.04	0.34
Loci_S5_41827195	GTTGGAATGCACGATCCTTTCATTTGCTTGAAGC CTTCAAGATTTTTTCTCCATGGAAACTGCA	0.27	0.02	0.20
Loci_S5_332355376	TCACTTTGAAGATTCAAGTGCAGGCGAGGAGTA AAGACCAGAGAGTGCTTACAAGTCGGCTGCA	0.09	0.01	0.05
Loci_S6_16493617	TGCAGAGTACGAGTACCTATCTCATACAACCAC GAACTGAAACGATGTATGTGTACAATCCAAT	0.31	0.03	0.27
Loci_S7_14150522	CATGCCATCGAGCAGAACATATTCGCCAGCTGTC TGTCACACCTGCAAGGAAAGCAAGTCTGCA	0.36	0.03	0.39
Loci_S7_24772605	TGCAGCATGCCCCTATCATGGTTTGGGAAGCAAT TGATGCCCCGCAGATGACATTTTTAAGAGG	0.36	0.03	0.39
Loci_S7_70434692	ATCCCGTGGCAGCATATTCAAAGATCGAATCTG AGCCGTCATCTTTCCCCGCCATTGCCCTGCA	0.25	0.02	0.18
Loci_S7_164428314	TGCAGAATTGACAGATGCATCAAAATTGGTAGC CGCTGAAGCTAACAATGCTCATGTTGATGTT	0.30	0.02	0.24
Loci_S7_164428314	TGCAGAATTGACAGATGCATCAAAATTGGTAGC CGCTGAAGCTAACAATGCTCATGTTGATGTT	0.30	0.02	0.24
Loci_S7_178923859	TCTTTCGCGACAACAAAAAGCATCGGGCGATCC AACTAGAGGCGGAGTTCAGGAACACTCTGCA	0.32	0.03	0.27
Loci_S8_511291591	TGCAGGTTTCATGGAGCTGCTCAAAGTCCTCAGT GGCCCTCACGGCAGCGTATACGTCTGGATT	0.12	0.01	0.07
Loci_S8_652811064	CTCCACATCAGCTTTTACGTAAAACTCCTATGTT ACTTTTCGGCATTTCCTATTTGACGCTGCA	0.37	0.04	0.44
Loci_S9_50810552	CATTTGTCCGTCCATACGTTAATGCTTGTCCCAT CCCCAACTCTCTGAATAATGCCTAGCTGCA	0.37	0.03	0.42
Loci_S9_50810552	CATTTGTCCGTCCATACGTTAATGCTTGTCCCAT CCCCAACTCTCTGAATAATGCCTAGCTGCA	0.37	0.03	0.42
Loci_S9_50810552	CATTTGTCCGTCCATACGTTAATGCTTGTCCCAT CCCCAACTCTCTGAATAATGCCTAGCTGCA	0.37	0.03	0.42
Loci_S9_50810552	CATTTGTCCGTCCATACGTTAATGCTTGTCCCAT CCCCAACTCTCTGAATAATGCCTAGCTGCA	0.37	0.03	0.42
Loci_S9_50810552	CATTTGTCCGTCCATACGTTAATGCTTGTCCCAT CCCCAACTCTCTGAATAATGCCTAGCTGCA	0.37	0.03	0.42
Loci_S9_96868734	TGCATACTACATGGATGGGTAAAGGCATTGTAA AGGCAGCATGCATGGCACTAGCATGACTGCA	0.37	0.03	0.45

Loci_S10_24957419	TGCAGTGATTTTATGCCAAGAAACAAGAGCACG TGCTGTAAATTTGCGCTTCTTTGGCCTTGTC	0.37	0.02	0.46
Loci_S10_201716835	ATCTCTACCTAACACGCCTCCAGCACTTCAACAG GAGAAGAAGAGCACCTCCATACCCCCTGCA	0.37	0.03	0.49
Loci_S10_202066152	GCACATACGACTTGCGGTGTTGGAGAGGTGGCT CTAGCTCCACGACTGCATCGGTGCCTCTGCA	0.37	0.04	0.49
Loci_S10_203231427	TGCAGCCCATCCCTCTGCACTTCCTCCAGGGTTT GGATGTGCTGTGCGGTGTCAACCCAACAAA	0.37	0.04	0.49
Loci_S11_4861934	TGCAGAAGCCAGGACTCCAGCCAGTGACATCAT GGAAATGTGAAAAGTTACCGCGCGCACACAC	0.36	0.02	0.39
Loci_S11_9212182	TGCAGTACATCATATTTCTGCTGGAAAGGGAGA AGCCTTCAATCTAATCAGAACTCATGACCAT	0.33	0.02	0.30
Loci_S11_9212182	TGCAGTACATCATATTTCTGCTGGAAAGGGAGA AGCCTTCAATCTAATCAGAACTCATGACCAT	0.33	0.02	0.30
Loci_S11_9212182	TGCAGTACATCATATTTCTGCTGGAAAGGGAGA AGCCTTCAATCTAATCAGAACTCATGACCAT	0.33	0.02	0.30
Loci_S11_9212182	TGCAGTACATCATATTTCTGCTGGAAAGGGAGA AGCCTTCAATCTAATCAGAACTCATGACCAT	0.33	0.02	0.30
Loci_S11_9212182	TGCAGTACATCATATTTCTGCTGGAAAGGGAGA AGCCTTCAATCTAATCAGAACTCATGACCAT	0.33	0.02	0.30
Loci_S11_9791147	CTCTACTACACAGCCTCTAATCGCATGTGTTTGT AGTACGGTAGGTGGGTACGCACTGGCTGCA	0.21	0.02	0.14
Loci_S13_77029577	TGCAGACCAGGTTAACGATCAACTTTCTCTCAAT AAAAAAATGTTAGCGATCAAAGCTGCTTGG	0.36	0.04	0.37
Loci_S13_77296972	TGCAGTCTGACGTACCCAGTGCTCCGCATCGATG ATTTCCTCGACTCTCTATATTCCTTCTCCA	0.35	0.04	0.34
Loci_S13_82914716	TATAGACTTTTTCTTCAAATCATTCCACACCGAT TATGCTTTCGCAAATTAAAGGCTGCCTGCA	0.37	0.04	0.47
Loci_S13_82914716	TATAGACTTTTTCTTCAAATCATTCCACACCGAT TATGCTTTCGCAAATTAAAGGCTGCCTGCA	0.37	0.04	0.47
Loci_S14_236729118	GCGCGCGGGTCTTGTTGATGGTGATGCCACCGA GCGACGACGAGTCACCGCCGAGAGCGCTGCA	0.27	0.02	0.20
Loci_S14_270906880	TGCAGTAGCCACAACTTGCAGCTTTAGCCGTGTG CGTGCATGTGTGTGAGAGGTCAGCAATTCA	0.13	0.01	0.08
Loci_S14_271965160	GTACGGAAGTAGTCGACCGTCGGCTTCTTCTTGC GCCACGCCTCGTAGGGCATCATTCCCTGCA	0.13	0.01	0.08

Loci_S15_143849002	GCGTCCACGTCGACATCACGGTTGCCAACGTTGC CATGCTCCGCTGCGCCGCGCTACCTGCA	0.18	0.02	0.11
Loci_S16_5034804	TGCAGGCCCACTCTGTGCCGCCCTGCTGCTGGCG ACGGCAGTCGTGCTCCTCGTGGTCGCCGCG	0.24	0.02	0.16
Loci_S16_8917119	GCCATGTCGTTGAACTACGGTTGCCATCTCGGAC AACTACAGTTGCCATGTTTGCTGAACTGCA	0.35	0.03	0.36
Loci_S16_19072856	GCGGTGACCGCGACCTCCAACTTGGCCCTGTCCG AGAGAAAGCGGAGCATCATGTTCCCCTGCA	0.34	0.03	0.33
Loci_S16_19072856	GCGGTGACCGCGACCTCCAACTTGGCCCTGTCCG AGAGAAAGCGGAGCATCATGTTCCCCTGCA	0.34	0.03	0.33
Loci_S16_19072856	GCGGTGACCGCGACCTCCAACTTGGCCCTGTCCG AGAGAAAGCGGAGCATCATGTTCCCCTGCA	0.34	0.03	0.33
Loci_S16_24998762	TGCAGCACACACCAGCAATTAAAATTTGCACAC CAAACCGTGCCACTATCTTAGCACTGAAAGG	0.33	0.03	0.29
Loci_S16_24998762	TGCAGCACACACCAGCAATTAAAATTTGCACAC CAAACCGTGCCACTATCTTAGCACTGAAAGG	0.33	0.03	0.29
Loci_S16_24998762	TGCAGCACACACCAGCAATTAAAATTTGCACAC CAAACCGTGCCACTATCTTAGCACTGAAAGG	0.33	0.03	0.29
Loci_S16_33226117	TGAACGCACCGAAGCCAACAATCGAAATCATAA AGCCATCAAATGCCGCGGGAGAGAGCCTGCA	0.28	0.02	0.21
Loci_S16_33226117	TGAACGCACCGAAGCCAACAATCGAAATCATAA AGCCATCAAATGCCGCGGGAGAGAGCCTGCA	0.28	0.02	0.21
Loci_S16_33226117	TGAACGCACCGAAGCCAACAATCGAAATCATAA AGCCATCAAATGCCGCGGGAGAGAGCCTGCA	0.28	0.02	0.21
Loci_S16_33226117	TGAACGCACCGAAGCCAACAATCGAAATCATAA AGCCATCAAATGCCGCGGGAGAGAGCCTGCA	0.28	0.02	0.21
Loci_S16_50275005	TGCAGAGTAGAAGGCATGAAGCGTACCATGGAG CCCCGCGTTATGAAGTTGGAGCTCGGTAACA	0.18	0.01	0.11
Loci_S16_50275005	TGCAGAGTAGAAGGCATGAAGCGTACCATGGAG CCCCGCGTTATGAAGTTGGAGCTCGGTAACA	0.18	0.01	0.11
Loci_S16_94542206	TGATTCTCATGTTGCTGCAAAATTCCCAGCAACA TCACCCACTGTCACTAGCAAAGCAGCTGCA	0.26	0.01	0.19
Loci_S16_143466155	ACTTGAGTTTCAACAACTTTTATGGAGAGATACC ACAATCAATCTGCAACCTGACGAACCTGCA	0.37	0.03	0.45
Loci_S16_150663555	TGCAGCGACCCGAAAAAGATTCAGGTGAGCCCT GGCGACTACTTTGGGCTACCAGCAGCCCAGT	0.18	0.01	0.11

Loci_S16_150663555	TGCAGCGACCCGAAAAAGATTCAGGTGAGCCCT GGCGACTACTTTGGGCTACCAGCAGCCCAGT	0.18	0.01	0.11
Loci_S16_150663555	TGCAGCGACCCGAAAAAGATTCAGGTGAGCCCT GGCGACTACTTTGGGCTACCAGCAGCCCAGT	0.18	0.01	0.11
Loci_S16_154516738	AAGACTCCACTAGACCATCCGAGAAAATTTGGC ATGCCTGCACGCAGTTCAGCTGAACCCTGCA	0.37	0.03	0.45
Loci_S17_5974923	TGCAGATCTCCGCAGCTGGCTTCACACCCAGGTC GTCAGCCACACGAACTGATAGACGGTAAAC	0.23	0.02	0.16
Loci_S17_5974923	TGCAGATCTCCGCAGCTGGCTTCACACCCAGGTC GTCAGCCACACGAACTGATAGACGGTAAAC	0.23	0.02	0.16
Loci_S17_6513799	CTCGTATCCACCTCTAGAGATGTCTTACTTCTGT ACTGACATTGCTCTAATTGTTAACTCTGCA	0.23	0.02	0.16
Loci_S17_164481606	TGCAGAATGAAGTACTATACGCTGTGCGCGCGC GCGTGTGCAGTCTAGGAGAAGAGGACGATGC	0.37	0.03	0.46
Loci_S17_164481606	TGCAGAATGAAGTACTATACGCTGTGCGCGCGC GCGTGTGCAGTCTAGGAGAAGAGGACGATGC	0.37	0.03	0.46
Loci_S17_176721264	GTCACTGGCACGACCGACATGACCATTTACCAC GGCCTGCCTGCACAAGAAACAGCTTCCTGCA	0.37	0.03	0.47
Loci_S17_176721264	GTCACTGGCACGACCGACATGACCATTTACCAC GGCCTGCCTGCACAAGAAACAGCTTCCTGCA	0.37	0.03	0.47
Loci_S17_180461356	TGCAGTGATATATATAACCATGGACTAGATGATA AAATTAGAGGAGTAGCTACGTGTAGCATCC	0.25	0.02	0.18
Loci_S18_20387611	TAGACCTTTACAACTAGCCATATTTACATACATA TGGTTGACATACTAGTAACTAGAGCCTGCA	0.36	0.04	0.38
Loci_S18_20387611	TAGACCTTTACAACTAGCCATATTTACATACATA TGGTTGACATACTAGTAACTAGAGCCTGCA	0.36	0.04	0.38
Loci_S18_111388404	GTGTCGGCATGACTGATTCTCCTATGCTACAACA TCACCCACTGTCACTAGCAAAGCAGCTGCA	0.37	0.03	0.46
Loci_S18_111388404	GTGTCGGCATGACTGATTCTCCTATGCTACAACA TCACCCACTGTCACTAGCAAAGCAGCTGCA	0.37	0.03	0.46
Loci_S19_15975054	TGCAGCTTCGCAAAAGAACGGCCATCGGGTCAA TAATGAAGGCAGGATAACATTTCGGTTTCTG	0.23	0.02	0.15
Loci_S19_15975054	TGCAGCTTCGCAAAAGAACGGCCATCGGGTCAA TAATGAAGGCAGGATAACATTTCGGTTTCTG	0.23	0.02	0.15
Loci_S19_18380157	TGCAGCTTGCACGTGCCCGAGCATTATCTTTGTC ACCTTGTCGGAGTTCTGTCTCCACATACCA	0.25	0.02	0.18

Loci_S19_78415889	TTTTCAACTTTCCATGTCTTCTGAGTCCAGTAAGT TTTATCAAATTCCATGTTAAATATCTGCA	0.37	0.04	0.41
Loci_S19_101352522	ATGACGTGGTGGCCAGGCGCTGGTTGCTGCCTCA CCCGACCCCAGCGTGCCCGAGGTGCCTGCA	0.36	0.03	0.39
Loci_S19_101352522	ATGACGTGGTGGCCAGGCGCTGGTTGCTGCCTCA CCCGACCCCAGCGTGCCCGAGGTGCCTGCA	0.36	0.03	0.39
Loci_S19_112027332	TGCAGGCCGTTTACCTAAGTCTCCACACACCTGT ACCCTACAGCTGGCCGCCACACGCCTGGAG	0.36	0.03	0.39
Loci_S19_112027332	TGCAGGCCGTTTACCTAAGTCTCCACACACCTGT ACCCTACAGCTGGCCGCCACACGCCTGGAG	0.36	0.03	0.39
Loci_S19_162307306	TGCAGTTGGTTCGTCTAGCCTCATCGTTGGTGTT CATCGGCACGTTCAAAGGGGAAGAAGACAG	0.26	0.02	0.19
Loci_S20_17646955	TGCAGATATATTCACATGCCTCAAGTTAATTGTA AGCCAGAACACTGAACATATAACCACATTC	0.26	0.01	0.19
Loci_S20_173057509	ACTTGGAGCACCTCTACCTTGATTTCTCTCGGGT GTTGAGTGCGTCGAAGGGTTCGTATCTGCA	0.37	0.05	0.49
Loci_S20_211863525	TATTATTATTGAGAAAGGAGCGGAGCGTATCAC ATGAGATGAACAACCGAATATCCTATCTGCA	0.37	0.03	0.43
Loci_S20_235418274	TTGTGATTAAAAGCCCGATCTACATAGTGAGCG GAAGTTCAGAAATAACACAAGAACATCTGCA	0.27	0.03	0.21
Loci_S20_251235369	GAAGGAAAGGCATGTTGGATGGCCAGTAAACTG TGCGAACGTAGCACAAAGCTAGCGGCCTGCA	0.37	0.04	0.42
Loci_S20_251235369	GAAGGAAAGGCATGTTGGATGGCCAGTAAACTG TGCGAACGTAGCACAAAGCTAGCGGCCTGCA	0.37	0.04	0.42
Loci_S20_251235369	GAAGGAAAGGCATGTTGGATGGCCAGTAAACTG TGCGAACGTAGCACAAAGCTAGCGGCCTGCA	0.37	0.04	0.42
Loci_S21_51496571	TTGAAACCGAAGAAAAACTAACAGCTACCAAGC AGAATGCAGGCGCTCCCTATTTAGCACTGCA	0.25	0.03	0.17
Loci_S21_55673430	TCGACGGTGTACAGCCACGGTGCTGTCGGAAAC TTGAGGCACGAGGATGAGCCAAATAACTGCA	0.30	0.03	0.25

Chapter 4 - Applying genomic selection for prediction of processing and end-use quality traits in Kansas hard red winter wheat breeding program

Sarah D. Battenfield, Jesse A. Poland, R. Chris Gaynor, R. Miller, and Allan K. Fritz

S. D. Battenfield and A. K. Fritz, Dep. of Agron., Kansas State Univ., 2004 Throckmorton Plant Sci. Ctr., Manhattan, KS, 66506; J.A. Poland, Dep. of Plant Pathology, Kansas State Univ., 4011 Throckmorton Plant Sci. Ctr., Manhattan, KS, 66506; R. C. Gaynor, The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, Midlothian, UK; R. Miller, Dep. of Grain Sci & Industry, Kansas State Univ., 3707 Throckmorton Plant Sci. Ctr., Manhattan, KS 66506.

Abbreviations: AVE, Average of all genomic selection models on mean-line basis; AVGDIA, Average kernel diameter; AVGWT, Average kernel weight; BAKEABS, Water absorption measured in baking; BAKETIM, Empirical baking time; BLUE, Best linear unbiased estimate; EN, Elastic net; FLRPRO, Flour protein; FLRYLD, Flour yield; GAUSS, Gaussian kernel; GBS, Genotyping-by-sequencing; GRNHRD, Grain hardness; GRNPRO, Grain protein; lBAKETIM, Log transformed baking time; LOFVOL, pup loaf volume; lMIXTIM, Log transformed mixing time; MIXABS, Water absorption measured from Mixograph; MIXTIM, Mixograph mix time; NIRS, Near infrared spectroscopy; PLSR, Partial least squares regression; TESTWT, Test weight; RF, Random forest; RRBLUP, Ridge regression best linear unbiased prediction; SKCS, Single kernel characterization system.

Abstract

Hard red winter wheat (*Triticum aestivum* L.), the predominant field crop of Kansas, is primarily used in making leavened bread. Regional quality standards for bread making are imperative for cultivar release. However, breeding for quality traits is often considered a secondary goal, largely due to amount of seed needed and expense of such testing. Without testing and selection, many undesirable materials tend to be advanced. Here we develop and validate whole genome prediction models for end-use quality phenotypes routinely generated by the Kansas State University hard red winter wheat breeding program in Manhattan, KS. With these predictions genomic selection (GS) can be applied at earlier stages and undesirable materials culled before implementing expensive yield and quality screenings. Prediction accuracy was tested using data from unbalanced yield trials from 2009 to 2014 (n = 472) in central Kansas evaluated for quality parameters: test weight, average kernel weight and diameter, grain and flour protein, mixing time, water absorption, and bread loaf volume. High-density markers were generated with genotyping-by-sequencing and SNPs were imputed. Prediction models were developed using ridge regression, Gaussian kernel, partial least squares, elastic net, and random forest models in R. In general, prediction accuracy increased over time as more data was available to train the model. Mean prediction accuracies (r) for quality parameters in cross validation ranged 0.2 (loaf volume) to 0.6 (Mixograph water absorption). Based on these prediction accuracies, we conclude that GS can be cautiously applied in the breeding program, but more model development is still needed for full integration.

Hard red winter wheat is the most widely grown crop in Kansas. The wheat from Kansas, and the Southern Great Plains, in general, is commonly considered as high-quality for wheat bread making. This is primarily due to the typically high protein content, dough strength, extensibility, and tolerance to overmixing which is common in commodity wheat from this region.

Wheat in the Great Plains can experience a variety of biotic and abiotic stresses resulting in yield reduction to various levels (Holman, *et al.*, 2011). Wheat quality can be highly impacted by these changes as high environment and genotype by environment interactions have been found for flour protein content in several locations across Nebraska (Graybosch, *et al.*, 1996). The variance in flour protein content in turn changes the quality profile for dough rheology and baking. Despite these environmental challenges and cost of screening, breeding progress have been made to increase the processing and end-use quality over time in the Great Plains (Cox, *et al.*, 1989).

Wheat tested as grain, flour, dough, and final products must be assessed to determine an overall best end-use product, for a given wheat cultivar or breeding line. Typically, hard grain with high protein and strong and extensible gluten is acceptable for making industrial pan bread, whereas soft grain with low protein and weak and extensible gluten is more acceptable for the cookies, cakes, and pastries (Peña, 2002). Regional breeding standards for hard red winter wheat quality traits in released cultivars are set by the Wheat Quality Council (HWW Quality Targets Committee, 2006). Meeting these parameters maintains the good quality for all wheat grown in this area and is an imperative of all individual released varieties.

Testing for wheat quality requires large amounts of seed, and has cost in addition to field testing. Due to the amount of materials in the breeding program to be tested and the amount of seed required, resources for testing these traits are often not utilized until late in the breeding program. Thus, breeding program resources will be spent on materials which may not pass the quality targets. Therefore, the objective of this research was to test the utility of genomic selection to predict processing and end-use quality in the Kansas State University wheat breeding program in Manhattan, so that predictions of wheat quality may be available earlier in the program.

Materials and Methods

Breeding Program Outline

There are two publicly funded wheat breeding programs in the state of Kansas working through Kansas State University (KSU). These programs are located in Manhattan and Hays, KS. The primary focus of this paper is the Manhattan breeding program, which produces hard red winter wheat varieties mostly for Central and Eastern Kansas.

The Kansas State University Manhattan wheat breeding program uses a selected bulk strategy in segregating generations, derives wheat pure lines at the F_5 stage, and then begins yield testing in the $F_{5:6}$. $F_{5:6}$ are tested in small plot, augmented yield tests, and processing and end-use quality is first tested for SKCS and Mixograph parameters. Results of these tests are typically available for selection for advancement of materials to replicated yield tests. Full bakes are not conducted until following the 7^{th} generation replicated yield tests. These results are typically not available until during the field season of the 8^{th} generation. By this time, the program has highly selected candidates based on agronomic performance, and then needs to conduct culling more heavily based on wheat quality.

The $F_{5:6}$ are tested in an unreplicated, small-plot (0.75 x 2.25 m) yield trial, referred to as the individual plant short row (IPSR). The lines grown here are investigated for overall appearance, disease resistance, augmented yield comparison, and test weight. After the lines have been filtered for these traits, the remaining entries are tested in the wheat quality lab for grain size, grain protein, then milled and screened for peak mix time and tolerance to overmixing using Mixograph dough recorder (National Manufacturing, Lincoln, NE). This step serves as a secondary filtration for entries which were questionable after the yield test, and likely only removes entries with no tolerance to overmixing or extreme mix times. If the line performed very well in the augmented yield test or had other target traits of interest, which will be handled in another cycle of breeding, the material stays in the program regardless of quality at this point.

The next stage of testing for the KSU wheat breeding program is larger (1.5 x 4.5 m) replicated preliminary yield tests or nurseries (PYN). These tests are grown also in an unreplicated, augmented design per location, however they are replicated in several locations across the state. The top performers of this test over several locations will be tested for the full panel of processing and end-use tests: SKCS, grain protein, milling Mixograph and baked into a pup loaf. The results of these tests do not return to the breeder until the following yield test is

planted, so all advancement decisions in this round are based on yield, agronomic performance, disease resistance, and other specifically targeted traits of interest.

Advanced yield nurseries (AYN) contain best materials at this stage from both Manhattan and Hays breeding programs. These tests are conducted in two replicate plots (1.5 x 4.5 m) in alpha lattice design for the Manhattan program managed locations, and three replicate plots (1.5 x 4.5 m) in a randomized complete block design for the Hays program managed locations. The final advanced stage of testing is the Kansas intrastate nursery (KIN) which is planted in all locations in replicated designs. Materials are typically tested in the KIN and other regional yield tests for several years before release decisions are made for wheat lines.

Genotypes

All materials in the PYN in 2011 and beyond from the KSU Manhattan wheat breeding program have been genotyped using genotyping-by-sequencing. Additionally, genotyping moved up one year in the breeding scheme starting in 2013, from that time forward all IPSR were genotyped. This results in a total of 6,134 materials genotyped in the KSU Manhattan breeding program to date.

Annually, as new nurseries were finalized, DNA was extracted from bulked leaf tissue using the BioSprint 96 DNA Plant Kit (Qiagen) with the BioSprint 96 Workstation (Qiagen). Genotyping-by-sequencing was conducted as in Poland, *et al.* (2012) using TASSEL (Bradbury, *et al.*, 2007) version 4 *de novo* pipeline to identify single nucleotide polymorphisms (SNPs). The SNPs were converted from the hap files to numeric (1, 0, -1, for homozygous major allele, heterozygous, and homozygous minor allele, respectively) using R package 'GSwGBS' (Gaynor, 2015). The genotype matrix was then filtered for a maximum of 20% missing within each marker, and maximum of 50% missing markers for each individual and mean marker imputation was conducted (Endelman, 2011).

Phenotypes

472 wheat lines yield tested between 2005 and 2014 were used in genomic selection for wheat quality. These lines represent all available entries with genotype and phenotype information from PYN testing and beyond. Since only best materials selected for advancement are tested for quality, these data are unbalanced replicates where some lines were replicated in more than one environment per year, or tested across multiple years. Best linear unbiased estimates (BLUEs) of the quality parameters were made to allow for one phenotype per line, per

trait for the GS modeling. BLUEs were determined using site-year, the location-year combination, as a fixed effect in linear modeling in R package 'lme4' (Bates, *et al.*, 2013).

Test weight per bushel (TESTWT) is the weight of grain required to fill a level Winchester bushel. This measurement is correlated with the flour yield from milling, and an important metric in grain sales. The regional goal for cultivars is greater than 60 lb bu^{-1} (HWW Quality Targets Committee, 2006). For this study, TESTWT was measured with a Seedburo Filling Hopper and Stand (Seedburo Equipment, Des Plaines, IL) using method 55-10.01 (AACC, 2000).

Individual wheat kernels were measured using the Single Kernel Characterization System (SKCS) for wheat kernel texture, SKCS 4100 (Perten Instruments, Inc., Springfield, IL). This instrument measures grain hardness, weight, and diameter on each kernel. Averages of 200 measured kernels are represented here. The method used was 55-31.01 (AACC, 2000). Almost all entries available for training were hard, and little deviation was found for hardness index, thus they were excluded from modeling. Regional targets for grain weight (AVGWT) and diameter (AVGDIA) are kernels greater than 30 mg and 2.4 mm, respectively (HWW Quality Targets Committee, 2006).

Grain protein (GRNPRO) was assessed using AACC method 46-30.01 (AACC, 2000). Measured using Diode Array 7200 NIR (Perten Instruments, Inc., Springfield, IL) and reported on a 12% moisture basis. Preferably, wheat contains great than 12% protein content at 12% moisture basis according to the HWW Quality Targets Committee (2006). Milling was conducted with Brabender Quadrumat Sr. (Brabender, Duisburg, Germany) and requires at least 1500 g seed for mixograph and bake tests for these tests. Flour protein (FLRPRO) and moisture determined by NIRS (Perten Instruments, Inc., Springfield, IL). This is reported at 14.5% moisture basis.

Dough development time (MIXTIM) and water absorption (MIXABS) were determined by Mixograph recording dough mixer (National Manufacturing, Lincoln, NE), method 54-40.02 (AACC, 2000). This test requires 50g of flour for each replicate. Regional targets prefer greater than 62% water absorption at 14% moisture basis and a peak mixing time between 4 and 8 minutes, for optimal adaptation to industrial baking processes (HWW Quality Targets Committee, 2006).

Test baking was conducted by AACC method 10-10.03 (AACC, 2000). This test requires 600 g flour. The water absorption and mixing time during baking preparation are recorded as BAKEABS and BAKETIM, respectively. Wheat quality standards for this region are greater than 62% water absorption at 14% moisture basis and 3-5 minute mixing time(HWW Quality Targets Committee, 2006). Loaf volume (LOFVOL) was measured by rape seed displacement. Local standards prefer loaf volume greater than 850 cc in released cultivars (HWW Quality Targets Committee, 2006).

GS Methods

Genomic selection prediction was conducted using the methods in Gaynor (2015). Briefly, R package 'GSwGBS' utilizes other R packages 'rrBLUP', 'pls', 'randomForest', and 'glmnet' to conduct various methods to solve for marker effects in a training population (Liaw and Wiener, 2002, Mevik and Wehrens, 2007, Friedman, *et al.*, 2009, Endelman, 2011). Methods in this package are ridge regression best linear unbiased predictor (RRBLUP), Gaussian kernel (GAUSS), partial least squares regression (PLSR), elastic net (EN), and random forest (RF). All prediction models are also averaged to prevent poor prediction accuracy in unknown years and characteristics (Gaynor, 2015). These methods are more fully described in chapter 2 of this dissertation and in Gaynor (2015).

Marker effects solved by the models are then tested on either a new population to be predicted, which was previously untested, or some portion of the previously tested materials in cross validation. Here we test GS on 472 wheat breeding lines with wheat bread making quality phenotypes were available from 2005-2014 (Figure 1). 251 of the entries were tested in 2014, and 221 entries were historical, or tested between 2005 and 2013. Historical entries were used to train the model, predicting for the 2014 entries, and vice versa. Additionally, cross validation was conducted where 80% of the entries were used to train the model and 20% were masked to test the model. In all testing methods correlations were made between predicted and empirical quality phenotypes.

Distributions of MIXTIM and BAKETIM did not follow a Gaussian distribution. These were log transformed prior to analysis. To demonstrate the log transformation these results are demonstrated as lMIXTIM and lBAKETIM.

Results and Discussion

Cross validation within all 472 entries resulted in significant predictions for all processing and end-use quality traits in the Kansas State University hard red winter wheat breeding program (Figure 2). GS methods were typically not significantly different in predictive ability; however RF has highest accuracies in mixing and baking time. While many methods demonstrate variable performance between the traits, AVE method performs stable and among the highest prediction correlation.

Forward prediction where one set of materials trains a model to predict different materials in a different year, however, is more indicative of a breeding program. Forward prediction accuracies were much lower than cross prediction for all traits. Significant predictions were found for AVGWT and AVGDIA were found in when historical material predicted 2014 entries for quality (Table). These significant prediction correlations were low, though, 0.261 and 0.176, respectively for AVGWT and AVGDIA.

In the reverse scenario when 2014 materials trained the model to predict historical lines more traits were significantly predictive: TESTWT, AVGWT, AVGDIA, MIXTIM, BAKETIM, LOFVOL, and the log transformed bake and mixing times. These significant models, however, still only predicted a low portion with correlations ranging from 0.12 to 0.29. In these tests the models fluctuated by trait which was most accurate, but significant differences were not seen in model performance.

Conclusions

Genomic selection for wheat end-use and processing quality has shown promise in the Kansas State University hard red winter wheat breeding program in cross validation testing. However, predictions across environments for processing and end-use quality are much less accurate and in many cases non-significant. Prediction models trained with the 2014 year where more materials were sampled from the same environment produced a more robust training model than the BLUEs from all unbalanced historical data.

Prediction model performance was variable by trait and material tested. Predictions made from the model averaging method demonstrate most stability over multiple traits in multiple years. This validates the results of Raftery, *et al.* (1997) and Claeskens and Hjort (2008) which indicate that the model averaging protects from model variability in predictions into unknown

situations. Breeders are constantly faced with unknown situations due to varying environment and genotype by environment impact, which is another reason this method is preferred.

Higher prediction accuracies were seen in the CIMMYT bread wheat breeding program for all traits in forward and cross prediction accuracy (Battenfield, Chapter 2). Possible explanations are fewer lines in the Kansas training population, possibility of less genetic variation for the traits of interest in the Kansas program, and more environment or genotype by environment impact unaccounted for in the Kansas program. Based on high accuracies of random forest method compared to the rrBLUP method in Kansas forward predictions, we can assume the models will continue increasing accuracy with more materials as this method works better with less data (unpublished results). Differing predictive abilities of the historical set of highly unbalanced data compared to the 2014 materials indicates that more wheat lines should be sampled from individual site-years to increase predictive accuracy of the model. Additionally, since Kansas breeds only for industrial pan bread in the less diverse winter background, it is possible that the genetic diversity may have an impact on the upper limits of attainable GS accuracy in Kansas. Finally, we know there are high impact of environment and genotype by environment for the quality traits in the Great Plains (Graybosch, *et al.*, 1996). Our methods used fixed effect BLUEs to account for these impacts, but models were not built to handle these factors well.

Currently, these models need more information before full implementation into the Kansas hard red winter wheat breeding program. However, the models have been utilized with cautious optimism. Here breeder recommendations have been made indicating lines greater than 1 or 2 standard deviations away from the mean of all predictions for a trait. Quality traits should remain within the parameters given by the HWW Quality Targets Committee (2006). Currently, middle predictions, average predictions plus or minus one standard deviation, represent the middle of the quality targets for most traits. Tails of the predicted distributions have materials which typically are also in the empirical tails, thus, we can select for the portion of the predicted curve as applicable to the trait of interest. This strategy will likely need to be reassessed as allele frequencies within the breeding program shift with selection.

Tables

Table 4-1: GS predictions using historical set to predict 2014 materials.

	RRBLUP	GAUSS	PLSR	ELNET	RF	AVE
	Correlation coefficient (r)					
TESTWT	0.016	0.062	0.028	0.031	0.097	0.042
AVGWT	0.254**	0.298**	0.231*	0.254**	0.190*	0.254**
AVGDIA	0.178	0.198*	0.155	0.179	0.190*	0.182
GRNPRO	0.052	0.092	0.039	0.092	0.177*	0.084
FLRPRO	-0.024	0.027	-0.071	-0.039	-0.008	-0.036
MIXABS	0.009	0.050	-0.034	0.008	-0.028	0.000
MIXTIM	0.079	0.069	0.084	0.081	0.161	0.097
BAKEABS	0.066	0.232**	0.093	0.074	0.133	0.109
BAKETIM	0.077	0.065	0.046	0.146	0.178*	0.112
LOFVOL	0.012	0.024	0.032	0.014	0.010	0.021
lBAKETIM	0.102	0.091	0.100	0.193*	0.231**	0.155
lMIXTIM	0.094	0.082	0.097	0.196*	0.211*	0.147

Table 4-2: GS predictions using 2014 materials to predict historical set.

	RRBLUP	GAUSS	PLSR	ELNET	RF	AVE
	Correlation coefficient (r)					
TESTWT	0.157**	0.161**	0.166**	0.118*	0.207***	0.187**
AVGWT	0.240***	0.286***	0.147**	0.163**	0.162**	0.221***
AVGDIA	0.146**	0.163**	0.114	0.121*	0.059	0.134*
GRNPRO	0.076	0.092	0.082	0.079	-0.050	0.069
FLRPRO	0.076	0.122	0.072	0.077	-0.036	0.073
MIXABS	-0.100	-0.090	-0.103	-0.055	-0.032	-0.085
MIXTIM	0.121*	0.119*	0.172**	0.191**	0.236***	0.190**
BAKEABS	0.013	0.024	0.033	0.011	-0.024	0.012
BAKETIM	0.114	0.127*	0.113	0.198***	0.239***	0.177**
LOFVOL	0.110	0.134*	0.086	0.155**	0.004	0.112
lBAKETIM	0.136*	0.165**	0.131*	0.260***	0.327***	0.224***
lMIXTIM	0.146*	0.147*	0.212***	0.255***	0.291***	0.254***

Figures

Figure 4-1: Number of entries from each year represented in genomic selection modeling

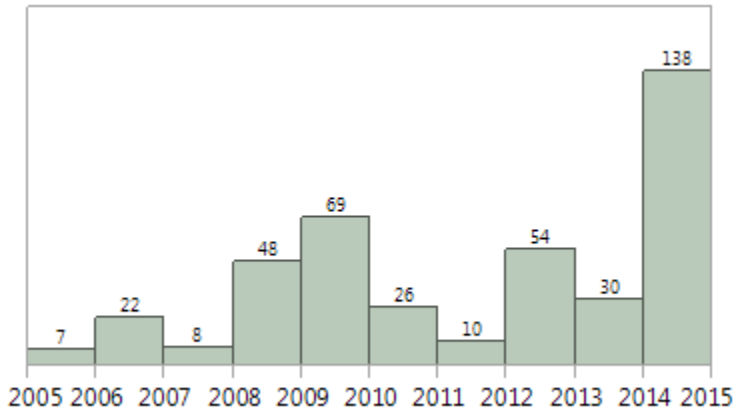
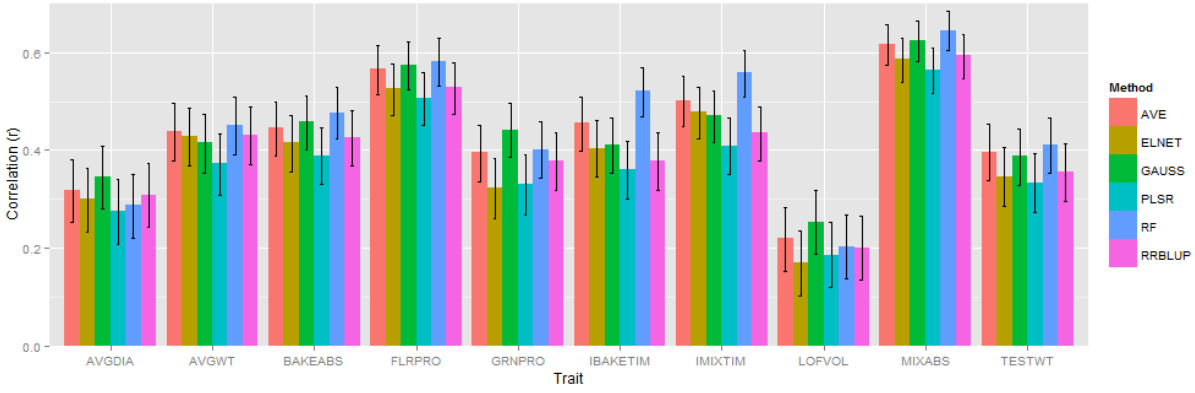


Figure 4-2: Cross validation correlations of genomic selection for all quality traits in 472 entries using 80% to train model and 20% to test. Model was randomly iterated 10 times.



Acknowledgements

Battenfield's PhD work was sponsored by Monsanto's Beachell-Borlaug International Scholars Program. The KSU breeding program and wheat quality testing are funded by the Kansas Wheat Alliance.

References

AACC. 2000. Approved Methods of the American Association of Cereal ChemistsAmer Assn of Cereal Chemists.

Abdel-Aal, E.-S.M. and P. Hucl. 2003. Composition and stability of anthocyanins in blue-grained wheat. Journal of Agricultural and Food Chemistry 51: 2174-2180.

Aisawi, K.A.B., M.P. Reynolds, R.P. Singh and M.J. Foulkes. 2015. The Physiological Basis of the Genetic Progress in Yield Potential of CIMMYT Spring Wheat Cultivars from 1966 to 2009. Crop Science 55: 1749. doi:10.2135/cropsci2014.09.0601.

Barlow, K., M. Buttrose, H. Simmonds and M. Vesk. 1973. The nature of the starch-protein interface in wheat endosperm.

Bates, D., M. Maechler, B. Bolker and S. Walker. 2013. lme4: Linear mixed-effects models using Eigen and S4. R package version 1.

Begum, F., D. Ghosh, G.C. Tseng and E. Feingold. 2012. Comprehensive literature review and statistical considerations for GWAS meta-analysis. Nucleic acids research: gkr1255.

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

Blandino, M., F. Marinaccio, P. Vaccino and A. Reyneri. 2015. Nitrogen Fertilization Strategies Suitable to Achieve the Quality Requirements of Wheat for Biscuit Production. Agronomy Journal 107: 1584-1594. doi:10.2134/agronj14.0627.

Bordes, J., C. Ravel, J. Le Gouis, A. Lapierre, G. Charmet and F. Balfourier. 2011. Use of a global wheat core collection for association analysis of flour and dough quality traits. Journal of Cereal Science 54: 137-147. doi:10.1016/j.jcs.2011.03.004.

Borghi, B., G. Giordani, M. Corbellini, P. Vaccino, M. Guermandi and G. Toderi. 1995. Influence of crop rotation, manure and fertilizers on bread making quality of wheat (Triticum aestivum L.). European journal of agronomy 4: 37-45.

Box, G.E.P.C.D.R. 1964. An Analysis of Transformation. Journal of the Royal Statistical Society 26: 211-252.

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

Branlard, G., J. Pierre and M. Rousset. 1992. Selection indices for quality evaluation in wheat breeding. Theoretical and Applied Genetics 84: 57-64.

Breiman, L. 2001. Random forests. Machine learning 45: 5-32.

Breseghello, F. and M.E. Sorrells. 2006. Association mapping of kernel size and milling quality in wheat (Triticum aestivum L.) cultivars. Genetics 172: 1165-1177. doi:10.1534/genetics.105.044586.

Bushuk, W. 1997. Wheat breeding for end-product use. Wheat: Prospects for global improvement. Springer. p. 203-211.

Cavanagh, C., M. Morell, I. Mackay and W. Powell. 2008. From mutations to MAGIC: resources for gene discovery, validation and delivery in crop plants. Curr Opin Plant Biol 11: 215-221. doi:10.1016/j.pbi.2008.01.002.

Chapman, J.A., M. Mascher, A. Buluç, K. Barry, E. Georganas, A. Session, et al. 2015. A whole-genome shotgun approach for assembling and anchoring the hexaploid bread wheat genome. Genome biology 16: 26.

Claeskens, G. and N.L. Hjort. 2008. Model selection and model averagingCambridge University Press Cambridge.

Cormier, F., M. Throude, C. Ravel, J. Gouis, M. Leveugle, S. Lafarge, et al. 2015. Detection of NAM-A1 Natural Variants in Bread Wheat Reveals Differences in Haplotype Distribution between a Worldwide Core Collection and European Elite Germplasm. Agronomy 5: 143-151. doi:10.3390/agronomy5020143.

Cox, T., M. Shogren, R. Sears, T. Martin and L. Bolte. 1989. Genetic improvement in milling and baking quality of hard red winter wheat cultivars, 1919 to 1988. Crop Science 29: 626-631.

Cox, T., J. Shroyer, L. Ben-Hui, R. Sears and T. Martin. 1988. Genetic improvement in agronomic traits of hard red winter wheat cultivars 1919 to 1987. Crop Science 28: 756-760.

Crossa, J., J. Burgueno, S. Dreisigacker, M. Vargas, S.A. Herrera-Foessel, M. Lillemo, et al. 2007. Association analysis of historical bread wheat germplasm using additive genetic covariance of relatives and population structure. Genetics 177: 1889-1913.

Crossa, J., L. Campos Gde, P. Perez, D. Gianola, J. Burgueno, J.L. Araus, et al. 2010. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186: 713-724. doi:10.1534/genetics.110.118521.

Crossa, J., P. Perez, J. Hickey, J. Burgueno, L. Ornella, J. Ceron-Rojas, et al. 2014. Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity (Edinb) 112: 48-60. doi:10.1038/hdy.2013.16.

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

Delcour, J. and R.C. Hoseney. 2010. Principles of cereal science and technology. status: published.

Edae, E.A., P.F. Byrne, S.D. Haley, M.S. Lopes and M.P. Reynolds. 2014. Genome-wide association mapping of yield and yield components of spring wheat under contrasting moisture regimes. Theor Appl Genet 127: 791-807. doi:10.1007/s00122-013-2257-8.

Endelman, J.B. 2011. Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. The Plant Genome Journal 4: 250. doi:10.3835/plantgenome2011.08.0024.

Faostat, F. 2013. Statistical Databases. Food and Agriculture Organization of the United Nations.

Friedman, J., T. Hastie and R. Tibshirani. 2009. glmnet: Lasso and elastic-net regularized generalized linear models. R package version 1.

Furtado, A., P. Bundock, P. Banks, G. Fox, X. Yin and R. Henry. 2015. A novel highly differentially expressed gene in wheat endosperm associated with bread quality. Scientific reports 5.

Garg, M., H. Singh, H. Kaur and H.S. Dhaliwal. 2006. Genetic Control of High Protein Content and Its Association with Bread-Making Quality in Wheat. Journal of Plant Nutrition 29: 1357-1369. doi:10.1080/01904160600830134.

Gautier, M.-F., M.-E. Aleman, A. Guirao, D. Marion and P. Joudrier. 1994. Triticum aestivum puroindolines, two basic cystine-rich seed proteins: cDNA sequence analysis and developmental gene expression. Plant molecular biology 25: 43-57.

Gaynor, R.C. 2015. GSwGBS: an R package Genomic Selection with Genotyping-by-Sequencing. Genomic Selection for Kansas Wheat. K-State Research Exchange.

Gerland, P., A.E. Raftery, H. Ševčíková, N. Li, D. Gu, T. Spoorenberg, et al. 2014. World population stabilization unlikely this century. Science 346: 234-237.

Giroux, M. and C. Morris. 1997. A glycine to serine change in puroindoline b is associated with wheat grain hardness and low levels of starch-surface friabilin. Theoretical and Applied Genetics 95: 857-864.

Graybosch, R.A., C.J. Peterson, D.R. Shelton and P.S. Baenziger. 1996. Genotypic and environmental modification of wheat flour protein composition in relation to end-use quality. Crop Science 36: 296-300.

Gupta, R. and K. Shepherd. 1990. Two-step one-dimensional SDS-PAGE analysis of LMW subunits of glutelin. Theoretical and Applied Genetics 80: 65-74.

Guzmán, C., A.S. Medina-Larqué, G. Velu, H. González-Santoyo, R.P. Singh, J. Huerta-Espino, et al. 2014. Use of wheat genetic resources to develop biofortified wheat with enhanced grain zinc and iron concentrations and desirable processing quality. Journal of Cereal Science 60: 617-622. doi:10.1016/j.jcs.2014.07.006.

Guzmán, C., G. Posadas-Romano, N. Hernandez-Espinosa, A. Morales-Dorantes and R.J. Pena. 2015. A new standard water absorption criteria based on solvent retention capacity (SRC) to determine dough mixing properties, viscoelasticity, and bread-making quality. Journal of Cereal Science.

Heffner, E.L., J.-L. Jannink, H. Iwata, E. Souza and M.E. Sorrells. 2011. Genomic Selection Accuracy for Grain Quality Traits in Biparental Wheat Populations. Crop Science 51: 2597. doi:10.2135/cropsci2011.05.0253.

Heffner, E.L., J.-L. Jannink and M.E. Sorrells. 2011. Genomic selection accuracy using multifamily prediction models in a wheat breeding program. The Plant Genome 4: 65-75.

Heffner, E.L., M.E. Sorrells and J.-L. Jannink. 2009. Genomic Selection for Crop Improvement. Crop Science 49: 1. doi:10.2135/cropsci2008.08.0512.

Himi, E. and K. Noda. 2005. Red grain colour gene (R) of wheat is a Myb-type transcription factor. Euphytica 143: 239-242.

Holman, J.D., A.J. Schlegel, C.R. Thompson and J.E. Lingenfelser. 2011. Influence of Precipitation, Temperature, and 56 Years on Winter Wheat Yields in Western Kansas. cm 10: 0. doi:10.1094/cm-2011-1229-01-rs.

HWW Quality Targets Committee. 2006. Recommended Quality Targets for Hard Red Winter Wheat.

International Wheat Genome Sequencing, C. 2014. A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. Science 345: 1251788. doi:10.1126/science.1251788.

IWGSC. 2014. A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. Science 345: 1251788. doi:10.1126/science.1251788.

Juliana, P., J.E. Rutkoski, J.A. Poland, R.P. Singh, S. Murugasamy, S. Natesan, et al. 2015. Genome-Wide Association Mapping for Leaf Tip Necrosis and Pseudo-black Chaff in Relation to Durable Rust Resistance in Wheat. The Plant Genome. doi:10.3835/plantgenome2015.01.0002.

Kaur, K., O. Lukow, K. Preston and L. Malcolmson. 2004. How well do early-generation quality tests predict flour performance? Canadian journal of plant science 84: 71-78.

Kuchel, H., P. Langridge, L. Mosionek, K. Williams and S. Jefferies. 2006. The genetic control of milling yield, dough rheology and baking quality of wheat. Theoretical and Applied Genetics 112: 1487-1495.

Langmead, B. and S.L. Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. Nature methods 9: 357-359.

Liaw, A. and M. Wiener. 2002. Classification and regression by randomForest. R news 2: 18-22.

Lillemo, M., F. Chen, X. Xia, M. William, R.J. Peña, R. Trethowan, et al. 2006. Puroindoline grain hardness alleles in CIMMYT bread wheat germplasm. Journal of Cereal Science 44: 86-92.

Liu, L., L. Wang, J. Yao, Y. Zheng and C. Zhao. 2010. Association mapping of six agronomic traits on chromosome 4A of wheat (Triticum aestivum L.). Molecular Plant Breeding 1.

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

Lorenz, A.J., S. Chao, F.G. Asoro, E.L. Heffner, T. Hayashi, H. Iwata, et al. 2011. Genomic Selection in Plant Breeding. 110: 77-123. doi:10.1016/b978-0-12-385531-2.00002-5.

Mascher, M., G.J. Muehlbauer, D.S. Rokhsar, J. Chapman, J. Schmutz, K. Barry, et al. 2013. Anchoring and ordering NGS contig assemblies by population sequencing (POPSEQ). The Plant Journal 76: 718-727.

Matus-Cadiz, M., P. Hucl, C. Perron and R. Tyler. 2003. Genotype× environment interaction for grain color in hard white spring wheat. Crop Science 43: 219-226.

McIntosh, R., K. Devos, J. Dubcovsky and W. Rogers. 2000. Catalogue of gene symbols for wheat: 2000 supplement. Wheat Information Service: 33-70.

McIntosh, R., Y. Yamazaki, J. Dubcovsky, W. Rogers, C. Morris, D. Somers, et al. 2013. MacGene 2012: catalogue of gene symbols for wheat.

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

Mevik, B.-H. and R. Wehrens. 2007. The pls package: principal component and partial least squares regression in R. Journal of Statistical Software 18: 1-24.

Miller, B., B. Hays and J. Johnson. 1956. Correlation of farinograph, mixograph, sedimentation, and baking data for hard red winter wheat flour samples varying widely in quality. Cereal Chemistry 33: 277-290.

Mir, R.R., N. Kumar, V. Jaiswal, N. Girdharwal, M. Prasad, H.S. Balyan, et al. 2012. Genetic dissection of grain weight in bread wheat through quantitative trait locus interval and association mapping. Molecular Breeding 29: 963-972. doi:10.1007/s11032-011-9693-4.

Miyamoto, T. and E. Everson. 1958. Biochemical and physiological studies of wheat seed pigmentation. Agronomy Journal 50: 733-734.

Morris, C.F. 2002. Puroindolines: the molecular genetic basis of wheat grain hardness. Plant molecular biology 48: 633-647.

Neumann, K., B. Kobiljski, S. Denčić, R.K. Varshney and A. Börner. 2010. Genome-wide association mapping: a case study in bread wheat (Triticum aestivum L.). Molecular Breeding 27: 37-58. doi:10.1007/s11032-010-9411-7.

Pasha, I., F. Anjum and C. Morris. 2010. Grain hardness: a major determinant of wheat quality. Food Science and Technology International: 1082013210379691.

Payne, P.I. and G.J. Lawrence. 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: 29-35.

Payne, P.I., M.A. Nightingale, A.F. Krattiger and L.M. Holt. 1987. The relationship between HMW glutenin subunit composition and the bread-making quality of British-grown wheat varieties. Journal of the Science of Food and Agriculture 40: 51-65.

Peña, R. 2002. Wheat for bread and other foods. Bread wheat improvement and production. Food and Agriculture Organization of the United Nations. Rome: 483-542.

Peña, R.J., A. Amaya, S. Rajaram and A. Mujeeb-Kazi. 1990. Variation in quality characteristics associated with some spring 1B/1R translocation wheats. Journal of Cereal Science 12: 105-112.

Peña, R.J., R. Trethowan, W.H. Pfeiffer and M.V. Ginkel. 2002. Quality (End-Use) Improvement in Wheat. Journal of Crop Production 5: 1-37. doi:10.1300/J144v05n01_02.

Poland, J.A., P.J. Brown, M.E. Sorrells and J.L. Jannink. 2012. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS One 7: e32253. doi:10.1371/journal.pone.0032253.

Poland, J.A. and T.W. Rife. 2012. Genotyping-by-Sequencing for Plant Breeding and Genetics. The Plant Genome Journal 5: 92. doi:10.3835/plantgenome2012.05.0005.

R Development Core Team. 2014. R: A lanugauge and environment for statistical computing. R Foundation for Statistical Computing.

Raftery, A.E., M. Kárný and P. Ettler. 2010. Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill. Technometrics 52: 52-66.

Raftery, A.E., D. Madigan and J.A. Hoeting. 1997. Bayesian model averaging for linear regression models. Journal of the American Statistical Association 92: 179-191.

Ragupathy, R., H.A. Naeem, E. Reimer, O.M. Lukow, H.D. Sapirstein and S. Cloutier. 2008. Evolutionary origin of the segmental duplication encompassing the wheat GLU-B1 locus encoding the overexpressed Bx7 (Bx7OE) high molecular weight glutenin subunit. Theor Appl Genet 116: 283-296. doi:10.1007/s00122-007-0666-2.

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

Rutkoski, J., J. Benson, Y. Jia, G. Brown-Guedira, J.-L. Jannink and M. Sorrells. 2012. Evaluation of Genomic Prediction Methods for Fusarium Head Blight Resistance in Wheat. The Plant Genome Journal 5: 51. doi:10.3835/plantgenome2012.02.0001.

Rutkoski, J.E., E.L. Heffner and M.E. Sorrells. 2010. Genomic selection for durable stem rust resistance in wheat. Euphytica 179: 161-173. doi:10.1007/s10681-010-0301-1.

Rutkoski, J.E., J.A. Poland, R.P. Singh, J. Huerta-Espino, S. Bhavani, H. Barbier, et al. 2014. Genomic Selection for Quantitative Adult Plant Stem Rust Resistance in Wheat. The Plant Genome 7: 0. doi:10.3835/plantgenome2014.02.0006.

Saghai-Maroof, M.A., K.M. Soliman, R.A. Jorgensen and R. Allard. 1984. Ribosomal DNA spacer-length polymorphisms in barley: Mendelian inheritance, chromosomal location, and population dynamics. Proceedings of the National Academy of Sciences 81: 8014-8018.

Singh, N., K. Shepherd and G. Cornish. 1991. A simplified SDS—PAGE procedure for separating LMW subunits of glutenin. Journal of Cereal Science 14: 203-208.

Sorrells, M.E., J.P. Gustafson, D. Somers, S. Chao, D. Benscher, G. Guedira-Brown, et al. 2011. Reconstruction of the synthetic W7984× Opata M85 wheat reference population. Genome 54: 875-882.

Terman, G., R. Ramig, A. Dreier and R. Olson. 1969. Yield-protein relationships in wheat grain, as affected by nitrogen and water. Agronomy Journal 61: 755-759.

Uauy, C., A. Distelfeld, T. Fahima, A. Blechl and J. Dubcovsky. 2006. A NAC gene regulating senescence improves grain protein, zinc, and iron content in wheat. Science 314: 1298-1301.

Wang, L., G. Li, R.J. Peña, X. Xia and Z. He. 2010. Development of STS markers and establishment of multiplex PCR for Glu-A3 alleles in common wheat (Triticum aestivum L.). Journal of Cereal Science 51: 305-312. doi:10.1016/j.jcs.2010.01.005.

Wang, L., X. Zhao, Z. He, W. Ma, R. Appels, R. Peña, et al. 2009. Characterization of low-molecular-weight glutenin subunit Glu-B3 genes and development of STS markers in common wheat (Triticum aestivum L.). Theoretical and Applied Genetics 118: 525-539.

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

Yu, J., G. Pressoir, W.H. Briggs, I. Vroh Bi, M. Yamasaki, J.F. Doebley, et al. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38: 203-208. doi:10.1038/ng1702.

Zheng, S., P.F. Byrne, G. Bai, X. Shan, S.D. Reid, S.D. Haley, et al. 2009. Association analysis reveals effects of wheat glutenin alleles and rye translocations on dough-mixing properties. Journal of cereal science 50: 283-290.