Breeding for β-glucan content in elite North American oat (*Avena sativa* L.) using molecular markers

by

Franco G. Asoro

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Plant Breeding

Program of Study Committee:
William D. Beavis, Co-major Professor
M. Paul Scott, Co-major Professor
Dianne Cook
Jack C.M. Dekkers
Maria Salas-Fernandez

Iowa State University

Ames, Iowa

2012

DEDICATION

I would like to dedicate this dissertation to my father Prudencio.

TABLE OF CONTENTS

ACKNOWLEDGEMENT

I would like to take this opportunity to express my sincere gratitude to those who helped me with various aspects of conducting this research and my graduate studies. I apologize in advance if I omit somebody. First, Dr. Jean-Luc Jannink, for his guidance, encouragement, understanding, patience and for being a great mentor. This research has benefited enormously from Dr. Jannink's research insights and extensive knowledge in quantitative genetics. My major advisors Dr. William Beavis and Dr. Paul Scott for providing me with the environment conducive for successful research in plant breeding, for valuable suggestions to improve this research and mentoring me in my graduate career. Dr. Dianne Cook, for her excellent teaching that generated my interests in data visualization and mining. Dr. Jack Dekkers and Dr. Maria Salas-Fernandez, for their willingness in guiding me towards the right direction and offering new perspectives for this research. Dr. Jode Edwards, for providing guidance in the analysis of data. George Patrick for his support in the field experiments and Adrienne Lauter for her assistance in the laboratory. The members of Beavis and Scott Labs for the fun and for sharing their ideas to improve my research. Mark Newell for his exceptional friendship since my first semester at Iowa State and support throughout our graduate studies and the conduct of this dissertation. My friends (Ates/Kuyas, Jan, Pom, Mapi and Fritzie) for providing encouragement. My parents, brothers and sisters for prayers and for encouragement. Finally, my wife Ruby Lynn for the love, support and encouragement throughout the conduct of this research and my graduate studies.

CHAPTER 1. GENERAL INTRODUCTION

Rationale and Significance

Breeding crops for nutritional value can help mitigate some diseases and increase the quality of human life. In maize (*Zea mays* L.), cultivars with high lysine and tryptophan levels have been released in Africa and South America (Cordova, 2000) to increase human intake of these essential amino acids. In rice (*Oryza sativa* L.), an effort is being devoted to develop cultivars with high vitamin A (Al-Babili and Beyer, 2005) to combat blindness in children. In oat (*Avena sativa* L.), breeding for β-glucan content has been a part of cultivar development programs for about two decades (Peterson, 1991) and few varieties have been publicly available (McMullen et al., 2005; Brian Rossnagel, pers.comm.). The oat grains produced in 2007 in North America mostly went into feed products (66%) while only smaller percentage (20%) went into human food consumption, (http://faostat.fao.org, verified 25 Jan 2012).

Mixed linkage (1-3, 1-4)-β-D glucan (referred to as β-glucan) is the soluble fiber component in oats and is found in the endosperm and aleurone portions of oat groats. Food agencies of different countries have approved the claim that β-glucan reduces blood cholesterol levels (Tiwari and Cummins, 2009; Butt et al., 2008; US Food and Drug Administration, 2010) and therefore can reduce the risk of coronary heart disease. In addition, β-glucan improves glycemic response, making it useful for treating diabetes (Bourdon et al., 1999). Prevalence of these diseases are high in developed countries, data collected in the United States alone showed that high LDL-cholesterol levels affected 33.5%

of adults (≥20 years old) 2005-2008 (Center for Disease Control, 2011a) while coronary heart diseases affected 6% of adults (≥18 years old) in 2010 (Center for Disease Control, 2011b). High prevalence was also detected for diabetes which affected 8.3% of children and adults in 2010 (http://www.diabetes.org, verified 26 January 2012). The nutritional benefits of β-glucan content found in oat food products can help decrease the occurrence of these diseases. One potential mechanism on how β-glucan reduces cholesterol is due to its viscosity that reduces cholesterol re-absorption (Butt et al., 2008, Uusitupa et al, 1997). The United States-Federal Drug Administration (US-FDA) recommends that at least 3 grams of β-glucan should be consumed daily from oatmeal or oatbran to have significant decrease in total serum cholesterol. Standard oat cultivars including varieties with claims of high fiber content contain about 4% to 6% β-glucan (Chernyshova et al., 2007; McMullen et al., 2005; http://wheat.pw.usda.gov/ggpages/UE-MOPN.html). Continous breeding for new oat varieties with higher level of β-glucan content can reduce the amount of oat products that needs to be consumed to provide sufficient amount of β-glucan. In addition, greater amount of this compound in the new varieties will add value to the oat crop.

Parallel to the discoveries of the nutritional benefits of β-glucan was the development of inexpensive, abundant and publicly available molecular markers for oat. These molecular marker data can be used to develop new genetic and breeding strategies with improved response to selection. Hence, the focus of this dissertation is to explore those strategies for improving β-glucan content in elite oat germplasm of North America.

Objectives

The objectives of this dissertation are:

1) To conduct genomewide association study for β-glucan content in elite oat varieties of the USA and Canada.
2) To assess the accuracy of genomic selection methods using empirical data from oat quantitative traits.
3) To compare the response to selection for β-glucan content using phenotypic, marker-assisted and genomic selection methods.

Literature Review

Oat Breeding

Oat is a self-pollinated crop that belongs to genus *Avena* which includes 25 other species in which the former is the most economically useful (Forsberg and Shands, 1989). Oat is an allohexaploid (6x=42) with a basic chromosome number of seven and composed of three genomes. The oat crop is distantly related to other small grain crops in the grass family such as rice (*Oryza sativa* L.), wheat (*Triticum aestivum* L.) and barley (*Hordeum vulgare* L.). Improvement of oat in North America is usually conducted by state universities and government research institutions in the US and Canada. Breeding objectives include grain yield, straw strength and yield, test weight, groat percentage, disease resistance, and seed quality traits. Cultivars of oat that are grown commercially are predominantly purelines (Forsberg and Shands, 1989), however, multilines were also used in areas with high disease occurrence (Frey et al., 1977).

The typical steps for oat cultivar development include parent selection, hybridization, inbreeding, and performance testing (Forsberg and Shands, 1989). When a recurrent selection scheme is used, recombination of selected parents and early generation testing is integrated in the program (Schipper and Frey, 1991). A one cycle per year recurrent selection program for oat improvement proposed by Frey et al. (1988) could be applied in β-glucan content improvement. In this method, hybridization and one season of generation advance are conducted in the greenhouse while hill-plot evaluation of early-generation lines (S0:1 or F2:3) is conducted in the field during summer. Advanced performance tests for oat in the US are conducted by a cooperative network of the United States-Department of Agriculture (USDA), state universities and public research institutions (http://wheat.pw.usda.gov/ggpages/UE-MOPN.html). These are grouped into the Uniform Early Oat Performance Nursery (UEOPN) and the Uniform Mid-season Oat Performance Nursery (UMOPN). Another advanced performance test for oat is conducted through the Quaker Uniform Oat Nursery (QUON). The QUON is a cooperative testing network for oats among northern US State Agricultural Experimental Stations, USDA-ARS and public breeding institutions in Canada.

Breeding for β-Glucan Content in Oats

Previous inheritance studies have revealed that β-glucan content in oat is governed primarily by additive gene action (Holthaus, 1996; Kibite and Edney, 1998; Chernyshova et al. 2007). This indicates that recurrent selection methods that exploit additive effects can be used to improve oat populations for increased β-glucan content. Previous studies reported low to medium broad sense heritability estimates for β-glucan content in oats, for example,

values of 0.55 (Holthaus et al., 1996), 0.45 to 0.58 (Kibite and Edney, 1998), and 0.39 (Chernyshova et al. 2007) on a plot basis have been estimated.

Previous experiments have demonstrated that phenotypic selection is effective for improving β-glucan content in oat. For example, Cervantes-Martinez et al. (2001) developed a broad-based population by intermating 23 oat lines and commercial cultivars with high β-glucan content for three generations followed by one cycle of recurrent selection for β-glucan content. In another study, Chernyshova et al. (2007) conducted population development by crossing high β-glucan parents with agronomically superior parents in the US using single seed descent to advance the lines.

Genotype by environment (G x E) interaction studies for β-glucan content have not been conclusive. Peterson et al. (2005) detected no significant GxE for oat β-glucan using 33 genotypes and nine environments. However, results from several studies suggested that β-glucan content is affected by rainfall, drought conditions, and nitrogen levels (Tiwari and Cummins, 2009). Regarding the effect of β-glucan on grain yield, a low level of correlation was identified between the two traits after a few cycles of selection for β-glucan (Cervantes-Martinez, 2002; Chernyshova et al., 2007). In addition, Peterson (2005) reported that β-glucan content has weak or insignificant correlations with other quality traits such as groat protein and oil percentage. These studies imply breeding for β-glucan content alone will not result in undesirable changes in other important traits of oat.

Linkage Disequilibrium

Gametic phase disequilibrium, also known as linkage disequilibrium (LD), is the non-independence of alleles at two loci in a population (Falconer and Mackay, 1996). Factors that

affect LD include mutation, migration, drift, selection, and population structure (Waugh et al., 2009). Based on previous studies, there is higher LD and population structure in elite cultivars than in non-elite germplasm (Gaut and Long, 2003; Flint-Garcia, 2002). The breeding methods used to generate new populations, in which the best performing lines in each generation are repeatedly crossed, could explain the high LD because of the limited effective recombination among individuals. These methods also result in population structure because some alleles are more prevalent in some subgroups (Breseghello and Sorrels, 2006; and Waugh et al., 2009). For example, different maize population types showed different patterns of LD decay - commercial inbred lines decay within 100 kb on average, diverse inbred lines decay within 2 kb, and open-pollinated landraces within 1 kb (Tenaillon et al., 2001; Remington et al., 2001; Ching et al., 2002). Likewise in barley, LD in diverse germplasm decays faster than in elite inbred lines (Abdurakhmonov and Abdukarimov, 2008). Results from these studies are in agreement with the theoretical expectation for LD ($[r^2 = 1/1+4N_ec]$, where r^2 is the LD, N_e is the effective population size and c is the recombination frequency), indicating that LD is repeatedly broken down by recombination (Sved, 1971).

Population Structure and Relatedness for Association Study

Population structure is a confounding factor in genome-wide association studies (GWAS). Population structure is the presence of differentially related individuals in a population (Waugh et al., 2009). In breeding programs, this occurs because population development is limited to parents with favorable alleles. Although population structure is common in a breeding program, one extreme case of it is the formation of heterotic groups in

hybrid breeding. Two methods for characterizing population structure using marker data are STRUCTURE (Pritchard et al. 2000) and Principal Component Analysis (PCA, Zhao et al. 2007). Both the probability of subgroup membership in STRUCTURE and the principal components in PCA can be used as covariates to adjust the phenotype of individuals for population structure. However, the PCA method is more practical to use for large data sets because of computational issues in STRUCTURE (Sneller at al., 2009). For PCA, each PC reflects an independent axis of relative ancestry (Patterson et al., 2006). Moreover, individuals with high absolute PC scores for a particular PC axis are correlated to that axis, while individuals with ancestry unrelated to that axis will have PC scores close to zero (Sneller et al., 2009).

In the mixed model analysis ($y=Xb + Zu +$ error, where y is the response, Xb is the fixed effects term, and Zu is the random term) used for GWAS, polygenic random effects are modelled based on the assumption that individuals who share many alleles resemble each other phenotypically (Zhao et al., 2007). In the mixed model notation, the covariance of polygenic effects is defined as $\text{Var}(u) = K\sigma^2_u$, (where K is the kinship matrix and σ^2_u is the genetic variance), which implies that the random deviation of individuals (u) from the population mean is constrained by genetic relationships or kinship. In other words, if the kinship of two individuals is high, then random effects of those two individuals should have similar values or a high covariance. Because of this genotype-phenotype covariance, many markers may appear to be associated with the trait when in fact they are just capturing relatedness (Myles et al., 2009). Therefore, accounting for kinship in the estimation of fixed marker effects reduces false positives and eliminates bias in marker effects (Kennedy et al., 1992). On the other hand, bias occurs in a simple model (excluding population structure)

because it assumes that individuals will not share alleles of other QTL influencing the trait resulting in correlated residuals (Kennedy et al., 1992). For association mapping purposes, these relationships can be accounted for by modelling the covariance among individuals as defined by the kinship matrix (Yu et al., 2006).

Marker-Assisted Selection

The term marker-assisted selection (MAS) can be used to (1) describe the use of markers in order to identify progenies with donor alleles similar to marker-assisted backcrossing (MAB), (2) denote selection of parents based on marker-based genetic distance or, (3) define the use of markers in selecting parents or progenies in populations for the highest breeding values (Lande and Thompson, 1990). In the implementation of MAS similar to Lande and Thompson (1990), important markers are first identified using traditional QTL mapping or GWAS. Marker effects due to QTL are then estimated and used in computing marker scores. Lastly, selection of the best genotypes based purely on marker scores or in combination with phenotype data is conducted. The marker score is defined as the sum of effects associated with a marker or QTL allele. As shown in the analytical experiment of Lande and Thompson (1990), using significant marker scores alone to denote the estimated breeding value of an individual is efficient only when the proportion of genetic variance explained by the markers is high. On the other hand, combining marker scores and phenotypic values in an index is efficient over phenotypic selection when the heritability of the trait is low and variance explained by markers is moderate to high (Lande and Thompson, 1990, Holland, 2004; Dekkers and Hospital, 2002). In this index, the phenotype reflects the collective effect of genes while the marker information provides the genetic values at the

QTL (Dekkers and Settar, 2004). Marker-assisted selection can be advantageous over phenotypic selection if cost of genotyping is low, the phenotype is difficult to measure and if MAS can be conducted in off-season locations (Holland, 2004; Hospital et al., 1997). However, hindrances in effective application of MAS in large breeding programs may include a lack of consistency of LD between marker and QTL across populations, (Xu and Crouch, 2008), QTL x environment interaction (Moreau et al., 2004), and overestimation of QTL effects ("Beavis effect") due to population size and significance testing issues (Beavis, 1994).

One alternative to traditional MAS is genomic selection (GS) or genome-wide selection proposed by Meuwissen et al. (2001) and defined as the use of all markers densely positioned across the genome to predict the total genetic value of an individual. Technically, GS can just be an extension of marker-based selection (i.e. MAS using only markers), except that markers are not pre-selected for significance. This means that significance testing for markers is not conducted and focuses only on defining the breeding values of individuals. Since all markers are included, GS can account for greater genetic variance resulting in better estimates of breeding values (Solberg et al., 2008). However, the accuracy of GS is affected by marker density, training population size, genetic relationships of training and selection candidates, and consistency of LD across populations (Meuwissen et al., 2001; Heffner et al., 2009; Zhong et al., 2009; Habier et al., 2007; Lorenz et al., 2011).

Dissertation Organization

This dissertation consists of six chapters. The first chapter is the general introduction that discusses the importance of β-glucan content and the general purpose of this research,

including a review of the relevant literature on the use of markers in plant breeding. The second chapter focuses on using GWAS to identify molecular markers associated with β-glucan content in elite oats. The third chapter investigates different training population designs and the value of genomic selection for the genetic improvement of various oat quantitative traits including β-glucan content. The fourth and fifth chapters compare the three selection methods (BLUP phenotypic selection, MAS and GS) in terms of response to selection for β-glucan content and maintenance of genetic variance. Lastly, the sixth chapter consists of the general conclusion which describes the brief outline of the research as a whole and the implications of results of all chapters taken jointly. Chapters two, three, four and five are organized to include the following sections: introduction, materials and methods, results, discussion and references.

References

Abdurakhmonov, I.Y., and A. Abdukarimov. 2008. Application of association mapping to understanding the genetic diversity of plant germplasm resources. International Journal of Plant Genomics, Article ID 574927, 18 pages. doi:10.1155/2008/574927.

Al-Babili, S., and P. Beyer. 2005. Golden Rice—Five years on the road—Five years to go?. Trends Plant Sci. 10:565–573.

Beavis, W.D. 1994 .The power and deceit of QTL experiments: lessons from comparative QTL studies, pp. 250–265 in Proceedings of the 49th Annual Corn and Sorghum Research Conference, edited by D. B. Wilkinson. American Seed Trade Association, Washington, DC.

Bourdon, I., W. Yokoyama, P. Davis, C. Hudson, R. Backus, D. Richter, B. Knuckles and B. Schneeman. 1999. Postprandial lipid, glucose, insulin, and cholecystokinin responses in men fed barley pasta enriched with β-glucan. Am J Clin Nutr. 69:55-63.

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

Cervantes-Martinez, C.T., K.J. Frey, P.J. White, D.M. Wesenberg, and J.B. Holland. 2001. Selection for greater β-glucan content in oat grain. Crop Sci. 41:1085–1091.

Chernyshova, A.A., P.J. White, M.P. Scott, and J.-L. Jannink. 2007. Selection for nutritional function and agronomic performance in oat. Crop Sci. 47:2330–2339.

Ching, A., K.S. Caldwell, M. Jung, M. Dolan, O.S. Smith, S. Tingey, M. Morgante, and A.J. Rafalski. 2002. SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. BMC Genet. 3:19.

Dekkers, J.C.M., and F. Hospital. 2002. Utilization of molecular genetics in genetic improvement of plants and animals. Nature Reviews: Genetics 3:22-32.

Dekkers, J.C.M., and P. Settar. 2004. Long-term selection with known quantitative trait loci. Plant Breeding Reviews. Wiley. Plant breeding reviews, Volume 24, Part 1, Long Term Selection: Maize. edited by Jules Janick. John Wiley&Sons, Inc.

Falconer, D.S., and T.F.C. Mackay. 1996. Introduction to quantitative genetics. 4th ed. Longman Technical and Scientific, Essex, UK.

Flint-Garcia, S.A., J.M. Thornsberry, and E.S. Buckler. 2003. Structure of linkage disequilibrium in plants. Annu. Rev. Plant Biol. 54:357–374.

Forsberg, R.A., and Shands, H.L. 1989. Oat breeding. In J. Janick (ed). Plant breeding reviews. Vol. 6, pp. 167 - 207. Portland, OR, USA, Timber Press.

Frey, K.J., J.K. McFerson, and C.V. Branson. 1988. A procedure for one cycle of recurrent selection per year with spring-sown small grains. Crop Sci. 28:855-856.

Gaut, B.S., and A.D. Long. 2003.The Lowdown on linkage disequilibrium. Plant Cell. 15:1502-1506.

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

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

Holland, J.B. 2004. Implementation of molecular markers for quantitative traits in breeding programs—Challenges and opportunities. In New directions for a diverse planet: Proc. Intl. Crop Sci. Congress, 4th,Brisbane, Australia. 26 Sept. – 1 Oct. 2004. The Regional Institute Ltd., Gosford, NSW, Australia.

Holthaus, J.F., J.B. Holland, P.J. White, and K.J. Frey. 1996. Inheritance of β-glucan content of oat grain. Crop Sci. 36:567–572.

Hospital, F., L. Moreau, F. Lacoudre, A. Charcosset, and A. Gallais, 1997. More on the efficiency of marker-assisted selection. Theor Appl Genet. 95: 1181–1189.

Kennedy, B.W., M. Quinton, and J.A.M. Vanarendonk. 1992. Estimation of effects of single genes on quantitative traits. J. Anim. Sci. 70: 2000–2012.

Kibite, S., and M.J. Edney. 1998. The inheritance of ß-glucan concentration in three oat (*Avena sativa* L.) crosses. Can. J. Plant Sci. 78:245–250.

Lande, R., and R. Thompson. 1990. Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124:743–756.

Lorenz, A., S. Chao, F. Asoro, E. Heffner, T. Hayashi, H. Iwata, K. Smith, M. Sorrels, and J.-L. Jannink. 2011. Genomic selection in plant breeding: knowledge and prospects. In: D. L. Sparks (Ed.), Advances in Agronomy, Academic Press, San Diego, CA USA. pp. 77-123.

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

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

Moreau, L., A. Charcosset, and A. Gallais. 2004. Experimental evaluation of several cycles of marker-assisted selection in maize. Euphytica 137:111–118.

Patterson N, A.L. Price, and D. Reich. 2006. Population structure and eigenanalysis. PLoS Genet 2: e190. doi:10.1371/journal.pgen.0020190.

Peterson, D.M., D.M. Wesenberg, and D.E. Burrup. 1995. β-glucan content and its relationship to agronomic characteristics in elite oat germplasm. Crop Sci. 35:965–970.

Pfeiffer, W.H., and B. McClafferty. 2007. HarvestPlus: Breeding crops for better nutrition. Crop Sci. 47: S-88-S-105.

Pritchard, J.K., M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945–959.

Remington, D.L., J.M. Thornsberry, Y. Matsuoka, L.M. Wilson, S.R. Whitt, J. Doebley, S. Kresovich, M.M. Goodman, and E.S. Buckler, IV. 2001. Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc. Natl. Acad. Sci. USA 98:11479–11484.

Schipper, H., and K.J. Frey. 1991. Observed gains from three recurrent selection regimes for increased groat-oil content of oat. Crop Sci. 31:1505-1510.

Solberg, T.R., A.K. Sonesson, J.A. Woolliams, and T.H.E. Meuwissen. 2008. Genomic selection using different marker types and densities. J Anim Sci 2008.86:2447-2454.

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

Sved, J.A. 1971. Linkage disequilibrium and homozygosity of chromosome segments in finite populations. Theor. Popul. Biol. 2:125–141.

Tenaillon, M.I., M.C. Sawkins, A.D. Long, R.L. Gaut, J.F. Doebley, and B.S. Gaut. 2001. Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). Proc. Natl. Acad. Sci. USA 98:9161–9166.

Tiwari, U., and E. Cummins. 2009. Factors influencing β-glucan levels and molecular weight in cereal-based products. Cereal Chemistry 86: 290–301.

Waugh, R., J.L. Jannink, G.J. Muehlbauer, and L. Ramsay. The emergence of whole genome association scans in barley. Current Opinion in Plant Biology. 12 (2):218-222.

Xu, Y., and J.H. Crouch. 2008. Marker-assisted selection in plant breeding: From publications to practice. Crop Sci. 48:391.

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

Zhao, K., M.J. Aranzana, S. Kim, C. Lister, C. Shindo, C. Tang, C. Toomajian, H. Zheng, C. Dean, P. Marjoram, and M. Nordborg. 2007. An arabidopsis example of association mapping in structured samples. PLoS Genet 3:e4.

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

CHAPTER 2. GENOMEWIDE ASSOCIATION STUDY FOR β-GLUCAN CONTENT IN NORTH AMERICAN ELITE OAT

A paper submitted to the journal Crop Science[1]

Franco G. Asoro[2,5], Mark A. Newell[2], M. Paul Scott[2,3] , William D. Beavis[2] and Jean-Luc Jannink[4]*

Abstract

Genome wide-association studies (GWAS) can be a useful approach to detect quantitative trait loci (QTL) controlling complex traits in crop plants. Oat (*Avena sativa* L.) β-glucan is a soluble dietary fiber and has been shown to have positive health benefits. We report a GWAS involving 446 elite oat breeding lines from North America genotyped with 1005 DArT markers and with phenotypic data from both historical and balanced two-year data. Association analyses accounting for pair-wise relationships and population structure were conducted using single marker tests and LASSO. Single marker tests yielded six and 15 significant markers for the historical and balanced data sets, respectively. The LASSO method selected 24 and 37 markers as the most important in explaining β-glucan content for the historical and balanced data sets, respectively. Comparisons of genetic location showed that 15 of the markers in our study were found on the same linkage groups as QTL identified

---

[1] Submitted to Crop Science Journal.

[2] Department of Agronomy, Iowa State University, Ames, IA, 50011.
[3] USDA-ARS, Corn Insects and Crop Genetics Research Unit, Ames, IA, USA 50011
[4] USDA-ARS, R.W. Holley Center for Agriculture and Health, Department of Plant Breeding and Genetics, Cornell University, Ithaca, New York, 14853, USA.
[5] Primary researcher and author.
*Corresponding author(jeanluc.jannink@ars.usda.gov).

in previous studies. Four of the markers co-localized to within 4 cM of three previously detected QTL, suggesting concordance between QTL detected in our study and previous studies. Two of significant markers were also adjacent to a β-glucan candidate gene in the rice genome. Our findings suggest that GWAS can be used for QTL detection for the purpose of gene discovery and for marker-assisted selection to improve β-glucan content in elite oat.

Introduction

Crop improvement for increased nutritional value is an important objective for breeding programs. In oat (*Avena sativa* L.), breeding for mixed-linkage-(1,3;1,4)-β-D-glucan (referred to as β-glucan) content has been an objective for over two decades in North America (Peterson, 1991). β-glucan is a soluble fiber component that is found in endosperm and in the aleurone layer of oat groat (Butt et al., 2008). Food agencies from Sweden, United Kingdom, Finland, and the Netherlands have approved the claim that β-glucan reduces blood cholesterol levels while the United States-Food and Drug Administration approved the claim that β-glucan decreases the risk of coronary heart disease (Tiwari and Cummins, 2009; FDA Health Claim 21CFR101.81). The reports on the positive health implications of oats when consumed as a whole grain are happening as plant breeding technologies are also rapidly evolving. Foremost are rapid and high density genotyping technologies (e.g, DArT markers, Tinker et al. 2009) and new statistical approaches to analyze the large amount of data that is being generated. The availability of high density marker data enables high resolution mapping of QTL controlling complex traits like β-glucan. Although traditional QTL mapping for β-glucan has been conducted in biparental oat populations (Kianian et al. 2000; De

Koeyer et al. 2004), Genomewide association studies (GWAS) has yet to be implemented for QTL detection in elite oat germplasm.

Genomewide association studies detect associations due to gametic phase disequilibrium between a marker allele and the causative QTL allele. Gametic phase disequilibrium, also known as linkage disequilibrium (LD), between loci is the non-independence of alleles at two loci in a population (Falconer and Mackay, 1996). The high-resolution mapping potential in GWAS relies on the availability of high-density genotyping technology and less LD in panels of unrelated lines than in biparental families. However, differential genetic relationships between individuals, in the form of different pedigree relationships or of subpopulation structure, can lead to false positives in GWAS. Thus, accounting for population structure and polygenic effects in association tests is important (Yu et al., 2006; Stich et al., 2008).

Single marker tests for GWAS like the unified mixed-model approach (Yu et al., 2006; Zhao et al., 2007) have been successfully implemented with a suitable correction for multiple testing. Given that complex traits are controlled by multiple QTL in concert, an alternative strategy would be to include all markers in a regression model. However, since the number of markers is usually larger than number of observations, applying an ordinary multiple regression model for variable selection would be impossible due to an insufficient number of degrees of freedom. One solution is to use penalty parameters in a linear model by employing a method such as the least absolute shrinkage and selection operator (LASSO, Tibshirani, 1996). Briefly, a penalized regression minimizes a function with two components: 1. the squared deviation between the phenotype and its prediction and 2. a penalty that increases with the magnitude of the regression coefficients. The LASSO solves the L1-norm

penalized regression, meaning that the penalty is the sum of the absolute values of the regression coefficients weighted by a parameter lambda: the larger the value of lambda the more markers with zero effect will be in the model. Therefore, LASSO can both do shrinkage and marker selection. This makes LASSO an attractive approach for GWAS since markers with small effects are shrunken to zero, resulting in a sparse model, while only markers with large effects are retained (Wu et al., 2009). One criticism of LASSO is that it may over-shrink markers with large effects, resulting in reduced prediction accuracy. However, this should not be a problem when the objective is merely to identify the associations of marker variability with trait variability (Ayers and Cordell, 2010).

Newell et al. (2010) explored the genome-wide LD in a world collection of oat germplasm and results suggested that GWAS in oats is feasible for QTL detection. Application of GWAS where elite germplasm is used as the association panel provides immediate inference for cultivar development programs (Breseghello and Sorrels, 2006). Consequently, markers that are identified can readily serve as a basis for selection in cultivar development (Bernardo, 2008). The use of breeding lines and cultivars also leads to the opportunity to use phenotypic data routinely collected for plant breeding purposes. However, the data from such programs are highly unbalanced, as a new set of lines is entered every year and only a few lines overlap between years. Although linear mixed-models are robust to this kind of situation, it would be beneficial to determine if a balanced data set from limited environments would be useful for GWAS. Using oat cultivars and breeding lines from the United States and Canada as GWAS panel, our objectives were to:

1. Assess population structure of elite oat lines as it relates to β-glucan content.

2. Apply GWAS using single and multiple marker model tests for β-glucan content.
3. Compare significant associations to previous β-glucan QTL studies and to rice candidate genes to develop a prioritized list for further study.

Materials and Methods

Genotype Data, Kinship and Population Structure

Seeds of 470 oat lines from breeding programs in the USA and Canada were planted in the greenhouse in January 2008. Plants were grown and leaf samples were collected for each entry for DNA extraction according to recommended protocols (Diversity Arrays Technology, 2011). Then the harvested seeds were grown in the field as increase hills from April to July 2008 at the Agronomy Farm, near Ames, IA.

Deoxyribonucleic acid samples of the 470 inbred lines were submitted to DArT PL for genotyping (Yarralumla, Australia) of which 446 produced high quality genotypic data. DArT marker redundancies were removed as described in Asoro et al. (2011). The genotypic data was used to compute the kinship matrix (K), defined as the proportion of common alleles shared by any oat line using the emma.kinship function in the emma (Kang et al., 2008) package implemented in the R software (R Development Core Team, 2011). A matrix estimating population structure, denoted as P, was calculated using principal components analysis (PCA) on the marker data and retaining the first five components through scree test (Cattell, 1966).

Analysis of Data from Uniform Performance Trials in North America

Phenotypic data for β-glucan from the Uniform Oat Performance Nurseries and the Quaker Uniform Oat Nurseries stored in the Graingenes 2.0 database (Carollo et al., 2005) was used for analysis. In addition, 18 lines with phenotypic data from previous research conducted at Iowa State University (Colleoni-Sirghie et al., 2004; Chernyshova et al., 2007), two lines with phenotypic data from North Dakota State University, and seven cultivars with phenotypic data from the National Plant Germplasm System (http://www.ars-grin.gov/npgs/acc/acc_queries.html verified Nov 16 2011) were included. Data from the same lines with different listed names were merged under one entry name and confirmed thru the Pedigree of Oat Lines (POOL) database (Tinker and Deyl, 2005). In total, the data consisted of 450 lines (446 genotyped lines plus 4 long term checks used for the phenotypic analysis) based on 2,909 observations and 129 environment combinations of test years (1994–2007) and locations in the US and Canada. The four long term checks were lines used in oat performance nurseries and were included to provide overlap across environments.

One common strategy to analyze highly unbalanced data sets is to employ a mixed-model approach and use the best linear unbiased prediction (BLUP) for each line as the response variable for GWAS (Zhang et al., 2009). However, BLUP values are shrunken towards the mean and the amount of shrinkage is dependent on the number of data points per individual. Because our data was highly unbalanced, differential shrinkage means that the trait would in effect be measured on a different scale for each observation, leading to reduced power and higher effect estimation error (Garrick et al. 2009). To avoid these shortcomings, we first fitted our data with the following mixed model:

$$y = mean + environment + oat\ lines + error$$

where *y* are the β-glucan observations (expressed in %), population *mean* and *environment* were considered fixed effects, and *oat lines* were considered random effects. The covariance matrix of *oat lines* was assumed proportional to the kinship matrix computed above. The mixed model was fitted using the kinship.BLUP function in rrBLUP package (Endelman, 2011). Raw phenotypes (*y*) were corrected for the fixed effects estimated from the model to derive the values for the observations corrected for environment. Finally, the sample mean of the corrected observations for each oat line was computed and used as the phenotypic value for GWAS (denoted y*). This value is referred to here as the OPN (oat performance nurseries) value. It measures β-glucan content without differential shrinkage despite large differences in replication across lines.

Analysis of β-glucan data from Ames 2009 and 2010

Balanced data sets for the elite lines came from field experiments that were conducted at the Agronomy Farm, ISU from April to July 2009 and April to July 2010. For 2009 and 2010, each hill plot consisted of seed collected from the 2008 field season. The source for each line in the 2008 field season was from the original seed source that was genotyped in January 2008 in the greenhouse. A total of 475 oat lines consisting of the 470 lines mentioned above plus checks were planted in two replicates using an incomplete block design where each incomplete block was composed of 25 hills arranged in a 5x5 grid. Heads were manually harvested and threshed after one week of drying in the field. Oats were dehulled using a Codema Laboratory dehuller (Codema LLC, MN) and grounded into flour in 15ml polycarbonate vials containing two 9.5 mm ball bearings (OPS Diagnostics LLC, Lebanon, NJ) using a reciprocating shaker (Talboys HT Homogenizer, Troemner, Thorofare,

NJ). B-glucan content (as percent on a dry-weight basis) was then measured using the streamlined mixed-linkage β-glucan enzymatic laboratory kit from Megazyme (Megazyme Inc., Wicklow, Ireland) that was improved for high-throughput analysis in a 96-well plate (Newell et al., in review).

To correct for fixed effects due to plate differences, the samples from a given incomplete block were analyzed on the same plate, thus confounding plate and incomplete block effects. The observations from two years were combined using the fixed-effects model:

*y = mean + year + replication + incomplete block (rep*year) + oat lines + error*

Statistical analysis was done using PROC MIXED in SAS Version 9.2 (SAS Institute, 2010) and least square means of *oat lines* were treated as the phenotypic values (y*) and referred to as the Ames values.

PK Model for Single Marker Association Analysis

The mixed model for each marker in the association analysis (Yu et al., 2006) is as follows:

$y^* = \mu + marker + population\ structure + polygenic\ effect\ of\ oat\ line + error$

where y^* is a vector of adjusted phenotypic data from either the OPN or the Ames data source, μ represents the population mean, *marker* is the fixed marker effect, *population structure* fixed effects are the first five PC scores, *polygenic effect of oat lines* is a random effect, and *error* is the random residual error. The variance of the polygenic effect is assumed to be equal to KV_A, where K is the kinship matrix of oat lines and V_A is the additive variance due to polygenic effects. The mixed linear model for association analysis was implemented by modifying the GWA function within the rrBLUP package. The modification was done to

include output for *p*-values, R^2 and marker effects. The false discovery rate (FDR; Benjamini and Hochberg, 1995) for multiple testing was applied to the *p*-values for marker effects from the PK model. We used a relaxed FDR of 0.33 to identify more markers that we subsequently filtered based on other criteria such as BLAST homology and comparison to previous biparental mapping studies for β-glucan content (Kianian et al. 2000; De Koeyer et al. 2004). Lastly, LD (measured as r^2) among the significant markers were calculated to determine if the significant markers were likely capturing effects from the same causal locus.

Mixed Model LASSO

A mixed model LASSO method proposed by Wang et al. (2010) for GWAS in plants was applied in this study. The R function 'amltest' (Dong Wang, personal communication) was modified so that marker effects would not be weighted. The objective of the mixed model LASSO is to estimate the marker and population structure effects that minimize the following equation:

$$(y^* - X^T\beta)^T V^{-1}(y^* - X^T\beta) + \lambda \sum_{j=1}^{p} |\beta|$$

where *X* is the matrix containing all markers and first five principal component axes as predictors, β is a vector of predictor effects, y* is the β-glucan data (response variable), and lambda (λ) is the penalty parameter. The V^{-1} is defined as $\sigma_g^2 ZKZ^T + \sigma_e^2 I$, where Z is the design matrix for observations, K is the kinship of all lines described above, and σ_g^2 and σ_e^2

are the genetic variance of oat lines and residual variance, respectively. As a note, the ordinary LASSO does not contain the V^{-1} term (Tibshirani, 1996).

The mixed model LASSO procedure proposed by Wang et al (2010) was applied in our study using the following:

1) The algorithm started from an ordinary LASSO on y* and the X matrix to reduce the number of variables. The first 50 predictors that entered into the LASSO solution path denoted as X_q were chosen.

2) Starting from a model with no markers up to the model with 50 markers, one marker was added every iteration based on the order in the LASSO solution path, by doing the following:

i. using the estimates of fixed effects from the previous iteration, variance components were calculated by maximum likelihood, the V^{-1} matrix was obtained, and the AIC value was calculated.

ii. y* is adjusted such that $\tilde{y} = V^{-1/2}y^*$ and the X matrix is adjusted in a similar fashion: $\tilde{X} = V^{-1/2}X$ where $\tilde{y} \sim N(\tilde{X}\beta, I_n)$. The ordinary LASSO was then applied to $\tilde{y}$ and $\tilde{X}$ using the LARS algorithm of Efron et al. (2004).

3) Lastly, the AIC was used to determine the final model and the algorithm was re-run with only the final set of markers to determine the marker effects. As a consequence of LASSO, the entry order of markers becomes important since once the marker enters the model it usually remains in the model (Wu et al., 2009). Entry order

can thus be used for ranking the markers in terms of their importance(Sung et al., 2009).

Comparison to Previous QTL Studies and BLAST Homology Search

Markers associated in this study were compared to previous QTL studies conducted in biparental populations. The location of the significant markers from this study and previously identified markers linked to β-glucan QTL from Kianian et al. (2000), De Koeyer et al. (2004) and Groh et al. (2001) were compared based on the updated Kanota x Ogle map (Tinker et al., 2009).

To determine whether our study and previous studies picked up some of the same causal loci, we tested whether the positions of our associated markers were more likely than a random sample of positions to fall on the same linkage groups as QTL from previous studies. The updated map is 1989 cM long while linkage groups on which β-glucan QTL have been identified total 828 cM. We therefore compared the fraction of our associated markers that were on linkage groups with previous β-glucan QTL with the binomial distribution with success probability 828 / 1989 = 0.416.

The nucleotide sequences of all oat DArT markers (Tinker et al., 2009) significant in this study plus the sequences of markers in perfect LD with those markers (Supplementary Table 1) were used in a BLASTn analysis against rice annotated sequences (Ouyang et al., 2007). The search was limited to an E-value cut off of 1×10^{-15}. The location of all rice cellulose synthase and cellulose synthase like genes were also determined. Then the location of the DArT marker sequences homologs and the rice candidate genes for β-glucan were then compared.

To establish a threshold value for proximity, a point was chosen at random in the rice genome (~370,000 kb) and the distance in kb between that point and the nearest rice candidate gene was determined. This process was conducted 1 million times to construct a distribution of distances under the null hypothesis that DArT marker homolog positions were random relative to rice candidate genes. The distance at the 5% quantile of this null distribution was 247 kb and was taken as the threshold value for adjacency to a rice candidate gene.

Results

β-glucan Content Data and Population Structure in Elite Oats

Phenotypic values for β-glucan content used for the association analysis from two datasets (OPN and Ames) were significantly correlated (r=0.71, Table 1). The two data sets had the same standard error of the mean of 0.03. The oat line variance was higher in Ames (0.45) than for the OPN data set (0.19) but the two had comparable residual variances (0.25 – 0.26) . The broad sense heritability was therefore higher in the Ames than in the OPN data source (0.63 and 0.43, respectively).

The clustering method proposed by Newell et al. (in review) resulted in five clusters using the k-means method. The number of lines for each cluster ranged from 66 oat lines in the Triple Crown Cluster to 105 in the Ogle Cluster (Table 2). Principal component analysis showed that the first five PCs explained 23% of the marker variation (data not shown). Visualization of the clusters in a scatter plot of PC1 vs PC2 and PC1 vs PC3 showed distinct separation of clusters with minimal overlap (Figure 1).

Means for β-glucan content per cluster (Table 2) showed significant differences based on ANOVA ($p<0.0001$). The AC Assiniboia Cluster had the lowest β-glucan (3.79 for OPN and 4.78 for Ames) while the Ogle Cluster had the highest (4.46 for OPN and 5.39 for Ames). The correlations of β-glucan content with PC1, PC2 and PC5 scores were significant at $p < 0.05$. For the OPN data set, the correlations with the PCs were -0.32, -0.16 and 0.25, respectively for the significant PCs. For the Ames data set, the correlations were -0.33, -0.11 and 0.22, respectively, for these PCs.

PK Single Test Association

The comparison of $-\log_{10}$(p-value) from the two data sets showed that there are more significant markers in the Ames data set for any nominal *p*-value cut-off (Supplementary Figure 1). An FDR of 0.33 resulted in six significant markers for the OPN data set and 15 for the Ames data set (Table 3). Out of 21 markers identified as significant from the two data sets, there were four common markers, oPt.12985, oPt.14067, oPt.16436 and oPt.18130. These four common markers have consistent direction of marker effects across datasets. The fraction of the phenotypic variance explained (R^2) by the significant markers ranged from 2.1 to 2.8 % for the OPN data and 1.7 to 3.2 % for the Ames data (data not shown) while the absolute marker effects ranged from 0.30 to 0.39 for the OPN data set and 0.26 to 0.47 for the Ames data set (Table 3).

Pairwise LD (r^2) values greater than 0.50 between significant markers were observed in the following marker pairs for the Ames data set: oPt.17611 and oPt.12985 with 0.89, oPt.11737 and oPt.14067 with 0.76, oPt.11737 and oPt.3063 with 0.73, oPt.14067 and oPt.3063 with 0.86, oPt.9329 and oPt.2635 with 0.86, and oPt.12704 and oPt.16436 with

0.65 (Supplementary Figure 3). There were no pairwise LD values greater than 0.50 between significant markers for the OPN data set.

Mixed Model LASSO Association

For model selection in mixed model LASSO, the lowest AIC corresponded to a model with the first 24 markers in the OPN data set and the first 37 markers in the Ames data set (Supplementary Figure 2 and 4, Table 4). There were 13 markers in common between the OPN and Ames data sets using the mixed model LASSO. The absolute effect of markers that were included in the model ranged from 0.003 to 0.18 for OPN and 0.004 to 0.17 for Ames. The marker with the largest and most consistent effect was oPt.18130 (0.18 and 0.17, respectively, for OPN and Ames). Furthermore, LD relationships among the markers identified using mixed model LASSO indicated that only one pair of markers, oPt.3063 and oPt.14067, were in high LD (r^2 = 0.86), which occurred only for the Ames data set (Supplementary Figure 5).

Markers Across Models and Datasets

All of the significant markers identified in the PK OPN were also significant in the mixed LASSO OPN. For Ames, 11 out of the 15 significant markers identified in the PK association were also significant in the mixed model LASSO (Tables 3 and 4, Figure 2). Only three markers were consistently detected across all datasets and models (oPt.14067, oPt.12985 and oPt.18130). It was also observed that there were 10, 16 and three markers that were unique to the mixed model LASSO OPN, mixed model LASSO Ames, and PK Ames, respectively. Altogether, the two models and datasets resulted in 51 unique markers that were further explored using various independent filters.

Comparison to β-Glucan QTL Mapping Studies and BLASTn Homology Search

To compare to previous QTL mapping studies, 24 out of the 51 markers significantly associated with β-glucan in this study were present in the updated Kanota x Ogle map (Tinker et al., 2009; Wight et al., 2003). None of the remaining 27 markers were in high LD (cut-off of r^2= 0.75) with any of the mapped markers (data not shown) so we did not seek to place them using LD. The 24 markers covered 15 linkage groups (Table 5). From the 24 markers, 15 were found on the same linkage groups as previously identified QTL (Table 5). The probability of 15 or higher from a binomial distribution with success probability of 0.416 and 24 trials is 0.03. We therefore rejected the null hypothesis that our associations were random relative to previous QTL identifications. Five of the 15 markers were located on linkage group 22_44_18. Three of those 15 markers mapped to within 1 cM of QTL found in Kianian et al. (2000).

A BLASTn homology search was conducted for all significant markers identified in this study. Thirteen out of 51 unique markers from all methods and datasets were found to have homology to a total of 34 rice genes. Six homologs were found in chromosome 1 of rice, four in chromosome 3, eight in chromosome 4, one each for chromosome 5, 6, 8and 12, two on chromosome 7 and lastly 10 hits were found in chromosome 11. The search showed that none of the markers reported in this study had a direct homology to any cellulose synthase gene families. To further filter the hits, the location of all cellulose synthase were searched in the database and compared to the location of DArT marker homologs (Figure 3). The comparison showed that closest distance was 63 kilo base pairs (kb) which was between the homolog (LOC_ Os03g59480) of oPt.12704 and cellulose synthase A catalytic subunit 2 (CesA2) (LOC_Os03g59340). This is followed by a homolog of oPt.8758

(LOC_Os03g58910) at 235 kb away from the same CesA2 gene. The distances of homologs of oPt.12704 and oPt.8758 to one of the candidate genes can be considered adjacent given that only 5% of random positions in the rice genome are within 247 kb or a candidate gene.

Discussion

β-glucan Content and Population Structure in the Elite Oat Association Panel

The inbred lines used in this study represent the wide spectrum of elite oat germplasm in North America. In particular, the association panel is composed of lines that were tested from 1994-2007 in Uniform Oat Performance Nurseries representing 15 breeding institutions in the US and Canada. Analysis of phenotypic data from historical (OPN) and balanced (Ames) datasets showed that these two data sets are highly correlated but show heterogeneity of variances, therefore data from OPN and Ames were analysed separately. The standard deviations (0.56, 0.69 for OPN and Ames, respectively) for β-glucan in this study were comparable to values found by Peterson et al. (2005) of 0.44 – 0. 59 for a group of elite cultivars. The level of heritability found in this study was similar to previous results (Holthaus et al., 1996).

The smaller broad sense heritability in the OPN relative to the Ames data can be attributed to the highly unbalanced data and / or to genotype by environment interaction. Variance components for this interaction were not estimable due to the OPN's unbalanced nature, so no direct comparion between the data sets could be made. It seems reasonable to assume, however, that Ames data came from more homogeneous environments. In this sense, data from the OPN represents *broad adaptation β-glucan content* which has lower genetic variance while Ames represents *narrow adaptation β-glucan content*.

Association studies have long been known to be sensitive to population structure (Kennedy et al., 1992). To explore population structure, we used both cluster analysis and PCA on the marker data. PC scatter plots classified with the clustering results indicated that the clusters were distinctly separated with little overlap. For example, PC1 separated the AC Assiniboia Cluster from the Baker and the Ogle clusters while PC3 separated the Ogle Cluster from the remaining clusters. It was also shown that the identified clusters differed for their β-glucan content means indicating the potential of structure to generate false-positive associations in the GWAS. The correlation of β-glucan content with a subset of the PCs also implies that the population structure effects explain some of the variation in β-glucan and can account for confounding effects.

Kinship among oat lines was also used in all GWAS models in our study to account for fine-grained relationship (Yu et al., 2006). Accounting for kinship among lines, measured as the covariance among observations, has been known to reduce false positives and eliminate bias in marker effects by accounting for the genetic background effects (Kennedy et al., 1992). One general cause of population structure confounding is that single-factors models are used to identify associations for traits that are multi-genic (Atwell et al. 2010). Explicitly multi-genic models therefore make sense and we evaluated their impact by contrasting a single marker test (Yu et al., 2006) and mixed-model LASSO (Wang et al., 2010).

Single Marker Tests

The 13 unique markers from OPN and Ames data sets had relatively low R^2 values ranging between 2 to 3%. In a previous QTL mapping study, six putative QTL were

identified in the Kanota x Ogle population (137 recombinant inbred lines) in which the five markers explained 2 to 5 % and one marker explained 12% variation in phenotypic data (Kianian et al., 2000). In general, the low R^2 value and many significant markers indicate that β-glucan is controlled by multiple loci with small additive effects (Holthaus et al., 1996). The high R^2 in biparental QTL mapping studies than GWAS panel may be explained by the higher LD and the reduced overall genetic variability in the former than the latter. On the other hand, the magnitude of individual marker effects in our study was comparable to Kianian et al. (2000). For example, the marker with the largest effect in this study (oPt.12985) can increase β-glucan content by 0.30 to 0.45%, similar to the 0.35% for a large effect marker identified by Kianian et al. (2000).

Pairwise linkage disequilibrium, measured as the r^2 between markers, indicated that some of the significant markers from the PK model for the Ames data set were in high LD. For example, the high LD between oPt.12985 and oPt.17611 can be explained by the fact they are located on the same region of linkage group 1_3_38_break in the Kanota x Ogle population (Table 6). The high LD relationships among oPt.12704, oPt.16436 and oPt.14317 likewise, are explained by their close proximity on linkage group 22_14_18 (Tinker et al., 2009).

Mixed Model LASSO

In this study, we used mixed-model LASSO as an alternative approach for QTL detection. In cases where correlation between predictors is present, the algorithm selects the best marker within a group of correlated markers and sets the effect of other predictors to zero (Ayers and Cordell, 2010). We decided to explore this method because theory indicates

that traits are controlled by multiple factors acting in concert. Instead of choosing a particular lambda for LASSO, we included an initial number of variables in the model and applied a goodness of fit test, the AIC, to decide the optimal number of markers in the model. Based on the AIC, the best model included the first 24 and 37 markers that entered into the model for the OPN and Ames data sets, respectively.

The distribution of absolute marker effects was comprised of many markers with zero effect, markers with near zero and few markers with large effects (Figure 3). The range of non-zero marker effects (0.003 – 0.18) suggested that the magnitude of effects can be used to determine markers that are likely associated with the trait. The low pairwise LD among the significant markers identified by LASSO confirms previous conclusions that LASSO results identify markers that are more independent (Sung et al., 2009).

Markers from Different Models and Datasets

The results in this study provide examples of advantages and disadvantages for single marker PK and mixed LASSO analyses. The single marker PK test is a popular method with many studies confirming its application in gene discovery and marker-assisted selection programs. However, this method will generate multiple hits for a single QTL when markers included are in high LD. The simplicity of application for the single marker test makes it a good initial method to explore associated markers. A major difference between the PK and mixed LASSO analyses is that the objective of the former is to perform hypothesis tests on every marker while that of the latter is to identify the best subset of markers in a model selection process. The two analyses are therefore complementary.

We also found that marker effects were heavily shrunk in mixed model LASSO compared to individual effects in the PK model. First, this is explained by the fact the LASSO method shrinks effects regardless of the dimension of the data (Tibshirani, 1996). Second, the fact that the LASSO model had more markers than the PK model may mean that in the PK model each marker maybe capturing more than one QTL, whereas markers in the LASSO model captured unique QTL (Ayers and Cordell, 2010). A positive outcome of such algorithm is that the markers with small effects can still be detected to be important to model β-glucan content. As a general recommendation, we propose to use both PK single marker test and mixed model LASSO to identify markers in genome-wide studies.

Comparison to QTL Studies

The most comprehensive genetic map in oat, Kanota x Ogle (Tinker et al., 2009), included only 24 out of the 51 markers identified in this study. These 24 markers were scattered across 15 of total 31 linkage groups of the Kanota x Ogle map, thus supporting the multigenic nature of β-glucan content in oat (Kianian et al., 2000; Orr and Molnar, 2008). The genetic location of 15 out of those 24 DArT markers corresponded to the same linkage groups of markers for β-glucan QTL identified by Kianian et al. (2000), De Koeyer et al. (2004) and Groh et al. (2001), a significantly greater number than expected by chance, indicating that our study detected some of the same signal as in bi-parental populations.

The genomic regions of four DArT markers in this study corresponded to the same regions of three QTL (cdo346A, cdo82, cdo1340) identified by Kianian et al. (2000). Two of the markers identified here (oPt.12985 and oPt.17611) co-localized within less than 1 cM of cdo346A – the marker with the largest effect QTL in the Kanota x Ogle population. This

implies that oPt.12985 and oPt.17611 might be detecting the same QTL given that these markers are also in high LD. Another associated marker, oPt.10823, mapped within 4 cM of a previously identified QTL (cdo82).

Five associated DArT markers (oPt.14317, oPt.12704, oPt.5064, oPt.16618, and oPt.16436) are close to a QTL from the Terra x Marion population De Koeyer et al. (2004). Three of these markers (oPt.14317, oPt.12704, and oPt.16436) had high LD with each other and mapped within 10 to 20 cM of cdo484A, the marker explaining the most variance in Terra x Marion (De Koeyer et al. 2004).

The rest of the markers in the study were more than 20 cM distant from previously detected QTL. At 20 cM, the expected LD in elite oat decays already to less than $r^2 = 0.05$, indicating that these markers will probably not be able to capture sufficient variance of β-glucan QTL to be identified (Newell et al., 2010). Therefore, those markers may be detecting separate QTL.

Rice BLAST Homologies

The *CslF* and *CslH* gene families have been previously shown to affect β-glucan synthesis in various species within the grass family (Burton et al., 2006; Doblin et al., 2009). Given the shared evolutionary history of species within the grasses, it is possible to identify the same gene families through comparative genomic methods. In this study, none of the significant markers were directly homologous to *CslF* or *CslH* gene families in rice. However, these gene families are not the only participants in β-glucan synthesis given that they interact with the whole carbohydrate synthesis network (Fincher, 2009). Therefore, we cannot rule out the possibility that the markers reported in this study could lead to QTL

controlling components of that metabolic network. For example, two of the significant markers (oPt.12704 and oPt.8758) in our study are adjacent to cellulose synthase A catalytic subunit 2 (*CesA*2), a gene which was identified to be co-expressed with *CslF*6 in transcriptional studies for barley (Burton and Fincher, 2009).

Implications for Marker-Assisted Selection

There is still a high discrepancy between QTL studies and application of these studies in MAS (Xu and Crouch, 2008). Attempts to breed for high β-glucan content using MAS has been initiated based on early QTL mapping studies. Orr and Molnar (2008) developed markers for β-glucan based on QTL identified in the Kanota x Ogle population (Kianian et al., 2000) and in the Terra x Marion population (De Koeyer et al., 2004). Because these populations were developed from parents chosen to be highly distinctive for their phenotype, it is possible that these QTL will be population specific and therefore less useful in the context of breeding programs (Bernardo, 2008). Since we used elite oat, the QTL that we found have higher probability of being valid across elite population. Finally, we note that our results confirm that β-glucan content is a polygenic trait (Holthaus et al., 1996; Chernyshova et al., 2007). For such traits, genomic selection, a method that predicts breeding values using all markers (Meuwissen et al., 2001) maybe employed in lieu of traditional MAS programs to increase β-glucan in oat (Asoro et al., 2011).

Our study can serve as an additional resource in understanding genetic mechanisms for this trait and enhancing the marker-assisted selection efforts toward the development of new oat cultivars with increased β-glucan content. The FDR cut-off of 0.33 and the LASSO model both use relaxed marker detection thresholds. However we implemented further

independent filters and we can prioritize the important QTL using these criteria. For the consistency of significance across methods and datasets, the important markers were: oPt.14067, oPt.12985 and oPt.18130. For close proximity to previous QTL, the important markers were: oPt.12985, oPt.17611, oPt.10823, oPt.4358, oPt.6974, oPt.14317 and oPt.12704. Finally, oPt.12704 and oPt.8758 were adjacent to β-glucan candidate genes. Since oPt.12985 and oPt.12704 were important for two criteria, they rise to the top of the list as candidates for further research.

Acknowledgments

This research was funded by the United States Department of Agriculture, National Institute of Food and Agriculture, grant 2008-55301-18746. We thank Adrienne Moran Lauter for laboratory work and George Patrick for field work.

References

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

Atwell S., Y.S. Huang, B.J. Vilhjalmsson, G.Willems, M.Horton, Y. Li, D.Meng, A. Platt, A.M. Tarone, T.T.Hu, R. Jiang, N.W. Muliyati, X. Zhang, M.A. Amer, I. Baxter, B. Brachi, J. Chory, C. Dean, M. Debieu, J. de Meaux, J.R. Ecker, N. Faure, J.M. Kniskern, J.D.G. Jones, T. Michael, A. Nemri, F. Roux, D.E. Salt, C. Tang, M. Todesco, M.B. Traw, D. Weigel, P. Marjoram, J.O. Borevitz, J. Bergelson, and M. Nordborg. 2010. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature 465:627-631.

Ayers, K.L., and H.J. Cordell. 2010. SNP selection in genome-wide and candidate gene studies via penalized logistic regression. Genetic Epidemiology 34: 879-891.

Benjamini, Y., and Y. Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. R. Statist. Soc. B 57:289–300.

Bernardo, R. 2008. Molecular markers and selection for complex traits in plants: Learning from the last 20 years. Crop Sci. 48:1649.

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

Burton, R.A., S.M. Wilson, M. Hrmova, A.J. Harvey, N.J. Shirley, A. Medhurst, B. A. Stone, E.J. Newbigin, A. Bacic, and G.B. Fincher. 2006. Cellulose synthase-like CslF genes mediate the synthesis of cell wall (1,3;1,4)-β-D-glucans. Science 311:1940–1942.

Burton, R.A., and G.B. Fincher. 2009. (1, 3; 1, 4)-β-D-Glucans in cell walls of the Poaceae, lower plants and fungi: a tale of two linkages. Molecular plant 2, no. 5: 873.

Butt M.S., M. Tahir-Nadeem, M.K.I. Khan, R. Shabir, and M.S. Butt. 2008. Oat: unique among the cereals. European Journal of Nutrition 47:68–79.

Carollo, V., D.E. Matthews, G.R. Lazo, T.K .Blake, D.D. Hummel, N.Lui, D.L. Hane, and O.D. Anderson. 2005. GrainGenes 2.0. An improved resource for the small-grains community. Plant Physiology 139:643-651.

Cervantes-Martinez, C.T., K.J. Frey, P.J. White, D.M. Wesenberg, and J.B. Holland. 2001. Selection for greater β-glucan content in oat grain.Crop Sci. 41:1085–1091. doi:10.2135/cropsci2001.4141085x.

Chernyshova A.A., P.J. White, M.P. Scott, and J-L Jannink. 2007. Selection for nutritional function and agronomic performance in oat. Crop Sci. 47:2330-2339.

Colleoni-Sirghie, M., J.-L. Jannink, and P.J. White. 2004. Pasting and thermal properties of flours from oat lines with high and typical amounts of β-glucan. Cereal Chem. 81:686-692.

De Koeyer, D.L., N.A. Tinker, C.P. Wight , J. Deyl , V.D. Burrows, L.S. O'Donoughue, A. Lybaert, S.J. Molnar , K.C. Armstrong, G. Fedak, D.M. Wesenberg, B.G. Rossnagel, and A.R. McElroy. 2004. A molecular linkage map with associated QTLsfrom a hulless covered spring oat population. Theor Appl Genet 108:1285–1298.

Doblin M.S., F. Pettolino, S.M. Wilson, R. Campbell, R.A. Burton, G.B. Fincher, E. Newbigin, and A. Bacic. 2009. A barley cellulose synthase-like *CSLH* gene mediates (1,3;1,4)-β-D-glucan synthesis in transgenic *Arabidopsis*. PNAS USA 106:5996-6001.

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

Falconer, D.S., and T.F.C. Mackay. 1996. Introduction to quantitative genetics. 4th ed. Longman Technical and Scientific, Essex, UK.

Fincher, G.B. 2009. Exploring the evolution of (1,3;1,4)-β-D-glucans in plant cell walls: comparative genomics can help! Current Opinion in Plant Biology 12, no. 2: 140-147.

Garrick, D.J., J.F. Taylor, and R.L. Fernando. 2009. Deregressing estimated breeding values and weighting information for genomic regression analyses. Genet. Sel. Evol. 41:55. doi:10.1186/1297-9686-41-55.

Groh S., A. Zacharias, S.F. Kianian, G.A. Penner, J. Chong, H.W. Rines, and R.L. Phillips. 2001. Comparative AFLP mapping in two hexaploid oat populations. Theoretical and Applied Genetics 102, no. 6-7: 876-884.

Holthaus, J.F., J.B. Holland, P.J. White, and K.J. Frey. 1996. Inheritance of β-glucan content of oat grain. Crop Sci. 36:567–572.

Kang H.M., N.A. Zaitlen, C.M. Wade, A. Kirby, D. Heckerman, M.J. Daly, and E. Eskin. 2008. Efficient control of population structure in model organism association mapping. Genetics 178:1709-1723.

Kennedy, B.W., M. Quinton, and J.A.M. Vanarendonk. 1992. Estimation of effects of single genes on quantitative traits. J. Anim. Sci. 70: 2000–2012.

Kianian S.F., R.L. Phillips, H.W. Rines, R.G. Fulcher, F.H. Webster, and D.D. Stuthman. 2000. Quantitative trait loci influencing β-glucan content in oat (Avena sativa, 2n=6x=42). Theor Appl Genet 101:1039-1048.

Lande R., and R. Thompson.1990. Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124:743-756.

Malosetti, M., C.G. van der Linden, B. Vosman, and F.A. van Eeuwijk. 2007. A mixed-model approach to association mapping using pedigree information with an illustration or resistanceto Phytophthora infestans in potato. Genetics 175:879–889.

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

Newell, M.A., D. Cook, N.A. Tinker, and J.-L. Jannink. 2010. Population structure and linkage disequilibrium in oat (Avena sativa L.): implications for genome-wide association studies. Theor. Appl. Genet. DOI: 10.1007/s00122-010-1474-7.

Orr, W., and S.J. Molnar. 2008. Development of PCR-based SCAR and CAPS markers linked to ß-glucan and protein content QTL regions in oat. Genome 51(6), pp. 421-425. doi: 10.1139/G08-026.

Ouyang, S., W. Zhu, J. Hamilton, H. Lin, M. Campbell, K. Childs, F. Thibaud-Nissen, R.L. Malek, Y. Lee, L. Zheng, J. Orvis, B. Haas, J. Wortman, and C.R. Buell. 2007. The TIGR Rice Genome Annotation Resource: improvements and new features. Nucleic Acids Research35:D883-D887.

Peterson, D.M. 1991. Genotype and environment effects on oat β- glucan concentration. Crop Sci. 31:1517–1520.

Peterson, D.M., D.M. Wesenberg, D.E. Burrup, and C. Erickson. 2005. Relationships among agronomic traits and grain composition in oat genotypes grown in different environments. Crop Science 45, no. 4: 1249. doi:10.2135/cropsci2004.0063.

R Development Core Team. 2011. R: A language and environment for statistical computing. Available at http://www.r-project.org (verified 18 Jan 2012). R Foundation for Statistical Comput., Vienna, Austria.

SAS Institute, 2010. SAS/STAT® 9.2 User's Guide. SAS Campus Drive, Cary, North Carolina 27513.

Stich, B., J. Mohring, H.-P. Piepho, M. Heckenberger, E.S. Buckler, and A.E. Melchinger. 2008. Comparison of mixed-model approaches for association mapping. Genetics 178:1745–1754.

Sung, Y. J., T.K. Rice, G. Shi, C.C. Gu, and D.C. Rao. 2009. Comparison between single-marker analysis using Merlin and multi-marker analysis using LASSO for Framingham simulated data. BMC Proceedings 3(Suppl 7):S27 doi: 10.1186/1753-6561-3-S7-S27.

Tibshirani R. 1996. Regression shrinkage via the lasso. J R Statis Soc 58:267–288.

Tinker, N.A., and J.K. Deyl. 2005. A curated internet database of oat pedigrees. Crop Science 45:2269-2272.

Tinker, N.A., A. Kilian, C.P. Wight, K. Heller- Uszynska, P. Wenzl, H.W.Rines, A. Bjornstad, C.J. Howarth, J.L. Jannink, J.M. Anderson, B.G.Rossnagel, D.D. Stuthman, M.E. Sorrells, E.W. Jackson, S. Tuvesson,F.L. Kolb, O. Olsson, L.C. Federizzi, M.L. Carson, H.W. Ohm, S.J.Molnar, G.J. Scoles, P.E. Eckstein, J.M. Bonman, A. Ceplitis, and T.Langdon. 2009. New DArT markers for oat provide enhanced mapcoverage and global germplasm characterization. BMC Genomics10:39. doi:10.1186/1471-2164-10-39

Tiwari, U., and E. Cummins.2009.Factors influencing β-glucan levels and molecular weight in cereal-based products, Cereal Chemistry 86**:** 290–301.

U.S. Food and Drug Administration. 2010. Health claims: Soluble fiber from certain foods and risk of coronary heart disease. Available at http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch. cfm?fr=101.81 (verified 19 May 2011). U.S. FDA, Silver Spring, MD.

Wang, D., K.M. Eskridge, and J. Crossa. 2010. Identifying QTLs and epistasis in structured plant populations using adaptive mixed LASSO. Journal of Agricultural Biological and Environmental Statistics 16: 1–15.

Wight, C.P., N.A. Tinker, S.F. Kianian, M.E. Sorrells, L.S. O'Donoughue,D. Hoffman, S. Groh, G.J. Scoles, C.D. Li, F.H. Webster, R.L. Philips,H.W. Rines, S.M. Livingston, K.C. Armstrong, G. Fedak, and S.J.Molnar. 2003. A molecular marker map in 'Kanota' × 'Ogle' hexaploidoat (*Avena* spp.) enhanced by additional markers and a robustframework. Genome 46:28–47. doi:10.1139/g02-099.

Wu, T.T., Y.F. Chen, T. Hastie, E. Sobel, and K. Lange. 2009. Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics (Oxford, England) 25, no. 6 (March 15): 714-21.

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

Yun, S.J., D.J. Martin, B.G. Gengenbach, H.W. Rines, and D.A. Somers. 1993. Sequence of a (1,3;1,4) β-glucanase cDNA from oat. Plant Physiology 103: 295-296.

Zhao, K., M.J. Aranzana, S. Kim, C. Lister, C. Shindo, C. Tang, C. Toomajian, H. Zheng, C. Dean, P. Marjoram, and M. Nordborg. 2007. An arabidopsis example of association mapping in structured samples. PLoS Genet 3:e4.

Zhang Z., E.S. Buckler, T.M. Casstevens, and P.J. Bradbury. 2009. Software engineering the mixed model for genome-wide association studies on large samples. Briefings in Bioinformatics 10:664-675.

Zhou, H., Sehl, M.E., Sinsheimer, J.S., and K. Lange .2010. Association screening of common and rare genetic variants by penalized regression. Bioinformatics (Oxford, England), 26(19), 2375-82. doi: 10.1093/bioinformatics/btq448.

List of Figures

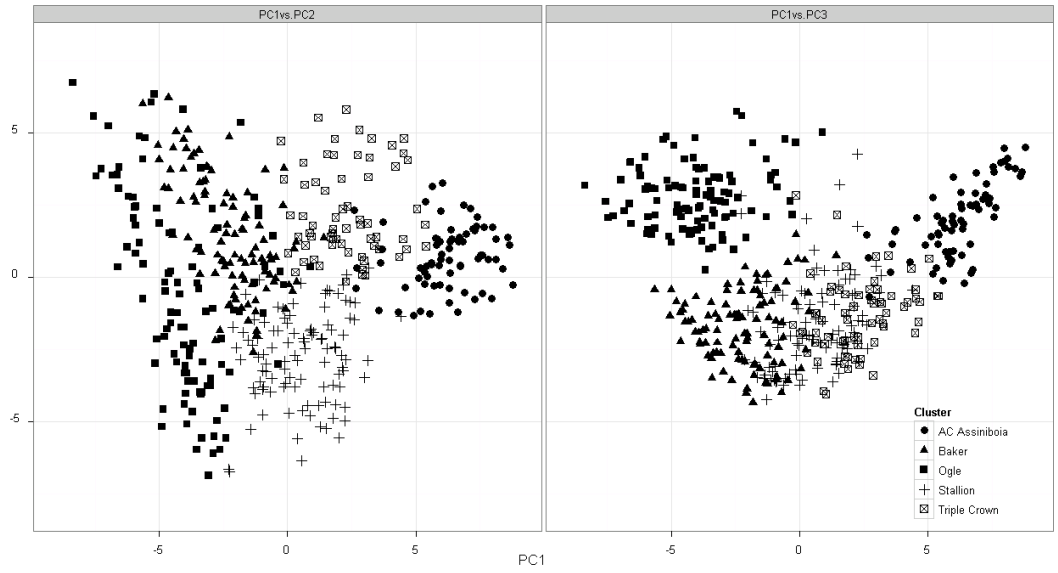


Figure 1. Scatter plot of principal components on the marker data where lines are assigned to various k-means clusters labelled according to a popular cultivar in each cluster, PC1 versus PC2 (left panel) and PC1 versus PC3(right panel).

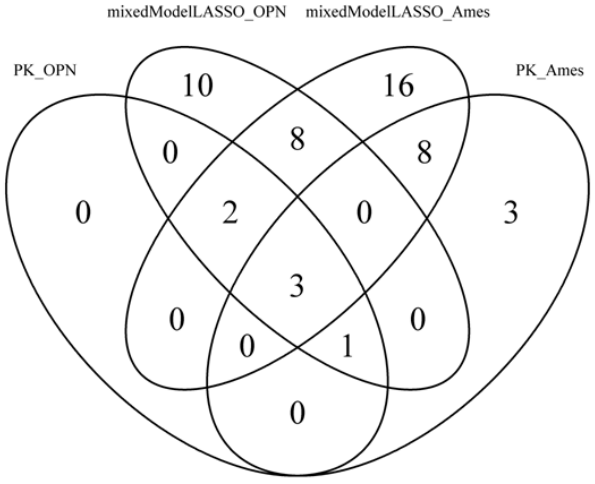


Figure 2. Venn diagram of markers identified in the PK and mixed model LASSO models for two datasets.

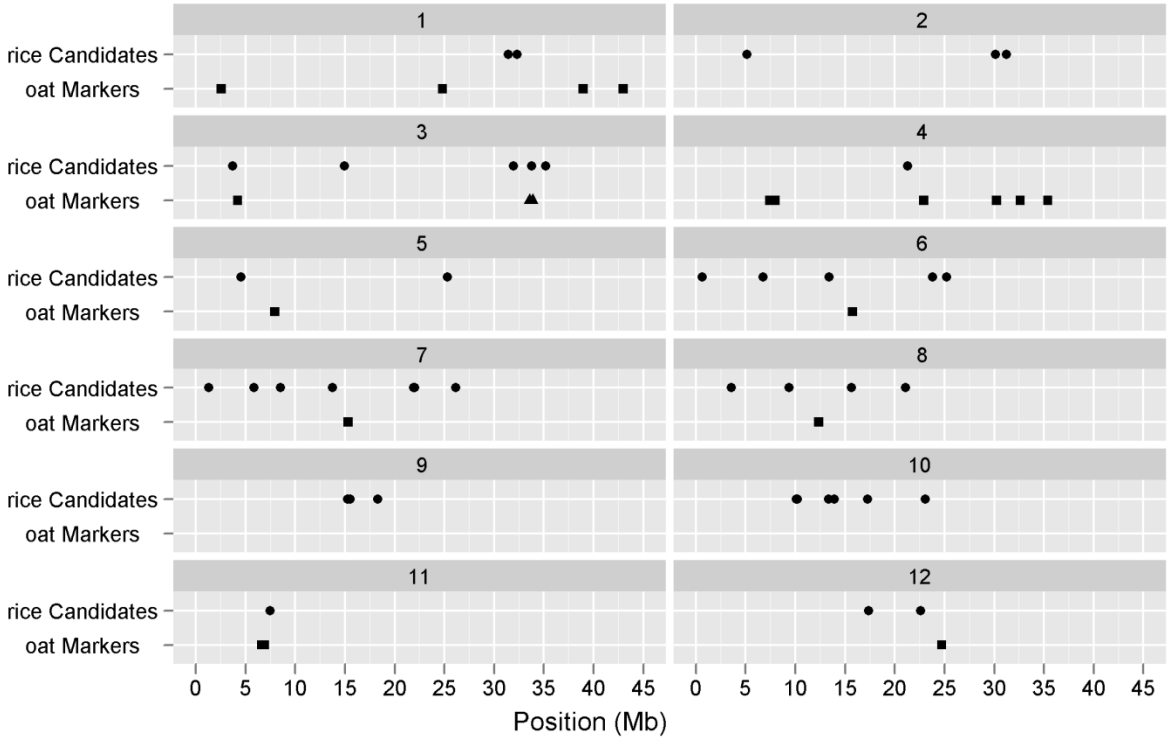
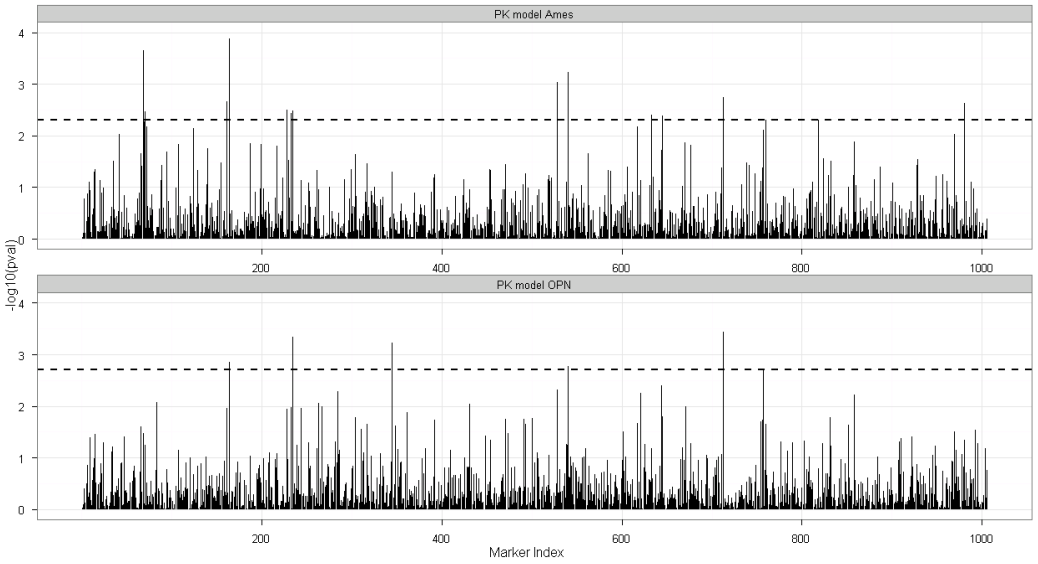
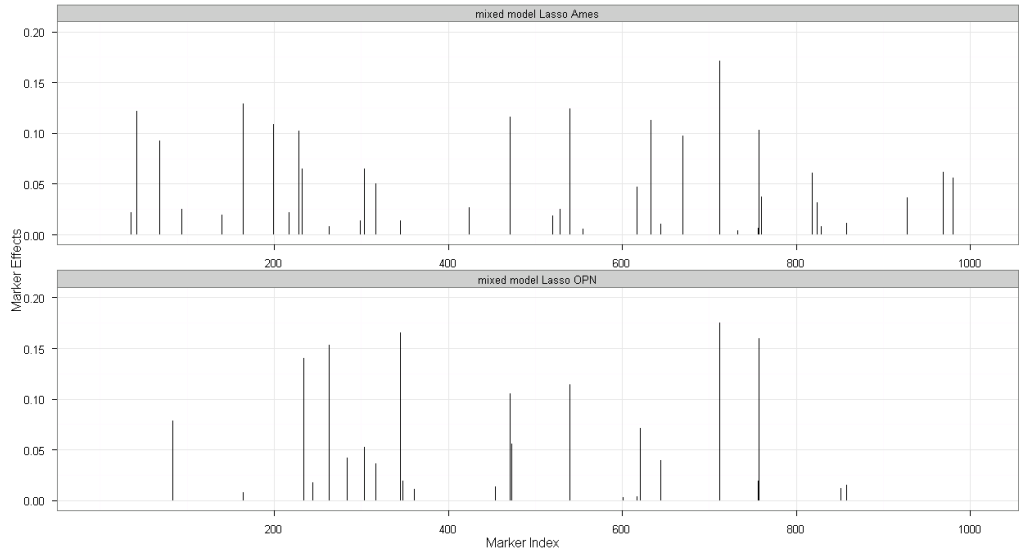


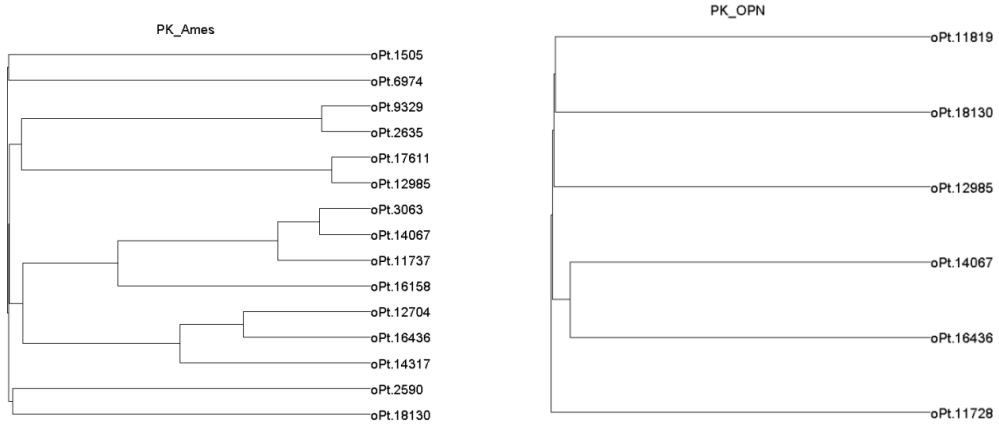
Figure 3. Comparison of locations of cellulose synthase gene families (round shape) and locations of DArT marker homologs (square and triangle) along the rice genome (x-axis). The triangle in chromosome 3 indicates that homologs are significantly close to rice candidate genes. The panels are named according to rice chromosome number. The x-axis is expressed as position in the rice genome in mega base pair.



Supplementary Figure 1. Manhattan plots of –log10 of p-values for PK model association tests for Ames (top panel) and OPN (bottom) data sets. The broken horizontal line is the threshold for FDR=0.33 in each data set.

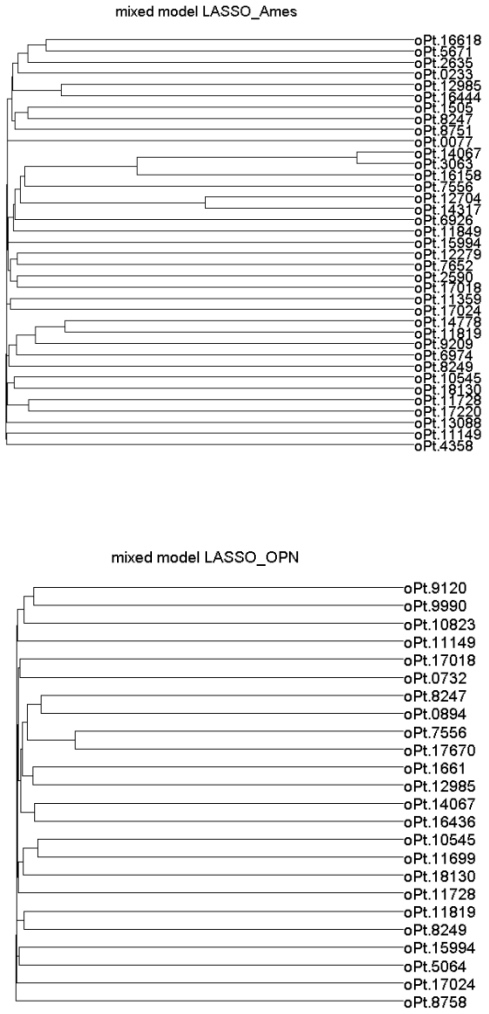


Supplementary Figure 2. Manhattan plots of marker effects from mixed model LASSO on Ames (top panel) and OPN (bottom) data sets.



Supplementary Figure 3. Cluster analysis using average linkage of pairwise LD values (r^2) of significant markers in PK model for Ames (left panel) and OPN (right panel) data sets. Scale is from 0 to 1 with rightmost as 1 indicating perfect LD.

Supplementary Figure 4. Akaike Information Criterion (AIC) values from mixed model LASSO on OPN (right panel) and Ames (left panel) data sets plotted against number of markers in the model.



Supplementary Figure 5. Cluster analysis using average linkage of pairwise LD values (r^2) important markers in mixed model LASSO models for Ames (top panel) and OPN (bottom panel) data sets. Scale is from 0 to 1 with rightmost as 1 indicating perfect LD.

List of Tables

Table 1. Summary statistics of β-glucan content (%) for two data sets.

Descriptive Data	OPN†	Ames
Mean	5.06	4.17
Minimum	3.15	2.32
Maximum	7.62	7.76
SE of Mean	0.03	0.03
Phenotypic Standard Dev	0.56	0.69
Oat line Variance ‡	0.19	0.45
Residual Variance ‡	0.25	0.26
H^2	0.43	0.63
Correlation of OPN and Ames	0.71	

† Oat performance nurseries.

‡Computed from original observed data where oat lines and residuals are the only random effects and both are assumed independently and identically distributed. The variances were significantly different from zero (p=<0.0001) based on Wald Z-test.

Table 2. β-glucan summary by cluster. Each clusters was named according to an oat variety in it.

Cluster	Number of Lines	Mean % BG (Std Dev)	
		OPN †	Ames
Baker	101	4.29 (0.56) ‡	5.09 (0.46) ‡
Ogle	105	4.46 (0.77)	5.39 (0.66)
AC Assiniboia	70	3.79 (0.45)	4.78 (0.32)
Stallion	104	4.09 (0.73)	5.00 (0.59)
Triple Crown	66	4.07 (0.65)	4.88 (0.39)

† Oat performance nurseries.

‡ Means of clusters are significantly different from each other based on ANOVA (p<0.0001).

Table 3. Significant markers from single marker test using the PK model for OPN and Ames data set at FDR of 0.33. Underlined markers are common between the two datasets.

OPN[†]			Ames		
	p-values	Marker Effects		p-values	Marker Effects
oPt.18130	0.0004	-0.38	oPt.12985	0.0001	0.45
oPt.16436	0.0005	0.39	oPt.2635	0.0002	0.45
oPt.11819	0.0006	-0.39	oPt.14067	0.0006	0.47
oPt.12985	0.0014	0.30	oPt.3063	0.0009	0.44
oPt.14067	0.0017	0.34	oPt.18130	0.0018	-0.41
oPt.11728	0.002	-0.34	oPt.17611	0.0022	0.38
			oPt.2590	0.0024	0.29
			oPt.14317	0.0031	-0.37
			oPt.16436	0.0033	0.41
			oPt.9329	0.0034	-0.34
			oPt.12704	0.0037	-0.39
			oPt.6974	0.0039	0.40
			oPt.11737	0.0042	-0.34
			oPt.1505	0.0049	0.26
			oPt.16158	0.0049	-0.36

[†] Oat performance nurseries

Table 4. Selected markers from mixed model LASSO based on Akaike Information Criterion, sorted based on their entry order in the model. Underlined markers are common between the two datasets.

	OPN†			AMES	
	Rank	Marker Effects		Rank	Marker Effects
oPt.11819	1	-0.165	oPt.18130	1	-0.172
oPt.18130	2	-0.175	oPt.14067	2	0.125
oPt.11728	3	-0.16	oPt.6926	3	0.122
oPt.8249	4	0.153	oPt.11728	4	-0.103
oPt.16436	5	0.14	oPt.17220	5	0.109
oPt.8758	6	0.071	oPt.11849	6	-0.097
oPt.14067	7	0.115	oPt.12985	7	0.129
oPt.11149	8	-0.106	oPt.14317	8	-0.102
oPt.0732	9	-0.078	oPt.6974	9	0.113
oPt.15994	10	0.053	oPt.11149	10	-0.116
oPt.17024	11	0.036	oPt.16158	11	-0.037
oPt.11699	12	0.056	oPt.2590	12	0.056
oPt.1661	13	-0.042	oPt.17024	13	0.051
oPt.8247	14	0.039	oPt.1505	14	0.061
oPt.10545	15	0.019	oPt.3063	15	0.025
oPt.9990	16	-0.02	oPt.15994	16	0.065
oPt.17018	17	0.015	oPt.7556	17	-0.047
oPt.5064	18	-0.018	oPt.2635	18	0.093
oPt.9120	19	-0.012	oPt.12704	19	-0.065
oPt.10823	20	0.011	oPt.4358	20	-0.025
oPt.0894	21	0.013	oPt.8751	21	-0.062
oPt.12985	22	0.008	oPt.11359	22	-0.037
oPt.7556	23	-0.004	oPt.14778	23	-0.032
oPt.17670	24	0.003	oPt.16618	24	0.022
			oPt.5671	25	0.019
			oPt.16444	26	0.022
			oPt.13088	27	0.019
			oPt.0077	28	-0.027
			oPt.0233	29	0.014
			oPt.11819	30	-0.014
			oPt.17018	31	0.012
			oPt.8247	32	0.011
			oPt.10545	33	0.006
			oPt.8249	34	0.008
			oPt.12279	35	-0.008
			oPt.7652	36	0.006
			oPt.9209	37	-0.004

† Oat performance nurseries

Table 5. Concordant genomic regions between β-glucan important mapped markers in elite oat association study and markers from published biparental QTL mapping studies. The genetic position is based on framework markers in the updated Kanota x Ogle map (KxO, Tinker et al., 2009).

Marker	KxO Linkage Group	Position (cM)	Distance from Previous β-glucan QTL	References
oPt.12985	1_3_38_break	0.5	0.4 cM from cdo346A	Kianian et al., 2001
oPt.17611	1_3_38_break	1	0.1 cM from cdo346A	Kianian et al., 2001
oPt.5671	1_3_38_X3	25		
oPt.17024	4_12_13	54	21.7 cM from cdo549B	Kianian et al., 2001
oPt.11819	5_30	107.5		
oPt.9990	6	15.6	70.5 cM from cdo82	Kianian et al., 2001
oPt.10823	6	90	3.9 cM from cdo82	Kianian et al., 2001
oPt.6974	7_10_28	71.5	5.3 cM from acacac236	Groh et al., 2001
oPt.6926	15	27		
oPt.16444	15	3		
oPt.2635	16_23	42.5		
oPt.9329	16_23	42.5		
oPt.0732	17	23	15.5 cM from cdo1340	Kianian et al., 2001
oPt.4358	17	38.5	0 cM from cdo1340	Kianian et al., 2001
oPt.17220	21_46_31_40	61		
oPt.14317	22_44_18	105.6	11.6 cM from cdo484A	De Koeyer et al., 2004
oPt.12704	22_44_18	106.5	12.5 cM from cdo484A	De Koeyer et al., 2004
oPt.5064	22_44_18	148.5	54.5 cM from cdo484A	De Koeyer et al., 2004
oPt.16618	22_44_18	73.5	20.5 cM from cdo484A	De Koeyer et al., 2004
oPt.16436	22_44_18	114	20 cM from cdo484A	De Koeyer et al., 2004
oPt.8249	24_26_34	53.4	31.4 cM from β-glucanase	Yun et al., 1993
oPt.1661	32	30	25 cM from cdo395A	De Koeyer et al., 2004
oPt.0233	36	20		
oPt.15994	37	11.4		

Supplementary Table 1. Significant markers with perfect LD (r^2=1) with other markers. These other markers were removed in the GWAS analysis but their sequences were used in BLAST analyses.

Significant Markers	Markers in perfect LD with significant markers
oPt.2635	oPt.13092
oPt.17611	oPt.1803, oPt.1881
oPt.12704	oPt.13230
oPt.5064	oPt.4939, oPt.9929
oPt.1661	oPt.17430, oPt.8886
oPt.9990	oPt.13262, oPt.16457, oPt.5874, oPt.9593, oPt.11536
oPt.0894	oPt.8270
oPt.11699	oPt.9224, oPt.14744, oPt.16537
oPt.8247	oPt.7806, oPt.10107, oPt.14536
oPt.14778	oPt.6784
oPt.11359	oPt.11866

CHAPTER 3. ACCURACY AND TRAINING POPULATION DESIGN FOR GENOMIC SELECTION ON QUANTITATIVE TRAITS IN ELITE NORTH AMERICAN OATS

A paper published in The Plant Genome[2]

Franco G. Asoro[2,5], Mark A. Newell[2], William D. Beavis[2], M. Paul Scott[2,3] and Jean-Luc Jannink[4]*

Abstract

Genomic selection (GS) is a method to estimate the breeding values of individuals by using markers distributed throughout the genome. We evaluated the accuracies of GS using data from five traits on 446 oat lines genotyped with 1005 Diversity Array Technology (DArT) markers and two GS methods (RR-BLUP and BayesCπ) under various training designs. Our objectives were to: 1) determine accuracy under increasing marker density and training population size; 2) assess accuracies when data is divided over time; and 3) examine accuracy in the presence of population structure. Accuracy increased as the number of markers and training size become larger. Including older lines in the training population

---

[2] Reprinted with permission from The Plant Genome 4:132-144.
[2] Department of Agronomy, Iowa State University, Ames, IA, 50011.
[3] USDA-ARS, Corn Insects and Crop Genetics Research Unit, Ames, IA, USA 50011
[4] USDA-ARS, R.W. Holley Center for Agriculture and Health, Department of Plant Breeding and Genetics, Cornell University, Ithaca, New York, 14853, USA.
[5] Primary researcher and author.
*Corresponding author(jeanluc.jannink@ars.usda.gov).

increased or maintained accuracy, indicating that older generations retained information useful for predicting validation populations. The presence of population structure affected accuracy: when training and validation subpopulations were closely related accuracy was greater than when they were distantly related, implying that LD relationships changed across subpopulations. Mixing less related subpopulations into a larger training set nevertheless improved accuracy. Across many different scenarios involving large training populations, the predictive ability of BayesCπ and RR-BLUP did not differ despite the conflicting assumptions of the two methods. This empirical study provided evidence regarding the application of GS to hasten the delivery of cultivars through the use of inexpensive and abundant molecular markers available to the public sector.

Introduction

The decreasing cost of high-density molecular markers allows saturation of crop genomes with genetic markers and offers an approach to predict genetic merit. These markers can help capture the effects of many quantitative trait loci (QTL) controlling polygenic traits regardless of location of the QTL in the genome by using linkage disequilibrium (LD), the non-random association of alleles at different loci (Falconer and Mackay, 1996). Meuwissen et al. (2001) proposed genomic selection (GS) based on prediction of the genetic value of individuals or the genomic estimated breeding values (GEBV) from high-density markers positioned throughout the genome. Because genomic selection includes all markers, major and polygenic effects can be captured, potentially explaining more genetic variance (Solberg et al., 2008). Therefore, the objective of GS is to predict the breeding value of each

individual instead of identifying QTL for use in a traditional marker assisted selection (MAS) program.

Selection methods can be evaluated by measuring accuracy, a major component of the response to selection equation, $R = ir\sigma_A$, where R is the response, i the selection intensity, r the accuracy and σ_A as the additive genetic standard deviation (Falconer and Mackay, 1996). As a general term in statistics, accuracy is the degree of similarity between the true value and the estimated value (Taylor, 1999). In crop selection programs, accuracy is defined as the correlation between the phenotype of the selected lines, i.e., selection units, and the phenotype transmitted to the progeny of the selected lines, i.e., response units (Holland et al. 2003). If the response population is composed of progeny of selected individuals, then accuracy is the correlation between the selection criterion and the true breeding value (TBV; Falconer and Mackay, 1996), since breeding values are by definition the mean of the progeny of individuals. If the selection criterion is the individual's phenotypic performance, r is equal to the square root of the heritability (Falconer and Mackay, 1996). In empirical cross-validation studies of GS, the TBV is unknown and to compute accuracy the TBV must be replaced by the traditional pedigree-based BLUP (best linear unbiased prediction) values, the least squares means from phenotypic evaluation, or some other appropriate phenotypic measurement (Garrick et al., 2009). The relationship between TBV and GEBV in the context of response to selection is explained in detail by Dekkers (2007).

Genomic selection in plant breeding has been studied in different types of populations. For example, GS has been used in narrow-based bi-parental populations (Lorenzana and Bernardo, 2009) and in broad-based populations like multi-lines of barley,

wheat, and maize (Zhong et al. 2009; Heffner et al., 2011; de los Campos., 2009). Regardless of the type of population used, the basic steps for implementation of genomic selection can be summarized in four steps: 1) designing training populations with complete phenotypic and genotypic data; 2) estimating marker effects in the training population; 3) calculating GEBV of new breeding lines with genotype data; and 4) selection (Heffner et al., 2009). Different methods exist to implement GS given the complexity of estimating marker effects to predict GEBV. These methods include ridge regression BLUP (RR-BLUP), and Bayesian-based methods such as BayesA, BayesB, BayesCπ and BayesLASSO (Meuwissen et al., 2001; Zhong et al., 2009; Kizilkaya et al., 2010; de los Campos, 2009). One important difference between RR-BLUP and the Bayesian methods is the prior distribution for the variance of marker effects: the former assigns equal variance to all markers while the latter allows unequal variances for markers. In numerous simulations and a few empirical studies of GS in both plants and animals, it has been shown that factors affecting accuracy include the genetic architecture of the trait, LD, genetic relationships between training and validation populations, marker density, training population size and heritability (Hayes et al., 2009; Zhong et al., 2009, Luan et al., 2009, Daetwyler et al., 2010, de Roos et al, 2009). In an empirical cross-validation study of bi-parental plant populations, Lorenzana and Bernardo (2009) demonstrated that accuracy increases with training population size. It was also shown that increasing the number of markers generally resulted in increased accuracy, but the increase was large only at low marker densities. For instance, in their study of grain protein content in the Steptoe x Morex-doubled haploid barley population, there was a clear increase in accuracy when changing from 64 to 128 markers, however accuracy did not change from 128 to 223 markers.

Population structure or differing levels of relatedness of individuals in a population can have an impact on genome-wide studies. It has been demonstrated that accounting for population structure avoids spurious associations in genome-wide association studies (GWAS) (Yu et al., 2006). In GS, while population structure is still relevant, the focus shifts to maintaining the accuracy across different subpopulations or germplasm groupings (Lorenz et al., 2010). In the simulation study of Toosi et al. (2010), accuracy was high when the training population and validation population belonged to the same breed of animals, but they also showed that there was no substantial decrease of accuracy when a multi-breed training population was used to estimate marker effects. In the empirical study of Hayes et al. (2009), GEBV predictions were more accurate within breed (eg. Jersey to Jersey) than across breeds (eg. Jersey to Holstein). However, when they used a multibreed training population (Jersey and Holstein) to predict purebred individuals (Jersey or Holstein), they found comparable accuracies as for the within breed predictions. Developing a multi-subpopulation training population is another way to increase training size and this approach may be important if subpopulations are small (de Roos et al., 2009). Although these studies suggest the importance of genetic relationships of the training and validation population, more importantly, they indicate that in the presence of population structure, LD should be consistent across subpopulations to maintain accuracy. This means that allelic effects estimated in one population should be predictive in another population (Lorenz et al., 2010). Such consistency of LD, however, requires higher marker densities (Meuwissen, 2009, Hamblin et al., 2010, Newell et al. 2010), and it is not clear if such densities are available for oat.

Currently there are few empirical studies of GS in crops. Thus, while simulations have shown that these methods have great potential, we do not know how well they will work in practice. Studies in several species and populations will be necessary to gain a general appreciation for investments in the marker density and training population size. As a case study, we evaluated the accuracies of GS for five traits in oats (grain β-glucan content, yield, heading date, groat percentage and plant height) from a public cooperative testing network in North America. The lines tested in the trials represent the breadth of alleles present in elite oat breeding populations, thus, they are a good sample for cross validation with potential impact in applied breeding programs. In this population, we assess the impact of marker density and training population size. This population is also structured so that we can present the first results in crops on the impact of structure on GS accuracy. Finally, RR-BLUP and BayesCπ have only been compared in simulation studies (Jannink, 2010) and here we provide a comparison using empirical data.

Materials and Methods

Phenotypic Data Analysis

The majority of phenotypic data for β-glucan percentage, yield, heading date, groat percentage, and plant height of oat breeding lines and cultivars included in this study came from the Uniform Oat Performance Nursery (UOPN) and the Quaker Uniform Oat Nursery (QUON) from 1994 to 2007 (http://wheat.pw.usda.gov/GG2/uopnquery.shtml). The UOPN is a cooperative testing network for oats among different US State Agricultural Experiment Stations and the USDA-ARS. The QUON is a cooperative testing network for oats among northern US State Agricultural Experimental Stations, USDA-ARS and public breeding

institutions in Canada. Data for β-glucan percentage was also included from research conducted by Chernyshova et al. (2007) and Colleoni et al. (2003). In total, there were 446 oat lines with β-glucan data and 421 lines with data for the four remaining traits. Data came from 129 environments (combination of years and locations) for β-glucan, 328 for days to heading, 278 for groat percentage, 354 for plant height and 388 for yield. Since not all of the lines were tested in the same environments, statistical analysis of this highly unbalanced data was conducted using PROC Mixed in SAS (SAS Institute, 2008), with environments considered fixed effects and oat lines as independently and identically distributed random effects. In this case, environments were considered as fixed effects to remove the effects of the mean of sets of environments on performance due to the fact that some lines were tested in few locations or some years only. As such, oat lines were treated as random effects as they are considered a sample of all possible oat genotypes. The best linear unbiased prediction (BLUP) for each line was used as its observed phenotypic value and denoted y*.

Marker Data, Relationship Matrix and Population Structure

Lines were planted in the Iowa State University Agronomy greenhouse in Spring 2008, leaf samples were collected for each entry and DNA was extracted according to the recommended protocol for DArT markers (www.diversityarrays.com). DNA samples were then sent to Diversity Arrays Technology (Yarralumla, Australia) for genotyping. DArT markers are a dominant marker system, thus for each of the 1295 markers, oat lines were scored for presence (1) or absence (0) of hybridization signal using a microarray platform (Tinker et al., 2009).

In order to eliminate redundant markers, sets of markers in perfect linkage disequilibrium (i.e., the squared correlation between marker scores was equal to 1) were identified. The marker with the lowest number of missing data points in each set was used in this study, resulting in 1005 markers.

To compute the marker-based relationship matrix, genotypic data points scored as absent (0) were recoded as -1, resulting in a data matrix of -1's and 1's. For each marker, missing values were replaced by the mean for that marker. The recoded marker matrix, **M,** was then used to compute the **MM'** matrix which was divided by 1005, scaling the relationship values from 0 to 1 in which the minimum value was 0.01 and the maximum value was 1.00. To account for population structure, principal component analysis (PCA) was applied to the relationship matrix. The first five PCs, which explained about 76% of variation in the marker data, were chosen based on the scree plot (Cattell, 1966). The corresponding five eigenvectors were used as fixed population structure covariates. Principal components have been used as another way to correct for population structure in GWAS and LD studies (Price et al., 2006; Stitch et al., 2008; Newell et al., 2010).

Methods of Genomic Selection and Prediction of GEBV

The general model used was : $y^* = \mu + \mathbf{Q}\nu + \mathbf{M}\alpha + e$ where y^* is the observed phenotypic value, μ is the intercept, $\mathbf{Q}\nu$ is a fixed effects term where **Q** is a matrix of the first five PC eigenvectors and ν is a vector of regression coefficients relating the first 5 PCs to the observed phenotype. The $\mathbf{Q}\nu$ term was excluded in the cluster-based training design (see below) because the clustering itself accounted for population structure. The $\mathbf{M}\alpha$ is a random effects term where **M** is the marker matrix and α is a vector of estimated marker effects.

Marker effects for RR-BLUP were simultaneously estimated and drawn from a normal distribution with equal variance, $N\left(0,\sigma_a^2\right)$ (Meuwissen et al, 2001). This method was implemented in the computer software R (R Development Core Team, 2009) using the emma package (Kang et al., 2008) and matrix algebra functions, in which the emma.MLE function was used to the estimate variance components $\sigma^2_{genetic}$ and σ^2_{error} and the shrinkage parameter $\sigma^2_{error}/\sigma^2_{genetic}$. The variance components and shrinkage parameter above were estimated in every sample of the training population. Finally, the shrinkage parameter computed above was incorporated in the mixed model equations to predict the marker effects.

For the BayesCπ method, described by RL Fernando (personal communication, June 2010), markers are represented as random effects (α) and are normally distributed when included in the model but equal to 0 when not included in the model with prior probability π. In contrast to BayesB (Meuwissen, 2001), the π parameter is estimated from the data. Further, the marker variance for BayesCπ, σ^2_a, is assumed *a priori* to be distributed as a scaled inverse chi-square as explained in detail in Kizilkaya et al. (2010). A total of 1000 burn-in and 4000 saved iterations of MCMC were used for BayesCπ in all designs. This method was implemented in R using code written by RL Fernando (personal communication, June 2010).

Marker effects estimated from RR-BLUP and BayesCπ were used to predict the estimated genotypic values for the validation population. The GEBV prediction model was: $GEBV = M\hat{\alpha}$, where **M** is the marker matrix and $\hat{\alpha}$ is are the estimated marker effects.

Design of Training and Validation Populations

In order to implement cross-validation for accuracy of GEBV, the observed phenotypic values (y*) for all lines were divided into training and validation data sets using three different methods:

1. Random Lines and Markers. Training populations were selected at random with the restriction that descendants of any individual in the validation population were excluded (to the extent possible given pedigree records available). We implemented this restriction because training populations will rarely contain descendants of selection candidates in practice and because descendants contain information about the Mendelian sampling term entering the breeding value of an individual (Falconer and Mackay, 1996), whereas collateral relatives will not. Including descendants would therefore bias accuracies upwards. Sets of 100, 200, and 300 lines were used as training populations while the remaining lines were used as validation populations with all 1005 markers retained. To determine the effect of marker density on accuracy, randomly selected sets of 300, 600, and 900 markers were used with a training population of 300 lines selected as describe above.
2. Testing Year-Grouping of Lines. Lines were grouped based on their first year of entry in the uniform nurseries. Years grouped as 1994-2003, 1998-2003 and 2001-2003 gave similar-sized training populations as for the randomly-selected lines, resulting in 292, 220 and 106 for β-glucan and 282, 213 and 99 for all other traits, respectively. To remove the effect of unequal training population sizes across traits, a random sample of 90 lines from 2001-2003, 180 lines from 1998-2003, and 270 lines from

1994-2003 were chosen as the final training population for 100 replicates. These training populations confound changes in size with changes in age. They do, however, answer the practically important question of the utility of increasing the training population size by adding older (historical) lines to the training population. To avoid confounding of training population size on training population age, another two sets of training population from 1994-1998 and 1998-2000 with 90 randomly selected lines each were also developed for comparison to the training population from 2001-2003. For all of these designs, the validation population consisted of lines from the 2004-2007 year grouping, which included 154 lines for β-glucan and 139 lines for the remaining traits.

3. Cluster-based Grouping of Lines. For grouping the oat germplasm, the relationship matrix among the 446 lines was converted to a distance matrix by subtracting the values from one. Hierarchical clustering using Ward's linkage was applied to the distance matrix and implemented using the hclust function in the computer software R (R Development Core Team, 2009). Three clusters were chosen for two reasons, 1) to maximize the number of individuals in each cluster, and 2) the clustering produced two more related clusters and one less related cluster (Supplemental Figure 1). The cluster dendrogram indicated that Cluster 2 (C2) and Cluster 3 (C3) are more highly related to each other than either is to Cluster 1 (C1). The clusters C1, C2, and C3 consisted of 130, 179, and 137 lines, respectively, for β-glucan, and 128,172 and 121 lines, respectively, for the other traits. A random sample of 120 lines from each cluster was used as the training population, while the other two clusters were used as validation populations. Additionally, to examine the effect of using combined clusters

and training population size in accuracy, random samples of 60 and 120 lines were taken from each of two clusters and combined to serve as 120 and 240 line training populations while the remaining cluster was used for the validation population.

For each of these designs, results were based on the average from 100 random replicates of the training populations.

Accuracy

Accuracy, calculated as the correlation of the observed (y*) and predicted breeding values (GEBV) in the validation sets was computed for each training design. Since population structure effects were in the model in the first two training designs, the accuracy was calculated to account for population structure effects in the y* vector by using the correlation (y* – **Q**ν, GEBV). This adjusted correlation will reflect the accuracy of GEBV excluding the variation due to population structure. The GEBV, with this adjustment, predicted within subpopulation or within cluster variation rather than all variation which combined within and between subpopulation variations.

Comparison of Accuracies

To compare how accuracy was affected by different GS methods, traits and training population designs, analysis of variance (ANOVA) was conducted for each training population-validation population design with the following model:

$$r = \mu + trait + method + design + trait*method + trait*design + method*design + error$$

where μ is the mean accuracy, the levels of *trait* are the five traits in this study; the levels of *method* are either BayesCπ or RR-BLUP, the levels of *design* depend on the design factor being analyzed (i.e., training population size, number of markers, year grouping, or cluster-based grouping), *trait*method*, *trait*design* , *method*design* are the main effect interaction terms and the *trait*method*design* interaction was considered the error term. We recognize that the ANOVA assumption of independence of errors is violated and thus p-values are not exact under the null hypothesis. The purpose of this ANOVA is not to test specific null hypotheses but simply to help quantify the relative magnitudes of the factors affecting accuracy.

Results

Randomly-selected Training Populations

In all cases, the factor with the strongest effect on accuracy was the trait being predicted (Table 1). Furthermore, this factor interacted in every case with aspects of training population design. In contrast to trait, the two methods we assessed had an impact only on the accuracies of training size but it never interacted with trait or training population design (Table 1).

In general, increasing the number of markers had a positive effect on prediction accuracy (Figure 1). Maximum accuracy was obtained at the highest density except for groat percentage. The highest increase in accuracy from 300 to 600 and from 600 to 900 markers were both obtained in yield using BayesCπ method with 0.05 and 0.03 increments, respectively. ANOVA suggested that not all traits responded equally to an increase in marker density, leading to an interaction between traits and marker density. In particular, groat

percentage reached a plateau in accuracy at 600 markers, while for the other traits accuracy continued to increase to the maximum of 900 markers (Figure 1).

For the standard deviations of accuracies computed from 100 random samples of the training population (data not shown), the values ranged across traits and marker densities between 0.06 to 0.08 for both RR-BLUP and BayesCπ.

Increasing the size of the training population also improved prediction accuracy (Figure 2). There were differences among the accuracies between traits (Table 1), with β-glucan as the trait with the highest accuracy and yield as the lowest. The accuracies across the three training sizes and traits ranged from 0.23 to 0.49 for BayesCπ and 0.16 to 0.49 for RR-BLUP. There was a steeper increase in accuracy when training population size increased from 100 to 200 than from 200 to 300 lines for all traits except yield (Figure 2). For instance, β-glucan gained 0.11 (BayesCπ) and 0.09 (RR-BLUP) from 100 to 200 lines, while there was only a 0.05 (BayesCπ) and 0.04 (RR-BLUP) increase in accuracy from 200 to 300 lines. For yield, the increase in accuracy was 0.01 (BayesCπ) and 0.05 (RR-BLUP) between 100 to 200 lines while it was 0.03 (BayesCπ) and 0.05 (RR-BLUP) when the training population was increased from 200 to 300 lines.

The standard deviations produced by BayesCπ were higher (0.08 – 0.10) across traits than RR-BLUP (0.04-0.06) when the training population size was 100, but were both within 0.04-0.08- across methods when the training population included 200 or 300 lines (data not shown).

Training Populations Constructed from Previous Generations

In practice, training sets will be comprised of previously developed breeding lines. To mimic this approach, the lines were divided based on their first year of entry in the uniform trials and grouped to obtain training population sizes of 90, 180, and 270 lines. Comparison of these training populations will indicate whether it is valuable to include older generations in order to increase the training population size. The ANOVA for this design (Test-Year, Table 1) indicated that there were differences among the accuracies from different training population sizes grouped according to year. Furthermore, there was also a trait by design interaction, caused primarily by the fact that some traits responded more to increased training population size than did others. The largest gain in accuracy was obtained for β-glucan, in which there was a gain of 0.17 (BayesCπ) and 0.19 (RR-BLUP) when the 1998-2003 training population was used instead of the 2001-2003 training population (Figure 3). The lowest gain in accuracy was observed for groat percent, in which there was minimal change in accuracy even when the 1994-2003 year grouping was used as the training population. We also found that using 1998-2003 as the training population produced a lower accuracy compared to when 2001-2003 was used as the training population for yield. This decrease in accuracy, however, was the only unequivocal decrease resulting from the addition of older phenotypic data to the training population. In other cases, accuracy was constant or increased.

To avoid confounding the effects of training population size and age of training population on prediction accuracy, 90 lines from 1994-1998, 1998 to 2000 and 2001-2003 were used as the training population. Results showed that most of the statistically not significant accuracies ($p>0.05$) came from 1994-1998 training population. In addition, for

this comparison there was also a large design by trait interaction (Table 1). The interaction came from two traits, days to heading and groat percent, for which older training populations led to lower accuracies than recent training populations while for the three other traits, older and recent training populations led to similar accuracies (Figure 4).

Training Populations Constructed from Different Subpopulations

To examine the effect of germplasm groupings on the accuracy of GEBV, clusters were used as the training population with a random set of 120 lines from each cluster while the remaining clusters were used as the validation population. Two clusters were also combined each time to form training population sets of 120 and 240 lines. Since C2 and C3 (C23) were more related to each other, they were treated as the "related training population" while C1 and C2 (C12) or C1 and C3 (C13) were treated as the "mixed training population". Accuracies for single cluster training populations and their combinations are presented in Figure 5 where each column of panels corresponds to the validation population. Most of the statistically not significant correlations ($p>0.05$) were observed when the validation population was C1, followed by C3 then by C2 (Figure 5). In this case the ANOVA showed differences between GS methods, and that the method interacted with trait (Table 2). This interaction arose because RR-BLUP was superior to BayesCπ for days to heading across all validation populations, and for plant height for the C2 and C3 validation populations, but the two methods performed similarly in all other cases.

Regarding the training population design, we were most interested to determine if related training populations outperformed unrelated training populations and how mixed training populations compared to single cluster training populations. Because there were trait

by design interactions (Table 1), these questions will need to be addressed trait by trait. C2 and C3 were more closely related to each other than either was to C1. We therefore expected better prediction when C2 or C3 served as the training population to predict the other than when C1 served to predict C2 or C3. Despite the trait by design interaction, this pattern is constant for every trait (Fig. 5, rightmost two columns: accuracy for "B120" is higher than accuracy for "A120"). In contrast, when C1 was the validation population, there was no reason that either C2 or C3 should generate more accurate predictions and there were generally only small differences between their accuracies across all traits (Fig. 5, leftmost column). We noted also that the highest accuracy in every trait for all 120-sized training populations involved C3 as either a single cluster or part of the mixed training sets (Fig.5, row-wise). Specifically, the C3 training population had the highest accuracy in β-glucan, groat percent and yield. In addition, the C23 and C13 training populations had the highest accuracy for days to heading and plant height, respectively.

With respect to the question of "mixed" training populations, the main issue is whether such a training population could generate more accurate predictions than that of the more accurate "pure" training population. The answer to this question varied by validation population and by trait, though overall it resulted in less accurate predictions. Nevertheless, this phenomenon occurred for days to heading for all validation populations and for plant height for the C2 and C3 validation populations (Fig. 5). If "mixed" means also a "bigger" training population (as would happen if the breeder already had data from from two subpopulations and combined them, as represented by the AB240 populations), then accuracies were generally higher than (or at least equal to) the most accurate pure training population. This improved accuracy occurred in every case except groat percentage for the

C1 and yield for the C2 and C3 validation populations. In general, there was higher gain of accuracy for the BayesCπ method than for RR-BLUP when the training population size was increased from 120 to 240 lines (Fig. 5).

Discussion

This study applied GS methods to empirical data gathered from long-term (1994-2007) multi-environment yield trials for oats in the United States and Canada. The impacts of marker density, training population size, and two GS methods on accuracy of GS were explored. Additionally, the effect of the age of the lines used in the training population and influence of population structure were investigated. Results of this study are encouraging regarding the use of GS in applied breeding programs even with the modest marker density of 1 marker for every 2 cM on average (1005 markers on a 1890 cM oat map, Wight et al. 2003). While accuracies that we found ranging from 0.27 to 0.50 for training populations of 300 individuals were fairly low, and might be insufficient for selection of lines as parents without any further phenotypic information, there are several reasons to believe that accuracies would be higher within breeding programs. First, oat lines in the UOPN are evaluated over a very broad range of environments, including environments outside of the target for which they were bred. Thus, for example “yield” as measured in this study might be better understood as “broad adaptation yield.” There will be less genetic variance for this broad adaptation yield than for the more narrow adaptation yield that most breeding programs target. Second, the phenotypic data came from highly unbalanced evaluations resulting in more error in the phenotypic observations. This error biases downward the estimated accuracy (Dekkers, 2007; Lorenz et al. 2011). Third, estimated accuracy would

have been higher if we had left the effects of structure in the prediction models. The reason for removing those effects is that we were more interested in performance relative to other lines in the same subpopulation than relative to lines in different subpopulations. Finally, we view the largest training population size that we used (300) as a still relatively modest training population.

Accuracy increased with increasing marker density. For β-glucan, days to heading, plant height, and yield, no plateau was reached indicating that more markers would be useful. For groat percentage, however, very minimal increase in accuracy was observed between 600 and 900 markers. It is unclear, however, why a plateau would be reached for some traits but not others. DArT markers may cluster in the oat genome (Tinker et al. 2009). If such clusters happen to coincide with QTL affecting a trait, then a lower marker number would be sufficient to tag all QTL for that trait. Perhaps such an effect occurred with groat percentage. The lower accuracy that was detected for lower marker densities than with higher densities may be explained by the smaller probability of LD between the markers and the QTL when there are fewer markers, hence only a smaller fraction of genetic variation can be explained (Solberg et al., 2008). Using the Kanota x Ogle comprehensive oat map size of 1890 cM (Wight et al., 2003), this data would indicate that on average there is one marker for every 7 cM when 300 markers are used. This assumes even distribution of markers across the genome, while there was one marker for every 2 cM when all the 1005 markers were used. Simulation (Calus et al., 2008) and empirical (Habier et al., 2010) studies have achieved high GS accuracies using data where average LD between adjacent markers (measured as r^2) was 0.20. Newell et al. (2010) explored genome-wide LD in oats and showed that to attain values

of r^2 = 0.20 between markers, one marker per cM was needed. These results indicate that we should still see improvements in accuracy up to at least 2000 markers.

There was increasing mean accuracy and lower standard deviations of accuracies with an increase in training population size. This implies that more lines are needed to improve estimates of marker effects and achieve higher accuracies for GS in oats. What is most remarkable about the increase in accuracy with the increase in training population size is that it showed little sign of reaching a plateau for any of the traits analyzed. We hypothesize that this arises from the high level of diversity for the population that we used (Figure 2). In any event, the result suggests that for training populations that cover several breeding programs, quite large populations will be valuable.

Meuwissen (2009) suggested that an increase in marker density should be coupled with higher training population size to result in higher accuracies. Given the available marker densities in this study, it is more important in the short term to increase the training population size rather than to increase marker density in order to increase GEBV accuracy.

Prediction Using Previous Generations as Training Populations

Making training populations based on their chronological entry on the uniform tests can mimic cultivar development processes, in which previous knowledge of the performance of lines can be used to predict future populations. In this kind of design, both LD and the genetic relationships between training population and selection candidates will contribute to accuracy. But since older generations could have a decreasing genetic relationship to recent generations (for this study see Supplemental Figure 2), the persistence of LD across generations will become more important to maintain accuracy (Habier et al., 2007). The

importance of a larger training population size was again emphasized in this design. For all traits that we examined, increasing the training population by adding older lines caused accuracy to either increase or, at least, remain constant (Fig. 3). The sole exception was yield for the period of 1998-2003, though, when adding even older lines, accuracy again increased. This observation of increased accuracy could be explained by the fact that even quite old lines (e.g., ones from 1994-1998) retained information to predict performance of recent lines (from 2004-2007, Fig. 4).

We compared equally-sized training populations that differed in age, and therefore in the time interval between the training and validation populations (Fig. 4). We expected that older training populations would lead to less accurate predictions. In simulation studies (Habier et al. 2007; Zhong et al. 2009) and in a study of Holstein bulls (Moser et al. 2009), when the training and validation populations were several generations removed, accuracy declined. This expectation only occurred for days to heading and groat percentage. Although oat is capable of going through three generations per year, there is a much slower effective generation time in oat breeding programs in which older inbreds may continue to be used as parents for a number of years. If breeding cycle time decreases in the future, through the use of early selection based on genomic prediction, we would no longer expect that such old training populations would retain as much relevant information.

Prediction of GEBV in Subpopulations

Most breeding programs have unique groupings of parents that are continuously adapted to produce better populations such as heterotic groups in hybrid breeding, or different market classes across a number of crops (e.g., feed versus malt barleys). In this

study, groupings in the population were determined by cluster analysis. Cluster 1 (C1) was composed mainly of oat lines from Canadian oat breeding programs while C2 and C3 were mostly from the US. Cluster analysis revealed that C1 was less related to C2 or C3.

As discussed above, the degree of relationship between the training and validation populations affects accuracy of GS (Habier et al., 2007; Hayes et al., 2009; Habier et al., 2010). This effect occurs whether divergence between training and validation populations arises from generations of descent or from population structure. Thus, for the most part, the C2 and C3 clusters predicted each other better than C1 predicted either one (Figure 5). These findings are similar to that reported by Hayes et al. (2009) for Jersey and Holstein breeds of cattle. This effect of degree of relationship on accuracy was also found within empirical data from four traits of German Holstein Friesian bulls (Habier et al. 2010).

We also found that mixing clusters can offer an alternative design for the training population. When less-related clusters were combined into training populations (i.e., C12 or C13) with the same size as the single clusters, the accuracy was better than the average accuracies for the two single clusters (e.g., average of C1 and C2 versus C12). Using a mixed-subpopulation or multi-breed training population has been explored in cattle by Hayes et al. (2009). Their study revealed that multi-breed training populations (i.e Jersey and Holstein) predicted purebred individuals (Jersey or Holstein) with comparable accuracies to the within breed prediction. In the simulation study conducted by de Roos et al. (2009) on training sets composed of two subpopulations (populations A and B), they showed that accuracy of prediction for selection candidates in A was higher if A and B were less divergent than when A and B were highly divergent. The empirical study of Daetwyler et al. (2010b) in sheep demonstrated that the breed of the selection candidates that was most

represented in multi-breed training populations achieved higher accuracies. Similar to what was found in this study, C12 or C13 training populations provided higher accuracy than C23 on average, because the former had related lines between the training and validation populations while the latter had training and validation populations that were less related.

Accuracy can be increased with higher marker density even if training sets and selection candidates are highly divergent (de Roos 2009). Meuwissen (2009) also suggested that in predicting unrelated individuals, a substantially larger training data set and a higher marker density are required to obtain high accuracies. These results lead to the recommendation that a single large mixed training population with a higher marker density would offer a better solution than multiple training populations, each serving one germplasm group. Higher marker density will help to increase the probability of finding markers that are in consistent LD with the same QTL across the different subpopulations (Daetwyler et al., 2010b). The focus of this strategy will be GS model building in which consistent historical LD across subpopulations is explored rather than just within-subpopulation LD.

We hypothesized that doubling the training set size would be less beneficial when the training population was composed of related individuals (e.g., C23) than when it was composed of unrelated individuals (e.g., C13). That effect was observed for β glucan, plant height, and yield, but not for days to heading and groat percent (data not shown). Results for increasing marker densities were likewise inconclusive. We believe a larger total experiment size would be needed to detect these effects.

Global Comparison of BayesCπ and RR-BLUP for all Training Designs

Training population designs used in this study found that neither GS method was consistently better in terms of accuracy. Simulation studies of Jannink (2010) showed that the difference of these two methods in terms of genetic gain were very small under low (0.20) and medium (0.50) heritabilities and varying training population size of 200 or 1000 However, in this study BayesCπ was consistently better or the same than RR-BLUP for days to heading across different marker density and randomly versus chronologically selected training populations (Figure 1-4). It was also observed that for small training set sizes (90 – 100 lines in our case), BayesCπ outperformed RR-BLUP in four out of five cases for randomly-selected training sets (Fig. 2) and in 13 out of 15 cases for chronologically selected training sets (Fig. 4). Similar results under small training population size were obtained by Meuwissen (2009) though conflicting observations on the performance of these types of models with small training sets have also been reported (Daetwyler, 2010; Habier et al., 2010).

Hayes et al. (2009) conceptualized the performance of multi-subpopulation training populations as dependent on the detection of ancestral LD that is common across subpopulations. This idea would suggest that methods that capture marker – QTL LD will be more effective than methods that model genetic relationships between the training and validation populations (see Habier et al. 2007 and Zhong et al. 2009 for a discussion of these two components of GS accuracy). Thus, we expected BayesCπ to outperform RR-BLUP in analyses where the training population came from a different subpopulation than the validation population, or where the training population was mixed. In fact, we observed the

opposite: RR-BLUP was better than BayesCπ in the cluster-based design for a training population comprised of 120 lines. We have no compelling explanation for this observation though we note that in these cross-subpopulation analyses, we could not include a term to account for population structure in the genomic prediction linear model. Failure of line clustering to account for all effects of subpopulation structure may therefore have played a role.

The difference in terms of average accuracy and standard deviations between BayesCπ and RR-BLUP decreased in larger training populations across different designs in this study. This was similar to the result of Meuwissen (2009) in which BayesB (related to BayesCπ) had similar accuracy with GBLUP (equivalent to RR-BLUP) when using larger training populations. These two methods differ in their assumptions of variance of marker effects; the former uses unequal variance for each marker while the latter assumes that all markers have equal variance. At constant heritability, RR-BLUP is insensitive to genetic architecture (i.e., the number of QTL and the distribution of their effects), while the accuracy of Bayesian methods improves as the number of QTL decreases and their effects increase (Luan et al. 2009; Daetwyler et al., 2010).

Implications for plant improvement programs

Accuracy as a component of response to selection can be used to predict the future gains using GS. As an example, accelerated breeding for β-glucan, a compound found in oats that has been shown to have positive health benefits (FDA Health Claim 21CFR101.81), can benefit from GS. B-glucan is a polygenic trait governed by genes with mainly additive effects and heritability ranging from 0.27 to 0.58 (Cervantez-Martinez et al., 2001). In a

typical phenotypic selection program, β-glucan content is evaluated every year from seeds of replicated plots during the summer season. In order to adapt a GS strategy for β-glucan improvement, in which there are two cycles of selection that can be done in one year (e.g., Jannink, 2010), an accuracy equal to ½ *h* may be enough to justify GS conducted twice a year. Assuming a heritability of 0.5 (h=0.71) versus a GS accuracy of *r*=0.5, GS will lead to around 40% more gain than phenotypic selection per unit time. Genomic selection, however, should be further validated in breeding programs with several generations to determine both advantages and disadvantages and modifications that could potentially maximize genetic gain. As mentioned, GS in plant breeding can be applied in broad-based populations like this study and Heffner et al. (2011) or in narrow-based populations like bi-parental populations (Lorenzana and Bernardo, 2009). Applications of GS with respect to these types of populations differ because of the extent LD: marker density requirements for biparental populations are much lower than for a set of lines with broad genetic diversity. Furthermore, population structure is of no concern in biparental populations since all individuals are equally related. Finally, the time requirement of GS model building will be greater in biparental populations due to the fact that every biparental population will need phenotypic data before model training (Heffner et al., 2011). Specific studies will need to be implemented to determine which GS process is best suited for the crop of interest.

Acknowledgments

This research was funded by the United States Department of Agriculture, National Institute of Food and Agriculture, grant 2008-55301-18746. We thank Adrienne Moran Lauter for laboratory work and Nick Tinker and Charlene Wight for providing oat pedigree data.

References

Calus, M.P.L, T.H.E. Meuwissen, A.P.W. de Roos, and R.F. Veerkamp. 2008. Accuracy of genomic selection using different methods to define haplotypes. Genetics 178:553-561.

Cattell, R.B. 1966. The scree test for the number of factors. Multivar Behav Res 1:245–276.

Cervantes-Martinez, C.T., K.J. Frey, P.J. White, D.M. Wesenberg, and J.B. Holland. 2001. Selection for greater β-glucan content in oat grain. Crop Sci. 41:1085–1091.

Chernyshova, A.A., P.J. White, M.P. Scott, and J.-L. Jannink. 2007. Selection for nutritional function and agronomic performance in oat. Crop Sci. 47:2330–2339.

Colleoni-Sirghie, M., J.-L. Jannink, I.V. Kovalenko, J.L. Briggs, and P.J. White. 2004. Prediction of ß-glucan concentration based on viscosity evaluations of raw oat flours from high-ß-glucan and traditional oat lines. Cereal Chem. 81:434-443.

Daetwyler, H.D., R. Pong-Wong, B. Villanueva, and J.A. Woolliams. 2010. The impact of genetic architecture on genome-wide evaluation methods. Genetics 185: 1021-1031.

Daetwyler, H.D. , J.M. Hickey, J.M. Henshall, S. Dominik, B. Gredler, J.H.J. van der Werf, and B.J. Hayes. 2010. Accuracy of estimated genomic breeding values for wool and meat traits in a multi-breed sheep population. Animal Production Science 50: 1004–1010.

de los Campos, G., H. Naya, D. Gianola, J. Crossa, A. Legarra, E. Manfredi, K. Weigel, and J.M. Cotes. 2009. Predicting quantitative traits with regression models for dense molecular markers and pedigrees. Genetics 182: 375 – 385.

de Roos A.P.W., B.J. Hayes, and M.E.Goddard. 2009. Reliability of genomic breeding values across multiple populations. Genetics 183: 1545 - 1553.

Dekkers, J.C.M. 2007. Prediction of response to marker-assisted and genomic selection using selection index theory. J Anim Breed Genet 124:331–341.

Falconer, D.S., and T.F.C. Mackay. 1996. Introduction to quantitative genetics. 4th ed. Longman Technical and Scientific, Essex, UK.

Garrick D.J., J.F. Taylor, and Fernando R.L. 2009. Deregressing estimated breeding values and weighting information for genomic regression analyses. Genet. Sel. Evol., 41: 55. doi:10.1186/1297-9686-41-55.

Habier D, J. Tetens, F.R. Seefried, P. Lichtner, and G. Thaller. 2010. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet Sel Evol.2010 42:5. doi: 10.1186/1297-9686-42-5.

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

Hayes, B.J., P.J. Bowman, A.C. Chamberlain, K. Verbyla, and M.E. Goddard. 2009. Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genet Sel Evol. 41:51. doi: 10.1186/1297-9686-41-51.

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

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

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

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

Jannink, J.-L. 2010. Dynamics of long-term genomic selection. Genetics Selection Evolution 2010 42:35.

Kang, H.M., N.A. Zaitlen, C.M. Wade, A. Kirby, D. Heckerman, M.J. Daly, and E. Eskin. 2008. Efficient control of population structure in model organism association mapping. Genetics 178: 1709-23.

Kizilkaya, K., R.L. Fernando, and D.J. Garrick. 2010. Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes. J Anim Sci 88: 544 - 551.

Lorenz, A., S. Chao, F. Asoro, E. Heffner, T.Hayashi, H. Iwata, K. Smith, M. Sorrels, and J.-L. Jannink. 2011. Genomic selection in plant breeding: knowledge and prospects. In: D. L. Sparks (Ed.), Advances in Agronomy, Academic Press, San Diego, CA USA. pp. 77-123.

Lorenzana, R.E., and R. Bernardo. 2009. Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theor. Appl. Genet. 120:151-161.

Luan, T., J. A. Woolliams, S. Lien, M. Kent, M. Svendsen, and T.H.E. Meuwissen. 2009. The accuracy of genomic selection in Norwegian Red cattle assessed by cross-validation. Genetics 183:1119-1126.

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

Meuwissen, T.H.E. 2009. Accuracy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping. Genet Sel Evol 41: 35.

Moser G., B. Tier, R.E. Crump, M. S. Khatkar, and H.W. Raadsma. 2009. A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. Genet Sel Evol 41:56 doi:10.1186/1297-9686-41-56.

Newell, M.A., D. Cook, N.A. Tinker, and J.-L. Jannink. 2010. Population structure and linkage disequilibrium in oat (*Avena sativa* L.): implications for genome-wide association studies. Theor. Appl. Genet. DOI: 10.1007/s00122-010-1474-7.

Price, A.L., N.J. Patterson, R.M. Plenge, M.E. Weinblatt, N.A. Shadick, and D. Reich. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38:904–909.

R Development Core Team. 2009. R: A language and environment for statistical computing. Available at ht tp://www.r-project.org (verified 23 Feb. 2011). R Foundation for Statistical Comput., Vienna, Austria.

SAS Institute. 2008. SAS/Stat User's Guide. SAS Institute, Cary, NC.

Solberg, T.R., A.K. Sonesson, J.A. Woolliams, and T.H.E. Meuwissen. 2008. Genomic selection using different marker types and densities. J Anim Sci 2008.86:2447-2454.

Stich, B., J. Mohring, H.-P. Piepho, M. Heckenberger, E.S. Buckler, and A.E. Melchinger. 2008. Comparison of mixed-model approaches for association mapping. Genetics 178:1745–1754.

Taylor, J.R. 1999. An introduction to error analysis: the study of uncertainties in physical measurements. University Science Books. pp. 128–129.

Tinker, N.A., A. Kilian, H.W. Rines, A. Bjornstad , C.J.Howarth , J.-L. Jannink, J.M. Anderson, B.G. Rossnagel, C.P. Wight, D.D. Stuthman, M.E. Sorrells, G.J.Scoles, , P.E. Eckstein, , H.W. Ohm, E.W. Jackson, S. Tuvesson, F.L. Kolb, S.J. Molnar, O. Olsson, M.L. Carson, A. Ceplitis, J.M. Bonman, L. Federizzi, and T. Langdon. 2009. New DArT markers for oat provide enhanced map coverage and global germplasm characterization. Biomed Central (BMC) Genomics. 10(39):1471-2164.

Toosi, A., R.L. Fernando, and J.C.M. Dekkers. 2010. Genomic selection in admixed and crossbred populations. J Anim Sci, January 1, 2010; 88(1): 32 - 46.

U.S. Food and Drug Administration. 2010. Health claims: Soluble fiber from certain foods and risk of coronary heart disease. Available at http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?fr=101.81 (verified 19 May 2011). U.S. FDA, Silver Spring, MD.

VanRaden, P.M., C.P. Van Tassell, G.R. Wiggans, T.S. Sonstegard, R.D. Schnabel, J.F. Taylor, and F.S. Schenkel. 2009. Invited review: Reliability of genomic predictions for North American Holstein bulls. J Dairy Sci 92:16-24.

Wight C.P., N. Tinker, S.F. Kianian, M.E. Sorrells, L.S. O'Donoughue, D.L. Hoffman, S. Groh, G.J. Scoles, C.D. Li, F.H. Webster, R.L.Phillips, Rines H.W., S.M. Livingston, K.C. Armstrong, G. Fedak, and S.J. Molnar. 2003. A molecular marker map in 'Kanota' x 'Ogle' hexaploid oat (Avena spp.) enhanced by additional markers and a robust framework. Genome 46:28-47.

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

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

List of Figures

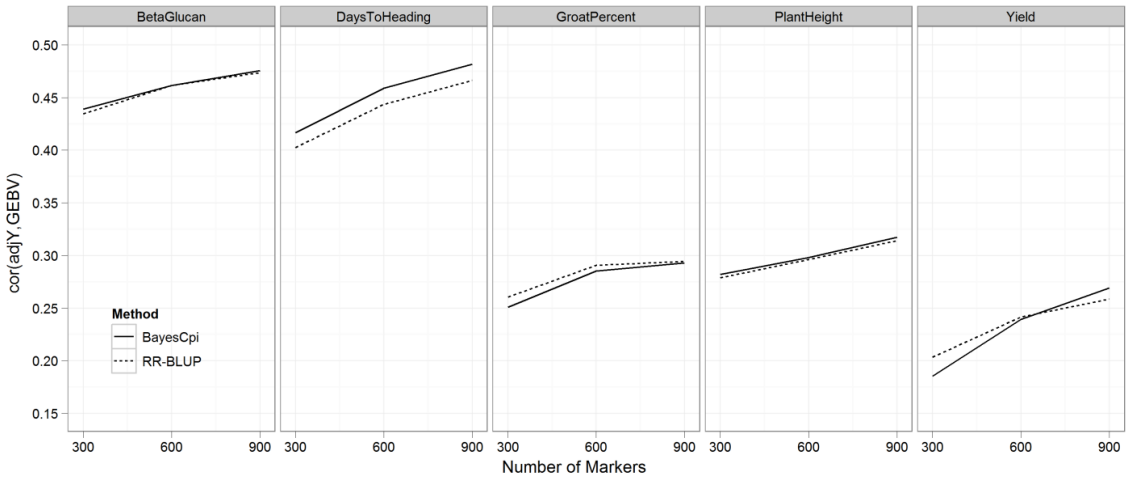


Figure 1. Average accuracies of two genomic selection methods for five traits computed from 100 replicates of randomly selected sets of 300, 600, and 900 markers (x-axis) included in the model and 300 randomly selected lines used as the training population. The y-axis is the correlation of population-structure adjusted phenotypic values and the genomic estimated breeding values (GEBV). All correlations shown were significant ($p < 0.05$).

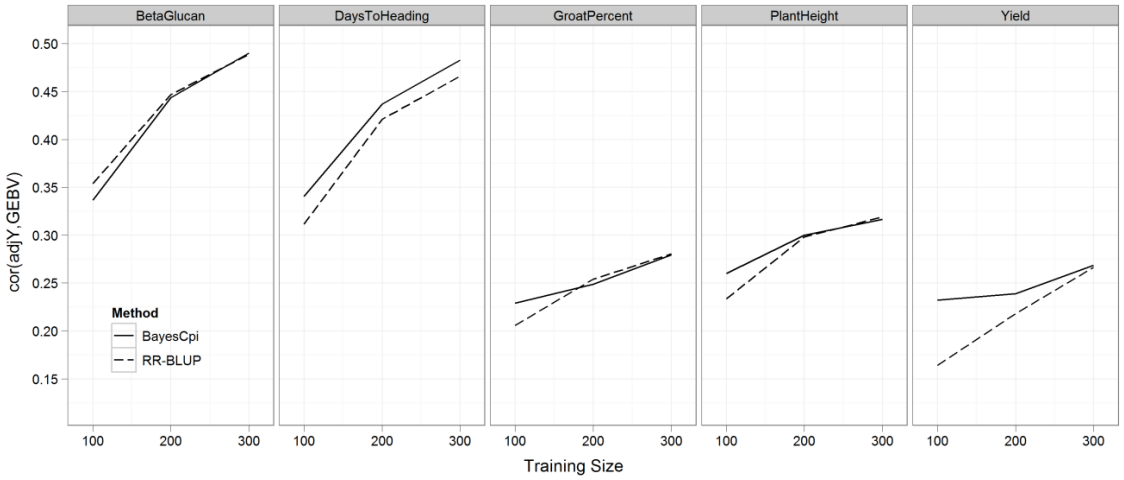


Figure 2. Average accuracies of two genomic selection methods for five traits computed from 100 replicates of randomly selected sets of 100, 200, and 300 lines as training populations (x-axis) with all 1005 markers included in the model. The y-axis is the correlation of population-structure adjusted phenotypic values and the genomic estimated breeding values (GEBV). All correlations shown were significant ($p < 0.05$).

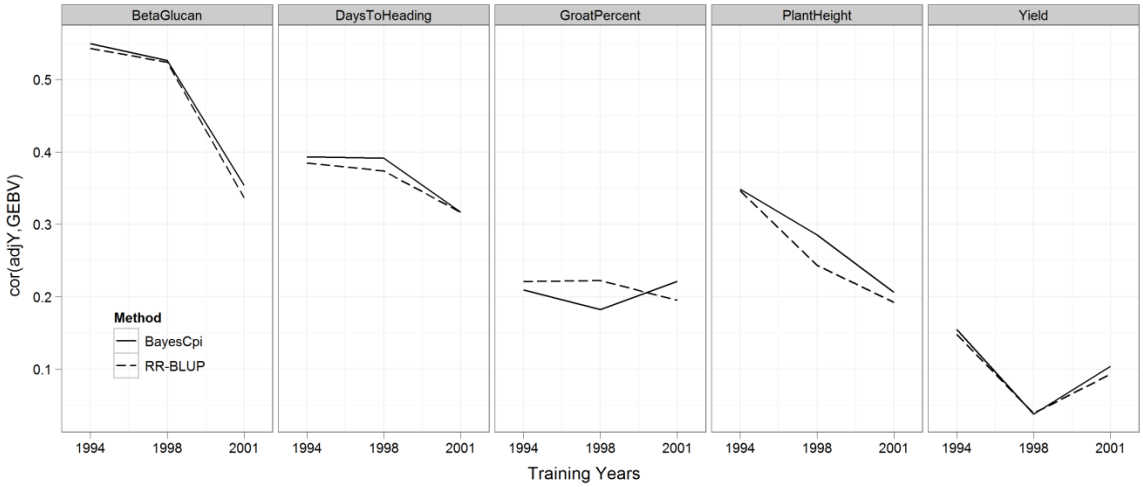


Figure 3. Accuracies for five traits and two genomic selection methods when lines developed during three time periods (1994-2003, 1998-2003, 2001-2003) were used as the training population to predict lines from 2004-2007. The x-axis shows only the beginning year of each period. The y-axis is the correlation of population-structure adjusted phenotypic values and the genomic estimated breeding values (GEBV). The minimum correlation that is significant ($p < 0.05$) is 0.14.

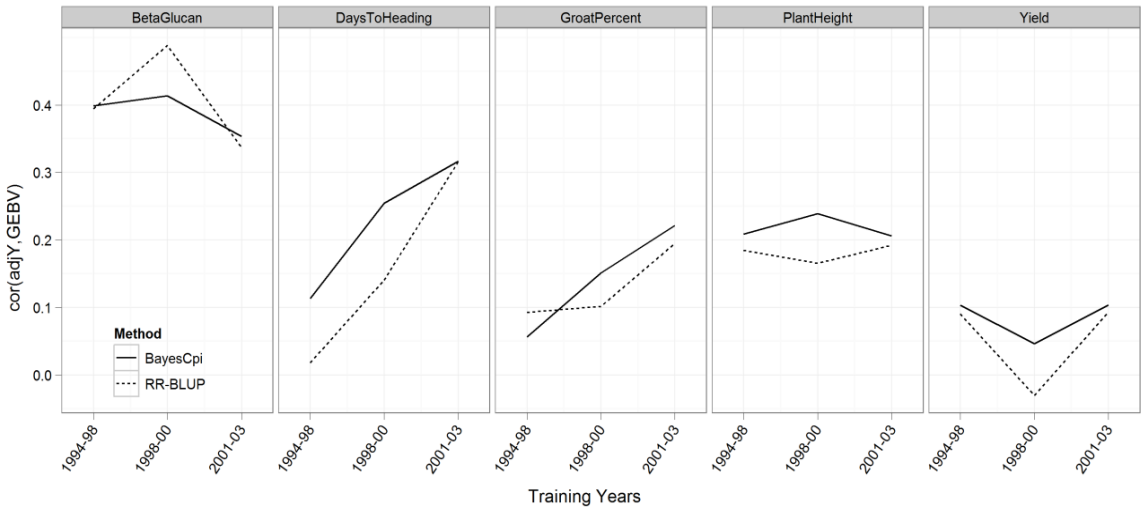


Figure 4. Accuracies for five traits and two genomic selection methods when training populations composed of 90 lines developed during three time periods (1994-1998, 1998-2000, 2001-2003; x-axis) were used to predict lines from 2004-2007. The y-axis is the correlation of population-structure adjusted phenotypic values and the genomic estimated breeding values (GEBV). The minimum correlation that is significant ($p < 0.05$) is 0.14.

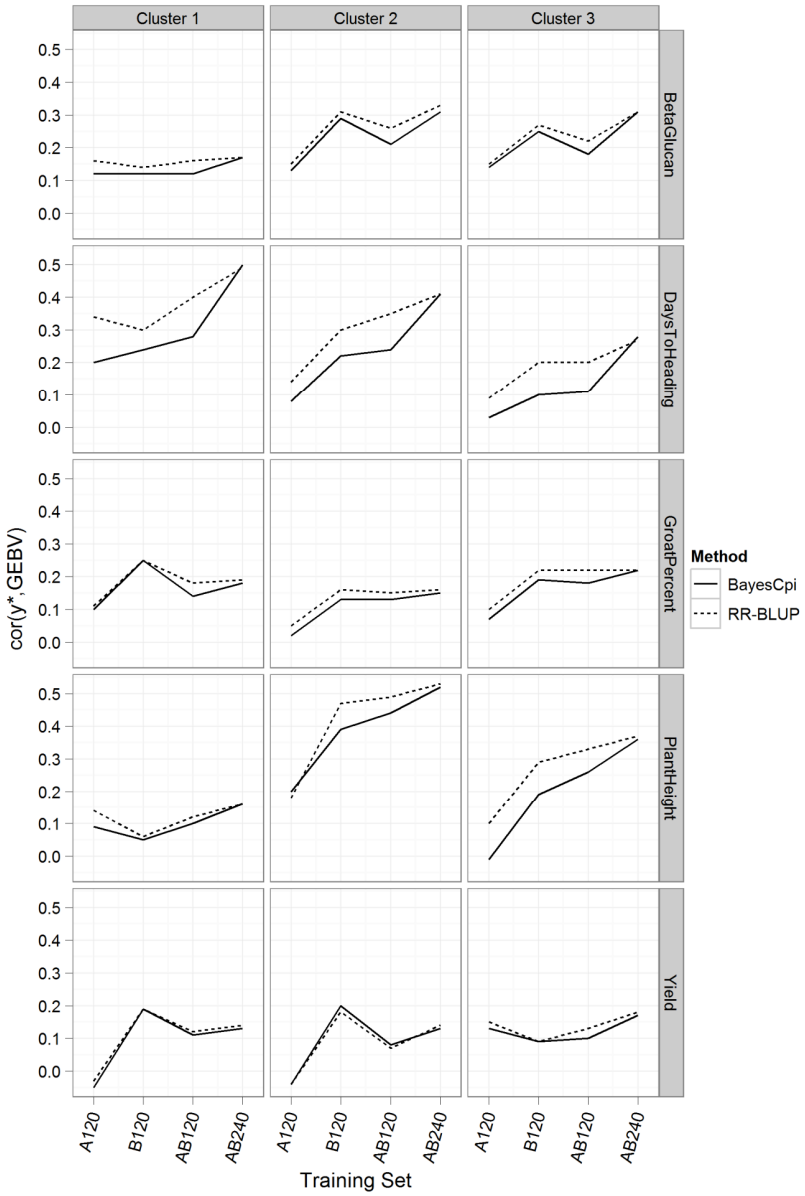
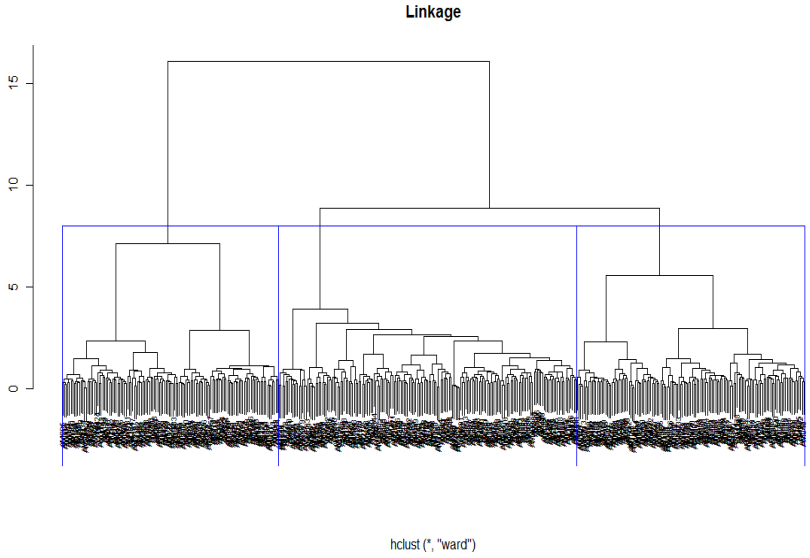
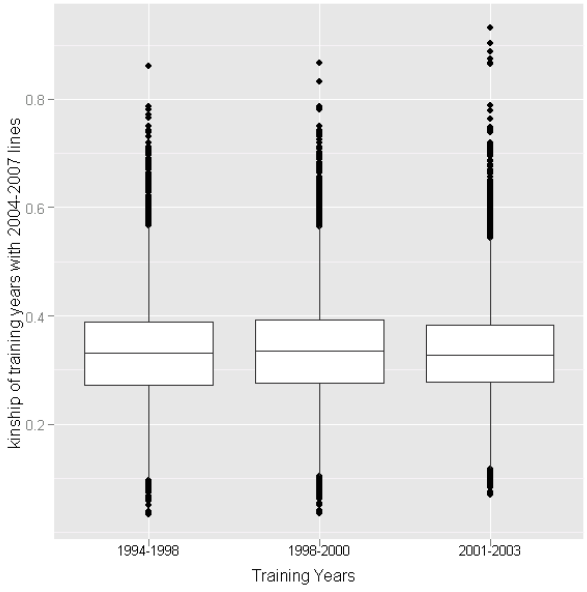


Figure 5. The accuracies of different training populations (x-axis) across traits (row panels) and validation populations (column panels). X-axis notation: The letter denotes the cluster from which lines were sampled for the training population, with A for the lower- and B for the higher-numbered cluster (e.g., for C2 as the validation population, A=C1, B=C3, and AB means equal representation of the two clusters). The number gives the training population size. The y-axis is the correlation of phenotypic values and the genomic estimated breeding values (GEBV). The minimum correlations that are significant ($p < 0.05$) are 0.15, 0.13, and 0.16 for validation populations C1, C2, and C3, respectively.



Supplemental Figure 1. Dendrogram from cluster analysis of 446 oat lines showing the three clusters. Clusters 1, 2, and 3 are depicted from left to right by blue rectangles.



Supplemental Figure 2. Boxplot of kinship (y-axis) of various training years with the validation years (2004-2007 lines). The kinship mean of 1994-1998 lines with validation population(VP,2004-2007) was lower than the kinship mean of 1998-2000 with VP (t-test p=0.002) or the kinship mean of 2001-2003 with VP (t-test p=0.069). In addition, the kinship means of 1998-2000 and 2001-2003 with VP were not significantly different (p=0.163) from each other.

List of Tables

Table 1. ANOVA p-values for factors affecting the accuracies when designing training population with different numbers of markers, number of lines, lines sampled deeper in time, and line of different ages.

Source of Variation	Df	Marker Density	Training Population Size	Training Population Depth	Training Population Age
Trait †	4	<0.0001	<0.0001	<0.0001	<0.0001
Method ‡	1	0.22	0.02	0.21	0.06
Design§	2	<0.0001	<0.0001	<0.0001	<0.01
Trait*Method	4	0.03	0.14	0.56	0.31
Trait*Design	8	<0.01	<0.01	<0.0001	<0.01
Method*Design	2	0.26	0.11	0.64	0.52
Error	8				
Total	29				

† Trait is the five traits (β-glucan, days to heading, groat percent, plant height or yield).

‡ Method is the two genomic selection models (RR-BLUP or BayesCπ).

§Design refers to different factors for each column of the table. Marker Density: number of markers (300, 600 or 900); training population size (100, 200 or 300 lines); training population depth (selection of increasing numbers of lines back in time from the periods 1994-2003, 1998-2003, or 2001-2003); training population age (selection of training populations of equal size from periods of increasing age 1994-1998, 1998-2000 and 2001-2003).

Table 2. ANOVA *p*-values for factors affecting the accuracies for different validation populations generated from three clusters of oat lines denoted by C3, C2 and C1.

Source of Variation	Df	C3 validation population	C2 validation population	C1 validation population
Trait[†]	4	<0.0001	<0.0001	<0.0001
Method[‡]	1	<0.0001	<0.01	<0.0001
Design[§]	2	<0.0001	<0.0001	<0.0001
Trait*Method	4	<0.01	0.03	<0.001
Trait*Design	8	<0.0001	<0.001	<0.0001
Method*Design	2	0.79	0.32	0.04
Error	8			
Total	29			

[†] Trait is the five traits (β-glucan, days to heading, groat percent, plant height and yield).
[‡]Method is the two genomic selection models (RR-BLUP and BayesCπ).
[§]Design in this table refers to three training populations of 120 lines sampled from clusters other than the corresponding to validation population. For example, Design levels for C3 validation population were the training populations: C1, C2, C12 at 120 lines.

Supplemental Table 1. Accuracies of clusters as training populations used in Figure 5.

Trait	Validation Set	Cluster 3				Cluster 2				Cluster 1			
	Training Set	Cluster 1	Cluster 2	Cluster 1 and 2		Cluster 1	Cluster 3	Cluster 1 and 3		Cluster 2	Cluster 3	Cluster 2 and 3	
	training population size	120	120	120	240	120	120	120	240	120	120	120	240
BGlucan	BayesCπ	0.14	0.25	0.18	0.31	0.13	0.29	0.21	0.31	0.12	0.12	0.12	0.17
	RR-BLUP	0.15	0.27	0.22	0.31	0.15	0.31	0.26	0.33	0.16	0.14	0.16	0.17
DaysToHeading	BayesCπ	0.03	0.1	0.11	0.28	0.08	0.22	0.24	0.41	0.2	0.24	0.28	0.5
	RR-BLUP	0.09	0.2	0.2	0.27	0.14	0.3	0.35	0.41	0.34	0.3	0.4	0.49
GroatPercent	BayesCπ	0.07	0.19	0.18	0.22	0.02	0.13	0.13	0.15	0.1	0.25	0.14	0.18
	RR-BLUP	0.1	0.22	0.22	0.22	0.05	0.16	0.15	0.16	0.11	0.25	0.18	0.19
PlantHeight	BayesCπ	-0.01	0.19	0.26	0.36	0.2	0.39	0.44	0.52	0.09	0.05	0.1	0.16
	RR-BLUP	0.1	0.29	0.33	0.37	0.18	0.47	0.49	0.53	0.14	0.06	0.12	0.16
Yield	BayesCπ	0.13	0.09	0.1	0.17	-0.04	0.2	0.08	0.13	-0.05	0.19	0.11	0.13
	RR-BLUP	0.15	0.09	0.13	0.18	-0.04	0.18	0.07	0.14	-0.03	0.19	0.12	0.14

[†] The minimum correlations that are significant ($p < 0.05$) are 0.15, 0.13, and 0.16 for validation populations C1, C2, and C3, respectively.

Appendix

Appendix 1. Pre-GS Analysis of Phenotypic Data in SAS.

Mixed Model: $y= Xb + Zu + e$

y=phenotypic data from unbalanced multi-environment trials

X=design matrix for environments

b=fixed environmental effects

Z=design matrix for oat lines

u=random oatline effects

For GS purpose in this study, $y^* = u +$ overall mean was treated as the observed value of each oatline.

CHAPTER 4. SELECTION METHODS FOR β-GLUCAN CONTENT IN ELITE OAT GERMPLASM: SHORT-TERM RESPONSE

A paper to be submitted to Crop Science

Abstract

β-glucan, a soluble fiber found in oat grain, has documented benefits in human health and nutrition and selection for higher levels of this compound is regarded as an important breeding objective. Recent advances in molecular marker technologies for oat present an opportunity to investigate new selection methods for polygenic traits such as β-glucan content. Our objectives in this study were (1) to compare genomic selection, marker-assisted selection and BLUP-based phenotypic selection in terms of short-term response to selection, and (2) to assess correlated response to selection for β-glucan content and other traits. Starting with a collection of 446 elite oat lines from North America, each selection method was replicated in two populations for two cycles of selection. The average β-glucan content increased by 2.09 to 2.31 % after two cycles of all selection methods from an average of 4.57 % in Cycle 0. The averages of marker-based selection methods after Cycle 2 were slightly greater than those of phenotypic selection. Moreover, the highest β-glucan progenies came from the marker-based selection methods, demonstrating the effectiveness of molecular markers in increasing the means and developing superior progenies. However, MAS for higher β-glucan content also resulted in a later heading date. Overall, the results of these experiments suggest that genomic selection is the superior method for selecting a polygenic complex trait like β-glucan content.

Introduction

Mixed-linkage (1→3), (1→4) β-D-glucan (commonly referred to as β-glucan) is a cell wall polysaccharide that can be found in high concentrations in the endosperm and aleurone layer of oat (*Avena sativa* L., Butt et al., 2008). This compound has been identified as the active component of soluble fiber that lowers blood serum cholesterol levels - a major risk factor for coronary heart disease (reviewed by Tiwari and Cummins, 2009; US-FDA, 2010). Therefore, improving the β-glucan content of oat is desirable for human health and nutrition (Peterson et al, 1991). Recently, a survey of β-glucan content in elite North American oat varieties showed that it ranges from 3.15 to 7.62 % on a dry weight basis (Asoro et al., in review). Research has shown that β-glucan content in oat is controlled by genes that behave additively, with heritability from 0.39 to 0.58 (Holthaus, 1996; Kibite and Edney, 1998; Chernyshova et al., 2007). Although β-glucan content may be influenced by environment, reports have shown that the ranking of varieties is mostly consistent across environments (Peterson, 1991; Cervantez-Martinez., 2001).

The development of more cost efficient and abundant genetic markers has significantly improved the genome coverage for oat (Tinker et al., 2009). This offers alternative ways to explore marker-assisted selection methods in oat breeding. One of these methods is to use genome-wide association studies (GWAS) to identify markers associated with QTL and incorporate those markers in predicting breeding values like in traditional marker-assisted selection (MAS, Lande and Thompson, 1990). A GWAS-based MAS can alleviate unintended interactions between major QTLs and other genetic background, a problem attributed to traditional MAS (Kennedy et al., 1992). Those interactions can be avoided in GWAS-MAS because of the use of large panel of representative germplasm as

populations for identifying QTL (Bernardo, 2008; Xu and Crouch, 2008; Heffner et al., 2011a). However, due to the fact that only a subset of markers is still used in GWAS-MAS, this method may capture only a portion of the total genetic variation and still retains the problem of overestimated marker effects (Beavis, 1994). Hence, GWAS-MAS can still result in a less accurate estimate of the breeding values. An alternative to MAS methods is genomic selection (GS) or genome-wide selection, which predicts the breeding values of individuals using all markers throughout the genome (Meuwissen et al., 2001). The use of all markers allows for selection based on a larger proportion of the genetic variation, therefore resulting to more accurate estimates of breeding values (Goddard and Hayes, 2007).

Comparative studies of GS, MAS, and phenotypic selection, as measured by components of response to selection ($R= ir\sigma_A$, where R is the response, i is the intensity, r is the accuracy of selection and σ_A is the additive genetic standard deviation; Falconer and Mackay, 1996), revealed that GS has a consistent advantage over the other methods. For example, in simulated breeding programs in maize, there was higher response to selection using GS than recurrent MAS (Bernardo and Yu, 2007; Mayor and Bernardo, 2009). In empirical cross-validation studies, GS had greater accuracy of selection than traditional MAS conducted among segregating progeny of biparental crosses and also greater than GWAS-MAS performed in a multi-family wheat population (Lorenzana and Bernardo, 2009; Heffner et al., 2011a; Heffner et al., 2011b). For comparison of GS and phenotypic selection, it was shown that accuracy of selection was higher in GS than phenotypic selection using pedigree information (Nelsien et al., 2009; Crossa et al., 2010). On the other hand, empirical studies in wheat multi-family populations have shown that the GS method was comparable only to phenotypic selection prediction accuracy (Heffner et al., 2011a). In oat, preliminary GS

studies for β-glucan content showed that GS with two cycles per year can have a 40% greater response than phenotypic selection assuming a heritability of 0.50 (Asoro et al., 2011). These studies support the idea that the GS advantage would likely come from the cumulative response generated by several selection cycles per year (Jannink, 2010; Mayor and Bernardo, 2009).

Despite the empirical cross-validation studies showing the advantages of using molecular markers, there is still a need to translate these results in actual breeding programs. The comparison of selection methods under a replicated selection programs can provide empirical validation before incorporating molecular marker strategies in larger breeding programs. A breeding program for β-glucan content in oat is appropriate for this kind of experiment because of its polygenic nature and the benefits of breeding for this trait. To make the comparison, we implemented BLUP-based phenotypic selection (Henderson, 1984), a GWAS-MAS with re-estimation of marker effects between cycles, and GS. Our specific objectives were to: (i) develop two oat populations for each of the three selection methods for β-glucan content; (ii) compare the short-term response and ability to develop superior progenies of the three selection methods over two cycles of selection and (iii) examine changes in correlated response in heading date and plant height of oats.

Materials and Methods

Marker Data for Cycles 0, 1 and 2

Oat lines and their progenies were planted in the Iowa State University Agronomy greenhouse in January 2008 (Cycle 0), January 2010 (Cycle 1) and January 2011(Cycle 2). Leaf samples were collected for each line and DNA was extracted according to

recommended protocol for Cycle 0 and 1 (Diversity Arrays Technology, Yarralumla, Australia). For Cycle 2, the DNA samples were extracted using a kit (QIAGEN, Valencia, CA). In every cycle, DNA from each line was sent to the Diversity Array Technology laboratory for genotyping.

Phenotypic Data of Base Population (Cycle 0)

The Cycle 0 population was composed of 446 lines from various oat breeding programs in North America. These lines were tested in the Uniform Oat Performance (UOPN) and the Quaker Uniform Oat Nurseries (QUON) for agronomic traits and other biochemical characters including β-glucan content from 1994 to 2007. Most β-glucan data (97%) in this study were from those nurseries and stored in Graingenes 2.0 database (Carollo et al., 2005). A small amount of β-glucan data were included from Chernyshova et al. (2007), Colleoni-Sirghie (2004), the Germplasm Resources Information Network, and the North Dakota State University Oat experiments (M. McMullen, personal communication).

Genomic Selection of 12 Parents for Cycle 1

The genotype matrix (M) in Cycle 0 was used to derive a marker-based relationship matrix equal to $MM'/\sum_k p_k(1-p_k)$ where p_k is the frequency of allele 1 in marker k computed using the Spagedi program (Hardy and Vekemans, 2002). Because there were negative relationship values, the resulting matrix was then scaled between 0 and 1. The same relationship matrix was also used to calculate the principal components using SAS PROC Princomp (SAS Institute, 2008). Only the eigenvectors of the first three PC axes (denoted as

P) were used in the association analysis because succeeding axes accounted for only a small proportion of the variation based on scree plot (results not shown).

To compute the genomic estimated breeding values (GEBV), a mixed model methodology was implemented in PROC MIXED using the following model:

$$y = Xb + Ej + Zu + e$$

where y is a vector of β-glucan values for each line from Cycle 0, b is the mean, j is a vector of random environmental effects, u is a vector of random polygenic effects, and e is a vector of residual errors. Observations for four long-term checks were also used to provide overlap across environments. The X, E and Z terms are the incidence matrices relating y to b, j and u, respectively. The variance of $u=KV_A$, where K is the marker-based relationship matrix and V_A is the additive variance due to polygenic effects derived using the REML option in PROC MIXED.

Cycle 0 lines were sorted based on random effects values and the highest 40 lines were selected. The marker-based relationship matrix of those 40 lines was then subjected to cluster analysis using Ward's linkage with 12 clusters in SAS PROC CLUSTER (SAS Institute, 2008). The line with the highest β-glucan content per cluster was selected for use in the final set, thus 12 parents were selected for this method.

MAS of 12 Parents for Cycle 1

To implement MAS, significant markers were first identified through association mapping. A two-stage association mapping was conducted because it was less computationally demanding and has produced results similar to a one-stage analysis (Stitch et al., 2008). First, a similar analysis to genomic selection described above was conducted

except that the 450 lines (446 lines plus four checks) were assumed to be independently and identically distributed (i.e. no relationship matrix among lines was included in the model). The solution for random effects of the 446 lines plus the grand mean was treated as the new observation for association mapping (Zhang et al., 2009). Second, the association test for β-glucan content was conducted using the TASSEL software (Bradbury et al., 2007) with the following model:

$$y^* = Xb + Ma + Ps + Zu + e,$$

where y^* is the vector of observations for β-glucan content as described above, b represents the mean, a is the marker effect, s is a vector of population structure effects, u is a vector of random polygenic effects, and e is a vector of residual error. The X, M, and Z are incidence matrices relating y to b, a, and u, respectively, while P is the matrix from PCA computed above relating s to y. The variance of $u=KV_A$, where K is the marker-based relationship matrix, and V_A is the additive variance due to polygenic effects. Using the p-values for each marker, six markers potentially controlling β-glucan content were identified using a false discovery rate (FDR) of 0.33 (Benjamini and Hochberg, 1995).

To estimate genetic effects, the six markers were included in a model that was analysed jointly using PROC MIXED in SAS with similar population structure and polygenic effects specifications as described above, except that the response variable was replaced by the original set of observations. The resulting marker and population structure effects plus the phenotypic values were then used to calculate an index (Lande and Thompson, 1990):

$$Index\ value = Ma + Ps + Phenotypic\ values,$$

where M is the genotype data matrix for the six markers and a is their corresponding estimated marker effect, P is the principal component eigenvectors matrix and s consisted of

the corresponding population structure effects. The phenotypic values were the y* values used in the association analysis model. The index values were then used to rank the 446 lines. Marker-based relationship matrix of the top 40 lines were subjected to cluster analysis using the same approach as mentioned for genomic selection to select 12 parents for use in creating cycle 1.

BLUP Phenotypic Selection of 12 Parents for Cycle 1

First, the pedigree record of each line was taken and confirmed from Pedigree of Oat Lines (POOL) database (Tinker and Deyl, 2005). Then, the pedigree data of 450 lines were used in the kinship software KIN (Tinker and Mather, 1993) to derive the coancestry matrix. The coancestry matrix ranged from 0 to 1, where 0 refers to two lines unrelated through pedigree and 1 as the diagonal representing perfect identity by descent. A mixed model method in SAS was used to determine the pedigree-based BLUP values of lines (Henderson, 1984). The model in this analysis was similar to genomic selection methodology except that the covariance matrix among lines was defined by the pedigree-based coancestry. The BLUP values were also sorted and the highest 40 lines were selected. The coancestry matrix of those 40 lines was subjected to cluster analysis using PROC CLUSTER in SAS (SAS Institute, 2008) with Ward's linkage and 12 clusters as options. The line with the highest β-glucan per cluster was selected to acquire the 12 parents for creating Cycle 1.

Recombination Scheme for Cycle 1 of each Selection Method

The 12 parents for Cycle 1 of each selection method were planted in December 2008 in the greenhouse. Two replications of a partial diallel (Kempthorne and Curnow, 1961) were

conducted for each selection method wherein each parent was crossed to four other parents without reciprocals to generate the F1 seeds. The replications of selection methods were denoted as GR1 or 2 for first and second replication of GS, MR1 or MR2 for MAS, PR1 or PR2 for BLUP phenotypic selection (Figure 1). The resulting 24 F1 crosses per replication were planted in September 2009 in the greenhouse to develop the F2 generation. Two seeds from each cross in the F2 generation were randomly obtained and grown from January to April 2010. Simultaneously, each F2 plant from the populations undergoing MAS and GS methods was genotyped using the protocol mentioned above. Seeds resulting from self-pollination of each plant (F2:3 progenies) were harvested separately and used for field evaluation in the Summer of 2010 (Supplementary Method 2). Phenotypic data were measured on each plot included: days to heading measured as the number of days from planting until 50% of tillers have panicles; plant height, measured in centimeters from ground to tip of the panicle; and β-glucan content. β-glucan content was measured using an enzymatic method in microplates (Newell et al., in review; Megazyme, Inc.) and was expressed as a percentage of beta-glucan on a dry weight basis.

GS, MAS and PS for 12 Parents of Cycle 2

For each selection method, the response variable was the entry effect values for each line computed from phenotypic measures obtained in Summer of 2010 (Supplementary Method 2). For GS, the selection was conducted by estimating the marker effects of all markers in both Cycle 0 and Cycle 1 using the RR-BLUP method (Meuwissen et al., 2001; Lorenz et al., 2011). For each individual in Cycle 1 populations (GR1 and GR2), the sum of effects from all markers was computed to estimate the breeding values.

For MAS, estimates of marker effects of six markers were computed with a mixed model using the F2:3 line effects data as response variables, marker identities as fixed effects and covariance matrix of F2:3 lines defined by the coancestry. Then, each marker effect was multiplied to the corresponding marker allelic states and summed across markers to compute the total marker scores (Lande and Thompson, 1990). An index containing both phenotypic and marker scores of F2:3 lines was developed, where the former had weight of 1.00 and the latter had a weight of 1.35 as described by Lande and Thompson (1990). Specifically, the weight of 1.35 for marker score was derived using the formula $b=[(1/h^2)-1]/(1-p)$, where the estimated h^2 for β-glucan was equal to 0.44 and p was the proportion of variance explained by the markers which was 0.06 based on the original association test.

For each BLUP phenotypic selection population (PR1 and PR2), the $F_{2:3}$ line effects were re-fitted as response variables in a mixed model where the pedigree-based coancestry of lines was used as a covariance matrix among the lines. The resultant estimated breeding values of the lines were ranked.

Because there were only 34-45 lines from the GS and MS populations with high quality marker data, a random selection of 36 lines were taken for all populations, including the PS populations. Estimated breeding values of randomly selected lines were ranked per population and the top 12 parents were selected as parents for Cycle 2. Finally, the 12 parents of each population were planted in the greenhouse in September 2010. A recombination scheme similar to Cycle 1 was conducted and two seeds from each cross were selected randomly at maturity to form 48 S0 lines for each population (Figure 1). S0 seeds from each population were planted in the greenhouse in January of 2011 for advancement from S0 to S1 (see Supplementary Method 3 for complete details).

Field Plot Design

The same field plot design was used for 2010 and 2011 evaluation (see Supplementary Method 2 for details of 2010). For each year, entries were evaluated in an incomplete block design with two replications, where blocks were nested within replications. However, the field evaluation for 2011 had additional set entries composed of random lines from each population in Cycle 1. Specifically, in 2011 the entries were comprised of a random sample of 24 lines from each population of Cycle1 (total of 144), the 48 random lines from Cycle 0, the 20 unique parents of Cycle 1 and four popular oat cultivars (total of 24), 288 S0:1 lines from Cycle 2, and five checks (IAN9N79-5-1-22, Baker, IA002130-2-2, Excel, and CDC Pro-Fi). For each incomplete block, the entries consisted of a random sample of six lines from Cycle 0, a random sample of three lines from each population of Cycle 1 (total of 24), a random sample of three lines from the parents of Cycle 1, six lines from each population of Cycle 2 (total of 36), and all of five checks. Each incomplete block was composed of 7 by 11 grid of hillplots. The experiment was grown at the Iowa State University Agronomy and Agricultural Engineering Field Research Center near Ames, IA from April to July 2011. Each hillplot was harvested by hand and threshed after one week of air-drying. The same set of data as Cycle 1 was gathered in this field evaluation.

Data Analysis for Comparison of Selection Methods

The following model was used to fit the combined data from 2010 and 2011 field evaluations using PROC Mixed in SAS:

$$y = Xb + Zu + error$$

where *y* was the data collected (β-glucan content, heading date and plant height). X is the design matrix for the following fixed terms: *grand mean + year + replication (year) + incomplete block (replication*year) + population* while *b* is the vector of corresponding effects. Z is the incidence matrix for the entries while *u* is the entry effect. The variance of $u=I\sigma^2_g$, where I is the identity matrix and σ^2_g is the genetic variance estimated from the data. Out of a total of 2000 plots over two years, 29 had missing data for β-glucan content due to non-germination or insufficient numbers of seeds for planting. The *population* term is composed of the combinations of the population itself and cycle (eg.GR1-Cycle 1, GR1-Cycle 2), parental lines for each population, checks, and the random Cycle 0 lines. These analyses were conducted to estimate lsmeans for each population of entries. Two models were compared using a goodness of fit test, the first assumed homogeneous variance and the second assumed heterogeneous variance between populations (e.g.GR1-Cycle 1, GR1-Cycle 2), checks and Cycle 0. Because the goodness of fit test showed that the model with heterogeneous variance performed better, the solution for fixed and random effects from this model was used in subsequent analyses. The BLUP of each oat line was computed as the combination of fixed and random effects BLUP= grand mean + population effect + oat entry effect.

Results

Marker-Trait Associations

The association analysis conducted in Cycle 0 demonstrated that the individual phenotypic variance explained by each of the six significant markers was close to one percent (not shown). The estimated effects ranged from 0.23 to 0.44 for Cycle 0, 0.04 to 0.24 for

Cycle 1, and 0.06 to 0.59 for Cycle 2 (Table 1). The favourable allele frequencies of the six markers ranged from 0.02 to 0.93 with an average of 0.27 for Cycle 0, 0.21 to 1.00 with an average of 0.56 for Cycle 1, and 0.34 to 0.96 with an average of 0.58 for Cycle 2.

Means of Populations for β-glucan Content

The combined analysis of data from years 2010 and 2011 displayed the grand mean of β-glucan content for this study was 4.83%. The mean of Cycle 0 was 4.57% and was significantly lower from the means of each of the Cycle 1 and Cycle 2 populations (Table 2). The means of populations in Cycle 1 were not significantly different from one another and ranged from 5.89 to 6.01%. These means were greater than Cycle 0 by 1.32 to 1.44 % β-glucan content (Table 3). In addition, the mean of the pooled set of 20 parents of Cycle 1 (5.72%) was not significantly different from the means of their progenies (i.e., all populations in Cycle 1).

The means of populations in Cycle 2 ranged from 6.66 to 6.88 and were significantly different from their respective population means in Cycle 1 (Table 2, Figure 2). These values were greater than their respective Cycle 1 means by 0.72 to 0.93 % β-glucan (Table 3). Over-all, the corresponding cumulative increase in β-glucan content ranged from 2.09 to 2.31% after two cycles of selection.

The individual population means in Cycle 2 were not significantly different from each other. However, the higher amount of β-glucan content (0.28%) in marker-aided selection populations (i.e. GR1, GR2, MR1 and MR2) than phenotypic selection populations in Cycle 2 was significant at p-value 0.08 (Table 2). The mean of parents for each population in Cycle 2 was not significantly different from their respective progenies.

Correlated Response to Selection

The responses in β-glucan content were accompanied by response in days to heading and plant height but these were significant only between Cycle 1 and 2 of MAS method (Table 4, Figure 2). Specifically, the increase in β-glucan content also resulted in an additional 3.38 mean days to heading date of progenies of MAS populations. However, the opposite occurred in plant height where selection for higher β-glucan content using MAS reduced the plant height by 5.61 cm.

Progeny Performance for β-glucan Content

The progenies, parents, and checks in the evaluation trial displayed a large range of β-glucan content BLUP values (3.86 to 9.06 %, not shown). The random sample of lines from Cycle 0 had β-glucan content ranging from 3.86 to 6.81, the progenies in Cycle 1 had values ranging from 4.38 to 8.33, and the progenies in Cycle 2 had values ranging from 6.05 to 8.11 (Figure 3). Of the 20 lines with the greatest β-glucan content, 11 of the lines were derived from the genomic selection populations, eight were from the marker-assisted selection populations, and one line was from the phenotypic selection populations (Table 5). In the top 20 lines, six were Cycle 1 progenies (of which three were parents of Cycle 2) and 14 were Cycle 2 progenies.

Discussion

Response to Selection for β-Glucan Content

The mean β-glucan content of every population in Cycle 1 was greater than the mean of Cycle 0 with percent response of 29-32 %. The first cycle of selection was conducted

based on data from oat growing regions of North America, but there were still substantial gains when evaluation was conducted only in Iowa, a stressful environment for oat (Cervantez-Martinez et al., 2001). These gains imply that β-glucan content could be stable across diverse environments. The responses detected were greater than the 4-11% reported by Cervantez-Martinez et al. (2001) after 1 cycle of selection for β-glucan content. This occurred despite a comparable proportion of parents selected in our study with 0.027 (12 parents /446 Cycle 0 lines) and in Cervantez-Martinez et al. (2001) with 0.024 (40 parents/1665 Cycle 0 lines). A possible reason for this dissimilarity could be the differences in the respective base populations. In our study the base population consisted of the breeding lines themselves (i.e mostly inbred cultivars) whereas in Cervantez-Martinez et al. (2001) the base population consisted of the progenies of random mated F1 crosses (S0 lines) from 23 breeding lines. Therefore, it is possible that their base population had a smaller genetic variance than our base population resulting in a lower response (Fehr, 1987; Bernardo, 2010).

The means of populations in Cycle 2 were greater than the means in Cycle 1. However, all populations in Cycle 2 had a lower rate of response than their counterpart in Cycle 1. One obvious reason for this is the fact that we used a lower selection intensity (i = 1.097, proportion = 0.33) for selecting parents of Cycle 2 than for selecting parents of Cycle 1 (i = 2.32, proportion = 0.027) in the previous cycle. Nonetheless, the results indicate different methods were effective in increasing β-glucan content in elite oat.

Comparison of Responses Across Selection Methods

The comparable performance of all populations in Cycle 1 across selection methods could be due to the similarity of some parents given that there were only 20 unique parents

utilized across methods. In Cycle 2, it was again observed that means based on pairwise comparisons of all populations were significantly ($p < 0.1$) different. In particular, a difference was detected in the contrast of means for the marker-aided methods (i.e. GS and MS) versus PS. Moreover, populations from GS and MS gave a consisently larger response than the PS populations. This suggests that the marker-aided methods produced more progenies with greater β-glucan content than that for PS populations. Because markers were not used to accelerate the breeding cycle in this study, the advantage of using markers in this study is due to the ability to identify parents with the best breeding values. Specifically for GS, this advantage is likely from the increased accuracy to model breeding values for β-glucan content and the use all marker effects regardless of their size. For MS, this advantage is likely due to the rapid increase in the frequency of favorable alleles in Cycle 1 that resulted in the parents of Cycle 2 with high breeding values.

Correlated Response to Selection

The correlated response of β-glucan content with other traits should be considered to avoid undesirable shifts during selection. In our study, correlated responses for heading date and plant height were detected only in the MAS method. In a previous phenotypic selection experiment by Cervantez-Martinez et al., (2002) it was shown that selection for β-glucan content in oat had no association with heading date while a reduction in plant height was detected after one cycle of selection. On the other hand, β-glucan content tended not to show correlation with heading date and plant height when evaluated in populations not undergoing selection (Holthaus et al, 1996). One potential reason for the occurrence of correlated response in MAS lies in the position on genetic maps of the QTL of β-glucan content,

heading date and plant height. For example, a β-glucan marker (oPt.8249) localized 14 cM away from major heading date QTL – bcd1968B and bcd1797D in linkage group 24_26_34 while another marker (oPt.7232) is located 3-6 cM away from another heading date QTL – cdo1467A, umn370, isu1755B, isu1364 in linkage group 17 (Holland et al., 1997; Tinker et al., 2009). Similarly, marker oPt.8249 is 8-14 cM away from plant height QTLs bcd1643A and umn220. Co-localizations of QTL for β-glucan, heading date and plant height were also observed by Kianian et al. (2001) and De Koeyer et al (2004). The proximity of these QTL in oat genome and the weight that was given to β-glucan content markers during MAS could have caused them to have correlated response. The lack of a correlated response with PS and GS is likely due to the polygenic in nature of selection in these approaches. Therefore, it is not expected that all QTL for β-glucan are associated in one direction with QTL for heading date and plant height.

Progeny Perfomance for β-Glucan Content

One way to judge the differences among populations or selection methods is their ability to produce the best progenies (Zhong and Jannink, 2007). We expect that one further cycle of selection would lead progenies of populations under marker-based methods to have higher means than progenies of population under BLUP phenotypic selection (PS). This follows because the possible selection differential of best parents from each population of GS and MAS could be higher than selection differential of best parents under BLUP phenotypic selection (Figure 3).

In this study, we noted that 19 progenies in the top 20 high β-glucan content entries came from GS and MS populations. From the examination of pedigrees of the top 20

progenies, 11 out of 20 of the original parents in Cycle 1 are in the pedigrees of the top 20 progenies (Table 5, Supplementary Table 1) Of those 11, five were developed by the Iowa State University oat breeding program. In addition, three of the 11 lines have relationships with one another, specifically ND030288 was derived from a cross between Hifi and IAN979-5-1-22. The high occurrence IAN979-5-1-22 either directly as parent or as part of ND030288 in the pedigrees of both Cycles 1 and 2 confirms that the former has favorable alleles for β-glucan content. Two other parents that are frequently present were IA95111 and AC Antoine (a line from a AAFC, Canadian breeding program), both could be valuable sources for β-glucan alleles that may not come from IAN979-5-1-22.

Breeding Implications

The one-year per cycle recurrent selection system (Frey et al., 1988) implemented in this study is seldom used in oat breeding, perhaps because of difficulties in making many crosses during the recombination stage. However, this work has demonstrated that recurrent selection is highly effective in achieving rapid gains and superior progeny for a trait such as β-glucan. The superior progenies that were developed during this experiment have been submitted to the National Small Grains Collection (Aberdeen, Idaho) for preservation and distribution (Supplementary Table 2). These top progenies can be tested for additional agronomic traits in advanced trials, and used in furtherstrategies for variety development and germplasm enhancement.

With regards to different selection methods, the advantage of MAS and GS over PS was minimal on a per cycle basis. The results from this selection program support the findings in cross-validation experiments of Heffner et al. (2011a; 2011b) in wheat, where MS

and GS accuracies were comparable to those of PS. We conclude that the impact of GS to increase response must come from conducting at least two cycles per year, which is not possible for phenotypic selection. In this scheme, one cycle is conducted at an off-season location and the other in the target environment with the addition of phenotypic data. The presence of progenies from GS and MAS methods among the top performing lines also suggests the superiority of these methods in cultivar development. However, index MAS may not have the advantage of GS given that phenotypic data collected during the summer season will still be needed to account for a polygenic effect (Dekkers, 2007). In addition, in our study, the use of few markers in MAS for β-glucan content resulted in a negative effect on oat maturity.

References

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

Beavis, W.D. 1994 .The power and deceit of QTL experiments: lessons from comparative QTL studies, pp. 250–265 in Proceedings of the 49th Annual Corn and Sorghum Research Conference, edited by D. B. Wilkinson. American Seed Trade Association, Washington,DC.

Benjamini, Y., and Y. Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B. 57:289-300.

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

Bernardo, R. 2008. Molecular markers and selection for complex traits in plants: Learning from the last 20 years. Crop Sci. 48:1649.

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

Bradbury P.J., Z. Zhang, D.E. Kroon, T.M. Casstevens, Y. Ramdoss, and E.S. Buckler. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 2007;23:2633–5.

Butt M.S., M. Tahir-Nadeem, M.K.I. Khan, R. Shabir, and M.S. Butt. 2008. Oat: unique among the cereals. European Journal of Nutrition 47 : 68–79.

Carollo, V., D.E. Matthews, G. R. Lazo, T. K .Blake, D.D. Hummel, N.Lui, D. L. Hane, and O.D. Anderson. 2005. GrainGenes 2.0. An improved resource for the small-grains community. Plant Physiology 139: 643-651.

Cervantes-Martinez, C.T., K.J. Frey, P.J. White, D.M. Wesenberg, and J.B.Holland. 2001. Selection for greater β-glucan content in oat grain.Crop Sci. 41:1085–1091. doi:10.2135/cropsci2001.4141085x.

Cervantes-Martinez, C.T., K.J. Frey, P.J. White, D.M. Wesenberg, and J.B. Holland. 2002. Correlated responses for greater β-glucan content in two oat populations. Crop Sci 42:730-738.

Chernyshova A.A., P.J. White, M.P. Scott, and J-L Jannink .2007. Selection for nutritional function and agronomic performance in oat. Crop Sci. 47:2330-2339.

Colleoni-Sirghie, M., J.-L. Jannink, and P.J. White. 2004. Pasting and thermal properties of flours from oat lines with high and typical amounts of β-glucan. Cereal Chem. 81:686-692.

Crossa, J., G. de los Campos, P. Perez, D. Gianola, G. Atlin, J. Burgueno, J.L. Araus, D. Makumbi, R. Singh, S. Dreisigacker, J. Yan, V. Arief,M. Banziger, and H.J. Braun. 2010. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics doi:10.1534/genetics.110.118521.

De Koeyer, D.L., N.A. Tinker, C.P. Wight, J. Deyl , V.D. Burrows, L.S. O'Donoughue, A. Lybaert, S.J. Molnar, K.C. Armstrong, G. Fedak, D.M. Wesenberg, B.G. Rossnagel, and A.R. McElroy. 2004. A molecular linkage map with associated QTLsfrom a hulless covered spring oat population. Theor Appl Genet 108:1285–1298.

Falconer, D.S., and T.F.C. Mackay. 1996. Introduction to quantitative genetics. 4th ed. Longman Technical and Scientific, Essex, UK.

Fehr, W. 1987. Principles of cultivar development. Volume 1. New York: Macmillan, 1987.

Forsberg, R.A., and H.L. Shands. 1989. Oat breeding. In J. Janick (ed). Plant breeding reviews. Vol. 6, pp. 167 - 207. Portland, OR, USA, Timber Press.

Frey, K.J., J.K. McFerson, and C.V. Branson. 1988. A procedure for one cycle of recurrent selection per year with spring-sown small grains. Crop Science 28:855-856.

Goddard, M., and B. Hayes. 2007. Genomic selection. J. Anim.Breed. Genet. 124:323–330.

Hardy, O.J., and X. Vekemans. 2002. SPAGeDi: A versatile computer program to analyse spatial genetic structure at the individual or population levels. Mol. Ecol. Notes 2:618–620.

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

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

Heffner, E.L., J.L. Jannink, H. Iwata, E. Souza, and M.E. Sorrells. 2011b. Genomic selection accuracy for grain quality traits in biparental wheat populations. Crop Sci. 51:2597–2606.

Holthaus, J.F., J.B. Holland, P.J. White, and K.J. Frey. 1996. Inheritance of β-glucan content of oat grain. Crop Sci. 36:567–572.

Jannink, J.-L. 2010. Dynamics of long-term genomic selection. Genetics Selection Evolution 2010 42:35.

Kempthorne, O., and R.N. Curnow. 1961. The partial diallel cross. Biometrics 17: 229-250.

Kianian, S.F., R.L. Phillips, H.W. Rines, R.G. Fulcher, F.H. Webster, and D.D. Stuthman. 2000. Quantitative trait loci influencing B-glucan content in oat (*Avena sativa*, 2n=6x=42). Theoretical and Applied Genetics 101:1039-1048.

Kibite, S., and M.J. Edney. 1998. The inheritance of β-glucan concentration in three oat (*Avena sativa* L.) crosses. Can. J. Plant Sci. 78:245–250.

Lande, R., and R. Thompson. 1990. Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124:743–756.

Lorenz, A., S. Chao, F. Asoro, E. Heffner, T.Hayashi, H. Iwata, K. Smith, M. Sorrels, and J.-L. Jannink. 2011. Genomic selection in plant breeding: knowledge and prospects. In: D. L. Sparks (Ed.), Advances in Agronomy, Academic Press, San Diego, CA USA. pp. 77-123.

Lorenzana, R.E., and R. Bernardo. 2009. Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theor. Appl. Genet. 120:151-161.

Mayor, P.J., and R. Bernardo. 2009. Genomewide selection and marker-assisted recurrent election in doubled haploid versus F2 populations. Crop Sci. 49:1719–1725.

McMullen, M.S., D.C. Doehlert, and J.D. Miller. Registration of 'HiFi' Oat. Published in Crop Sci. 45:1664.

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

Nielsen, H.M., A.K. Sonesson, H.Yazdi, and T.H.E. Meuwissen. 2009. Comparison of accuracy of genome-wide and BLUP breeding value estimates in sib based aquaculture breeding schemes. Aquaculture 289: 259–264.

Peterson, D.M. 1991. Genotype and environment effects on oat glucan concentration. Crop Sci. 31:1517–1520.

Podlich, D.W., C.R.Winkler, and M. Cooper. 2004. Mapping as you go: An effective approach for marker-assisted selection of complex traits. Crop Sci. 44: 1560-1571.

SAS Institute. 2008. SAS/Stat user's guide. SAS Institute, Cary, NC.

Stich, B., J. Mohring, H.-P. Piepho, M. Heckenberger, E.S. Buckler, and A.E. Melchinger. 2008. Comparison of mixed-model approaches for association mapping. Genetics 178:1745–1754.

Tinker, N.A., and D.E Mather. 1993. KIN: Software for computing kinship coefficients. J Hered. 84:238.

Tinker, N.A., and J.K. Deyl. 2005. A curated internet database of oat pedigrees. Crop Science 45:2269-2272.

Tinker, N.A., A. Kilian, H.W. Rines, A. Bjornstad, C.J. Howarth, J. Jannink, J.M. Anderson, B.G. Rossnagel, C.P. Wight, D.D. Stuthman, M.E. Sorrells, G.J. Scoles, P.E. Eckstein, H.W. Ohm, E.W. Jackson, S. Tuvesson, F.L. Kolb, S.J. Molnar, O. Olsson, M.L. Carson, A. Ceplitis, J.M. Bonman, L. Federizzi, and T. Langdon. 2009. New DArT markers for oat provide enhanced map coverage and global germplasm characterization. Biomed Central (BMC) Genomics. 10(39):1471-2164.

Tiwari, U., and E. Cummins. 2009. Factors influencing β-glucan levels and molecular weight in cereal-based products. Cereal Chemistry 86:290–301.

Xu, Y., and J.H. Crouch. 2008. Marker-assisted selection in plant breeding:From publications to practice. Crop Sci. 48:391.

Zhang Z., E.S. Buckler, T.M. Casstevens, and P.J. Bradbury. 2009. Software engineering the mixed model for genome-wide association studies on large samples. Briefings in Bioinformatics 10:664-675.

Zhao, K., M.J. Aranzana, S. Kim, C. Lister, C. Shindo, C. Tang, C. Toomajian, H. Zheng, C. Dean, P. Marjoram, and M. Nordborg. 2007. An arabidopsis example of association mapping in structured samples. PLoS Genet 3:e4.

Zhong, S., and Jannink, J. 2007. Using QTL results to discriminate among crosses based on their progeny mean and variance. Genetics. 177:567-576.

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

List of Figures

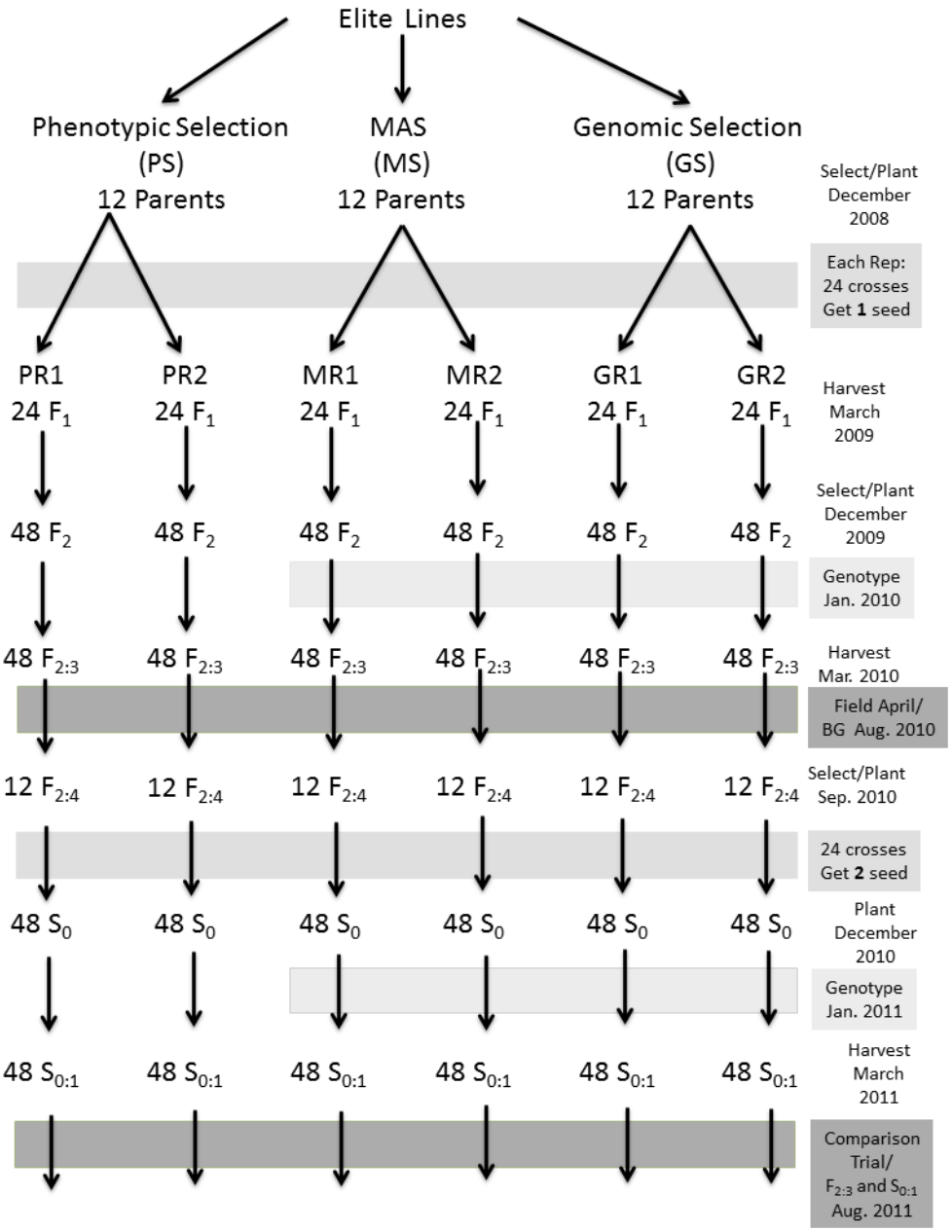


Figure 1. Timeline and selection scheme used in developing high β-glucan lines and for comparing selection methods. For populations, the first letter stands for selection method, the second and third means replication number. Specifically, PR1 and PR2 stand for first and second replication of BLUP phenotypic selection replication 1, MR1 and MR2 stand for first and second replication of marker-assisted selection while GR1 and GR2 stand for first and second replication of genomic selection, respectively.

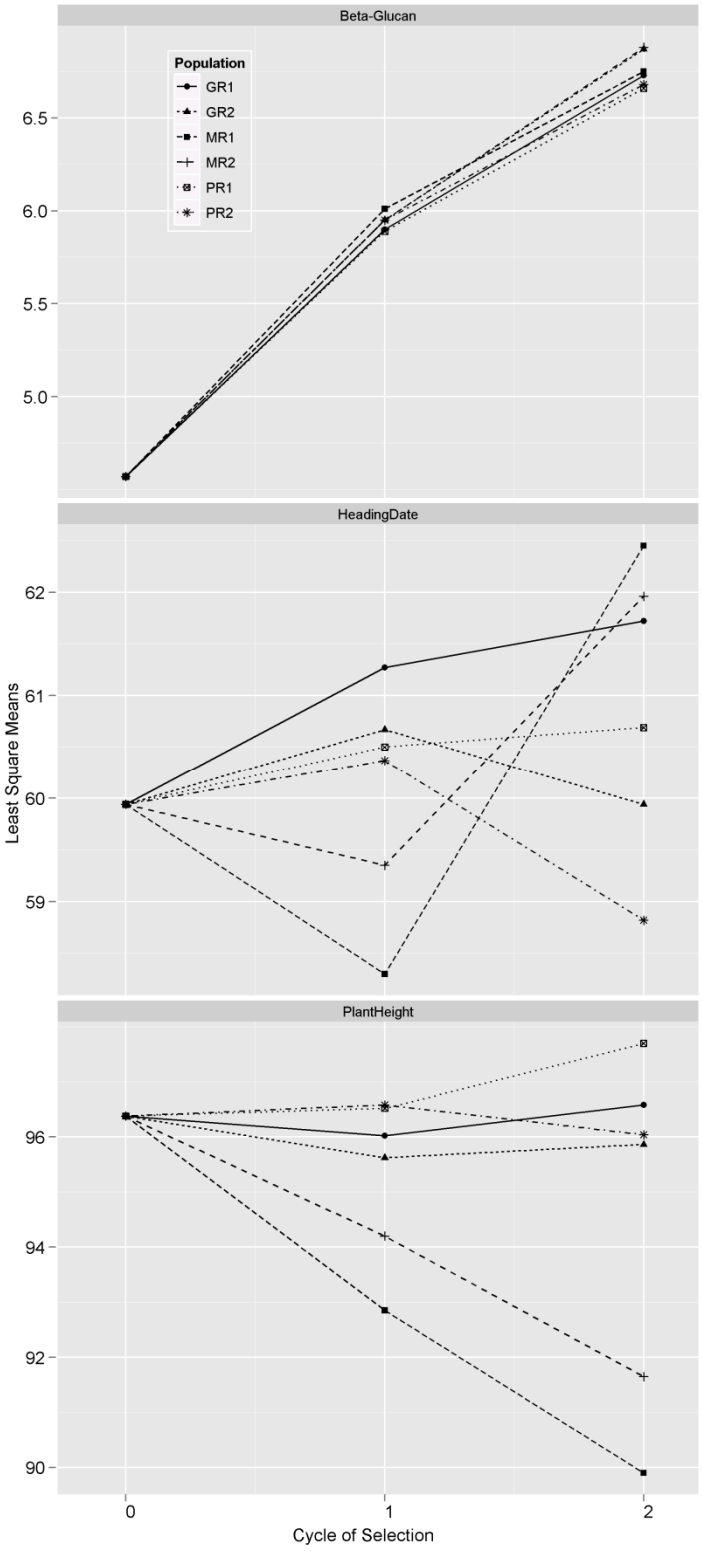


Figure 2. Least-square means of different populations per cycle of selection for three traits.

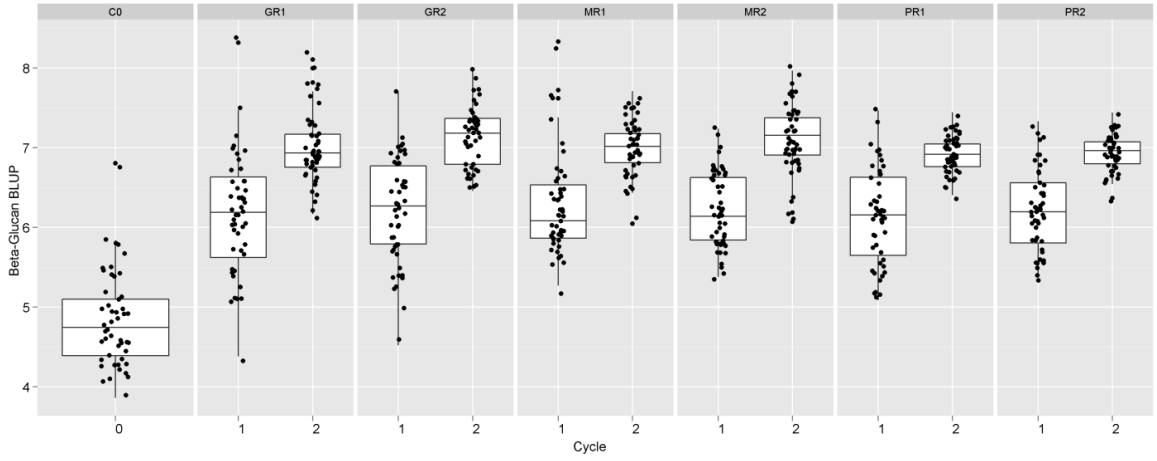


Figure 3. Boxplots representations of Best Linear Unbiased Prediction values (y-axis, computed as BLUP= grand mean + population mean + entry effect) for β-glucan content of Cycle 0 lines and progenies of Cycle 1 and Cycle 2. Each panel represents population name cycles are presented on the horizontal axis.

List of Tables

Table 1. Frequency of favourable alleles of selected markers for β-glucan content for two cycles of selection. Estimated genetic effects were computed from multiple regression models.

Marker[†]	Cycle 0		Cycle 1		Cycle 2	
	Effect	Allele frequency	Effect	Allele frequency	Effect	Allele frequency
oPt-11819	-0.39	0.06	-0.05	0.70	NA	NA
oPt-14067	0.42	0.08	0.24	0.45	0.59	0.96
oPt-18130	-0.30	0.07	0.04	0.30	NA	NA
oPt-18282	-0.23	0.45	0.13	0.70	-0.06	0.41
oPt-8249	0.44	0.02	-0.14	0.21	0.34	0.34
oPt-11728	-0.32	0.93	NA[‡]	NA	NA	NA
oPt-7232			-0.11	0.58	-0.39	0.61

[†] All marker effects were significant ($p<0.05$) in Cycle 0. The effect of oPt-14067 was significant in Cycle 1 while oPt-7232 was significant in Cycle 2.

[‡]NA means the marker was not included in the DArT genotyping report.

Table 2. Least square means for β-glucan and their standard errors for populations and parental lines.

Populations[†]	Cycle[‡]	Estimate	Standard Error
Base	0	4.57	0.12
GR1	1	5.90	0.13
GR2	1	5.95	0.12
MR1	1	6.01	0.11
MR2	1	5.95	0.10
PR1	1	5.89	0.12
PR2	1	5.95	0.11
GR1	2	6.73	0.11
GR2	2	6.87	0.11
MR1	2	6.75	0.10
MR2	2	6.88	0.11
PR1	2	6.66	0.09
PR2	2	6.68	0.08
Cycle 1 Parents[§]	1	5.72	0.17
GR1 Parents	2	6.75	0.21
GR2 Parents	2	6.82	0.22
MR1 Parents	2	6.56	0.19
MR2 Parents	2	6.81	0.21
PR1 Parents	2	6.83	0.17
PR2 Parents	2	6.27	0.17

[†] PR1 and PR2 stand for first and second replication of BLUP phenotypic selection replication 1, MR1 and MR2 stand for first and second replication of marker-assisted selection while GR1 and GR2 stand for first and second replication of genomic selection, respectively.

[‡] The test of contrast of population means showed that Cycle 0 is significantly different from each of the populations in Cycle 1 and Cycle 2. The Cycle 1 populations are also significant different from Cycle 2 populations ($p<0.0001$). Test of contrast between selection methods showed that GS and MS populations means are significantly different (p-value = 0.08) from PS population means.

[§]Parents are not significantly different from their respective progenies.

Table 3. Response to selection for β-glucan content and standard error of difference for each population.

Population	Cycle	Mean Response[†]	SE[‡]	% Over Previous Generation
GR1	1	1.33	0.18	29.1
GR2	1	1.38	0.17	30.2
MR1	1	1.44	0.16	31.5
MR2	1	1.38	0.16	30.2
PR1	1	1.32	0.17	28.9
PR2	1	1.38	0.16	30.2
GR1	2	0.83	0.17	14.1
GR2	2	0.92	0.17	15.5
MR1	2	0.74	0.15	12.3
MR2	2	0.93	0.15	15.6
PR1	2	0.77	0.15	13.1
PR2	2	0.72	0.14	12.1

[†] The mean responses were all significant ($p<0.0001$) from their respective previous populations. For example, the mean response of GR1 in Cycle 1 is the difference between Cycle 1 progenies and Cycle 0 lines. The mean response under GR1 in Cycle 2 is the mean difference between Cycle 2 and Cycle 1 progenies.

[‡] SE is the standard error of the difference between the two populations under consideration.

Table 4. Estimate of differences and their standard error between cycles of each selection method.

Selection Method	β-Glucan Content (%)	Heading Date (days)	Plant Height (cm)
	Cycle 1 - Cycle 0		
GS	1.35±0.15***	1.03±0.78	-0.56±1.41
MS	1.41±0.14***	-1.11±0.81	-2.86±1.51
PS	1.35±0.14***	0.50±0.80	0.17±1.43
	Cycle 2 - Cycle 1		
GS	0.87±0.12***	-0.14±0.71	0.41±0.95
MS	0.84±0.11***	3.38±0.63***	-2.75±1.31*
PS	0.75±0.10***	-0.68±0.72	0.32±0.93

Table 5. Top 20 lines from combined analysis of Cycle 1 and Cycle 2 data. The β-glucan content of each entry was computed as grandmean + population effect + line effect.

Line Code	Cycle	Pop	Pedigree	BG%
IA10203	Cy1	GR2	IAN979-5-1-22 /ND030288-1	9.06
IA10119	Cy1	MR1	IA95111 /AC Antoine -1	8.33
IA10078	Cy1	GR1	IA95258 /IAN979-5-1-22 -2	8.32
IA10194	Cy1	GR2	IA95111 /HiFi PI633006-2	8.30
IA11081	Cy2	GR1	IA95258 /ND030288-1-B //IA95258 /ND030288-2-B	8.11
IA11089	Cy2	GR1	IA95111 /AC Antoine -2-B //ND030288 /IA03146-6 -2-B	8.00
IA11228	Cy2	GR2	AC Antoine /MN95170 -2-B //AC Antoine /IA03146-6 -1-B	7.97
IA11147	Cy2	MR2	IA95111 /ND030288-1-B //IA91524-1-5-1 /ND030288-1-B	7.97
IA11164	Cy2	MR2	IL97-6202 /IA03146-6 -1-B //IL97-6202 /IAN979-5-1-22 -2-B	7.86
IA11214	Cy2	GR2	IAN979-5-1-22 /ND030288-1-B //IA03146-6 /IAN979-5-1-22 -2-B	7.85
IA11161	Cy2	MR2	IL97-6202 /IA03146-6 -1-B //IA95111 /IA91524-1-5-1 -1-B	7.83
IA11090	Cy2	GR1	IA95111 /AC Antoine -2-B //ND030288 /IA03146-6 -2-B	7.82
IA11151	Cy2	MR2	IA03146-6 /IAN979-5-1-22 -1-B //Reeves /IAN979-5-1-22 -2-B	7.79
IA11166	Cy2	MR2	IA95111 /IA91524-1-5-1 -1-B //IL97-6202 /IAN979-5-1-22 -2-B	7.78
IA11212	Cy2	GR2	IAN979-5-1-22 /ND030288-1-B //IAN979-5-1-22 /ND030288-2-B	7.77
IA11213	Cy2	GR2	IAN979-5-1-22 /ND030288-1-B //IA03146-6 /IAN979-5-1-22 -2-B	7.76
IA11211	Cy2	GR2	IAN979-5-1-22 /ND030288-1-B //IAN979-5-1-22 /ND030288-2-B	7.73
IA10120	Cy1	MR1	IA95111 /AC Antoine -2	7.72
IA11114	Cy2	MR1	IA95111 /MN95170 -1-B //ND030288 /IA03146-6 -1-B	7.71
IA10005	Cy1	PR1	IA03146-6 /IA95258 -1	7.71

Underlined linecodes are parents of Cycle 2.

Supplementary Table 1. Oat lines used in different selection methods and their pedigree information.

Parents in Cycle 1	Selection Method†			Pedigree
98P04-AT1	GS			Gem/AC_Medallion
AC Antoine	GS	MS	PS	Terra/Marion
Baker PI642412	GS	MS	PS	Blaze/Vista
Clinton CIav3971			PS	Richland/GreenRussian/2/Bond
HiFi PI633006	GS		PS	ND90141/ND900118
IA00020-12-3	GS		PS	Wabasha/IL94-784
IA03146-6	GS	MS	PS	WIX7509-5-1/IA95172-1-4-17
IA91524-1-5-1		MS		IL85-6183-1/2/Starter/3/IAB709-98-1/ Porter/2/Premier/5/IAD227-32-6/Ogle/2/ IL75-5681/3/Starter/4/MO07929
IA94190-10-1			PS	MN3_BGline/IAP307/2/MN4_BGline/IAP307/3/IA91530-1
IA95111	GS	MS		IA94134-2/IA94084-2
IA95258	GS		PS	IA94148-5/IA94031-1
IA97115-1			PS	IA91462-4-1-6/Brawn
IAN979-5-1-22	GS	MS	PS	MO07929/IL85-6183-1
IL00-7070		MS		IL95-4774/WIX6165-6
IL90-4950			PS	IL83-7646/ND810106
IL97-6202		MS		IL86-1956/IL91-7827
MN95170	GS	MS		MN86228/2/P7869D1/MN88231
ND030288	GS	MS	PS	HiFi/IAN979-5-1-22
Reeves		MS		SD87672/3/IL75-3402/2/Trucker/ ND810106/5/IAN111-5/3/Spear/Kelsey/2/Dumont/4/NO820-3
WIX8254-2	GS	MS		P8652A1-X-10-11-3/WIX6356-1

† GS means genomic selection, MS means marker-assisted selection, PS means phenotypic selection.

Supplementary Table 2. Oat lines submitted to National Small Grains Collection, Aberdeen, ID.

Oat Line Code	PI Code	Pedigree
IA10194	PI 664539	IA95111 (oat) /HiFi PI633006
IA10203	PI 664540	IAN979-5-1-22 (oat) /ND030288
IA10078	PI 664541	IA95258 (oat) /IAN979-5-1-22 (oat)
IA10119	PI 664542	IA95111 (oat) /AC Antoine (oat)
IA11081	PI 664543	IA95258 (oat) /ND030288@-1-B //IA95258 (oat) /ND030288@-2-B
IA11089	PI 664544	IA95111 (oat) /AC Antoine (oat)@-2-B //ND030288 /IA03146-6 (oat)@-2-B
IA11147	PI 664545	IA95111 (oat) /ND030288@-1-B //IA91524-1-5-1 (oat) /ND030288@-1-B
IA11164	PI 664546	IL97-6202 (oat) /IA03146-6 (oat)@-1-B //IL97-6202 (oat) /IAN979-5-1-22 (oat)@-2-B
IA11214	PI 664547	IAN979-5-1-22 (oat) /ND030288@-1-B //IA03146-6 (oat) /IAN979-5-1-22 (oat)@-2-B
IA11228	PI 664548	AC Antoine (oat) /MN95170 (oat)@-2-B //AC Antoine (oat) /IA03146-6 (oat)@-1-B

Supplementary Method 1. Removing redundant markers in Cycle 0

First, markers were merged if they had less than 1% difference in marker scores (Tinker et al., 2009). From 1295 markers, 848 merged markers were defined. Second, merged markers were split if they contained markers that belonged to different contigs in the sequence assembly described by Tinker et al. (2009). Eleven merged markers were split in this way, resulting in 859 bins. Finally, merged markers were split if they contained markers that mapped more than 2 cM apart from each other based on the Kanota x Ogle map (Tinker et al., 2009). Seven merged markers were split in this way, resulting in 866 bins. For each merged marker, a consensus score was calculated as the most common allele among the markers.

Supplementary Method 2. Field Plot Design and Data Analysis for Cycle 1

The experimental design was an incomplete block with two replications. The entries were randomly selected from 288 F2:3 lines from six populations developed from the greenhouse, 48 Cycle 0 plants, 24 unique parents of all breeding programs, and four checks. The checks included IAN9N79-5-1-22, Baker, IA002130-2-2, and Excel. Each of the two replications was composed of six sets. For each set, the entries consisted of a random group of eight lines from each population (48 F2:3), a random group of eight lines from Cycle 0, a group of four random parents, and all four checks. Each set was planted in one incomplete block composed of an 8 by 8 grid of hillplots. The experiment was grown at the Agronomy and Agricultural Engineering Field Research Center near Ames, IA from April to July 2010. Each hill plot was then harvested by hand and threshed after one-week of air-drying.

Oat samples from each hillplot were dehulled using an air-pressure dehuller (Codema, Eden Prairie, MN) to recover the groats. The groat samples were powdered using a homogenizer (Talboys HT Homogenizer, Troemner, Thorofare, NJ). Then β-glucan content was assayed for each sample using an enzymatic method adapted for microplates as described in detail by Newell et al. (in review). The following model was used:

β-glucan= replication + incomplete block (replication) + population + oat entries + error

where oat entries and error terms were considered random effects and the remaining terms as fixed effects. The effects for oat entries were sorted per population. Then the effect of each oat entry was treated as the new response variable for selection of parents for Cycle 2.

Supplementary Method 3. Recombination Scheme for Cycle 2

In September of 2010, the 12 parents of each breeding program were randomly assigned entry numbers in their respective crossing block. Seeds were planted in the greenhouse and a partial diallel was implemented where each parent was crossed to four other parents, thus 24 crosses were completed. Two seeds from each cross were selected randomly at maturity to form 48 S0 lines for each population. The 48 S0 seeds of each population were planted in January of 2011 for advancement from S0 to S1. Leaf samples for DNA extraction was conducted for MS and GS S0 plants and processed as mentioned. Lastly, the S0:1 seeds were harvested at maturity for each plant and used in a field trial.

CHAPTER 5. SELECTION METHODS FOR β-GLUCAN CONTENT IN ELITE OAT GERMPLASM: CHANGE IN GENETIC VARIANCE

A paper to be submitted to Crop Science

Abstract

One determinant of response to selection is genetic variance for the trait of interest. Empirical data on the changes in genetic variance under existing and new selection strategies like BLUP phenotypic selection, marker assisted selection and genomic selection are not yet available. In this study we (i) assessed the genetic variance for β-glucan content that exists in a collection of elite oat lines from North America and (ii) evaluated the changes in genetic variance under the different selection strategies in an actual breeding program. The results showed that the estimated genetic variance for β-glucan content was mostly composed of additive effects, as explained by the pedigree-based relationship or marker-based relationship data. The estimated genetic variances decreased after two cycles of selection, although the magnitude of reduction was different for the three selection strategies. The highest reduction in genetic variance was obtained for BLUP phenotypic selection, which may be explained by the greater chance of co-selection of sibs with pedigree-based BLUP. Both MAS and GS maintained genetic variance but the latter had lower coancestry among progenies. Overall, we found that marker-based selection methods maintained greater genetic variance than did the BLUP phenotypic selection, potentially assuring greater future selection gains.

Introduction

In recent decades there has been considerable research to develop alternative methods for selection to accelerate the development of cultivars and improve breeding populations. Fundamental to those efforts is the quest for methods that can improve the response to selection or increase accuracy of estimates of breeding values (Falconer and Mackay, 1996). One of the earliest results of those efforts was the use of best linear unbiased prediction (BLUP) methodology for estimating breeding values with additional information from pedigree data or coancestry (Henderson, 1984). Earlier applications of BLUP in plant breeding have shown that it can result in higher percentages of superior crosses in soybean relative to traditional mid-parent value (Panter and Allen, 1995) and it can predict performance of untested single crosses of maize using relationship data from relatives (Bernardo, 1996). However, pedigree-based relationships do not account for random Mendelian segregation within families. This segregation is important because, when using inbred lines, as is common in plant breeding, a given progeny will not necessarily contain an equal contribution from its two parents.

Another alternative method is marker-assisted selection (MAS) or marker-assisted recurrent selection (MARS), which is based on identification of significant markers for QTL controlling the trait (Lande and Thompson, 1990). The significant markers for QTL can be identified in biparental linkage mapping studies or from genomewide association mapping studies (GWAS, Yu et al., 2006). However, the process of significance testing conducted for markers can lead to the “Beavis Effect,” that is the overestimation of marker effects (Beavis, 1994). Moreover, the use of the significant markers only accounts for the breeding values due to those few QTL, resulting in a smaller proportion of genetic variance explained (Hayes and

Goddard, 2010). Because of these limitations, traditional MAS for polygenic traits may not be appropriate (Bernardo, 2008). Recently, genomic selection (genomewide selection or genomewide prediction) has been proposed as an alternative to MAS methods (Meuwissen et al., 2001; Heffner et al, 2009). Genomic selection (GS) uses all markers distributed across the genome to predict estimated breeding values of individuals. This method captures more of the genetic variance, leading to more accurate estimated breeding values (Hayes and Goddard, 2010). Furthermore, the use of all markers traces the Mendelian segregation for each QTL and prediction of breeding values within families is feasible (Daetwyler et al., 2007).

Numerous simulations and a few empirical studies have shown that GS has higher accuracy of selection than MAS and BLUP-based phenotypic selection and it has the ability to accelerate genetic gain (Meuwissen et al., 2001; Dekkers et al., 2007; Lorenzana and Bernardo, 2009; Jannink, 2010). However, there are few studies about the impact of these selection methods on maintaining genetic variation (Daetwyler et al., 2007). Change in genetic variance that is attributed to selection could either be caused by factors such as changes in allele frequencies, level of inbreeding or coancestry, negative LD or the Bulmer effect (Robertson, 1960; Hill and Robertson, 1966; Bulmer, 1971; Sorensen and Kennedy, 1984; Falconer and Mackay, 1996). In a simulation study by Bastiaansen et al. (2012), the genetic variance for a polygenic trait decreased after short-term selection but had similar magnitude for GS and BLUP phenotypic selection. In addition, both Bastiaansen et al. (2012) and Daetwyler et al. (2007) have shown that there was higher inbreeding of animals undergoing phenotypic BLUP selection than those animals under GS. These studies suggest

that higher coancestry of selected lines can affect the genetic variance of succeeding cycles of selections.

Studies on the differences of these selection methods on a short-term basis in an actual plant breeding program have not been reported. Inferences from those comparisons are important because response to selection is dependent on the level of genetic variance for any trait (Falconer and Mackay, 1996). Recently, we conducted a small scale selection program for β-glucan content in oat using BLUP phenotypic selection (PS), GS and MAS. β-glucan has been identified as the active component of soluble fiber in oat that lowers blood serum cholesterol levels – a major risk for heart disease (reviewed by Butt et al., 2008). Previous research on β-glucan content indicated that this trait is controlled by many genes acting in an additive manner (Holthaus et al., 1996; Chernyshova et al, 2007; Cervantes-Martinez et al., 2001). Our specific objectives therefore were to (i) examine estimated genetic variances and heritability for β-glucan content in a collection of elite oat germplasm in North America; (ii) assess short-term changes in genetic variance for β-glucan content for the three selection methods and (iii) examine changes in the magnitude of coancestry among progenies within the different selection methods.

Materials and Methods

Cycle 0 Genetic Parameters

The Cycle 0 population in this study was composed of 446 oat lines from various oat breeding institutions in the USA and Canada which were evaluated in cooperative performance nurseries from 1994 to 2007. The phenotypic data including β-glucan content from evaluation trials are stored in the Graingenes database (Carollo et al., 2005) and herein

referred to as OPN data (oat performance nurseries). This type of data is highly unbalanced because not all lines are tested at the same sites and every year some entries are dropped and new entries are added. To generate balanced data for β-glucan content, oat lines and check cultivars were evaluated in Ames in the summer of 2009 and 2010. An incomplete block design with two replications was used in each year. Each incomplete block, nested within replications was arranged in 5 by 5 grids of hill plots. At maturity, heads were harvested manually and threshed after one of week of air drying. Samples of seeds from each hill plot was dehulled and ground for a β-glucan content assay (Newell et al., in review) and this data is referred to as Ames data.

The Cycle 0 lines were genotyped in Spring of 2008 and the resulting genotype matrix composed of 446 lines and 866 markers (M) was used to derive a marker-based relationship matrix equal to $MM'/\sum_k p_k(1-p_k)$ where p_k is the frequency of allele 1 in marker k computed using the Spagedi program (Hardy and Vekemans, 2002). A pedigree-based relationship among lines was also estimated using the kinship software KIN (Tinker and Mather, 1993), which is based on the probability that alleles at a locus are identical by descent (Malecot, 1948). The pedigree records were taken from the Pedigree of Oat Lines Database (POOL; Tinker and Deyl, 2005). The values in the coancestry matrix ranged from 0 to 1, where 0 refers to two lines unrelated through pedigree and 1 is the diagonal representing a perfect identity by descent. To compare the similarity of marker-based and pedigree-based relationships of lines in Cycle 0, a Mantel Test (Sokal and Rohlf, 1995) was conducted between the two relationship matrices.

To investigate genetic variance parameters in Cycle 0, a mixed model was implemented using PROC Mixed in SAS:

$$y = Eb + Zg + error$$

where *y* is the β-glucan content observation, *Eb* term is the fixed environmental effects and *Zu* term is the oat line effects, assumed to be random effects based on the sample from all possible North American oat lines. For the OPN data set, *y* consisted of 2909 observations from 129 location-year environments, while the *Eb* term is composed only of environmental effects. For the Ames data set, *y* consisted of 1950 β-glucan content measurements generated from 2009 to 2010, while the *Eb* term is composed of *grand mean + year + rep (year) + incomplete block (rep*year)* terms. For both data sets, E and Z are the incidence matrices relating *b* and *g* to *y*, respectively. The *b* is the fixed effects while *g* is the random polygenic effects of the oat entries. The variance of $g = 2KV_A$, where K is the marker-based relationship matrix or the pedigree-based relationship, and V_A is the additive variance estimated from the Cycle 0 data, 2K defines the additive relationship matrix and allowed us to estimate the additive genetic variance. On the other hand, treating the oat lines as unrelated to each other provides an estimate of only the genetic variance. Broad sense heritability was estimated as the ratio of *genetic variance / (genetic variance + residual variance),* the narrow sense heritability was estimated as the *additive genetic variance / (additive genetic variance + residual variance).*

Implementation of GS, MAS and BLUP-PS

The selection programs began in 2008 as discussed in detail in Asoro et al. (in review). Each selection program was replicated twice. Briefly, each selection method started

from the same Cycle 0 data using phenotypic data from oat performance nurseries (OPN) and marker data for 866 DArT markers. The first cycle of selection for parents under GS was conducted by using marker-based relationships among lines to estimate breeding values. Selection of parents for the second cycle of GS was conducted by estimating the marker effects jointly and taking the sum of effects of all markers for each individual to estimate the breeding values. For MAS, the important markers were first identified using association mapping and false discovery rates of 0.33. Then for each cycle of MAS, an index developed from marker scores from the effects of six β-glucan content markers and phenotypic values were applied to define the estimated breeding values of individuals. For selection of parents under BLUP phenotypic selection (PS), both cycles used pedigree-based coancestry as a covariance matrix to compute the estimated breeding values.

The random lines from Cycle 0 and the progenies in Cycle 1 of all populations were evaluated in the field from April to July 2010 (see Asoro et al. for details). The progenies of Cycle 2 together with random samples from both Cycle 1 and Cycle 0 were evaluated in the field from April 2011 to July 2011. For both years, samples from the harvested seeds were dehulled and the ground samples were used in β-glucan content assay using an enzymatic method designed for 96-well plates (Newell et al, in review).

Field Evaluation and Statistical Analysis for Comparison of Selection Methods

The field plot design was an incomplete block design with two replications. The entries were comprised of: a random sample of 24 lines from each population of Cycle1 (24 x 6 populations=144 lines), the 48 random lines from Cycle 0, the 20 unique parents of Cycle 1 plus four popular oat cultivars (total of 24), 288 S0:1 lines from Cycle 2, and five checks

(IAN9N79-5-1-22, Baker, IA002130-2-2, Excel, and CDC Pro-Fi). For each incomplete block, the entries consisted of a random sample of six lines from Cycle 0, a random sample of three lines from each population of Cycle 1 (total of 24), a random sample of three lines from the parents of Cycle 1, six lines from each population of Cycle 2 (total of 36), and all of five checks. Each incomplete block was planted in a 7 by 11 grid of hillplots. The lines were grown at the Iowa State University Agronomy and Agricultural Engineering Field Research Center near Ames, IA from April to July 2011. Over-all, Cycle 0 and Cycle 1 populations were evaluated in two years while Cycle 2 was evaluated only for one year.

Combined data from 2010 and 2011 were analyzed using PROC Mixed in SAS with following model:

$$y = Xb + Zu + error$$

where y was the β-glucan content of entries in the field evaluations, b is the vector of effects for following fixed terms: *grand mean + year + replication (year) + incomplete block (replication*year) + population* while X is the design matrix relating b to y and u is the random term for entry effects while Z is the incidence matrix relating u to y. The variance of $u = I\sigma^2_g$, where I is the identity matrix and σ^2_g is the genetic variance estimated from the data.

Significance Test for Differences in Genetic Variances

For significance tests of the difference of genetic variances, the likelihood ratio test was used, which assumes that the difference between the –2 REML log-likelihood of the full model and the reduced model has a chi-square distribution (Saxton, 2004). Then the p-value associated with the chi-square distribution was reported, with degrees of freedom defined by the difference in parameters of models under comparison. To systematically compare the

genetic variances, indicator variables were developed for different groups of entries, including different populations (Table 2). For example, to test the hypothesis that variance of the GS method is the same as the variance of the PS method under Cycle 2, an analysis was first conducted assuming heterogeneous variance for groups of entries: Cycle 0, GS Cycle 1, MS Cycle 1, PS Cycle 1, GS Cycle 2, MS Cycle 2, PS Cycle 2 and checks. A subsequent analysis was conducted where GS Cycle 2 and PS Cycle 2 were in the same group. The difference in -2 REML Log Likelihood values for the two analyses was calculated and the p-value associated was taken with one degree of freedom because there were eight parameters in the first test and seven in the second test.

Coancestry in Cycle 1 and Cycle 2 Progenies

Because of incomplete pedigree information among lines in Cycle 0, the marker-based relationship among lines in Cycle 0 was used as the reference for assessing coancestry of succeeding cycles of progenies. Subsequently, pedigree records of Cycle 1 and Cycle 2 were used to derive the coancestry of the progenies using the tabular method (Bernardo, 2010). The average coancestry among progenies within population of each cycle was estimated as the average among all pairs of lines, where each pairwise value is equivalent to an inbreeding coefficient of hypothetical progenies derived from each pair.

Results

Estimates of Genetic Parameters in Cycle 0

The Mantel Test for relatedness between marker-based and pedigree-based relationship of lines from Cycle 0 showed that two the matrices (Figure 1) were positively

correlated ($r = 0.43$, $p < 0.0001$). The regression coefficient of marker-based on pedigree-based relationship was 0.45, with an intercept of 0.20.

Estimated genetic and additive genetic variances for β-glucan content were lower in the OPN than in the Ames data set across different relationship matrices (Table 1). On the other hand, the residual variances were consistent (0.25-0.27) across covariance matrices and data sets. These resulted in lower broad and narrow sense heritability for OPN than Ames data sets. The narrow sense heritability values estimated using markers and pedigree data were comparable within each data set. Fit statistics based on AIC (Akaike Information Criterion) showed that analysis with the covariance matrix based on marker data provided a better fit to the data across data sets, followed by analysis using pedigree data, and lastly the analysis assuming independent and identical distribution of observation (Table 1).

Individual Genetic Variances of Populations in Selection Experiments

All subsequent analyses were based on the evaluation of progenies from selection experiments conducted from 2010 to 2011 (Figure 1 in Asoro et al. in review). Individual comparisons of populations showed that genetic variances of populations in Cycle 1 ranged from 0.38 (PR2-Cy1) to 0.66 (GR1-Cy1) and from 0.12 (PR2-Cy2) to 0.37(GR2-Cy2) in Cycle 2 (Table 2; Figure 2). Estimated genetic variances of populations from Cycle 0 to Cycle 2 were all significantly greater than zero (Table 2). The reduction in genetic variance for every population from Cycle 1 to Cycle 2 ranged from 0.03 to 0.34, which corresponded to reductions by 9% (MR2) to 70% (PR1).

Comparison of Genetic Variances across Selection Methods

Estimated genetic variances between Cycle 0 and Cycle 1 within each selection method indicated a non-significant reduction (cut-off p=0.01, Table 3). A significant reduction was detected only between Cycle 1 and 2 of the PS populations (cut-off p=0.01). The genetic variances within Cycle 1 were not significantly different between selection methods. For Cycle 2, only the variance between GS and PS populations were significantly different from one another (GS > PS, cut-off p=0.01).

Average Coancestry

The average coancestry among progenies in the GS Cycle 1 was 0.41 (GR1=0.41 and GR2=0.41) and increased to 0.48 (GR1=0.49 and GR2=0.46) in Cycle 2 (Figure 3). Similarly, the average coancestry in MS Cycle 1 was also 0.41 (MR1=0.41 and MR2=0.41) and increased to 0.49 (MR1=0.49 and MR2=0.48) in Cycle 2. On the other hand, the average coancestry for PS Cycle 1 was 0.43 (PR1=0.43 and PR2=0.43) and increased to 0.50 (PR1=0.49 and PR2=0.52) in Cycle 2.

Discussion

Comparison of Pedigree-based and Marker-based Relationships

We have employed marker-based relationships and pedigree-based relationships in the implementation of Cycle 1 for GS and PS, respectively. Although the Mantel Test showed that the two types of relationship were related, the marker-based relationship varied substantially around pedigree-based relationship, indicating that the former detected deviations from expected relationships through pedigree. The same was also observed by

Wolc et al. (2011) on the comparison between pedigree-based and marker-based relationship of layer chickens. As an example of deviations of marker-based relationships from pedigree-based relationships in our data, there were 15 pedigree-based relationship values that were exactly one (rightmost side of Figure 1) but those values varied from 0.31 to 0.81 according to marker-based relationships. Examination of pedigree records showed that these 15 pairwise values came from lines with the same parents.

Estimates of Genetic Parameters for β-Glucan Content in Cycle 0

Baseline information for estimates of genetic variance components of a trait is important to determine changes and for designing breeding strategies. In this study, the genetic variance and broad sense heritability detected for β-glucan content were comparable to previous studies (Holthaus et al., 1996; Humphreys and Mather, 1996). However, the small estimated genetic variance detected in the OPN data set may represent broad adaptation variance, while the estimated variances in the Ames data set may represent only the narrow adaptation variance. Moreover, the genetic variance for the complete set of Cycle 0 that was planted in Ames was comparable to the genetic variance of a random sample of Cycle 0 that was planted in the selection experiments in this study. This indicates that the baseline genetic variance information for Cycle 0 that was used for further analysis in the selection study was reliable.

The analysis of Cycle 0 lines with polygenic effects modelled by marker-based relationship (RA-BLUP, Zhong et al., 2009) yielded a better fit than using only pedigree based relationship (Piepho et al., 2008; Bauer et al., 2008). This means that the marker data was able to detect differences among lines due to both pedigree and information arising from

Mendelian sampling (Daetwyler et al., 2007). Therefore, the marker data provided a realized set of relationships (Zhong et al., 2009) that better reflected the differences in breeding values for β-glucan content in Cycle 0. Our result is in agreement with previous comparisons of marker-based and pedigree based relationships, for example, Wolc et al. (2011) in layer chickens and by Nielsen et al. (2009) in aquaculture-based populations.

The results from using marker or pedigree-based relationships also showed that estimates of additive genetic variance comprises most of the genetic variance for β-glucan content. This study agrees with the selection study results conducted by Cervantes-Martinez et al. (2001) and by Chernyshova et al (2007) and lend support to the suggestion of Hill et al. (2008) that genetic variance for polygenic traits are due primarily to additive effects. The results from this study further indicated that assuming additivity of β-glucan QTL effects to define the total genetic values of individuals is appropriate. In other words, genomic selection models that ignore interactions among markers should be sufficient in implementing GS for β-glucan content.

Estimated Changes in Genetic Variance

Response to selection is dependent on the genetic variance of the selected parents (Falconer and Mackay, 1996). In our study, all selection methods resulted in a reduction of variance for β-glucan content from Cycle 0 to Cycle 2. This is in agreement with Cervantes-Martinez et al. (2001). Given that we conducted only two cycles of selection, and assuming that β-glucan content is controlled by polygenic effects, changes in allele frequencies of all QTL could be too small to alter genetic variance (Falconer and Mackay, 1996, p.201). Another reason for a reduction in genetic variance from Cycle 1 to Cycle 2 may be the

"Bulmer Effect," where selection results in negative LD between genes controlling the trait (Bulmer, 1971). In other words, because Cycle 1 is a product of selection in Cycle 0, selection might have caused β-glucan QTL to create repulsion-phase LD among QTL. In turn, this led to a lower variance for Cycle 2. Although recombination is known to breakdown LD, the limited diallel crossing conducted among parents in our study probably had little effect on LD. Therefore, the "Bulmer Effect" may still play a role in the reduction of variance (van der Werf and de Boer, 1990).

Genetic Variance and Coancestry

Although all methods reduced genetic variace, the magnitude of decrease was not the same across the three selection methods. Greater coancestry among lines and therefore inbreeding among individuals can contribute to reduction in genetic variance (Sorensen and Kennedy, 1984). In our study, the magnitude of decrease of genetic variance might be explained by the differences of buildup of coancestry of the various methods. For instance, the higher coancestry of progenies detected in PS populations could be explained by the fact that BLUP -based PS can increase the chance of co-selection of sibs as parents (Sonesson et al., 2005). In our case, the coselection of sibs with similar breeding values for β-glucan content could have eventually led to lower genetic variance for β-glucan content in Cycle 2 of PS (Supplementary Figure 1). This can lead to fixation of alleles for β-glucan content, resulting in minimal long term gains from selection. A higher probability of co-selection of sibs happens in the BLUP PS method (Henderson, 1975) because pedigree information does not account for segregation terms, resulting in a higher correlation of estimated breeding values within families (Daetwyler et al. et al., 2007). On the other hand, GS can account for

Mendelian segregation, which can lead to a reduced correlation of estimated breeding values within families (Daetwyler et al. et al., 2007). Therefore, the use of markers can reduce the build-up of related selected individuals in a breeding program, resulting in a less reduction of genetic variance. The similar level of coancestry between MAS and GS progenies in Cycle 1 might be explained by the fact that both methods used marker-based relationships in the final selection of diverse parents.

Breeding Implications

The recurrent selection implemented in this study is similar to advanced cycle breeding conducted in cultivar development breeding programs for maize, wheat and barley (Yu and Benardo, 2004). In advanced cycle breeding, the best performing lines are continually used to produce the next generation, resulting to progenies belonging to the same genetic background. It could be expected that recent generations will have a higher proportion of fixed alleles and can have lower genetic variance (Yu and Bernardo, 2004). However, the use of molecular markers can delay fixation of those alleles, as in the case of GS.

Although GS can provide a rapid increase in genetic gain for β-glucan content by factors such as multiple cycles per year (Asoro et al., 2011) and greater selection intensity, these factors may also lead to faster rate of loss of genetic diversity. The loss in genetic diversity can eventually lead to decreased genetic variance for the trait of interest and lower gain from selection (Robertson, 1960). In this case, a strategy which introgresses unrelated germplasm could be employed jointly with GS (Odegard et al. 2009; Bernardo, 2009). Another strategy would be a selection criterion that weights the low-frequency favorable

alleles more heavily to avoid losing them. This approach could sustain gains from selection and limit the loss of genetic variance (Jannink, 2010). Therefore, implementing GS in large breeding programs will require strategies that will balance rapid genetic gain and preserve genetic variation in elite breeding populations.

References

Bastiaansen, J.W.M., A. Coster, M.P.L. Calus, J.A.M. van Arendonk, and H. Bovenhuis. 2012. Long-term response to genomic selection: effects of estimation method and reference population structure for different genetic architectures. Genetics Selection Evolution 2012, 44:3 doi:10.1186/1297-9686-44-3.

Bauer, A.M., T.C. Reetz, and J. Léon. 2006. Estimation of breeding values of inbred lines using best linear unbiased prediction (BLUP) and genetic similarities. Crop Sci. 46:2685–2691.

Beavis, W.D. 1994 .The power and deceit of QTL experiments: lessons from comparative QTL studies, pp. 250–265 in Proceedings of the 49th Annual Corn and Sorghum Research Conference, edited by D. B. Wilkinson. American Seed Trade Association, Washington,DC.

Bernardo, R. 1996. Best linear unbiased prediction of the performance of crosses between untested maize inbreds. Crop Sci. 36:872–876.

Bernardo, R. 2008. Molecular markers and selection for complex traits in plants: Learning from the last 20 years. Crop Sci. 48:1649.

Bernardo, R. 2009. Genomewide selection for rapid introgression of exotic germplasm in maize. Crop. Sci 49:419- 425.

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

Bulmer, M.G. 1971. The effect of selection on genetic variability. Am. Nat. 105:201–211.

Butt M.S., M. Tahir-Nadeem, M.K.I. Khan, R. Shabir, and M.S. Butt . 2008. Oat: unique among the cereals. European Journal of Nutrition 47 : 68–79.

Carollo, V., D.E. Matthews, G.R. Lazo, T.K .Blake, D.D. Hummel, N.Lui, D. L. Hane, and O.D. Anderson. 2005. GrainGenes 2.0. An improved resource for the small-grains community. Plant Physiology 139: 643-651.

Cervantes-Martinez, C.T., K.J. Frey, P.J. White, D.M. Wesenberg, and J.B.Holland. 2001. Selection for greater β-glucan content in oat grain.Crop Sci. 41:1085–1091. doi:10.2135/cropsci2001.4141085x.

Chernyshova A.A., P.J. White, M.P. Scott, and J-L Jannink .2007. Selection for nutritional function and agronomic performance in oat. Crop Sci. 47:2330-2339.

Daetwyler, H.D., B. Villanueva, P. Bijma, and J.A. Woolliams. 2007. Inbreeding in genome-wide selection. J. Anim. Breed. Genet. 124, 369-376.

Dekkers, J.C.M. 2007. Prediction of response to marker-assisted and genomic selection using selection index theory. J Anim Breed Genet 124:331–341.

Falconer, D.S., and T.F.C. Mackay. 1996. Introduction to quantitative genetics. 4th ed. Longman Technical and Scientific, Essex, UK.

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

Hardy, O.J., and X. Vekemans. 2002. SPAGeDi: A versatile computer program to analyse spatial genetic structure at the individual or population levels. Mol. Ecol. Notes 2:618–620.

Hayes, B., and M. Goddard. 2010. Genome-wide association and genomic selection in animal breeding. Genome 53(11):876-83.

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

Henderson, C.R. 1975. Best linear unbiased estimation and prediction under a selection model. Biometrics 31: 423–447.

Henderson C.R. 1984. Applications of linear models in animal breeding. Univ. Guelph, Guelph, Ontario, Canada.

Hill, W.G., and A. Robertson. 1966. The effect of linkage on limits to artificial selection. Genet Res 8: 269–294.

Hill W.G., M.E. Goddard, and P.M. Visscher. 2008. Data and theory point to mainly additive genetic variance for complex traits. PLoS Genet 4(2): e1000008. doi:10.1371/journal.pgen.1000008.

Holthaus, J.F., J.B. Holland, P.J. White, and K.J. Frey. 1996. Inheritance of β-glucan content of oat grain. Crop Sci. 36:567–572.

Humphreys, D.G., and D.E. Mather. 1996. Heritability of β-glucan, groat-percentage, and crown rust resistance in two oat crosses. Euphytica 91:359–364.

Jannink, J.-L. 2010. Dynamics of long-term genomic selection. Genetics Selection Evolution 2010 42:35.

Lande, R., and R. Thompson. 1990. Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124:743–756.

Lorenzana, R.E., and R. Bernardo. 2009. Accuracy of genotypic value predictionsfor marker-based selection in biparental plant populations. Theor.Appl. Genet. 120:151–161. doi:10.1007/s00122-009-1166-3

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

Nielsen, H.M., A.K. Sonesson, H.Yazdi, and T.H.E. Meuwissen. 2009. Comparison of accuracy of genome-wide and BLUP breeding value estimates in sib based aquaculture breeding schemes. Aquaculture 289: 259–264.

Odegard J., M.H. Yazdi, A.K. Sonesson, and T.H.E. Meuwissen. 2009. Incorporating desirable genetic characteristics from an inferior into a superior population using genomic selection. Genetics 181:737-45.

Panter D.M., and F.L. Allen. 1995. Using best linear unbiased predictions to enhance breeding for yield in soybean. 1. Choosing parents. Crop Sci. 35:397–405.

Piepho, H.P., J. Mohring, A.E. Melchinger, and A. Buchse. 2008. BLUP for phenotypic selection in plant breeding and variety testing. Euphytica 161:209–228.

Robertson, A.1960. A theory of limits in artificial selection. Proc R Soc Lond B Biol Sci 153: 234–249.

SAS Institute. 2008. SAS/Stat user's guide. SAS Institute, Cary, NC.

Saxton, A.M. 2004. Genetic analysis of complex traits using SAS. SAS Institute, Inc. Cary, N.C.

Sokal, R.R., and F.J. Rohlf. 1995. Biometry, 3rd edn. New York: Freeman.

Sonesson, A.K., B. Gjerde, and T.H.E. Meuwissen. 2005. Truncation selection for BLUP-EBV and phenotypic values in fish breeding schemes. Aquaculture 243:61–68.

Sorensen, D.A., and B.W. Kennedy. 1984. Estimation of genetic variance from unselected and selected populations. J. Anim. Sci. 59, 1213-1221.

Tinker, N.A., and D.E Mather. 1993. KIN: Software for computing kinship coefficients. J Hered. 84:238.

Tinker, N.A., and J.K. Deyl. 2005. A curated internet database of oat pedigrees. Crop Science 45:2269-2272.

Tinker, N.A., A. Kilian, H.W. Rines, A. Bjornstad, C.J. Howarth, J. Jannink, J.M. Anderson, B.G. Rossnagel, C.P. Wight, D.D. Stuthman, M.E. Sorrells, G.J. Scoles, P.E. Eckstein, H.W. Ohm, E.W. Jackson, S. Tuvesson, F.L. Kolb, S.J. Molnar, O. Olsson, M.L. Carson, A. Ceplitis, J.M. Bonman, L. Federizzi, and T. Langdon. 2009. New DArT markers for oat provide enhanced map coverage and global germplasm characterization. Biomed Central (BMC) Genomics. 10(39):1471-2164.

van der Werf, J.H., and I.J. de Boer.1990. Estimation of additive genetic variance when base populations are selected. Journal of Animal Science 68:3124-3132.

Wolc, A., C. Stricker, J. Arango, P. Settar, J.E. Fulton, N.P. O□Sullivan, R. Preisinger, D. Habier, R. Fernando, D.J. Garrick, S.J. Lamont, and J.C.M. Dekkers. 2011. Breeding value prediction for production traits in layer chickens using pedigree or genomic relationships in a reduced animal model. Genetics Selection Evolution 43: no. 1: 5.

Yu, J., and R. Bernardo. 2004. Changes in genetic variance during advanced cycle breeding in maize. Crop Science 44:405-410

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

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

List of Figures

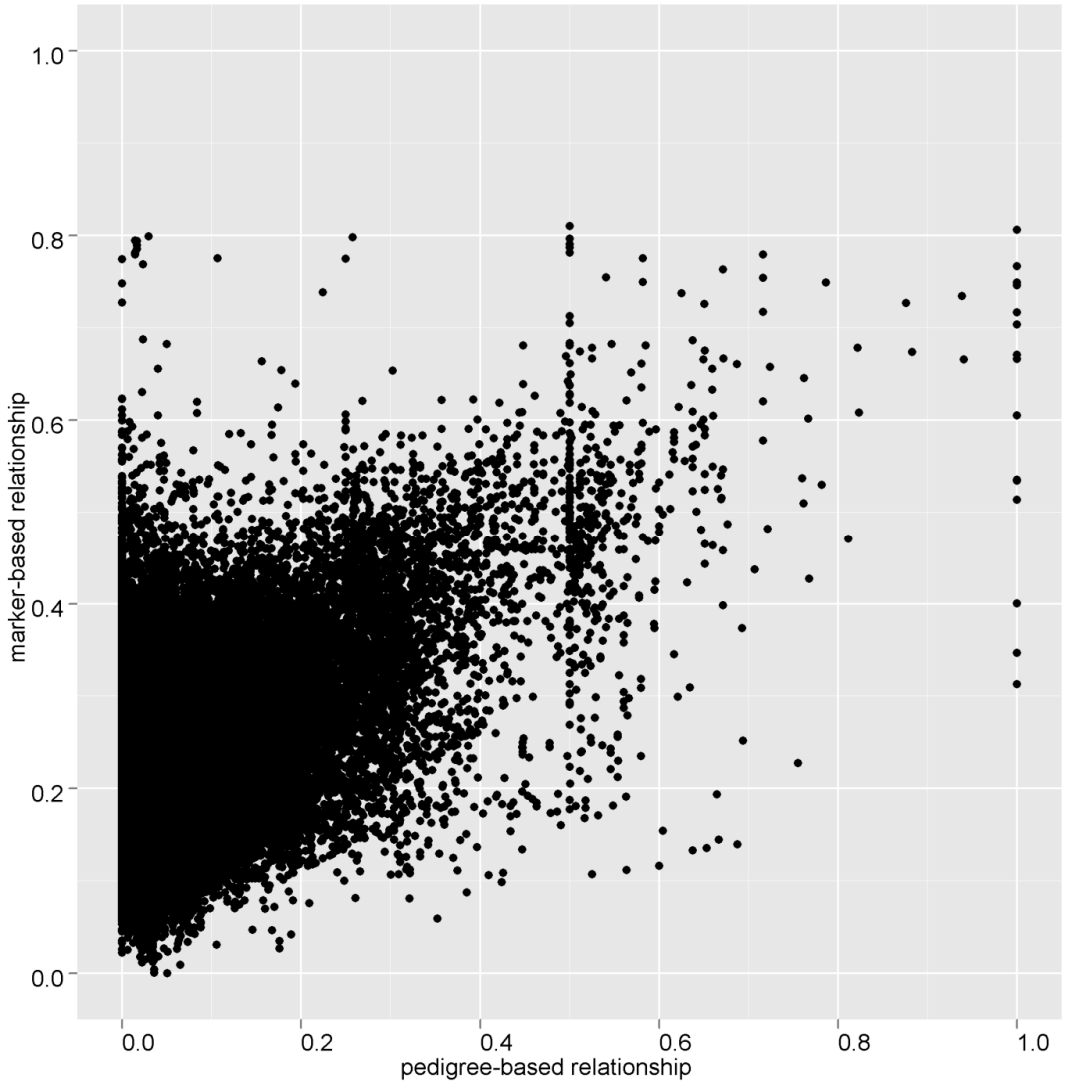


Figure 1. Scatter plot of pairwise relationship among Cycle 0 lines based on pedigree relationships (x-axis) and marker-based relationships (y-axis).

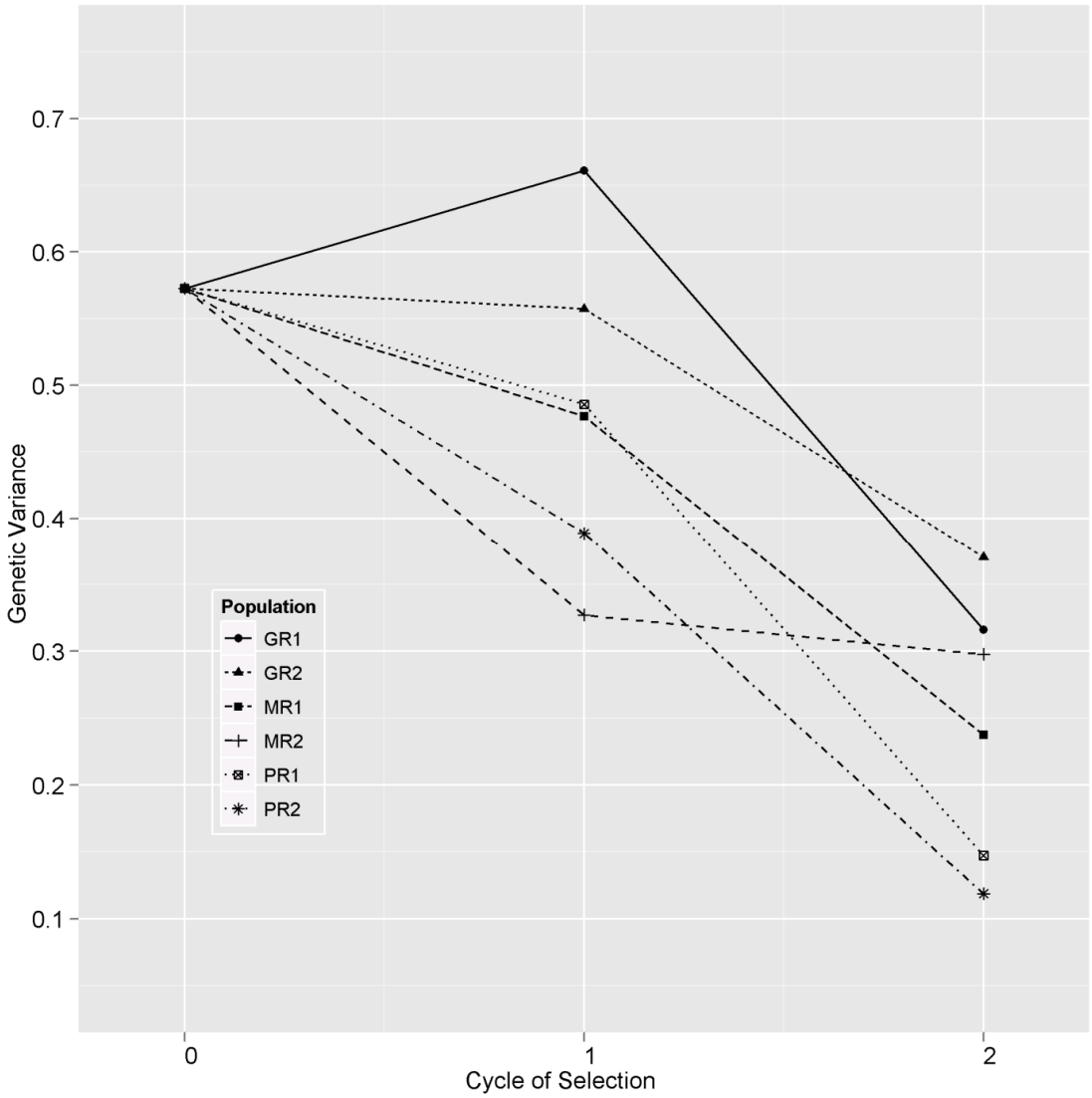


Figure 2. Genetic variances for β-glucan content of populations from Cycle 0 to Cycle 2. Each line represents a different population.

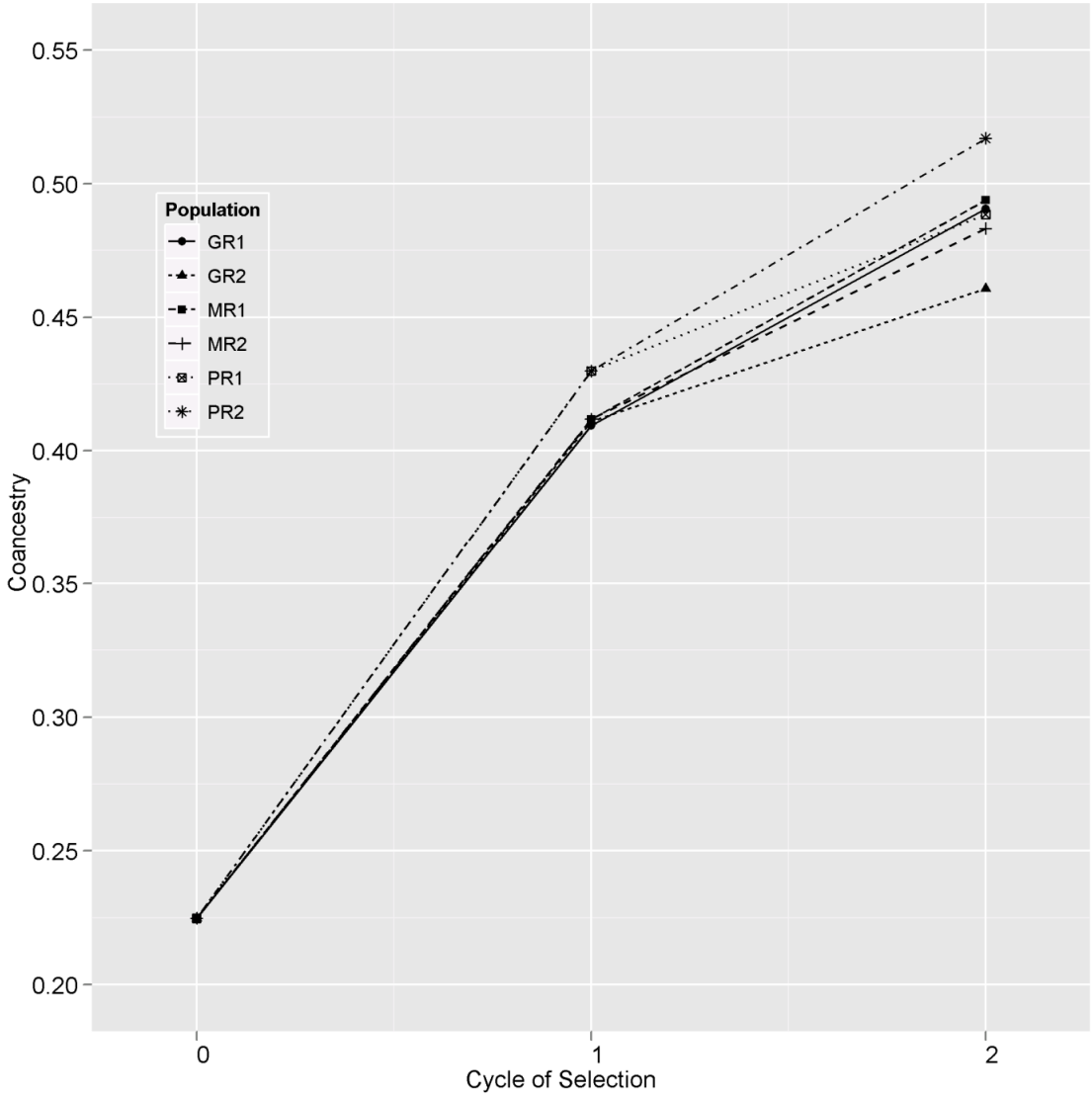
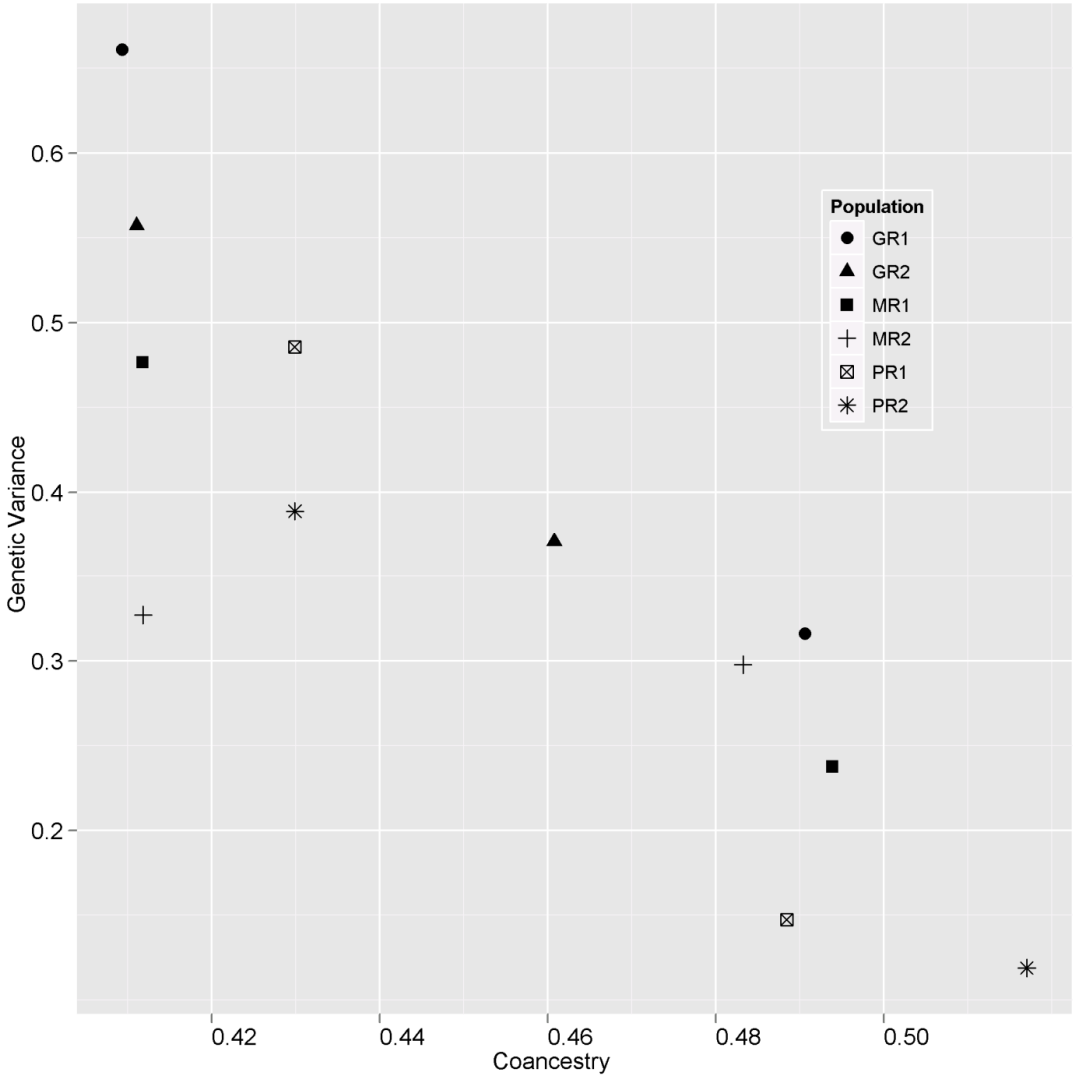


Figure 3. Average coancestry per population from Cycle 0 to Cycle 2. The reference coancestry of parents of Cycle 1 was all based on marker-based relationship data. Then the crossing or pedigree data of Cycle 1 progenies were used to compute coancestry among them. The coancestry of parents of Cycle 2 were based on the on Cycle 1, then the crossing data of Cycle 2 progenies were used to compute coancestry among them.



Supplementary Figure 1. Scatter plot of co-ancestry and genetic variance for β-glucan content. All points to the left of 0.44 coancestry are Cycle 1 progenies and those at the right are Cycle 2 progenies.

List of Tables

Table 1. Genetic parameter estimates for Cycle 0 lines using different covariance matrices.

Parameters	Identity	Pedigree-based	Marker-based
		OPN	
Genetic Variance	0.20	NA	NA
Additive Genetic	NA	0.12	0.13
Residual Variance	0.25	0.25	0.25
H^2 or h^2	0.45	0.32	0.34
Akaike Information Criterion	5363.9	5267.6	5203.3
		Ames	
Genetic Variance	0.45	NA	NA
Additive Genetic	NA	0.23	0.28
Residual Variance	0.26	0.27	0.26
H^2 or h^2	0.63	0.47	0.52
Akaike Information Criterion	4029.9	3924.5	3848.5

†H^2 is for data that used the identity matrix to define covariance among lines.

Table 2 . Estimated variances for each population and over-all residual variance.

Group	Estimate	SE	p-value	Reduction from Cycle 1 to Cycle 2	% Reduction
GR1Cy1	0.66	0.17	<.0001		
GR1Cy2	0.32	0.10	0.00	0.34	52
GR2Cy1	0.56	0.15	<.0001		
GR2Cy2	0.37	0.11	0.00	0.19	34
MR1Cy1	0.48	0.13	<.0001		
MR1Cy2	0.24	0.09	0.00	0.24	50
MR2Cy1	0.33	0.10	0.00		
MR2Cy2	0.30	0.10	0.00	0.03	9
PR1Cy1	0.49	0.13	0.00		
PR1Cy2	0.15	0.07	0.01	0.34	70
PR2Cy1	0.39	0.11	0.00		
PR2Cy2	0.12	0.06	0.02	0.27	69
Cycle 0	0.57	0.11	<.0001		
Residual	0.40	0.02	<.0001		

† p-value is the test for variance greater than zero.

Table 3. Contrast of estimated genetic variances of selection methods based on likelihood ratio test.

Selection Methods	Difference in Variance	p-value	Difference in Variance	p-value
	Cycle 0 vs Cycle 1		Cycle 1 vs Cycle 2	
GS	-0.04	0.7518	0.27	0.0359
MS	0.17	0.2059	0.14	0.1797
PS	0.14	0.3173	0.30	0.0006
	Cycle 1		Cycle 2	
GS vs PS	0.17	0.2059	0.21	0.0091
GS vs MS	0.21	0.1213	0.06	0.4028
MS vs PS	-0.04	0.7518	0.13	0.0736

† p-value is computed based on the difference between -2 REML Log Likelihood of homogeneous variance and heterogeneous variance assumption among selection methods.

CHAPTER 6. GENERAL CONCLUSIONS

The main objective of this dissertation was to investigate genetic and breeding strategies in oat for β-glucan content – a complex trait with positive human health benefits. This research presented empirical evidence that new plant breeding technologies can be used to improve β-glucan content. Several β-glucan QTL identified using both the single-marker and multiple-marker test GWAS methods were reported in Chapter 2. Some of these QTL corresponded to loci identified in previous traditional linkage mapping studies. However, none of the QTL had a large effect confirming that β-glucan content is controlled by many QTL with small effects. From a breeding perspective, this result suggests that phenotypic information will still be needed to realize a large response to selection if a GWAS-marker-assisted selection (GWAS-MAS) strategy is implemented. Because of the polygenic nature of β-glucan content, this also predicts that an alternative strategy, namely genomic selection, which uses data from all molecular markers, will be more efficient as a selection method for β-glucan content. This new method is addressed in Chapter 3, where the empirical exploration of GS models shows that it can help accelerate gains in selection not only for β-glucan content but also for other complex traits in oat. However, the accuracy of GS for those traits is influenced by size of training population, marker density, and the genetic relationships between training and selection candidates. One way to compare GWAS-MAS and GS strategies is to use them in actual recurrent selection programs for β-glucan content. Therefore, Chapter 4 and Chapter 5 present data on the comparison of GWAS-MAS, GS strategies, and BLUP-based phenotypic selection with regards to their ability to increase β-

glucan content in an actual recurrent selection program. The results of a two-cycle breeding program indicate that both marker-based approaches have higher response than pedigree-based phenotypic selection. However, GS can be the most effective in terms of response because the breeding cycle can be accelerated in this selection method with minimal reduction in genetic variance.

Overall, the knowledge generated in this study suggests that GWAS for β-glucan can be used primarily for gene discovery and understanding the genetic architecture of this trait. The polygenic nature of β-glucan implies that:

a) The identification of markers to be used in MAS can be performed by LASSO. The LASSO method provides an alternative approach that could be less stringent than traditional single marker tests but can explain more genetic variation.

b) Ridge regression BLUP-based genomic selection is more appropriate than Bayesian-based genomic selection in selecting for high β-glucan lines and could be a better choice for any other trait with similar genetic architecture.

The research questions that we have confronted in this dissertation can also be asked in any plant breeding data. However, there are still studies that can be added in these experiments, for example, we can include the phenotypic BLUP method of breeding value prediction in Chapter 3 for comparison purposes if there are no markers. It would also be beneficial if more cycles of selection can be added in the breeding program under Chapter 4 to know if there will be a plateau in response. In addition, there is also a need for future research on whether the GS methods and additive models that we are currently using can still be improved or modified. On a larger scale, better phenotyping technologies could be needed to accurately estimate marker effects.

Finally, the oat lines developed in this study can be valuable breeding materials to be incorporated into oat breeding programs for the genetic improvement of β-glucan content. These new oat lines have β-glucan content that are close to twice the amount present in standard oat varieties (4-6%). Using a GS selection strategy that can be conducted twice a year, oat cultivars with more than 10% β-glucan can be developed in a short period of time. In summary, this research addressed the stated objectives by using basic and applied research methods. The results obtained in this study will benefit the oat research and breeding community as well as plant and animal breeders that employ the methods presented here.