A CANADIAN OAT GENOMIC SELECTION STUDY INCORPORATING GENETIC AND ENVIRONMENTAL INFORMATION

A Thesis Submitted to the College of
Graduate and Postdoctoral Studies
In Partial Fulfillment of the Requirements
For the Degree of Doctor of Philosophy
In the Department of Plant Sciences
University of Saskatchewan Saskatoon

By

Bo Gui

PERMISSION TO USE

Head of the Department of Plant Sciences

College of Agriculture and Bioresources

University of Saskatchewan

51 Campus Drive

Saskatoon, Saskatchewan S7N 5A8, Canada

OR

Dean

College of Graduate and Postdoctoral Studies

University of Saskatchewan

116 Thorvaldson Building, 110 Science Place

Saskatoon, Saskatchewan S7N 5C9, Canada

ABSTRACT

Oat (*Avena sativa* L.) is an important crop in Canada that has been seeded on an average of 3.3 million acres over the past five years. It is considered a healthy cereal due to the presence of beta-glucan in the grain, which been shown to reduce the risk of heart disease, as well as being a good source of protein that is rich in globulins. Identifying new breeding strategies that can improve breeding efficiency in oat is important for future progress in this crop. To this end, genomic and environmental factors, along with their interactions, were examined to determine what contributed to variation in important oat traits. This information was then used to develop genomic selection (GS) models that can be used in oat breeding programs.

In the first study, 305 elite oat breeding lines grown in the Western Cooperative Oat Registration Trial (WCORT) from 2002 to 2014 were used to investigate important factors for genomic selection model building. The influence of phenotypic data, genotyping platforms, statistical model, marker density, population structure, training population size and trait heritability were assessed. It was determined that the machine learning model Support Vector Machine and the additive linear model rr-BLUP offered the best overall prediction accuracies. Prediction accuracy increased when using the iSelect Oat 6K SNP chip, as the marker number increased, with larger training population size and with traits that were more heritable.

In the second study, environmental and correlated agronomic variables, along with their inter-relationships, that contributed to variation in yield and grain β-glucan content in oat lines was investigated. A hypothesized structural equation model (SEM) that included variables related to environmental and phenotypic traits was created and tested against observed yield data. Significant paths were identified to explain yield variation (59%-76%) among the three oat varieties. A similar approach was taken for β-glucan in which significant paths were found which explained 16%-41% of the variation in β-glucan. Results from this study suggest that a longer period to heading and maturity, and a taller stature were the three phenotypic traits that most positively influence yield. Limited precipitation before maturity, high temperatures during heading and grain filling were the three environmental variables that contributed to decreased yield. Precipitation and July temperature were the two most important environmental variables that influenced β-glucan, while maturity was the most important trait affecting β-glucan, although the direction of effect for maturity varied by oat variety.

In the third study, additional information was added into the previous GS models to determine if prediction could be improved. Genotype, environment and their interaction were used to conduct genomic selection for yield. Four mega-environments were identified from Ward's hierarchical clustering using the significant environmental variables identified in the second study. It was found that using individual locations to represent environment provided more accuracy compared to using mega-environments. The reaction norm model was also tested which allowed significant environmental variables to be incorporated as a covariance matrix in the model. Including an environmental covariance matrix and interaction terms increased prediction accuracy compared to models with only genotype main effects. Multiple trait GS did not provide better prediction accuracy for most the traits.

In the final study, GS was used to predict the GEBVs of two populations, a biparental derived population and a population consisting of elite breeding lines from several different breeding programs. Higher predication accuracy was found in the elite breeding line population which was likely due to the closer genetic relationship between it and the training population. Finally, random selection and genomic selection were compared in the two populations. Genomic selection out-performed random selection in the elite breeding population, but not in the bi-parental population. Again, the poor performance of GS in the bi-parental population was best explained by the unrelatedness between it and the training population.

Taken together, these studies provided deeper insight into how GS could be applied in oat breeding programs.

To Maxwell and Ryan Derpak:

We are all limited by time, the time we live in and the time we have. But we have to try.

ACKNOWLEDGEMENTS

I want to give my deepest thanks to my supervisor Dr. Aaron Beattie for his guidance, patience and friendship over the course of my Ph.D. I could not have finished this work without his support and understanding during this difficult period of time. Dr. Beattie is not only my supervisor in research, but also a great influence in my life.

I would like to thank my advisory committee, Dr. Yuguang Bai (Head, Department of Plant Sciences), Dr. Tom Warkentin, (Graduate Chair), Dr. Kirstin Bett, Dr. Curtis Pozniak, Dr. Fiona Buchanan (Department of Animal Sciences, University of Saskatchewan), Dr. Mika Asai-Coakwell (Department of Animal Sciences, University of Saskatchewan) and Dr. Juan Diego Hernandez Jarquin (Department of Agronomy and Horticulture, University of Nebraska-Lincoln) for serving as the external examiner. I would like to say a special thanks to Dr. Curtis Pozniak for his guidance in helping to understand plant breeding and quantitative genomics. I would like to say a special thanks to Dr. Kirstin Bett for the advice she provided to me, not only as my committee member, but also as a female scientist, a wife, and a mother.

I would also like to thank our research team in the Barley and Oat Breeding Program at the Crop Development Centre for their help and support: Peter Eckstein, Mark Colin, Dr. Xiangmin Zhang, and Lindsay Griffith. I also want to give my great appreciation to the field crew, who worked very hard seeding, weeding and harvesting my field plots, and providing the raw data.

I would like to acknowledge the help from Dr. Rosalind Buekert, Dr. Brian Fowler, Dr. Eric Lamb and Dr. Steve Shirtliffe for my structural equation modelling; Dr. Yan Yan, Dr. Lingling Jin and Dr. Yinghong Shi for their help with coding in R. A special thanks goes to Dr. Yong Bao for genomic selection modelling help. I also would like to thank Dr. Kirby Nilsen for being the “extra hand” in my project and helping me explore the possibilities of reference genome-assisted marker selection.

I would also like to personally acknowledge the funding provided by the Department of Plant Sciences Devolved Graduate Scholarship, the Harold Horner Scholarship and the Gerhard Rakow Legacy Award.

I want to express my great appreciation to my family. I am thankful that my mom put my education first when there were very limited resources when I grew up. I am extremely grateful that my parents flew from China in March 2020 despite the COVID-19 pandemic to help me take care of my family so I can finish writing this thesis. I also want to express my gratitude to my husband, Jordan Derpak, for being my buddy in the battlefield of life and letting me know the other possibilities in the future.

In the end, I would like to express special thanks to Department of Plant Sciences and my supervisor Dr. Beattie for accepting me as a Ph.D student. It is through this degree that I found my limitation. It was a painful and dark period of my life when I discovered the inadequacy of my research ability, and it took a tremendous amount of time to gather my courage to keep on going. However, I felt the responsibility to write what I have learned for this thesis so the oat community can learn from my success and failure. It is with this thesis that I realized that we are not only bound by our own limitations, but also the time we are living in. My intention to model and discover the important attributes to yield variation was right-minded, but my limited research capability, the inadequate amount of information about plant growth and development and the environment, along with the rudimentary stages of statistical models for GS led to a moderate thesis. I understand the effort was needed and the necessity of research like this to move GS a little further into the future. Despite my feeling like an ant pushing a boulder uphill in this project, I still see a great future in genomic selection and believe it will provide great service in breeding in the near future when we are not limited by data, and machine learning algorithms are more developed to capture the essence in the field. I am glad I pushed as far as I could. The best is yet to come.

COVID-19 pandemic was a challenge and a blessing in disguise. It gave me more time to reflect and realize that even with extreme obstacles, I still wanted to finish this thesis, imperfect as it might be. Critical thinking and the freedom to choose for myself despite the background noise have guided me on this bumpy road. May they never die.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AE	Across environment
AFLP	Amplified fragment length polymorphism
BC	Bayes Cpi
BG	β-glucan
BGLR	Bayesian Generalized Linear Regression
BLUP	Best linear unbiased prediction
BP	Breeding population
BRR	Bayesian ridge regression
CDC	Crop Development Centre, University of Saskatchewan
CDNA	Complimentary DNA
CORE	Collaborative Oat Research Enterprise
DArT	Diversity array technology
DH	Double haploidy
DNA	Deoxyribonucleic acid
EBVs	Estimated breeding values
EC	Environmental covariates
GBS	Genotyping by sequencing
GEBVs	Genomic estimated breeding values
G×E	Genotype × environment
GRT	Groat percentage
GS	Genomic selection
GWAS	Genome-wide association study
h^2	Narrow sense heritability
H^2	Broad sense heritability
HD	Days to heading
LD	Linkage disequilibrium
LR	Linear regression
MCMC	Markov Chain Monte Carlo
MAS	Marker-assisted selection
MAT	Days to maturity
M×E	Marker × environment
MWK	Thousand kernel weight
NGS	Next-generation sequencing
NJ	Neighbour joining
NN	Neural networks
PCA	Principal component analysis
PH	Plant height
PLP	Plumpness
PRO	Protein
QTL	Quantitative trait locus

RAPD	Random amplified polymorphic DNA
RHSK	Reproducing kernel Hilbert space
RIL	Recombinant inbred lines
RF	Random forest
RFLP	Restriction fragment length polymorphism
rr-BLUP	Ridge regression best linear unbiased prediction
SCAR	Sequence characterized amplified regions
SE	Single environment
SEM	Structural equation model
SNP	Single nucleotide polymorphism
SSR	Simple sequence repeat
SVM	Support vector machine
TBV	True breeding value
VP	Validation population
THN	Thins
TP	Training population
TWT	Test weight
WCORT	Western Cooperative Oat Registration Trial
YLD	Grain yield

CHAPTER 1. GENERAL INTRODUCTION

1.1 Background

Oat (*Avena sativa* L.) is an important cereal crop that is grown world-wide, primarily in cool temperate climates. Its production ranks seventh in the world, behind maize, rice, wheat, barley, sorghum and millet (FAOSTAT, 2018). Oats were widely grown and used mainly as animal feed but demand drastically decreased as farming became mechanized. While the world production of oat decreased from 35 million tonnes to 20 million tonnes between 1994 and 2018, Canada has remained a major producer of premium oats for the worldwide oat market, consistently contributing more than 10% of the world's high-quality oats over the past 25 years (FAOSTAT, 2018). Oat remains an attractive option for Canadian producers because of its low input costs compared to other crops, such as canola, and its high demand, including new uses in products, such as oat milk or as a source of plant-based protein.

Being the largest oat exporter in the world, Canada exports increased by 6% to 2.62 million tonnes in 2019-2020 (Statistics Canada, 2020). The majority of Canadian oats are exported to the US, destined for the high-quality food market which recognizes oat as a healthy cereal, due in part to the presence of β-glucan. Numerous nutrition studies revealed oats can reduce the risk of coronary heart disease and diabetes as a result of the β-glucan which lowers blood cholesterol and intestinal absorption of glucose (Behall et al. 2006; Butt et al. 2008; DeGroot, 1963; Wolever et al. 2011).

Oat yields in Canada are among the highest in the world, averaging above 3,000 kg ha^{-1} (FAOSTAT, 2018). High yield results from a combination of improved agronomic practices and superior cultivars from oat breeding. Traditional oat breeding mainly relies on phenotypic selection based on visual, field and laboratory evaluation. The beginning of the breeding process traditionally involves crossing two or three parents, F1 seeds are then collected, and the subsequent generations (F2-F5) are grown and bulked until homozygosity where limited numbers (200-500) of single plants are selected based on their phenotypic appearance in the field. Oat lines derived from the single plant selections then undergo multi-year, multi-location yield testing in the field, quality testing in the lab and are released as cultivars if requirements are met. The breeding cycle from crossing to release of a cultivar is usually 12-15 years. This process requires a great deal of resources, time and labour. Moreover, it is problematic at the single plant selection stage as the variability for low heritability traits is more-or-less randomly sampled, rather than being selected towards a desired direction. Thus, this conventional method fails to maximize genetic gain because of the non-heritable environmental influences on the response to selection for low heritability traits.

By contrast, marker-assisted selection (MAS) using DNA markers has been effective in improving simply inherited traits, such as crown rust resistance and β-glucan content in oats (Asoro, 2012; McCartney et al. 2011). Selection using genetic markers holds several advantages over traditional phenotypic selection: firstly, MAS can be performed on single plants in earlier generations; secondly, selection can be done more accurately in a lab without the influence of environment; thirdly, selection can be done for traits in the absence of the corresponding selection pressure like disease or lodging. Restriction-fragment length polymorphisms (RFLPs), amplified fragment length polymorphisms (AFLPs), random amplified polymorphic DNA (RAPDs), diverse arrays technology (DArTs) and sequence characterized amplified regions (SCARs) were marker systems commonly used to detect marker gene associations and for maker development, despite the limited number of available markers and low throughput nature of these marker systems. Recently, single nucleotide polymorphisms (SNPs) have become the most predominant molecular markers for plant genetic studies because of: 1) their abundance in both genic and non-genic regions across the entire genome, 2) their allele-discriminatory nature, and 3) their amenability to high throughput genotyping platforms (Rimbert et al. 2018). DNA markers for crown rust resistance (i.e. *Pc91*, *Pc94*, *Pc45*, *Pc98*, APR) were developed based on SNPs and widely implemented in oat breeding programs in Canada for resistance screening in early segregating F2-F4 generations (Aung et al. 1996; Chong et al. 2004; Gnanesh et al. 2014; McCartney et al. 2011).

The traditional methods of MAS assume the average marker effects on phenotypes are well established and the favourable allele is known by the user (Bernardo and Charcosset, 2006; Charmet et al. 1999; Hospital et al. 1997). These approaches perform well for major-gene traits. For quantitative traits controlled by many genes and heavily influenced by the environment, locus identification and estimation of marker effects are problematic. To solve these problems, Meuwissen et al. (2001) proposed the concept called genomic selection (GS), where genome-wide markers were integrated in prediction models to capture alleles with minor effects. Advanced statistical models were used to estimate the effects of all marker loci simultaneously, even for markers that were not “significant”. In GS, two separate groups of individuals are involved. The first group is called the training population (TP), consisting of individuals/lines with both phenotypic and genotypic data. Through statistical models, allele effects of all loci are simultaneously estimated. The other group, the testing or breeding population, are lines with only genotypic information, which can be used to calculate genomic estimated breeding values (GEBVs) to assist selection in the absence of phenotypic data. This evolutionary development shifts how selections could be done for quantitative traits. Replication of lines are no longer needed as long as the alleles were replicated within and across environments. Selection might be done on single plants during early generations when conditions (such as replicated environments and sufficient seed) do not exist. This would shorten the breeding cycle significantly and redirect resources from the field and lab to large scale single plant genotyping. The number of individuals tested could potentially increase exponentially, especially in the information era when genotyping has improved dramatically, at a decreased and more feasible cost. In theory, shortened breeding cycles and more intense selection pressure on much larger populations would enhance genetic gain greatly.

The GS process might be executed in three general steps: first, creation of an initial genomic model for target traits using genotypic and phenotypic data available for a training population and validation of the model; secondly, application of the genetic model to derive GEBV for individuals within a related breeding population; and lastly, re-evaluation of the genomic model based on phenotypic data arising from individuals selected by the genetic model, as well as, additional genotypic and phenotypic data representing new elite and novel germplasm. Several simulation and empirical studies illustrated the success of GS based on GEBVs in animal and plant breeding scenarios (Lorenzana and Bernardo, 2009; Meuwissen et al. 2001; Storlie and Charmet, 2013; VanRaden et al. 2009; Zhong et al. 2009). Important factors influencing GS included the number of markers, statistical models, population size, heritability, linkage dis-equilibrium and population structure (Calus, 2010; Lorenzana and Bernardo, 2009; Luan et al. 2009; Meuwissen et al. 2001; Solberg et al. 2008; VanRaden et al. 2009; Moser et al. 2010; Heffner et al. 2011b; Asoro et al. 2011; Hamblin et al. 2011; Jannink et al. 2010; Combs and Bernardo, 2013; Wientjes et al. 2013).

The development of oat varieties with better nutritional properties for consumers, improved agronomic performance and disease resistance for Canadian growers, and enhanced milling properties for processing facilities would help oat breeders to improve oat varieties more efficiently. The goal of this thesis is to investigate both genetic and environmental factors that could be used to develop genomic selection strategies for Canadian oat breeding.

1.2 Research hypotheses

1. Genotyping-by-sequencing (GBS) will provide a greater amount of unbiased marker information and higher marker density than the iSelect oat 6K SNP Chip (6K), which will increase the accuracy of the phenotypic prediction model,
2. Different statistical models will provide different prediction accuracy of the GS models,
3. Greater marker density and higher heritability traits will produce higher prediction accuracy in GS models,
4. A larger training population which is genetically related to the validation population will improve GS model accuracy,
5. Incorporating environmental parameters will improve the accuracy of the GS model,
6. Lines selected based on GEBV will show more genetic improvement compared to lines selected using random selection.

1.3 Objectives

1. To compare the accuracy of GS models created from two genotyping systems, the iSelect oat 6K SNP Chip and genotyping-by-sequencing,
2. To evaluate the effect of marker density, heritability, population size and different prediction models on the accuracy of GS in a training population,
3. To identify influential environmental factors for yield and β-glucan,
4. To include environmental parameters in the GS model and evaluate their effects on prediction accuracy in a training population,
5. To classify mega-environments and test GS models with G × E interaction,
6. To include additional information such as correlated traits in GS and test prediction accuracy,
7. To test GS in two different types of oat populations,

8. To compare the performance of random selection versus GS method in the two different oat populations.

CHAPTER 2. LITERATURE REVIEW

2.1 Oat as a crop

Oat (*Avena sativa* L.) is an important cereal crop that is grown world-wide. Its production ranks seventh in the world, behind maize, rice, wheat, barley, sorghum and millet (Table 2.1) (FAOSTAT, 2019). More than 10% of world oat production is contributed by Canada, predominantly within Saskatchewan, Alberta and Manitoba. Since 1995, Saskatchewan has been the largest oat producer (Statistics Canada, 2019). It produced 1.96 million tonnes or approximately 53% of total Canadian production on average from 2015 to 2020 (Statistics Canada, 2020).

Table 2. 1 World grain production (in millions of Tonnes) of principle cereal crops from 2010-2019 (FAOSTAT, 2019).

Crop	**Year**									
	2010	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**
Maize	851.7	886.7	875.0	1016.2	1039.2	1052.1	1127.0	1164.4	1147.6	1148.5
Rice	701.1	726.4	736.6	742.5	742.5	745.9	751.9	769.8	782.0	755.5
Wheat	640.8	696.9	673.7	710.4	728.7	741.6	748.4	773.5	734.1	765.8
Barley	123.3	132.8	132.2	143.5	145.1	147.8	145.9	149.1	141.4	159.0
Sorghum	60.2	56.8	57.3	61.9	68.3	66.0	63.7	57.7	59.3	57.9
Millet	32.8	27.1	26.6	26.4	28.4	28.2	27.7	28.4	31.0	28.4
Oat	19.7	22.6	21.2	23.8	22.8	23.3	23.7	26.1	23.1	23.1
Rye	11.9	13.1	14.5	16.7	15.2	13.8	13.4	13.8	11.3	12.8

Oat is predominantly consumed as a feed or forage for livestock in many regions. Oat grain is also a good source of protein, fiber and minerals, which make it attractive for human consumption. Oatmeal, oat flakes, oat bran and oat flour are common forms in which oat is produced for food purposes. The chemical composition of oat grain is similar to other cereal grains, other than it has a higher oil and protein content (Table 2.2) (reviewed by Belitz and Schieberle, 2009).

Table 2. 2 Grain chemical composition of principle cereal crops.

Crop	Protein (%)[1]	Lipid (%)	Carbohydrate (%)	Fiber (%)	Minerals (%)
Maize	9.3	3.9	65.1	9.8	1.3
Rice	7.5	2.4	75.1	2.2	1.2
Wheat	11.9	2.2	60.4	13.5	1.5
Barley	10.7	2.1	64.2	9.9	2.3
Millet	10.7	4.2	69.8	3.9	1.6
Oat	12.8	7.2	56.5	9.8	2.9
Rye	9.6	1.7	61.5	13.4	1.9

[1] Percentages normalized to 12% grain moisture (modified from Belitz and Schieberle, 2009).

Oat is also known to have relatively high amounts of soluble fiber in the form of β-glucan (2-8%), which is comparable to barley (3-7%) and much higher than wheat and rye (0.5-2%) (Marlett, 1993; Welch, 1995; reviewed by Belitz and Schieberle, 2009). Breeding efforts to lower the oil content and increase β-glucan content have produced cultivars grown in Canada that normally contain less than 7.0% oil and more than 5.0% β-glucan (Beattie, personal communication). Studies have shown that diets containing oat products, such as oatmeal and oat bran, were effective in lowering blood cholesterol in animals and humans, which reduced the risk of heart disease (Anderson et al. 1991; Bridges et al. 1992; Chen et al. 1981). These effects were attributable to β-glucan which was specifically indicated to lower plasma cholesterol and reduce the risk of heart disease in numerous studies (Chen et al. 1981; Klopfenstein and Hoseney, 1987; Jennings et al. 1988; Liatisa et al. 2009; Queenan et al. 2007). As a result, health claims for oat exist in both Canada (Health Canada, 2010) and the United States (FDA, 1997) indicating that the soluble fiber contained in oats has the ability to reduce the risk of heart disease. Oat is also used in cosmetic products, as it has both soothing and sun-blocking properties (Potter et al. 1997; Singh and Belkheir, 2013; Sur et al. 2008).

2.1.1 Global and Canadian oat production and challenges

Although world oat production gradually dropped to less than 30 million tonnes due to mechanization replacing horse-based power in agriculture, production has stabilized since 2000, averaging 24.2 million tonnes (FAOSTAT, 2001-2019). Over the past 60 years (1961-2019), Russia, Canada and USA have been the top three oat producing countries in the world. Over the past 5 years, Canada has produced 3.68 million tonnes of oats per year (Statistics Canada 2020). Western Canada accounted for 92% of the production with Saskatchewan being the top producing province, followed by Manitoba, Alberta and British Columbia. Top milling oat cultivars grown in 2020 in Western Canada include CS Camden, AC Morgan, Summit and CDC Arborg (Canadian Grain Commission, 2020).

The reduction in available cultivated land due to urbanization, competition from other crops and difficulties resulting from climate change make consistent oat production challenging. As such, high yielding oats with high-quality is in ever greater demand. Good production of high-quality oats can be achieved by a combination of factors: 1) breeding for improved disease resistance, lodging resistance, increased leaf area, and photosynthetic capacity (Forsberg, 1986), 2) improved qualities such as groat percentage, kernel weight and dietary fibre content, and 3) optimizing agronomic management such as using high quality seed, controlling seeding timing and density, use of herbicides and fungicides, and good crop rotation to manage soil and diseases (Forsberg and Reeves, 1995). Climate change may also impact the production of oats in different ways. For example, the increasing carbon dioxide content in the atmosphere may increase production in C3 plants like oats (Tester and Langridge, 2010) and may also benefit oat production in countries at higher latitudes by extending the growing season. On the other hand, photorespiration and night-time transpiration may increase due to the increased temperatures, offsetting any gains from increased carbon dioxide. Extreme weather conditions, including sudden storms, extended heatwaves and drought, or excessive moisture are also a risk to oat production. These risk factors, along with new disease and pest pressures, could hinder the production of high quality, high-yielding oats and therefore, collectively form part of the focus of oat breeding activities.

Even with the above-mentioned obstacles, successful oat breeding strategies have led to average yields more than 3,000 kg/ha, making Canada one of the highest-yielding production areas in the world (FAOSTAT 2019).

2.1.2 Conventional oat breeding

Breeding objectives are divided into three general categories: agronomy, disease and pest resistance, and quality. More specific objectives for spring planted milling oat production in North America include: yield, resistance to lodging, acceptable maturity, shorter stature, disease resistance (resistance to crown rust, stem rust, BYDV, FHB and smut), milling quality (high groat percentage and low groat breakage), physical grain characteristics (white hulls, high plumpness, low thins, high test weight, high kernel weight), and nutritional grain characteristics (higher β-glucan content, lower oil content) (reviewed by Valentine et al. 2011). Most oat breeding programs involve hybridization of 2-4 parents to initiate population development, followed by bulk, modified pedigree-bulk or single seed descent breeding through early generations, with single-plant selection for high heritability traits (e.g. crown rust resistance) possible. The subsequent generations, typically at the F5-F6 generations, allow selection for plant stature, straw strength and maturity. The F6-F7 generation is typically a single replication, single location test for which physical and nutritional grain characteristics can be tested. More advanced generations beyond this point are grown in replicated, multi-location trials over several years to evaluate lower heritability traits like yield. Final stages of the breeding process may involve cooperative evaluation by expert committees with registration and commercialization being the ultimate endpoint of the breeding process. Double haploidy (DH) has been successfully utilized in certain cereal crops, such as wheat and barley, to help shorten the breeding time by inducing the plant to reach homozygosity in one generation (Guha and Maheshwari, 1964; Kasha and Kao, 1970; De Buyser et al. 1987). Unfortunately, application of the DH method in oat has met with limited success via the maize pollen method (Rines and Dahleen, 1990; Marcińska et al. 2013). Work by Tanhuanpää et al. (2008) was able to produce an anther culture-derived DH population for cultivated oat, and more recent work by Ferrie et al. (2014) saw some success producing DHs using microspores. However, the use of this technique remains limited.

Conventional phenotypic oat breeding methods have two significant drawbacks. Firstly, the loss of genetic variability at earlier generations when only high heritability traits are selected based on visual evaluation of single plants, which limits genetic variability remaining for subsequent selection of lower heritability traits, and secondly, an inability to select for low heritability traits (i.e. yield) at earlier generations because of inadequate amount of seed needed for replicated, multiple environment testing. Molecular markers are seen as a means to alleviate these two issues.

2.1.3 Molecular-assisted breeding in oats

One of the goals of plant breeding is to achieve maximum genetic gain, which is a function of response to selection over time. Genetic gain obtained using conventional breeding is compromised not only by the prolonged duration of a breeding cycle, but also by the reduced (and random) genetic variability produced after single plant selection which leads to a reduced response to selection. Molecular markers could be used to complement phenotypic selection of target traits in such a manner that traits could be evaluated in the absence of phenotypic data and genetically inferior genotypes could be discarded at earlier generations. Moreover, molecular makers shed insight on the genetic makeup of individuals and may allow for optimized parental selection at earlier generations (thus shortening the breeding cycle).

Cost, efficiency, accuracy and reproducibility are the major criteria determining the utility of markers. Restriction fragment length polymorphism (RFLPs), random amplified polymorphic DNA (RAPDs) and amplified fragment length polymorphism (AFLPs) are used less frequently because they are labor-intensive, costly and time-consuming. Simple sequence repeat (SSR) markers are considered relatively efficient, robust and cost-effective. Unfortunately, only limited numbers of SSR have been discovered in oat (Li et al. 2000; Pal et al. 2002, Becher, 2007). Among DNA markers, single nucleotide polymorphisms (SNPs) are considered the most useful due to their abundance, co-dominant nature, amenability to high-throughput technology and increasing ease of discovery due to the decreasing costs of sequencing (reviewed by Mammadov et al. 2012).

Before the era of high-throughput sequencing was applied to oat, a limited number of SNP markers were identified in oat genetic studies. Some of the few examples include SNPs linked to *Pc68* and *Pc94* crown rust resistance genes, which were developed from RFLP and AFLP markers (Chen et al. 2006; Chen et al. 2007), and SNPs linked to the *Dw6* dwarfing gene which was developed from RAPD and SSR markers (Tanhuanpää et al. 2006). However, work over the past decade has seen a dramatic increase in the genetic resources and markers available to oat breeders, and the possibilities of utilizing markers to improve genetic gain are coming closer to reality.

2.2 Genetic resources for oat research

Cultivated oat is a self-pollinated hexaploid (2n=6x=42) with 7 pairs of chromosomes present in each of three diploid sub-genomes (AA, CC, DD), which was domesticated from the wild hexaploid oat *A. sterilis* L. that is thought to have originated from hybridization between a CCDD tetraploid and an AA diploid (Zhou et al. 1999; Yan et al. 2016). An alternate theory proposes that hexaploid oat may have formed by hybridization between an ancestral CC genome diploid with an ancestral AA genome diploid to form a tetraploid species (represented by species such as *A. magna*, *A. murphyi*, or *A. insularis*). Such tetraploids would then have hybridized again with an ancient AA genome diploid to create the current hexaploidy arrangements of chromosomes (Chew et al. 2016). The genome is estimated to be 11.3-14.0 Gb in size (Bennett and Smith, 1976; Walstead et al. 2018). Genomic studies in oats have been hindered due to its large genome size, complexity of the genome and a scarcity of sequence data (Oliver et al. 2013). Compared to other polyploids, diploid relatives of oat offer limited value when constructing hexaploid oat maps because of a lack of colinearity due to chromosomal rearrangements among the oat sub-genomes of hexaploid oat (Jellen et al. 1994). Despite these obstacles, in June 2020, the first hexaploid oat genome was published using the CDC breeding line OT3098 (Pepsico, 2020), thus providing exciting possibilities for further genetic research in oats.

2.2.1 Genetic mapping of hexaploid oat

Genetic maps are essential for oat breeding because they help provide an understanding of the proximity of different genes, which informs a breeder about the potential of recombining certain traits. In addition, maps help determine the genomic location of new genes and may reveal potential allelic relationships to current genes. Within a species like oat where defined chromosomal rearrangements (e.g., the 7C-17A reciprocal translocation) exist within a certain portion of the gene pool, a genetic map can help a breeder understand which genes they should expect to be linked.

The first oat genetic map contained 561 RFLP loci across 38 linkage groups with a map size of 1,482 cM and was developed using a RIL (recombinant inbred line) population derived from the cross Kanota × Ogle (KO) (O'Donoughue et al. 1995). An updated KO map with 29 linkage groups and a map size of 1,890 cM was subsequently published by Wight et al. (2003). Additional mapping efforts were made in several other oat crosses (e.g. Ogle × TAM O-301 map containing 441 loci with 34 linkage groups and a map size of 2,049 cM), but comparative locus positions among different genetic maps remained a challenge due to the lack of common markers between mapping populations (Portyanko et al. 2001; Rines et al. 2006). To address this issue, high throughput diversity array technology (DArT) markers were developed from a panel of 60 oat varieties of global origin. This result not only produced an enhanced KO map with an additional 1,010 DArT markers (Tinker et al. 2009), but a common genotyping platform was created to allow cross referencing between mapping populations.

The most recent and successful attempt to create a well-saturated oat genetic map was made by the Collaborative Oat Research Enterprise (CORE), a team of oat researchers from North and South America, Europe and Australia. Using root, shoot, embryo and pistil tissue collected from 20 genotypes of diverse origin from within the cultivated gene pool, a large cDNA library was created and sequenced (using Roche 454 sequencing) from which 11,000 high confidence *in silico* candidate SNP loci were identified (Oliver et al. 2013). Using this information, cDNA-based SNPs predicted with highest confidence (2,606 in total) were combined with other known markers (for a total of 3,072 SNPs) and incorporated into a pilot Illumina (San Diego, CA) GoldenGate assay. Among these SNPs, 985 SNPs were mapped across six bi-parental RIL mapping populations. These six genetic maps, along with previous maps, were used to create the first physically-anchored oat consensus map, resolving 21 linkage groups with a total distance of 1,839 cM (Oliver et al. 2013).

2.2.2 iSelect oat 6K SNP chip

Another important output from the CORE project was the development of a high throughput genotyping tool, the iSelect oat 6K BeadChip (Tinker et al. 2014). Initially, 5,743 SNPs were identified from four different sources: 1) cDNA libraries representing four tissue types from 20 hexaploid oat cultivars, 2) DArT sequences from 24 hexaploid oat cultivars, 3) sequences from tetraploid oat genomic reductions, and 4) GBS sequences from RIL progeny from the ‘Ogle’ × ‘TAM-O-301’ population. In total, 4,975 SNPs survived the quality evaluation process and were successfully converted to the Infinium assay. About 3,500 SNPs were polymorphic when validated across 1,100 lines from six mapping populations and a set of diverse oat cultivars. The Infinium iSelect oat 6K SNP chip was recently used to genotype a segregating population that was used for QTL discovery of an adult plant crown rust resistance (Nanjappa et al. 2014), seedling crown rust resistance (Esvelt Klos et al. 2017) and stem rust (Kebede et al. 2020).

2.2.3 Genotyping-by-sequencing in oat research

Removing repetitive regions to reduce genome complexity and increasing the targeting efficiency of lower copy regions are important for genetic research in species with complex genomes (Gore et al. 2007; Gore et al. 2009). Genotyping-by-sequencing is another popular way of sequencing, which uses restriction enzymes to reduce the complexity of the genome to be sequenced by removing the repetitive regions. The reduced genome size thus allows multiple samples (96-384) to be sequenced in parallel through the use of DNA barcoded adapters which uniquely identify each sample (Elshire et al. 2011). Genotyping-by-sequencing thus provides a relatively low-cost method of generating a large amount of SNP information across multiple genotypes. Furthermore, GBS does not require prior knowledge, such as a reference genome sequence, to discover SNPs, which makes it suitable for researching understudied crops (Elshire et al. 2011). This technology was utilized in oat to identify numerous SNPs from EST libraries derived from different tissue types which were then incorporated into a high-throughput genotyping platform (Oliver et al. 2011 & 2013).

As a result of the cost-effectiveness and marker abundance, GBS has been widely used in plant genetic mapping studies (Baird et al. 2008; Poland et al. 2012a), breeding programs (Poland et al. 2012b; Crossa et al. 2013: Lado et al. 2013), and diversity studies (Elshire et al. 2011; Fu, 2012; Lu et al. 2012). For example, GBS identified 41,371 SNP markers across a set of 254 advanced wheat breeding lines at a cost of approximately $20 per sample (Poland et al. 2012b). An oat study evaluated GBS (using a the *PstI-MspI* two enzyme method) in seven bi-parental populations and 2,664 diverse inbred lines (Huang et al. 2014). A total of 45,117 loci were discovered and placed on the oat consensus map. This work provided a positional reference for these markers which will be valuable in future studies that use GBS. More than 70,000 loci were later added to the consensus map by Bekele et al. (2018) in a *de novo* GBS analysis of 64-base tag-level haplotypes, in which 241,224 SNPs were discovered in 4,657 accessions of cultivated oat. The ever-decreasing cost of GBS should allow breeders to consider genotyping a larger percentage of their breeding populations, along with a greater number of lines within each population, for use in genomic selection strategies.

2.3 Genomic selection

The concept of genome selection (GS) was first introduced by Meuwissen et al. (2001). Genomic selection attempts to estimate the effect of markers across the entire genome on a target trait phenotype, instead of relying on one or a few predefined markers (Heffner et al. 2009; Jannink et al. 2010), There is no requirement to further investigate the function of genes underlying or close to the associated markers, but it is certainly a worthwhile endeavor in order to begin understanding the genetic/biochemical pathways underlying the traits of interest. The ultimate application of GS according to Lorenz et al. (2011) was to "chip a seed, extract its DNA, discard or select it as a parent of the next generation".

Genomic selection has been applied in animal breeding to improve genetic gain and has shown much success (Goddard and Hayes, 2007; Hayes et al. 2009). Genomic selection has been widely implemented in the dairy cattle breeding industry. For example, GS has shortened the generation interval up to 4 years over progeny testing in dairy bull breeding (Schaeffer 2006). In a simulation study, Meuwissen et al. (2001) predicted the estimated breeding value with an accuracy near 0.85 using only marker information. Prediction accuracy ranged from 0.44 - 0.79 for traits with low to moderate heritability (0.04 to 0.5) when predicted values and true breeding values were compared in Holstein bulls (VanRaden et al. 2009). In pigs, a 23-91% genetic gain in maternal traits was observed by Lillehammer et al. (2011) using GS, as compared to 7% genetic gain achieved by traditional progeny testing.

As a result of this success in animal breeding, GS is being actively explored within plant breeding and is believed to hold much potential for increasing genetic gain per breeding cycle (reviewed by Heffner et al. 2009; Cabrera-Bosquet et al. 2012). The process can be divided into three general components (Fig. 2.1): 1) creation of an initial genomic model for target traits using genotypic and phenotypic data available from a training population and validation of the model, 2) application of the genetic model to derive GEBV for individuals within a related breeding population for which only genotypic information is available, and 3) re-evaluation of the genomic model based on phenotypic data arising from individuals selected by genetic model, as well as, additional genotypic and phenotypic data representing new elite and novel germplasm (Heffner et al. 2009; Jannink et al. 2010; Heffner et al. 2011a).

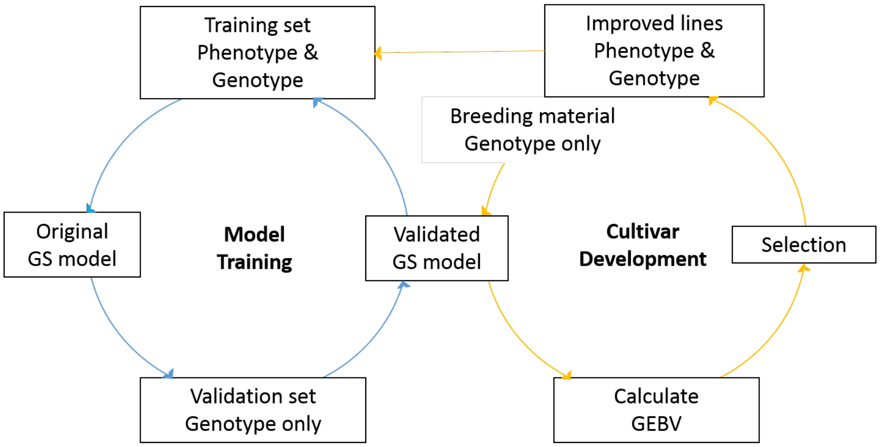


Fig. 2. 1 Generalized genomic selection scheme.

2.3.1 Design of training and cross-validation populations

Estimation of individual SNP effects on a trait of interest is a crucial step in GS. A population which is used to estimate these SNP effects using both phenotypic and genotypic information is termed a training population (Meuwissen et al. 2001). Two predominant population types with different genetic structures have been used to create GS training populations. Bi-parental populations have been used in maize, Arabidopsis and barley (Lorenzana and Bernardo, 2009; Heffner et al. 2011b) while multi-line (association mapping) populations with greater genetic diversity have also been used in barley (Zhong et al. 2009), maize (Crossa et al. 2010) and wheat (de los Campos et al. 2009; Crossa et al. 2010; Heffner et al. 2011b). The historical recombination present in multi-line populations captures a large number of recombination events in comparison to the limited number produced in bi-parental crosses (Flint-Garcia et al. 2003). While good from the resolution stand-point, these populations require a larger number of markers to identify the smaller LD blocks. By contrast, the non-random mating bi-parental-derived populations contain a limited number of recombination events which results in more extensive linkage disequilibrium (Doerge, 2002; Holland, 2007; Smith et al. 2008; Zhu et al. 2008), allowing fewer markers to be used to tag the LD blocks (Lorenzana and Bernardo, 2009). In addition, different populations will have different population structures which can affect the prediction accuracy in GS (Saatchi et al. 2011; Riedelsheimer et al. 2013; Wray et al. 2013).

In GS, validation populations are used to test the accuracy of the prediction model. Phenotypic and genotypic information are collected for the individuals in the validation population, but only genotypic information is used to calculate GEBV. Accuracy can be estimated using the correlation coefficient between the GEBV and true breeding value (TBV) of the validation population (Meuwissen et al. 2001). When applying the GS model created within the training population to the validation population, accuracy was greater when the relationship between the two populations was closer (Asoro et al. 2011). Therefore, in a training population containing multi-lines with greater genetic variation, LD needs to be consistent between the training and validation population so that allelic effects estimated in one population are predictive in the other (Lorenz et al. 2011). Therefore, higher marker densities are a requirement to keep LD consistent between training and validation populations in order to maintain accuracy of the prediction model (Hamblin et al. 2010; Meuwissen, 2009; Newell et al. 2011).

2.3.2 GS statistical models to estimate breeding values

Different statistical models with different assumptions have been used to estimate marker effects in genomic selection (reviewed by Lorenz et al. 2011). With high throughput genotyping technology, the number of markers (p) obtained for GS is much higher than the number of phenotypic observations (n). This is known as the "large p, small n problem" in model building. To solve this problem, different statistical strategies can be implemented, such as shrinkage methods, variable selection methods, and kernel and machine learning methods.

Ridge regression best linear unbiased prediction (rr-BLUP) (Meuwissen et al. 2001; Whittaker et al. 2000) is one of the shrinkage methods that assume marker effects are normally distributed with fixed common variance. All marker effects are included in the model and shrunken to the same degree towards zero, implying that the observed phenotype is controlled by many loci with small effects. Bayesian shrinkage regression methods (Calus et al. 2008; Meuwissen et al. 2001; Ter Braak et al. 2005; Xu, 2003) were developed to relax the constraint assumption of rr-BLUP. The BayesA method draws each marker effect from a normal distribution with its own variance, allowing each marker to be shrunken toward zero, but to different degrees. The BayesB method, one of the variable selection methods, accepts the probability that a marker might not have an effect at all and might reflect the underlying genetic structure in situations where genetic variance was found at only a few loci (Meuwissen et al. 2001). Reproducing kernel Hilbert spaces (RKHS) regression, a kernel method, has also been used in GS (Gianola et al. 2006; Scholkopf et al. 2004). A kernel function, combined with an additive genetic model, converts predictor variables to distances among observations to produce a matrix, which can be used in a linear model. Marker scores are treated as input data, and converted to input space, which is further converted into feature space by applying a kernel function.

There is no single best model for all species, populations and traits because of the differences of genetic architecture among them (Lorenz et al. 2011). For instance, rr-BLUP functions well when addressing traits controlled by many loci with small effects (Buckler et al. 2009; Lorenz et al. 2011). BayesB may be more suitable when several QTLs with larger effects are responsible for most of the genetic variation (Anderson et al. 2001; Munkvold et al. 2009; Lorenz et al. 2011). Other models, such as RKHS regression, could be considered when non-additive effects are revealed in some populations (Dudley and Johnson, 2009; Lorenz et al. 2011). Machine learning models such as Neural Networks (NN), Random Forest (RF) and Support Vector Machine (SVM) were also used in GS to avoid linear modelling from genotype to phenotype and were considered crucial for modelling more complex traits (Azodi et al. 2019).

2.3.3 Accuracy of GS predictions

To understand the accuracy of genetic estimated breeding values (GEBVs) and GS models, one must fully understand the quantitative genetic concepts and relationship among GEBVs, empirically estimated breeding values (EBVs), and true breeding values (TBVs) (Falconer and Mackay, 1996; reviewed by Lorenz et al. 2011).

The prediction accuracy of the GEBVs can be obtained by estimating the correlation r(GEBVs: EBVs) where GEBVs are the genetic estimated breeding values predicted by GS models using genotypic information and EBVs are empirically estimated breeding values, or, phenotypic values. The correlation between GEBVs and EBVs offers an estimate of selection accuracy (Falconer and Mackay, 1996).

The accuracy of GS can be calculated as the correlation r(GEBVs: TBVs) where GEBVs are genetic estimated breeding values and TBVs are true breeding values. Because there is no way to directly measure r(GEBVs: TBVs), r(GEBVs: EBVs) can be used to estimate the correlation between GEBVs and TBVs under the assumption that r(GEBVs: EBVs)=r(GEBVs: TBVs) × r(EBVs: TBVs) where GEBV=TBV+e1 and EBV=TBV+e2, where e1 and e2 are unrelated residuals. Therefore, some environments should be avoided when collecting phenotypic data for training and validation populations because one common error will be generated from genotype by environment (G × E) interaction in both GEBVs and EBVs.

When the above assumption holds, the accuracy of GS models, r(GEBVs: TBVs), can be calculated as the proportion r(GEBVs: EBVs)/r(EBVs: TBVs). r(GEBVs: EBVs), as mentioned above, is the correlation between GEBVs and phenotypic values. r(EBVs: TBVs) is equal to the square root of the heritability (h) within the validation set (Falconer and Mackay, 1996).

GS accuracies of 0.73-0.84 using simulation were achieved in a training population size of 2,200 individuals containing 6 QTLs with the assumption of additive gene action and 0.5 heritability (Meuwissen et al. 2001). Simulation accuracies of 0.64-0.69 were obtained using a training population of 1,000 individuals with 13 QTLs under the same assumptions (Habier et al. 2007). Similar simulation accuracies of 0.61-0.62 were achieved using 1,040 markers and a training population containing 500 lines derived by randomly mating 42 two-row barley founders under the assumption of additive gene action and 0.4 heritability (Zhong et al. 2009).

GS prediction accuracies of 0.44-0.79 using empirical data were discovered for different traits with heritability ranging from 0.04-0.50 in a training population consisting of 3,567 Holstein bulls and genotyped with 28,416 SNP markers (VanRaden et al. 2009). Prediction accuracies of 0.3-0.8 were found in other empirical GS studies in cattle (Habier et al. 2010; Hayes et al. 2009; Luan et al.

2009; Moser et al. 2009; Su et al. 2010; Verbyla et al. 2009). A GS study using 446 elite North American oat lines was conducted and accuracies of 0.14-0.55 was observed for five different traits (Asoro et al. 2011). Compared to traditional marker-assisted selection, an average of 28% higher accuracy for 13 traits was achieved using GS prediction in a winter wheat study, with higher accuracies found in traits with higher heritability (Heffner et al. 2011a).

2.3.4 Factors affecting the accuracy of prediction models

The number of markers, population size, heritability, linkage disequilibrium and the presence of population structure (LD changes among populations) are known to influence the accuracy of GS models (e.g. Asoro et al. 2011; Calus, 2010; Hamblin et al. 2011; Heffner et al. 2011b; Jannink et al. 2010; Meuwissen et al. 2001; Moser et al. 2010; Lorenzana and Bernardo, 2009; Luan et al. 2009; Solberg et al. 2008; VanRaden et al. 2009; Wientjes et al. 2013).

2.3.4.1 Marker density effect on the accuracy of prediction models

In general, lower marker densities result in lower GS accuracy (Moser et al. 2010). In a winter wheat study, decreasing marker number from 1,158 to 384 caused a slight decrease in GS prediction while decreasing to 192 led to a dramatic 10% accuracy decrease (Heffner et al. 2011a). When marker number was increased from 300 to 900, no plateau of GS accuracy was observed in four of five traits studied in oat, suggesting that more than 900 markers would be useful (Asoro et al. 2011). For the remaining trait, little accuracy gain was found when increasing marker number from 600 to 900 (Asoro et al. 2011). In a study conducted by Combs and Bernardo (2013), increasing marker number from low to moderately high density increased the prediction accuracy and plateaued at moderately high density. In several empirical GS studies in plant species, marker number showed little effect on prediction accuracies until the density became very low (Lorenzana and Bernardo, 2009). In GS cattle studies, masking as many as 75% of the original markers affected GEBV accuracy minimally (Luan et al. 2009; VanRaden et al. 2009). Solberg et al. (2008) studied the effect of prediction accuracy when marker densities was scaled to the effective population size. In their study, a density increase from 0.25Ne/morgan to 2Ne/morgan led an increase of accuracy from 0.63 to 0.83, where 1 Ne marker/morgan means 100 markers per morgan if the effective size (Ne) is 100. A similar accuracy increase was found when SNP marker density was increased from 1Ne/morgan to 8Ne/morgan. To conclude, increasing marker density would increase GS accuracy until marker saturation occurs in the population (Combs and Bernardo, 2013).

2.3.4.2 Training population size effect on the accuracy of prediction models

Compared to marker density, training population size had a larger impact on the prediction accuracy in plant studies compared to animal studies (Asoro et al. 2011; Heffner et al. 2011a; Lorenzana and Bernardo, 2009; Luan et al. 2009; VanRaden et al. 2009). A linear relationship was found between training population size and prediction accuracy in bulls (VanRaden et al. 2009). A 20% accuracy increase was observed when increasing the training population size from 48 to 332 in an Arabidopsis study, with the biggest increase (10%) occurring when training population size increased from 48 to 96 (Lorenzana and Bernardo, 2009). Similar accuracy increase was found when training population size increased from 100 to 300 for five different traits in oat (Asoro et al. 2011). Increasing the training population size from 96 to 198 led to a 11% prediction accuracy increase, while increasing to 288 gave rise to a 30% accuracy increase (Heffner et al. 2011a).

In the study of Wientjes et al. (2013), the effect of training population size on linkage disequilibrium and family relationship was revealed. A smaller training population size led to a larger family relationship impact on prediction accuracy, while larger training population size led to a larger LD impact on prediction accuracy.

2.3.4.3 Heritability effect on the accuracy of prediction models

More accurate predictions were obtained when the trait of interest exhibits higher heritability (Moser et al. 2010; Heffner et al. 2011a). Heritability and QTL number determine the adequate marker density and training population size for GS studies. Five-fold greater marker density was needed to obtain equivalent accuracies when comparing traits controlled by 10 versus 1,000 loci (Hayes et al. 2009). Research showed that when dealing with low heritability traits, sufficiently large training population size is needed to obtain prediction accuracy (Combs and Bernardo, 2013; Lorenz et al. 2009; Solberg et al. 2008).

2.3.4.4 Linkage disequilibrium and population structure effect on the accuracy of prediction models

GS assumes the same structure exists in the training population, validation population and the population from which the tested individuals are derived. The consistency of LD between markers and QTLs in the training population and the validation population (or the tested individuals) was important to maintaining the prediction accuracy (de Roos et al. 2009). Poor accuracy was expected if the training population and validation population (or tested individuals) were distantly related (de Roos et al. 2009). Better accuracy was found when close genetic relationship existed between tested individuals and the training population (Calus, 2010; Clark et al. 2012; Solberg et al. 2008; VanRaden et al. 2009; Wientjes et al. 2013). To maintain good prediction accuracy, a set of diverse genotypes and phenotypes needs to be included in the training population (Calus, 2010).

2.3.5 Genomic selection with environmental information

Genomic selection was adopted in plant breeding slowly as the initial GS models failed to capture the realities observed during the plant breeding process, that is, a genotype might have different phenotypes in different locations and years. Multiple environment testing offers more accurate estimates of genetic effect and understanding of genotype-by-environment (G × E) interaction, a crucial component in understanding genotype stability and cross-over interaction. In earlier GS studies involving crops, replicated line means obtained from multiple environments were used for genomic prediction (Crossa et al. 2015; Lopez-Cruz et al. 2015). However, traits with higher G × E interaction have poorer prediction accuracy when compared to traits that are stable across multiple environments (Heffner et al. 2011a). When incorporating spatial variation as a covariate using moving-means, the accuracy of GS was increased using rr-BLUP (Lado et al. 2013). Different strategies to include G × E in GS models have been investigated and resulted in increased prediction accuracy in different crops (Cuevas et al. 2017; Heslot et al. 2014; Jarquín et al. 2014; Jarquín et al. 2017; Pérez-Rodríguez et al. 2015; Sukumaran et al. 2017; Technow et al. 2015). Heslot et al. (2014) integrated environmental co-variates and a crop growth model into genomic prediction models in an attempt to improve accuracy. In their study, 2,437 elite winter wheat lines were gown in 44 environments over 6 years. Environment was characterized using daily weather data and incorporated in prediction models. Performance was predicted using environment effect, genotype effect and sensitivity of each genotype to a stress covariate generated from weather data using a crop model. Increased accuracy (11%) was observed in predictions in unobserved environments when weather data were available. In another simulation study, a crop growth model which included plant population, daily temperature and solar radiation, and several genetically controlled physiological traits including total leaf number, area of largest leaf, solar radiation use efficiency and thermal units to maturity, was directly added into the estimation of whole genome marker effects in whole genome prediction using Approximate Bayesian Computation to improve prediction accuracy (Technow et al. 2015). Another means to incorporate G × E interaction is to model marker by environment (M × E) interactions in GS models. Prediction accuracy was reported as being higher in M × E GS models when compared to cross-environment models (Crossa et al. 2015; Lopez-Cruz et al. 2015). Recent research also indicated that reaction norm models could be used to include G × E interaction in GS models to increase prediction accuracy (Jarquín et al. 2014; Jarquín et al. 2017; Pérez-Rodríguez et al. 2015; Sukumaran et al. 2017). For example, environment covariances (ECs) were included as covariance functions, along with molecular markers and their interaction, in reaction norm models in GS and increased accuracy was reported (Jarquín et al. 2014). Jarquín et al. (2014) incorporated 68 ECs derived from five wheat phenology phases grown in northern France into GS models to model interaction, which explained 16% of within-environment yield variance. A 17% to 34% increase was discovered in comparison to models which only used main effects. A similar study by the same group investigated more complex models that included line effects, environment effects, genotypic effects, site effects, and interactions between them (Jarquín et al. 2017). These reaction norm models which included interactions provided a 16% to 82% increase in prediction accuracy. Other information, such as pedigree, has be included in reaction norm models and increased predictions were observed (Pérez-Rodríguez et al. 2015; Sukumaran et al. 2017).

CHAPTER 3. GENOMIC SELECTION FOR NORTH AMERICAN OAT BREEDING USING EMPIRICLE DATA

3.1 Introduction

Oat (*Avena sativa* L.) is an important world cereal crop that has been seeded annually on 9.6 million hectares with an annual production of 23.4 million metric tons over the past five years (FAOSTAT, 2019). It is considered a healthy cereal due to several nutritional compounds found within the grain, including β-glucan. β-glucan is a soluble fiber that has been shown to lower plasma cholesterol and reduce the risk of heart disease (Queenan et al. 2007; Liatis et al. 2009). This has resulted in health claims being established in both Canada (Health Canada, 2010) and the United States (U.S. FDA, 1997). Oat grain also contains a number of antioxidant compounds, including the polyphenolic avenanthramides, which have anti-inflammatory effects that may protect against coronary heart disease (Meydani, 2009). Oat contains 12-20% protein which is rich in globulins and contains more lysine and threonine than other cereals, providing a better balance of essential amino acids (Klose and Arendt 2012). Finally, oat can be consumed by most people suffering from celiac disease and is thus considered to be gluten-free (Peraaho et al. 2004).

Oat remains an attractive option for Canadian producers because of its low input costs compared to other crops, such as canola or corn, and its high demand, including new uses in products such as oat milk or as a source of plant-based protein. Canada also has one of the highest oat yielding countries in the world, exceeding 3,000 kg ha-1 (FAOSTAT, 2018), as a result of improved agronomic practices and the release of improved cultivars through oat breeding. As a result of these desirable attributes, oat has been seeded on an average of 3.3 million acres over the past five years with exports of 3.3 million metric tonnes (Statistics Canada, 2020) worth over $1B CND in 2020 from sales of raw oat and milled oat products (Strychar, 2021).

Identifying new breeding strategies that can improve breeding efficiency in oat is important for future progress in this crop. Most oat breeding programs generally follow a standard process involving the hybridization of 2-4 parents to generate genetic variability which is followed by inbred line development (with or without selection) at which point single-plant selection for higher heritability traits (e.g. crown rust resistance, plant height) is performed. Subsequent generations are grown in replicated, multi-location trials over several years to evaluate lower heritability traits like yield. While this process performs well, it suffers from two significant drawbacks. Firstly, the loss of genetic variability at earlier generations when only high heritability traits are selected based on visual evaluation of single plants limits genetic variability remaining for subsequent selection of lower heritability traits, and secondly, an inability to select for low heritability traits (i.e. yield) at earlier generations because of inadequate amounts of seed needed for replicated, multiple environment testing. Molecular breeding is seen as a means to alleviate these two issues.

Molecular breeding, in the form of marker-assisted selection (MAS), has been used complementarily with traditional phenotypic selection to improve simply inherited traits. This has worked well in oat for marker-tagged traits, such as adult plant crown rust resistance (Lin et al. 2014), seedling crown rust resistance (McCartney et al. 2011; Gnanesh et al. 2015; Zhao et al. 2020) and stem rust (Kebede et al. 2020). However, MAS was unsuitable for low heritability traits and thus an alternative strategy, termed genomic selection (GS), was developed to assist selection for these traits (Meuwissen et al. 2001). The fundamental basis of GS relies on utilizing genome-wide markers and integrating them into prediction models that attempt to capture all alleles, including those with minor effects (Meuwissen et al. 2001). Genomic selection not only allows selection to be done on single plants at early generations when limited seed prevents testing in replicated environments, it also attempts to prevent random sampling of genetic variability for low heritability traits at these early generations, but instead selected towards a desired direction which increases genetic gain. As a result, the breeding cycle should be theoretically shortened significantly, and resources redirected from the field to large scale single plant genotyping in the lab. Numerous simulation and empirical studies have demonstrated the potential of GS based on GEBVs in animal and plant breeding scenarios (Ankamah-Yeboah et al. 2020; Haile et al, 2020; Juliana et al. 2020; Lorenzana and Bernardo, 2009; Lozada et al. 2019; Meuwissen et al. 2001; Storlie and Charmet, 2013; Tessema et al. 2020).

Phenotypic data evaluation, genotyping methods, marker density, GS models, population size, training population size and heritability of traits are considered important non-environment related factors affecting prediction of genomic estimate breeding values (GEBVs) within breeding populations (Meuwissen et al. 2001; Lorenzana and Bernardo, 2009; Luan et al. 2009; Solberg et al. 2008; VanRaden et al. 2009; Calus, 2010; Moser et al. 2010; Heffner et al. 2011b; Asoro et al. 2011; Hamblin et al. 2011; Jannink et al. 2010; Combs and Bernardo, 2013; Wientjes et al. 2013). Additionally, precision in phenotypic data collection is considered a key component in training a GS model (Cabrera-Bosquet et al. 2012). Different genotyping methods will provide different marker numbers and affect marker density in GS models. In empirical and simulation studies, the decrease in accuracy was small and noticeable when SNP numbers decreased to a few hundred (Asoro *et al.* 2011; Tsai et al. 2016; Correa et al. 2017; Gutierrez et al. 2018; Robledo et al. 2018; Vallejo et al. 2018; Palaiokostas et al. 2019). For instance, decreasing marker numbers from 900 to 300 led to small decreases in prediction accuracy for five different traits (decrease <0.1) in 446 oat lines (Asoro *et al.* 2011). Parametric and non-parametric GS models capture different genetic effects and are more or less suitable for different traits depending on their genetic architecture (Buckler et al. 2009; Lorenz et al. 2011; Anderson et al. 2001; Munkvold et al. 2009; Dudley and Johnson, 2009; Azodi et al. 2019). Parametric model rr-BLUP considers all markers with a common variance and might be used for complicated traits controlled by a large number of minor effect QTLs (Endelman 2011). Bayesian models tend to be sensitive to the number of QLTs and could provide more accuracy with qualitative traits (Habier et al. 2007; Daetwyler et al. 2010; Wang et al. 2015). Semi-parametric and non-parametric models offered additional modelling of non-additive effects for complicated traits, especially when the population was small (Gianola et al. 2006; González-Camacho et al. 2018; Azodi et al. 2019; Merrick and Carter, 2021). After comparing 17 GS models, Merrick and Carter concluded that there was not much difference among parametric and non-parametric GS models for predicting seedling emergence of wheat and recommended the use of rr-BLUP or c-BLUP when dealing with complex traits with unknown architecture. Finally, increasing training population size has been shown to increase prediction accuracy in numerous GS studies (Asoro et al. 2011; Meuwissen et al. 2001; Saatchi et al. 2010).

The objective of this study was to evaluate numerous critical factors, including phenotypic data, genotyping methods, marker density, GS models, population size, training population size and trait heritability, that are important for the creation of a GS model. An unbalanced data set of oat lines grown in the prairie provinces of western Canada over a thirteen-year period was used as the data source. This dataset was selected as it is typical of that generated in breeding programs.

Understanding how various factors impact the ability to accurately predict GEBV will help oat breeders create more effective and efficient GS models that should improve genetic gain.

3.2 Materials and Methods

3.2.1 Phenotypic data

Phenotypic data was obtained for oat lines grown in the Western Cooperative Oat Registration Trial (WCORT) from 2002-2014. Thirty-three advanced lines (along with 3 checks) from different breeding programs in North America were grown each year at 9-14 locations. The test was randomized and planted as a 6 × 6 lattice with three replications. Test locations were located within the four soil zones of western Canada: zone 1 (Black Soil), zone 2 (Black and Gray Soils), zone 3 (Brown Soil) and zone 4 (Irrigated). The twelve traits assessed were yield, days to heading, days to maturity, plant height, test weight, thousand kernel weight, grain plumpness, grain thinness, groat percentage, grain protein content, grain oil content, and grain β-glucan content. Descriptions for these traits are provided in Table 3.1. Genetic correlations between traits were analyzed used Meta-R software (v. 6.0.4) (Gregorio et al. 2015).

Properly estimating the phenotypic value based on observed values is crucial in GS, especially when dealing with unbalanced field data. Best linear unbiased prediction (BLUP) estimation using mixed models results in a smaller mean squared error and is thought to provide a more accurate representation of individual performance values across different environments (Piepho et al. 2008). Best linear unbiased predictions were thus calculated for each trait. The following mixed model was used (Equation 3.1):

$$Y_{ijk}=\mu+g_i+l_j+y_k+g{:}l_{ij}+g{:}y_{ik}+l{:}y_{jk}+\varepsilon_{ijk} \qquad \text{(Equation 3.1)}$$

where Y_{ijk} is the value of the phenotypic trait for the ith genotype (g_i) evaluated in the jth location (l_j), during the kth year (y_k); μ represents the intercept; g_i, l_j, and y_k are the effects of genotype, location and year, respectively; $g{:}l_{ij}$, $g{:}y_{ik}$, $l{:}y_{jk}$ are the interactions between genotype and location, genotype and year, and location and year, respectively; ε_{ijk} represents the residuals. The lme4 package (Bates et al. 2012) in R (v. 4.0.1) was used to fit the mixed model (R Core Team, 2020). The lmer function within the lme4 package was used for model fitting and the ranef function was used to call BLUPs for each line. Phenotypic means of each line were also calculated by using averages of each line reported in individual years.

Table 3. 1 Description of the twelve traits measured on the 305 oat lines grown in the WCORT from 2002-2014.

Trait	Units	Description
Yield (YLD)	Kg/Ha	Grain weight adjusted to 12% moisture
Days to heading (HD)	days	Duration from sowing to time of 50% panicle emergence
Days to maturity (MAT)	days	Duration from sowing to 50% peduncle color loss
Plant height (HT)	cm	Height of entire plant measured from soil level to top of the panicle
Test weight (TWT)	Kg/Hl	The volume of grain determined using a Cox funnel
Thousand kernel weight (MKW)	g	Weight of a thousand kernels
Plumpness (PLP)	%	The percentage of grain remaining above a 2.2 mm x 190 mm slotted sieve
Thins (THN)	%	The percentage of grain falling below a 2.0 mm x 190 mm slotted sieve
Groat (GRT)	%	The percentage of groat remaining in a 50 g sample after removal of the hull from intact oat seed using a Codema Dehuller
Protein (PRO)	%	Analyzed by Combustion Nitrogen Analysis on a LECO Model FP-428 CAN analyzer and reported on a dry weight basis
Oil (OIL)	%	Analyzed by AOAC Method 922.06 and reported on a dry weight basis
β-glucan (BG)	%	Analyzed by AACC Method 32-23 and reported on a dry weight basis

3.2.2 Genotypic data

Genomic DNA of lines was extracted using the Qiagen DNeasy Plant Mini Kit from 5-day coleoptiles. A total of 305 lines grown in the WCORT from 2002-2014 were genotyped using the iSelect oat 6K SNP Assay (Tinker et al. 2014). SNPs that displayed 2-3 clusters, a minor allele frequency >0.05, and missing data <0.5 were called using GenomeStudio Diploid version.

Genotyping by sequencing was also conducted on the 305 WCORT lines. Library preparation used the double digest (*PstI-MspI*) GBS method (Huang et al. 2014). Haplotype derived SNPs were discovered using the method of Bekele et al (2018). Missing genotypic markers were replaced with numeric population mean of the markers using the a.mat function in the rr-BLUP package.

3.2.3 Calculation of heritability

When lines were grown in the same environments in replicates in a given year, trait heritability in a single year was calculated across environments (e) and replicates (r) using Equation 3.2:

$$h^2 = \frac{\sigma^2{}_g}{\sigma^2{}_g + \frac{\sigma^2{}_{ge}}{e} + \frac{\sigma^2{}_e}{er}} \qquad \text{(Equation 3.2)}$$

where $\sigma^2{}_g$, $\sigma^2{}_{ge}$ and $\sigma^2{}_e$ are the genotypic (additive), genotypic × environment, and residual variances, respectively, e is the number of environments in which the lines are tested each year, and r is the number of replicates per environment. As replicated data was not available for all years or all the traits, only plant height, days to heading, days to maturity and yield in 2003, 2004, 2005, 2008, and 2009 were calculated. The purpose of this analysis was to offer a comparison to the heritability calculated in an unbalanced dataset where genotypes changed each year.

Trait heritability for lines grown in multiple environments using unbalanced data was estimated using Equation 3.3:

$$H_C{}^2 = 1 - \frac{\overline{V}\,\text{BLUP}}{2\sigma^2{}_g} \qquad \text{(Equation 3.3)}$$

where $\overline{V}$ BLUP is the mean variance of two different BLUPs and $\sigma^2{}_g$ is the genotypic variance (Cullis et al. 2006).

3.2.4 Population structure of WCORT lines

Genotypic data generated from 6K chip and GBS were used independently to understand population structure within the WCORT population. Missing genotypic markers were replaced with numeric population means for each missing marker. Genetic structure of WCORT lines was tested in Structure (v. 2.3.4; Pritchard et al. 2000) using a subset of 40 unlinked markers (>50cM). The number of subpopulations (K) was estimated from 2 to 9 with 20 replicates. Markov Chain Monte Carlo (MCMC) cycles were set to 100,000 with 10,000 cycles as a burn-in period. The optimum number for K was obtained using Structure Harvester (v. 0.6.94; Earl and vonHoldt, 2012) according to the highest ΔK value.

Principle component analysis (PCA) and a neighbor joining (NJ) phylogenetic tree were generated using the GAPIT package (v. 2.0) in R (Tang et al. 2016). Scree plots indicating the variance explained by each principal component (PC) were used to identify the "elbow" indicating the number of important PC. The linkage disequilibrium function in JMP Genomics was used to generate the square of correlation coefficients between pairwise markers. In order to estimate LD decay, a fitted curve for a nonlinear regression model was generated using Equation 3.4:

$$\mathrm{r}^2 = \frac{10+\mathrm{d}}{(2+\mathrm{d})(11+\mathrm{d})} \times \left[1 + \frac{(3+\mathrm{d})(12+12\mathrm{d}+\mathrm{d}^2)}{\mathrm{n}(2+\mathrm{d})(11+\mathrm{d})}\right] \qquad \text{(Equation 3.4)}$$

where d is the genetic distance between pairwise markers (in cM) and n is the population size (Hill and Weir, 1988). The critical value of r^2 was calculated using the 95th percentile of unlinked r^2 values (pairwise marker distance > 50cM).

3.2.5 Genomic selection models

Six different GS models were tested, including the ridge-regression best linear unbiased prediction (rr-BLUP) (Bernardo and Yu, 2007; Endelman, 2011; Meuwissen et al. 2001) and Bayes Cπ (BCPi) (Habier et al. 2011) parametric models, the semi-parametric model reproducing kernel Hilbert Spaces regression (RHSK) and the support vector machine (SVM) (Long et al. 2011), neural networks (NN) and random forest (RF) (González-Recio and Forni, 2011) non-parametric machine learning models. The rr-BLUP, BCPi, RHSK, SVM, NN and RF models were implemented with the packages "rrBLUP" (v. 4.6.1) (Endelman, 2019), "BGLR" (v.1.0.8) (Perez and de los Campos, 2018), “brnn” (v.0.8) (Perez-Rodriguez and Gianola, 2020), “Kernlab” (v 0.9-29)(Karatzoglou et al. 2019), “brnn” (v.0.8) (Rodriguez and Gianola, 2020) and “randomForest” (v 4.6-14) (Breiman et al. 2018) in R, respectively (R Development Core Team, 2020). For BCPi, a total of 10,000 burn-ins and 50,000 saved Markov-Chain Monte Carlo (MCMC) iterations were used with pi set to 0.7. “rbfdot” kernel and “eps-svr” type were implemented in SVM, while 500 trees and 4 branches were tested for RF. For the remaining models, parameters were left at default settings in each R package. BCPi was not examined using GBS data due to the significant computation time required.

3.2.6 Cross validation

In order to test the accuracy of the GS models, 5-fold cross validation was used. In each iteration, the population was divided into 5 folds, 4 of which were used as training sets and one as a validation set. In the training set, genotypic and phenotypic data were used to estimate marker effects, which was subsequently used to calculate the GEBV for lines in the validation set using genotypic data only. GEBV and empirical phenotypic values were used to calculate correlation coefficients which represented the prediction accuracy. Once each subset was used as a validation set this represented one iteration. A total of 100 iterations were run and the mean of correlation coefficients was calculated and reported as prediction accuracy.

3.2.7 Marker density and training population size

The effect of marker density was examined on GS prediction accuracy by randomly sampling 100, 200, 300, 400, 700, 1400, 2500 SNPs from the filtered 6K SNP dataset. Similarly, 100, 500, 1000, 10,000, 25,000, 48,000 SNPs were randomly selected for GBS dataset. Five-fold cross validation was conducted using the “rrBLUP” GS model. To avoid sampling bias for markers, 100 random subsampling processes were done at each marker density and averaged to represent the prediction accuracy for each marker density.

To evaluate the training population size effect on GS prediction accuracy, 61, 122, 183, 244 lines were randomly drawn and used as the training population. All markers were included in the prediction models. To avoid sampling bias, 100 random sub-samples were done at each population size and averaged to represent the prediction accuracy for each training population size.

3.3 Results and Discussion

3.3.1 Analysis of phenotypic and genotypic data

Summary statistics and distributions based on the entry means for the 12 traits measured on the 305 WCORT lines are shown in Table 3.2 and Fig. 3.1a. Yield, days to maturity, height, protein, oil and β-glucan data were fairly symmetrically distributed as their skewness scores were between -0.5 and 0.5. Days to heading, days to maturity, height, protein, oil and β-glucan were fairly normally distributed, having kurtosis scores between -0.5 and 0.5. Due to the unbalanced nature of the dataset, BLUPs were estimated for the 305 WCORT lines. Trait distributions for BLUPs are shown in Fig.3.1b.

A total of 2,587 high quality SNPs were produced from the iSelect oat 6K SNP chip after the filtration process. Of these, 2,211 SNPs were distributed across all 21 linkage groups (Fig. 3.2a). A total of 51,411 SNPs were discovered by GBS with 12,864 located on the consensus map across the 21 linkage groups (Fig. 3.2b).

3.3.2 Heritability estimates for balanced and unbalanced data

Heritability estimates for yield, heading date, maturity, and plant height were calculated using balanced, replicated data from 2003, 2004, 2005, 2008 and 2009 (Table 3.3). Heritability for the other traits and the remaining years could not be obtained as replicated data were not available. Heritability for each trait varied from year to year (0.66-0.87 for height, 0.40-0.90 for days to heading, 0.58 to 0.70 for days to maturity and 0.44 to 0.76 for yield) reflecting changes due to different genotypes being tested and environmental conditions in these years. Heritability, being the proportion of variation contributed by genetic factors, is a dynamic concept that is not a constant in a population or across populations (reviewed by Visscher et al. 2008). Environmental change, genetic variation change from changing individuals in populations and the method of measurements and calculation could all contribute to different heritability estimates (Piepho, 2008; Visscher et al. 2008; Schimidt et al. 2019). Heritability estimates for all 12 traits using the unbalanced data set containing 305 lines over 12 years is listed in Table 3.4. Days to maturity had the lowest heritability value (0.40) within the unbalanced data set, suggesting the trait was most influenced by the environment. This value was also lower than the estimate obtained using individual years in the balanced data set which indicates a large portion of variance was due to year-to-year differences. Yield, days to heading, test weight, plant height, thins, groat percentage, plumpness and protein had moderate heritability (from 0.51 to 0.80). Kernel weight, β-glucan and oil were very heritable (above 0.89), implying less environmental influence on these traits. Heritability estimates have been shown to vary based on factors such as the population, environment and calculation method used (Pixley 1990; Chauhan 2018; Yan 2021). For instance, heritability of oat test weight ranged from 0.39 to 0.97 (Pawlisch and Shands 1962; Wesenberg and Shands 1973; Branson and Frey, 1989; McFerson 1987; Smith 1988; Pixley 1990; Holland and Munkvold 2001). Yan (2020) reported a wide range (0.36 to 0.95) of yield heritability when estimated in different years, areas, and replicates. Heritability estimates of β-glucan content and groat percentage were low in two oat crosses (<0.45) (Humphreys and Mather 1996) while heritability of heading date also differed in two oat crosses (0.68 vs. 0.46) (Mazurkievicz et al. 2019).

Table 3. 2 Summary statistics calculated from entry means for the 12 traits measured on the 305 oat lines grown in the WCORT from 2002-2014.

	YLD (Kg/Hl)	HD (days)	MAT (days)	HT (cm)	TWT (Kg/Hl)	MKW (g)	PLP (%)	THN (%)	GRT (%)	PRO (%)	OIL (%)	BG (%)
N	4403	2400	3771	4070	4228	4228	2397	2397	2393	2362	2362	2080
Mean	6120.9	59.1	96.6	100.4	50.8	36.6	82.1	5.0	69.8	15.3	7.7	5.0
SD	2075.2	6.2	10.1	15.7	4.0	6.1	14.1	5.4	5.3	2.1	1.1	0.6
Median	5963.2	58.0	96.4	100.3	51.1	36.2	86.8	3.2	70.6	15.1	7.8	5.0
Min	1590.1	45.5	72.0	54.0	30.8	18.5	15.3	0.1	45.3	9.6	4.9	3.4
Max	12683.9	80.0	132.0	143.9	62.9	74.2	99.3	56.6	83.6	24.6	12.6	7.6
Skewness	0.2	0.7	0.2	-0.1	-0.5	0.7	-1.8	2.8	-0.7	0.4	0.2	0.5
Kurtosis	-0.6	0.2	0.0	-0.4	0.7	1.8	3.4	11.6	0.6	0.3	0.5	0.1

Note: YLD=yield, HD=days to heading, MAT=days to maturity, HT=plant height, TWT=test weight, MKW=thousand kernel weight, PLP=plumpness, THN=thins, GRT=groat, PRO=protein, OIL=oil, BG= β-glucan

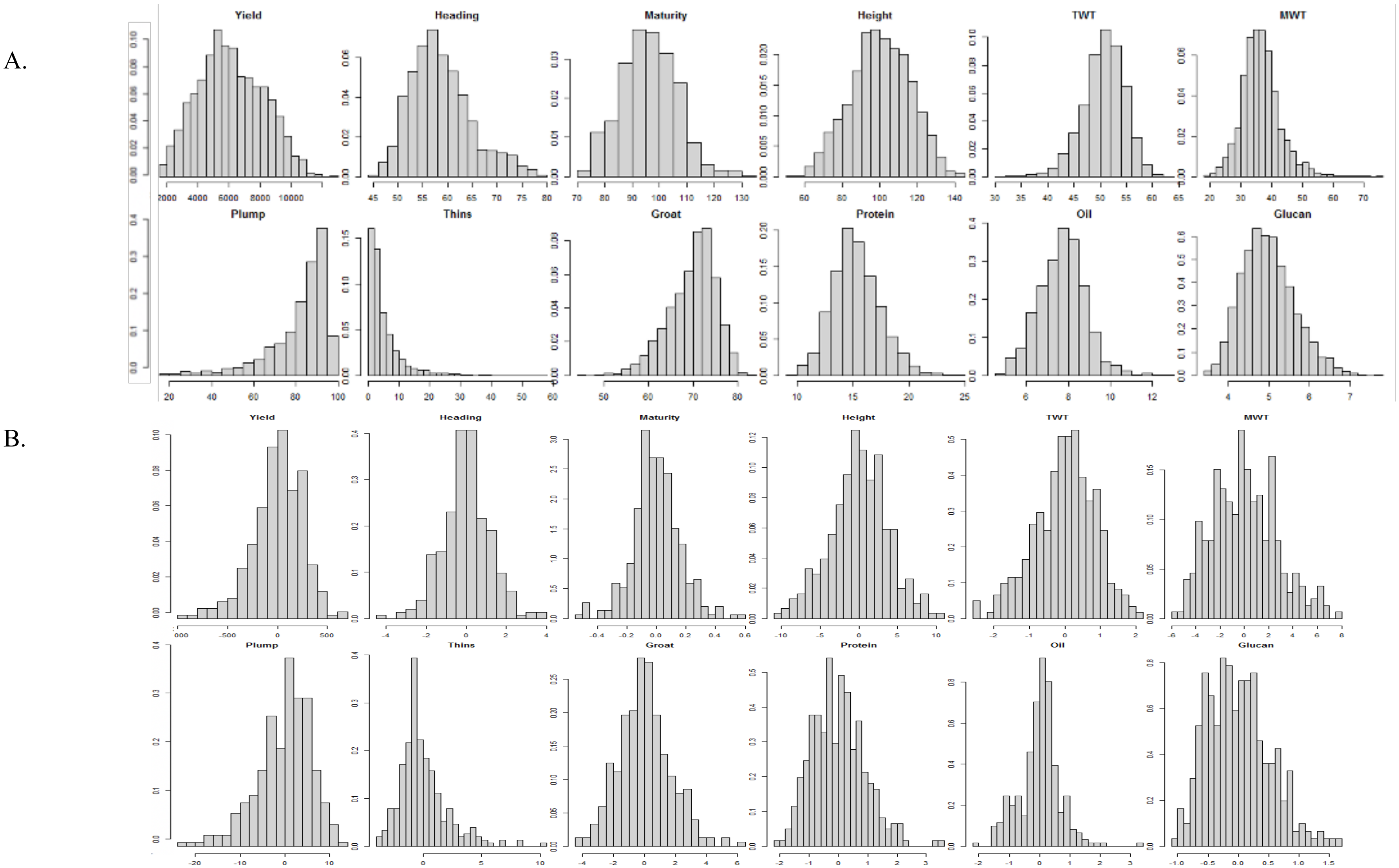


Fig. 3. 1 Distributions calculated for the 12 traits measured on the 305 oat lines grown in the WCORT from 2002-2014. Distributions are presented for entry means (A) and BLUPs (B).

Table 3. 3 Heritability estimates for height, days to heading, days to maturity and yield based on balanced data obtained on 36 lines grown in the WCORT in each of five different years.

Trait	**2003**	**2004**	**2005**	**2008**	**2009**
YLD	0.44	0.76	0.76	0.75	0.67
DH	0.68	0.40	0.40	0.90	0.79
MAT	0.70	0.60	0.60	0.70	0.58
HT	0.80	0.66	0.66	0.87	0.82

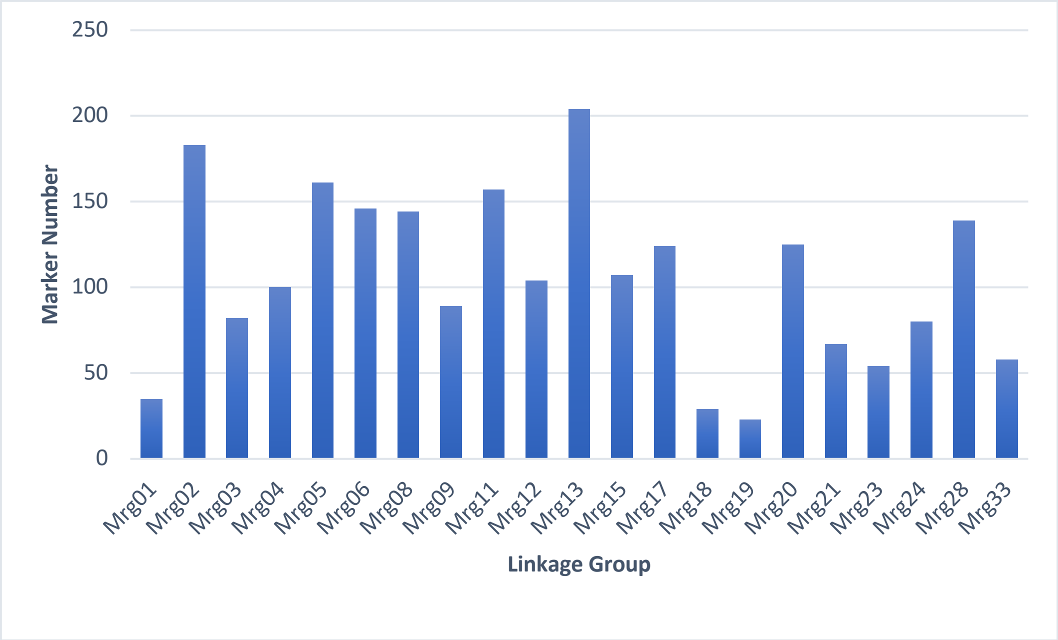
Note: YLD=yield, HD=days to heading, MAT=days to maturity, HT=plant height

Table 3.4 Heritability estimates for 12 traits based on unbalanced data obtained on 305 oat lines grown in the WCORT from 2002-2014.

Trait	**Heritability**
YLD	0.51
DH	0.60
MAT	0.40
HT	0.68
TWT	0.62
MKW	0.89
PLP	0.79
THN	0.71
GRT	0.71
PRO	0.80
OIL	0.95
BG	0.94

Note: YLD=yield, HD=days to heading, MAT=days to maturity, HT=plant height, TWT=test weight, MKW=thousand kernel weight, PLP=plumpness, THN=thins, GRT=groat, PRO=protein, OIL=oil, BG= β-glucan

A.



B.

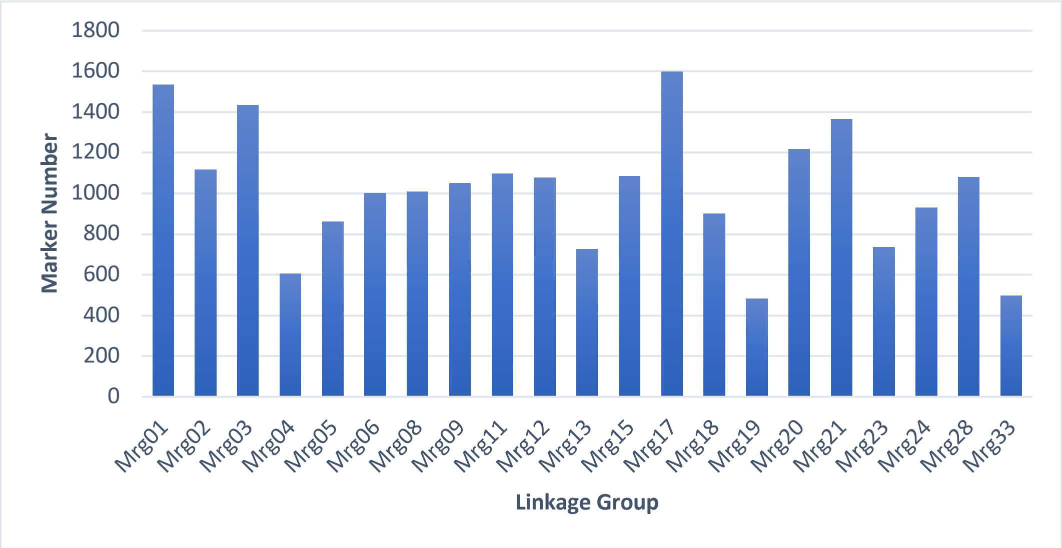


Fig. 3. 2 Distribution across the 21 oat linkage groups of the 2,211 6K (A) and 12,864 GBS (B) SNP markers determined to be segregating in the 305 oat lines grown in the WCORT from 2002-2014. Mrg: merge group.

3.3.3 Population structure of WCORT lines

The optimal number of subpopulations was estimated as three according to the value of ΔK calculated in Structure (Fig. 3.3). Scree plots which displayed the variance explained by each principal component (PC) indicated the "elbow" occurred at the third PC for both marker data sets. These three PCs explained 21% of the genetic variance in the WCORT population using genetic data generated from the 6K SNPs, comparable to the 21.5% explained variance from the GBS SNPs (Figs. 3.4a and 3.5a). Mild population structure was discovered using data from both genotyping methods. No clear separation of subpopulations based on common ancestors among the WCORT lines was apparent (Figs. 3.4b and 3.5b). Although the number of subpopulations was the same for both genotyping methods, individuals within each subpopulation changed (Figs. 3.4b and 3.5b). More distinctive separation among the subpopulations was found with the GBS dataset, likely due to more comprehensive coverage of the genome with the larger number of markers. The neighbor joining method was used to create a phylogenetic tree (Figs. 3.4c and 3.5c). Similar to the PCA plots, individuals within groups changed based on the marker dataset. For instance, CDC Dancer and AC Morgan were placed in the same subgroup, as were Leggett and Ronald, using the 6K SNP dataset. However, the GBS dataset placed AC Morgan and Leggett in the same subgroup while CDC Dancer and Ronald were in distinct groups. Population structure is specific to a given population, being dependent upon the variation present in the population. In the study of 2,043 commercial spring oat breeding lines from Finland, the first two PC explained 8.8% and 4.5% of the total variation in the population, suggesting no clear clustering among the breeding lines composing the population (Haikka et al. 2019). By contrast, population structure was more apparent a population of 759 historic oat cultivars and landraces, with 73% of the lines fitting within one of distinct three clusters, and first three principal components accounting for 38.8% of the variation (Winkler et al. 2016). In this current study, only mild population structure was discovered. This is due to two main causes. First, a large amount of genetic admixture with the WCORT population exists due to the common practice of exchanging germplasm between the breeding programs that contributed lines to the population. Secondly, the need for lines in the WCORT to be adapted to the western Canadian environment and meet a defined set of quality and disease traits would also tend to push the germplasm towards a more uniform genetic structure.

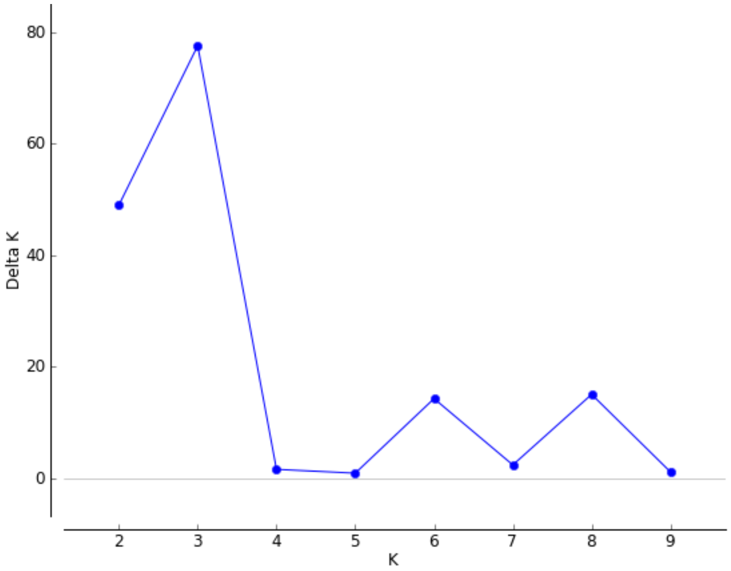


Fig. 3. 3 Evaluation of population structure in the 305 lines composing the WCORT population. Sub-populations (K) from 2-9 were evaluated. The ΔK peak value at K=3 suggested the presence of three subpopulations.

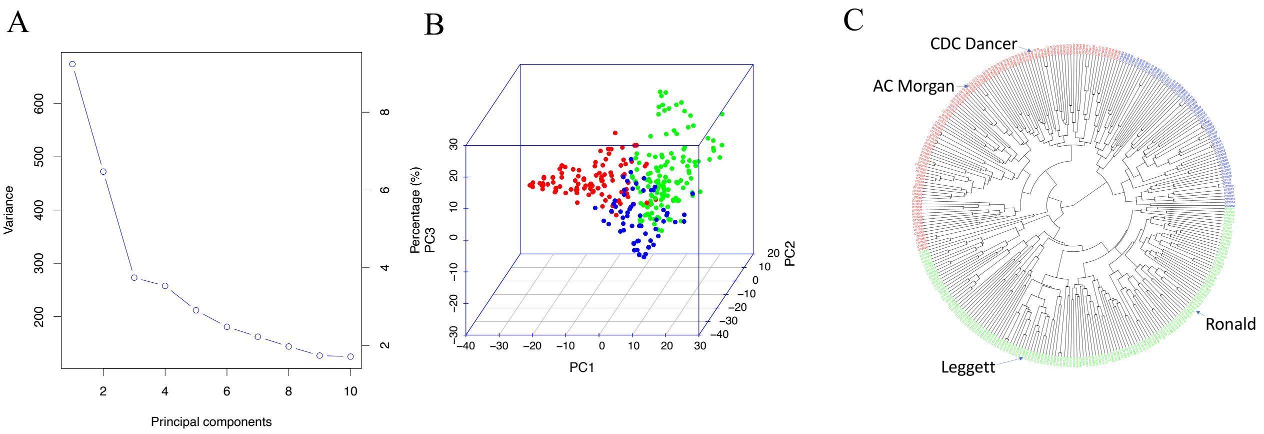


Fig. 3. 4 Evaluation of population structure within the 305 lines grown in the WCORT using 2,587 SNPs from the 6K chip. Variance explained by principal components (A), distribution of lines with the three subpopulations determined by PCA analysis (B) and a phylogenetic tree created by the neighbor-joining method (C).

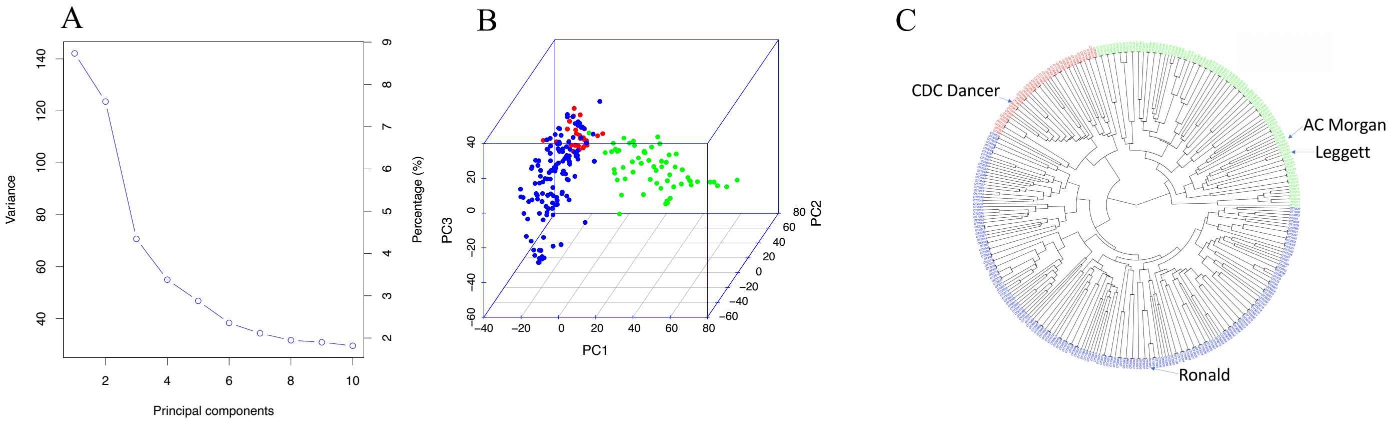


Fig. 3. 5 Evaluation of population structure within the 305 lines grown in the WCORT using 12,864 GBS SNPs. Variance explained by principal components (A), distribution of lines with the three subpopulations determined by PCA analysis (B) and a phylogenetic tree created by the neighbor-joining method (C).

LD decay plots for the WCORT lines was estimated as the intersection between the critical value line and the non-linear regression line (Figs. 3.6a and 3.6b). LD decay was estimated as 12.1 cM using the 6K SNP dataset and 10.0 cM using the GBS SNP dataset. It was interesting to note that both data sets provided fairly similar estimates despite the much larger number of GBS SNPS used. The 6K SNP dataset appeared to adequately capture the underlying genetic structure of the population and the extra markers in the GBS SNP dataset provided largely redundant information. In either case, such extensive LD is typical of elite breeding germplasm that have been selected for adaptation to specific production regions and for specific disease and quality traits. In a genome-wide association study (GWAS) using oat germplasm consisting of 1,205 oat lines, much more rapid LD decay was discovered (1.0 cM at r^2 =0.2) (Newell et al. 2011). The composition of the two study populations might explain the difference in LD decay. The WCORT lines in this study are advanced breeding lines that possess a narrower genetic variation, for the reasons mentioned before, which would thus possess larger LD blocks than those in the GWAS population studies (Newell et al. 2011; Canales et al. 2021).

3.3.4 Phenotypic datasets, genotypic datasets and GS model effects on GS prediction accuracy

Mean of lines across different environments and BLUPs were compared as phenotypic datasets with respect to their impact on GS prediction accuracy. In general, GS models using the phenotypic dataset calculated as BLUPs gave better predictions than the dataset based on line means across environments (Figs. 3.7a vs 3.7b and Figs. 3.7c vs 3.7d). Among the 12 traits, the difference between BLUP and mean-based prediction accuracies varied, but predictions were higher for all traits using BLUPs, except for plant height and test weight, both of which displayed minimal difference between the two phenotype datasets. BLUPs have been reported to be a better phenotypic representation to reflect true breeding values of individual lines, especially when phenotypes were unbalanced (Piepho et al. 2008; Molenaar et al. 2018).

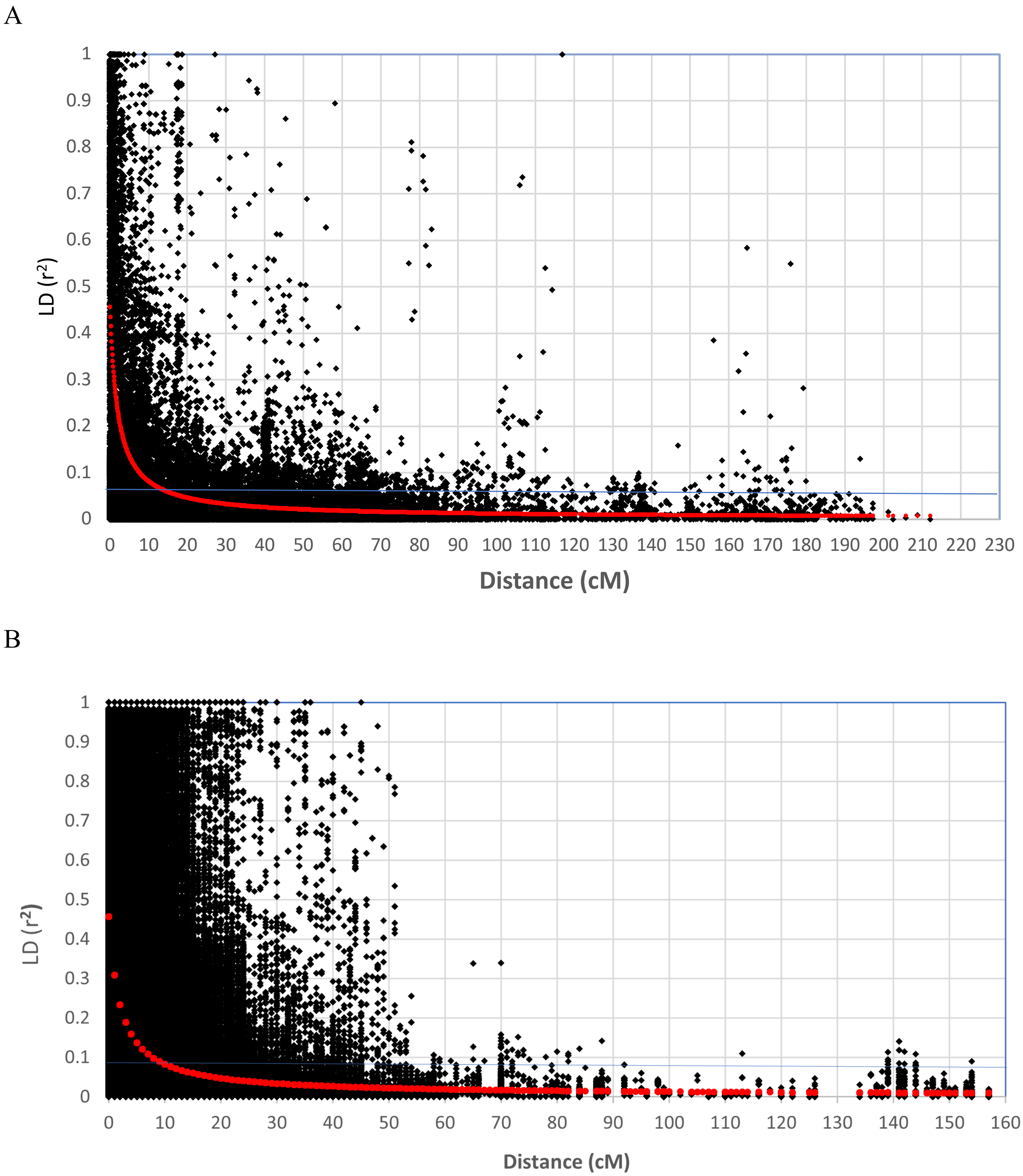


Fig. 3. 6 Genomic-wide average LD decay estimated using the 6K SNP dataset (A) and the GBS SNP dataset (B). The fitted non-linear regression line (red) indicated LD decay whereas the critical value line (blue) describes the 95th percentile of unlinked r^2 values.

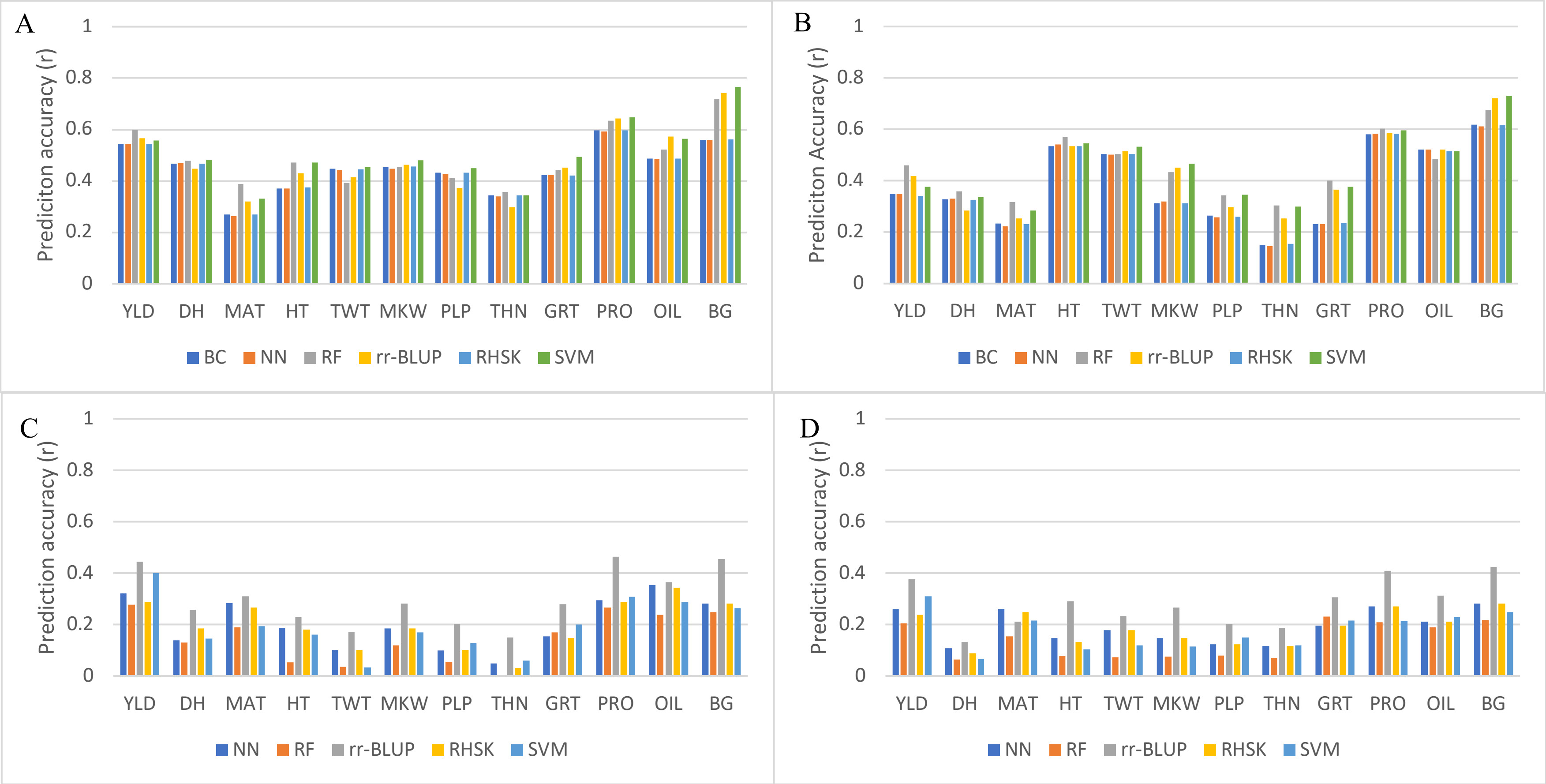


Fig. 3. 7 Genomic selection model prediction accuracy in the WCORT population using different models, phenotypic data sets and marker data sets for 12 traits. Comparison of GS models using the BLUP phenotype dataset and 6K SNPs (A), the mean phenotype dataset and 6K SNPs (B), the BLUP phenotype dataset and GBS SNPs (C) and mean phenotype dataset and GBS SNPs (D). Note: YLD=yield, HD=days to heading, MAT=days to maturity, HT=plant height, TWT=test weight, MKW=thousand kernel weight, PLP=plumpness, THN=thins, GRT=groat, PRO=protein, OIL=oil, BG= β-glucan, BC=Bayes Cpi, NN=neural networks, RF=random forest, rr-BLUP=ridge regression-best linear unbiased prediction; RHSK= Reproducing kernel Hilbert space, SVM=support vector machine.

Genomic selection models based on the 6K SNP dataset provided better predictions than GS models based on the GBS SNP dataset (Figs. 3.7a vs 3.7c and Figs. 3.7b vs 3.7d). Despite the higher density and better genome coverage provided by the GBS dataset, the prediction accuracies decreased dramatically for all traits, implying there was perhaps an oversaturation of correlated but irrelevant markers in the model. This rejects our hypothesis that prediction accuracy would increase with more markers. The missing data from GBS might be another reason that GBS did not offer better prediction. In the future, practical haplotype graph (PHG) could be utilized for more accurate imputation of missing GBS data and might lead to better GS prediction as has been observed in the literature (Franco et al. 2020; Jensen et al. 2020). The oat 6K chip is primarily composed of SNPs derived from cDNA sequences and thus likely offer more relevant information related to functional genes that will impact the traits of interest (Tinker et al. 2014). Based on the current study it would appear that phenotypic datasets based on BLUPs and a smaller, but genetically relevant set of SNPs (in this case represented by the 6K SNP dataset) provided the best prediction accuracy.

Among the 12 traits, prediction accuracy was considered good (r>0.5) for yield, oil, protein and β-glucan. Higher predictions may have resulted from the higher heritability of oil, protein and β-glucan (Fig. 3.8). A strong relationship between GS prediction accuracy and heritability has been reported (Combs and Bernardo, 2013; Heffner et al. 2011; Saatchi et al. 2010). In the current study the moderate heritability of yield (h_c^2=0.51) was predicted better than expected when compared to other moderately heritability traits, such as days to heading or plant height. Combs and Bernardo (2013) also observed that root lodging, a low heritability trait, had the second highest prediction accuracy in a maize biparental population. The higher accuracy for yield observed in the current study might be due to intensive yield testing in different environments (to increase heritability), which would result in more accurate estimation of breeding values using BLUPs. Other traits had relatively low prediction accuracy (0.3<r<0.5). Days to maturity had the lowest prediction accuracy (r=0.32), which corresponded with its low heritability (lowest among the 12 traits). The remaining traits displayed relatively high heritability (h_c^2>0.6) and lower prediction accuracy (r<0.5). A similar lack of correlation was found when wheat grain softness (h^2=0.88) displayed low prediction accuracy (r=0.37) (Heffner et al. 2011).

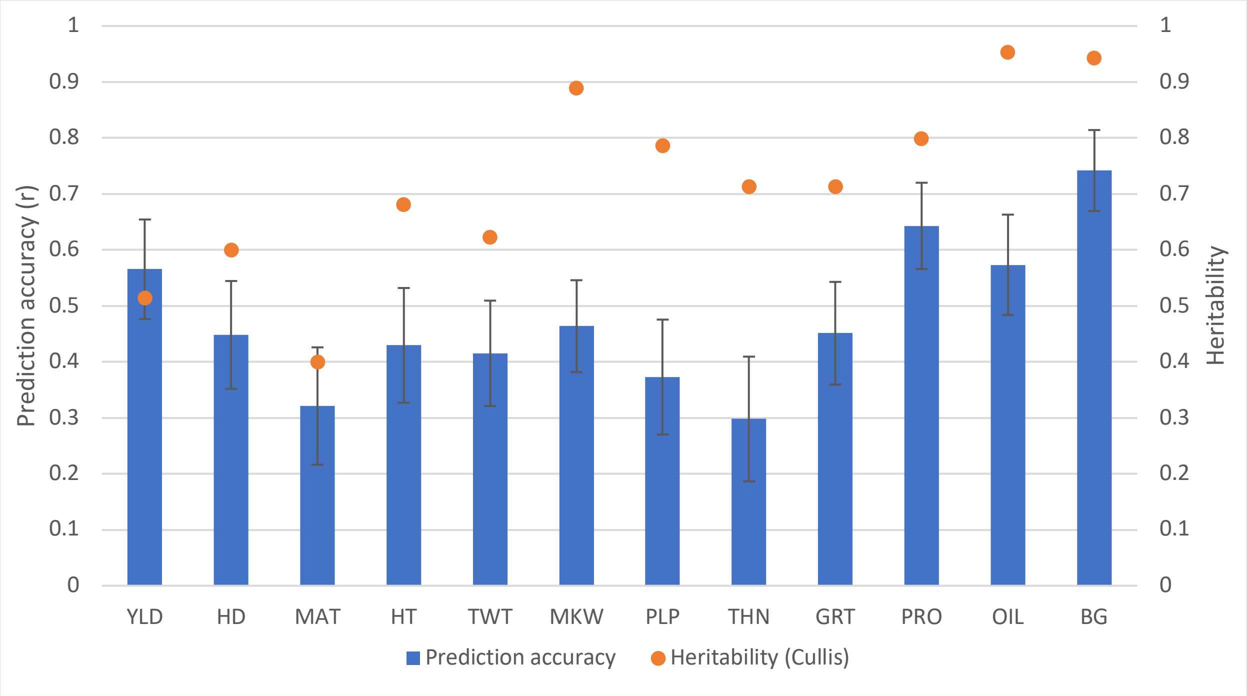


Fig. 3. 8 Comparison of GS model prediction accuracy and heritability estimates for 12 traits measured in the WCORT population. The GS model was created using the additive linear model rr-BLUP based on phenotypic BLUPs and the 6K SNP dataset. Note: YLD=yield, HD=days to heading, MAT=days to maturity, HT=plant height, TWT=test weight, MKW=thousand kernel weight, PLP=plumpness, THN=thins, GRT=groat, PRO=protein, OIL=oil, BG= β-glucan.

Among the six GS models evaluated using the different phenotypic and genotypic datasets, the Random Forest and SVM machine learning models, and the additive linear rr-BLUP model out-performed the more sophisticated and computationally intensive Bayesian method and the RHSK method. This finding corresponds with several GS studies in maize, oat and barley for disease and agronomic traits which showed only slight differences in prediction accuracy among different statistical GS models (Asoro et al. 2011; Lorenzana and Bernardo, 2009; Lorenz et al. 2012; Rutkoski et al. 2012). The parametric models (rr-BLUP and BayesCpi) differ in the assumption of how marker effects contribute to observed variance. Homogeneous distribution of marker effects across loci is assumed in rr-BLUP models, whereas Bayes Cpi allow heterogeneity of markers effects and the estimation of the probability π that a SNP marker has an effect (Goddard and Hayes 2007; Pérez et al. 2010; Colombani et al. 2011). Supervised machine learning models, such as SVM, can solve complex problems even with limited observations and a large number of predictors (Montesinos-López et al. 2019). It is interesting to note out the machine learning models showed promise in this study, suggesting their future potential in solving complicated associations between data and outcomes. It is believed that machine (deep) learning algorithms could capture nonlinear patterns from different sources than do conventional genome-based methods and could be widely used in GS assisted breeding (Montesinos-Lopez et al. 2021), although the decision-making process might remain unclear.

3.3.5 Marker density and training population size effect on GS prediction accuracy

Marker density effect was evaluated using the rr-BLUP package and phenotypic BLUPs for both and 6K SNP and GBS SNP datasets. Prediction accuracy improved with increasing 6K SNP marker number and plateaued once the marker number reached between 400 to 700 markers (Fig. 3.9). The relatively low number of markers required to maximize prediction accuracy may be due by the extensive LD identified in this population (LD decay was 12.1 cM). Similar results were observed with the GBS SNP marker dataset in which prediction accuracy increased up to 500 markers for all traits except yield (Fig. 3.10). In a study evaluating marker density effects in dairy cattle GS models, a low-density assay with evenly spaced SNPs across the genome offered good prediction accuracy (Moser et al. 2010). Heffner et al. (2011) discovered that decreasing marker number from 1,158 to 384 in a winter wheat study only slightly decreased GS prediction accuracy, but further reduction to 192 markers led to a dramatic 10% accuracy drop. GS prediction accuracy did not reach plateau for four out of five oat traits when marker numbers ranged from 300 to 900, indicating more markers may have been required in the investigated population (Asoro et al. 2011). As LD decay varies in different populations, the number of markers needed for GS should be evaluated in this context as it is critical that sufficient markers be assayed to tag all LD blocks within the genome. Using haplotypes, as defined by clusters of linked SNPs, increased GS prediction accuracy as genomic relationships among lines in the population were better estimated (Hess et al. 2017; Won et al. 2020). In a soybean GS study, a 4% prediction accuracy increase was observed when one SNP per haplotype block plus non-haplotype forming SNPs were used to select grain yield, as compared to using random or equally distanced SNPs (Ma et al. 2015). Several other factors, such as effective population size and genetic architecture of the trait, should be taken into account when considering marker density required. For example, increasing marker density without increasing the effective population size might result in poor GS predictions as inaccurate marker effects result from the lack of phenotypic variation and collinearity among markers (Muir, 2007). Training population size plays a significant role on GS model prediction accuracy. Increasing training population size has been shown to increase prediction accuracy in numerous GS studies (Asoro et al. 2011; Meuwissen et al. 2001; Saatchi et al. 2010). For example, an average 30% decrease in prediction accuracy among 13 winter wheat traits was observed when the training population size decreased from 288 to 96 individuals (Heffner et al. 2011). In the recent GS study in lentil, it was concluded that training population size should not be small when across population GS was conducted (Haile et al. 2020). In this study, increasing training population size increased prediction accuracy for all traits (Fig. 3.11). As population size increased from 61 to 244 individuals, the prediction accuracy increased from 25% to 100% among the different traits (Fig. 3.12).

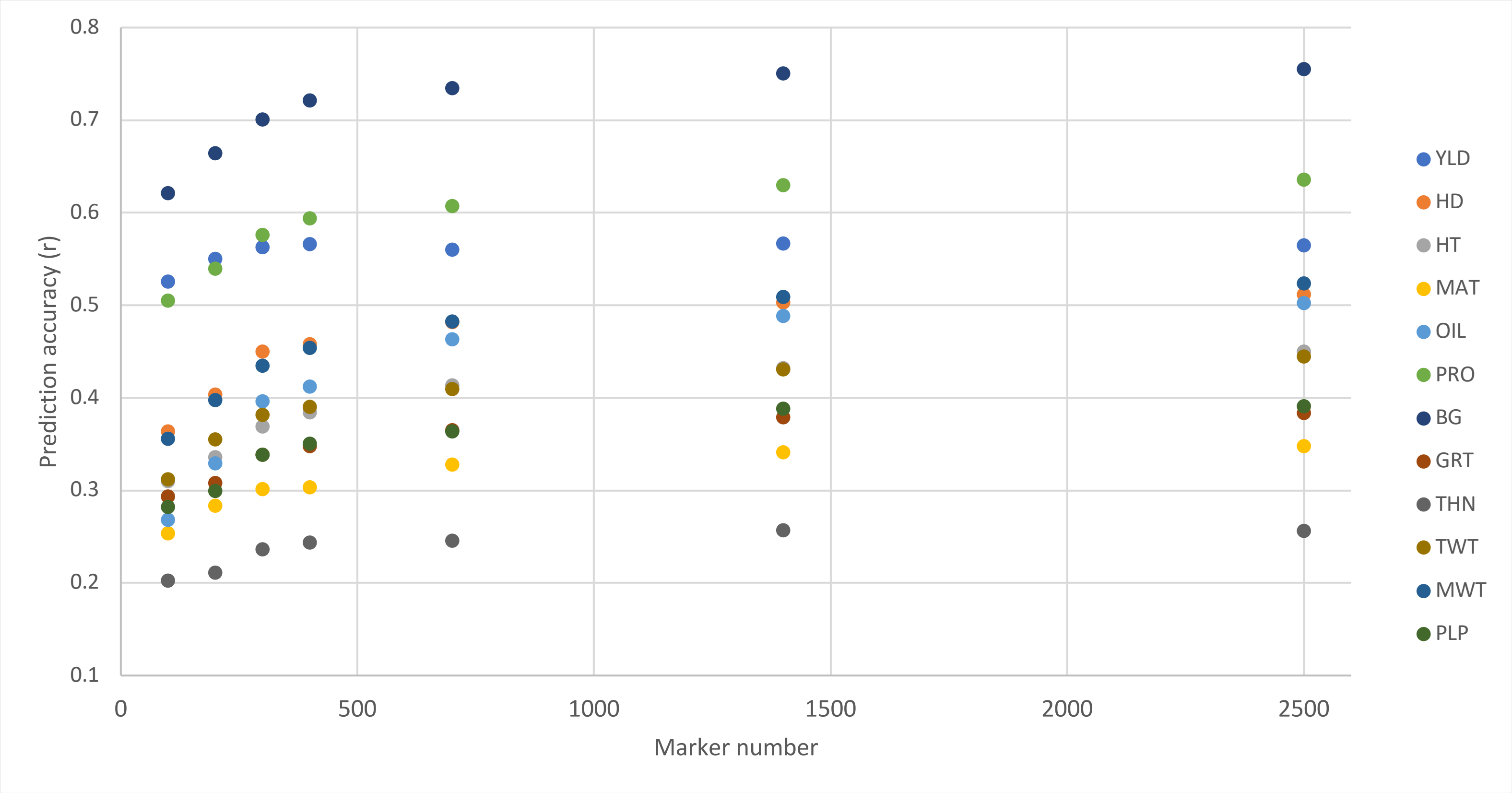


Fig. 3. 9 Marker density effect on prediction accuracy for 12 traits measured on the 305 oat lines grown in the WCORT from 2002-2014. Markers were randomly selected, and 100 sampling iterations were used to assess prediction accuracy. The additive linear model rr-BLUP was used for creation of the GS model based on phenotypic BLUPs and the 6K SNP dataset. Note: YLD=yield, HD=days to heading, MAT=days to maturity, HT=plant height, TWT=test weight, MKW=thousand kernel weight, PLP=plumpness, THN=thins, GRT=groat, PRO=protein, OIL=oil, BG= β-glucan.

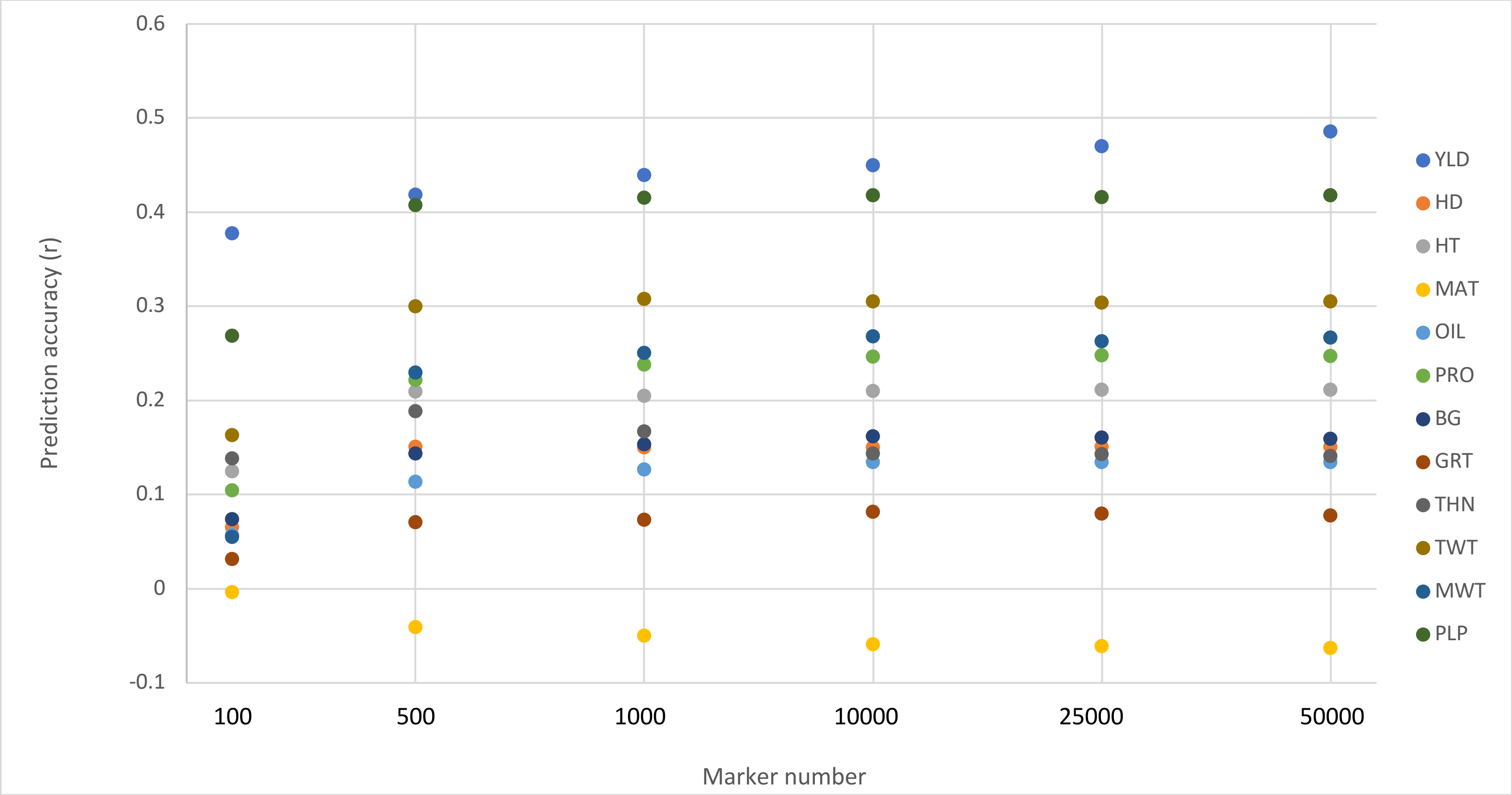


Fig. 3. 10 Marker density effect on prediction accuracy for 12 traits measured on the 305 oat lines grown in the WCORT from 2002-2014. Markers were randomly selected, and 100 sampling iterations were used to assess prediction accuracy. The additive linear model rr-BLUP was used for creation of the GS model based on phenotypic BLUPs and the GBS SNP dataset. Note: YLD=yield, HD=days to heading, MAT=days to maturity, HT=plant height, TWT=test weight, MKW=thousand kernel weight, PLP=plumpness, THN=thins, GRT=groat, PRO=protein, OIL=oil, BG= β-glucan.

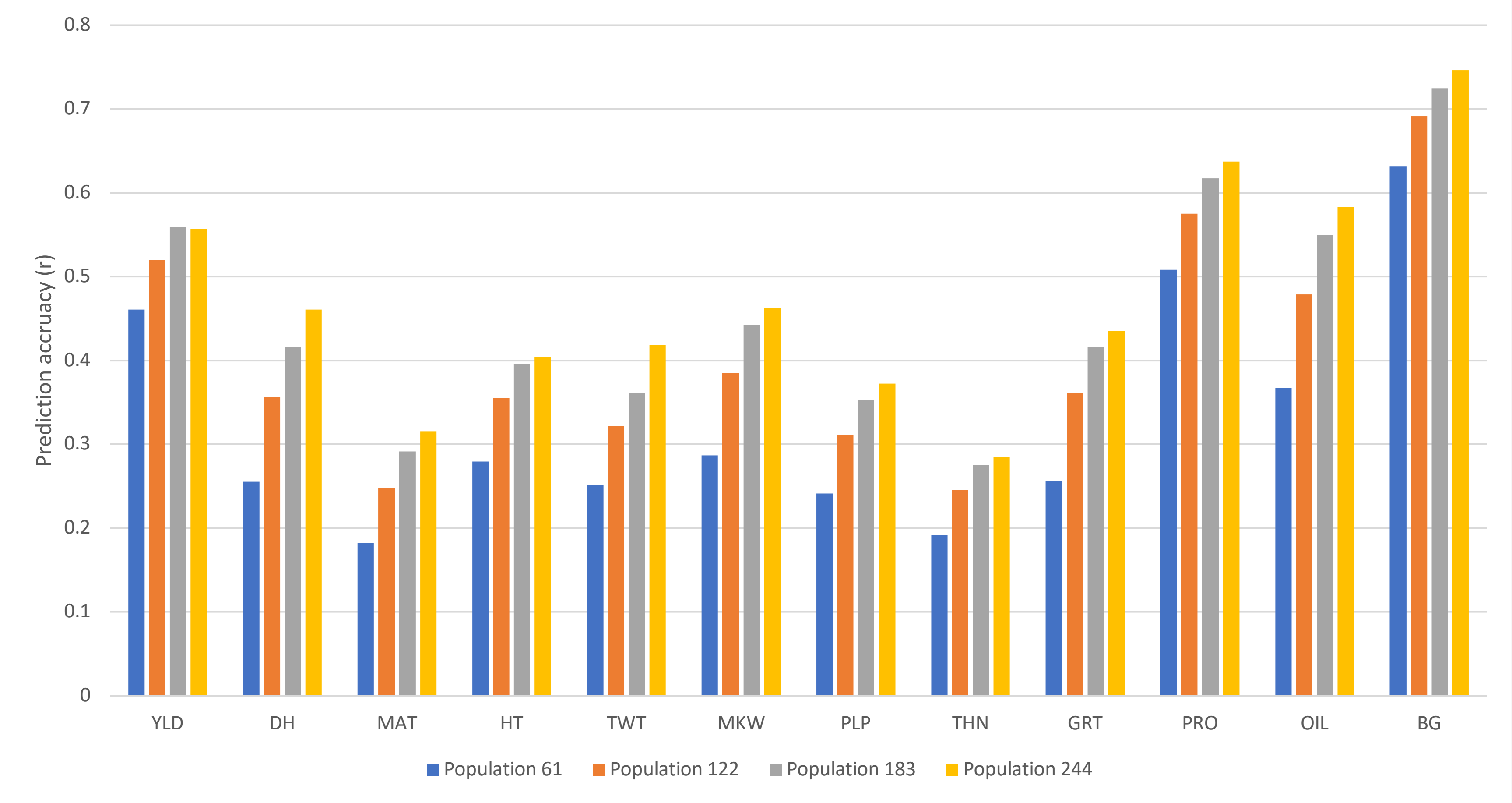


Fig. 3. 11 Training population size effect on prediction accuracy for 12 traits measured on the 305 oat lines grown in the WCORT from 2002-2014. The additive linear model rr-BLUP was used for creation of the GS model based on phenotypic BLUPs and the 6K SNP dataset. Note: YLD=yield, HD=days to heading, MAT=days to maturity, HT=plant height, TWT=test weight, MKW=thousand kernel weight, PLP=plumpness, THN=thins, GRT=groat, PRO=protein, OIL=oil, BG= β-glucan.

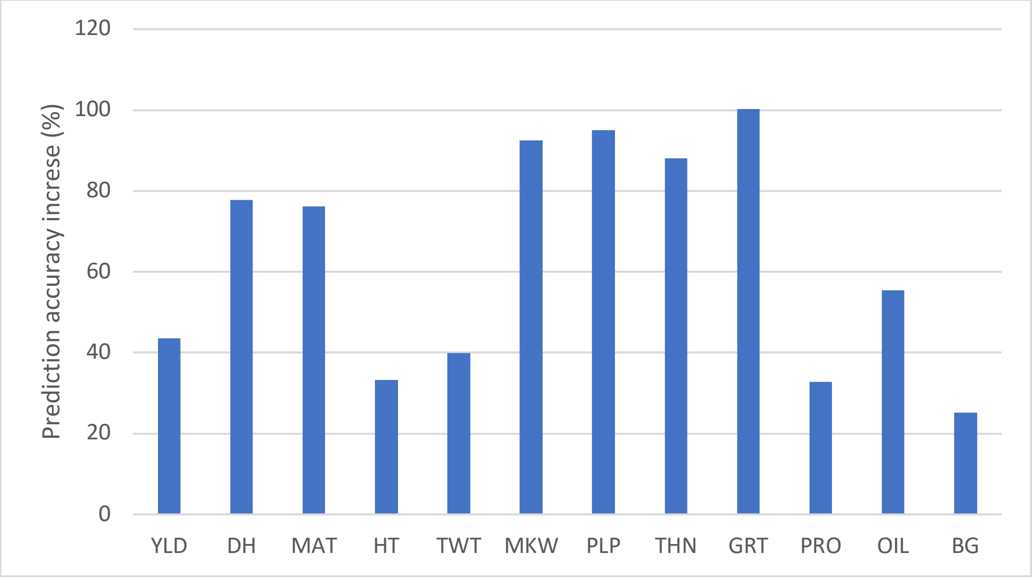


Fig. 3. 12. Increase in prediction accuracy for 12 traits measured on the 305 oat lines grown in the WCORT from 2002-2014 when training population size increased from 61 to 244. The additive linear model rr-BLUP was used for creation of the GS model based on phenotypic BLUPs and the 6K SNP dataset. Note: YLD=yield, HD=days to heading, MAT=days to maturity, HT=plant height, TWT=test weight, MKW=thousand kernel weight, PLP=plumpness, THN=thins, GRT=groat, PRO=protein, OIL=oil, BG= β-glucan.

3.4. Conclusion

In this study, different factors that affect GS model prediction accuracy were examined. Firstly, entry-mean and BLUPs were compared as phenotypic datasets. The BLUP phenotypic dataset resulted in better prediction accuracy using the current unbalanced dataset in which different oat lines were tested in different environments. This is significant given the similarity between the WCORT population and typical populations developed by plant breeders with their programs. Secondly, marker datasets obtained from 6K SNP and GBS SNP datasets were compared with the former offering better GS model prediction accuracy. Given the extensive LD (>10 cM) discovered in the elite lines comprising the WCORT population, it is probable that the high density GBS SNP dataset resulted in co-linearity among many markers, which in combination with an inadequate amount of phenotypic variation to match such a large number of markers, would have created "noise" in the model and thus reducing the accuracy of the marker effects in the model. With the availability of the OT3098 reference genome (*Avena sativa* – OT3098 v1, PepsiCo, https://wheat.pw.usda.gov/GG3/graingenes_downloads/oat-ot3098-pepsico), improvement in the GBS SNP dataset might be achieved by selecting meaningful polymorphisms that are genic and distributed to capture haplotype blocks. Thirdly, increasing marker density did improve prediction accuracy to a certain point for most traits (up to 500 markers for the GBS SNP dataset and between 400 to 700 markers for the 6K SNP dataset). Marker numbers above this number did not improve prediction accuracy for any trait, except for yield with the GBS SNP dataset in which prediction improvements continued to be made up to the maximum number of SNPs evaluated (50,000). Population structure, LD decay, haplotype blocks and genetic architecture of the traits should be investigated to determine appropriate marker density needed for GS. Fourthly, correlation was discovered between the heritability of a trait and the prediction accuracy, with several high heritability traits like protein, oil and β-glucan predicting well and lower heritability traits like days to maturity predicting poorly. Replicated multi-environment testing for low heritability traits, such as yield, did provide a better dataset to improve GS prediction accuracy as it increased heritability, thus providing better estimation of phenotypic values of individual lines. Finally, increasing training population size offered better prediction accuracies for all traits.

This study revealed some basic, but important factors to consider when conducting GS in oat. We started with some simple questions but ended up more complicated ones such as “What population can the GS model be used to predict within?” and “What are the boundaries within which we can trust the predictions from the GS model we developed?”. These questions point out the significance of careful evaluation of the studied trait (genetic architecture, heritability, correlated traits), the composition of the TP and the relationship between the TP and the BP before predication occurs so that proper estimation of phenotypes, appropriate marker selections and suitable TP design is done in order to have the most effective use of GS.

CHAPTER 4. MODELLING YIELD AND Β-GLUCAN IN CANADIAN OAT USING ENVIRONMENTAL VARIABLES AND CORRELATED TRAITS

4.1 Introduction

Oat is considered a healthy cereal due to a number of nutritional compounds found within the grain, including β-glucan. β-glucan is a soluble fiber that has been shown to lower plasma cholesterol and reduce the risk of heart disease (Wilson et al. 2004; Othman et al. 2011). This has resulted in health claims being established in both Canada (Health Canada, 2010) and the United States (FDA, 1997) based on oats products which contain a minimum 3g/100g β-glucan or 0.75g oat β-glucan per serving. This is a critical breeding target for oat breeders due to the importance to oat processors and end-users. Yield is another critical objective for oat breeders as it affects the financial viability of growing the crop for producers. Average yields within Canada over the past five years exceed 3,000 kg/ha, making Canada one of the highest-yielding oat production areas in the world (FAOSTAT, 2018). Improving the yield performance and β-glucan content of oat cultivars remains critical to maintain the health attributes and competitiveness of this crop. Yield performance and β-glucan content are controlled by both environmental and genetic factors and are inter-correlated with other traits. Heritability of these two traits is dynamic and is dependent on the genetic material and environments in which they are evaluated (Eagles, 1975; Holthaus et al. 1996; Kibite and Edney, 1998; Yan, 2020). While this can make modeling challenging, understanding these influential factors has important applications. For example, understanding the impact of environmental factors on β-glucan can help grain buyers to target growing regions that have shown favorable conditions throughout the growing season to produce higher β-glucan content in oat grown in those regions. Also, critical environmental factors related to a given trait can be used by breeders by incorporating these variable as co-factors in genomic selection models.

Numerous studies have examined these two traits within different populations and environments (Adegoke and Frey, 1987; Andersson and Börjesdotter, 2011; Doehlert et al. 2001; Henry, 1987). Genotypic effects on yield and β-glucan variation in oat is well described (Aman and Graham, 1987; Doehlert et al. 2001; Hurt et al. 1988), but at the genetic level, identifying QTLs for complex traits like yield remains difficult. Siripoonwiwat et al. (1996) identified approximately 40 QTLs linked to yield using the Kanota × Ogle (KO) population, but Beer et al. (1997) failed to validate them within a diverse set of 64 oat genotypes. De Koeyer et al. (2004) found two QTLs affecting grain yield in the Terra × Marion (TM) population, none of which were identified by Siripoonwiwat et al. (1996) or Beer et al. (1997). Interestingly, one yield QTL identified by De Koeyer et al. (2004) resided close to a QTL found in a previous study (De Koeyer et al. 2001) associated with plant height. This finding highlighted the possible interrelation between yield and plant height. Overall, these QTL studies suggested the difficulty of identifying consistent QTLs across different environments and genotypes.

Environmental factors, like heat and drought stress during flowering cause metabolic and developmental changes, resulting in a shortened life cycle, reduced light perception and photosynthesis and therefore yield loss (reviewed by Barnabás et al. 2008). Moving seeding date to early May from June led to improved quality and yield in Canadian eastern prairie regions (May et al. 2004). This might result from favorable growing environmental conditions and crown rust reduction. Earlier heading days led to more green leaves at anthesis, resulting in increased evapotranspiration and less reduced yield when facing post-heading heat stress in wheat (Tewolde et al. 2006). Prior studies indicate that β-glucan content tends to accumulate to a higher concentration in oat grain under warmer, drier conditions and to a lower degree when the growing season is cooler and wetter (Brunner and Freed, 1994; Güler, 2003; Lim et al. 1992; Saasatamoinen, 1995).

Structural equation modeling (SEM) is a multivariate path analysis approach that has been used to understand possible cause-and-effect relationships for complex crop traits that are affected by environmental factors, agronomic practices and genetic factors which may be expressed through other plant characteristics (Guillen-Portal et. al. 2006, Dhungana et al. 2007, Vargas et al. 2007, Kozak et al. 2008). Lamb et al. (2011) demonstrated the ability to analyze a complex network of inter-correlated variables using SEM in the analysis of a multi-site, multi-year oat yield trial. In his research, interrelationship among precipitation, seed size, seeding density, plant density, panicle density and yield were examined. Seeding density had insignificant effects on yield, even though strong effects were found on plant and panicle density. Precipitation has significant effects on yield, both directly and indirectly via plant density. A similar SEM study proved useful when modelling grain yield and its related traits in winter wheat (Mądry et al. 2015). Using five predictors, including early and late precipitation, fertilizer nitrogen level, thousand kernel weight and kernel nitrogen concentration, 67% of the variation in spring wheat yield was modelled using SEM (He et al. 2013). Other examples of SEM in plant sciences research included the work of Dhungana et al. (2007), Guillen-Portal et. al. (2006), Vargas et al. (2007) and Kozak et al. (2008). Dhungana et al. (2007) were able to explain 74% of yield genotype-by-environment interaction (GEI) variation via SEM, where direct and indirect effects of covariates were discovered. Spikes per square meter GEI possessed the highest direct effect on yield GEI variation and interactions with environmental variables such as temperature and precipitation were revealed. The objectives of this study were to identify the environmental variables and correlated phenotypic traits which impact yield and β-glucan content in oat, and to identify and test possible paths via SEM that incorporate these variables in a biologically sound model. The results of this study will provide a broader understanding of the direct and indirect influences of environmental variables on yield and β-glucan content.

4.2 Materials and Methods

Yield and β-glucan data collected from 2002-2014 as part of the Western Cooperative Oat Registration Trial (WCORT) were used in this study. The WCORT is the standardized trialing test used to collect data that forms the basis of the oat variety registration system in western Canada. As part of the trial operating procedures, three check varieties (AC Morgan, CDC Dancer and Leggett) were grown at each test site alongside candidate test lines in a 6 × 6 lattice design with three replications at 9-14 locations each year across western Canada. The WCORT locations from which phenotypic data was collected for this study are listed in Table 4.1 and the phenotypic traits included in the SEM models were days to heading, days to maturity, plant height, β-glucan and yield, as described in Table 3.1. Over the 13-year period used for this study a total of 305 candidate test lines were grown alongside the three check varieties.

Table 4. 1 Description of tests sites used in the Western Cooperative Oat Registration Trial (WCORT) from 2002-2014.

Test site	Code	Soil zone	Latitude	Longitude	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Beaverlodge, AB	BVR	2	55°12'N	119°27'W	1	1	1	1	1		1	1		1	1		1
Brandon, MB	BRA	1	49°50'N	99°57'W		1	1	1		1		1			1	1	1
Clive, AB	CLV	2	52°48'N	113°45'W	1	1											
Dawson Creek, BC	DAC	2	55°76'N	120°24'W										1	1	1	1
Fort Vermillion, AB	FTV	2	58°23'N	116°0'W	1	1	1		1	1	1	1	1	1			
Glenlea, MB	GLE	1	49°37'N	97°08'W		1											
Indian Head, SK	INH	1	50°33'N	103°39'W	1	1	1	1	1	1	1	1		1	1	1	1
Lacombe, AB	LAC	2	52°28'N	113°52'W	1	1	1	1	1	1	1	1	1	1	1	1	1
Lethbridge, AB	LET	4	49°42'N	112°50'W		1	1	1	1	1	1	1	1				
Melfort, SK	MEL	1	52°52'N	104°35'W	1	1		1			1	1	1		1	1	1
Morden, MB	MOR	1	49°11'N	98°06'W		1	1	1					1	1	1	1	1
Neapolis, AB	NPL	2	51°38'N	113°52'W			1	1									
North Battleford, SK	NBF	2	52°50'N	108°17'W			1										
Portage, MB	POR	1	49°54'N	98°16'W	1	1	1		1	1	1	1	1	1	1	1	
Regina, SK	REG	3	50°27'N	104°35'W	1	1		1	1		1	1	1	1			
Kernen, SK	KER	3	52°07'N	106°38'W		1	1	1	1	1	1	1		1	1	1	1
Scott, SK	SCT	3	52°21'N	108°50'W			1			1	1						
Swift Current, SK	SWC	3	50°17'N	107°41'W	1	1	1	1	1	1	1	1	1	1	1	1	1
Watrous, SK	WAT	3	51°40'N	105°28'W							1	1	1	1	1		
Wilkie, SK	WIL	2	52°24'N	108°41'W				1									
				Total:	9	14	13	12	9	9	12	12	9	11	11	9	9

For all locations in this study, data for environmental variables were directly collected or calculated from Environment Canada (https://weather.gc.ca/) (Environment Canada, 2020). A description of the environmental variables included in the SEM models are listed in Table 4.2, while a more extensive list of variables initially evaluated is provided in Appendix A. Briefly, the cumulative precipitation amounts for the September to April period, the September to May period, September to June period and September to July period were calculated to assess the impact of moisture accumulated prior to the growing season. Additionally, the monthly precipitation totals for each month from May to August were calculated, as were the cumulative precipitation values over the May to June, May to July and May to August periods, to assess the impact of moisture accumulated during the growing season. Temperature variables were assessed only for the growing season (May through August). These variables consisted of the monthly maximum (average of daily maximum temperature for that month), minimum (average of daily minimum temperature for that month) and mean (average of maximum and minimum temperature) temperatures. Cumulative sunlight hours were collected and calculated from National Research Council Canada (http://www.nrc-cnrc.gc.ca/eng/services/sunrise/).

Pearson correlation analyses were conducted among yield, β-glucan, days to heading, days to maturity, plant height and the environmental variables indicated in Table 4.2 and Appendix A using R (v. 3.3.2). The 305 lines grown over the 12 years of the WCORT were used to obtain better insight into the influence of these variables on one another and determine which variables to include in the hypothesized structural equation models for yield and β-glucan. Correlation was arbitrarily classified into three categories: high ($r>0.5$), moderate ($0.25<r<0.5$) and weak ($0.1<r<0.2$).

Additional environmental variables were included in the hypothesized SEM based on information in the literature demonstrating their influence on different plant development stages or phenotypic traits.

Table 4. 2 Description of the environmental variables evaluated in the SEM model.

Variable	**Units**	**Description**
Seeding date	days	Number of days after May 1
Precipitation	mm	Monthly rain and/or snow accumulation
Mean temperature	°C	Average of monthly minimum and maximum temperatures
Maximum temperature	°C	Average of daily maximum temperatures throughout a month
Cumulative sunlight	hr	Number of sunshine hours from May 1 to August 31

The Lavaan package (v. 0.5-23.1097) within the statistical package R (v. 3.3.2) was used for SEM model testing. Goodness of fit was assessed by comparison of the variance-covariance matrix between hypothesized models and fitted observed models via the χ^2 test. Root mean square error of approximation (RMSEA) and comparative fit index (CFI) were also used to determine the goodness of fit of the hypothesized models.

4.3 Results and Discussion

4.3.1 Hypothesized SEM models for yield and β-glucan

Among the traits measured on the WCORT lines across the different environments, higher correlation was found between days to heading and maturity (r=0.44, n=2092), days to maturity and yield (r=0.36, n=3806) and plant height and yield (r=0.62, n=3105). Correlation analyses between these traits and environmental factors showed that days to heading was correlated with seeding date (r=-0.72, n=2502) and cumulative sunlight hours (r=-0.32, n=2400). Days to maturity was correlated with July and August precipitation (r=0.48, n=3878), July mean temperature (r=-0.70, n=3842) and August mean temperature (r=-0.60, n=3842). Plant height was correlated with May to July precipitation (r=0.41, n=4015), cumulative sunlight hours (r=-0.32, n=2400) and July mean temperature (r=-0.30, n=4104). Yield was correlated with May to July precipitation (r=0.22, n=4258) and July maximum temperature (r=-0.37, n=4376).

Thermal time is known to influence plant growth and development. Therefore, June mean temperature was chosen as an indirect indicator of heat units received before heading, June mean and July mean temperature were selected as factors influencing plant height, and July and August temperatures were selected as indirect indicators of heat units received between days to heading and days to maturity. Water availability during plant development has significant effects on stomatal conductance, viable leaf area, root/shoot ratio, photosynthesis, anthesis, grain filling, and thus on yield (Barnabás, et al. 2008). May and June precipitation were proxy indicators of water intake before heading, July and August precipitation was a proxy indicator of water intake between days to heading and days to maturity, and May to July precipitation was a proxy indicator of water intake impacting plant height and yield. The causal relationship between high temperatures during anthesis/grain filling and yield reduction in crops like wheat is well established (Tashiro and Wardlaw, 1990, Stone and Nicolas, 1995; Wardlaw et al. 2002). Therefore, July maximum temperature was added as a variable directly influencing yield to capture heat stress during anthesis and grain filling. These variables were combined with the highly correlated environmental factors identified in the WCORT population and used as explanatory variables in the hypothesized path model for yield (Fig. 4.1a).

Relatively strong correlations were found between β-glucan and days to maturity (r=-0.24, n=1726), July mean temperature (r=0.28, n=2115), and July precipitation (r=-0.17, n=2115). Lacking any other meaningful information pertaining to traits or environmental variables influencing β-glucan, SEM models were developed to test a hypothesized path model between β-glucan, days to maturity, July Mean Temperature and July Precipitation (Fig. 4.1b).

4.3.2 Correlation analysis

Correlation coefficients among the target variables (yield and β-glucan) and explanatory traits and environmental variables in the hypothesized SEM models were assessed using data obtained from the three check varieties to determine how well the larger set of data from the WCORT lines predicted information from the three individual oat check varieties. Similar correlation trends were observed between the WCORT lines and the three oat check varieties (Table 4.3). Most of the correlation coefficients for the WCORT lines were very similar to the mean correlation coefficients of the three checks, indicating the hypothesized SEM model should be valid for the three check varieties. One exception was the stronger negative correlation observed between sunlight hours and plant height in the WCORT lines (-0.32) compared to the three check cultivars (-0.14 to -0.08).

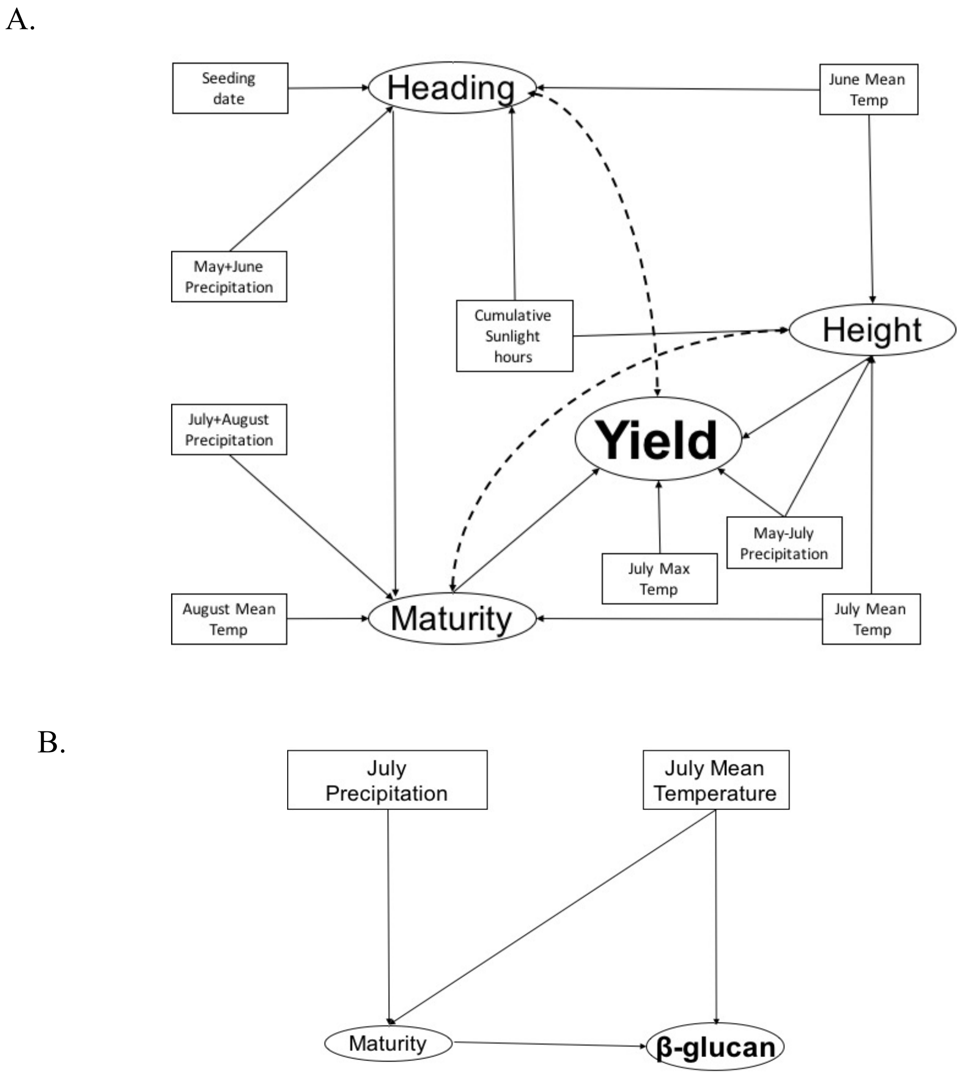


Fig. 4. 1 Hypothesized path models for yield (A) and β-glucan (B).

Table 4. 1 Correlation coefficients between days to heading, days to maturity, plant height, yield and β-glucan and their respective influential traits and environmental variables for three check varieties and 305 lines grown in the WCORT from 2002-2014.

Days to Heading	Seeding Date[1]	Sunlight Hours[2]	June Mean Temperature[3]	May and June Precipitation[4]
CDC Dancer	-0.76	-0.36	-0.24	0.18
AC Morgan	-0.76	-0.39	-0.22	0.09
Leggett	-0.74	-0.31	-0.25	0.01
WCORT Lines	-0.72	-0.32	-0.27	0.10
Days to Maturity	Days to Heading	July Mean Temperature[5]	August Mean Temperature[6]	July and August Precipitation[7]
CDC Dancer	0.34	-0.70	-0.64	0.48
AC Morgan	0.33	-0.72	-0.63	0.47
Leggett	0.35	-0.72	-0.50	0.47
WCORT Lines	0.42	-0.70	-0.60	0.48
Plant Height	June Mean Temperature	July Mean Temperature	Sunlight Hours	May to July Precipitation
CDC Dancer	-0.17	-0.27	-0.14	0.48
AC Morgan	-0.15	-0.29	-0.08	0.44
Leggett	-0.08	-0.34	-0.08	0.30
WCORT Lines	-0.19	-0.30	-0.32	0.41
Yield	Days to Maturity	Plant Height	July Max Temperature[8]	May to July Precipitation
CDC Dancer	0.30	0.68	-0.42	0.22
AC Morgan	0.51	0.62	-0.46	0.15
Leggett	0.26	0.67	-0.42	0.23
WCORT Lines	0.36	0.62	-0.37	0.22
β-glucan	Days to Maturity	July Mean Temperature	July Precipitation[9]	
CDC Dancer	-0.41	0.43	-0.30	
AC Morgan	-0.54	0.51	-0.33	
Leggett	-0.15	0.47	-0.18	
WCORT Lines	-0.24	0.28	-0.17	

[1] Number of days between seeding and May 1

[2] Cumulative sunlight hours from May 1 to August 31

[3] Average of the maximum and minimum June temperature

[4] Cumulative precipitation for May and June

[5] Average of the maximum and minimum July temperature

[6] Average of the maximum and minimum August temperature

[7] Cumulative precipitation for July and August

[8] Average maximum July temperature

[9] Total precipitation in July

Days to heading

The correlation between days to heading and seeding date, cumulative sunlight hours, and June mean temperature were consistent across the three oat check varieties. Days to heading and seeding date had the strongest negative correlation (r=-0.76 to -0.74), indicating that earlier seeding did not lead to earlier heading. Growing season sunlight hours showed a moderate negative correlation with heading date (r=-0.36 to -0.31). This observation was reflected in the latitude of the growing location with more northerly locations being associated with fewer days to heading. Indeed, being a long day crop, oats flower faster when given longer photoperiod hours among different cultivars (Wiggans and Frey, 1955). June mean temperature was weakly correlated with heading (r=-0.16 to -0.12), indicating the role of temperature was not significant. The lack of relationship between days to heading and May and June precipitation in this study (r=-0.08 to 0.09) also indicated that daylength is the predominate driver of flowering.

Days to maturity

Maturity had consistent correlation with days to heading, July mean temperature, August mean temperature, and July and August precipitation across the three oat check varieties. Moderate correlation was found between maturity and heading (r=0.33 to 0.35), suggesting that there was a consistent relationship across genotypes with respect to the growth period spent pre- and post-flowering. July and August mean temperature had strong negative correlations with maturity (r=-0.72 to -0.70 and r=-0.64 to -0.50, respectively). Higher temperatures during anthesis and grain filling have been shown to accelerate maturity and senescence in spring wheat (Hatfield and Prueger, 2015; reviewed by Yang and Zhang, 2006). The higher temperatures appear to increase enzyme activities and metabolic processes which accelerates progression during these final phases of the plant's lifecycle (reviewed by Dupont and Altenbach, 2003). July and August precipitation had a moderate correlation with maturity (r=0.30 to 0.48), demonstrating the importance water availability has on plant development. The combined stresses of higher temperature and less water available are known to cause plants to mature earlier as an avoidance strategy (reviewed by Chaves et al. 2003).

Plant height

Plant height was influenced by growing season cumulative sunlight hours, June and July mean temperature, and May to July precipitation. The correlation between growing season sunlight hours and plant height was weak among the three check varieties (-0.14 to -0.08) and somewhat higher (r=-0.32) with the WCORT lines. Hours of sunlight has been shown to have a significant influence on the duration of vegetative growth in crops like safflower and cotton (Ozkaynak, 2013).

Negative correlations were observed for June (r=-0.17 to -0.08) and July (r=-0.34 to -0.27) temperatures, both factors being indicators of heat stress. May to July precipitation had a mild positive effect on plant height (r=0.3 to 0.44), this factor being an indicator of drought stress. During a hot, dry growing season, plants tend to direct more energy from shoot development to root development in order to improve water uptake from the soil. This result confirms the effect that higher temperatures and drought have on plant development (Cairns et al. 2013; reviewed by Lipeic et al. 2013). To conserve water during extremely hot days, plants can close stomata to prevent water loss, which would slow down photosynthesis and promote photorespiration (reviewed by Bueckert et al. 2013). As a result, slower aboveground vegetative growth and shorter plant stature are expected.

Yield

A high positive correlation was found between yield and plant height (r=0.62-0.68). Plant height can be considered an indirect indicator of above ground biomass where photosynthetic activity takes place. Benaragama (2011) studied seven different oat lines and found that taller oat plants tended to have wider and longer flag leaves. Flag leaves produce the greatest proportion of post-anthesis assimilates in barley and wheat (Carr and Wardlaw, 1965; Thorne, 1974), and their removal causes a significant reduction in oat grain yield (Frey, 1962). Height might also reflect total biomass at anthesis, as it was highly correlated with oat grain yield in a study by McMullan et al. (1988). Correlations between plant height and leaf area have also been observed in oats, which would further support the relationship between height and yield (Djanaguiraman et al. 2020). Plant height was also positively correlated to competitive ability in wheat (Huel and Hucl 1996; Cosser et al. 1997). Greater interception of photosynthetically active radiation allowed taller wheat plants to compete with weed more efficiently (Wicks et al. 1986; Champion et al. 1998). The shortest cultivars had the least yield with the largest weed growth among Canadian spring wheat (Huel and Hucl 1996). It was possible that taller oat plants had more competitive ability under weed pressure and obtained greater yield. A significant correlation was found between plant height and grain yield in oats when lodging did not occur (Sampson, 1976; Tibelius and Klinck, 1985).

Correlation between days to maturity and yield was moderate and varied among the three oat checks (r=0.26-0.51). The positive association between maturity and yield has been reported previously in oat (Kaufmann 1961) and is due to the increase in the number of photosynthetic hours accumulated during the growing season, and thus carbon accumulation (Kaufmann 1961). Association between growth duration and gain yield were discovered in rice, but other factors such as dryness of the growing season, nitrogen and spacing were also important on grain yield (Vergara et al. 1966). Weaker association (r=0.36) between days to maturity and yield was observed in resynthesized *Brassica napus* (Karim et al. 2014). Longer maturity days lead to increased yield in summer maize grown in Huanghuaihai Plain of China (Yang et al. 2019). The poorer correlations with CDC Dancer may result from its relatively earlier maturity, thus it is unable to take advantage of a longer growing period at some locations, unlike moderate maturing varieties like AC Morgan. At the other end of the maturity spectrum, a later maturing variety (like Leggett) may not have time to complete grain filling in some shorter growing season locations, thus affecting yield. July maximum temperature had a consistent negative correlation with yield among the check varieties (r=-0.46 to -0.42).

Heat stress can affect yield by impacting the development of reproductive organs, male and female gametes and the grain-filling process. Higher temperature around anthesis and early grain fill can lead to pollen and embryo abortion, leading to less yield (Stone and Nicolas, 1994; Farooq et al. 2011). For example, temperatures above 30°C for a duration of two days during flowering time caused abortion of buds, flowers, and young seeds in pea (Karr et al. 1959; Guilioni et al. 1997). A decrease in the duration of reproductive growth and yield in pea under drought conditions has also been found when there are over 20 days in a growing season with a temperature above 28°C (Bueckert et al. 2015). In the same study, irrigation and precipitation mitigated the adverse effect of the heat stress. In wheat, high temperature stress (>30°C) during anthesis damaged the viability of pollen grains, thus affecting fertilization and causing a reduction in seed set and yield (Saini and

Aspinall, 1982; Ferris et al. 1998). May to July precipitation had a weak correlation with yield (r=0.15-0.23). This variable can be impacted by other considerations like spring soil moisture content, timing of precipitation and potential for both too much and too little precipitation negatively influencing yield (Kutcher et al. 2010; He et al. 2013; Meng et al. 2017).

β-glucan

β-glucan content in oat grain was associated with July mean temperature, July precipitation and maturity. Warm, dry environments tend to produce higher β-glucan content while cooler, damp environments result in lower β-glucan levels (Brunner and Freed, 1994; Güler, 2003; Saasatamoinen, 1995). The correlation between β-glucan and July mean temperature was high and consistent among the check varieties (r=0.43-0.51). However, genotypic differences appear to be more important with respect to how maturity and July precipitation may impact β-glucan. The correlation with both these environmental variables was lower with Leggett compared to the other two checks. Leggett shows significantly higher amounts of grain β-glucan than the other two checks so it may be more sensitive to environmental variation that impacts grain filling. For example, an environment which leads to later maturity may decrease grain filling which has a dilution effect on β-glucan. As mentioned above, Leggett is a later maturing variety.

4.3.3 Observed structural equation modelling for yield

Observed structural equation models were fit for each oat variety against the hypothesized yield model shown in Fig. 4.1a. The fitted yield models for each variety based on observed data indicated that the hypothesized model was consistent with the observed data and was a reasonable description of yield formation (Table 4.4). The fitted SEMs for yield for each oat variety are shown in Figs. 4.2a-c.

Table 4. 2 Statistical tests evaluating the goodness of fit between the hypothesized models for yield and β-glucan and the SEM based on observed data.

Trait	Variety	Observations	Chi square	P value	CFI	RMSEA
YLD	CDC Dancer	34	31.90	0.20	0.98	0.06
	AC Morgan	34	20.99			
	Leggett	25	28.86			
BG	CDC Dancer	34	0.43	0.43	1.00	0.00
	AC Morgan	34	0.24			
	Leggett	24	2.11			

Note: YLD=yield, BG= β-glucan

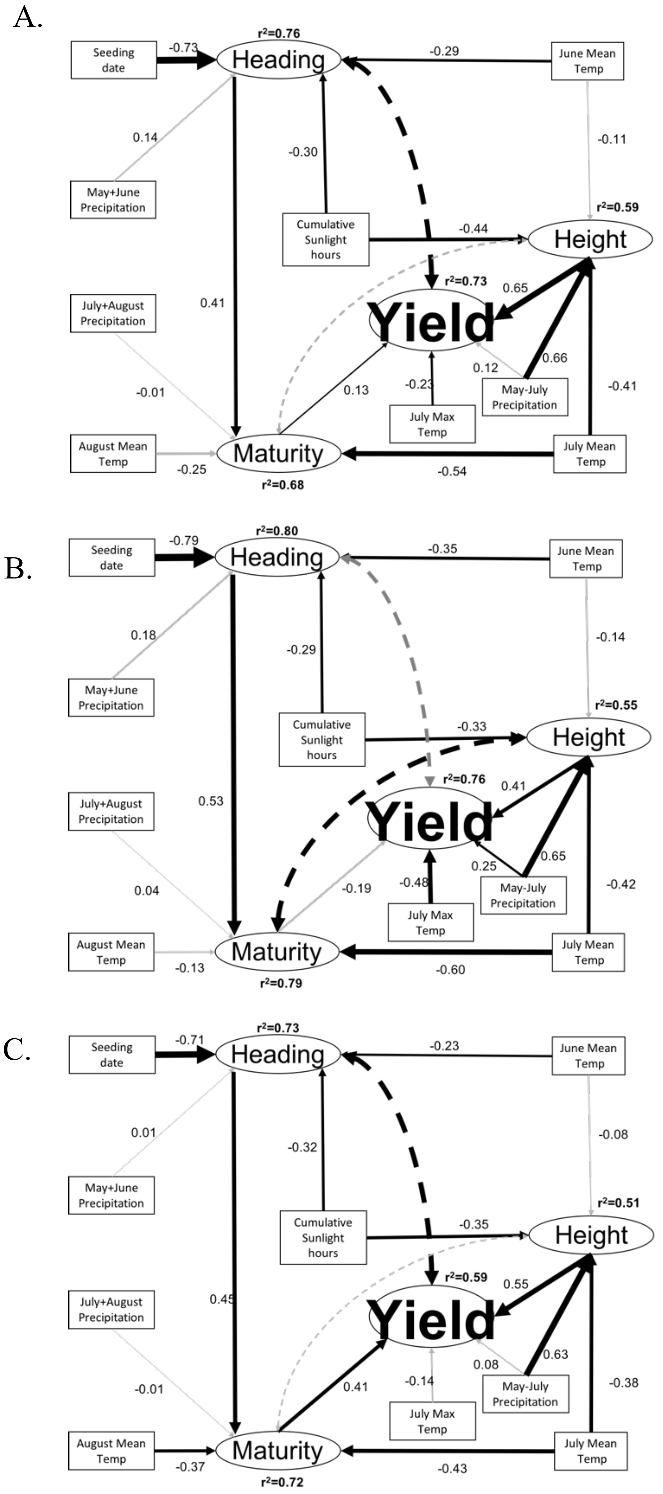


Fig. 4. 2 Fitted observed structural equation models for yield for each of the three oat check lines CDC Dancer (A), AC Morgan (B) and Leggett (C). A significant ($p<0.05$) path is shown as solid lines with the width proportional to the magnitude of the standardized coefficients. Covariance paths are shown with broken lines. Non-significant paths ($p > 0.05$) are shown in gray.

Paths to days to heading

The three oat check cultivars are well-adapted to western Canada, where long day-length photoperiods are favourable for fewer days to heading. The fewer days to heading among the three oat checks suggested that genes related to photoperiod sensitivity and vernalization insensitivity existed in the genome of these adapted oat varieties (Kibite and Menzies, 2000; Fetch et al. 2007; Mazurkievicz et al. 2019). In the study of Mazurkievicz et al (2019), variation in days to heading in two oat bi-parental populations was mainly explained by the genetic differences derived from parents in response to photoperiod and temperature. In this study, we eliminated the genetic variation by analyzing the three checks individually in different environments and solely focused on environmental factors. Significant direct paths to days to heading were established with seeding date, June mean temperature, and cumulative sunlight hours for all three check varieties. Later seeding dates, cooler June temperatures and growing sites located further north led to longer days to heading. The path with May and June precipitation was insignificant, suggesting precipitation is not a critical variable that impacts heading. Coefficient of multiple determination (r^2) for heading was 0.73-0.80, suggesting a good fit.

Paths to maturity

Direct paths with days to heading and July mean temperature were significant for maturity. Longer days to heading days and cooler July temperatures resulted in a longer maturity period. Indirectly, seeding date and June mean temperature negatively affect maturity via their paths with days to heading. The relationship between delayed seeding and the decrease in days to maturity and grain yield was also reported previously for the eastern prairies of western Canada (May et al. 2004). The path between July-August precipitation and maturity were insignificant. August mean temperature affected maturity in Leggett but remained insignificant for the other two checks. As mentioned previously, the relatively later maturity of Leggett may be relevant to this variable, that is, it is able to benefit from the longer growing season implied by cooler August temperatures. The coefficient of multiple determination for maturity was 0.68-0.79 which again indicted a good fit.

Paths to plant height

Direct significant paths to height included July mean temperature, cumulative sunlight hours and May-July precipitation. Warmer July temperatures, more sunlight during the growing season and reduced precipitation led to shorter statue in the three checks. The path from June temperature was not significant. The coefficient of multiple determination for plant height was 0.51-0.59, suggesting some important variables impacting plant height are still not being accounted for. Variations in the duration of phenophases was explained by thermal time accumulated between development phases and its interaction with photoperiod and vernalization in wheat (Shaykewich, 1995). As vernalization was not a requirement for cultivated oat, temperature, daylength, variety and their interaction could the contributing factors on height variation (Kirby and Appleyard, 1986; Brouwer and Flood, 1995; Slafer and Rawson, 1995). Oat varieties might also display variation in plant height due to the presence of dwarfing genes that are less affected by environment (Milach et al. 1997; Zhao et al. 2018).

Paths between yield and other variables

This study revealed the importance of correlated traits and environmental variables on explaining field variations of yield and β-glucan via structural equation models. The inability of the models to fully explain the variability for yield and β-glucan indicates that not all relevant factors impacting these traits were captured. It is also important to keep in mind that conclusions based on the three oat cultivars used in this study may not apply universally to all oat varieties, especially those which may carry genes that impact important traits like daylength sensitivity or plant height (vie dwarfing genes), or those adapted to environments different than those in western Canada. However, the similar relationships observed in the larger WCORT population did offer some confidence in the results from this study.

Paths between plant height and yield were significant for all three oat varieties, while paths to maturity, July maximum temperature, and May-July precipitation were variety-dependent. The maturity path to yield was not significant for Leggett. As indicated above, a later maturing variety like Leggett may not have time to complete grain filling in some shorter growing season locations, thus affecting its predictive relationship to yield. July maximum temperature had a significant negative effect on yield for CDC Dancer and Leggett, but not for AC Morgan, while May-July precipitation was only relevant to yield formation in Leggett. These different results implied that genetic differences between the varieties may be impacting their interaction with the environment. Seeding date and June mean temperature have indirect negative effects on yield via their impact on days to heading, July mean temperature has an indirect negative impact on yield via its effect on plant height and maturity, while cumulative sunlight hours had an indirect negative effect on yield due to its association with both plant height and days to heading.

Overall, the coefficient of multiple determination for yield was 0.59-0.76, indicating a moderate to good incorporation of important variables impacting yield formation was achieved.

4.3.4 Observed structural equation modelling for β-glucan

Observed structural equation models were fit for each oat variety against the hypothesized β-glucan model shown in Fig. 4.1b. The fitted β-glucan models for each variety based on observed data indicated that the hypothesized model was consistent with the observed data (Table 4.4). The fitted SEMs for β-glucan for each oat variety are shown in Figs. 4.3a-c.

Direct paths with July mean temperature and July precipitation were significant for maturity. As with yield, cooler July temperatures resulted in a longer maturity period, while more precipitation in July had a small impact on extending maturity. Significant direct paths with July mean temperature and maturity were observed for β-glucan. Higher July mean temperature decreased β-glucan across all three varieties. This may be explained by the effect temperature may have on yield. July temperature was seen to have a negative direct effect on yield (Fig. 4.2) and higher yield can have a dilution effect on β-glucan content. Thus, the direct impact of warmer July temperatures would be to increase β-glucan. The paths between maturity and β-glucan were inconsistent, with standardized coefficients being negative for CDC Dancer and AC Morgan and positive for Leggett. While the negative relationship between maturity and β-glucan may be explained by the dilution effects presumed to occur with greater yield resulting from longer maturity time, the positive relationship observed with AC Morgan is harder to understand.

Overall, the coefficient of multiple determination for β-glucan was 0.16-0.41, indicating a low to moderate incorporation of important variables impacting β-glucan formation was achieved.

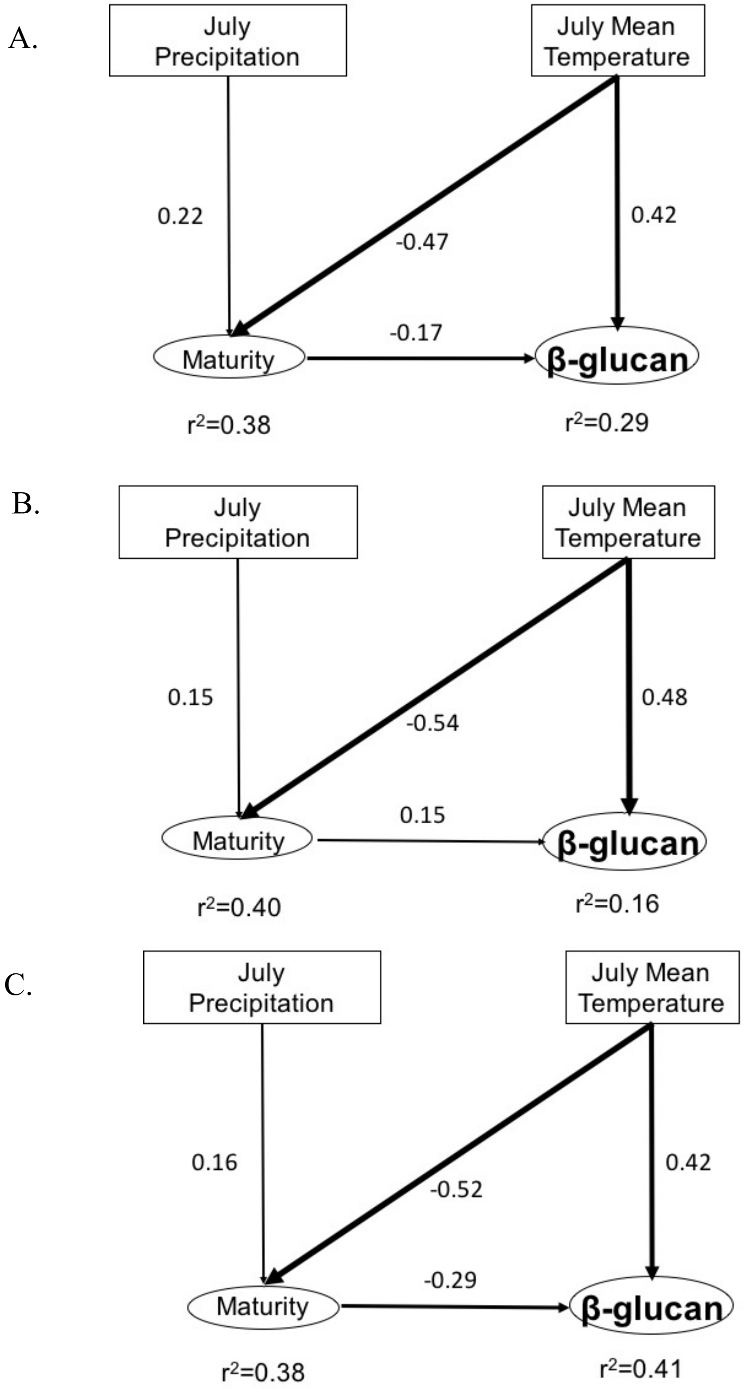


Fig. 4. 3 Fitted observed structural equation models for β-glucan for each of the three oat check lines CDC Dancer (A), AC Morgan (B) and Leggett (C). A significant ($p<0.05$) path is shown as solid lines with the width proportional to the magnitude of the standardized coefficients.

4.4 Conclusion

Oat yield and β-glucan content are the end result of a series of biochemical processes influenced by the inherent genetics of a variety, as well as external environmental influences. Identifying paths involving inter-correlated phenotypic traits and environmental factors that are influential to yield formation and grain β-glucan content will have important uses, such as identifying oat growing regions that have experienced beneficial environmental conditions that will maximize grain β-glucan, an important end-use trait who concentration in grain is critical to meeting health claims associated with consumer products. Secondly, understanding correlated traits and environmental factors related to yield or grain β-glucan allow them to be used as co-factors in genomic selection models developed by oat breeders.

In this study structural equation models that incorporated different phenotypic traits and environmental variables were developed for oat grain yield and β-glucan content. A hypothesized model for yield was developed by assessing correlations among 36 environmental variables that captured information related to temperature, precipitation, sunlight hours and seeding dates, and three phenotypic traits that captured phenological information and plant stature. Significantly correlated variables, along with additional variables thought to be relevant based on literature, were placed into the hypothesized model and subsequently compared to a fitted model based on observed data for three oat varieties. The yield model displayed direct path effects related to heading date, maturity, plant height and July max temperature, while indirect path effects captured data from mean temperature in June, July and August, May-July precipitation and seeding date. Path effects were consistent across three oat varieties and the coefficient of multiple determination ranged from 0.59-0.76, indicating a fairly good incorporation of relevant variables impacting yield formation occurred. The β-glucan model displayed direct path effects related to maturity and July mean temperature, while indirect path effects captured data from July precipitation. The path effect associated with July mean temperature was consistent across oat varieties but varied for maturity. The coefficient of multiple determination was lower than for yield, ranging from 0.16-0.41, indicating a low-moderate incorporation of important environmental variables impacting this trait.

CHAPTER 5. IMPROVING GENOMIC SELECTION WITH ENVIRONMENTAL AND CORRELATED TRAIT INFORMATION

5.1 Introduction

Oat *(Avena sativa* L.) is the seventh largest production crop worldwide with an average of 23.4 million metric tonnes produced annually over the past five years (FAOSTAT, 2019). Production has stabilized since 2009, in part due to the health benefits associated with consuming oat. These health benefits are associated with the presence of β-glucan, which has been shown to reduce the risk of heart disease (Queenan et al. 2007, Liatisa et al. 2009), and avenanthramides, a unique class of antioxidants which may also protect against heart disease (Meydani et al. 2009). The 12-20% protein content of oat (Klos and Arendt, 2012), which can be consumed by most people suffering from Celiac disease and is thus considered "gluten-free" (Peraaho et al. 2004), and its neutral flavour and colour profile have placed it at the top of new sources to supply the increasing demand for plant-based proteins (Givaudan, 2019).

Oat varieties are commonly released genetically uniform inbreds. The breeding process involves hybridization of parental lines to generate genetic variability which is then followed by inbred line development (with or without selection) at which point single-plant selection for higher heritability traits (e.g. crown rust resistance, plant height) is performed. The final generations are grown in replicated, multi-location trials over several years to evaluate lower heritability traits like yield. While this process is effective, the most significant deficiency is the fact that the single plant selection stage essentially samples variation for lower heritability traits in a random manner, thus genetic gain is not realized to a fuller potential.

Genomic selection (GS) is a tool that utilizes genomic-trait information to increase genetic gain in breeding programs, and it is particularly useful when applied to quantitative traits which are not only controlled by genotype, but also by the environment and their interaction. Genomic selection attempts to estimate the effect of markers across the entire genome on a target trait phenotype, instead of relying on one or a few predefined markers (Heffner et al. 2009; Jannink et al. 2010), and is believed to hold much potential for increasing genetic gain per breeding cycle (reviewed by Heffner et al. 2009; Cabrera-Bosquet et al. 2012). The basic principle of the method entails, 1) creation of an initial genomic model for target traits using genotypic and phenotypic data available from a training population and validation of the model, 2) application of the genetic model to derive GEBV for individuals within a related breeding population for which only genotypic information is available, and 3) re-evaluation of the genomic model based on phenotypic data arising from individuals selected by genetic model, as well as, additional genotypic and phenotypic data representing new elite and novel germplasm (Heffner et al. 2009; Jannink et al. 2010; Heffner et al. 2011a).

Many factors are known to be critical in the development of GS models with good prediction accuracy. Marker density, haplotype block structure, training population size, statistical model used and the genetic relationship between the training and breeding populations are some of the factors that have been examined (e.g. Meuwissen et al. 2001; Lorenzana and Bernardo, 2009; Luan et al. 2009; Solberg et al. 2008; VanRaden et al. 2009; Calus, 2010; Moser et al. 2010; Heffner et al. 2011b; Asoro et al. 2011; Hamblin et al. 2011; Jannink et al. 2010; Combs and Bernardo, 2013; Wientjes et al. 2013). These aspects of GS model development are focused on the genetic and statistical components of the model, but an equally important aspect of model development entails understanding the impact of environment and genotype-by-environment (G×E) interaction, especially for many important, but low heritability traits, like yield.

Evaluation of lower heritability traits is typically done by conducting multi-environment field trials to evaluate the stability of different genotypes. To incorporate the reality of environment and G×E interaction within GS models, with the goal to improve their prediction accuracy over models originally based on single environment or entry means across environments, several different strategies have been attempted (Burgueño et al. 2012; Crossa et al. 2015; Dawson et al. 2013; Jarquín et al. 2014; Lopez-Cruz et al. 2015; Jarquín et al. 2017; Haile et al. 2020). For example, environment was incorporated in a factor analysis model in which genetic and environmental covariance matrices were included (Burgueño et al. 2012). Heslot et al. (2014) incorporated stress covariates selected from a larger set of daily weather variables into GS models for winter wheat using an ensemble learning method. Reaction Norm models are another way to account for environmental factors and G×E in GS, where interactions between large numbers of markers and environmental variables (i.e. high dimensional data sets) are modelled using covariance functions (Jarquín et al. 2014; Jarquín et al. 2017; Roorkiwal et al. 2018; Haile et al.

2020). Lopez-Cruz et al. (2015) used a marker × environment (M×E) G-BLUP model to model G×E, which was extended by Crossa et al. (2015) by using a prior probability distribution (aka priors) that produces shrinkage and variable selection using Bayesian models. These studies showed that including environment and/or G×E in GS models offered better, or equivalent prediction accuracy compared to single environment models. However, prediction accuracy was noticeably higher for lines that were previously tested in correlated environments upon which the GS model was built, as compared to newly developed lines that had never been evaluated (Burgueño et al. 2012; Jarquín et al. 2014; Lopez-Cruz et al. 2015; Jarquín et al. 2017; Haile et al. 2020).

Strategies that include additional information, such as pedigree and marker significance, in GS models to improve prediction accuracy have been attempted (Burgueño et al. 2012; Crossa et al. 2015; Jarquín et al. 2014; Lopez-Cruz et al. 2015; Pérez-Rodríguez et al. 2015; Jarquín et al. 2017; Lopes et al. 2017; Montesinos- Lopez et al. 2019; Ankamah-Yeboah et al. 2020). One such strategy, called multi-trait GS, incorporates information from genetically correlated traits to the trait being predicted in the GS model and has been shown to provide better prediction accuracy than single-trait methods, especially for traits with low heritability (Jiang et al. 2015; Schulthess et al. 2016). The multi-trait whole genome prediction method was shown to be more accurate and robust than single-trait methods in simulation studies and those using empirical data in mice (Jiang et al. 2015). Schulthess et al. (2016) demonstrated an improvement in prediction accuracy for the lower heritability trait of protein content using multi-trait GS when the training population size was very low (40 lines).

The objectives of this study were to: 1) evaluate the impact on prediction accuracy of four different models that incorporate environmental information into GS models, and 2) to examine if including correlated traits in a multi-trait GS model would increase predication accuracy.

5.2 Materials and Methods

5.2.1 Incorporating environmental information in GS models

Four different approaches to include environment and G×E interaction into GS models were evaluated for their impact on prediction accuracy for yield. The models tested were termed single environment model (SE), across environment model (AE), M×E model (M×E) and Reaction Norm model (RNM). The 2,587 filtered 6K SNP dataset described in Chapter 3 was used as the genotypic data for these models. Prediction accuracies for the SE, AE and M×E models were evaluated using two different definitions of environment. Firstly, yield prediction accuracy of these three models were evaluated for five individual sites distributed across the western Canadian oat growing region. The sites were Beaverlodge, AB, Lacombe, AB, Melfort, SK, Indian Head, SK and Portage, MB. A subset of 161 WCORT lines (1,231 datapoints) grown in all five sites was used for these analyses. Secondly, yield prediction accuracy of the three models were evaluated for four oat western Canadian mega-environments. The mega-environments were identified among all 125 sites within the 2002-2014 period, regardless of year, by Ward's hierarchical clustering based on the significant environmental variables identified in Chapter 4 that impacted yield. These variables were growing season precipitation, cumulative sunlight hours, June mean temperature, July maximum temperature and July mean temperature. A subset of 119 WCORT lines (1,984 data points) grown in all four mega-environments was used for these analyses.

For the SE model, each site was treated as an individual environment and the effect of year was ignored. Marker effects were estimated for each individual environment and a linear model was used in which phenotype was regressed on markers separately for each environment. The model is described in Equation 5.1:

$$y_j = \mu_j + X_j\beta_j + \varepsilon_j \qquad \text{(Equation 5.1)}$$

where μ_j is the intercept, X_j is a matrix of marker centered and standardized genotypes, β_j is a vector of marker effects and ε_j is a vector of model residuals. Normal distribution and independence were assumed for marker effects and model residuals: $\beta_j \sim N\left(0, I\sigma^2_{\beta_j}\right)$ and $\varepsilon_j \sim N\left(0, I\sigma^2_{\varepsilon_j}\right)$. The model was fit using the R package BGLR (v. 1.0.8; Perez and de los Campos, 2018).

For the AE model, marker effect was considered constant across environments and estimated using all phenotypes across environments. The matrix notation for this model is described in Equation 5.2:

$$\begin{bmatrix} y_1 \\ \vdots \\ y_j \end{bmatrix} = \begin{bmatrix} 1\mu_1 \\ \vdots \\ 1\mu_j \end{bmatrix} + \begin{bmatrix} X_1 \\ \vdots \\ X_j \end{bmatrix} \beta + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_j \end{bmatrix} \qquad \text{(Equation 5.2)}$$

where $\begin{bmatrix} 1\mu_1 \\ \vdots \\ 1\mu_j \end{bmatrix}$ is a vector of intercepts, $\begin{bmatrix} X_1 \\ \vdots \\ X_j \end{bmatrix}$ is a matrix of marker centered and standardized genotypes for each of the j environments, β is a vector of marker effects estimated across all environments and ε_j is a vector of model residuals. Normal distribution is assumed for marker effect $\beta \sim N(0, I\sigma^2_\beta)$ and model residuals $\varepsilon_j \sim N\left(0, I\sigma^2_{\varepsilon_j}\right)$. The model was fit using the R package BGLR (v. 1.0.8).

For the M×E model, an interaction term between marker and environment that borrowed information across environments yet allowed for the marker effects to change across environments, was used. This model is described in Equation 5.3

$$\begin{bmatrix} y_1 \\ \vdots \\ y_j \end{bmatrix} = \begin{bmatrix} 1\mu_1 \\ \vdots \\ 1\mu_j \end{bmatrix} + \begin{bmatrix} X_1 \\ \vdots \\ X_j \end{bmatrix} b_0 + \begin{bmatrix} X_1 & 0 & & \cdots & 0 \\ 0 & & & \cdots & 0 \\ & & \ddots & & \\ \vdots & \vdots & & & \vdots \\ 0 & 0 & & \cdots & X_j \end{bmatrix} \begin{bmatrix} b_1 \\ \vdots \\ b_j \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_j \end{bmatrix} \qquad \text{(Equation 5.3)}$$

where $\begin{bmatrix} 1\mu_1 \\ \vdots \\ 1\mu_j \end{bmatrix}$ is a vector of intercepts, $\begin{bmatrix} X_1 \\ \vdots \\ X_j \end{bmatrix}$ is a matrix of marker centered and standardized genotypes for each of the j environments, b_0 is main marker effects estimated across all environments, $\begin{bmatrix} b_1 \\ \vdots \\ b_j \end{bmatrix}$ is the deviation of main marker effect due to M×E interaction and $\boldsymbol{\varepsilon}_j$ is a vector of model residuals. Marker effects, M×E interaction effects and model residuals are assumed to be normally distributed: $b_0 \sim N(0, I\sigma^2_{b_0})$, $b_j \sim N\left(0, I\sigma^2_{b_j}\right)$ and $\varepsilon_{ij} \sim N\left(0, I\sigma^2_{\varepsilon_j}\right)$. The model was fit using R package BGLR (v.1.0.8).

A RNM model was tested in which interactions between markers and individual environmental variables were used to model G×E using covariance functions (Jarquín et al. 2014 and 2017). Jarquín et al. (2014) developed a RNM method based on G-BLUP and include interaction effects between markers and environmental covariates (ECs) using covariance structures. We evaluated a RNM using the 305 WCORT lines grown between 2002-2014 as described in Chapter 3. This dataset comprised 9-13 sites per year over 13 years for a total of 125 individual environments and a total of 4,132 yield observations. The 2,587 filtered 6K SNP data set described in Chapter 3 were used for these models. The five significant ECs that were identified in the SEM study in Chapter 4 were used. These were May to July precipitation, cumulative sunlight hours during growing season, June mean temperature, July mean temperature, and July maximum temperature.

The basic model, is equivalent to the G-BLUP model and is described in Equation 5.4

$$y_j = \mu + \mathrm{g}_j + \varepsilon_j \qquad \text{(Equation 5.4)}$$

where y_j is the value of the phenotypic trait for the jth genotype; g_j is an approximation of the true genetic value of the jth line estimated using marker-derived genomic relationship matrix; ε_j represents the residuals.

Other models were built on top of the basic G model where random effects of site ($\mathrm{S_k}$), year ($\mathrm{Y_t}$), weather covariates ($\mathrm{w_{ij}}$) derived from environmental variables (ECs), environment being the combination of site and year (E_i) and their interactions with markers were added to the basic G model to develop the RNMs listed in Table 5.1. Normal distribution and independence were assumed for each effect and model residuals. The most complicated RNM in this study was the GWE-G×E-G×W, where main effects of marker, environment, weather covariates, plus the interaction between markers and environment, and the interaction between markers and ECs were included.

Table 5. 1 Ten Reaction Norm models used to incorporate various descriptors of environment and their interactions.

Model	**Equation**
G	$y_j = \mu + \mathrm{g}_j + \varepsilon_j$
GE	$y_{ij} = \mu + \mathrm{g}_j + E_i + \varepsilon_{ij}$
GS	$y_{kj} = \mu + \mathrm{g}_j + S_k + \varepsilon_{kj}$
GY	$y_{tj} = \mu + \mathrm{g}_j + Y_t + \varepsilon_{tj}$
GW	$y_{ij} = \mu + \mathrm{g}_j + w_{ij} + \varepsilon_{ij}$
GE-G×E	$y_{ij} = \mu + \mathrm{g}_j + E_i + \mathrm{g}E_{ij} + \varepsilon_{ij}$
GS-G×S	$y_{kj} = \mu + \mathrm{g}_j + S_k + \mathrm{gS}_{kj} + \varepsilon_{kj}$
GY-G×Y	$y_{tj} = \mu + \mathrm{g}_j + Y_t + \mathrm{gY}_{tj} + \varepsilon_{tj}$
GW-G×W	$y_{ij} = \mu + \mathrm{g}_j + w_{ij}+ \mathrm{g}w_{ij} + \varepsilon_{ij}$
GWE-G×E-G×W	$y_{ij} = \mu + +\mathrm{g}_j + E_i + +w_{ij} + \mathrm{g}E_{ij} + \mathrm{g}w_{ij}+\varepsilon_{ij}$

Note: Normal distribution and independence were assumed for each effect and model residuals in every equation.

Prediction accuracy was estimated using 20 random partitions of the population as described by Crossa et al. (2016). In each partition, 70% of the lines were selected as the training population and the remaining 30% acted as the validation population to calculate prediction accuracy. Models were fit in the training population and prediction accuracy was calculated by computing the correlation between phenotypic values and GEBV. The mean of the correlations within the 20 partitions was calculated and reported as prediction accuracy. Two cross validation (CV) methods were used to mimic the selection scenarios in plant breeding programs: CV1 reflects the situation in which newly developed lines that have not been grown in any environment. The CV2 method mimics the problem that some lines may only be grown in a subset of environments. To select the validation population, lines representing 30% of all observations (lines × environment) were selected at random and then one environment per line was subsequently picked at random. For each CV technique and training-validation partition, inferences and predictions were based on 12,000 iterations collected from posterior distributions after discarding 2,000 samples as a burn-in.

5.2.2 Incorporating multiple trait information in GS models

Multi-trait GS was tested using the Bayesian multiple-trait multiple-environment (BMTME) model (Montesinos-López et al. 2019). BLUPs calculated from all 305 WCORT lines for each trait were used as multivariate phenotypic responses in the BMTME model. The 2,587 filtered 6K SNP data set described in Chapter 3 were again used as the genotypic data for this model and a design matrix for the genetic effects was generated using marker information. This data was used along with genetic correlation data calculated between traits to assess the variance-covariance structure among traits and genotypes. Finally, the genetic correlations between traits were used to cluster the traits into three sub-groups using Ward's hierarchical clustering method in Meta-R (Ward, 1963; Alvarado et al. 2020). The BMTME package in R was used to assess the BMTME model for each of the 12 traits described in Chapter 3 and the three trait sub-groups identified by Ward's clustering method.

5.3 Results and Discussion

5.3.1 Single environment, across environment and M×E models to predict oat yield

Four mega-environments were identified using significant environmental variables identified in Chapter 4 that impact yield (Fig. 5.1). GS models for oat yield using single environment (SE), across environment (AE) and M×E interaction models were compared for their prediction accuracies at five individual sites and within the four mega-environments using two different cross validation methods. On average, better prediction accuracy was observed for individual sites as compared to the mega-environments using either of the cross-validation methods (Table 5.2). However, it was worth noting that the individual sites were predicted using a larger population (161 lines) compared to the mega-environments (119 lines) which may have led to the better prediction accuracy. The SE model provided higher or equivalent prediction accuracy compared to the AE or M×E model for both the individual site (r=0.33, 0.27 and 0.31 for SE, AE and M×E, respectively) and mega-environments (r=0.21, 0.15, 0.16 for SE, AE and M×E, respectively) when using the CV1 method. The AE and M×E models showed comparable and significant increases in prediction accuracy compared to the SE model when using the CV2 method (Table 5.2). Prediction accuracy for the AE model averaged 0.59 (r= 0.38-0.78) across individual sites and 0.52 (r= 0.42-0.71) for mega-environments, while prediction accuracy for the M×E model averaged 0.61 (r=0.38-0.77) across individual sites and 0.51 (r=0.42-0.68) for mega-environments. Crossa et al. (2016) also observed that predictions with CV1 were lower than CV2 for grain yield. Because the AE and M×E models provided similar prediction results with both CV methods, it appears that including a G×E interaction is not an effective means to improve prediction accuracy of oat yield. This observation is in agreement with that of Dawson et al. (2013) in which no increase in prediction accuracy was found when a G×E interaction was included in GS models used to predict wheat yield within an unbalanced data set spanning 17 years, similar to the dataset used in this study. Similar observations were found by Haile et al. (2020) where incorporation of G×E interaction only increased prediction accuracy for low heritability traits (h^2<0.30), but had no effect for higher heritable traits (h^2>0.30) in three lentil populations.

It was interesting to note the differences in prediction accuracies between the two cross validation methods. For the SE model, CV1 and CV2 produced similar results. This is likely because the SE model does not borrow information from other environments which is the basis (and advantage) of the CV2 method, and thus the SE model could not make use of this strategy. The AE and M×E models performed better with the CV2 method as they incorporate information from multiple environments, which is analogous to the methodology of borrowing information used in the CV2 method (Burgueño et al. 2012).

Lopez-Cruz et al. (2015) revealed that the M×E model offered better prediction accuracy when a subgroup of environments with positive correlations between them were used. In this study, lines were not balanced in locations or mega-environments due to the nature of the WCORT dataset, thus correlations between locations or mega-environments could not be calculated.

In general, it appears that the AE model provided the simplest and highest prediction accuracy when used with the CV2 method. The M×E model was more complex without offering better accuracy.

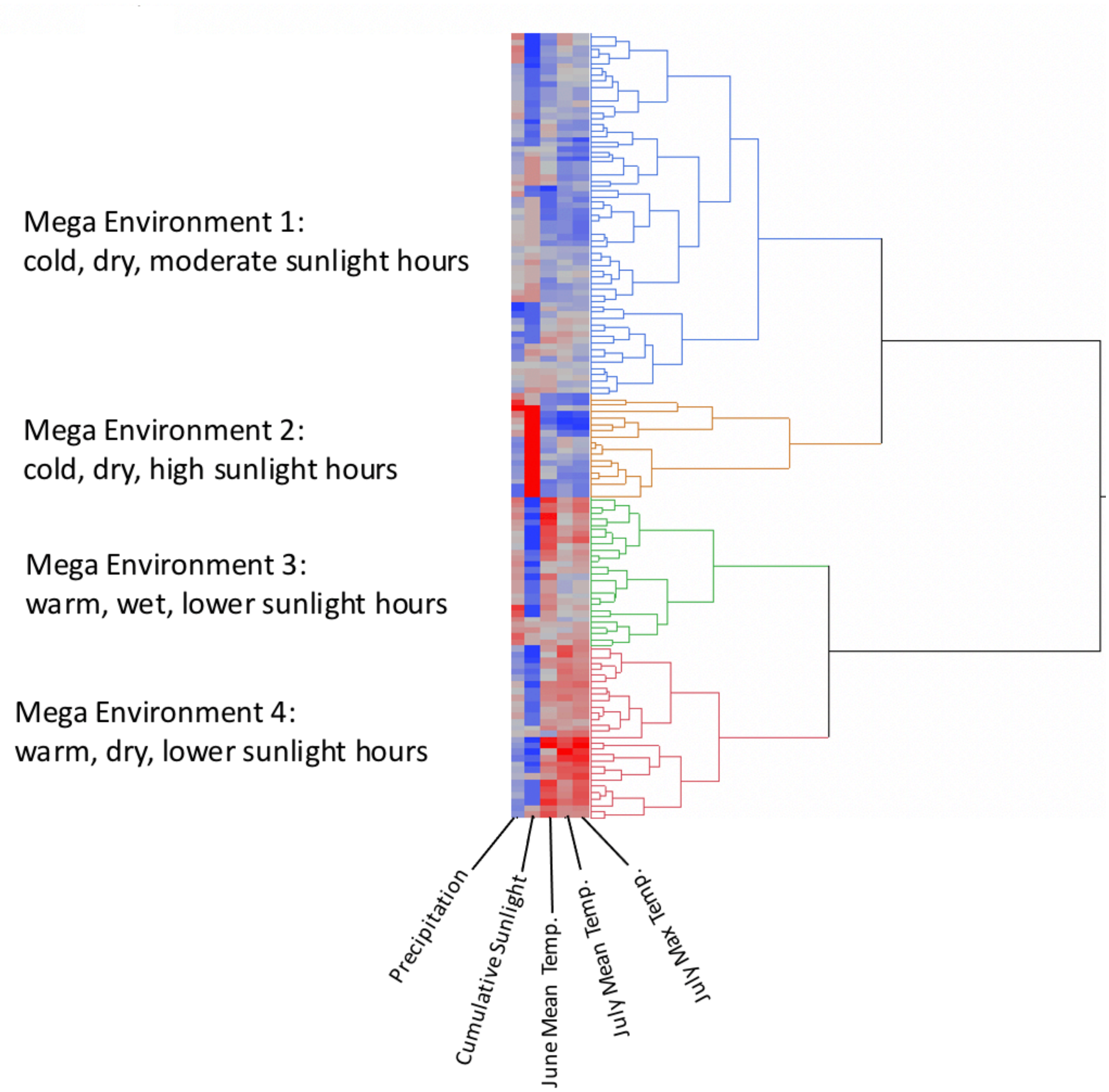


Fig. 5. 1 Clustering of WCORT sites into four mega environments using Ward's Hierarchical clustering based on the environmental variables May to July precipitation, cumulative sunlight hours, June mean temperature, July mean temperature and July maximum temperature.

Table 5.2 Prediction accuracy using individual sites and mega environments as target environments in single environment (SE), across environment (AE) and marker × environment (M×E) models.

	CV1			**CV2**		
Environment*	SE	AE	M×E	SE	AE	M×E
Individual Sites as Environments						
Beaverlodge, AB (237)	0.23	0.15	0.18	0.18	0.38	0.38
Indian Head, SK (251)	0.33	0.35	0.35	0.30	0.78	0.77
Lacombe, AB (279)	0.32	0.25	0.30	0.32	0.53	0.56
Melfort, SK (220)	0.38	0.30	0.34	0.36	0.67	0.70
Portage, MB (245)	0.40	0.31	0.36	0.43	0.61	0.65
Mean	0.33	0.27	0.31	0.32	0.59	0.61
Mega-environments as Environments						
Mega 1 (696)	0.20	0.12	0.14	0.18	0.52	0.50
Mega 2 (546)	0.20	0.12	0.15	0.23	0.42	0.43
Mega 3 (482)	0.33	0.25	0.27	0.33	0.71	0.68
Mega 4 (260)	0.09	0.11	0.10	0.08	0.44	0.42
Mean	0.21	0.15	0.16	0.20	0.52	0.51

*the number of oat lines grown in each environment is indicated in brackets.

5.3.2 Reaction Norm Models to predict oat yield

Prediction accuracy was lower for CV1 across all ten RN models (0.15 to 0.26) compared to CV2 (0.37 to 0.53) (Table 5.3). This agreed with numerous GS studies (Burgueño et al. 2012; Crossa et al. 2015; Jarquín et al. 2014 and 2017; Lopez-Cruz et al. 2015; Pérez-Rodríguez et al. 2015). Like the prior three models, it appears that the superior performance of the CV2 method is due to its ability to borrow of information from other environments (Burgueño et al. 2012; Crossa et al. 2015; Jarquín et al. 2014 and 2017).

Reaction Norm models that included main effects for genotype (G), site (S), year (Y), weather covariates (W) and their interactions were tested in this study. The inclusion of different variables to represent the effect of environment on yield offered little to no benefit when using the CV1 method (Table 5.3). For instance, including main effects related to environment, site, year or weather covariates did not provide better accuracy (r=0.16-0.19) compared to a model that only included a genotype main effect (r=0.21) (Table 5.3). Similarly, models that included interaction terms did not improve accuracy (r=0.15-0.19), except for the GW-G×W model which slightly increased accuracy to 0.26 (Table 5.3).

Table 5.3 Prediction accuracy for oat yield measured on 305 lines grown in the WCORT from 2002-214 using Reaction Norm models and two different cross validation methods.

Model	**CV1**	**CV2**
G	0.21	0.38
GE	0.18	0.49
GS	0.19	0.38
GY	0.16	0.39
GW	0.17	0.38
GE-G×E	0.18	0.49
GS-G×S	0.19	0.37
GY-G×Y	0.15	0.42
GW-G×W	0.26	0.44
GWE-G×E-G×W	0.18	0.53

When assessing prediction accuracy results using the CV2 method, some models which included environmental variables showed improvement over the model which only included genotype. Inclusion of the 125 environments (in the GE model) as a main effect was effective at increasing accuracy by 28% (r=0.49) comparing to the genotype model (r=0.38). However, adding a G×E interaction to the GE model did not increase accuracy further (r=0.49). Models which included site, year and weather covariates as main effects did not increase prediction accuracy (r=0.38-0.39) compared to the genotype main effect model. Inclusion of a G×S interaction to the GS model did not increase accuracy, while marginal increases in accuracy were observed when a G×Y interaction was added in the GY model (10% increase) and when a G×W interaction was included in the GW model (15% increase). The highest accuracy came from the most complex model in this study. When the three main effects of environment (E), genotype (G) and weather covariates (W) were included in a model with G×W and G×E interaction terms, then an increase in accuracy of 40% was observed compared to the simple genotype main effect model (Table 5.3).

These results agreed with the study of Jarquin et al. (2014) where inclusion of interaction terms between genotype and environment and environmental covariates (ECs) led to a 35% increase in prediction accuracy compared to models with only main effects when using an unbalanced data set consisting of 139 wheat lines grown in 340 environments. Similarly, they also found that the GE model performed nearly as well as the GE-GxE model. Mota et al. (2020) also observed that inclusion of a G×E interaction term along with main effects in RNM models increased prediction accuracy within a population of Brazilian beef cattle. The results of this study indicate that inclusion of meaningful environmental information and interaction of this information with genotypic data is crucial to increase prediction accuracy in GS models.

5.3.3 Multiple trait genomic selection

A multiple trait GS model was examined to determine if prediction accuracy for the 12 oat traits could be improved when compared to traditional single trait GS models. Genetic correlations among the 12 oat traits studied ranged from 0.00 (days to heading and oil) to -0.95 (plumpness and thins) (Table 5.4) and three main groups were identified among these twelve traits (Fig. 5.2), with thins, protein and β-glucan comprising Cluster A, days to maturity, days to heading, yield and height forming Cluster B, and kernel weight, test weight, plumpness, oil and groat percentage forming Cluster C.

Prediction accuracies for the 12 oat traits using the multiple trait GS model was similar or poorer than the corresponding single trait GS model, except with groat percentage and thins for which 7% and 13% increases in accuracy were observed, respectively (Fig. 5.3).

In an effort to improve prediction accuracies using multiple trait data, only data from traits within the same cluster was used in GS models, as opposed to using data from all 12 traits. Although this method did show better prediction accuracies than multi-trait GS for yield, days to heading, plant height, days to maturity, groat percentage, plumpness, protein, prediction accuracy was still equal to or poorer than single trait GS, except for groat percentage, protein and plumpness for which prediction accuracy was 57%, 27% and 69% higher, respectively (Fig. 5.3).

Table 5. 4 Genetic correlations among 12 traits using measured in 305 lines grown in the WCORT from 2004-2012 phenotypic observations from individual lines.

Traits	YLD	DH	MAT	HT	TWT	MKW	PLP	THN	GRT	PRO	OIL
DH	0.58										
MAT	0.58	0.45									
HT	0.60	0.45	0.31								
TWT	0.16	0.04	0.19	0.50							
MKW	0.51	-0.08	0.24	0.43	-0.15						
PLP	0.25	-0.34	0.08	-0.20	0.02	0.65					
THN	-0.39	0.16	-0.11	-0.25	-0.11	-0.55	-0.95				
GRT	-0.06	-0.34	0.32	-0.22	0.25	0.12	0.38	-0.34			
PRO	-0.49	-0.34	-0.35	-0.29	0.06	-0.06	-0.01	0.02	-0.04		
OIL	0.18	0.00	0.28	0.13	0.19	-0.03	-0.07	0.12	0.16	0.07	
BG	-0.23	-0.15	-0.20	-0.19	-0.05	-0.14	-0.32	0.30	-0.16	0.36	0.16

Note: YLD=yield, HD=days to heading, MAT=days to maturity, HT=plant height, TWT=test weight, MKW=thousand kernel weight, PLP=plumpness, THN=thins, GRT=groat, PRO=protein, OIL=oil, BG= β-glucan

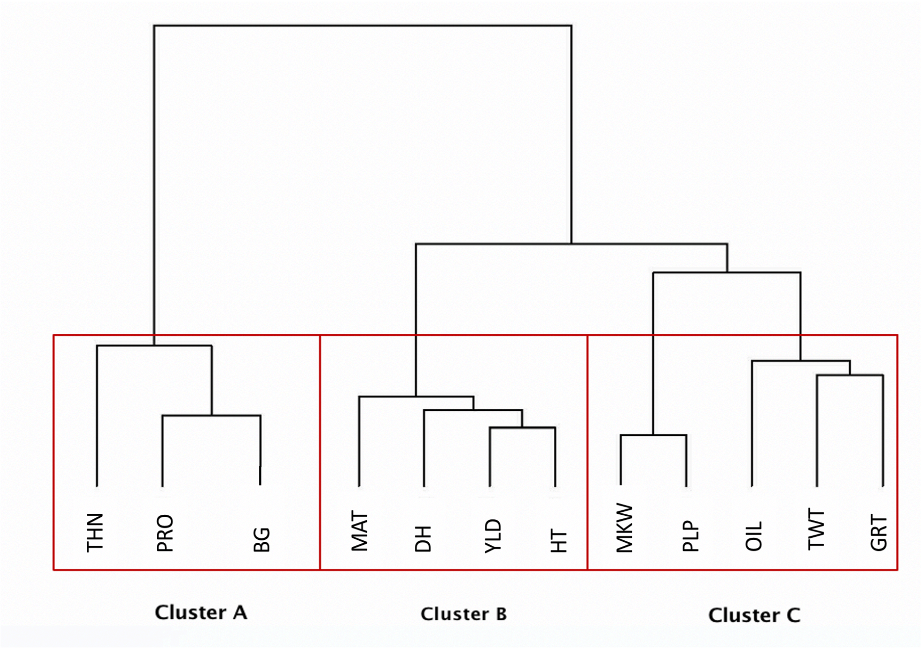


Fig. 5. 2 Clustering of 12 traits measured on 305 lines grown in the WCORT from 2002-2014 based on genetic correlations using Ward's hierarchical clustering method. Note: YLD=yield, HD=days to heading, MAT=days to maturity, HT=plant height, TWT=test weight, MKW=thousand kernel weight, PLP=plumpness, THN=thins, GRT=groat, PRO=protein, OIL=oil, BG= β-glucan.

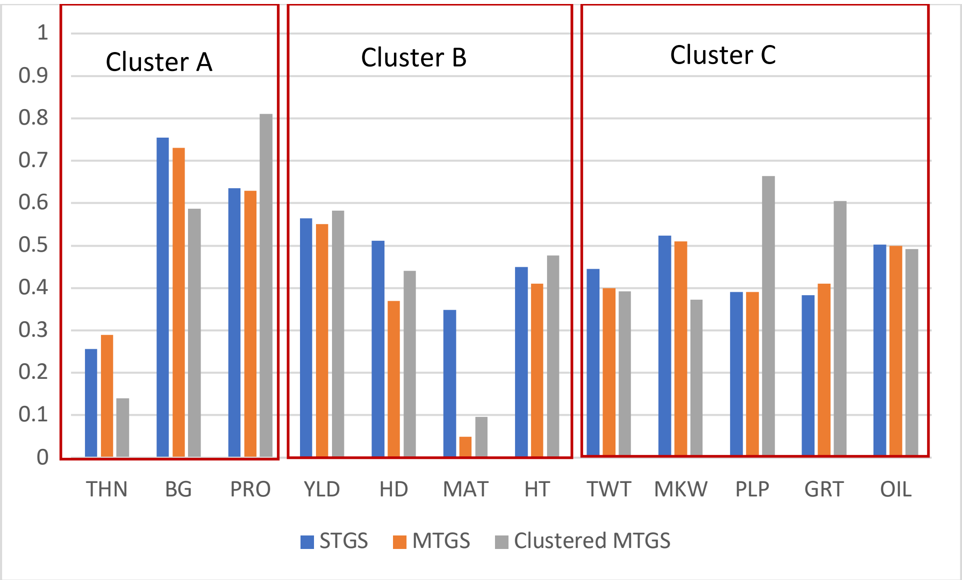


Fig. 5. 3 Comparison of prediction accuracy for 12 traits measured on lines grown in the WCORT from 2004-2012 using single trait (STGS), multiple trait (MTGS) and clustered GS models. Note: YLD=yield, HD=days to heading, MAT=days to maturity, HT=plant height, TWT=test weight, MKW=thousand kernel weight, PLP=plumpness, THN=thins, GRT=groat, PRO=protein, OIL=oil, BG= β-glucan.

Simulation studies indicate that multiple trait models are less effective compared to single trait models when there is less genetic correlation between traits (Jia and Jannink, 2012). In addition, higher heritability traits (i.e. h^2 = 0.5-0.8) showed little difference in prediction accuracy between STGS and MTGS models and did not improve as genetic correlation between traits increased. Only when a very low heritability (i.e. h^2=0.1) trait, controlled by a lower number of QTL (i.e. 20 versus 200), did the MTGS show any meaningful increase in prediction accuracy (Jia and Jannink, 2012). Importantly, improved prediction accuracy for a low heritability trait is only realized when information from well correlated and highly heritability traits is available for the MTGS model (Calus and Veerkamp, 2011; Guo et al. 2014; Jiang et al. 2015). Schulthess et al. (2016) confirmed these predictions to some degree for two high heritability traits in rye, protein content and grain yield (h^2=0.83 and 0.84, respectively), which showed little difference in prediction accuracy between STGS and MTGS. When heritability for protein content was lower (h^2=0.68) in a second rye population, MTGS only performed better when the training population size for the predicted trait was very low (40 lines), with no difference observed when training populations were the same between the correlated and predicted traits (Schulthess et al. 2016).

In the current study it was found that single trait GS offered equal or increased prediction accuracy in comparison to MTGS, even with the presence of genetic correlation among traits. This is likely due to the heritability of the traits evaluated in this study. Eleven of the 12 traits had heritability values exceeding 0.5, essentially equivalent to the high heritability trait in the Jia and Jannink (2012) study. Genetic architecture of the traits might be another reason that the single trait GS model offered better or similar results than the multiple trait GS model. Jia and Jannink (2012) reported that MTGS was less effective as the number of QTLs controlling a trait increased. The quantitative nature of the traits evaluated in this study, along with the diverse population in which they were measured, would certainly lead to the expectation that many different genes are responsible for influencing trait values.

5.4 Conclusion

In this study additional information related to environment (e.g. site, year, mega-environments, and environmental variables), G × E interactions and correlated traits were included in GS models to determine if this additional information would increase GS prediction accuracy. Different CV methods which mimicked different breeding scenarios were also evaluated. The SE model was comparable in prediction accuracy to the AE and M×E model when using CV1 which simulated the situation when newly developed lines have never been grown in any environment. In contrast, the AE and M×E models produced similar and better prediction accuracy than the SE model when using CV2 which borrows information from environments where lines have been tested in order to predict untested environment. When using individual sites to represent environment, improved prediction accuracy was found compared to using mega-environments that had been identified based on significant environmental variables discovered in Chapter 4. Reaction Norm models which included additional main effect information, such as individual environment, site (across years), year, and weather variables, along with interaction terms, offered no improvement compared to the base model with only a genetic main effect when using CV1. However, certain models with environment and interaction terms (GE and GWE-G×E-G×W) provided better prediction accuracy when using CV2. Finally, inclusion of multiple and correlated trait data in the GS model was also tested, but no significant improvement was discovered for most of the traits.

It is important to note that environment plays a significant role on lower heritability, quantitative traits like yield such that environment-specific GS models targeting specific growing regions that share similar weather, soil or daylength characteristics, might be a viable strategy to develop cultivars for different regions. In addition, with the increasing amount of weather and environmental data collected, deep learning algorithms which excel at capturing unnoticed/undefined relationships within large datasets, could offer improvements in GS models that attempt to incorporate environmental data.

CHAPTER 6. EVALUATION OF AN OAT GENOMIC SELECTION MODEL USING EMPIRICAL DATA

6.1 Introduction

Evaluating different factors known to influence GS prediction accuracy, such as statistical model (Goddard, 2009; Habier et al. 2011;Meuwissen et al. 2001; Yi and Xu, 2008), marker density (Heffner et al. 2010), marker platform (Solberg et al. 2008), population structure (Habier et al. 2007), training population (TP) size (Asoro et al. 2011; Heffner et al. 2011) and environment interactions (Dawson et al. 2013; Jarquin et al. 2014), provides useful information towards the development of the most effective model. However, cross validation (CV) of any GS model provides inadequate information on the performance of the GS model in external breeding populations (BP) since the CV process randomly splits the population used to develop the model into TP and validation populations (VP). Indeed, overestimation of prediction accuracy can occur if the VP is more related to the TP than the BP on which the GS model will be used (Wray et al. 2013). Ensuring that the design of the TP is relevant to the targeted BP, and that the CV methods are similar to the selection scenarios within the targeted breeding programs, are important criteria to ensure that GS is meaningful (Daetwyler et al. 2013). Relatedness between the TP and BP has been shown repeatedly to be a crucial factor to ensure satisfactory GS prediction (Clark et al. 2012; Habier et al. 2007 and 2010; Riedelsheimer et al. 2013). When TPs were not related to BP, accuracies were diminished (Charmet et al. 2014; Crossa et al. 2014; Riedelsheimer et al. 2013; Windhausen et al. 2012) because allelic combinations within the breeding population were infrequently represented in the TP (reviewed by Bassi et al. 2016). In a simulation study by Habier et al. (2007), linkage disequilibrium (LD) between markers and QTL, and the ability of markers to accurately describe the genetic relationships within BP were important factors in realizing GS prediction accuracy. The objectives of this study were to: 1) test the prediction accuracy of the GS model developed in Chapter 3 in two different breeding populations using empirical data and, 2) determine potential reasons behind prediction differences in the two breeding populations.

6.2. Materials and Methods

6.2.1 GS model evaluation in a biparental population

The first population used to assess the GS model was a biparental population consisting of 158 F4:7 RILs derived from the cross ‘CDC Dancer’ × ‘AC Morgan’ (DM). The population was grown in a two replicate lattice design at the Kernen Crop Research Farm (KCRF) located near Saskatoon, SK (52°09’06”N, 106°31’41”W, 486m, orthic dark brown, clay-clay loam) from 2014-2017. ‘AC Morgan’ was registered in 2000 and was developed from the cross OT526 × OT763 by the Lacombe Research Centre, Agriculture and Agri-Food Canada (Lacombe, Alberta) (Kibite and Menzies, 2001). It is a medium maturing variety with high yield and desirable grain features including high protein, low oil, low hull percentage, and plump kernels (Kibite and Menzies, 2001). ‘CDC Dancer’ was derived from the cross OT344 × OT269 and was registered in 2000 by the Crop Development Centre (Saskatoon, Saskatchewan). It is a medium maturing variety, with lower yield than ‘AC Morgan’, but it has excellent grain and milling quality. Data was collected for yield, days to heading, groat percentage, kernel weight, oil, protein, thins and test weight (as described in Table 3.1). Best linear unbiased predictions (BLUPs) were used for the phenotypic data. Genotyping data was obtained using the iSelect oat 6K SNP Chip (Tinker et al. 2014).

The rr-BLUP statistical model was used to create the GS model. Two different TPs were used. First, all 305 lines grown in the WCORT from 2002-2014, as described in Chapter 3, were used. Second, a subset of the 305 lines consisting of 267 WCORT lines grown at KCRF from 2002-2014 were used. Prediction accuracy was calculated as the correlation coefficient between observed values and GEBVs. Principle component analysis on the DM population was conducted using the 6K SNP dataset with the R package ggfortify (v.0.4.11). LD decay was estimated using the method described in Chapter 3, section 3.2.4.

As genotypes were grown in the same location every year, trait heritability was calculated using Equation 6.1:

$$h^2 = \frac{\sigma^2{}_g}{\sigma^2{}_g + \frac{\sigma^2{}_{gy}}{n} + \frac{\sigma^2{}_e}{nr}} \qquad \text{(Equation 6.1)}$$

, where $\sigma^2{}_g$, $\sigma^2{}_{gy}$ and $\sigma^2{}_e$ are the genotypic (additive), genotypic × year, and residual variances, respectively, n is the number of years in which the lines were tested, and r is the number of replicates per environment.

6.2.2 GS model evaluation in an association population

The second population used to assess the GS model consisted of 48 elite breeding lines that were grown in the WCORT from 2015 to 2017 (WCORT 15-17). Data were collected for yield, days to heading, height, maturity, kernel weight and test weight (as described in Table 3.1). Best linear unbiased predictions (BLUPs) were calculated for each trait as described in Chapter 3. Genotyping data was obtained using the iSelect oat 6K SNP Chip (Tinker et al. 2014) and a set of high-quality markers were created using the same filtering process described in Chapter 3. The rr-BLUP statistical model was used to create the oat GS model using the 305 lines grown in the WCORT from 2002-2014 as the TP. Prediction accuracy was calculated as the correlation coefficient between observed values and GEBVs. Principle component analysis was conducted using the 6K SNP dataset with the R package "ggfortify" (v.0.4.11) (Horikoshi et al. 2020). LD decay was estimated using the method described in Chapter 3, section 3.2.4 and heritability was estimated using Equation 4.3 described in Chapter 3, section 3.2.3.

6.2.3 Performance comparison of random selection and genomic selection

Random selection and genomic selection for yield were compared in the DM and WCORT 15-17 populations. The best 15% of lines within each population were selected using each method. For random selection, the population was randomly selected using the sample function in R package "dplyr" (v.1.0.5; Wickham et al. 2018). To avoid sampling bias, this process was repeated 20 times and the mean result was reported. Genomic estimated breeding values derived from the oat GS model were used for genomic selection. The mean values for each trait of the top 15% lines were calculated for comparison.

6. 3 Results and Discussion

6.3.1 GS evaluation in the DM population

Prediction accuracy in the DM population was poor when the 305 line WCORT population was used as the TP. Prediction values in the DM population for all traits was below 0.27, significantly lower than prediction values obtained for these traits previously in the TP (Fig. 6.1). There are several possible reasons for the low prediction accuracies observed. GS accuracy can be influenced by several factors, including but not limited to, heritability of the traits being predicted (Combs and Bernardo, 2013), the extent of LD (Heffner et al. 2010), marker type (Solberg et al. 2008), statistical model used (Spindel et al. 2015), marker density (Heffner et al. 2010), G×E interaction between the TP and BP target environments (Dawson et al. 2013; Jarquin et al. 2014), genotype imputation method (Rutkoski et al. 2013) and critically, the genetic relationship between the TP and BP (Habier et al. 2007). Heritability of yield, protein, groat percentage, kernel weight and oil content were slightly lower in the DM population, thins and test weight were significantly lower in the DM population and days to heading was slightly higher in the DM population as compared to the TP (Table 6.1). As such, the lower heritability might be a possible explanation for the poorer prediction accuracies for thins and test weight, but unlikely to explain the poor results observed for the other traits.

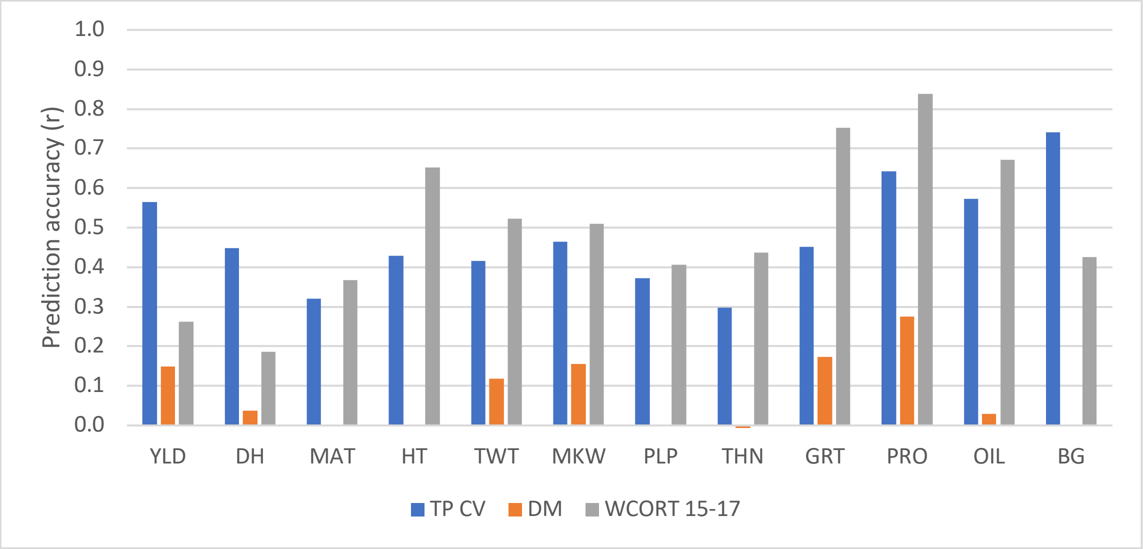


Fig. 6. 1 Prediction accuracy of the GS model in the DM population (684 markers), the WCORT 15-17 population (1,152 markers) and TP (2,587 markers). Note: TP CV=Cross validation of training population, YLD=yield, HD=days to heading, MAT=days to maturity, HT=plant height, TWT=test weight, MKW=thousand kernel weight, PLP=plumpness, THN=thins, GRT=groat, PRO=protein, OIL=oil, BG= β-glucan.

Table 6. 1 Heritability estimates obtained for traits in the training population (TP), the DM population grown from 2014-2017 and the WCORT lines grown from 2015-2017.

Trait	**TP**	**DM**	**WCORT 15-17**
YLD	0.51	0.41	0.75
HD	0.60	0.75	0.94
MAT	0.40	-	0.88
PH	0.68	-	0.93
TWT	0.62	0.36	0.78
MKW	0.89	0.78	0.45
PLP	0.79	-	0.74
THN	0.71	0.32	0.78
GRT	0.71	0.64	0.85
PRO	0.80	0.75	0.97
OIL	0.95	0.89	0.68
BG	0.94	-	0.95

Note: YLD=yield, HD=days to heading, MAT=days to maturity, HT=plant height, TWT=test weight, MKW=thousand kernel weight, PLP=plumpness, THN=thins, GRT=groat, PRO=protein, OIL=oil, BG= β-glucan

A similar degree of LD is required between the TP and BP to achieve good prediction accuracy in GS (Bassi et al. 2016). LD in the DM population was twice as large (20 cM) compared to the TP (10 cM) (Fig. 6.2a). The larger LD in the DM, typical when comparing bi-parental populations to association populations, was due to the lower number of recombination events observed in such populations. The larger LD in the DM population would have resulted in poor marker effect estimation for loci which had no or little recombination between then in the DM population. Increasing marker density could potentially increase GS prediction accuracy according to Heffner et al. (2010), however Riedelsheimer et al. (2013) pointed out that marker numbers only minimally affected prediction accuracy in bi-parental maize populations once every QTL had at least one marker in LD. Instead of marker numbers, markers distributed across the genome based on haplotype blocks might have more influence on the prediction accuracy.

The most likely explanation for the poor prediction accuracy in the DM population is the poor genetic relationship between the DM and TP (Fig. 6.3a). PCA analysis revealed very limited overlap between the two populations. This conclusion is supported by numerous other studies. For example, the lack of prediction accuracy in a simulation study using bi-parental cattle populations was mainly due to the lack of genetic relationship captured by markers in LD with QTLs (Habier et al. 2010). In a GS study in sorghum, prediction accuracy was poor when a low genomic relationship between TP and BP was discovered (Hunt et al. 2018). Higher prediction accuracy was observed when the training data had a strong relationship to the tested data in a training population optimization study in wheat and rice (Isidro et al. 2015). Better prediction could be achieved when BP and TP are better related through the inclusion of parents, siblings and other relatives (Daetwyler et al. 2014; Riedelsheimer et al. 2013; Zhao et al. 2013). In this study, despite using a TP that included the parents CDC Dancer and AC Morgan, prediction was still poor suggesting that the lack of relatedness of the TP and DM population was too extreme. In a simulation study, Meuwissen (2009) reported greatly increasing TP size and marker density could help achieve reasonable accuracy in unrelated BPs. However, these conditions were not available in this study. Similar linkage phases between available markers and QTL in both BP and TP are needed to obtain proper marker effect estimation (Goddard and Hayes, 2007). It is likely that this requirement was not met with the training and DM population.

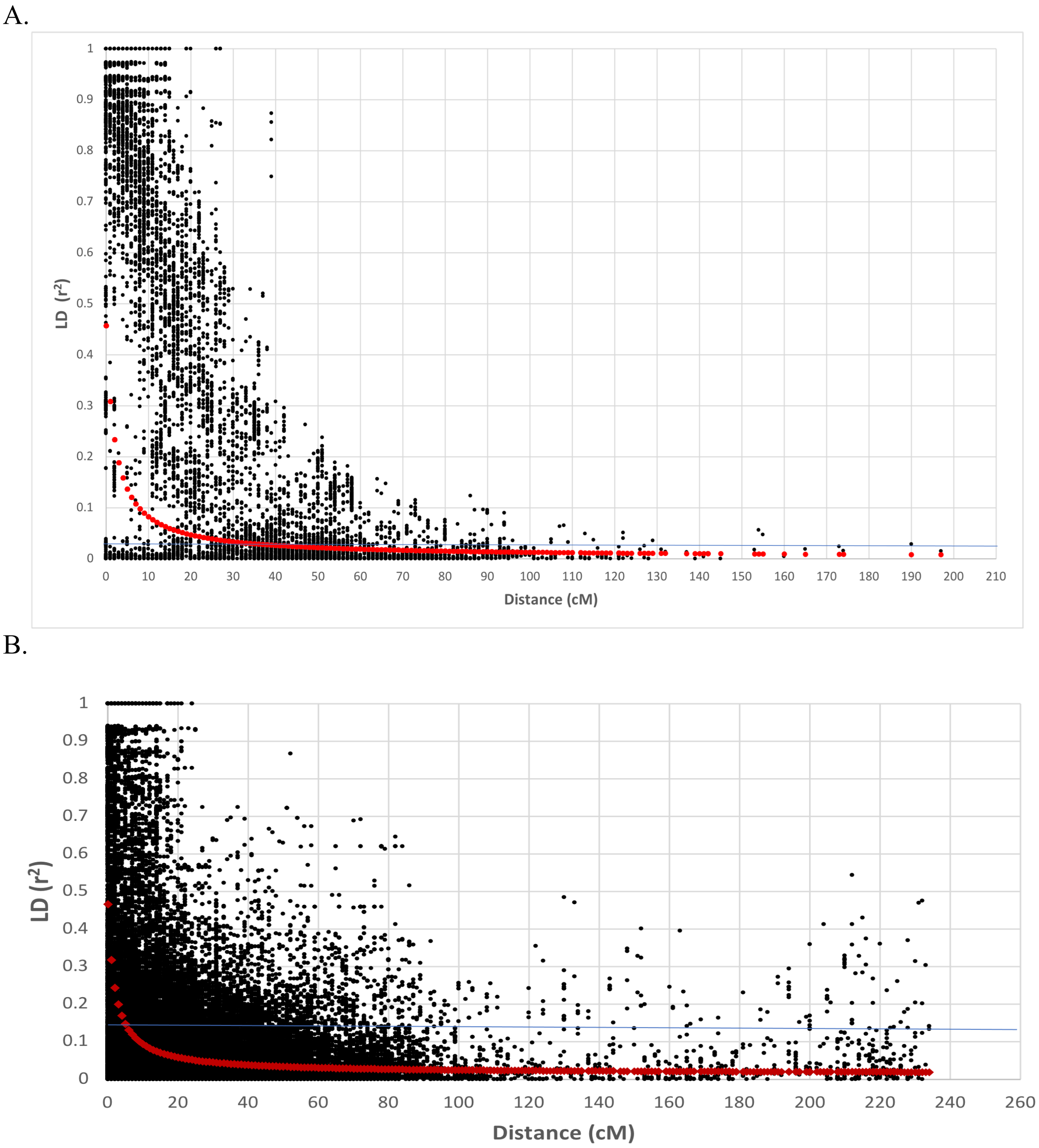


Fig. 6. 2 Genome-wide LD decay in the DM population (A) and WCORT 15-17 population (B). The fitted non-linear regression line (red) indicates LD decay, whereas the critical value line (blue) describes the 95th percentile of unlinked r^2 values.

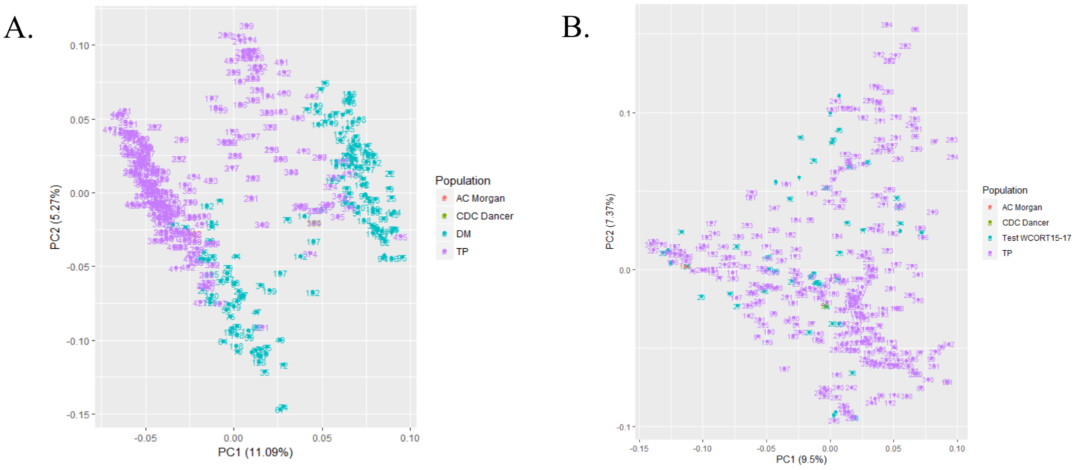


Fig. 6. 3 PCA analysis of the DM population (based on 684 markers) (A) and the WCORT 15-17 (based on 1,152 markers) (B) showing their relationships to the TP.

To determine if the different environments in which the TP was grown impacted the prediction accuracy in the DM population, a second TP was used in which only data from KCRF was used to create the GS model. However, this GS model provided worse prediction accuracy than the first model for all traits except for thins (Fig. 6.4). This finding agreed with the observation in Chapter 5 which showed that single environment GS models provided the poorest prediction accuracy when no information could be borrowed from other environments.

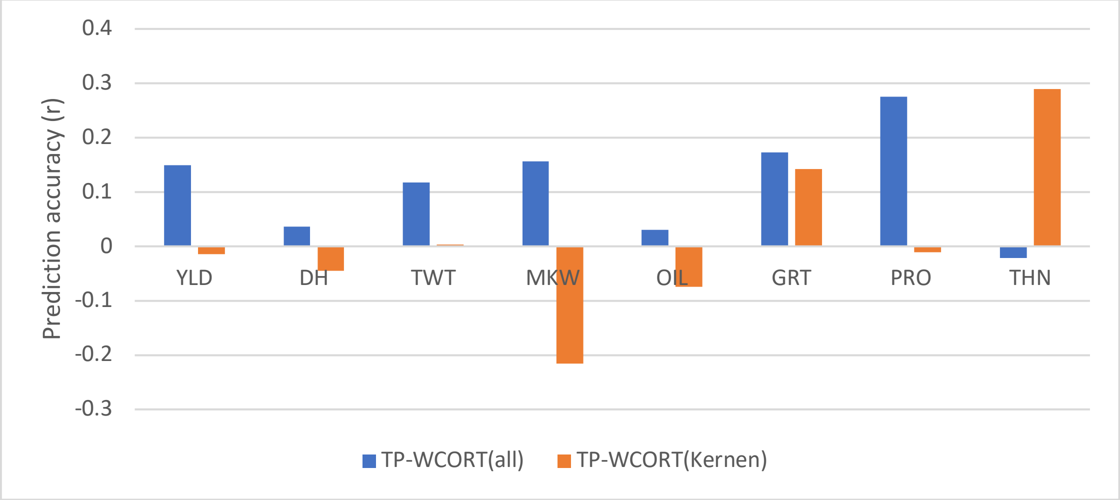


Fig. 6. 4 Comparison of GS prediction accuracy (6K+BLUP) in the DM population when the training population consisted of either the 305 WCORT lines grown at multiple locations from 2002-2014, or 267 WCORT lines grown at KCRF from 2003-2009 and 2011-2014. Note: YLD=yield, HD=days to heading, MAT=days to maturity, HT=plant height, TWT=test weight, MKW=thousand kernel weight, PLP=plumpness, THN=thins, GRT=groat, PRO=protein, OIL=oil, BG= β-glucan.

6.3.2 GS evaluation in the WCORT 15-17 population

Prediction accuracy within the WCORT 15-17 population was much improved over the DM population for all traits (Fig. 6.1). The higher prediction values are likely the result of two differences between the WCORT 15-17 and DM populations, that is, LD and its genetic relationship to the TP. The LD value for the WCORT 15-17 population was 7 cM (Fig. 6.2b), much closer to the value observed in the TP population, and as mentioned above this may have allowed for better estimation of marker effects. More significantly, the WCORT 15-17 population appeared to have much better genetic relationship with the TP (Fig. 6.3b). This important factor would mean that the GS model developed in the TP would be much more appropriate in the WCORT 15-17 population, as compared to the DM population.

6.3.3 Performance comparison of random selection and genomic selection

Random selection (RS) was conducted to mimic breeder selections at the single plant selection stage, which is arguably random for quantitative traits like yield. The mean of ten random selections offered higher selection performance for yield in the DM population compared to GS (Table 6.2). This is a carry-on effect resulting from the poor prediction accuracy of the GS model in the DM population. In the WCORT 15-17 population, GS performed better than RS for yield, even exceeding the best result from random selection (Table 6.2). The better performance of GS in this population is a result of the much better prediction accuracy obtained using the GS model. Indeed, GS might lead to improved performance and genetic gain in crop breeding. In a wheat GS simulation study, genetic gain tripled when using GS compared to phenotypic selection (PS) (Tessema et al. 2020). In an empirical study in soft red winter wheat, GS using rr-BLUP model was implemented on two biparental population using a subset of lines from a diverse panel as the TP. Response to selection for grain yield was compared among GS, PS, GS+PS and RM and showed that GS+PS provided the most response to selection, followed by PS, GS, and RM at selection intensity of 10% (Lozada et al. 2019). Another 2 year study involving wheat stem rust quantitative resistance showed that GS and PS resulted in equal rates of short-term gain, but GS reduced genetic variance more drastically (Rutkoski et al. 2015). These studies confirmed the potential of GS, but more validation and further studies are needed.

Table 6. 2 Comparison of random selection and genomic selection for yield within the DM and WCORT 15-17 populations using the oat GS model trained on the 305-line WCORT population.

	Breeding Population	
Selection Method	**DM**	**WCORT 15-17**
Random	4.60 (-10.8 to 34.6)[1]	6.43 (-87.33 to 50.40)
Genomic Selection	-2.72	60.54

[1] values were calculated as the mean BLUP of selected lines. The range of mean BLUPs is also indicated to the random selection method to show different possible outcomes from this method.

6.4 Conclusion

This study evaluated the potential of an oat GS model to improve genetic gain within two different breeding populations. Results from the biparental DM population showed that prediction accuracies for all eight traits evaluated were much poorer that were observed from cross validation in the 305 WCORT TP. Attempts to improve predictions by restricting the TP to data from the same location that the DM population was grown were not successful, again demonstrating that single environment-based GS models perform poorer when no information can be borrowed from other environments. In contrast, prediction accuracy was superior in the WCORT 15-17 population for 9 of the 12 traits evaluated in comparison to the results from cross validation in the TP. Not surprisingly, subsequent comparison of random selection versus GS showed that GS performed better than RS in the 15-17 WCORT population, but not in DM population.

The result from this study highlights the potential of GS to improve genetic gain within a breeding program. However, before conducting GS in a breeding program it is important to be reminded about the importance of evaluating certain aspects of the TP and BP, including the heritability and genetic architectures of traits of interests, LD in both populations, and the genetic relationship between the TP and BP. Given that many GS strategies may not include obtaining information about trait heritability, evaluation of LD and genetic relationship can serve as good indicators regarding the potential of a GS model to perform well in a given BP.

CHAPTER 7. GENERAL DISCUSSION AND CONCLUSIONS

7.1Investigation of genetic information important to genomic selection models

In Chapter 3 numerous critical factors, including phenotypic data, genotyping methods, marker density, GS models, population size, training population size and trait heritability, that are important for the creation of a GS model were evaluated in an unbalanced data set typical of that generated in breeding programs. We examined the performance of six GS prediction models (rr-BLUP, BCPi, RHSK, SVM, NN and RF) on 12 traits important to oat breeding programs. A five-fold cross validation scheme was implemented, and the prediction accuracy varied among traits from 0.39 to 0.72 (Fig. 3.7). The differences in prediction accuracy observed resulted from several factors, including heritability of the traits, genotyping method, phenotyping data format, and statistical prediction models. Overall, traits with higher heritability had better prediction accuracy (Fig. 3.8). The finding corresponded to previous literature (Combs and Bernardo, 2013; Heffner et al. 2011; Saatchi et al. 2010). With regard to the genotyping platform used, it was observed that the 6K SNP dataset offered improved prediction accuracy in comparison to the GBS SNP dataset (Fig. 3.7). Advantages of the 6K SNP dataset might be that it encompassed SNPs located in functional genes as most of the SNPs were derived from cDNA sequences (Tinker et al. 2014). Using BLUPs to represent phenotype provided better or equal results as entry-means and it is suggested they be used for unbalanced datasets (Fig. 3.7). BLUPs have been recommended in the literature when dealing with unbalanced data sets in plant breeding programs (Piepho et al. 2008). The rr-BLUP and SVM statistical models generally provided better results than the other models, although for many traits the various statistical models performed similarly, with some models performing better for some traits. This finding corresponded with past work that there is no one best prediction model in GS, especially when dealing with traits with different genetic architectures (Montesinos-López et al. 2018). The observed improved prediction accuracy using more advanced machine learning models, even though the decision mechanism behind the model was not clear, suggests the application of artificial intelligence in GS model creation would be a fruitful area of future research.

Factors such as marker density and training population size also influenced prediction accuracy. Increasing density above 500 markers did not provide noticeable improvement on prediction (Figs.

3.9 and 3.10) which is likely a result of the higher LD we found in this population (Fig. 3.6). The slow LD decay is likely due to the elite nature of the lines in the WCORT population that have been selected to perform agronomically in a defined geographic region and are required to meet certain disease and quality parameters. Increasing the training population size (from 61 to 244) improved the prediction accuracy by 18% to 80% among traits (Fig. 3.12). Overall, higher heritability, increased training population size and the quality rather than quantity of markers were three leading causes for improved prediction accuracy. In the future, the oat reference genome (PepsiCo, 2020) and practical haplotype graphs (described by Franco et al. 2020; Jensen et al. 2020) might be of assistance for imputation of missing GBS data in oats so that markers with good quality can be used in GS model. Combined with marker selection based on haplotype, GS using GBS markers might achieve better prediction accuracy than observed in this study.

Overall, this study provided an understanding of some of the important factors influencing GS models in oat. However, the prediction accuracy obtained using genetic information only might be improved by inclusion of environmental information which plays a large role in traits with low heritability, such as yield. We therefore examined important environmental factors influencing yield in Chapter 4 and investigated various GS models that incorporated environmental factors. The prediction values obtain in this study were limited to the training population used so validation of the model in actual breeding populations was assessed in Chapter 6.

7.2 Environmental factor selection and structure equation modeling for yield

Genomic selection was developed to assist selection in crop breeding, especially for traits that are resource intensive to evaluate and heavily influenced by environment, such as yield. In Chapter 4 significant and meaningful environmental factors that contributed to the variation observed in oat yield were investigated. The difficulty of modelling yield lies in the complexity of the trait, which interacts with both environment and other traits. Therefore, structural equation modelling (SEM) was implemented to model the direct and indirect effects of environmental and correlated trait variables. SEM offers a comprehensive statistical way to test the significance of hypothesized relationships among a system of variables (Hoyle, 1995). SEM requires an initial theoretical model supported by literature and scientific theory to create hypothesized relationships among variables which is then assessed using data (Fan et al. 2016). Correlation analysis was carried in the WCORT population out to identify traits important to yield formation, among which, days to heading, days to maturity and plant height were chosen due to their impact on vegetative and reproductive growth (Kaufmann 1961; May et al. 2004; Tibelius and Klinck, 1985; Wiggans and Frey, 1955). Similarly, correlation analysis was conducted with environmental factors related to precipitation and temperature due to their importance on crop production (Tashiro and Wardlaw, 1990, Stone and Nicolas, 1995; Wardlaw et al. 2002; Barnabás, et al. 2008). Initial theoretical models were developed based on literature and correlation results and goodness of fit was assessed by comparison of the variance-covariance matrix between hypothesized models and fitted observed models.

Although the WCORT population was helpful to identify traits correlated to yield, lines included in the population varied each year which made it difficult to assess the genetic and G×E components of yield. Therefore, three oat check varieties grown in the WCORT each year were chosen and evaluated individually in this study to eliminate the genetic and G×E variation so that only environmental factors and effects could be focused on. Significant paths were discovered in the SEM (Fig. 4.2) with high coefficients of determination (>0.7) for days to heading, days to maturity and yield which indicated that the variation observed in field could be largely explained by the proposed model. The relatively low coefficient of determination (<0.6) for plant height suggested that other contributing factors are missing from the model. Ultimately, seeding date, correlated traits (days to heading, days to maturity and plant height) and environmental factors (May to July precipitation, cumulative sunlight hours during the growing season, June mean temperature, July mean temperature, July max temperature) were determined to be most relevant to yield formation in each of the three oat check varieties. The relationship between these factors and yield among the three oat check varieties mirrored that of the larger WCORT population (Table 4.3).

This study revealed the importance of environmental factors, such as precipitation and temperature, and correlated trait interactions on yield formation. The inability to fully explain the variation in yield may have been due to limitations in the data available. Precise and high-throughput field phenotyping in the future could capture more additional and more frequent environmental and phenotypic data throughout the growing season that could improve the model

(Chawade et al. 2019, Zhao et al. 2019). As the era of big data continues to grow in the field of agriculture, additional information about plant growth and the environment will lead to new modelling methods, perhaps again with the aid of artificial intelligence. This will offer a broader, yet deeper view, on how these factors affect plant growth and yield in a systematic and comprehensive way.

7.3 Inclusion of environment, G×E and correlated trait information in GS models

One of the proposed benefits of GS is to predict important traits with lower heritability, such as yield. After examining various genetic and statistical factors that influence GS model prediction, Chapter 5 evaluated different methods to include environment, G×E and correlated trait information into GS models to improve prediction accuracy.

When considering the influence of environment, one has several options to categorize what is meant by environment. For example, sites across years, years, site-years combination and mega-environments (defined by common environmental variables) are possibilities. In Chapter 4, significant environmental variables that impact oat yield were defined, these were May to July precipitation, cumulative sunlight hours, June mean temperature, July mean temperature and July maximum temperature. These five environmental variables were used to cluster 125 environments into 4 mega-environments (Fig. 5.1). The use of individual sites versus mega-environments to represent environment were evaluated were then evaluated against three different models based on the method of Lopez-Cruz et al. (2015). These models were the single environment (SE), across environment (AE) model and marker by environment (M×E) models. Finally, two different cross-validation (CV) methods were compared, CV1 mimics the situation where prediction was made for newly developed lines that have not been tested in any environment, whereas CV2 mimics the situation where prediction was made for lines in field trials that have been tested in some environments, but not others (Burgueño et al. 2012). For both cross-validation methods, improved prediction accuracy was observed when individual sites were treated as environment as opposed to mega-environments (Table 5.2). The difference may be due to the larger population used when treating sites to represent environment in comparison to the population used for the mega-environments (161 vs 119 lines). When using CV1, SE gave comparable prediction accuracy as AE or M×E, as no borrowing of information from other environments was allowed with this method. However, when using CV2 the AE and M×E models revealed significant prediction accuracy increase in comparison to the SE model. The AE and M×E models had the ability to incorporate information from multiple environments which allowed them to take advantage of the borrowed information in the CV2 method (Burgueño et al. 2012). However, no obvious difference was found between the AE and M×E models. This agreed with the finding of Dawson et al. (2013) where no increase in prediction accuracy was observed when G×E interaction was included in GS models for wheat yield using an unbalanced dataset containing 17 years of field trial data. Similarly, incorporation of G×E interaction did not increase prediction accuracy for traits with heritability above 0.3 in a lentil GS study (Haile et al. 2020).

The last method used to incorporate environment and G×E information was the Reaction Norm model (RNM) in which covariance matrices related to genetic marker, environment, site, year, and weather were calculated, and their main effects and interactions were evaluated as described by Jarquín et al. (2014) (Table 5.1). The various RNM were cross validated with the same CV1 and CV2 methods. Little to no benefit was observed in GS models that included different main effects and interaction terms related to environment as compared to the basic genotype (G) model when using the CV1 method (Table 5.3). However, some models (GE and GWE-G×E-G×W) showed good improvement over the basic model when using the CV2 method. This finding supports reports by Jarquin et al. (2014) and Mota et al. (2020) where inclusion of a G×E interaction term along with main effects in RNM increased prediction accuracy. In Jarquin et al. (2014), very similar findings were discovered where EGW-G × WG × E offered the best accuracy over other models with main effects (EL, EG, ELW, EGW) or models with interactions (EGW-G × E, EGW-G × W) in CV2.

GS models with multiple traits were also tested to determine if prediction accuracy improved when borrowing information from other traits. Single trait GS, multiple trait GS and clustered multiple trait GS were compared. The single trait GS model performed as well, or better than multiple trait GS or clustered multiple trait GS for 9 of the 12 traits studied. In the study of Jia and Jannink (2012), multiple trait GS failed to provide better prediction accuracy when there was less genetic correlation between traits. This is unlikely to be the case in the current study as relatively good correlations existed among the traits, especially within their own cluster (Fig. 5.2

and Table 5.4). In addition, higher heritability traits (i.e. h^2 = 0.5-0.8) showed little difference in prediction accuracy between STGS and MTGS models and did not improve as genetic correlation between traits increased. Only traits with extremely low heritability (i.e. h^2=0.1) and a low number of explanatory QTLs showed improvement in prediction accuracy with MTGS, and only when correlated traits with high heritability were included in the model (Calus and Veerkamp, 2011; Jia and Jannink, 2012; Guo et al. 2014; Jiang et al. 2015). Most of traits in this study had heritability values exceeding 0.5, which could be the reason why no significant improvement in prediction accuracy was observed with the MTGS model.

This is the first study to include environment and G×E in an oat GS model and supports the idea that inclusion of meaningful environment information and interaction with genotype can increase prediction accuracy. Although it may be hard to implement models with environment and G×E interactions in breeding programs, it does offer some insights about the prediction power of GS in different scenarios. For instance, if one were to predict newly developed lines that have never been tested, we can expect a lower prediction accuracy, whereas a higher accuracy would be possible if a limited amount of performance data from one or a few environments was available. Additionally, because prediction accuracy was better when using individual sites to represent environments, as opposed to mega-environments, environment-specific GS models targeting specific growing regions that share similar weather, soil or daylength characteristics, might be a viable strategy to develop cultivars for different regions. Understanding the heritability and correlation among traits being predicted is important, and within the context of the breeding population used in this study, MTGS models are unlikely to be helpful due to the extreme, and perhaps unrealistic, set of conditions under which any benefits of the MTGS model can be realized. Finally, with the ever-increasing amount of weather and environmental data collected, deep learning algorisms which excel at capturing unnoticed/undefined relationships within large datasets, could offer improvements in GS models that attempt to incorporate environmental data.

7.4 GS model evaluation in two breeding populations

Cross validation within a training population offers limited insight as to how GS models might perform in different breeding populations that are generated within breeding programs every year. Therefore, Chapter 6 tested an oat GS model on two different breeding populations, a biparental

(DM) population consisting of 158 F4:7 RILs grown at one location from 2014-2017 and a diverse population consisting of 48 elite WCORT lines grown at multiple sites from 2015 to 2017. Prediction accuracy in the DM population was poor (<0.27) for all traits studied and lower than accuracies obtained in the original 305 WCORT TP. In an effort to improve the prediction accuracies, only TP data from the same site the DM population was grown was used. This produced poorer prediction which highlighted that GS models which can borrow information from more environments are superior to those which only use data from a single environment. Prediction accuracy in the WCORT 15-17 population was better than the original TP for 9 of the 12 traits. The poorer performance of the GS model in the DM population is in part due to the lack of consistent LD between the TP and BP, which is known to be essential for good predictions (Bassi et al. 2016). The larger LD in the DM population, an indicator of lower frequency of recombination, could also lead to improper marker effect estimation for loci within LD blocks (Fig. 6.2). Indeed, LD was one of the most significant contributing factors to prediction accuracy in situations where only weak genetic relationship between TP and BP exists (Liu et al. 2015). In this case, higher marker density, missing in the DM population, was needed to capture the LD between SNPs and QTL. The primary important factor influencing the poor prediction accuracy was the genetic relationship between the TP and DM populations. A very weak relationship was discovered between these two populations (Fig. 6.3a) and is likely the biggest reason for the poor prediction accuracies given the importance of having a close genetic relationship between the TP and BP (Habier et al. 2007；Habier et al. 2010; Zhong et al. 2009). Similar heritability of traits between TP and BP has also been noted to create better prediction accuracy (Hayes, 2009), but for 6 of the 8 traits evaluated heritability were similar so the factor may not be as critical a consideration for good prediction accuracy. By contrast, the WCORT 15-17 population demonstrated a much more similar LD and genetic relationship to the TP (Fig. 6.3.b) and these two characteristics would explain the superior prediction accuracies observed in the population. These results revealed the importance of evaluating the TP and BP before GS prediction, including careful examination of LD and genetic relationships between the two populations. This should be possible within breeding programs as the cost of genotyping continues to diminish. One might envision genotyping several dozen possible candidate populations, assessing LD and structure relative to the TP and selecting a subset of populations on which to apply GS.

The performance of random selection versus GS were also evaluated in these two populations by evaluating mean values of selected lines from both methods. Random selection mimicked the essentially random nature of selections for low heritability traits like yield that occurs in breeding programs during the single plant selection stage. As would be expected from the very poor prediction accuracies obtained in the DM population, GS failed to give better results than random selection. On the other hand, GS in the WCORT 15-17 population provided much better selection results than random selection. Again, this would be the direct result of the better prediction accuracies observed in this population. This experiment revealed the potential of GS to improve genetic gain, but only when properly applied within the context of there being a strong genetic relationship between the TP and BP.

7.5 GS for oat breeding, a guide to its application

This thesis began as an investigation into GS with the intention of assessing how it may be incorporated into an oat breeding program. Several fundamental concepts relevant to GS model accuracy, such as the importance of heritability, training population size, statistical models, marker quality and phenotypic format were assessed and provided guidance to breeders. From this work it was determined that to obtain good prediction power, one must keep several factors in mind before conducting GS: 1) the data associated with the TP is essential. Proper estimation of breeding values requires phenotypic data in the form of BLUPs and appropriate marker coverage of the genome (i.e. QTLs) associated with the traits of interest, 2) similar LD between the TP and BP and 3) proper composition of the TP to achieve close genetic relationship between the TP and BP. These three factors will determine the overall success of a GS-assisted breeding program.

Given the improvement in prediction accuracies observed when borrowing data from many environments (i.e., the AE model), versus site or mega-environment-specific data (i.e. the SE model), it appears that collecting as much phenotypic data as possible to build GS models is the best method to improve accuracy. It was also notable that prediction accuracies increased when the target region, that is, specific sites versus mega-environments, became more focussed. It leads to the question that if one selects lines based on a GS model built around lines grown at a specific site (but with borrowing of information from other sites), will lines selected by this model be site/regionally adapted, or will they also perform well across a wider geography? Typically having varieties with wide adaptation is desired, so a GS model that has high prediction accuracy, but results in regionally adapted varieties may not be desirable. This is a question requiring more investigation.

The studies also brought up several considerations for further implementation of GS in oat breeding. The most relevant consideration is at which stage in the breeding program should GS be applied so that time and resources efficiencies are gained and response to selection increased. Two directions might be considered. First, conduct cross-specific GS in which the same (bi-parental) population acts as both TP and BP. This strategy requires extensive work in earlier generations, including seed increase of a subset of the population (i.e. the TP) to obtain field data for low heritability traits like yield, so that a GS model can be developed. This is then followed by genotyping and creation of GEBVs for many lines from the same population (the BP) that have not been field tested. One might also concurrently begin the process of seed increase on the BP while field testing of the TP is occurring. This would allow the selected lines with the highest GEBVs from the BP to be more quickly field tested in replicated, multi-location field trials. Ultimately, this method would need to be assessed based on the competing considerations of the cost associated with increasing many lines, most of which will be discarded, versus the decreased time required to evaluate the GS-selected lines. If concurrent seed increase is not done, then the time lag associated with increasing selected lines from the BP must be outweighed by a vastly superior genetic gain realized through GS. The main advantage of this approach is that the close genetic relationship between TP and BP is guaranteed, which maintain high prediction accuracies.

Secondly, one could create a large, universal TP composed of thousands of breeding lines with phenotypic from multiple sites/years and good genotypic data. Before applying GS in the breeding program, a preliminary investigation of the genetic relationship and LD that exists between the TP and many potential bi-parental BP would be conducted. A subset of lines from the TP with closer relationship to a given BP could be selected for model fitting and marker effect estimation. This strategy is beneficial as the TP can grow over time and acts as a repository of "historical" breeding program information. Given that most varieties emanate from crosses between parents with several years of data associated with them, it would be reasonable to assume that an appropriate subset of lines from a large TP could be identified to create a GS model appropriate to many populations with a breeding program.

Importantly, these studies indicated that even the most basic GS model, one that only incorporates genetic information, can improve genetic gain when applied to an appropriate BP. This was demonstrated in the final chapter using the WCORT 15-17 population. This rudimentary model, applied at the single plant selection stage with no additional information being collected on the population, is a simple application of the GS process that is easily applied to any breeding program with the only additional cost being that of genotyping. This alone makes GS worth pursuing in oat breeding programs.

REFERENCES

Adegoke A.O., Frey K. J. (1987). Grain yield response and stability for oat lines with low-, medium-, and high-yielding ability. Euphytica 36: 121–127.

Administration U.S Federal Food and Drug (1997). Food labeling: health claims; oats a coronary heart disease. Final rule. Fed Regist 62: 3583–3601.

Alvarado G., Rodríguez F., Pacheco A., Burgueño J., Crossa J., Vargas M., Pérez-Rodríguez P., Lopez-Cruz M. (2020). META-R: A software to analyze data from multi-environment plant breeding trials. Crop J 8(5).

Aman P., Graham H. (1987). Mixed-linked beta-(1-3), (1-4)-D-glucans in the cell walls of barley and oats--chemistry and nutrition. Scand J Gastroenterol Suppl 129: 42-51.

Anderson J. A., Stack R. W., Liu S., Waldron B. L., Fjeld A. D., Coyne C., Moreno-Sevilla B., Fetch J. Mitchell, Song Q. J., Cregan P. B., Frohberg R. C. (2001). DNA markers for Fusarium head blight resistance QTLs in two wheat populations. Theor Appl Genet 102(8): 1164-1168.

Anderson J.W., Gilinsky N.H., Deakins D.A., Smith S.F., O'Neal D.S., Dillon D.W., Oeltgen D.W. (1991). Lipid responses of hypercholesterolemic men to oat-bran and wheat-bran intake. Am J Clin Nutr 54: 678-683.

Andersson A. A. M., Börjesdotter D. (2011). Effects of environment and variety on content and molecular weight of β-glucan in oats. J Cereal Sci 54(1): 122-128.

Ankamah-Yeboah T., Janss L. L., Jensen J. D., Hjortshøj R. L., Rasmussen S. K. (2020). Genomic selection using pedigree and marker-by-environment interaction for barley seed quality traits from two commercial breeding programs. Front Plant Sci 11: 539.

Asoro F. G. (2012). Breeding for β-glucan content in elite North American oat (*Avena sativa* L.) using molecular markers, Iowa State University, IA, USA

Aung T., Chong J., Leggett M. (1996). The transfer of crown rust resistance gene Pc94 from a wild diploid to cultivated hexaploid oat. Proceedings of 9th European and Mediterranean Cereal Rust & Powdery Mildews Conference. Lunteren, Netherlands.

Azodi C. B., Bolger E., McCarren A., Roantree M., de Los Campos G., Shiu S. H. (2019). Benchmarking parametric and machine learning models for genomic prediction of complex traits. G3 9(11): 3691-3702.

Baird N. A., Etter P. D., Atwood T., Currey M., Shiver A., Lewis Z., Selker E., Cresko W. A., Johnson E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One 3(10): e3376.

Barnabás B., Jäger K., Fehér A. (2008). The effect of drought and heat stress on reproductive processes in cereals. Plant Cell Envon 31: 11-38.

Bassi F. M., Bentley A. R., Charmet G., Ortiz R., Crossa J. (2016). Breeding schemes for the implementation of genomic selection in wheat (Triticum spp.). Plant Sci 242: 23-36.

Beer S.C., Siripoonwiwat W., O'Donoughue L.S., Souza E., Matthews D., Sorrells M.E. (1998). Association between molecular markers and qualitative traits in an oat germplasm pool: Can we infer linkages? J Agric Genomics 3. https://wheat.pw.usda.gov/jag/papers97/paper197/jqtl1997-01.html

Behall K. M., Scholfield D. J., Hallfrisch J. G. (2006). Barley β-glucan reduces plasma glucose and insulin responses compared with resistant starch in men. Nutr Res 26(12): 644-650.

Bekele W. A., Wight C. P., Chao S., Howarth C. J., Tinker N. A. (2018). Haplotype-based genotyping-by-sequencing in oat genome research. Plant Biotechnol J 16(8): 1452-1463.

Belitz H.D., Grosch W., Schieberle P. (2009). Cereals and cereal products. Food chemistry, 4th Edition. Berlin, Springer.

Benaragama D. (2011). Enhancing the competitive ability of oat (*Avena sativa* L.) cropping systems. University of Saskatchewan, Saskatchewan, Canada.

Bennett M. D., Smith J. B. (1976). Nuclear DNA amounts in angiosperms: progress, problems and prospects. Ann Bot 95(1): 45–90.

Bernardo R., Charcosset A. (2006). Usefulness of gene information in marker-assisted recurrent selection: A simulation appraisal. Crop Sci 46(2): 614-621.

Branson C. V., Frey K. J. (1989). Recurrent selection for groat oil content in oat. Crop Sci 29(6): 1382-1387

Breiman L., C. Adele, L. Andy, W. Matthew. (2018). RandomForest: Breiman and Cutler's random forests for classification and regression. doi:10.1023/A:1010933404324

Bridges S.R., Anderson J.W., Deakins D.A., Dillon D.W., Wood C.L. (1992). Oat bran increases serum acetate of hypercholesterolemic men. Am J Clin Nutr 56: 455-459.

Brouwer J., G. Flood. R. (1995). Aspects of oat physiology. The oat crop: production and utilization. W. R.W. London, UK, Chapman & Hall.

Brunner B. R., Freed R. D. (1994). Oat grain β-glucan content as affected by nitrogen level, location, and year. Crop Sci 34(2): cropsci1994.0011183X003400020031x.

Buckler E.S., Holland J. B., Bradbury P. J., Acharya C. B., Brown P. J., Browne C., Ersoz E., Flint-Garcia S. A., Garcia A., Glaubitz J. C., Goodman M. M., Harjes C., Guill K., Kroon D. E., Larsson S., Lepak N. K., Li H., Mitchell S. E., Pressoir G., Peiffer J. A., Rosas M.O., Rocheford T. R., Romay M. C., Romero S., Salvo S., Villeda H. S., Sofia da Silva H., Sun Q, Tian F., Upadyayula N., Ware D., Yates H., Yu J., Zhang Z., Kresovich S., McMullen M. D. (2009). The genetic architecture of maize flowering time. Science 325(5941): 714-718.

Bueckert R. A., Clarke J. M. (2013). Review: Annual crop adaptation to abiotic stress on the Canadian prairies: Six case studies. Can J Plant Sci 93(3): 375-385.

Bueckert R. A., Wagenhoffer S., Hnatowich G, Warkentin T, D. (2015). Effect of heat and precipitation on pea yield and reproductive performance in the field. Can J Plant Sci 95(4): 629-639.

Burgueño J., de los Campos G., Weigel K., Crossa J. (2012). Genomic prediction of breeding values when modeling genotype× environment interaction using pedigree and dense molecular markers. Crop Sci 52(2): 707-719.

Butt M. S., Tahir-Nadeem M., Khan M. K. I., Shabir R., Butt M. S. (2008). Oat: unique among the cereals. Eur J Nutr 47(2): 68-79.

C. Potter R., M. Castro J., C. Moffatt L. (1997). Oat oil compositions with useful cosmetic and dermatological properties. Patent:US5620692A

Cabrera-Bosquet L., Crossa J., von Zitzewitz J., Serret M., Luis Araus J. (2012). High-throughput phenotyping and genomic selection: the frontiers of crop breeding converge. J Integr Plant Biol 54(5): 312-320.

Cairns J. E., Crossa J., Zaidi P. H., Grudloyma P., Sanchez C., Araus J. L., Thaitad S., Makumbi D., Magorokosho C., Bänziger M., Menkir A., Hearne S., Atlin G. N. (2013). Identification of drought, heat, and combined drought and heat tolerant donors in maize. Crop Sci 53(4): 1335-1346.

Calus M., Veerkamp R. F. (2011). Accuracy of multi-trait genomic selection using different methods. Genet Sel Evol 43: 26.

Calus M. P. L., Meuwissen T. H. E., de Roos A. P. W., Veerkamp R. F. (2008). Accuracy of genomic selection using different methods to define haplotypes. Genetics 178(1): 553-561.

Canada Health. (2010). "Oat Products and Blood Cholesterol Lowering." from https://www.canada.ca/en/health-canada/services/food-nutrition/food-labelling/health-claims/assessments/products-blood-cholesterol-lowering-summary-assessment-health-claim-about-products-blood-cholesterol-lowering.html.

Canada National Research Council (2020). from http://www.nrc-cnrc.gc.ca/eng/services/sunrise/

Canada Statistics. (2019). from https://www.statcan.gc.ca/eng/start.

Canada Statistics. (2020). from https://www.statcan.gc.ca/eng/start.

Canales F. J., Montilla-Bascón G., Bekele W. A., Howarth C. J., Langdon T., Rispail N., Tinker N. A., Prats E. (2021). Population genomics of Mediterranean oat (*A. sativa*) reveals high genetic diversity and three loci for heading date. Theor Appl Genet 134(7): 2063-2077.

Carr D.J., Wardlaw I.F. (1965). Supply of photosynthetic assimilates to grain from flag leaf and ear of wheat. Aust J Biol Sci 18(4): 711.

Champion G. T., Froud-Williams R. J., Holland J. M. (1998). Interactions between wheat (*Triticum aestivum* L.) cultivar, row spacing and density and the effect on weed suppression and crop yield. Ann Appl Biol 133(3): 443-453.

Charmet G., Robert N., Perretant M., Gay G., Sourdille P., Groos C., Bernard S., Bernard M. (1999). Marker-assisted recurrent selection for cumulating additive and interactive QTLs in recombinant inbred lines. Theor Appl Genet 99(7-8): 1143-1148.

Chauhan C., Singh S. K. (2018). Genetic variability, heritability and genetic advance studies in oat (Avena sativa L.). Int J Chem Stud 7: 992-994.

Chaves M. M., Maroco J. P., Pereira J. S. (2003). Understanding plant responses to drought-from genes to the whole plant. Funct Plant Biol 30(3): 239-264.

Chawade A., van Ham J., Blomquist H., Bagge O., Alexandersson E., Ortiz R. (2019). High-throughput field-phenotyping tools for plant breeding and precision agriculture. Agron 9(5): 258.

Chen G., Chong J., Gray M., Prashar S., J. Douglas P. (2006). Identification of single-nucleotide polymorphisms linked to resistance gene *Pc68* to crown rust in cultivated oat. Can J Plant Pathol 28(2): 214-222.

Chen G., Chong J., Prashar S., Procunier J. D. (2007). Discovery and genotyping of high-throughput SNP markers for crown rust resistance gene *Pc94* in cultivated oat. Plant Breed 126(4): 379-384.

Chen W. J. L., Anderson J. W., Gould M.R. (1981). Effects of oat bran, oat gum and pectin on lipid metabolism of cholesterol-fed rats. Nutr Rep Int 24: 1093-1098.

Chew P, Meade K., Hayes A., Harjes C., Bao Y., Beattie A. D, Puddephat I., Gusmini G., Tanksley S. D. (2016). A study on the genetic relationships of Avena taxa and the origins of hexaploid oat. Theor Appl Genet 129(7): 1405-1415.

Chong J., Reimer E., Somers D., Aung T., Penner G. (2004). Development of sequence-characterized amplified region (SCAR) markers for resistance gene Pc94 to crown rust in oat. Can J Plant Pathol 26: 89-96.

Clark S. A., Hickey J. M., Daetwyler H. D., van der Werf J. H. J. (2012). The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. Genet Sel Evol 44(1): 4-4.

Colombani C., Legarra A., Croiseau P., Fritz S., Guillaume F., Ducrocq V., Robert-Granié C. (2011). Bayes Cpi versus GBLUP, OLS regression, sparse PLS and elestic net methods for genomic selection in french dairy cattle. J Dairy Sci (96): 1-17.

Combs E., Bernardo R. (2013). Genome wide selection to introgress semidwarf maize germplasm into u.s. corn belt inbreds. Crop Sci 53(4): 1427-1436.

Commission Canadian Grain. (2020). "Quality of Western Canadian Oats 2020." from https://grainscanada.gc.ca/en/grain-research/export-quality/cereals/oats/2020/preliminary/.

Correa K., Bangera R., Figueroa R., Lhorente J., Yáñez J. (2017). The use of genomic information increases the accuracy of breeding value predictions for sea louse (Caligus rogercresseyi) resistance in Atlantic salmon (Salmo salar). Gen Sel Evol(49): 15.

Crossa J., Beyene Y., Kassa S., Perez P., Hickey J. M., Chen C., de los Campos G., Burgueno J., Windhausen V. S., Buckler E., Jannink J. L., Lopez Cruz M. A., Babu R. (2013). Genomic prediction in maize breeding populations with genotyping-by-sequencing. G3 3(11): 1903-1926.

Crossa J., Campos G., de los Pérez P., Gianola D., Burgueño J., Araus J. L., Makumbi D., Singh R. P., Dreisigacker S., Yan J., Arief V., Banziger M., Braun H.-J. (2010). Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186(2): 713-724.

Crossa J., de los Campos G., Maccaferri M., Tuberosa R., Burgueño J., Pérez-Rodríguez P. (2016). Extending the marker× environment interaction model for genomic-enabled prediction and genome-wide association analysis in durum wheat. Crop Sci 56(5): 2193-2209.

Cuevas J., Crossa J., Montesinos-López O. A., Burgueño J., Pérez-Rodríguez P., de Los Campos G. (2017). Bayesian genomic prediction with genotype × environment interaction kernel models. G3 7(1): 41-53.

Cullis B. R., Smith A. B., Coombes N. E. (2006). On the design of early generation variety trials with correlated data. J Agric Biol Environ Stat 11(4): 381.

Daetwyler H., Pong-Wong R., Villanueva B., Woolliams J. A. (2010). The impact of genetic architecture on genome-wide evaluation methods. Genetics 185(3): 1021-1031.

Database. Food and agriculture organization of the United Nations. FAOSTAT Statistical. (2018). from http://www.fao.org/faostat/en/#data/qc.

Database. Food and agriculture organization of the United Nations. FAOSTAT Statistical. (2019). from http://www.fao.org/faostat/en/#data/qc.

Dawson J., Endelman J., Heslot N., Crossa J., Poland J., Dreisigacker S., Manès Y., Sorrells M. E., Jannink J.-L. (2013). The use of unbalanced historical data for genomic selection in an international wheat breeding program. Field Crops Res 154: 12-22.

De Buyser J., Henry Y., Lonnet P., Hertzog R., Hespel A. (1987). 'Florin': a doubled haploid wheat variety developed by the anther culture method. Plant Breed 98(1): 53-56.

de Groot A., Luyken R., Pikaar N. A. (1963). Cholesterol-lowering effect of rolled oats. Lancet 2(7302): 303-304.

De Koeyer D. L., Tinker N., Wight C., Burrows V., O'Donoughue L., Lybaert A., Molnar S., Armstrong K., Fedak G., Wesenberg D., Rossnagel B., McElroy A. (2004). A molecular linkage map with associated QTLs from a hulless × covered spring oat population. Theor Appl Genet 108: 1285-1298.

de los Campos G., Naya H., Gianola D., Crossa J., Legarra A., Manfredi E., Weigel K., Cotes J. M. (2009). Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182(1): 375-385.

de Roos A. P. W., Hayes B. J., Goddard M. E. (2009). Reliability of genomic predictions across multiple populations. Genetics 183(4): 1545-1553.

Dhungana P., Eskridge K. M., Baenziger P. S., Campbell B. T., Gill K. S., Dweikat I. (2007). Analysis of genotype-by environment interaction in wheat using a structural equation model and chromosome substitution lines. Crop Sci 47(2): 477-484.

Djanaguiraman M., Narayanan S., Erdayani E., Prasad P. V. V. (2020). Effects of high temperature stress during anthesis and grain filling periods on photosynthesis, lipids and grain yield in wheat. BMC Plant Bio 20(1): 268.

Doehlert D., McMullen M., Hammond J. (2001). Genotypic and environmental effects on grain yield and quality of oat grown in North Dakota. Crop Sci 41.

Doerge R. W. (2002). Mapping and analysis of quantitative trait loci in experimental populations. Nat Rev Genet 3(1): 43-52.

Dudley J. W., Johnson G. R. (2009). Epistatic models improve prediction of performance in corn. Crop Sci 49(3): 763-770.

Dupont F. M., Altenbach S. B. (2003). Molecular and biochemical impacts of environmental factors on wheat grain development and protein synthesis. J Cer Sci 38(2): 133-146.

Eagles H. A. (1975). The heritability and environmental specificty of adaptation parameters for grain and straw yields of oats, Iowa State University, IA, USA.

Earl D. A., vonHoldt B. M. (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv Genet Resour 4(2): 359-361.

Elshire R.J., C. Jeffrey, Glaubitz J.C., Sun Q., Poland J.A., Kawamoto K., Buckler E.S., Mitchell S.E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6: e19379.

Endelman J. B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome 4(3).

Environment Canada (2020) from https://weather.gc.ca.

Falconer D. S., Mackay T. F. C. (1996). Introduction to quantitative genetics, 4th edition. Longman Science and Technology, Harlow, UK.

Fan Y., Chen J., Shirkey G., John R., Wu S. R., Park H., Shao C. (2016). Applications of structural equation modeling (SEM) in ecological studies: an updated review. Ecol Process 5(1): 19.

Farooq M., Wahid A., Lee D.-J., Cheema S. A., Aziz T. (2010). Drought stress: comparative time course action of the foliar applied glycinebetaine, salicylic acid, nitrous oxide, brassinosteroids and spermine in improving drought resistance of rice. J Agron Crop Sci 196(5): 336-345.

Ferris R., Ellis R. H., Wheeler T. R., Hadley P. (1998). Effect of high temperature stress at anthesis on grain yield and biomass of field-grown crops of wheat. Ann Bot 82(5): 631-639.

Fetch J. M., Duguid S., Brown P., Chong J., Fetch T. G., Haber S., Menzies J., Ames N., Noll J., Aung T., Stadnyk K. (2007). Leggett oat. Can J Plant Sci 87: 509.

Flint-Garcia S. A., Thornsberry J.M., Buckler E. (2003). Structure of linkage disequilibrium in plants. Annu Rev Plant Biol 54(1): 357-374.

Forsberg R. A. (1986). World status of oats and biological constraints to increased production. proceedings of the second international oats conference. Aberystwyth, U.K.: 241-246.

Forsberg R. A., Reeves D. L. (1995). Agronomy of oats. The Oat Crop. 223-251.

Franco J., Gage J., Bradbury P., Johnson L., Miller Z., Buckler E., Romay M. (2020). A maize practical haplotype graph leverages diverse NAM assemblies. bioRxiv: 2020.2008.2031.268425.

Frey K. J. (1962). Influence of leaf-blade removal on seed weight of oats. Iowa State J Sci 37: 17-22.

Gianola D., Fernando R. L., Stella A. (2006). Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics 173(3): 1761-1776.

Givaudan (2019). The future of plant proteins: potential game changers. White paper.

Gnanesh B. N., McCartney C. A., Eckstein P. E., Fetch J. W. M., Menzies J. G., Beattie A. D. (2015). Genetic analysis and molecular mapping of a seedling crown rust resistance gene in oat. Theor Appl Genet 128(2): 247-258.

Goddard M. E., Hayes B. J. (2007). Genomic selection. J Anim Breed Genet 124(6): 323-330.

González-Camacho J., Ornella L., Pérez-Rodríguez P., Gianola D., Dreisigacker S., Crossa J. (2018). Applications of machine learning methods to genomic selection in breeding wheat for rust resistance. Plant Genome 11: 1-15.

Gore M., Bradbury P., Hogers R., Kirst M., Verstege E., van Oeveren J., Peleman J., Buckler E., van Eijk M. (2007). Evaluation of target preparation methods for single-feature polymorphism detection in large complex plant genomes. Crop Sci 47(S2): S-135-S-148.

Gore M., Wright M., Ersoz E., Bouffard P., Szekeres E., Jarvie T., Hurwitz B., Narechania A., Harkins T., Grills G., Ware D., Buckler E. (2009). Large-scale discovery of gene-enriched SNPs. Plant Genome 2(2): 121-133.

Guha S., Maheswari S. C. (1964). In vitro production of embryos from anthers of Datura. Nature 204: 497.

Guilioni L., Wery J., Tardieu F. (1997). Heat stress-induced abortion of buds and flowers in pea: is sensitivity linked to organ age or to relations between reproductive organs? Ann Bot 80(2): 159-168.

Guillen-Portal F., Stougaard R., Xue Q., Eskridge K. M. (2006). Compensatory mechanisms associated with the effect of spring wheat seed size on wild oat competition. Crop Sci 46 (2): 935-945.

Güler M. (2003). Nitrogen and irrigation effects on β-Glucan content of wheat grain. Acta Agric Scand B 53(3): 156-160.

Guo G., Zhao F., Wang Y., Zhang Y., Du L., Su G. (2014). Comparison of single-trait and multiple-trait genomic prediction models. BMC Genet 15(1): 30.

Gutierrez A., Symonds J., King N., Steiner K., Bean T., Houston R. (2020). Potential of genomic selection for improvement of resistance to ostreid herpesvirus in Pacific oyster (*Crassostrea gigas*). Anim Genet 51(2): 249-257.

Habier D., Fernando R. L., Dekkers J. C. M. (2007). The impact of genetic relationship information on genome-assisted breeding values. Genetics 177(4): 2389-2397.

Habier D., Tetens J., Seefried F. R., Lichtner P., Thaller G. (2010). The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet Sel Evol 42: 5.

Haikka H., Manninen O., Hautsalo J., Pietilä L., Jalli M., Veteläinen M. (2020). Genome-wide association study and genomic prediction for *Fusarium graminearum* resistance traits in Nordic oat (*Avena sativa* L.). Agron 10: 174.

Haile T. A., Heidecker T., Wright D., Neupane S., Ramsay L., Vandenberg A., Bett K. E. (2020). Genomic selection for lentil breeding: empirical evidence. Plant Genome 13(1): e20002.

Hamblin M. T., Buckler E. S., Jannink J.-L. (2011). Population genetics of genomics-based crop improvement methods. Trends Genet 27(3): 98-106.

Hamblin M. T., Close T. J., Bhat P. R., Chao S., Kling J. G., Abraham K. J., Blake T., Brooks W. S., Cooper B., Griffey C. A., Hayes P. M., Hole D. J., Horsley R. D., Obert D. E., Smith K. P., Ullrich S. E., Muehlbauer G. J., Jannink J.-L. (2010). Population structure and linkage disequilibrium in U.S. barley germplasm: implications for association mapping. Crop Sci 50(2): 556-566.

Hatfield J. L., Prueger J. (2015). Temperature extremes: Effect on plant growth and development. Weather Clim Extremes 10: 4-10.

Hayes B. J., Bowman P. J., Chamberlain A. J., Goddard M. E. (2009). Invited review: genomic selection in dairy cattle: progress and challenges. J Dairy Sci 92(2): 433-443.

Heffner E. L., Jannink J.-L., Iwata H., Souza E., Sorrells M. E. (2011b). Genomic selection accuracy for grain quality traits in biparental wheat populations. Crop Sci 51(6): 2597-2606.

Heffner E. L., Jannink J.-L., Sorrells M.E. (2011a). Genomic selection accuracy using multifamily prediction models in a wheat breeding program. Plant Genome 4(1): 65-75.

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

Henry R. J. (1987). Pentosan and (1 → 3),(1 → 4)-β-Glucan concentrations in endosperm and wholegrain of wheat, barley, oats and rye. J. Cereal Sci. 6(3): 253-258.

Heslot N., Akdemir D., Sorrells M. E., Jannink J. L. (2014). Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. Theor Appl Genet 127(2): 463-480.

Hess M., Druet T., Hess A., Garrick D. J. (2017). Fixed-length haplotypes can improve genomic prediction accuracy in an admixed dairy cattle population. Genet Sel Evol 49(1): 1-14.

Hill W. G., Weir B. S. (1988). Variances and covariances of squared linkage disequilibria in finite populations. Theor Popul Biol 33(1): 54-78.

Holland J. B. (2007). Genetic architecture of complex traits in plants. Curr Opin Plant Biol 10(2): 156-161.

Holland J. B., Munkvold G. P. (2001). Genetic relationships of crown rust resistance, grain yield, test weight, and seed weight in oat. Crop Sci 41: 1041.

Holthaus J., Holland James, White Pamela, Frey K. (1996). Inheritance of β-Glucan content of oat grain. Crop Sci 36.

Horikoshi M., Tang Y., Dickey A., Genié M., Thompson R., Seltzer L. (2020). ggfortify: Data visualization tools for statistical analysis results. R package version 0.4.11.2020.

Hospital Frédéric, Moreau L., Lacoudre F., Charcosset A., Gallais A. (1997). More on the efficiency of marker-assisted selection. Theor Appl Genet 95(8): 1181-1189.

Hoyle R. H. (1995). Structural equation modeling: Concepts, issues, and applications. Thousand Oaks, CA, US, Sage Publications, Inc.

Huang Y.-F., Poland J.A., Wight C.P., Jackson E.W., Tinker N.A. (2014). Using Genotyping-by-sequencing (GBS) for genomic discovery in cultivated oat. PLoS One 9(7): e102448.

Huel D. G., Hucl P. (1996). Genotypic variation for competitive ability in spring wheat. Plant Breed 115(5): 325-329.

Humphreys D. G., Mather D. E. (1996). Heritability of β-glucan, groat percentage, and crown rust resistance in two oat crosses. Euphytica 91(3): 359-364.

Hunt C. H., van Eeuwijk F. A., Mace E. S., Hayes B. J., Jordan R. (2018). Development of genomic prediction in sorghum. Crop Sci 58(2): 690-700.

Hurt H.D., Mathews R., Ink Shews, R. (1988). Biomedical considerations of oat dietary fiber and beta-glucans. International Oat Conference, Lund, Sweden.

Isidro J., Jannink J.-L., Akdemir D., Poland J., Heslot N., Sorrells M. E. (2015). Training set optimization under population structure in genomic selection. Theor Appl Genet 128(1): 145-158.

Jannink J.-L., Lorenz A. J., Iwata H. (2010). Genomic selection in plant breeding: from theory to practice. Brief Funct Genomics 9(2): 166-177.

Jarquín D., Crossa J., Lacaze X., Du Cheyron P., Daucourt J., Lorgeou J., Piraux F., Guerreiro L., Pérez P., Calus M., Burgueño J., de los Campos G. (2014). A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theor Appl Genet 127(3): 595-607.

Jarquín D., da Silva C., Gaynor R., Poland J., Fritz A., Howard R., Battenfield S., Crossa J. (2017). Increasing genomic-enabled prediction accuracy by modeling genotype x environment interactions in Kansas wheat. Plant Genome10(2).

Jellen E. N., Gill B. S., Cox T. S. (1994). Genomic in situ hybridization differentiates between A/D-and C-genome chromatin and detects intergenomic translocations in polyploid oat species (genus Avena). Genome 37(4): 613-618.

Jennings C.D., Boleyn K., Bridges S.R., Wood P.J., Anderson J.W. (1988). A comparison of the lipid-lowering and intestinal morphological effects of cholestyramine, chitosan, and oat gum in rats. Exp Biol Med 189(1): 13-20.

Jensen S., Charles J., Muleta K., Bradbury P., Casstevens T., Deshpande S., Gore M., Gupta R., Ilut D., Johnson L., Lozano R., Miller Z., Ramu P., Rathore A., Romay M., Upadhyaya H., Varshney R., Morris G., Pressoir G., Buckler E., Ramstein G. (2020). A sorghum practical haplotype graph facilitates genome-wide imputation and cost-effective genomic prediction. Plant Genome 13(1): e20009.

Jia Y., Jannink J. L. (2012). Multiple-trait genomic selection methods increase genetic value prediction accuracy. Genetics 192(4): 1513-1522.

Jiang J., Zhang Q., Ma L, Li J, Wang Z, Liu J. F. (2015). Joint prediction of multiple quantitative traits using a Bayesian multivariate antedependence model. Heredity 115(1): 29-36.

Juliana P., Singh R., Braun H.-J., Huerta-Espino J., Crespo-Herrera L., Payne T., Poland J., Shrestha S., Kumar U., Joshi A. K., Imtiaz M., Rahman M. M., Toledo F. H. (2020). Retrospective quantitative genetic analysis and genomic prediction of global wheat yields. Front Plant Sci 11(1328).

Karatzoglou A., Smola A., Hornik K. (2019). kernlab, the kernel based machine learning Lab. (R package version 0.9-29).

Karim M., Siddika A., Tonu N., Hossain D., Meah M., Kawanabe T., Fujimoto R., Okazaki K. (2014). Production of high yield short duration Brassica napus by interspecific hybridization between B. oleracea and B. rapa. Breed Sci 63(5): 495-502.

Karr E. J., Linck A. J., Swanson C. A. (1959). The effect of short periods of high temperature during day and night periods on pea yields. Am J Bo 46(2): 91-93.

Kaufmann M. L. (1961). Yield-maturity relationships in oats. Can J Plant Sci 41(4): 763-771.

Kebede A. Z., Admassu-Yimer B., Bekele W. A., Gordon T., Bonman J., Babiker E., Jin Y., Gale S., Wight C. P., Tinker N. (2020). Mapping of the stem rust resistance gene *Pg13* in cultivated oat. Theor Appl Genet 133(1): 259-270.

Kebede A. Z., Bekele W.A., Fetch J. W., Beattie A. D., Chao S., Tinker N. A., Fetch T. G., McCartney C. A. (2020). Localization of the stem rust resistance gene *pg2* to linkage group Mrg20 in cultivated oat (*Avena sativa*). Phytopathology 110(10): 1721-1726.

Kibite S., Menzies J. (2001). AC Morgan oat. Can J Plant Sci 81: 85-87.

Kirby E. J. M., Appleyard M. (1986). Cereal development guide. Stoneleigh, Great Britain.

Klopfenstein C.F., Hoseney R.C. (1987). Cholesterol lowering effect of beta-glucan enriched bread. Nutr Rep Int 36: 1091-1098.

Klos K. E., Yimer B. A., Babiker E. M., Beattie A. D., Bonman J. M., Carson M. L., Chong J., Harrison S. A., Ibrahim A. M. H., Kolb F. L., McCartney C. A., McMullen M., Fetch J. M., Mohammadi M., Murphy J. P., Tinker N. A. (2017). Genome-wide association mapping of crown rust resistance in oat elite germplasm. Plant Genome 10(2).

Klose C., Arendt E. K. (2012). Proteins in oats; their synthesis and changes during germination: a review. Crit Rev Food Sci Nutr 52(7): 629-639.

Kozak M., Bocianowski J., RybiŃSki W. (2008). Selection of promising genotypes based on path and cluster analyses. J Agric Sci146: 85-92.

Kutcher H. R., Warland J., Brandt S. (2010). Temperature and precipitation effects on canola yields in Saskatchewan, Canada. Agric For Meteorol 150: 161-165.

Lado B., Matus I., Rodríguez A., von Zitzewitz J. (2013). Increased genomic prediction accuracy in wheat breeding through spatial adjustment of field trial data. G3 3(12): 2105–2114.

Lamb Eric G., Shirtliffe Steven J., May William E. (2011). Structural equation modeling in the plant sciences: An example using yield components in oat. Can J Plant Sci 91(4): 603-619.

Li C. D., Rossnagel B. G., Scoles G. J. (2000). The development of oat microsatellite markers and their use in identifying relationships among Avena species and oat cultivars. Theor Appl Genet 101(8): 1259-1268.

Liatisa S., Tsapogasa P., Chalab E., Dimosthenopoulosa C., Kyriakopoulosa K., Kapantaisb E., Katsilambrosa N. (2009). The consumption of bread enriched with betaglucan reduces LDL-cholesterol and improves insulin resistance in patients with type 2 diabetes. Diabetes Metab 35(2): 115–120.

Lillehammer M., Meuwissen T. Sonesson, A. K. (2011). A comparison of dairy cattle breeding designs that use genomic selection. J Dairy Sci 94(1): 493-500.

Lim H.S., White P. J., Frey K. J. (1992). Genotypic effects on beta-glucan content of oat lines grown in two consecutive years. Cereal Chem. 69(3): 262-265.

Lin Y., Gnanesh B. N., Chong J., Chen G., Beattie A. D., Mitchell Fetch J. W., Kutcher H. R., Eckstein P. E., Menzies J. G., Jackson E. W., McCartney C. A. (2014). A major quantitative trait locus conferring adult plant partial resistance to crown rust in oat. BMC Plant Bio 14(1): 250.

Lipiec J., Doussan C., Nosalewicz A., Kondracka K. (2013). Effect of drought and heat stresses on plant growth and yield: A review. Int Agrophys 27(4): 463-477.

Liu S., Vallejo R. L, Palti Y., Gao G., Marancik D. P, Hernandez A. G, Wiens G. D. (2015). Identification of single nucleotide polymorphism markers associated with bacterial cold water disease resistance and spleen size in rainbow trout. Front genet 6: 298.

Lopes M., Bovenhuis H., Hidalgo A., van Arendonk J., Knol E., Bastiaansen J. (2017). Genomic selection for crossbred performance accounting for breed-specific effects. Genet Sel Evol 49(1): 51.

Lopez-Cruz M., Crossa J., Bonnett D., Dreisigacker S., Poland J., Jannink J. L., Singh R. P., Autrique E., de los Campos G. (2015). Increased prediction accuracy in wheat breeding trials using a marker × environment interaction genomic selection model. G3 5(4): 569-582.

Lorenz A. J., Chao, S., Asoro, F. G., Heffner, E. L., Hayashi, T., Iwata, H., Smith, K. P., Sorrells, M. E. and Jannink, J. (2011). Genomic selection in plant breeding: knowledge and prospects. Adv Agron 110: 77-124.

Lorenzana R. E., Bernardo R. (2009). Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theor Appl Genet 120(1): 151-161.

Lozada D. N., Mason R. E., Sarinelli J. M., Brown-Guedira G. (2019). Accuracy of genomic selection for grain yield and agronomic traits in soft red winter wheat. BMC Genet 20(1): 82.

Lu F., Lipka A.E., Elshire R.J., Glaubitz J., Cherney J., Casler M., Buckler E.S., Costich D. (2012). Characterization of the genetic diversity of switchgrass using genotyping by sequencing. Plant and Animal Genome XX, San Diego, CA, USA.

Luan T., Woolliams J. A., Lien S., Kent M., Svendsen M., Meuwissen T. H. E. (2009). The accuracy of genomic selection in Norwegian red cattle assessed by cross-validation. Genetics 183(3): 1119-1126.

Ma J. Q., Huang L., Ma C. L., Jin J. Q., Li C. F., Wang R. K., Zheng H. K., Yao M. Z., Chen L. (2015). Large-scale SNP discovery and genotyping for constructing a high-density genetic map of tea plant using specific-locus amplified fragment sequencing (SLAF-seq). PLoS One 10(6): e0128798.

Mądry W., Studnicki M., Rozbicki J., Golba J., Gozdowski D., Pecio A., Oleksy A. (2015). Ontogenetic-based sequential path analysis of grain yield and its related traits in several winter wheat cultivars. Acta Agric Scand B Soil Plant Sci 65(7): 605-618.

Marcińska I., Nowakowska A., Skrzypek E., Czyczyło-Mysza I. (2013). Production of double haploids in oat (Avena sativa L.) by pollination with maize (Zea mays L.). Cent Eur J Biol 8(3): 306-313.

Marlett J.A. (1993). Comparisons of dietary fiber and selected nutrient compositions of oat and other grain fractions. St. Paul, American Association of Cereal Chemists.

May W. E., Mohr R. M., Lafond G. P., Johnston A. M., Stevenson F. C. (2004). Effect of nitrogen, seeding date and cultivar on oat quality and yield in the eastern Canadian prairies. Can J Plant Sci 84(4): 1025-1036.

Mazurkievicz G., Ubert I., Krause F., Nava I. (2019). Phenotypic variation and heritability of heading date in hexaploid oat. Crop Breed Appl Biotechnol 19: 436-443.

McCartney C. A., Stonehouse R. G., Rossnagel B. G., Eckstein P. E., Scoles G. J., Zatorski T., Beattie A. D., Chong J. (2011). Mapping of the oat crown rust resistance gene *Pc91*. Theor Appl Genet 122(2): 317-325.

Mcferson J. K. (1987). Three selection strategies that utilize recurrent selection to increase protein yield in oats.

Mcmullan P. M., McVetty P. B. E., Urquhart A. A. (1988). Dry matter and nitrogen accumulation and redistribution and their relationship to grain yield and grain protein in oats. Can J Plant Sci 68(4): 983-993.

Meng Q., Wang H., Yan P., Pan J., Lu D., Cui Z., Zhang F., Chen X. (2017). Designing a new cropping system for high productivity and sustainable water usage under climate change. Sci Rep 7: 41587-41587.

Merrick L., Carter A. (2021). Comparison of genomic selection models for exploring predictive ability of complex traits in breeding programs. bioRxiv: 2021.2004.2015.440015.

Meuwissen T. H., Hayes B. J., Goddard M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. Genetics 157(4): 1819-1829.

Meuwissen T. H. E. (2009). Accuracy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping. Genet Sel Evol 41(1): 35.

Meydani M. (2009). Potential health benefits of avenanthramides of oats. Nutr Rev 67(12): 731-735.

Milach S., Rines H., Phillips R. (1997). Molecular genetic mapping of dwarfing genes in oat. Theor Appl Genet 95: 783-790.

Molenaar H., Boehm R., Piepho H.-P. (2018). Phenotypic selection in ornamental breeding: it's better to have the BLUPs than to have the BLUEs. Front Plant Sci 9: 1511-1511.

Montesinos-López O. A., Montesinos-López A., Pérez-Rodríguez P., Barrón-López J., Martini J., Fajardo-Flores S., Gaytan-Lugo L., Santana-Mancilla P., Crossa J. (2021). A review of deep learning applications for genomic selection. BMC Genomics 22(1): 19.

Moser G., Khatkar M., Hayes B., Raadsma H. (2010). Accuracy of direct genomic values in Holstein bulls and cows using subsets of SNP markers. Genet Sel Evol 42(1): 37.

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

Mota L. F. M., Fernandes Jr G. A., Herrera A. C., Scalez D. C. B., Espigolan R., Magalhães A. F. B., Carvalheiro R., Baldi F., Albuquerque L. G. (2020). Genomic reaction norm models exploiting genotype × environment interaction on sexual precocity indicator traits in Nellore cattle. Anim Genet 51(2): 210-223.

Muir W. M. (2007). Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. J Anim Breed Genet 124(6): 342-355.

Munkvold J. D, Tanaka J., Benscher D., Sorrells M. E. (2009). Mapping quantitative trait loci for preharvest sprouting resistance in white wheat. Theor Appl Genet 119(7): 1223-1235.

Nanjappa J., McCartney C., Lin Y., Beattie A. D., Chong J., Fetch J. M., Kutcher R., Eckstein P.E., Jackson E., Menzies J., Parkin I. (2014). Poster: A major qtl for adult plant crown rust resistance in oat. Plant and Animal Genome XXII Conference. San Diego, CA, USA.

Newell M. A., Cook D., Tinker N. A., Jannink J. L. (2011). Population structure and linkage disequilibrium in oat (*Avena sativa* L.): implications for genome-wide association studies. Theor Appl Genet 122(3): 623-632.

O'Donoughue L. S., Sorrells M. E., Tanksley S. D., Autrique E., Deynze A., Van,Kianian S. F., Phillips R. L., Wu B., Rines H. W., Rayapati P. J., Lee M., Penner G. A., Fedak G., Molnar S. J., Hoffman D., Salas C. A. (1995). A molecular linkage map of cultivated oat. Genome 38(2): 368-380.

Oliver R., Lazo G., Lutz J., Rubenfield M., Tinker N., Anderson J., Wisniewski Morehead N., Adhikary D., Jellen E., Maughan P., Brown Guedira G., Chao S., Beattie A. D, Carson M., Rines H., Obert D., Bonman J. M., Jackson E. (2011). Model SNP development for complex genomes based on hexaploid oat using high-throughput 454 sequencing technology. BMC Genomics 12(1): 77.

Oliver R.E., Tinker, N.A., Lazo, G.R., Chao, S., Jellen, E.N., Carson, M.L., Rines, H.W., Obert, D.E., Lutz, J.D., Shackelford, I., Korol, A.B., Wight, C.P., Gardner, K.M., Hattori, J., Beattie, A.D., Bjornstad, A., Bonman, J.M., Jannink, J.L., Sorrells, M.E., Brown-Guedira, G.L., Fetch, J.W.M., Harrison, S.A., Howarth, C.J., Ibrahim, A., Kolb, F.L., McMullen, M.S., Murphy, J.P., Ohm, H.W., Rossnagel, B.G., Yan, W., Miclaus, K.J., Hiller, J., Maughan, P.J., Redman Hulse, R.R., Anderson, J.M., Islamovic, E., Jackson, E.W. (2013). SNP discovery and chromosome anchoring provide the first physically-anchored hexaploid oat map and reveal synteny with model species. PLoS One 8(3): 58068.

Othman R. A., Moghadasian M. H., Jones P. J. (2011). Cholesterol-lowering effects of oat beta-glucan. Nutr Rev 69(6): 299-309.

Özkaynak E. (2013). Effects of air temperature and hours of sunlight on the length of the vegetation period and the yield of some field crops. Ekoloji 22: 58-63.

Pal N., Sandhu J. S., Domier L. L., Kolb F. L. (2002). Development and characterization of microsatellite and RFLP-derived PCR markers in oat. Crop Sci 42(3): 912-918.

Palaiokostas C., Vesely T., Kocour M., Prchal M., Pokorova D., Piackova V., Pojezdal L., Houston R. (2019). Optimizing genomic prediction of host resistance to koi herpesvirus disease in carp. Front Genet 10: 543.

Pawlisch P. E, Shands H. L. (1962). Breeding behavior for bushel weight and agronomic characters in early generations of two oat crosses. Crop Sci 2(3): 231-237.

PepsiCo (2020). Avena sativa – OT3098 v1.
https://wheat.pw.usda.gov/GG3/graingenes_downloads/oat-ot3098-pepsico

Peräaho M., Collin P., Kaukinen K., Kekkonen L., Miettinen S., Mäki M. (2004). Oats can diversify a gluten-free diet in celiac disease and dermatitis herpetiformis. J Am Diet Assoc 104(7): 1148-1150.

Peräaho M., Kaukinen K., Mustalahti K., Vuolteenaho N., Mäki M., Laippala P., Collin P. (2004). Effect of an oats-containing gluten-free diet on symptoms and quality of life in coeliac disease. A randomized study. Scand J Gastroenterol 39(1): 27-31.

Pérez-Rodríguez P., Crossa J., Bondalapati K., De Meyer G., Pita F., de los Campos G. (2015). A pedigree-based reaction norm model for prediction of cotton yield in multienvironment trials. Crop Sci 55(3): 1143-1151.

Pérez-Rodríguez P., Flores-Galarza S., Vaquera Huerta H., Valle-Paniagua D., Montesinos-López O. A., Crossa J. (2020). Genome-based prediction of Bayesian linear and non-linear regression models for ordinal data. Plant Genome 13(2):

Piepho H. P., Möhring J., Melchinger A. E., Büchse A. (2008). BLUP for phenotypic selection in plant breeding and variety testing. Euphytica 161(1): 209-228.

Pixley K. (1990). Inheritance of test weight and associated traits of oat (*Avena sativa* L.) Iowa State University, IA, USA.

Poland J. A., Brown P. J., Sorrells M. E., Jannink J.-L. (2012b). Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS One 7(2): e32253.

Poland J.A., Endelman J., Dawson J., Rutkoski J., Wu S., Manes Y., Dreisigacker S., Crossa J., Sánchez-Villeda H., Sorrells M. E., Jannink J.-L. (2012b). Genomic selection in wheat breeding using genotyping-by-sequencing. Plant Gen 5(3): 103-113.

Poppitt S. D. (2007). Soluble fibre oat and barley beta-glucan enriched products: can we predict cholesterol-lowering effects? Br J Nutr 97(6): 1049-1050.

Portyanko V. A., Hoffman D. L., Lee M., Holland J. B. (2001). A linkage map of hexaploid oat based on grass anchor DNA clones and its relationship to other oat maps. Genome 44(2): 249-265.

Queenan K. M., Stewart M. L., Smith K. N., Thomas W., Fulcher R. G., Slavin J. L. (2007). Concentrated oat beta-glucan, a fermentable fiber, lowers serum cholesterol in hypercholesterolemic adults in a randomized controlled trial. Nutr J 6: 6-6.

R. Sampson D. (1971). Additive and nonadditive genetic variances and genotypic correlations for yield and other traits in oats. Can J Genet Cytol 13(4): 864-872.

Riedelsheimer C., Endelman J. B., Stange M., Sorrells M. E., Jannink J.-L., Melchinger A. E. (2013). Genomic predictability of interconnected biparental maize populations. Genetics 194(2): 493-503.

Rines H. W., Dahleen L. S. (1990). Haploid oat plants produced by application of maize pollen to emasculated oat florets. Crop Sci 30(5): 1073-1078.

Robinson G. (1991). That BLUP is a good thing: the estimation of random effects. Stat Sci 6(1): 15-32.

Robledo D., Matika Oswald, Hamilton Alastair, Houston Ross D. (2018). Genome-wide association and genomic selection for resistance to amoebic gill disease in Atlantic salmon. G3 8(4): 1195-1203.

Rutkoski J., Poland J., Jannink J.-L., Sorrells M. (2013). Imputation of unordered markers and the impact on genomic selection accuracy. G3 3(3): 427-439.

Rutkoski J., Singh R.P., Huerta-Espino J., Bhavani S., Poland J., Jannink J.L., Sorrells M.E. (2015). Genetic gain from phenotypic and genomic selection for quantitative resistance to stem rust of wheat. Plant Genome 8(2): plantgenome2014.2010.0074.

Kibite S., Edney M. (1998). The inheritance of β-glucan concentration in three oat (Avena sativa L.) crosses. Can J Plant Sci 78(2): 245-250.

Saastamoinen M. (1995). Effects of environmental factors on the β-Glucan content of two oat varieties. Acta Agric Scand B 45(3): 181-187.

Saatchi M., McClure M.C., McKay S.D., Rolf M. M., Kim J., Decker J. E., Taxis T. M., Chapple R. H., Ramey H. R., Northcutt S. L., Bauck S., Woodward B., Dekkers J. C. M., Fernando R. L., Schnabel R.D., Garrick D. J., Taylor J. F. (2011). Accuracies of genomic breeding values in American Angus beef cattle using K-means clustering for cross-validation. Genet Sel Evol 43(1): 40-40.

Saini H. S., Aspinall D. (1982). Abnormal sporogenesis in wheat (*Triticum aestivum* L.) induced by short periods of high temperature. Ann. Bot. 49(6): 835-846.

Schaeffer L. R. (2006). Strategy for applying genome-wide selection in dairy cattle. J Anim Breed Genet 123(4): 218-223.

Schmidt P., Hartung J., Rath J., Piepho H.-P. (2019). Estimating broad-sense heritability with unbalanced data from agricultural cultivar trials. Crop Science 59(2): 525-536.

Scholkopf B., Tsuda K., Vert J.-P. (2004). Kernel methods in computational biology. Cambridge, MA, USA, MIT Press.

Schulthess A. W., Wang Y., Miedaner T., Wilde P., Reif J., Zhao Y. (2016). Multiple-trait- and selection indices-genomic predictions for grain yield and protein content in rye for feeding purposes. Theor Appl Genet. 129 (2):273-287

Shaykewich C. F. (1995). An appraisal of cereal crop phenology modelling. Can J Plant Sci(75): 329–341.

Singh R., De S., Belkheir A. (2013). Avena sativa (Oat), a potential neutraceutical and therapeutic agent: an overview. Crit Rev Food Sci Nutr 53(2): 126-144.

Siripoonwiwat W., O'Donoughue L.S., Wesenberg D., Hoffman D., Barbosa-Neto J.F., Sorrells M. E. (1996). Chromosomal regions associated with quantitative traits in oat. J. Agric. Genomics 2.

Slafer G. A., Rawson H. M. (1995). Base and optimum temperatures vary with genotype and stage of development in wheat. Plant Cell Environ 18(6): 671-679.

Smith J. S. C., Hussain T., Jones E. S., Graham G., Podlich D., Wall S., Williams M. (2008). Use of doubled haploids in maize breeding: implications for intellectual property protection and genetic diversity in hybrid crops. Mol Breed 22(1): 51-59.

Solberg T. R., Sonesson A. K., Woolliams J. A., Meuwissen T. H. E. (2008). Genomic selection using different marker types and densities. J Anim Sci 86(10): 2447-2454.

Spindel J., Begum H., Akdemir D., Virk P., Collard B., Redoña E., Atlin G. N., Jannink J.-L., McCouch S. R. (2015). Genomic selection and association mapping in rice (oryza sativa): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. PLoS Genet 11(2): e1004982.

Stone P. J., Nicolas M. E. (1995). A survey of the effects of high temperature during grain filling on yield and quality of 75 wheat cultivars. Aust J Agric Res 46(3): 475-492.

Storlie E., Charmet G. (2013). Genomic selection accuracy using historical data generated in a wheat breeding program. Plant Genome 6(1): plantgenome2013.2001.0001.

Strychar R. (2021) Canada oat and oat exports. Oatinformation.com (March 7, 2021).

Su G., Guldbrandtsen B., Gregersen V. R., Lund M. S. (2010). Preliminary investigation on reliability of genomic estimated breeding values in the Danish Holstein population. Journal of dairy science 93(3): 1175-1183.

Sukumaran S., Jarquin D., Crossa J., Reynolds M. (2018). Genomic-enabled prediction accuracies increased by modeling genotype × environment interaction in durum wheat. Plant Genome 11(2): 170112.

Sur R., Nigam A., Grote D., Liebel F., Southall M. D. (2008). Avenanthramides, polyphenols from oats, exhibit anti-inflammatory and anti-itch activity. Arch Dermatol Res 300(10): 569-574.

Tang Y., Liu X., Wang J., Li M., Wang Q., Tian F., Su Z., Pan Y., Liu D., Lipka A.E., Buckler E., Zhang Z. (2016). GAPIT Version 2: An enhanced integrated tool for genomic association and prediction. Plant Genome 9(2): plantgenome2015.2011.0120.

Tanhuanpää P., Kalendar R., Laurila J., Schulman A. H., Manninen O., Kiviharju E. (2006). Generation of SNP markers for short straw in oat (*Avena sativa* L.). Genome 49(3): 282-287.

Tashiro T., Wardlaw I.F. (1990). The Effect of High Temperature at Different Stages of Ripening on Grain Set, Grain Weight and Grain Dimensions in the Semi-dwarf Wheat 'Banks'. Ann Bot 65(1): 51-61.

Technow F., Messina C. D., Totir L. R., Cooper M. (2015). Integrating crop growth models with whole genome prediction through approximate Bayesian computation. PLoS One 10(6): e0130855.

ter Braak C. J. F., Boer M. P., Bink M. C. A. M. (2005). Extending Xu's Bayesian model for estimating polygenic effects using markers of the entire genome. Genetics 170(3): 1435-1438.

Tessema B. B., Liu H., Sørensen A. C., Andersen J. R., Jensen J. (2020). Strategies using genomic selection to increase genetic gain in breeding programs for wheat. Front Genet 11: 578123.

Tester M., Langridge P. (2010). Breeding technologies to increase crop production in a changing world. Science 327(5967): 818-822.

Thorne G. N. (1974). Physiology of grain yield of wheat and barley. Rothamsted Experimental Station Report for 1973. 2: 5-25.

Tibelius A.C., Klinck H.R.(1986). Development and yield of oat plants grown from primary and secondary seeds. Can J Plant Sci (66): 299-306.

Tinker N., Kilian A., Wight C., Heller-Uszynska K., Wenzl P., Rines H., Bjornstad A., Howarth C., Jannink J-L., Anderson J., Rossnagel B., Stuthman D., Sorrells M., Jackson E., Tuvesson S., Kolb F., Olsson O., Federizzi L., Carson M., Ohm H., Molnar S., Scoles G., Eckstein P., Bonman J. M., Ceplitis A., Langdon T. (2009). New DArT markers for oat provide enhanced map coverage and global germplasm characterization. BMC Genomics 10(1): 39.

Tinker N. A., Chao S., Lazo G. R., Oliver R. E., Huang Y.-F., Poland J. A., Jellen E. N., Maughan P. J., Kilian A., Jackson E. W. (2014). A SNP genotyping array for hexaploid oat. Plant Genome 7(3): plantgenome2014.2003.0010.

Tsai H.-Y., Hamilton A., Tinch A., Guy D., Bron J., Taggart J., Gharbi K., Stear M., Matika O., Pong-Wong R. (2016). Genomic prediction of host resistance to sea lice in farmed Atlantic salmon populations. Genet Sel Evol 48(1): 1-11.

Tshiro T., Wardlaw I. F.(1990). The effect of high temperature at different stages of ripening on grain set, grain weight and grain dimensions in the semi-dwarf wheat 'banks'. Ann. Bot. 65(1): 51-61.

Vallejo R., Silva R., Evenhuis J., Gao G., Liu S., Parsons J., Martin K., Wiens G., Lourenco D., Leeds T. (2018). Accurate genomic predictions for BCWD resistance in rainbow trout are achieved using low-density SNP panels: Evidence that long-range LD is a major contributing factor. J Anim Breed Genet 135(4): 263-274.

VanRaden P. M., Van Tassell C. P., Wiggans G. R., Sonstegard T. S., Schnabel R. D., Taylor J. F., Schenkel F. S. (2009). Reliability of genomic predictions for North American Holstein bulls. Journal of dairy science 92(1): 16-24.

Verbyla K.L., Hayes B.J., Bowman P.J., Goddard M.E. (2009). Accuracy of genomic selection using stochastic search variable selection in Australian Holstein Friesian dairy cattle. Genet Res 91(05): 307-311.

Visscher P. M., Hill W. G., Wray N. R. (2008). Heritability in the genomics era — concepts and misconceptions. Nat Rev Genet 9(4): 255-266.

Walstead R. N., Whaley A., Reid R., Vallejo V., Brouwer C., Schlueter J. (2018). Assembly and annotation of the hexaploid oat genome. Plant & Animal Genome Conference XXVI, San Diego, USA.

Wang X., Yang Z., Xu C. (2015). A comparison of genomic selection methods for breeding value prediction. Sci Bull 60(10): 925-935.

Ward J. H., Hook M. E. (1963). Application of an hierarchical grouping procedure to a problem of grouping profiles. Educ Psychol Meas 23(1): 69-81.

Wardlaw I., Blumenthal C., Larroque O., Wrigley C. (2002). Contrasting effects of chronic heat stress and heat shock on kernel weight and flour quality in wheat. Funct. Plant Biol. 29: 25-34.

Welch R.W. (1995). Oats in human nutrition and health. The oat crop: Production and utilization. W. R.W. London, Chapman and Hall: 433–471.

Wesenberg D. M., Shands H. L. (1973). Heritability of oat caryopsis percentage and other grain quality components. Crop Sci 13(4): cropsci1973.0011183X001300040026x.

Whitehead A., Beck E. J., Tosh S., Wolever T. M. (2014). Cholesterol-lowering effects of oat beta-glucan: a meta-analysis of randomized controlled trials. Am J Clin Nutr 100(6): 1413-1421.

Whittaker J. C., Thompson R., Denham M. C. (2000). Marker-assisted selection using ridge regression. Genet Res 75(2): 249-252.

Wickham H., François R., Henry L., Müller K. (2018). dplyr: A grammar of data manipulation (v.1.0.5).

Wicks G. A., Ramsel R. E., Nordquist P. T., Schmidt J. W., Challaiah. (1986). Impact of wheat cultivars on establishment and suppression of summer annual weeds. Agron J 78(1): 59-62.

Wientjes Y. C. J., Veerkamp R. F., Calus M. P. L. (2013). The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. Genetics 193(2): 621-631.

Wiggans S. C., Frey K. J. (1955). Photoperiodism in oats. Proc Iowa Acad Sci 62(1): 125-130.

Wilson T. A., Nicolosi R. J., Delaney B., Chadwell K., Moolchandani V., Kotyla T., Ponduru S., Zheng G. H., Hess R., Knutson N., Curry L., Kolberg L., Goulson M., Ostergren K. (2004). Reduced and high molecular weight barley beta-glucans decrease plasma total and non-HDL-cholesterol in hypercholesterolemic Syrian golden hamsters. J. Nutr. 134(10): 2617-2622.

Windhausen V. S., Atlin G. N., Hickey J. M., Crossa J., Jannink J.-L., Sorrells M. E., Raman B., Cairns J. E., Tarekegne A., Semagn K., Beyene Y., Grudloyma P., Technow F., Riedelsheimer C., Melchinger A. E. (2012). Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. G3 2(11): 1427-1436.

Winkler L. R., Bonman J., Chao S., Admassu Yimer B., Bockelman H., Esvelt Klos K. (2016). population structure and genotype–phenotype associations in a collection of oat landraces and historic cultivars. Front Plant Sci 7(1077).

Wolever T. M., Gibbs A. L., Brand-Miller J., Duncan A. M., Hart V., Lamarche B., Tosh S. M., Duss R. (2011). Bioactive oat β-glucan reduces LDL cholesterol in Caucasians and non-Caucasians. Nutr J 10: 130.

Won S., Park J., Son J., Lee S., Park B., Park M., Park W., Chai H., Kim H., Lee J., Lim D. (2020). Genomic prediction accuracy using haplotypes defined by size and hierarchical clustering based on linkage disequilibrium. Front Genet 11(134).

Wray N. R., Yang J., Hayes B. J., Price A. L., Goddard M. E., Visscher P. M. (2013). Pitfalls of predicting complex traits from SNPs. Nature reviews. Genetics 14(7): 507-515.

Xu S. (2003). Theoretical basis of the Beavis effect. Genetics 165(4): 2259-2268.

Yan H., Bekele W. A., Wight C. P., Peng Y., Langdon T., Latta R. G., Fu Y. B, Diederichsen A., Howarth C. J., Jellen E. N., Boyle B., Wei Y., Tinker N. A. (2016). High-density marker profiling confirms ancestral genomes of Avena species and identifies D-genome chromosomes of hexaploid oat. Theor Appl Genet 129(11): 2133-2149.

Yan W. (2021). Estimation of the optimal number of replicates in crop variety trials. Front Plant Sci 11(2231).

Yang J., Li Y., Cao H., Yao H., Han W., Sun S. (2019). Yield-maturity relationships of summer Maize from 2003 to 2017 in the Huanghuaihai Plain of China. Sci Rep 9(1): 11417.

Yang J., Zhang J. (2006). Grain filling of cereals under soil drying. New Phytol. 169(2): 223-236.

Zhao C., Zhang Y., Du J., Guo X., Wen W., Gu S., Wang J., Fan J. (2019). Crop phenomics: current status and perspectives. Front Plant Sci 10(714).

Zhao J., Kebede A. Z., Menzies J., Paczos-Grzęda E., Chong J., Mitchell Fetch J., Beattie A. D, Peng Y.-Y., McCartney C. (2020). Chromosomal location of the crown rust resistance gene *Pc98* in cultivated oat (*Avena sativa* L.). Theor Appl Genet 133(4): 1109-1122.

Zhao J., Tang X., Wight C., Tinker N., Jiang Y., Yan H., Ma J., Lan X., Wei Y., Ren C. (2018). Genetic mapping and a new PCR-based marker linked to a dwarfing gene in oat (*Avena sativa* L.). Genome 61(7): 497-503.

Zhao Y., Zeng J., Fernando R. L., Reif J. C. (2013). Genomic prediction of hybrid wheat performance. Crop Sci 53(3): 802-810.

Zhong S., Dekkers J., Fernando R. L., Jannink J. (2009). Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study. Genetics 182(1): 355-364.

Zhou M., Robards K., Glennie-Holmes M., Helliwell S. (1999). Oat lipids. J Am Oil Chem Soc 76(2): 159-169.

Zhu C., Gore M., Buckler E. S., Yu J. (2008). Status and prospects of association mapping in plants. Plant Gen 1(1): 5-20

APPENDIX

Table A. Correlation coefficients between yield and correlated traits, seeding date and influential environmental variables for three check varieties and 305 lines grown in the WCORT from 2002-2014.

	MAT	**HT**	**BG**	**Seeding Date**	**MayPPT**	**JunPPT**	**JulPPT**	**AugPPT**
WCORT Lines	0.36	0.68	-0.15	0.03	0.05	0.10	0.32	0.06
CDC Dancer	0.30	0.62	-0.36	0.00	0.03	0.11	0.31	0.08
AC Morgan	0.51	0.67	-0.42	-0.04	-0.06	0.06	0.34	0.08
Leggett	0.26	0.66	-0.34	0.06	-0.05	0.06	0.38	0.13
	MayJunPPT	**May-Jul PPT**	**May-Aug PPT**	**Sep-Apr PPT**	**Sep-May PPT**	**Sep-Jun PPT**	**Sep-Jul PPT**	**Sep-Aug PPT**
WCORT Lines	0.13	0.22	0.25	-0.17	-0.10	-0.03	0.08	0.10
CDC Dancer	0.13	0.22	0.25	-0.14	-0.09	-0.01	0.10	0.12
AC Morgan	0.03	0.15	0.19	-0.22	-0.19	-0.12	0.00	0.03
Leggett	0.08	0.23	0.29	-0.22	-0.18	-0.10	0.05	0.09
	MayMax	**MayMin**	**MayMean**	**MayDiff**	**MayHDD**	**MayCDD**		
WCORT Lines	0.09	-0.12	-0.03	0.23	-0.01	-0.11		
CDC Dancer	0.12	-0.11	0.00	0.24	-0.01	-0.09		
AC Morgan	0.06	-0.21	-0.10	0.31	0.13	-0.13		
Leggett	0.11	-0.24	-0.10	0.33	0.09	0.01		
	JunMax	**JunMin**	**JunMean**	**JunDiff**	**JunHDD**	**JunCDD**		
WCORT Lines	-0.25	-0.11	-0.20	-0.13	0.17	-0.19		
CDC Dancer	-0.27	-0.12	-0.21	-0.13	0.15	-0.21		
AC Morgan	-0.29	-0.25	-0.30	-0.01	0.19	-0.25		
Leggett	-0.16	-0.07	-0.12	-0.08	0.08	-0.12		
	JulMax	**JulMin**	**JulMean**	**JulDiff**	**JulHDD**	**JulCDD**		
WCORT Lines	-0.37	-0.26	-0.35	-0.15	0.32	-0.30		
CDC Dancer	-0.42	-0.27	-0.38	-0.19	0.32	-0.32		
AC Morgan	-0.46	-0.39	-0.46	-0.09	0.37	-0.39		
Leggett	-0.42	-0.26	-0.37	-0.23	0.30	-0.30		
	AugMax	**AugMin**	**AugMean**	**AugDiff**	**AugHDD**	**AugCDD**		
WCORT Lines	-0.16	-0.17	-0.18	-0.01	0.09	-0.26		
CDC Dancer	-0.17	-0.15	-0.18	-0.06	0.11	-0.23		
AC Morgan	-0.25	-0.30	-0.30	0.01	0.24	-0.35		
Leggett	-0.16	-0.06	-0.12	-0.11	0.07	-0.19		

Table B. Correlation coefficients between β-glucan and correlated traits, seeding date and influential environmental variables for three check varieties and 305 lines grown in the WCORT from 2002-2014

	MAT	**HT**	**YLD**	**Seeding Date**	**MayPPT**	**JunPPT**	**JulPPT**	**AugPPT**
WCORT Lines	-0.24	-0.05	-0.15	0.10	0.06	-0.01	-0.10	0.00
CDC Dancer	-0.41	-0.37	-0.36	0.06	-0.14	0.02	-0.30	0.09
AC Morgan	-0.54	-0.19	-0.42	0.23	0.06	-0.05	-0.34	-0.01
Leggett	-0.15	-0.22	-0.34	0.04	0.10	0.00	-0.20	0.25

	MayJunPPT	**May-Jul PPT**	**May-Aug PPT**	**Sep-Apr PPT**	**Sep-May PPT**	**Sep-Jun PPT**	**Sep-Jul PPT**	**Sep-Aug PPT**
WCORT Lines	0.04	-0.01	-0.01	0.07	0.09	0.08	0.04	0.04
CDC Dancer	-0.03	-0.16	-0.10	-0.05	-0.09	-0.06	-0.15	-0.11
AC Morgan	0.02	-0.14	-0.13	0.10	0.12	0.09	-0.04	-0.03
Leggett	0.10	-0.02	0.06	0.27	0.28	0.26	0.16	0.21

	MayMax	**MayMin**	**MayMean**	**MayDiff**	**MayHDD**	**MayCDD**
WCORT Lines	0.15	0.20	0.20	-0.10	-0.20	0.10
CDC Dancer	0.33	0.32	0.36	-0.07	-0.35	0.35
AC Morgan	0.17	0.25	0.24	-0.14	-0.26	0.21
Leggett	-0.21	0.06	-0.07	-0.23	-0.10	-0.09

	JunMax	**JunMin**	**JunMean**	**JunDiff**	**JunHDD**	**JunCDD**
WCORT Lines	0.22	0.21	0.23	-0.05	-0.13	0.19
CDC Dancer	0.46	0.26	0.38	0.11	-0.15	0.30
AC Morgan	0.43	0.46	0.49	-0.17	-0.24	0.38
Leggett	0.27	0.11	0.20	0.12	-0.02	0.19

	JulMax	**JulMin**	**JulMean**	**JulDiff**	**JulHDD**	**JulCDD**
WCORT Lines	0.28	0.24	0.28	0.00	-0.20	0.26
CDC Dancer	0.55	0.27	0.43	0.28	-0.26	0.41
AC Morgan	0.56	0.40	0.51	0.10	-0.34	0.47
Leggett	0.48	0.40	0.47	0.03	-0.27	0.47

	AugMax	**AugMin**	**AugMean**	**AugDiff**	**AugHDD**	**AugCDD**
WCORT Lines	0.21	0.21	0.23	0.00	-0.21	0.22
CDC Dancer	0.33	0.27	0.32	0.09	-0.24	0.40
AC Morgan	0.21	0.23	0.24	-0.03	-0.21	0.24
Leggett	0.08	0.20	0.16	-0.17	-0.16	0.17

Table C. Correlation coefficients between plant height and correlated traits, seeding date and influential environmental variables for three check varieties and 305 lines grown in the WCORT from 2002-2014.

	MAT	**YLD**	**BG**	**Seeding Date**	**MayPPT**	**JunPPT**	**JulPPT**	**AugPPT**
WCORT Lines	0.28	0.62	-0.05	0.10	0.25	0.28	0.24	0.07
CDC Dancer	0.24	0.67	-0.37	0.09	0.30	0.34	0.26	0.07
AC Morgan	0.26	0.62	-0.19	0.06	0.27	0.28	0.28	0.07
Leggett	0.23	0.66	-0.22	0.07	0.13	0.22	0.21	0.18

	MayJunPPT	**May-Jul PPT**	**May-Aug PPT**	**Sep-Apr PPT**	**Sep-May PPT**	**Sep-Jun PPT**	**Sep-Jul PPT**	**Sep-Aug PPT**
WCORT Lines	0.39	0.41	0.43	-0.02	0.10	0.24	0.32	0.32
CDC Dancer	0.45	0.48	0.47	-0.02	0.12	0.29	0.37	0.36
AC Morgan	0.40	0.44	0.45	-0.04	0.09	0.23	0.33	0.33
Leggett	0.30	0.30	0.41	-0.07	0.01	0.14	0.23	0.26

	MayMax	**MayMin**	**MayMean**	**MayDiff**	**MayHDD**	**MayCDD**
WCORT Lines	0.01	0.07	0.04	-0.07	-0.12	-0.12
CDC Dancer	0.01	0.07	0.05	-0.07	-0.14	-0.15
AC Morgan	0.02	0.07	0.05	-0.07	-0.12	-0.11
Leggett	-0.02	0.00	-0.01	-0.01	-0.08	0.00

	JunMax	**JunMin**	**JunMean**	**JunDiff**	**JunHDD**	**JunCDD**
WCORT Lines	-0.35	0.01	-0.19	-0.38	0.15	-0.23
CDC Dancer	-0.37	0.02	-0.17	-0.40	0.13	-0.26
AC Morgan	-0.33	0.02	-0.15	-0.36	0.11	-0.24
Leggett	-0.26	0.02	-0.08	-0.30	0.09	-0.15

	JulMax	**JulMin**	**JulMean**	**JulDiff**	**JulHDD**	**JulCDD**
WCORT Lines	-0.36	-0.14	-0.30	-0.30	0.29	-0.23
CDC Dancer	-0.35	-0.11	-0.27	-0.31	0.25	-0.23
AC Morgan	-0.38	-0.13	-0.29	-0.32	0.24	-0.25
Leggett	-0.42	-0.19	-0.34	-0.36	0.34	-0.25

	AugMax	**AugMin**	**AugMean**	**AugDiff**	**AugHDD**	**AugCDD**
WCORT Lines	-0.09	-0.07	-0.09	-0.04	-0.01	-0.18
CDC Dancer	-0.08	-0.05	-0.07	-0.05	-0.03	-0.17
AC Morgan	-0.10	-0.05	-0.09	-0.07	-0.01	-0.18
Leggett	-0.11	0.04	-0.04	-0.16	-0.04	-0.12

Table D. Correlation coefficients between days to maturity and correlated traits, seeding date and influential environmental variables for three check varieties and 305 lines grown in the WCORT from 2002-2014.

	YLD	HT	BG	Seeding Date	MayPPT	JunPPT	JulPPT	AugPPT
WCORT Lines	0.36	0.28	-0.24	-0.18	-0.07	0.09	0.44	0.48
CDC Dancer	0.30	0.24	-0.41	-0.21	-0.09	0.11	0.44	0.48
AC Morgan	0.51	0.26	-0.54	-0.2	-0.14	0.08	0.44	0.47
Leggett	0.26	0.23	-0.15	-0.18	-0.13	0.16	0.5	0.47

	MayJunPPT	May-Jul PPT	May-Aug PPT	Sep-Apr PPT	Sep-May PPT	Sep-Jun PPT	Sep-Jul PPT	Sep-Aug PPT
WCORT Lines	0.02	0.23	0.28	-0.16	-0.16	-0.1	0.06	0.12
CDC Dancer	0.03	0.24	0.28	-0.13	-0.14	-0.07	0.09	0.14
AC Morgan	-0.03	0.19	0.23	-0.19	-0.21	-0.15	0.01	0.07
Leggett	0.03	0.29	0.29	-0.21	-0.22	-0.13	0.07	0.09

	MayMax	MayMin	MayMean	MayDiff	MayHDD	MayCDD
WCORT Lines	-0.29	-0.46	-0.43	0.21	0.42	-0.24
CDC Dancer	-0.3	-0.46	-0.43	0.2	0.41	-0.24
AC Morgan	-0.3	-0.5	-0.46	0.25	0.44	-0.28
Leggett	-0.26	-0.56	-0.5	0.33	0.52	-0.19

	JunMax	JunMin	JunMean	JunDiff	JunHDD	JunCDD
WCORT Lines	-0.56	-0.51	-0.61	-0.01	0.62	-0.46
CDC Dancer	-0.58	-0.53	-0.63	0	0.64	-0.48
AC Morgan	-0.55	-0.56	-0.63	0.06	0.64	-0.48
Leggett	-0.53	-0.56	-0.62	0.13	0.65	-0.48

	JulMax	JulMin	JulMean	JulDiff	JulHDD	JulCDD
WCORT Lines	-0.63	-0.59	-0.70	-0.16	0.66	-0.61
CDC Dancer	-0.64	-0.59	-0.70	-0.15	0.67	-0.61
AC Morgan	-0.65	-0.62	-0.72	-0.13	0.68	-0.64
Leggett	-0.65	-0.62	-0.72	-0.14	0.67	-0.64

	AugMax	AugMin	AugMean	AugDiff	AugHDD	AugCDD
WCORT Lines	-0.5	-0.49	-0.60	-0.13	0.51	-0.5
CDC Dancer	-0.53	-0.51	-0.64	-0.14	0.54	-0.53
AC Morgan	-0.52	-0.53	-0.63	-0.11	0.53	-0.53
Leggett	-0.41	-0.35	-0.50	-0.06	0.41	-0.41

Table E. Correlation coefficients between days to heading and correlated traits, seeding date and influential environmental variables for three check varieties and 305 lines grown in the WCORT from 2002-2014.

	YLD	HT	MAT	Seeding Date	MayPPT	JunPPT	JulPPT	AugPPT
WCORT Lines	-0.02	0.10	0.42	-0.61	0.11	0.27	-0.08	0.29
CDC Dancer	-0.06	0.12	0.38	-0.66	0.18	0.37	-0.09	0.31
AC Morgan	-0.08	0.02	0.37	-0.65	0.12	0.29	-0.13	0.31
Leggett	-0.21	0.04	0.43	-0.63	0.17	0.17	-0.18	0.16
	MayJunPPT	**May-Jul PPT**	**May-Aug PPT**	**Sep-Apr PPT**	**Sep-May PPT**	**Sep-Jun PPT**	**Sep-Jul PPT**	**Sep-Aug PPT**
WCORT Lines	0.10	0.01	0.09	-0.0355	-0.0267	0.0349	-0.0153	0.0482
CDC Dancer	0.18	0.07	0.14	-0.0335	-0.0062	0.0839	0.025	0.0843
AC Morgan	0.09	-0.02	0.06	-0.0686	-0.0577	0.0089	-0.0626	0.0017
Leggett	0.01	-0.12	-0.10	0.0062	0.0225	0.0113	-0.0828	-0.0691
	MayMax	**MayMin**	**MayMean**	**MayDiff**	**MayHDD**	**MayCDD**		
WCORT Lines	-0.24	-0.20	-0.24	0.00	0.21	-0.12		
CDC Dancer	-0.23	-0.17	-0.23	-0.02	0.19	-0.13		
AC Morgan	-0.23	-0.21	-0.24	0.02	0.20	-0.12		
Leggett	-0.31	-0.19	-0.29	-0.10	0.25	0.00		
	JunMax	**JunMin**	**JunMean**	**JunDiff**	**JunHDD**	**JunCDD**		
WCORT Lines	-0.36	-0.11	-0.27	-0.25	0.12	-0.07		
CDC Dancer	-0.37	-0.03	-0.24	-0.32	0.08	-0.06		
AC Morgan	-0.32	-0.06	-0.22	-0.25	0.10	-0.02		
Leggett	-0.27	-0.17	-0.25	-0.08	0.04	-0.12		
	JulMax	**JulMin**	**JulMean**	**JulDiff**	**JulHDD**	**JulCDD**		
WCORT Lines	0.00	-0.07	-0.04	0.09	0.01	-0.06		
CDC Dancer	0.05	-0.01	0.02	0.08	-0.05	-0.02		
AC Morgan	0.06	-0.03	0.02	0.11	-0.03	-0.01		
Leggett	0.03	-0.08	-0.03	0.15	0.00	-0.06		
	AugMax	**AugMin**	**AugMean**	**AugDiff**	**AugHDD**	**AugCDD**		
WCORT Lines	-0.01	-0.02	-0.02	0.01	0.00	-0.06		
CDC Dancer	0.00	-0.01	-0.01	0.01	-0.03	-0.07		
AC Morgan	0.03	0.02	0.02	0.02	-0.04	-0.02		
Leggett	0.11	0.05	0.09	0.08	-0.07	0.07		

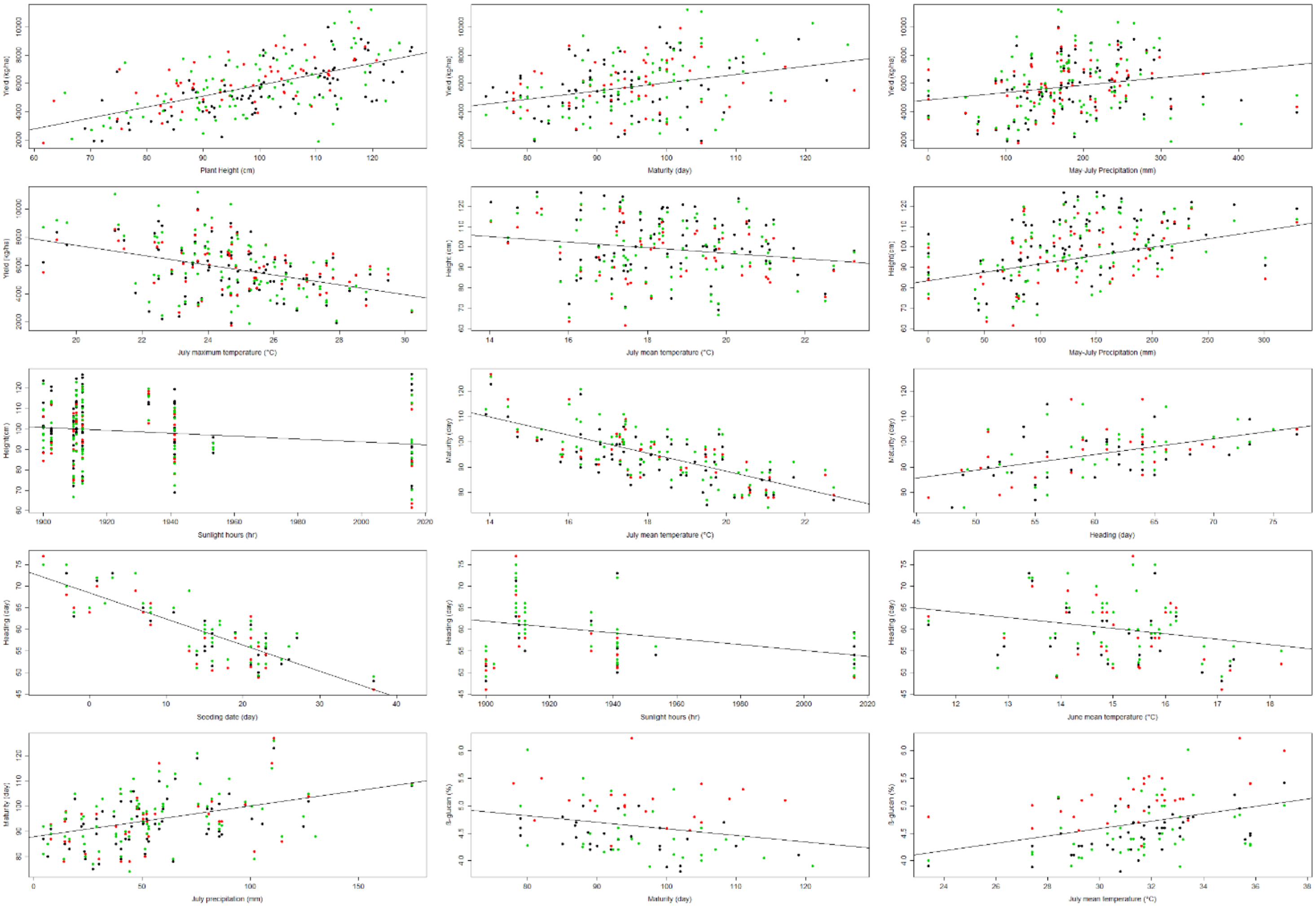


Figure A. Bivariate scatterplots showing the relationships among yield, β-glucan and significant path variables used in structural equation models developed for yield and β-glucan for each of three oat varieties (CDC Dancer: red dots, AC Morgan: green dots and Leggett: black dots).

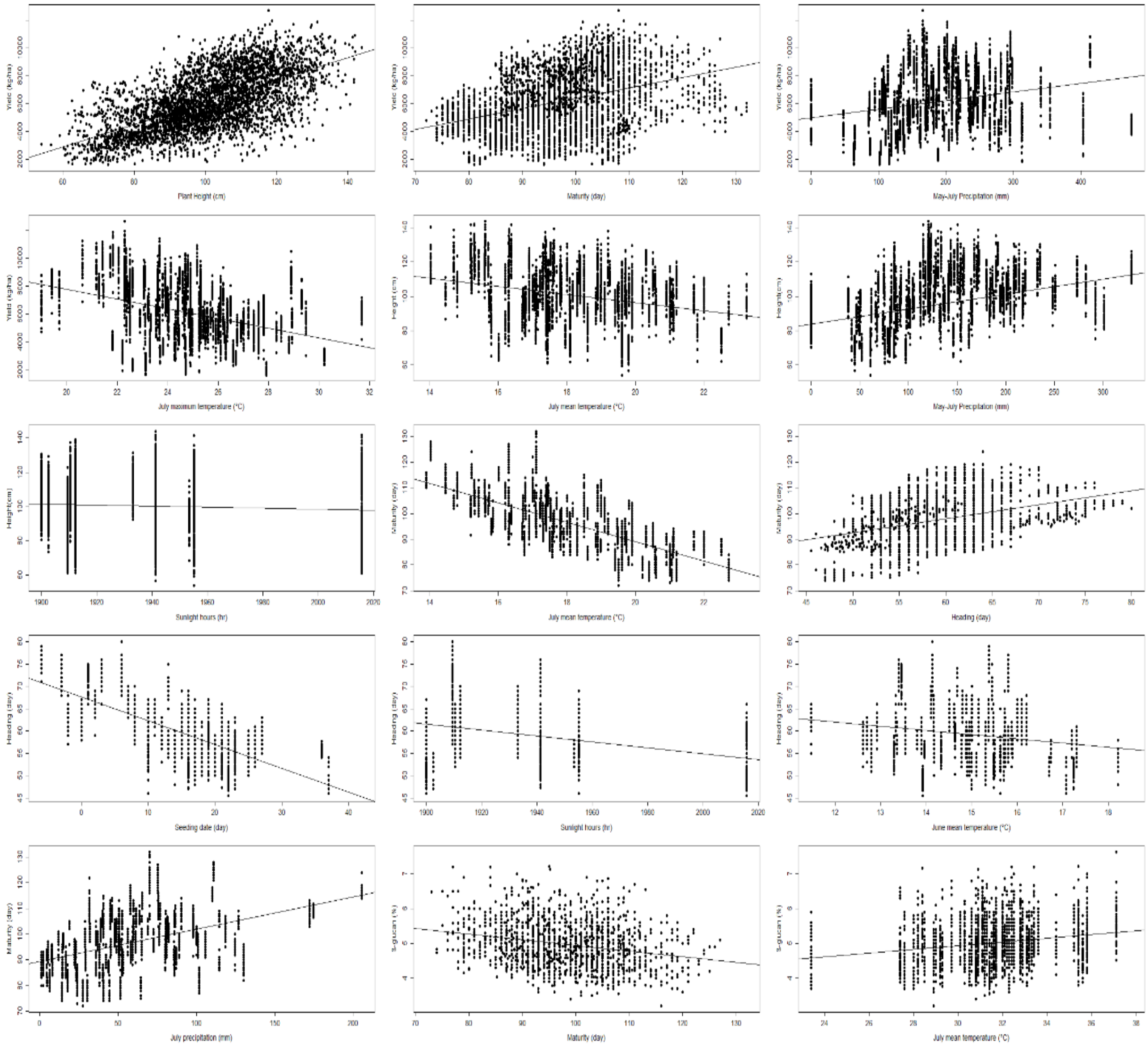


Figure B. Bivariate scatterplots show the relationships among yield, β-glucan and significant path variables used in structural equation models developed for yield and β-glucan in the 305 lines of the WCORT population.