DEVELOPMENT OF GENOMIC TOOLS FOR ACCELERATED BREEDING OF CRESTED WHEATGRASS [*Agropyron cristatum* (L.) Gaertn.]

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

KIRAN BARAL

PERMISSION TO USE

In presenting this thesis/dissertation in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Request for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Plant Sciences
University of Saskatchewan
51 Campus Drive
Saskatoon, SK, 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

Crested wheatgrass [*Agropyron cristatum* (L.) Gaertn.], particularly valued as forage crop for its early spring vigor providing high quality, highly palatable forage before the native forage grasses. However, the nutritive value of the forage starts declining after heading and becomes less palatable to livestock. The objectives of this thesis research were: 1) To assess the genetic diversity and population structure of crested wheatgrass using genome-wide Single Nucleotide Polymorphism (SNP) markers generated using Genotyping-by-Sequencing (GBS); 2) To develop a high-resolution linkage map in an intraspecific F_1 mapping population of crested wheatgrass; and 3) To study the prediction ability of genomic selection models in crested wheatgrass. Breeding initiatives for the development of high yield, high palatable and late maturing cultivars begins with uncovering of the extent of genetic diversity available and its utilization in the breeding program. Molecular characterization of un-sequenced plant species with complex genomes is now possible by GBS using recent next generation sequencing technologies (NGS). SNP markers were used to assess the genetic diversity present in 192 genotypes from 12 tetraploid lines of crested wheatgrass. The model-based Bayesian analysis revealed four major clusters of the samples assayed. The diversity analysis revealed 15.8% of SNP variation residing among the 12 lines. The principal coordinates analysis and dendrogram were able to distinguish four lines of Asian origin from Canadian cultivars and breeding lines. With an attempt to utilize SNP markers towards molecular breeding, a study to develop genetic linkage maps in *Agropyron cristatum* utilizing segregating F_1 mapping population developed from intraspecific cross of two diploid elite cultivars grouped 678 SNP markers into seven linkage groups. The linkage map generated here could be saturated for quantitative trait loci (QTL) mapping and marker assisted selection (MAS). However, MAS is less effective for improvement of traits controlled by a large number of small effect QTLs. Genomic selection (GS) is gaining attention towards improving genetic gain of quantitative traits. Genomic selection study showed moderate prediction accuracies (in range of 0.20–0.41) for traits such as ADF, TPP, CD, PH, LW and FRG suggesting that it is possible to implement GS. The additive GS models in this study were similar in prediction suggesting prediction ability is less influenced by the model preference. Increasing SNP densities did not improved prediction ability except for few traits suggesting the traits could have been influenced by large number of small effect QTLs such that GBS approach adopted was unable to capture the effect across the genome. However, GS in crested wheatgrass can be improved with refining structure and size of training population, field experiments to accurately measure phenotypic records and

utilize GS models to incorporate genotype-by-environment and gene interactions/ epistasis, for this comparison of parametric models with non-parametric models in GS could provide insight.

ACKNOWLEDGEMENTS

I extend my sincere gratitude to my co-supervisors Drs. Bruce E. Coulman, Yong-Bi Fu and Bill Biligetu for the opportunity, insightful guidance, support, and motivation throughout my study and research. I am thankful for the extensive knowledge, serious attitude, and commitment of the team to accomplish my research and make this study a success. Special thanks to my committee members, Drs. Bunyamin Tar'an, Helen Booker, Xiao Qiu for their valuable suggestions. Special thank goes to Dr. Steve Larson (United States Department of Agriculture) for accepting to be my external examiner. I extend my gratitude to the chair Dr. Tom Warkentin and Dr. Pierre Hucl for chairing the session on the day of defense.

Sincere thanks to Gregory Peterson, Carolee Horbach and Cheryl Duncan (Agriculture and Agri-Food Canada) for the technical support during sample preparation, DNA extraction, library preparation, bioinformatics and greenhouse experiments. My thanks are also extended to Dashnyam Byambatseren and Ninh Cao of the University of Saskatchewan forage breeding group for their technical help during field and lab experiments. I appreciate faculty and staff of the Department of Plant Sciences and the support I received from my fellow graduate students at the Department of Plant Sciences and summer students. Special thanks go to Surendra Bhattarai, Dr. Nitya Nanda Khanal, Hu Wang, Amarjargal Gungaabayar, Dr. Fangqin Zeng, and Samuel Tandoh of the forage breeding group for their assistance and resourcefulness. I also appreciate Drs. Kirstin E. Bett, Amidou N'Diaye, Amit Deokar, Teketel Haile and Lester Young for their advice and support in linkage mapping and genomic selection projects.

I am grateful for the financial support provided by the Beef Cattle Research Council for the project. I also greatly appreciate Bursary and Scholarships awarded by the Department of Plant Sciences. I am also thankful to Nepal Agricultural Research Council for granting me study leave.

Finally, I would like to offer special thanks to my beloved wife Sapana Acharya and charming daughter Aakrisha Baral for their patience, love and support during my studies, my parents Keshab Raj Baral and Yam Kumari Baral, sister Kabita Baral Adhikari for your support, advice and encouragement throughout this journey.

DEDICATION

To my late grandmother Goma Devi Baral who taught me morals, values and compassion. To my teachers (Gurus) for their guidance.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

°C	Degree Celsius
AAFC	Agriculture and Agri-Food Canada
ADF	Acid detergent fiber
AFLP	Amplified fragment length polymorphism
AMOVA	Analysis of molecular variance
ANGSD	Analysis of Next Generation Sequencing Data
BLUP	Best linear unbiased prediction
bp	Base pair
CbyT	Character by taxa
CD	Clump diameter
cM	Centimorgans
CP	Cross pollinating
CP	Crude protein
DM	Dry matter
DNA	Deoxyribonucleic acid
DTH	Days to heading
ESTs	Expressed sequence tags
ESV	Early spring vigor
FRG	Fall regrowth
GBLUP	Genomic best linear unbiased prediction
GBS	Genotyping-by-Sequencing
GDD	Growing degree days
GEBVs	Genomic estimated breeding values
GS	Genomic selection
LASSO	Least absolute shrinkage and selection operator
LD	Linkage disequilibrium
LG	Linkage Group
LOD	Logarithm of odds
LW	Leaf width

MAS	Marker assisted selection
Me	number of independent chromosomal segments
ML	Maximum likelihood
NCBI	National center for biotechnology information
NDF	Neutral detergent fiber
Ne	Effective population size
NGS	Next generation sequencing
NJ	Neighbor-joining
N_{QTL}	number of QTL
PCoA	Principal coordinates analysis
PCR	Polymerase chain reaction
PGRC	Plant gene resources of Canada
PH	Plant height
QTL	Quantitative trait loci
RAPD	Randomly amplified polymorphic DNA
RFLP	Restriction fragment length polymorphism
RGAH	Regrowth after harvest
RRBLUP	Ridge regression linear unbiased prediction
SCARs	Sequence characterized regions
SNP	Single Nucleotide Polymorphism
SRA	Sequence read archive
SSR	Simple Sequence Repeat
STSs	Sequence tagged sites
TPP	Tillers per plant
UNEAK	Universal network enabled analysis kit
UPGMA	Unweighted pair group method, with arithmetic mean
USDA-ARS	United states department of Agriculture- Agricultural research service

1. INTRODUCTION

Canada produces forages in all agricultural regions. Approximately 3.7 million hectares of alfalfa (*Medicago sativa* L.) and alfalfa mixtures, 2 million hectares of tame hay and fodder crops, 165,000 hectares of forage seed, 5.1 million hectares of tame pasture and 14.3 million hectares of native rangeland accounted for around 39% of Canada's total farm area in the year 2016 (Statistics Canada 2016). Saskatchewan is recognized as a major producer of grains and oilseeds; however, the province is also an important producer of forages with approximately 32% of total farm area dedicated to forage production. In Saskatchewan, approximately 1.1 million hectares of alfalfa and alfalfa mixtures, 357,000 hectares of tame hay and fodder crops, 42,000 hectares of forage seeds, 2 million hectares of tame pasture and 4.6 million hectares of native rangeland were reported in the most recent census (Statistics Canada 2016). Other agricultural crops are produced on a smaller land area than forage crops. However, valuation of direct and indirect economic benefits of forage crop production is difficult given its diverse nature. Forage crops are used in diverse ways ranging from amenity or turf, soil conservation, seed production, domestic and export hay production, and grazing or stored feed for livestock. The most important use is feed for cattle, which rely on pasture and hay/silage of perennial and annual forage crops. It is important that Canadian cattle producers have provision of high yielding, high quality and well adapted forage cultivars to improve the economics of the cattle industry.

Crested wheatgrass, [*Agropyron cristatum* (L.) Gaertn.] is one among the important perennial forage species in Canada with desirable characteristics such as drought resilience, winter hardiness, extensive fibrous root system, easy to establish and high forage productivity (Rogler and Lorenz 1983). It is one of the important commodities among the tame hay and fodder crop grown in about 1.16 million hectares in western Canada (Statistics Canada 2016). It is ideal feed for cattle, sheep, horses and other livestock being particularly valuable in providing early season, nutritious and palatable forage (Li et al. 2004).

Conventional forage breeding is a long-term endeavor requiring 10 or more years to develop a new variety with improved productivity, disease resistance and adaptation to local climatic conditions. There is a need to develop genomic tools to improve genetic gain and reduce the time required for a selection cycle in forage breeding. Next generation sequencing could be utilized to identify the extent of underlying genetic diversity in various crested wheatgrass populations and to determine the most diverse germplasm for its use in a breeding program. Further, by generating high density molecular markers, high resolution linkage maps could be developed for the purpose of exploring genetic backgrounds and improving the complex traits in crested wheatgrass. Finally, a genomic selection approach based on the molecular markers and phenotypic data could be carried out to improve complex quantitative traits such as forage yield of crested wheatgrass. Genomic selection estimates the breeding value of individuals using all available molecular markers and corresponding phenotypic data using a statistical model (Meuwissen et al. 2001; de los Campos et al. 2009; Crossa et al. 2010; Heffner et al. 2011; Vitezica et al. 2011). In this thesis research it was hypothesized that: 1) The Genotyping-by-sequencing (GBS) application will generate large number of SNP markers in non-model autopolyploid crested wheatgrass; 2) High density genome-wide SNP markers will be useful tool for studying genetic diversity and population structure within this species; 3) SNP markers generated in crested wheatgrass can provide a higher resolution linkage map for crested wheatgrass than other genetic markers in F_1 mapping population of intraspecific cross; and 4) Combined genotypic and phenotypic data generated for crested wheatgrass predicts breeding values more precisely than only phenotypic information and improves selection accuracies in forage breeding. The objectives of this thesis research were: 1) To assess the genetic diversity and population structure of crested wheatgrass using genome-wide SNP markers generated using GBS application; 2) To develop a high-resolution linkage map in an intraspecific F_1 mapping population of crested wheatgrass using SNP markers; and 3) To study the prediction ability of genomic selection models in crested wheatgrass using SNP markers.

2. LITERATURE REVIEW

2.1. Forage breeding and genetics

2.1.1. *Taxonomy, biology and genetics of crested wheatgrass*

Crested wheatgrass [*Agropyron cristatum* (L.) Gaertn.] is a species within the crested wheatgrass complex which have a P genome (Asay and Jensen, 1996; Chen et al. 2013), and exists in diploid (*2n=2x*=14), tetraploid (*2n=4x*=28) and hexaploid (*2n=6x*=42) forms (Mellish et al. 2002; Copete et al. 2018).

Crested wheatgrass has a finely branched fibrous root system extending up to a meter in depth. The inflorescence of crested wheatgrass is a spike on culms of 20—90 cm height growing in a dense bunch. Dense spikelets (awned or awnless) are oblong, laterally compressed and arranged along the rachis to form flat spikes, tapering towards the tip. They contain about 3 to 8 overlapping fertile florets per spikelet. Seeds are covered by glumes lemma and palea that are generally tapering to the tip or with short awns. There are many leaves both at the base and along the stem. The leaf blades are slender, extended, flat and pubescent on the upper surface (Ogle, 2006; Clayton et al., 2015).

Crested wheatgrass, a C_3 perennial bunchgrass is very winter hardy and tolerant to drought (Henderson and Naeth 2005; Vaness and Wilson 2007). It is native to temperate frigid and desert regions of Eurasia, and was introduced to the University of Saskatchewan in 1911 and distributed throughout western Canada in 1927 (Smoliak and Bjorge 1981). It was extensively used for revegetation of abandoned and overgrazed rangelands of the prairies in 1930's (Smoliak and Bjorge 1981). Crested wheatgrass has acclimatized to the dry Canadian prairies as an important pasture and hay species. Early studies indicated that the nutrient content and feed value of crested wheatgrass has been considered equal or better than *Agropyron tenerum* Vas. (Western rye grass), *Phleum pratense* (timothy) and *Bromus inermis Leyss.* (smooth bromegrass) (Kirk 1932). It can withstand close grazing and trampling by animals but is not adapted to areas with high water tables (Smoliak and Bjorge 1981; Yu et al. 2012a). Some crested wheatgrass species have been found to be resistant to diseases of wheat (*Triticum aestivum* L.), thus, they are potential genetic

resources for wheat improvement (Sharma et al. 1984; Dong et al. 1992). Crested wheatgrass is an outcrossing species with high degree of self-incompatibility; thus, individual plants are highly heterozygous and heterogeneous.

2.1.2. *Forage breeding, issues and challenges*

Humans have been making selection decisions for the improvement of desirable traits in cultivated plants for food and feed since thousands of years. This unconscious selection based on high yielding, disease free and easy to harvest plants has laid the foundation of modern crop diversity and plant breeding. Nevertheless, conscious effort towards selection of superior forage species is a recent approach dating back to late 18th century (Casler 1997; Casler and Vogel 1999) where selection was based on plant phenotypes as enhanced plant vigor, lower disease symptoms, lower senescence and acceptability to livestock, but without laboratory techniques to evaluate forage quality. Forage breeding for varietal improvement and cultivar development formally began in the late 19th century with selection for morphological, phenological, disease, persistence and vigor related traits (Casler 1997). Towards the early 20th century, breeding efforts in *Agropyron* species were initiated with the introduction, improvement and release of first variety "Fairway" in 1932 (Elliott and Bolton 1970). Breeding initiatives in crested wheatgrass involved selection of improved lines from plant introductions, the study of chromosome pairing and reproduction (Dewey 1974), interploidy breeding (Asay 1986) and hybridization to improve morphological and quality traits (Asay 1990). Dewey (1974) concluded from several studies that same basic genome that exists in the polyploidy series (diploid, tetraploid and hexaploid) of crested wheatgrass can be considered as a single gene pool for mining various trait to combine at a desired ploidy.

Conventional forage breeding and improvement began with the exploitation of natural variability within ecotypes, and with the introduction of materials to improve the breeding pool. Crested wheatgrass breeding to date has been based on the phenotypic evaluation of target traits. Many such traits are quantitative in nature, controlled by many genes of small effect. Improvement of population depends on

the heritability of the trait. Evaluation of such traits is carried out in replicated trials under different environments. An effective method of population improvement of perennial outcrossing species such as crested wheatgrass involves recurrent selection with or without progeny testing (Conaghan and Casler 2011). This breeding method improves the population mean for a trait by increasing the frequency of favourable alleles in the population through repeated selection of best performing plants and release of an improved population as a synthetic cultivar (Vogel and Lamb 2007). Each selection cycle is completed after a new population is formed by crossing the best plants from an existing population. Phenotypic recurrent selection is a widely adopted approach for the improvement of perennial outcrossing forages (Vogel and Pedersen 1993; Wilkins and Humphreys 2003; Conaghan and Casler 2011; Resende et al. 2013), where phenotypically superior individual plants are selected for crossing to create a new population for subsequent selection cycles. Crossing can be done either to produce half sibs through open pollination of the selected plants and bulking of the harvested seeds or to produce full sibs through paired cross among the selected plants and bulking of harvested seeds from each cross. In genotypic recurrent selection parent plants are selected based on the performance of their progenies (Casler and Brummer 2008). Progeny testing (half-sib or full-sib families) evaluates the breeding value of parents with the superior parents being selected for the development of subsequent populations. Recurrent selection results in the development of synthetic or open pollinated cultivars.

Casler and Brummer (2008) reviewed and reported low progress in yield improvement of forage crops, except for a few exceptional incidences. Forage breeding programs not only select for yield, but also improve plant persistence, disease resistance and forage quality. This takes 3 to 5 years of field evaluation for a single selection cycle in spaced-planted nurseries and requires at least 10-15 years to test, develop and register a new forage cultivar through conventional breeding approaches (Casler 1997; Wilkins and Humphreys 2003; Tester and Langridge 2010; Resende et al. 2013). Genetic gains resulting from recurrent selection are low in such traits. Moreover, selection of plants in spaced-planted nurseries which do not

accurately represent the intergenotypic competition on the sward density field conditions under which forages are grown may also limit breeding progress (Casler and Brummer 2008; Resende et al. 2013). Breeding systems that are based on phenotype alone require considerable effort and several evaluations within and among multiple environments. Recent developments in genetic analysis such as marker discovery and genotyping offer the potential for marker assisted selection or genomic selection for faster and more efficient delivery of results in breeding programs.

2.2. Additive and non-additive genetic effects

Many traits of interest that a plant breeder attempts to improve are quantitatively inherited and controlled by many genes at different loci. Phenotypic variation in a quantitative trait is controlled by both additive and non-additive effects of alleles, and their interactions with environment. In a breeding program involving sexually reproducing species, additive genetic effects are of prime significance as they are heritable (White et al. 2007). Thus, current breeding programs for improving quantitative traits emphasize additive effects (Muñoz et al. 2014). Non-additive genetic effects such as dominance, overdominance and epistasis, do not accumulate over generations, and simply manifest the phenomenon of heterosis at each generation, however, both additive and non-additive effects contribute towards total genetic value (Falconer and Mackay 1996; Muñoz et al. 2014). Adaptation and persistence of perennial forage species are determined by the capacity of populations to accumulate favorable alleles and their stability over a wide range of environments, both within and between years, compared to single year environmental variation in annual crops. Determination of genetic value of a trait of interest relies on the estimation of genetic variance components which is better achieved using genetic markers than a conventional approach of pedigree and phenotypic records (Meuwissen et al. 2001), which are sometimes inaccurate or unavailable in out crossing species.

2.3. Genetic markers and their use in forage breeding

Variation in genomic DNA sequences with the ability to differentiate individuals, populations or species are genetic markers. Every genetic marker has a definite region in the genome of concern, and the differences in genotypes that each marker embodies can be tested for association with a phenotypic trait of interest (Semagn et al. 2006a). DNA based markers have facilitated genomic analysis and molecular breeding of crops. Recent advances in next-generation sequencing technologies and computation methods has transformed the quantity of marker identification. This has permitted its application in developing highly efficient single nucleotide polymorphism (SNP) markers for various downstream analysis including genetic diversity studies, linkage mapping and genomic selection for accelerated crop breeding and selection (Singh and Singh 2015).

2.3.1. *Molecular markers*

Genetic markers can be grouped in two sets: a) Classical markers and b) DNA markers. Morphological, biochemical and cytological markers come under classical markers, while DNA markers are further categorized as sequence based, non-PCR, PCR markers based on polymorphism detection methods (Collard et al. 2005; Semagn et al. 2006a). DNA marker technology available for plant breeding include Restriction Fragment Length Polymorphisms (RFLP) (Botstein et al. 1980), Amplified Fragment Length Polymorphisms (AFLPs) (Vos et al. 1995), Randomly Amplified Polymorphic DNA (RAPD) (Williams et al. 1990), Simple Sequence Repeats (SSR) (Salimath et al. 1995), sequence tagged sites (STSs), Expressed Sequence Tags (ESTs) (Adams et al. 1991), Sequence Characterized Regions (SCARs) (Paran and Michelmore 1993) and Single Nucleotide Polymorphisms (SNPs) (Lander 1996). These markers are favored for their absence of environmental influence and ease of detection irrespective of plant growth stage. SNPs are genetic markers in which there is a difference in a nucleotide at a specific genomic region between individuals. They are functionally essential molecular markers making them easy to assay and interpret,

giving them advantage over other markers for genetic diversity assessment, characterization and sub-genome differentiation (Collard et al. 2005; Ganal et al. 2009; He et al. 2014; Singh and Singh 2015).

The relationship, diversity and ploidy level of species in the genus *Agropyron* were studied using morphological, taxonomic and cytological information (Dewey and Asay 1982; Asay et al. 1986). Cytological research has been carried out to explain variation in *A. cristatum*, *A. imbricatum and A. pectinatum* based on the presence and absence of B chromosomes (Asghari et al. 2007). The protein marker gliadin was used to identify substantial genetic diversity among different species of *Agropyron* (Chen et al. 2013). A study on inter-population relationship and diversity within and among populations from *Agropyron* species using AFLPs found the majority of the variation within populations, but also there were significant AFLP differences among the populations (Mellish et al. 2002). Genetic characterization of selected crested wheatgrass lines has been accomplished with the use of SSR markers (Che et al. 2008, 2011, 2015). Recently, SNP markers has been used to assess the genetic diversity present within and among 192 genotypes of 12 tetraploid crested wheatgrass lines (Baral et al. 2018).

2.3.2. *Next generation sequencing and SNP identification*

With recent advancement in sequencing technology, marker identification and number of useful markers have changed from fragment-based DNA markers to sequence based single nucleotide polymorphism (SNP) identification. Introduction of next generation sequencing technologies and computational pipelines have allowed sequencing, genotyping and marker discovery in a short time frame at low cost (Shendure and Ji 2008; Metzker 2010; Stapley et al. 2010; He et al. 2014). Next generation sequencing generates millions of short DNA sequence reads through parallel sequencing and imaging techniques (Shendure and Ji 2008). In brief, all next generation sequencing technologies include processes of library preparation, sequencing and imaging, genome alignment and assembly (Metzker 2010). Next generation sequencing introduces genotyping-by-sequencing (GBS) (Elshire et al. 2011), an approach based on genome reduction with restriction enzymes for the construction of highly multiplexed reduced

representation libraries for SNP marker discovery and genotyping without the requirement of a reference genome. GBS is being widely used in germplasm characterization and diversity studies, linkage mapping and plant breeding (Poland and Rife 2012a; Fu and Yang 2017; Li et al. 2017, 2018; Baral et al. 2018; Paudel et al. 2018).

Agropyron cristatum belongs to the tribe Triticeae, however, genetic resources available in crested wheatgrass are limited (Zhang et al. 2015a; Li et al. 2017; Zeng et al. 2017a, 2017b; Baral et al. 2018). GBS has been effective in marker discovery, genetic diversity study, genetic mapping, quantitative trait locus (QTL) analysis, genome wide association studies and genomic selection studies in species with varying ploidy levels (Li et al. 2014; Su et al. 2017; Vining et al. 2017; Hussain et al. 2017; Goonetilleke et al. 2018). Recently, studies in intermediate wheatgrass (*Thinopyrum intermedium* L.*)* (Kantarski 2015), northern wheatgrass (*Elymus lanceolatu ssp. Lanceolatuss* (Scribn. & J. G. Sm.) Gould) (Li et al. 2018), alfalfa (*Medicago sativa* L.) (Annicchiarico et al. 2015; Biazzi et al. 2017; Jia et al. 2018), switchgrass (*Panicum virgatum* L.) (Lipka et al. 2014) and crested wheatgrass (*Agropyron cristatum*) (Baral et al. 2018) among others highlighted the potential of GBS application in perennial forage genetic and breeding research.

2.3.3. *Linkage mapping*

A genetic linkage map is essentially a schematic representation of assembly and arrangement of genetic markers on a chromosome based on their frequency of recombination. The linkage relies on the concept that markers segregate through chromosomal rearrangement during meiosis. Recombination frequency between the markers can be estimated from the mapping population (progeny) created from a crossing experiment. Based on the recombination frequency, markers can be assigned to linkage groups using statistical procedure (Collard et al. 2005; Semagn et al. 2006b; Baxter et al. 2011). Linkage mapping with a high level of genome coverage forms the foundation of genome organization and QTL identification, marker-assisted breeding, comparative genomics of important species, a framework for a physical map and map-based cloning (Semagn et al. 2006b). Development of a comprehensive genetic linkage map is vital for

detecting the genome location of traits to facilitate crop improvement (Yu et al. 2012). Tracking segregation of molecular markers in a mapping population and their linear alignment based on pair-wise recombination frequencies makes it possible to develop a linkage map, which can be further used to identify genomic regions responsible in trait expression and detection of marker trait associations. Well-developed linkage maps could identify useful QTLs for use in marker assisted selection (Sieper and Chen 1998). They also help in the application of genomics through fine mapping, map-based cloning and development of tightly linked markers for marker assisted selection. The upsurge in sequence data, with concurrent drop in cost for sequencing, have made the development of high-density genetic linkage maps attainable for species with few or no genetic resources. The possibility to generate high-density genetic linkage maps for species of interest facilitates additional studies including QTL mapping, marker assisted selection, gene cloning and evolutionary studies (Kantarski 2015). However, limited genomic resources limits the application of genomic mapping in crested wheatgrass.

The first genetic linkage maps with 14 linkage groups for mapping populations of allotetraploid *Elymus wawawainsis* (J.Carlson & Barkworth) (Snake River Wheatgrass) and *Elymus lanceolatus* (Scribn. & J.G. Sm.) Gould (thickspike wheatgrass) were developed by (Mott et al. 2011) with EST-based SSR and STS markers. With the advent of technologies for marker discovery, attempts have been made to develop a genetic linkage map of crested wheatgrass using interspecific crosses (Yu et al. 2012; Zhang et al. 2015b; Zhou et al. 2018). The advent of NGS and GBS has the potential to develop marker data and genotyping of large numbers of lines in a mapping population derived from intraspecific crosses of *A. cristatum* at reduced cost, generating sufficient SNPs information in this outcrossing forage species for the development of linkage map.

2.3.4. *Methods of linkage map development*

Linkage mapping begins with the selection of divergent parents for the development of a mapping population. The parents should exhibit clear genetic differences for one or more traits of interest. The

parents should not be genetically wide resulting in sterile progenies or high levels of segregation distortion (Semagn et al. 2006b). Mapping populations in self-pollinating species are developed from parents that are highly homozygous (inbred). Cross pollinating (outcrossing) species do not tolerate inbreeding. Hence, two-way pseudo-testcross, half-sib and full-sib families resulting from designed crosses have been suggested for mapping in outcrossing species (Semagn et al. 2006b).

The accuracy of linkage maps is affected by types and sizes of mapping populations, marker types, statistical procedures and computer packages used in the construction of genetic maps (Ferreira et al. 2006). Ferreira et al. (2006) reported that population size with the lowest number of individuals provide several fragmented linkage groups and inaccurate locus order, thus a total of 200 individuals are required to construct reasonably accurate linkage maps for all population types. Another step in linkage analysis is determination of the sufficient number of polymorphic markers. In general, high level of polymorphism exists in cross pollinated species (Semagn et al. 2006b). Genotyping of the parents and the mapping population with high throughput technology yields high density markers useful for the linkage analysis. The marker information from the parents are inherited to the progenies. The marker data must be carefully inspected for all possible errors that could result from missing data, incorrect genotype coding, incorrect marker ordering or typographical errors. Hackett and Broadfoot (2003) from a simulation study reported that considerable influence can occur on the correct order and length of the maps even with minor typing errors. Similarly, large amount of missing information in the marker data introduces lack of information about the recombination events thus, reducing length and accuracy of maps (Hackett and Broadfoot 2003). Mapping functions are used to determine the relative positions of markers in a chromosome which relates the calculated recombination frequency to a map distance expressed as centimorgans (cM) where one cM represents the distance between two markers with 1% recombination rate (Semagn et al. 2006b). Haldane and Kosambi functions are widely used mapping functions. The map distance obtained from Haldane mapping are inflated than obtained for Kosambi, as the former assumes absence of interference between

crossovers (Semagn et al. 2006b). These two mapping functions can be interchanged to one another for estimating map distance (Kantarski, 2015).

Linkage maps of highly heterozygous species have been reported with OneMap (Margarido et al. 2007), FsLinkageMap (Tong et al. 2010) and JoinMap (Van Ooijen 2006). Monte Carlo multipoint maximum likelihood algorithm in Joinmap4.1 (Van Ooijen 2011) facilitates linkage analysis in full-sib families of outcrossing species.

2.3.5. *Current issues of linkage map development*

Molecular marker development and application in major crops as wheat, maize, rice, barley have changed the crop development process and enabled marker detection for trait of interests. However, marker development, gene mapping and quantitative trait loci (QTL) analysis in most forage grasses remains far behind the major crops. Conventional markers such as RAPD, AFLP, SSRs etc. are expensive and challenging to interpret into useful sequence-based markers (Baxter et al. 2011). Hence, a conventional mapping approach is limited by the type of marker technology adopted for the development of markers, associated cost for the generation of markers, the number of markers available to create linkage maps and the resolution of map thus created (Baxter et al. 2011). Parents used to create a mapping population have significant effect on the length of linkage map, density of markers distributed in linkage groups and in the resolution of final map created. Thus, the parents used must be distinct (Zhang et al., 2000; Yang et al., 2004); however, genetically distant parents may limit chromosome pairing and recombination, thereby reducing the recombination rate between linked loci, leading to segregation disorder. AFLP and RAPD markers used to develop the genetic linkage map of crested wheatgrass were unevenly distributed across the linkage map and generated gaps of more than 8cM in LG2, LG4 and LG5, including a gap of more than 20 cM in LG5 alone (Yu et al. 2012).

The limitations of conventional marker technologies could be avoided with the use of sequence derived SNPs for linkage mapping to anchor the marker positions identified in the genomes of model grass species by means of bioinformatics tools (Studer 2012).

2.3.6. *Marker assisted selection*

Marker assisted selection (MAS) is an indirect selection process based on the detection of markers linked to the genomic region of a trait of interest. The markers linked to the trait of interest could be functional markers (markers localized in the gene of interest), markers genetically associated to the trait of interest, or trait affected by QTLs (Francia et al. 2005). This approach facilitates selection of a desired trait in an early generation through the selectin of markers, thus accelerating the breeding process by reducing the required breeding time (Collard and Mackill 2008). With the availability of different types of DNA based markers and linkage maps, it is possible to select the markers (directly or indirectly) that are closely linked to the traits of interest and overcome the challenges of conventional breeding based on phenotypic selection (Francia et al. 2005). Most of the traits in forage species display continuous phenotypic variations which are controlled by quantitative trait loci (QTL). The QTLs involved could be determined using an approach called QTL analysis in a bi-parental mapping population using the method known as QTL mapping. This detects genomic regions (QTLs) that are statistically significantly associated to the trait of interest (Paterson et al. 1988). Recent developments in NGS and marker discovery platforms has enabled marker detection and subsequent use in the development of high-resolution genetic maps with increased genome coverage for QTL studies in a biparental mapping populations (Talukder and Saha 2017). However, in a polyploid perennial forage like crested wheatgrass, self-incompatibility and allogamy precludes the development of inbred lines. Thus, heterozygous parents are crossed to produce segregating F1 progeny on which QTL studies are implemented. QTL study on such populations results in poor resolution of the QTL positions (Bourke et al. 2018). Moreover, most of the quantitative traits are governed by a large number of genes with small effects which are difficult to detect through QTL analysis, unlike

large effect QTLs governed by a smaller number of genes (Francia et al. 2005). Recent advances in marker discovery and genotyping has enabled QTL studies in forages crop species including perennial ryegrass (*Lolium perenne* L.) (Shinozuka et al. 2012), meadow fescue (*Fescuta pratensis* Huds.) (Alm et al. 2011), alfalfa (*Medicago sativa* L.) (Adhikari et al. 2018) and switchgrass (*Panicum virgatum* L.) (Poudel et al. 2019).

2.3.7. *Genomic selection: Concepts and status of genomic selection in forage breeding*

Forage breeding programs have made moderate genetic gains per decade for major traits such as forage yield, quality and persistence in grasses (Wilkins and Humphreys 2003). Much of the gain has been from phenotypic rather than pedigree information (Henderson 1984) and selection of superior individuals has relied on assessed breeding values. Successful application of MAS in plant breeding has resulted in gene introgression or gene pyramiding for the improvement of traits that are controlled by a few large effect QTLs in major crops. However, lack of efficient marker systems, costs associated with marker generation, and identification of only the large effect QTLs in designed biparental mapping populations has limited the application of MAS in forage breeding. Genomic selection (GS) is a form of MAS that attempts to overcome some of the limitations of MAS. In contrary to MAS that uses markers that are statistically associated to a trait above a given threshold, GS utilizes all the information in a larger set of genome wide markers such as SNPs. These explain the total additive genetic variance and predict the genomic estimated breeding values (GEBV) for individuals without phenotypic records by simultaneously estimating the effect of all marker loci across the entire genome, hence capturing all the QTLs affecting the trait of interest (Meuwissen et al. 2001; Heffner et al. 2010; Jannink et al. 2010; Resende et al. 2014; Varshney et al. 2017). GS predicts the GEBVs with enough accuracy for selection without the requirement of phenotypic evaluation thereby reducing the cost and time associated with evaluation (Habier et al. 2007). GS envisages GEBVs for individuals with marker information in a prediction population (selection candidates) using a statistical prediction model developed using marker, pedigree and phenotypic information of the training population for any trait of importance in a breeding program (Meuwissen et al.

2001; Jannink et al. 2010; Heffner et al. 2011b). The training population used to train GS models must be genetically related to the breeding population. Application of GS for selection in a breeding program relies on the prediction accuracy of the genomic selection models. Prediction accuracies are estimated on a set of individuals that have both phenotypic and genotypic records termed as a validation population. Then, correlation between the GEBVs estimated using a GS model and the true breeding value (TBV) or the phenotypic record of the individuals in the validation population is the prediction accuracy of the GS model. Recent development in molecular genetics and bioinformatics has overcome the limitation of lack of marker information in forage species with limited or no prior marker information and have made GS a promising new approach for improving the genetic gain, especially for complex traits with low heritability. This is achieved through an increase in selection accuracy and average genetic gain per year, decrease in cost of evaluation per genotype and a reduction in number of breeding cycles (Heffner et al. 2010; Jannink et al. 2010; Resende et al. 2012, 2014; Lipka et al. 2014; Jia et al. 2018).

2.3.8. *Genomic selection models*

In the GS approach, genome-wide marker information is analyzed along with phenotype information to estimate the effect of each marker on the expression of trait phenotype. With increasing availability of low-cost high-density genome-wide markers in crops with no prior marker information, now it is possible to implement the latest crop breeding techniques such as GS to achieve better prediction and selection of candidates with improved genetic gain with reduced cost and time. Regression of phenotypic values on all available markers and its use to predict the breeding values as a sum of all marker effects is the basic idea of genomic selection (de Los Campos et al. 2013). The increasing availability of large number of genome-wide SNP markers may provide a challenge as the number of markers (p) exceeds the number of individuals with phenotypic record (n) by several fold resulting in overfitting of the regression model and low accuracy of the GEBVs (Jannink et al. 2010; Lorenz et al. 2011; de Los Campos et al. 2013; Pérez and de Los Campos 2014). Complexities associated with computational and statistical challenges due to

large-p with small-n regressions, and genetic mechanisms involving gene interactions can be resolved with the implementation of whole genome regression models performing variable selection, shrinkage of estimates, or combination of both (de Los Campos et al. 2013). Genomic selection aims to estimate marker effects using various statistical models to calculate GEBVs (phenotype). The GS models implemented to estimate the GEBVs differ primarily in the prior assumptions of variance of markers effects. To estimate the marker effects, the various statistical models available can be broadly categorized into shrinkage, variable selection, kernel and dimension reduction models or methods (Lorenz et al. 2011). The assumptions about the variance of marker effects of some of the models are highlighted in the following section.

2.3.8.1. *Ridge regression best linear unbiased prediction*

Ridge regression linear unbiased prediction (RRBLUP), also known as ridge regression is among the first and most popular method implemented in GS. It was first proposed by Whittaker et al. (2000) for MAS in biparental mapping populations. Later, Meuwissen et al. (2001) implemented this method to calculate the best linear unbiased predictor estimates of all markers simultaneously considering markers as random effects. This model assumes all marker effects are normally distributed with mean zero and common marker effect variance thus, shrinking marker effects towards zero (Lorenz et al. 2011), though the marker effects vary considerably from each other. Such an assumption of equal marker effect variance is unrealistic, and the shrinkage of the marker effect variance towards zero is ineffective for large markers effects, however this approach is appropriate for markers with small effects (Heffner et al. 2009; Lorenz et al. 2011).

2.3.8.2. *Bayesian methods*

The shortcoming of the equal marker effect and common variance assumed in RRBLUP is overcome with the implementation of models to relax these assumptions using Bayesian regression models (Hayes 2007). In Bayesian estimations, marker variances are treated more realistically and variance for each

marker can differ as it is estimated separately for each marker, and the variances are assumed to follow a specified prior distribution (Meuwissen et al. 2001). The assumptions of the Bayesian approach support our knowledge about the chromosome containing QTLs and their effects, where some segments contain large effect QTLs, some small effect QTLs, and some with no QTL (Hayes 2007). Various Bayesian models are available for GS, each differing in assumption about the prior distribution of variances of marker effect.

Here, the assumptions of the Bayesian models implemented in this thesis are described. Three Bayesian models, BayesA and BayesB, first proposed by Meuwissen et al. (2001), and BayesCπ (Habier et al. 2011) were used for the GS study in this thesis. BayesA assumes that all markers have non-zero effects (π=0) and the prior distribution of a marker effect is assumed to be normal with a mean of zero and a marker specific variance (Habier et al. 2011) allowing variable degrees of shrinkage of each marker towards zero. The variance associated with the effect of each marker is in turn sampled from a scaled inverse chi-square distribution (Gianola et al. 2009; Lorenz et al. 2011; Habier et al. 2011). BayesB assumes some markers to have zero effects and variances, while other marker effects are greater than zero with a scaled inverse chi-square distribution of their variances (Habier et al. 2011). BayesCπ differs from BayesA and BayesB in consideration about π, where BayesCπ assumes π to be unknown and inferred from the data (Habier et al. 2011). The prior assumptions (priors) used in BayesA have scaled-t distribution (Meuwissen et al. 2001; Pérez and de Los Campos 2014), while the mixture of two priors in BayesB and BayesCπ can have scaled-t distribution in BayesB (Meuwissen et al. 2001) or Gaussian in BayesCπ (Habier et al. 2011).

2.3.9. *Genomic selection in forage breeding*

Considerable work in GS has been conducted in animal breeding (Hayes and Goddard 2010) after being first proposed by Meuwissen et al. (2001). The potential of this approach in plants is discussed by Heffner et al. (2009; Jannink et al. (2010); and Varshney et al. (2017). The potential and issues related to GS in forage crop breeding have been reviewed by Hayes et al. (2013), Resende et al. (2014) and Talukder and Saha (2017). Recent advancement in NGS technology has enabled sequencing of multiple individuals for

marker discovery and genotyping at reduced cost. With this, the breeding programs for perennial forage crops can benefit from the application of GS. A few studies using GS has been carried out in forage crop species. Prediction accuracies for morphological and forage quality traits in switchgrass ranged from -0.08 to 0.52 with similar prediction accuracies among the three genomic selection models RRBLUP, Elastic net, and least absolute shrinkage and selection operator (LASSO) (Lipka et al. 2014). The prediction accuracy for the total biomass yield was reported to range from 0.21 to 0.66 in a genomic selection study in alfalfa (Li et al. 2015). Prediction accuracy ranged from 0.32 to 0.35 for biomass yield in alfalfa in two different reference populations (Annicchiarico et al. 2015). Moderate and similar prediction accuracies in a range of 0.3 to 0.4 were reported for GS models RRBLUP, BayesB and Bayesian Lasso for alfalfa forage quality traits (Biazzi et al. 2017). A genomic prediction study in alfalfa for agronomic and quality traits reported accuracy to range from 0.0021 to 0.6485 (Jia et al. 2018). Prediction ability of GS models ranged from 0.07 to 0.43 for herbage accumulation and 0.40 to 0.52 for days to heading in perennial ryegrass (Faville et al. 2018).

2.3.10. *Factors affecting prediction ability of genomic selection models*

For a successful application of GS, the fundamental aspects of population and quantitative genetics theory, and resource allocation and cost benefit analysis play a significant role. Simulation and empirical studies have highlighted that the accuracy of prediction in GS is influenced by factors related to effective population size (N_e) of the training population and the marker density as suggested by the extent of LD (Calus et al. 2008; Heffner et al. 2011a; Lorenz et al. 2011; Combs and Bernardo 2013); the size, composition and phenotyping accuracy of the training population (Heffner et al. 2011a; Lorenz et al. 2011; Combs and Bernardo 2013; de los Campos et al. 2015); the heritability and the genetic architecture of the trait of interest (Lorenz et al. 2011; Combs and Bernardo 2013); and the data analysis process (Daetwyler et al. 2012).

3. Research component 1: Genotyping-by-Sequencing Enhances Genetic Diversity Analysis of Crested Wheatgrass [*Agropyron cristatum* (L.) Gaertn.]

This chapter has been published in the International Journal of Molecular Sciences.

Baral, K., Coulman, B., Biligetu, B., and Fu, Y.-B. 2018. Genotyping-by-Sequencing Enhances Genetic Diversity Analysis of Crested Wheatgrass [*Agropyron cristatum* (L.) Gaertn.]. Int. J. Mol. Sci. 19(9): 2587 doi:10.3390/ijms19092587.

The sequence data cleaning, formatting, marker discovery and subsequent analysis such as Analysis of molecular variance, principal coordinate analysis, neighbor-joining analysis, UPGMA clustering was carried out by Kiran Baral. Structure analysis was completed with the help of Dr. Yong-Bi Fu. The manuscript was prepared by Kiran Baral. Assessment of genetic variation is prerequisite for the development of new varieties. Genetic diversity assessment at individual level help to identify and include diverse genotypes in a breeding program. Discripancy between genotypes is more accurately and easily determined at reduced cost and time with the use of genetic markers. This help towards accelerated breeding through early detection of discripency in populations of crop as crested wheatgrass where genotypes are morphologically identical though geneticlly diverse.

3.1. Abstract:

Molecular characterization of un-sequenced plant species with complex genomes is now possible by genotyping-by-sequencing (GBS) using recent next generation sequencing technologies. This study represents the first use of GBS application to sample genome-wide variants of crested wheatgrass [*Agropyron cristatum* (L.) Gaertn.] and assess the genetic diversity present in 192 genotypes from 12 tetraploid lines. Bioinformatic analysis identified 45,507 single nucleotide polymorphism (SNP) markers in this outcrossing grass species. The model-based Bayesian analysis revealed four major clusters of the samples assayed. The diversity analysis revealed 15.8% of SNP variation residing among the 12 lines, and 12.1% SNP variation present among four genetic clusters identified by the Bayesian analysis. The principal

coordinates analysis and dendrogram were able to distinguish four lines of Asian origin from Canadian cultivars and breeding lines. These results serve as a valuable resource for understanding genetic variability, and will aid in the genetic improvement of this outcrossing polyploid grass species for forage production. These findings illustrate the potential of GBS application in the characterization of non-model polyploid plants with complex genomes.

Keywords: genotyping-by-sequencing; Agropyron; genetic diversity; genetic structure; SNP

3.2. Introduction

Genotyping-by-sequencing (GBS) is a powerful genomic approach for identification of genetic variation on a genome-wide scale for genetic diversity analysis of non-model plants ((Fu and Peterson 2011; Peterson et al. 2012; Peterson et al. 2014). This approach produces high-density, low-cost genotypic information without the requirement for a reference genome sequence (Poland and Rife 2012b). The detailed GBS approach in plant diversity analysis is described in Peterson et al. (2014). In brief, the GBS analysis involves five major steps: (1) genome complexity reduction with restriction enzyme; (2) barcoding the seared genomic DNAs with indexed adaptors; (3) high-throughput sequencing of barcoded DNA fragments; (4) identification of genetic variants through a bioinformatics analysis of de-multiplexed reads; and (5) a genetic diversity analysis of sequenced samples based on sample-by-variant matrix. The GBS application, despite being a powerful approach, has certain limitations, including many missing data points, uneven genome coverage, complex bioinformatics, and issues related to polyploidy (Poland et al. 2012; Huang et al. 2014; Fu et al. 2016; Fu and Yang 2017). To overcome these limitations, a GBS-based pipeline, called Haplotag, was developed by Tinker et al. (2016), which can generate tag-level haplotype and single nucleotide polymorphism (SNP) data for polyploid organisms. This approach has been successfully applied in the study of diploid and polyploid genomes in oat (*Avena sativa*) (Yan et al. 2016; Bekele et al. 2018; Al-Hajaj et al. 2018) and genetic diversity analysis of northern wheatgrass (*Elymus lanceolatus* ssp. *Lanceolatus*) (Li et al. 2018).

Crested wheatgrass [crested wheatgrass; *Agropyron cristatum* (L.) Gaertn.] is one of the perennial species of the genus *Agropyron* that comprises 10–15 species in a polyploid series of diploid (2n = 2x = 14), tetraploid (2n = 4x = 28) and hexaploid (2n = 6x = 42) forms with the P genome (Dewey, 1984; Asay et al. 1992). *Agropyron* species are native to temperate-frigid grassland and sandy soils of Eurasia (Rogler and Lorenz 1983; Dewey 1984; Chen et al. 2013), and were first introduced to Canada in 1911 (Rogler and Lorenz 1983). Crested wheatgrass is the most important commercial species of the crested wheatgrass complex in Canadian grasslands (Mellish et al. 2002). It is characterized by an extensive root system, making it drought tolerant and winter hardy. crested wheatgrass is considered an important pasture grass for early spring grazing, providing highly palatable and nutritious forage (Looman and Heinrichs 1973). This species is easy to establish, has strong competitive ability, tolerates insect predation, provides high forage yield, and can be managed for multiple harvests in a season (Looman and Heinrichs 1973; Rogler and Lorenz 1983; Asay and Jensen 1996). It performs well on marginal lands and semi-desert environments to moist moderately saline soils (Looman and Heinrichs 1973; Asay and Jensen 1996). Due to these features, this species can be used for land reclamation of abandoned croplands, burnt and degraded areas, as well as in erosion control (Zlatnik 1999). It has persisted as a high yielding species compared to native forage species, even in 20- to 40-year-old pastures, despite heavy grazing and trampling (Hull and Klomp 1966; Looman and Heinrichs 1973). In addition, crested wheatgrass is also known to possess traits of interest, including disease resistance, tolerance to abiotic stress, and high yield, which have been utilized in wheat and barley breeding (Sharma et al. 1984; Dong et al. 1992; Wu et al. 2006; Ochoa et al. 2015; Zhang et al. 2015a). The palatability and nutrient content of crested wheatgrass declines after anthesis, and it becomes less desirable for summer grazing (Looman and Heinrichs 1973). Thus, a goal of present crested wheatgrass breeding programs is to develop later maturing cultivars that would maintain nutritive value into the summer grazing season. Development of high forage-quality, late-maturing crested wheatgrass cultivars is limited by the relatively long varietal development process, few studies to assess genetic variability of the germplasm, and lack of an

effective marker system for marker-assisted and/or genomic selection/breeding. Recent RNA-seq studies in crested wheatgrass have identified flowering time related genes and flowering related differentially expressed genes (Zeng et al. 2017b, 2017a). This emphasizes the need for genetic diversity studies of crested wheatgrass for the management and utilization of proper genetic resources in a breeding program as exogamous perennial forage species are often morphologically comparable, though they are genetically highly heterogeneous and heterozygous (Forster et al. 2001; Che et al. 2008). An adequate level of genetic diversity is crucial for both germplasm adaptation and the long-term sustainability of plant communities (Rogers and Montalvo 2004).

Attempts have been made to assess genetic variability within and among the genus *Agropyron* using molecular markers like amplified fragment length polymorphism (AFLP) (Mellish et al. 2002) and simple sequence repeat (SSR) markers (Che et al. 2008, 2011, 2015). The revealed variabilities have allowed for better understanding of the extent of diversity present in the genus. However, these marker systems are unable to provide high resolution of genetic diversity and population structure information to understand the ancestry and microevolution of the populations. Research is needed to assess molecular characteristics of crested wheatgrass for plant breeding. The molecular characterization is now more feasible than before with the advanced sequencing technology and reduced cost to acquire informative markers such as SNPs in non-model polyploid crested wheatgrass plants. Recent GBS studies in polyploid plants (Yan et al. 2016; Li et al. 2018) demonstrate the likelihood that GBS will unveil genetic variability on a genome-wide scale in crested wheatgrass plants, and characterize crested wheatgrass germplasm for breeding and genetic research.

This study was conducted with the objective to apply GBS in combination with the Universal Network Enabled Analysis Kit (UNEAK) and the Haplotag pipelines to (1) identify genome-wide SNP markers; (2) assess the genetic diversity present in 12 lines of *A. cristatum*; and (3) assess whether the GBS application is useful in the genetic diversity analysis of complex polyploid plants.

3.3. Materials and Methods

3.3.1. *Plant Materials*

The study material comprised 12 tetraploid crested wheatgrass lines consisting of six breeding lines, three cultivars, and three genebank accessions (Table 3.1). These accessions were acquired from USDA-ARS plant germplasm system, Plant Gene Resources of Canada (PGRC), and the joint forage breeding program of the University of Saskatchewan and Agriculture and Agri-Food Canada (AAFC). For ease of interpretation, all the acquired material will be referred to as lines, rather than accessions, in this study. Seeds of each line were grown for six weeks in the greenhouse at the Saskatoon Research and Development Centre, AAFC, under the following growth conditions: 16 h photoperiod at 22 °C and 8 h dark at 16 °C. Young leaf tissues were collected from 16 randomly selected plants for each of the lines and stored at −80 °C prior to DNA extraction. A total of 192 genotypes from the 12 tetraploid lines, listed in Table 3.1, were used for bioinformatics and genetic diversity analyses.

3.3.2. *Genotyping-by-Sequencing*

For each of the 192 genotypes, DNA was extracted from 0.1 g finely ground tissue following the protocols of NucleoSpin® Plant II Kit (Macherey-Nagel, Bethlehem, PA, USA), and was eluted in a 1.5 mL Eppendorf tube with Elution Buffer. NanoDrop 8000 (Thermo Fisher Scientific, Waltham, MT, USA) was used to measure the quality of the DNA by comparing the 260 and 280 nm absorptions. DNA samples were further quantified through the Quant-iTTM PicoGreen® dsDNA assay kit (Invitrogen, Carisbad, CA, USA) and diluted to 60 ng/μL with 1× TE buffer prior to sequencing analysis.

A genetic diversity-focused GBS (gd-GBS) protocol by Peterson et al. (2014) was used for the preparation of multiplexed GBS libraries. In brief, for each library, 200 ng purified genomic DNA was first digested with the restriction enzyme combination *Pst*I and *Msp*I (New England Biolabs, Whitby, ON, Canada). Ligation of pair of enzyme-specific adapters onto the 5′ and 3′ ends of the restriction fragments by T4 ligase was subsequently carried out to all samples universally. Then, the ligation fragments were

purified by an AMPure XP kit (Beckman Coulter, Brea, CA, USA). Following the purification, Illumina TruSeq HT multiplexing primers specific to each adapter consisting of Illumina index sequence and flow cell annealing complementary sequences were added through PCR amplification. The amplicon fragments were further quantified, concentrated, and pooled to form 4 subgroups of 12 samples each. The samples in the subgroups were pre-selected using a Pippin Prep instrument (Sage Science, Beverly, MA, USA) for an insert size range of 250–450 bp, before pooling the samples into a library. Each pooled library was diluted to 6 pM, and denatured with 5% of sequencing-ready Illumina PhiX Library Control (Illumina, San Diego, CA, USA) that can serve for calibration. Sequencing was completed using an Illumina MiSeq Instrument with paired-ends of 250 bp in length. MiSeq runs generated 384 FASTQ sequence files from 192 genotypes of 12 lines (one forward and one reverse for each of 192 genotypes). All the raw pair-end sequencing data in FASTQ format were deposited into the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) with accession number SRP115373 as part of the larger sequencing effort to enhance crested wheatgrass breeding (Li et al. 2017). The sequencing information for all 192 assayed samples is described in the online Supplementary Material, Section A.

3.3.3. *Bioinformatics Analysis*

Bioinformatic analysis began with sequence (FASTQ) data cleaning, using Trimmomatic version 0.36 (Bolger et al. 2014) to remove any sequenced-through Illumina adapters, low quality sequence (sliding window of 10 bases, average Phred of 20), and fragments under 64 bases long.

As the UNEAK-GBS pipeline (Lu et al. 2013) only considers sequences of 64 bp (after barcode removal) with an intact 5-base *Pst*I residue (TGCAG) at the beginning, each FASTQ file of 250 bp was first split into three fragment sets with a custom Perl script *fastq184CutandCode-Pst.pl*. The first set comprised the first 64 bases with the *Pst*I residual restriction site, and the next two sets each with 59 base portions and an added 5-base *Pst*I residue. The script also provided an arbitrary barcode sequence (CATCAT) at the start of each sequence fragment, since the UNEAK pipeline expects to deconvolute barcoded sequence reads which are

not already separated by sample. The three 70-base-long fragments formed, thereafter, were independent, as their relationship was not preserved. Each fragment set was recognized by the UNEAK-GBS pipeline (Lu et al. 2013), and was passed into UNEAK as an independent dataset.

Each fragment set (70 bases long) thus formed was analyzed with UNEAK and the Haplotag pipelines (Tinker et al. 2016), resulting in the analysis of a total of 177 bases (3 sets with 59 bases each after removal of arbitrary sequence and 5 base *Pst*I residue) of genetic sequence. Online Supplementary Material, Section B, describes the procedures to run UNEAK. Two types of meta data files—a single mergedAll.txt (all tags observed more than 10 times) and a set of individual tagCount files (one per sample) needed for the Haplotag pipeline—were generated from the UNEAK run.

Haplotag was run with the parameters and filtering threshold settings described in the HTinput.txt file, and generated a matrix of samples by SNP loci (online Supplementary Material, Section B). A set of tag-level haplotypes ("HTgenos") are first generated by Haplotag, followed by a set of SNP data derived from these haplotypes ("HTSNPgenos"). These two data types are technically redundant, so choosing one of them relies on the implementation and preference of software. In the present study, most (97.5%) haplotypes were found to contain only a single SNP; thus, we decided to analyze the SNP dataset for simplicity and compatibility with downstream analysis software.

The character by Taxa (CbyT) program supplied by N. Tinker was used to generate a filtered SNP file. In brief, Haplotag generated three separate "HTSNPGenos" files, which were merged before running CbyT. The "minimum presence" value in CbyT was set to 80%, 70%, 60%, and 50% for 20%, 30%, 40%, and 50% missing data, respectively. A SNP-by-sample matrix in the output files was used in further analyses. Additional descriptions of the SNP data matrix and the custom Perl and Shell scripts are available in the online Supplementary Material, Section A. Analyses from FASTQ file separation to SNP generation were conducted using Microsoft Windows 7 64-bit OS with an Intel (R) Xeon (R) CPU E5-2623 v3 @ 3.00 GHz (8 threads) and 32 GB RAM.

3.3.4. *Genetic Diversity Analysis*

The diversity analysis was based on 45,507 SNP markers, with 50% or less missing values in 192 genotypes from 12 crested wheatgrass lines. Data analysis began with calculation of the minor allele frequency and the extent of missing SNP data with Microsoft Excel®. Thereafter, diversity analyses at the individual and line levels were carried out.

Three types of diversity analysis were performed at individual genotype level. First, genetic structure of 192 crested wheatgrass genotypes was examined using a model-based Bayesian method implemented in the program STRUCTURE version 2.2.3 (Pritchard et al. 2000; Falush et al. 2007). Linux server with 60 core parallel computing was used to run the STRUCTURE program, where each population subgroup (K = 1–9) was run 20 times, using an admixture model with 10,000 replicates each for burn-in and during the analysis. Based on (1) a plot of likelihood of these models, (2) the rate of change in the second derivative (ΔK) between successive K values (Evanno et al. 2005), and (3) the consistency of group configuration across 20 runs, the final population subgroups were determined. For a given population subgroup (K) with 20 runs, the run having the highest likelihood value was chosen to assign the posterior membership coefficients to each sample. These posterior membership coefficients were used to create a graphical bar plot. The size and formation of each optimal cluster with respect to population were evaluated. Second, a neighbor-joining (NJ) analysis of the 192 genotypes was conducted using MEGA version 7.0.14 (Kumar et al. 2016) based on the dissimilarity matrix obtained from R routine AveDissR (Yang and Fu 2017; R development core team 2018), and a radiation tree was displayed. Third, a PCoA of all 192 genotypes was also done using the R routine AveDissR (Yang and Fu 2017; R development core team 2018) to assess genetic distinctness and redundancy, and to assess the genotype associations, plots of the first two resulting principal components were generated. For comparison, the resulting NJ trees and PCoA plots were individually labeled for the inferred structures.

Genetic variation present among the 12 lines was evaluated with AMOVA using Arlequin version 3.5 (Excoffier and Lischer 2010) on 45,507 markers. In addition, the pairwise genetic distances were computed and line-specific F_{st} values (inbreeding coefficient) for each line (Weir and Hill 2002) were generated to infer the reduction in heterozygosity. To inspect the genetic variation among the clusters identified from the STRUCTURE analysis, additional AMOVA was performed. Unweighted pair group method, with arithmetic mean (UPGMA) dendrogram based on pairwise genetic distances among the 12 lines obtained from AMOVA, were generated using MEGA version 7.0.14 (Kumar et al. 2016), to evaluate line differentiation and distinctness.

To estimate the influence of missing SNP data on the genetic diversity analysis, four datasets of 272; 1884; 10,738; and 45,507 SNPs representing 20%, 30%, 40%, and 50% of missing SNPs (M20%, M30%, M40%, and M50%) were attained for the 192 genotypes, respectively. For each dataset, the among-line variance from AMOVA and the optimal number of genetic clusters from STRUCTURE were obtained and compared among the four datasets of varying percentages of missing data.

3.4. Results

3.4.1. *SNP Discovery and Characterization*

The Miseq run of 192 genotypes from 12 crested wheatgrass lines (Table S1) generated approximately 87.8 million raw forward (R1) sequence reads of 250 bp. The number of raw forward sequence reads per sample ranged from 190,606 to 775,160 with an average of 457,279. Combined UNEAK and Haplotag analysis at the 20%, 30%, 40%, and 50% level of missing data generated 227; 1,884; 10,738; and 45,507 SNPs, respectively across the 192 genotypes. In addition, this analysis also generated many metagenomic files associated with the SNP discovery, which are described and accessible in the online Supplementary Materials. The distribution of the minor allele frequency in 45,507 SNPs' data ranged from 0.025 to 0.5, and exhibited a steady decline of minor alleles with increased occurrence of frequencies from 0.075 to 0.5 (Figure 3.1A). Likewise, there were more SNPs at the higher percentages of missing data (Figure 3.1B).

Table 3.1 List of the 12 crested wheatgrass (*A. cristatum*) lines used in the study.

Lines	CN Number [a]	Alternative Identification [a]	Origin	Type
Kirk	CN108662	PI 536010	Canada	Cultivar
AC-Goliath	CN108673		Canada	Cultivar
NewKirk		FOR552	Canada	Cultivar
Vysokij 9	CN30995	PI 370654	Siberia, Former Soviet Union, Omsk region	Genebank line
Karabalykskij 202	CN31068	PI 326204	Kazakhstan, Former Soviet Union, Kustanai region	Genebank line
PGR 16830	CN43478		Kazakhstan	Genebank line
S8959E		FOR917	Siberia/Canada	Breeding line
S9491		S9491	Canada	Breeding line
S9514		S9514	Canada	Breeding line
S9516		S9516	Canada	Breeding line
S9544		S9544	Canada	Breeding line
S9556		S9556	Canada	Breeding line

[a] CN number is the line identification in Plant Gene Resources of Canada, Agriculture, and Agri-Food Canada (AAFC), while the alternative identifications, including FOR or S, are from the joint forage breeding program of the University of Saskatchewan and AAFC, and PI is from plant inventory book, National Germplasm Resources Laboratory, USA.

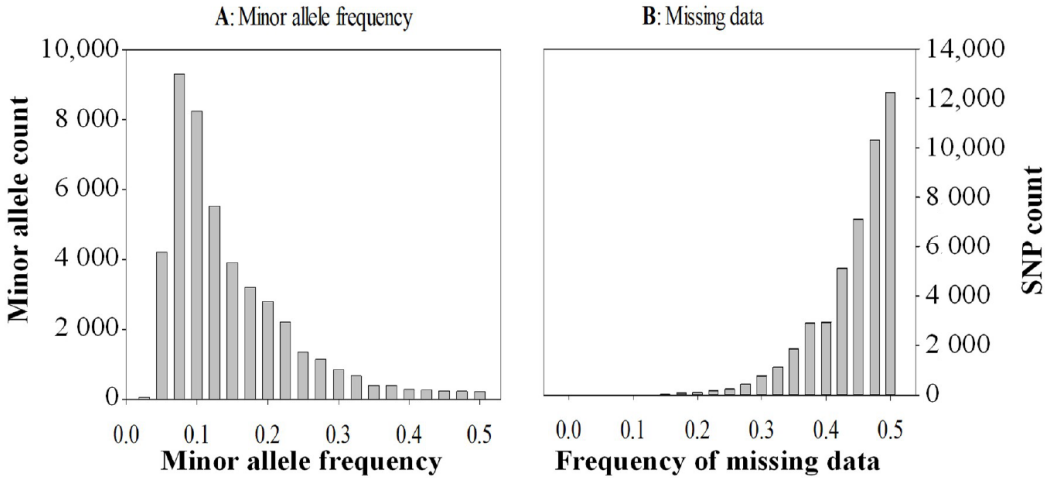

Figure 3.1 The minor allele frequency distribution (A) and the frequency of missing data (B) for 45,507 SNP markers in 192 genotypes of 12 crested wheatgrass lines.

3.4.2. *Genetic Structure and Relationship*

The genetic structure estimated for 192 genotypes from 12 crested wheatgrass lines without consideration of prior population information in the STRUCRURE analysis revealed four optimal clusters (Figure 3.2A) with strong support from change in LnP(K) variance (Figure 3.2B) and the largest delta K

value (Figure 3.2C). Cluster 1 (red in color) consisted of 17 genotypes (16 from Vysokij 9 and one from S8959E). Cluster 2 (green in color) had 22 genotypes (16 from S9491 and 6 from S9514). Cluster 3 (blue in color) was the largest cluster, with 95 genotypes from seven lines. Cluster 4 (yellow in color), with 58 genotypes from five lines, was the second largest cluster. The neighbor-joining (NJ) tree was in agreement with clusters obtained from the STRUCTURE analysis (Figure 3.3). However, there existed some discrepancies, as some members of cluster 4 (yellow in color) were spread into cluster 2 (green in color) and cluster 3 (blue in color).

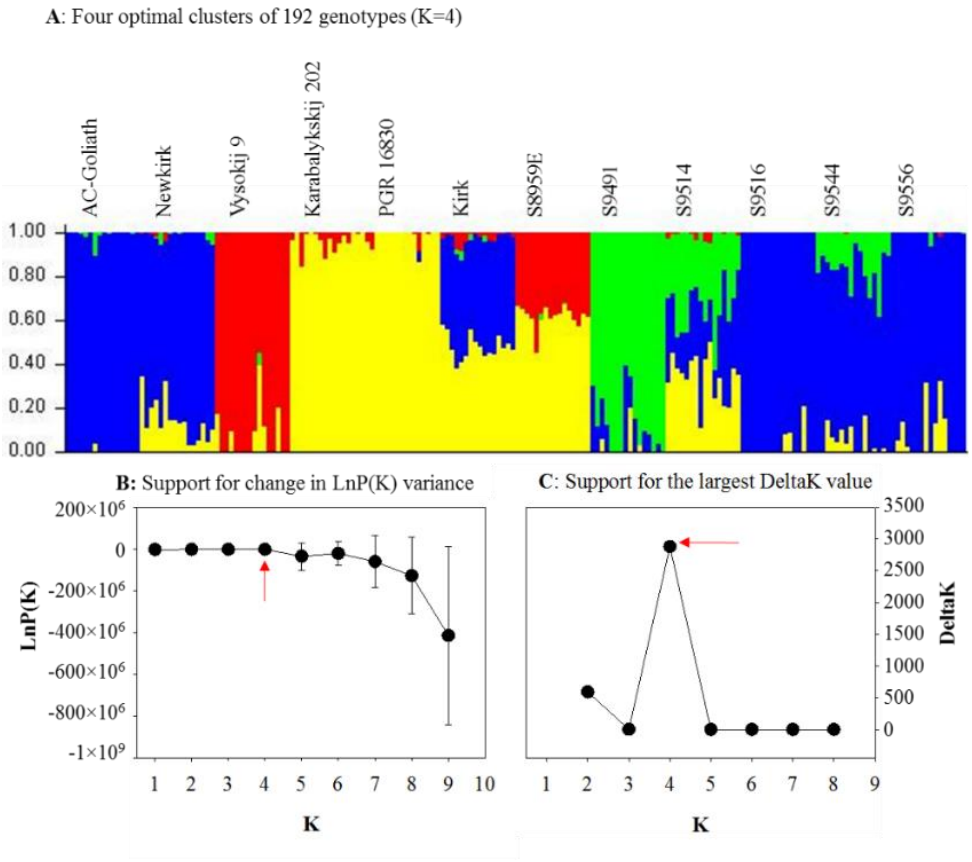

Figure 3.2 Four genetic clusters of 192 genotypes of the 12 crested wheatgrass lines inferred by STRUCTURE based on 45,507 SNP markers. (A) The mixture coefficients of 192 genotypes with K = 4, presented in the original order of genotypes from 12 lines (see Table 3.1 for line label); (B) support from the LnP(K) estimation; (C) support from the estimation of the largest value of the delta K = mean. (|Ln″(K)|)/sd (LnP(K)).

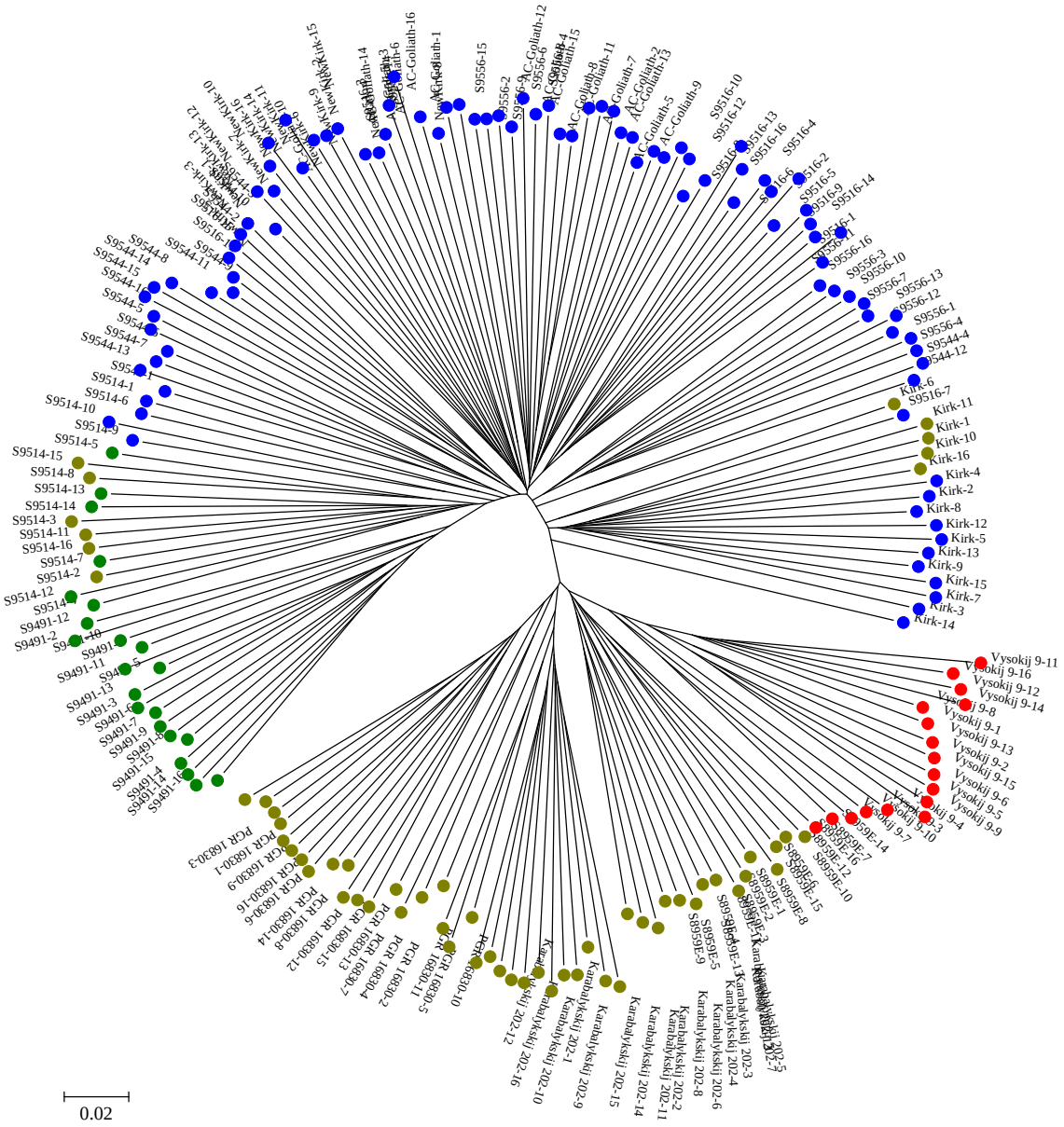

Figure 3.3 Genetic relationship of 192 genotypes of the 12 crested wheatgrass lines as revealed by neighbor-joining clustering with the 45,507 SNP markers. Each genotype is numbered after its line label. Each node for a genotype is represented with colored circle followed by genotype name. Red, green, blue, and yellow represent plants in Clusters 1, 2, 3, and 4, inferred from the STRUCTURE analysis (Figure 3.2A), respectively.

The principal coordinates analysis (PCoA) revealed that the genetic relationship of 192 genotypes (Figure 3.4A) was not in accordance to the Bayesian inferences from the STRUCTURE analysis. The clusters II, III, and IV identified by the Bayesian inferences appeared to overlap and became undistinguishable with PCoA. However, the PCoA plot was able to distinguish four lines Karabalykskij 202 (from Kazakhstan), PGR 16830 (from Kazakhstan), Vysokij 9 (from Russia) and S8959E (selected from Vysokij 9) from the rest of the lines (Figure 3.4B). We also observed lines S9516, S944 and S9556 from cluster 3 (blue in color from the model-based Bayesian analysis) were more dispersed than other breeding lines and cultivars, likely indicating the larger genetic diversity present in those breeding lines (Figure 3.4B).

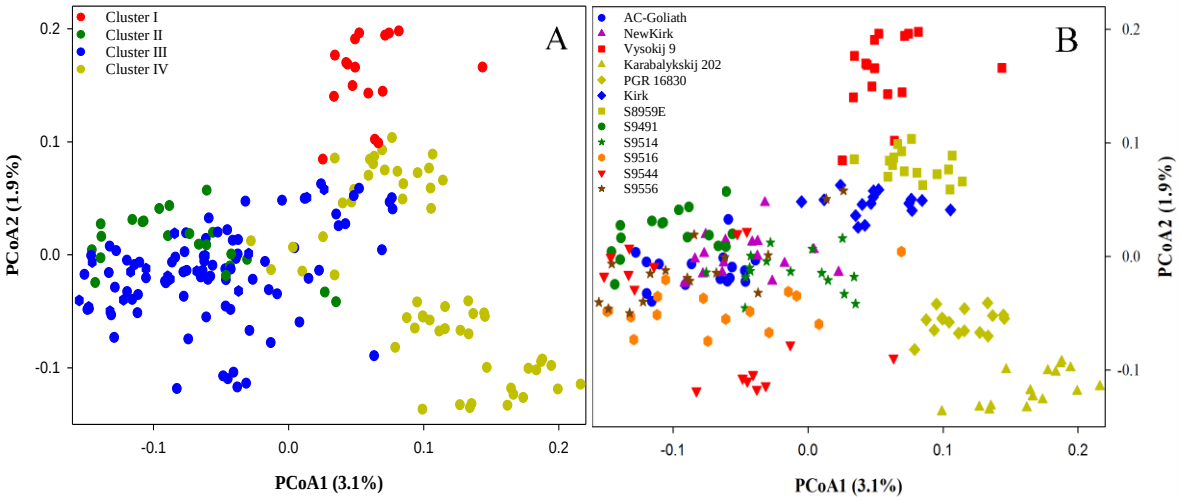

Figure 3.4 Genetic relationship of 192 genotypes of the 12 crested wheatgrass lines as revealed by principal coordinates analysis (PCoA) with the 45,507 SNP markers. Two panels are identical, but in the left panel (A) each genotype is labelled with colored circles representing the clusters obtained from the STRUCTURE analysis, while the right panel (B) labels genotypes for 12 lines.

3.4.3. *Genetic Differentiation*

The analysis of molecular variance (AMOVA) revealed that most of the SNP variations were present within the lines (84.2%), while much smaller variations reside among lines (15.8%) or among the four Bayesian clusters (12.07%) (Table 3.2). Line-specific F_{st} was also estimated from AMOVA for each line

as the weighted variation among individual plants within a line to observe the extent of inbreeding. They were obtained in the range of 0.56 (in line S9491) to 0.64 (in the cultivar Kirk) with mean of 0.60 (Figure 3.5B). The pairwise genetic distance among the 12 lines ranged from 0.055 (between AC-Goliath and S9544) to 0.32 (between Karabalykskij 202 and S9491) with an average distance of 0.15.

Table 3.2 Results of the analysis of molecular variance for two models of genetic structure (12 lines and four clusters from the STRUCTURE analysis) based on 45,507 SNP markers.

Model/Source of Variation	df	Sum of Squares	Variance Explained	Variance (%)[a]
12 lines				
Among lines	11	101,048.8	246.0	15.8
Within lines	372	488,598.0	1313.4	84.2
Four clusters from STRUCTURE				
Among clusters	3	54,736.5	193.3	12.1
Within clusters	380	534,910.3	1407.7	87.9

[a] These variances were statistically significant from zero at $P < 0.0001$.

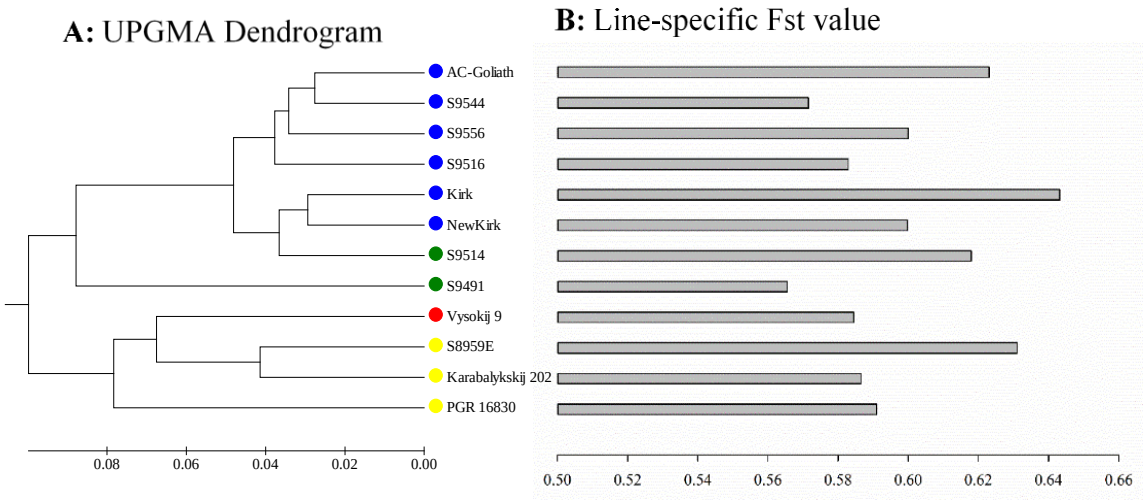

Figure 3.5 Genetic diversity and genetic relationships of the 12 crested wheatgrass lines. Left panel (A) shows their genetic relationship in the unweighted pair group method with arithmetic mean (UPGMA) dendrogram based on the Phi statistics obtained from the AMOVA. The right panel (B) displays the line-specific F_{st} values for the 12 lines.

The dendrogram based on AMOVA showed the grouping of the 12 crested wheatgrass lines into three genetically distinct clusters at the Phi statistic of 0.08 or more (Figure 3.5A). The dendrogram grouped the lines from Kazakhstan and Russia in one distinct cluster. The second distinct cluster consisted of the single line S9491. The largest of all is the third cluster, with seven lines consisting of cultivars and breeding lines from Canada.

3.4.4. *Effects of Missing Data on Diversity Analysis*

The optimal numbers of genetic clusters inferred from STRUCTURE analyses with respect to the extent of missing data from M20%, M30%, M40%, and M50% datasets provided 4, 6, 6, and 4 optimal clusters, respectively (Figure 3.6A). Comparing the proportions of SNP variance residing among the 12 lines inferred from the AMOVA analysis showed 24.6%, 20.3%, 17.8%, and 15.8% for M20%, M30%, M40%, and M50%, respectively (Figure 3.6B).

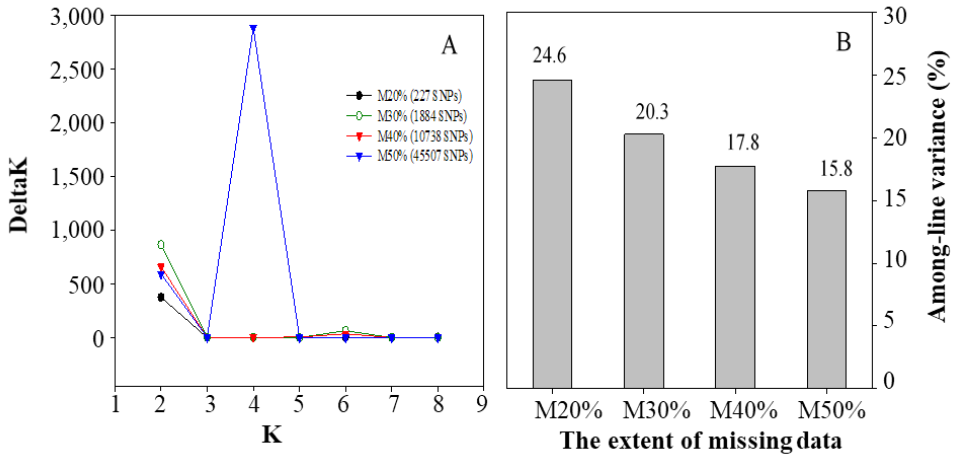

Figure 3.6 The impact of missing SNP data on the inferences of STRUCTURE and AMOVA analysis. The left panel (A) shows the four optimal clusters obtained from the STRUCTURE analyses at the missing level of M20% and M50%, and six clusters at M30% and M40%. The right panel (B) shows the SNP variances, ranging from 24.6 to 15.78%, inferred from AMOVA analyses residing among 12 lines at the increasing level of missing values from M20% to M50%, respectively.

3.5. Discussion

This study utilized the gd-GBS application, in combination with Haplotag pipeline, for the first time in crested wheatgrass, to generate a data matrix of 192 genotypes × 45,507 SNP markers, and captured genome-wide genetic variants to evaluate the genetic diversity present in tetraploid crested wheatgrass. The diversity analysis revealed 15.8% of SNP variation residing among the 12 lines and the model-based Bayesian analysis identified four major clusters of the assayed samples. These research outputs are not only useful for understanding the genetic diversity of crested wheatgrass and for its breeding, but also are encouraging for molecular characterization of non-model polyploid plants.

The revealed patterns of genetic diversity are interesting. First, the model-based Bayesian approach in the STRUCTURE identified four major clusters of the assayed genotypes, while the distance-based approaches like PCoA and UPGMA identified three major clusters; however, the neighbor-joining analysis was in accordance with the result from STRUCTURE analysis. Following the pedigree of the assayed genotypes (Table S1), we could infer that the model-based Bayesian analysis and neighbor-joining analysis were able to genetically infer population substructure—an outcome of probable processes such as genetic drift, migration, mutation, and selection—more distinctly than distance-based approaches. Results also showed most of the genotypes grouped together within their lines, revealing that different lines were distinct. The STRUCTURE analysis (Figure 3.2A), neighbor-joining analysis (Figure 3.3), PCoA (Figure 3.4B), and UPGMA dendrogram (Figure 3.5A) revealed the genetic distinctness of lines Karabalykskij 202, PGR 16830, S8959E, and Vysokij 9. S8959E is a breeding line in the Saskatoon program, but it is a selection from Russian gene bank line Vysokij 9. Although it has been recurrently selected for vigorous growth and plant type, it has not been inter-pollinated with any other lines, explaining its distinctness from other Canadian cultivars/breeding lines. However, STRUCTURE revealed all genotypes, except one (S8959E-14; Figure 3.2A) from line S8959E, showing high affinity with the line from Kazakhstan. This is also supported by UPGMA clustering (Figure 3.5A), while neighbor-joining analysis revealed the relatedness of lines from

Russia. These findings will serve as valuable information for the genetic improvement of crested wheatgrass for forage production.

Our analysis showed high within-line genetic variation (Table 3.2) of assayed crested wheatgrass lines, which is in agreement with studies on highly outcrossing species (Hamrick and Godt 1989). Overall, our genetic diversity results are in accordance with diversity studies of crested wheatgrass reported by Mellish et al. (2002) using AFLP markers and (Che et al. 2008) and (Che et al. 2011, 2015) using SSR markers. The somewhat higher among population variation (15.8%) observed in the present study may partly be due to narrower genetic base of eight of the breeding lines/cultivars relative to the three gene bank lines and one line of Russian origin (S8959E). Most of the Canadian cultivars and breeding lines shared one or more common parents in their genetic background (Table S1), and they have gone through many cycles of recurrent selection for vigor and yield. Thus, there has probably been a slight reduction in heterozygosity as indicated by the generally higher inbreeding coefficients (Figure 3.5B). The distinctness of the lines S8959E, Vysokij 9, Karabalykskij 202, and PGR 16830 can be attributed to their Asian origin and absence of inter-pollination with Canadian cultivars/lines and selection under Canadian conditions, except for the recurrent selection of line S8959E, mentioned above. Thus, the cultivars/breeding lines likely have reduced the within-line variation, while diverging more from the unselected Asian lines, explaining some increase of the among-line variation. Further research is needed on the utilization of the genetic variability of these lines with focus on morpho-physiological studies, adaptation, and their utilization in breeding programs. Likewise, the distinctness of the line S9491 in the UPGMA analysis (Figure 3.5A) is attributed to its synthesis from seven different lines/cultivars from breeding programs in Saskatoon and Logan, Utah, USA. The line S9514 was directly selected from S9491, which explains why these two lines clustered (green cluster) together in the STRUCTURE analysis (Figure 3.2) and neighbor-joining analysis (Figure 3.3). However, the Canadian cultivar "Kirk" developed partly from a plant introduction from a botanical garden in Finland (University of Turku) in 1968 showed shared pedigree with some or all of the Kazakhstan lines

based on model-based Bayesian clustering (Figure 3.2A) and neighbor-joining analysis (Figure 3.3). While the origin of the plant introduction from the University of Turku remains unknown, it can be reasoned that this original introduction may have common genetic background with some of the Kazakhstan lines based on Bayesian clustering.

It was observed that the extent of reduction in heterozygosity, as explained by F_{st}, was more in cultivars than most of the breeding lines. Two cultivars "AC-Goliath" and "Kirk" had lower diversity as indicated by higher inbreeding coefficient (F_{st} values) (Figure 3.5B), perhaps because of being synthesized from the inter-pollination of fewer genotype than many of the breeding lines. Also, most of the breeding lines included cultivars "Kirk", "AC-Goliath", and other sources, in their pedigrees. The cultivar "Newkirk" was selected from progenies of crosses between "Kirk" and "AC-Goliath". However, the inbreeding coefficient of "Newkirk" was lower than the parental cultivars, indicating a higher level of heterozygosity. The three breeding lines S9516, S9544, and S9556 showed high within-line genetic diversity according to greater dispersal of these lines on PCoA (Figure 3.4B), higher within line variation (92.2%) as explained by a separate AMOVA (Data not shown), and lower line-specific F_{st} (Figure 3.5B). This greater genetic diversity could be attributed to inclusion of diverse germplasm sources during their synthesis (Table S1). The high within-line variability suggests that there is sufficient genetic variation in all lines in this study to make progress from selection. Inclusion of germplasm from the Asian lines in the breeding program to inter-pollinate with Canadian cultivars/breeding lines will increase diversity.

Results from the assessment of the impact of missing data on genetic diversity exhibited some discrepancies (Figure 3.6). However, the discrepancies were not too large to be unacceptable. With increasing the threshold of missing data, up to 50% it was possible to include more loci for the assignment of individuals to the clusters. The clustering with SNPs at 50% missing data followed the pedigree of the assayed samples (Table S1). The robustness of the result to the level of missing data selected indicates that uncertainty introduced by the missing data is offset by the increasing number of data points.

Our gd-GBS application has identified thousands of genome-wide SNP markers to assess the extent of genetic diversity in the non-model polyploid crested wheatgrass with no prior genomic information. These results demonstrated the technical feasibility and effectiveness of GBS to sample genome-wide genetic variability in other perennial grass species with complex genomes. High resolution plant genetic diversity analysis, with 45,000 SNP markers spread over a genome, is more informative than with relatively few markers, like AFLP and SSR used in previous studies (Hamrick and Godt 1989; Pritchard et al. 2000; Mellish et al. 2002; Fu and Peterson 2011; Al-Hajaj et al. 2018). Also, the experimental cost for sampling genome-wide variants in this study was roughly $12,000, suggesting the feasibility of a wider application of GBS to characterize other perennial polyploid grass species. The results of the present study, along with those published in northern wheatgrass and wild oat (Li et al. 2018; Al-Hajaj et al. 2018), demonstrate the utility of GBS in molecular characterization of non-model plants with complex ploidy and genetic structures.

3.6. Conclusions

With the application of GBS, it has been possible to generate 45,507 SNP markers for a diversity analysis of crested wheatgrass. The variation residing among these 12 lines of crested wheatgrass was found to be 15.8%. Further analysis grouped the assayed samples into four genetic clusters, and revealed the genetic distinctness of two cultivars each from Kazakhstan and Russia, respectively. These results can enhance parental selection for increased genetic variation and improved offspring performance in crested wheatgrass breeding. The findings in this study can also aid in the application of GBS in the characterization of non-model plants with complex genomes.

Supplementary Materials: Supplementary materials can be found at http://www.mdpi.com/1422-0067/19/9/2587/s1.

Chapter Connecting Statement

Generation of genome-wide SNP markers in crested wheatgrass is feasible with the GBS application. These markers have shown potential in explaining the variation present within and among the genotypes of crested wheatgrass lines. The present study shows the SNP markers discovered with the GBS application successfully clustered crested wheatgrass genotypes based on their origin and pedigree and also provided information on the possible origin of cultivar "Kirk". However, the use of these markers for Marker Assisted Selection (MAS) is yet to be determined. In the following chapter SNP markers generated using the GBS application will be studied for their potential in linkage mapping in an intraspecific F_1 mapping population of crested wheatgrass.

4. Research component 2: Development of linkage maps of crested wheatgrass [*Agropyron cristatum* (L.) Gaertn.] using genotyping-by-sequencing

4.1. Abstract:

Crested wheatgrass *[Agropyron cristatum (L.)* Gaertn.] is an important species in seeded forage fields in the Great Plains region of Canada providing high quality forage for early grazing during spring. Development and use of genetic markers to construct a high density-linkage map for understanding the genetic potential of crested wheatgrass has been limited by the large size of its genome and polyploidy. Genotyping-by-sequencing (GBS), which does not require a reference genome to generate single nucleotide polymorphisms (SNP) markers, generated 86,172 SNPs in bi-parental F1 mapping populations derived from an intraspecific cross of two diploid outcrossing cultivars of *A. cristatum* "Fairway" and "Parkway". Genetic maps were constructed using 678 SNP markers distributed among seven linkage groups spanning 1259.76 cM. The average distance between adjacent markers was 1.85 cM. This is the first intraspecific genetic map of crested wheatgrass using SNPs developed from the GBS approach.

Keywords: Genotyping-by-sequencing, crested wheatgrass, mapping population, linkage map

---

4.2. Introduction

Crested wheatgrass [*Agropyron cristatum* (L.) Gaertn.], an important perennial forage crop native to temperate-frigid grasslands and sandy soils of Eurasia (Rogler and Lorenz 1983; Dewey 1984), is a common seeded forage species in Canada. It consists of diploid (2n = 2x = 14), tetraploid (2n = 4x = 28) and hexaploid (2n = 6x = 42) cytotypes with the P genome (Dewey 1984; Asay and Jensen 1996) with an average DNA content (2C) of 14.12, 28.02 and 39.48 pg, respectively (Tandoh 2019). Said et al. (2018) reported the genome size (1C) of crested wheatgrass to be about 6 Gbp. Crested wheatgrass is an important component of Canadian livestock feeds for its highly nutritious and palatable forage, and high productivity, outperforming native forage grasses. Crested wheatgrass serves as an excellent source of forage in early spring before the native forage species start growing, exhibits excellent winter hardiness,

drought tolerance, and resistance to trampling (Looman and Heinrichs 1973). This perennial cool-season grass has remained productive even in stands of more than 50 years in age (Hull and Klomp 1966; Looman and Heinrichs 1973). It is important to understand the genetics of crested wheatgrass as the P genome has been reported to harbor many superior traits such as disease resistance, tolerance to abiotic stress and high yield (Sharma et al. 1984; Dong et al. 1992; Ochoa et al. 2015; Zhang et al. 2015a). The productivity and quality of crested wheatgrass declines rapidly following inflorescence emergence (heading) and this species becomes less desirable for summer grazing as livestock avoid consuming it. Later-heading cultivars are desirable to maintain the quality and palatability of crested wheatgrass pastures into the summer period. Detailed information on the composition and structure of the crested wheatgrass genome would aid plant breeding efforts. The development and use of genetic markers for the construction of a high density-linkage map covering the entire genome will permit the mapping of genes and quantitative trait loci (QTL) through identification of closely linked markers in the periphery of loci contributing to phenotypic variation of important traits in crested wheatgrass. However, the large size of the P genome and polyploidy in crested wheatgrass precludes the development of an effective marker system that facilitates in developing a linkage map dense enough for marker assisted selection (MAS).

Next-generation-sequencing permits genotyping-by-sequencing (GBS), an approach for generating dense marker coverage in experimental populations without a reference genome (Fu and Peterson 2011; Peterson et al. 2012; Peterson et al. 2014; Poland and Rife 2012). The GBS approach is primarily a genome reduction approach and has been described in detail by Peterson et al. (2014). In brief, the GBS approach is delivered following five major steps: (1) reduced genome complexity using restriction enzymes; (2) ligation of indexed adaptors for barcoding the seared fragments; (3) sequencing of barcoded fragments; (4) genetic variant identification with bioinformatics analysis; and (5) subsequent downstream analysis for linkage mapping with sample-by-variant matrix. The downstream analysis is

complicated following GBS, which generates large numbers of missing data, provides uneven coverage of the genome, involves complex bioinformatics, and has issues related to polyploidy (Poland et al. 2012; Huang et al. 2014; Fu and Yang 2017).

Construction of high-density linkage maps with the use of molecular markers spanning the entire genome and subsequent fine mapping of quantitative trait loci (QTLs) are crucial for dissecting molecular basis of trait variation, MAS, comparative genomic studies and genome sequence assembly (Chen et al. 2014; Wang et al. 2015). Genotyping-by-Sequencing is capable of producing thousands of SNP markers for developing linkage maps in previously unstudied species and cultivated crops whether they are diploids, polyploids or heterozygous outcrossing species (Li et al. 2014; Su et al. 2017; Vining et al. 2017; Hussain et al. 2017; Goonetilleke et al. 2018). Molecular markers in crested wheatgrass have mainly been used to study genetic diversity and genetic characterization (Mellish et al. 2002; Che et al. 2008, 2011, 2015; Baral et al. 2018) and for the identification of flowering time related and differentially expressed genes (Zeng et al. 2017a, 2017b). In the past few years, marker development and linkage mapping in *Agropyron* species has resulted in three genetic linkage maps in populations produced from wide crosses. Yu et al. (2012) identified 152 AFLP and 23 RAPD markers on seven linkage groups, developing the first linkage map of *Agropyron*. Zhang et al. (2015b) developed the second linkage map in *Agopyron* with 1023 SNP markers generated from specific-locus amplified fragment sequencing (SLAF-seq) distributed over seven linkage groups. More recently, Zhou et al. (2018) developed a linkage map for crested wheatgrass with 913 SNP markers in seven linkage groups using the wheat 660K SNP array. These three linkage maps developed were based on mapping populations derived from interspecific cross between diploid *A. cristatum* L. (Gaertn.) and *A. mongolicum* Keng. High-throughput SNP markers discovery utilizing next-generation sequencing and GBS has potential to develop high resolution and comparable genetic linkage maps among mapping population in a cost-effective way (Huang et al. 2009). Such maps coupled with precise mapping of QTLs will facilitate in the detection of candidate genes governing morphological,

phenological and nutritive value traits, in crested wheatgrass with unknown genetic architect, thus will accelerate its breeding.

The present study attempts to sequence diploid crested wheatgrass with no prior sequence information generating SNPs for developing linkage maps in bi-parental F_1 mapping populations derived from intraspecific crosses of *A. cristatum*. This differs from the recent linkage maps produced for crested wheatgrass in that it uses a population generated from crosses of elite crested wheatgrass cultivars, rather than interspecific crosses. This will provide a framework for further marker assisted breeding efforts and quantitative analysis of complex traits in crested wheatgrass. This experiment was developed from the hypothesis that GBS will generate large number of SNP markers in biparental F_1 mapping populations of crested wheatgrass facilitating the develoment of linkage map.

4.3. Materials and Methods

4.3.1. *Crested wheatgrass germplasm and genetic stocks*

Two full-sib F_1 mapping population developed by reciprocal crossing of two diploid cultivars of crested wheatgrass "Fairway" and "Parkway" was used for marker development and genetic map construction. The diploid cultivars "Fairway" and "Parkway" were developed by Agriculture and Agri-food Canada, Saskatoon, Saskatchewan and released in 1932 and 1969, respectively (Elliott and Bolton 1970). The cultivar "Parkway" was selected from "Fairway" (Asay 1986). Individual plants of each cultivar were randomly selected to be used as parents to create the mapping populations. As crested wheatgrass is an outcrossing species, the parental plants were assumed to be heterozygous and genetically different. Reciprocal cross involving one "Fairway" and one "Parkway" plant was made in a green house in 2015 to create two mapping populations, with each parent serving as the female parent of one population and male parent of the other population. The seeds from the two crosses were harvested, cleaned and stored separately prior to their seeding. Seeds of each cross were germinated in germination tray and later planted in six-inch pots. They were grown for six weeks in the greenhouse at the Saskatoon

Research and Development Centre, AAFC, under a 16 h photoperiod at 22 °C and an 8 h dark period at 16 °C. Young leaf tissues were collected from 94 randomly selected progeny plants from each mapping population and the two parents, and stored at −80 °C prior to DNA extraction. A total of 190 genotypes from the mapping populations were used for bioinformatics and linkage mapping.

4.3.2. *DNA isolation and library construction*

This study took advantage of a reduced representation library protocol involving a two-enzyme double digest to scan regions of the crested wheatgrass genome from a biparental population and generate large number of SNPs utilizing a GBS approach outlined in Peterson et al. (2014). Briefly, parental and F_1 progeny total DNA was extracted from 0.1 g finely ground leaf tissue following the protocol of NucleoSpin® Plant II Kit (Macherey-Nagel, Bethlehem, PA, USA) and was eluted in a 1.5-ml Eppendorf tube with Elution Buffer. The DNA quality was measured with a NanoDrop 8000 (Thermo Fisher Scientific, Waltham, MT, USA) by comparing the 260 and 280-nm absorption. DNA samples were further quantified through the Quant-iTTM PicoGreen® dsDNA assay kit (Invitrogen, Carisbad, CA, USA). Subsequently, DNA was diluted to 50 ng/μL, with1×TE buffer prior to sequencing analysis.

For each library, 200 ng of purified genomic DNA was double-digested with a combination of restriction enzyme i) *HinfI* and *HpyCH4IV* (New England Biolabs, Whitby, ON, Canada) in the year 2016 and ii) *Pst*I and *Msp*I (New England Biolabs, Whitby, ON, Canada) in the year 2017 thus creating two sets of libraries for each enzyme combinations. Enzyme specific adapters containing specific priming site for Illumina Hiseq chemistry were ligated to 5′ and 3′ ends of the restriction fragments universally. The ligated reaction mix was cleaned with Agencourt AMPure XP Beads (Beckman Coulter, Brea, CA, USA) to remove unligated adapters. Following the cleaning, PCR amplification of the fragments with adapter specific indexed primers (consisting of Illumina index sequence and flow cell annealing complementary sequences) was completed. Four amplicons were pooled and concentrated using Zymo Research (Irvine, CA, USA) DNA clean and concentrator-5 kit. 10 μL each of three samples were again pooled prior to size

selection with an electrophoresis instrument Pippin Prep (Sage Science, Beverly, MA, USA) for fragment size of 100-400 bp before pooling the samples into a library. Each pooled library was diluted to 6 pM and denatured with 5% of sequencing-ready Illumina PhiX Library Control (Illumina, San Diego, CA, USA) that can serve for calibration. Sequencing was completed using an Illumina Hiseq2500. Libraries for each set of enzyme combinations were prepared separately. Illumina Hiseq2500 runs generated 380 FASTQ sequence files from 190 genotypes in the F_1 mapping population (one forward and one reverse for each of 190 genotypes).

4.3.3. *Bioinformatics analysis*

DNA sequences in FASTQ data were cleaned with Trimmomatic v0.36 (Bolger et al. 2014) to remove any sequenced-through Illumina adapters, low quality sequence (sliding window of 10 bases, average Phred of 20), and fragments under 40 bases long.

Cleaned FASTQ data files were analyzed to obtain unique sequences with Fastx_collapser (Gordon 2010) followed by *de novo* assembly of contigs using Minia (Salikhov et al. 2014). Minia was run with a k-mer size of 47, the minimum k-mer abundance of 6 and genome size of 1,000,000,000. This generated twelve contig files in each mapping population, with each enzyme combination. The twelve contig files generated were merged with Minia (with minimum k-mer abundance of 9) to create a consensus contig file which served as a reference to call the SNP markers in each mapping population. Also, all 24 contig files from both the mapping populations were merged with Minia (with minimum k-mer abundance of 18) to create another consensus contig file. This will be referred to as combined contig file here after to differentiate it from consensus contig created from the above method and the merged contig file generated in the following section and used as a reference for SNP calling in each mapping population. This was performed for each enzyme combination. Likewise, Fastq files and combined contig files obtained with *HinfI* and *HpyCH4IV,* and *Pst*I and *Msp*I enzyme combinations were merged to create merged data. Bowtie2 was used to merge the combined contig files obtained from both enzyme

combinations (*HinfI* and *HpyCH4IV*, and *Pst*I and *Msp*I) with the removal of duplicates to get a merged contig files which served as a reference for SNP calling with the merged data.

Another bioinformatics pipeline, Analysis of Next Generation Sequencing Data (ANGSD)(Korneliussen et al. 2014) that employs a probabilistic approach by using genotype likelihood calculations was used for more accurate SNP genotyping from the DNA sequences with low genome coverage in diploid species. In this study, SNP calling with ANGSD utilized consensus contig files as reference for each mapping population in each enzyme combination while calling for individual enzyme combinations. Likewise, combined contig file was the reference for both mapping populations in each enzyme combination while SNP calling in combined calls. Whereas, for SNP calling in merged data, merged contig file was the reference. SNP data were generated for each progeny genotype and the parents. The SNP marker information for the missing markers was reconstructed using probabilistic PCA, a PCA-based imputation method, using the freely available R package "pcaMethods" (Stacklies et al. 2007; Fu 2014).

The mapping population was grouped into two sets based on the parents. The first mapping population had Parkway as the female parent while Fairway was the female parent for second mapping population. Based on the genotype of parents, SNP allele frequency and expected segregation patterns were used to filter the marker data. The null hypothesis for a chi-square goodness-of-fit test for markers heterozygous in only one parent was a 1:1 segregation of homozygotes to heterozygotes, while for markers heterozygous in both the parents the segregation ratio was 1:2:1 of homozygote, heterozygote, and homozygote, respectively. Thus, the markers were grouped into two parental marker datasets for linkage mapping.

4.3.4. *Mapping code assignment*

The SNP markers were coded following the coding scheme of the CP population type of JoinMap 4.1 (Van Ooijen 2011). Three kinds of segregation types were involved, including markers that were heterozygous only in the female (lmxll), only in the male (nnxnp), or in both parents (hkxhk). JoinMap 4.1 determined the phase of markers in the population.

4.3.5. *Linkage map construction*

Joinmap 4.1 (Van Ooijen 2011) with population option CP (cross pollinating full-sib population) was used for map construction. The multipoint maximum likelihood model implemented in JoinMap 4.1 facilitates the generation of an integrated map from a segregating F_1 mapping population generated by crossing two outcrossing parents. The marker type heterozygous in both parents (hk × hk) serves as the anchor for map integration. Markers were grouped according to the independence of logarithm of odds (LOD) in each mapping population with an LOD threshold of 2–25 increasing with an increment of LOD 1 and maximum recombination frequency at 0.40. Seven linkage groups were chosen from the Grouping (tree) tab of population node. Marker ordering and map distance calculation were done using Maximum Likelihood algorithm (ML) and the Haldane function respectively. Removal of the markers with the highest nearest-neighbor stress (N.N. Stress) followed by re-ordering of linkage group was repeated until the markers with a N.N. Stress level greater than 3 cM were removed. Markers with more relaxed thresholds on segregation distortion were added to each map. Each population's final map was created using the final map node from each framework map's linkage groups. Within a grouping node, markers were added to the linkage group as follows: Ungrouped markers were assigned using SCL threshold 10 LOD from the grouping node SCL worksheet. From the grouping node SCL worksheet all "excluded" markers (i.e markers with node= excluded) were moved back to Group 0. This step was repeated using progressively increasing SCL threshold steps of 15 LOD, 20 LOD and 25 LOD. Then, markers within each seven linkage groups were ordered using ML algorithm, as well as markers in each linkage group from

the framework map as fixed orders. The Haldane function was used for map distance calculation. Markers having high N.N. Stress (cM) values (>4) were excluded from the final map. Marker ordering and exclusion were repeated until N.N. Stress values were below 4. Linkage maps were drawn using Mapchart (Voorrips 2002).

4.4. Results

4.4.1. *SNP markers from genotyping-by-sequencing*

The bioinformatic pipeline ANGSD produced 96904, 116115, 102145, 122084, 88703, 108312, and 196,098 contigs, respectively for Fairway mapping population with restriction enzyme combination *HinfI* and *HpyCH4IV*, and *Pst*I and *Msp*I; Parkway mapping population with restriction enzyme combination *HinfI* and *HpyCH4IV*, and *Pst*I and *Msp*I; Fairway and parkway mapping population combined for enzyme combination *HinfI* and *HpyCH4IV*, and *Pst*I and *Msp*I; and contig files with enzymes *HinfI* and *HpyCH4IV* and contig files with *Pst*I and *Msp*I merged (merged data). SNP calling based on these contigs as reference produced SNPs in a range of 30,135 to 88,373 (Table 4.1). The monomorphic and incorrectly genotyped markers were removed prior to mapping. The summary of the total number of SNP, SNPs heterozygous in female, male and both parents, final number of SNPs used for linkage mapping and total number of SNPs grouped to seven linkage groups are provided in Table 4.1. The number of SNPs generated were less with the use of *Pst*I and *Msp*I enzyme combination (Table 4.1). There was no improvement in SNP generation by using the consensus contigs from both the mapping populations with enzyme combination *Pst*I and *Msp*I whereas, combined contig generated around fifteen thousand more SNPs than calling individually with enzyme combination *HinfI* and *HpyCH4IV* (Table 4.1).

4.4.2. *Component maps*

Mapping population and linkage maps were named after the female parents Parkway and Fairway. Ten component maps were constructed which included two Parkway and two Fairway maps with consensus contig files for each enzyme combinations (*HinfI* and *HpyCH4IV,* and *Pst*I and *Msp*I) as reference in each mapping population (4 maps); two Parkway and two Fairway maps with combined contig files from each enzyme combinations (*HinfI* and *HpyCH4IV,* and *Pst*I and *Msp*I) as reference (4 maps) and a Parkway-merged and a Fairway-merged map with merged contig as reference (2 maps) (Table 4.1). The mapping population Parkway with enzyme combination *Pst*I and *Msp*I and consensus contig files as reference for SNP calling generated 34,691 SNPs however, had the lowest number of filtered markers used for mapping (Table 4.1). It had 424 SNP markers which included 238 SNPs heterozygous for the female parent (Parkway), 106 SNPs heterozygous in male parents (Fairway) and 80 SNPs heterozygous in both parents (Table 4.1). The Faiway-merged (*HinfI* and *HpyCH4IV,* and *Pst*I and *Msp*I) population that used merged contig as reference for SNP calling produced 79,432 SNPs of which 3,428 SNPs were obtained for mapping after filtering. Of these, only 678 SNPs markers were mapped to seven linkage group (Figure 4.1, Table 4.1 and Table 4.2).

4.4.3. *Linkage groups*

Marker distribution into seven linkage groups, map distance for each linkage group and total map length were obtained for Parkway and Fairway (*HinfI* and *HpyCH4IV,* and *Pst*I and *Msp*I) with consensus contig as reference for each population and enzyme combination; Parkway combined and Fairway combined (*HinfI* and *HpyCH4IV,* and *Pst*I and *Msp*I) with combined contigs as reference in each population and enzyme combinations; and Parkway-merged and Fairway-merged (*HinfI* and *HpyCH4IV,* and *Pst*I and *Msp*I) with merged contigs as reference (Table 4.2). Of these, the linkage map Fairway merged had the highest number of SNPs (678) distributed into seven linkage groups. The first linkage

Table 4.1 Information on SNPs obtained and mapped to seven linkage groups of crested wheatgrass

Female parent (Enzymes)	Number of SNPs	SNPs heterozygous in female parent lm x ll	Heterozygous in female parent (1:1)	SNPs heterozygous in male parent nn x np	Heterozygous in male parent (1:1)	SNPs heterozygous in female and male parents hk x hk	Heterozygous in both parents (1:2:1)	SNPs used for Mapping	SNPs grouped to linkage groups
Fairway (*HinfI* and *HpyCH4IV*)	55,793	3,511	1,078	3,080	1,007	19,313	234	2,319	212
Fairway (*Pst*I and *Msp*I)	30,135	604	125	947	376	10,594	23	524	135
Fairway combined (*HinfI* and *HpyCH4IV*)	70,112	3,991	1,393	3,324	1,135	21,197	206	2734	330
Fairway combined (*Pst*I and *Msp*I)	38,931	776	263	2,637	852	11,726	33	1148	430
Faiway-merged (*HinfI* and *HpyCH4IV* **&** *Pst*I and *Msp*I**)**	**86,172**	**5,210**	**1,589**	**5,686**	**1,664**	**29,779**	**223**	**3,428**	**678**
Parkway (*HinfI* and *HpyCH4IV*)	52,772	3,251	1,147	3,886	1,073	18,800	193	2,413	222
Parkway (*Pst*I and *Msp*I)	36,157	862	238	419	106	12,563	80	424	147
Parkway combined (*HinfI* and *HpyCH4IV)*	70,112	3,324	947	3,991	1083	21197	84	2,114	238
Parkway combined (*Pst*I and *Msp*I)	38,931	2,637	841	776	192	11,726	36	1069	286
Parkway-merged (*HinfI* and *HpyCH4IV* & *Pst*I and *Msp*I)	88,373	2,965	606	3,035	646	32,550	346	1,598	139

Table 4.2 Information on distribution of SNP markers into seven linkage groups and the map distance in crested wheatgrass

Female parent/Enzymes	LG1	LG2	LG3	LG4	LG5	LG6	LG7	Map distance cM
Fairway *HinfI* and *HpyCH4IV* (SNPs)	81	33	37	12	15	22	21	
Map length (cM)	400.1	232.45	201.67	136.11	129.44	158.71	104.38	1362.86
Fairway *Pst*I and *Msp*I (SNPs)	24	22	24	11	38	5	11	
Map length (cM)	129.23	74.32	46.01	36.29	91.90	53.92	51.88	483.55
Fairway combined *HinfI* and *HpyCH4IV* (SNPs)	78	91	61	37	33	14	16	
Map length (cM)	528.71	373.02	363.23	176.60	250.60	117.69	104.34	1924.19
Fairway combined *Pst*I and *Msp*I (SNPs)	115	54	56	124	16	44	21	
Map length (cM)	222.16	138.66	143.08	254.53	59.47	228.48	74.54	1120.92
Faiway-merged (*HinfI* and *HpyCH4IV* **&** *Pst*I and *Msp*I**) (SNPs)**	**255**	**172**	**13**	**42**	**85**	**90**	**21**	**678**
Map length (cM)	**312.66**	**280.84**	**127.3**	**112.75**	**143.81**	**209.75**	**72.65**	**1259.76**
Parkway *HinfI* and *HpyCH4IV* (SNPs)	88	46	52	16	16	2	2	
Map length (cM)	559.45	256.23	306.25	120.89	168.26	1.08	43.30	1455.46
Parkway *Pst*I and *Msp*I (SNPs)	33	63	9	11	22	3	6	
Map length (cM)	207.71	165.1	37.31	32.20	99.71	37.01	61.76	640.8
Parkway combined call *HinfI* and *HpyCH4IV* (SNPs)	84	41	28	46	14	12	13	
Map length (cM)	436.13	196.56	202.11	266.83	89.61	82.06	100.50	1373.8
Parkway combined *Pst*I and *Msp*I (SNPs)	50	35	20	60	58	40	23	
Map length (cM)	147.94	79.11	54.95	137.63	194.21	127.38	79.91	821.13
Parkway-merged (*HinfI* and *HpyCH4IV* & *Pst*I and *Msp*I) (SNPs)	41	22	27	9	19	13	8	
Map length (cM)	112.41	120.08	203.86	64.55	228.08	156.99	98.96	984.93

group (LG1) in this map had 255 SNPs distributed along a map length of 312.66 cM with an average distance of 1.23 cM among the SNPs. The second linkage group (LG2) had 172 SNPs distributed along a map length of 280.84 cM with an average distance of 1.63 cM among the SNPs. The third linkage group (LG3) had 13 SNPs distributed along a map distance of 127.3 cM with an average distance of 7.49 cM among the SNPs. The fourth linkage group (LG4) with a total map distance of 112.75 cM had 42 SNPs distributed with an average distance of 2.68 cM among the SNPs. Linkage group five (LG5) had 85 SNPs distributed in a map length of 143.81 cM with an average distance of 1.69 cM among the SNPs. Linkage group six (LG6) had 90 SNPs arranged on a map length of 209.75 cM and the average distance among the markers was 2.33 cM. The seventh linkage group (LG7) had 21 SNPs distributed along a map length of 72.65 cM with an average distance of 3.46 cM among the SNPs (Figure 4.1).

4.5. Discussion

High-density genetic linkage maps are fundamental for genomic and genetic study of individual species. This study utilized the gd-GBS application, in combination with the ANGSD pipeline to generate SNP markers ranging from 30,135–88,373 in F_1 mapping populations from an intraspecific cross of heterogeneously heterozygous diploid crested wheatgrass cultivars. The genome-wide SNP markers generated were filtered and classified into three segregation patterns (lm x ll, nn x np, hk xhk) and used for subsequent analysis in linkage mapping. This linkage mapping analysis generated 10 component maps. The SNP markers in a range of 135 to 678 were arranged into seven linkage groups in these component maps. Of these, the component map Fairway-merged was densest with 678 SNP markers distributed in seven linkage groups with a total map distance of 1259.76 cM. This map had an average distance of 1.86 cM among the SNPs. These data are a good resource for future linkage mapping, genome selection and genome-wide association studies. Compared to previous linkage maps, these research outputs are useful for understanding that genetic linkage maps are possible to develop from F_1 mapping populations from intraspecific crosses in outcrossing species like crested wheatgrass. In addition, this research highlights the

potential of low cost high-throughput gd-GBS application to provide SNPs information in species lacking reference genome.

A genetic linkage map is fundamental for the studies of the genetic structure and marker-trait association (Semagn et al. 2006b). In this study, the SNP markers generated were assigned into the three segregation patterns following an outcross pollinated (CP) scheme for outcrossing species. The segregation pattern lm x ll, where the female parent is heterozygous and male parent is homozygous had 5,210 SNPs of which only 1,589 were segregating in 1:1 ratio; nn x np, where the female parent is homozygous and male parent is heterozygous had 5,686 SNPs of which 1,664 were segregating in 1:1 ratio and hk x hk, where both the male and the female parents are heterozygous had 29,779 SNPs but only 223 were segregating in 1:2:1 expected ration of segregation. Out of 79,432 SNPs generated, 40,675 SNPs were in the three segregation patterns of which 3,428 SNPs were in the expected ratio of segregation for mapping. However, the total number of mapped SNPs into seven linkage group was 678. High-throughput SNP markers are liable to genotyping errors introduced through missing genotypes, unanticipated double recombination, segregation distortion and allele switching which if not taken into consideration potentially inflate the map distance and map order (Cartwright et al. 2007). This reduction in the total numbers of markers for mapping and only 600–700 of them being mapped to seven linkage group could be accounted for the choice of parents, whereby "Parkway" is a direct selection from "Fairway". The cultivars "Fairway" and "Parkway" were found to cluster together in a study using AFLP markers (Mellish et al. 2002). The number of heterozygous markers (hk type) indicates that the parents chosen were diverse; however, the total number of heterozygous loci available for mapping was low. The choice of such closely related parents reduces the number of recombination events. Also, the choice of restriction enzymes did not offer substantial change in the numbers of markers available for mapping. This may be due to the inability of the restriction enzyme combinations to sample across the large genome size of crested wheatgrass. In addition to this, the markers discovered may not be regularly dispersed along the chromosomes and do not adequately cover the

genome (Brown 2002). A similar study in Napiergrass (*Cenchrus purpureus*) mapped 1,913 SNPs out of 287,093 SNPs identified with GBS approach to 14 linkage groups with an average distance of 0.73 cM between markers (Paudel et al. 2018). Marker density of 1.86 cM in our study is comparable to the marker density obtained for bread wheat and oil palm genetic maps developed with GBS approach crops (Pootakham et al. 2015; Hussain et al. 2017). Generally, linkage mapping assumes the two parents used to create mapping populations are distinct for certain characteristics, but the two parents used in the present study were related. This relatedness among the parents used resulted in fewer regions along the chromosomes that are heterozygous which could be used for linkage mapping. However, crested wheatgrass being an outcrossing species has loci along the chromosomes that are heterozygous permitting linkage mapping even if parents are related, as in the present study. With the inclusion of diverse parents (not closely related) we would get more heterozygous loci that would segregate in the mapping populations; thus, marker density would likely be increased. Linkage mapping in orchardgrass (*Dactylis glomerate* L.), using biparental F_1 mapping population from two diverse parents, mapped 2,510 markers into seven linkage groups spanning 715.77 cM (Zhao et al. 2016). Linkage mapping in intermediate wheatgrass with six full sib mapping population and one self-derived family resulted in the development of a consensus map with 10,029 markers distributed in 21 linkage groups covering a map distance of 5061 cM, with an average distance of 0.5 cM between markers (Kantarski et al. 2017). A similar approach as adopted in intermediate wheatgrass will generate additional markers that could be used to saturated current crested wheatgrass map.

It is of interest to compare our linkage map, the first map developed from an intraspecific cross with *A. cristatum* to genetic linkage maps developed by Zhang et al. (2015) and Zhou et al. (2018) using an interspecific cross of two diploid *Agropyron* species, one of which was *A. cristatum*. Our linkage map represents the third densest map produced in crested wheatgrass, the densest map being the one with 1023 markers with a total of 907.8 cM (Zhang et al. 2015) and, the second densest map with 913 markers spanning

839.7 cM (Zhou et al. 2018). Both of these denser maps were created from interspecific crosses of plants of the same parents *A. mongolicum* x *A. cristatum* to create an F_1 mapping population. The genetic linkage maps available in crested wheatgrass cannot be considered dense, given the large size of the genome, such that several regions from the genome may have remained undetected because of the available techniques of marker discovery (Semagn et al. 2006b). Our map would likely have been denser if our parents would have been more diverse (having more heterozygous regions); however, the interspecific cross used in the other *Agropyron* studies did not produce a genetic map substantially denser than our map and not nearly as dense as maps determined for the other grasses mentioned above. Perhaps the *Agropyron* species do not have as much genetic variation as other perennial grasses. We also attempted to combine the two parental integrated maps (Fairway and Parkway) generated in two mapping populations with each enzyme combination and merged data; however the seven linkage groups from the parental maps did not share enough markers for map integration in most of the linkage groups and we were unable to develop a consensus map. The missing data and genotyping errors which are found in GBS applications and the use of ML algorithm and Haldane mapping function is likely to result in an inflated map distance (Hackett and Broadfoot 2003).

The genetic linkage map developed from F_1 mapping population arising from intraspecific cross provides a platform for QTL fine mapping of economically important forage traits in this outcrossing species. Further studies should be performed using more informative mapping population from diverse parents which will generate large number of markers that can be used to saturate the current linkage map. In addition to this, the linkage map developed could be used for comparative studies.

4.6. Conclusions

The Fairway mapping population with merged contigs from both enzyme combinations grouped 678 SNPs into seven linkage groups with a total map distance of 1259.76 cM and represents the first linkage map in crested wheatgrass using an intraspecific F_1 as mapping population. Importantly, this is the first linkage map of crested wheatgrass utilizing SNP markers generated using the Genotyping-By-Sequencing approach. This linkage map represents the third densest map produced in crested wheatgrass.

Genotyping-by-sequencing has demonstrated that it is possible to obtain genome-wide SNP markers from an intraspecific biparental F_1 mapping population in outcrossing crested wheatgrass for linkage map development. However, the development of a denser map which could be utilized for QTL mapping and MAS requires the choice of more diverse parents to increase heterozygosity and also could likely be achieved through inclusion of more than one full-sib mapping population. Similarly, careful selection of restriction enzyme combinations and the GBS protocol may also facilitate map saturation.

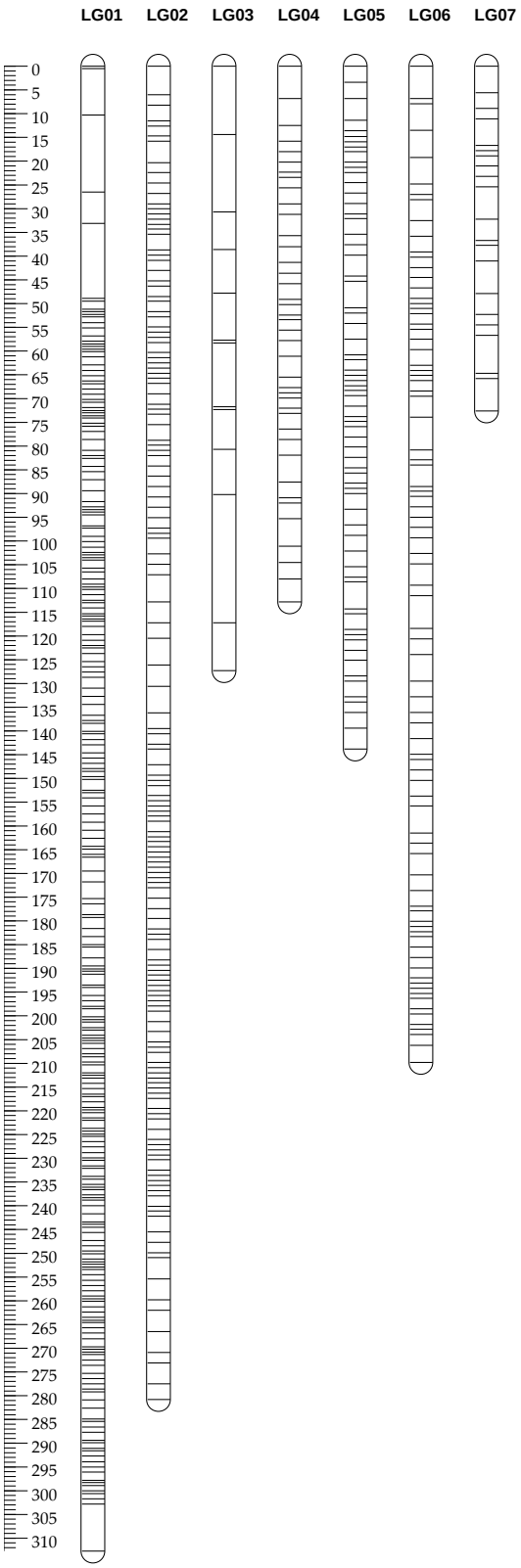

Figure 4.1. Distribution of single nucleotide polymorphism (SNP) markers in seven linkage groups of crested wheatgrass. The left scale plate represents the genetic distance (centimorgan as unit).

Chapter Connecting Statement

The current availability of low-cost high-throughput marker discovery techniques such as GBS, has offered an abundance of available polymorphic markers. Marker information of such abundance has diverse applications including genotypic discrimination, high-density linkage mapping, fine mapping of quantitative trait loci and detection of candidate genes of complex traits. QTL detection will accelerate crop breeding through marker assisted selection. However, quantitative traits that are of interest to breeders are difficult to improve through MAS owing to limitations such as small population size and lack of power to detect and estimate small effect QTLs. Concurrent reduction in cost and time for marker discovery has offered a genomic selection (GS) approach that can overcome some of the limitations of MAS, even for the crops with no prior sequence information. The following chapter reports on a project to investigate the possibility of GS application in crested wheatgrass breeding.

5. Research component 3: Accelerating breeding of crested wheatgrass [*Agropyron cristatum* (L.) Gaertn.] through genotyping-by-sequencing and genomic selection.

5.1. Abstract:

Crested wheatgrass [*Agropyron cristatum* (L.) Gaertn.] is a perennial forage species providing high quality, highly palatable forage for early grazing. The forage quality declines rapidly after heading. Its breeding is challenging because of polyploidy and outcrossing nature, requiring long breeding cycle for trait improvement, and difficulty in developing molecular markers. This has limited breeding of late maturing crested wheatgrass germplasms with adequate quality for extended summer grazing. Genomic selection (GS), a method to predict the phenotypic performance of genotypes using genome-wide markers and methods of quantitative genetics, has potential for improving traits in perennial forage crops. Genotyping-by-sequencing (GBS), using next generation sequencing, has enabled the development of whole-genome markers in non-model polyploid plants and was used in this study on crested wheatgrass. Bioinformatics analysis identified 827, 3,616, 14,090 and 46,136 single nucleotide polymorphism (SNP) markers at 20%, 30%, 40% and 50% missing marker levels. Five additive genomic selection (GS) models (Bayes A, Bayes B, Bayes Cπ, GBLUP and RRBLUP) were implemented to assess the prediction ability of models for morphological and three quality traits in crested wheatgrass. Moderate prediction ability was obtained for leaf width, plant height, clump diameter, and tillers per plant with a range of 0.26 to 0.41 for genotypes evaluated at Saskatoon. Moderate prediction ability was obtained for acid detergent fiber, fall regrowth and plant height, with a range of 0.20 to 0.43 for genotypes evaluated at Swift Current. Similar prediction ability of the five genomic models (parametric models) indicated choice of model had no effect in the prediction ability. Similar prediction ability of the models at each SNP density inferred the traits of unknown genetic architecture in crested wheatgrass might be under the influence of large number of small effect genes. However, low to moderate prediction ability observed in this study suggests the traits might not only be controlled by simple additive genes but also by polygenes or complex interactions between

genes and the environment. This study represents the first use of GBS application to sample genome-wide variants and study the ability of GS models to predict phenotypic values of morphological and three quality traits in crested wheatgrass lines consisting of diploids and tetraploids. These findings illustrate the potential of GBS application in genomic selection of non-model polyploid plants with complex genomes. A repeat of the experiment with refining field experiment and phenotyping, increasing population size, comparing GS models (parametric and non-parametric models) that account for interactions of genes, environments and genotype by environment might improve the prediction ability in crested wheatgrass.

Key words: Genotyping-by-sequencing, genomic selection, crested wheatgrass, prediction ability

---

5.2. Introduction

Crested wheatgrass [crested wheatgrass; *Agropyron cristatum* (L.) Gaertn.], found as diploid, tetraploid and hexaploid in form (Dewey, 1984; Asay et al. 1992), is an important pasture grass species in western Canadian grasslands as it displays the desirable characteristics of early spring growth, high palatability and high nutritive value. It is also valued for its drought tolerance, and winter hardiness (Looman and Heinrichs 1973). Its persistence and competitiveness have continued to provide higher yields than native forage species, even in 20 to 40-year-old pastures, irrespective of heavy grazing and trampling (Hull and Klomp 1966; Looman and Heinrichs 1973). Crested wheatgrass is also known to possess resistance to disease, tolerance to abiotic stress, and high yielding traits, which are important to, and have been utilized in, wheat and barley breeding (Sharma et al. 1984; Dong et al. 1992; Ochoa et al. 2015; Zhang et al. 2015). High palatability and nutrient content in crested wheatgrass decline rapidly post-heading. Thus, a crested wheatgrass breeding objective is to develop later maturing cultivars that would maintain yield and nutritive value into the summer grazing season. However, varietal development/improvement of crested wheatgrass is a long process through phenotypic evaluation and selection leading to slow rate of genetic gain. Several factors such as highly outcrossing nature due to self-incompatibility and prevalence of high level of genetic diversity in crested wheatgrass species limits the fixation of desirable genes. Morphological and quality traits

are influenced by complex genes and their interactions, and the environment. This makes trait evaluation difficult through current phenotypic methods. Most importantly, lack of an effective marker system for marker-assisted and/or genomic selection/breeding is a constraint. The heterogenous and heterozygous nature of the population and genetic complexity of traits limits the application of marker assisted selection. Genetic gain in traits can be improved through accelerated breeding of this perennial species by utilizing novel breeding strategies such as genomic selection (GS) that associates DNA marker variation to phenotypic variation with statistical models.

Availability of molecular markers for non-model polyploid plants like crested wheatgrass is a recent development. Next generation sequencing technologies have offered genome-wide markers for crops with no prior sequence information. Genotyping-by-sequencing (GBS) is one such powerful genomic approach for identification of genome-wide SNPs of non-model plants (Fu and Peterson 2011; Peterson et al. 2012; Peterson et al. 2014). This approach produces high-density, low-cost genotypic information without the requirement for a reference genome sequence (Poland et al. 2012). Peterson et al (2014) described a detailed GBS approach. In brief, the GBS analysis involves five major steps: (1) genome complexity reduction with restriction enzymes; (2) barcoding the seared genomic DNAs with indexed adaptors; (3) high-throughput sequencing of barcoded DNA fragments; (4) identification of genetic variants through a bioinformatics analysis of de-multiplexed reads; and (5) genomic selection application of sample-by-variant matrix. Irrespective of the robustness of the GBS application, many missing data points, uneven genome coverage, complex bioinformatics, and issues related to polyploidy limits this application (Poland et al. 2012; Huang et al. 2014; Fu and Yang 2017). These drawbacks can be overcome with a GBS-based pipeline, called Haplotag, developed by Tinker et al (2016), which can generate tag-level haplotype and single nucleotide polymorphism (SNP) data for polyploid organisms.

Current forage breeding programs rely on recurrent phenotypic selection (with or without progeny testing), alone or in combination with pedigree information, to assess the genetic merit (breeding value) of

individuals (Wilkins and Humphreys 2003; Conaghan and Casler 2011; Coulman and Jefferson 2013). Such a method alone is unable to predict the genetic merit accurately. With the advent of genetic markers, marker assisted selection (MAS), an approach that relies on small number of DNA markers and their association with quantitative trait loci (QTL), has demonstrated improved prediction accuracy compared to phenotypic selection alone (Lande and Thompson 1990). However, weak association between the markers and traits across different genetic backgrounds, and the small proportion of variation explained by the small numbers of markers used to trace major QTLs, limits its application (Meuwissen et al. 2001; Hayes and Goddard 2010). Genomic selection, proposed by Meuwissen et al (2001), overcomes some limitations of MAS. This approach utilizes all the genome-wide markers simultaneously along with phenotypic data in a training population to predict breeding value (genomic estimated breeding value, GEBV) for individuals in a test population with only the genotypic data and assumes that all QTLs will be in linkage disequilibrium (LD) with at least one marker (Meuwissen et al. 2001; Crossa et al. 2017). Simulations and empirical studies have confirmed that GS can significantly accelerate breeding programs, and improve genetic gain compared to phenotypic selection or QTL approaches utilizing fewer resources (Heslot et al. 2012; Crossa et al. 2017). Genomic selection possesses greater potential for perennial forage breeding which is a long-term endeavor. The potential of GS has been recently described for forage crop breeding and has been demonstrated in alfalfa (*Medicago sativa* L.) (Annicchiarico et al. 2015; Li et al. 2015; Jia et al. 2018), intermediate wheatgrass (*Thinopyrum intermedium* L.) (Zhang et al. 2016), switchgrass (*Panicum virgatum* L.) (Lipka et al. 2014; Ramstein et al. 2016; Fiedler et al. 2018), and perennial ryegrass (*Lolium perenne* L.) (Grinberg et al. 2016; Faville et al. 2018; Pembleton et al. 2018). Yet, utilization of available marker systems in crested wheatgrass has been limited to the study of the genetic relationship of breeding lines, ecotypes and species of crested wheatgrass (Mellish et al. 2002; Che et al. 2008, 2011, 2015; Baral et al. 2018), linkage mapping (Yu et al. 2012; Zhang et al. 2015b; Zhou et al. 2018), and identification of flowering time related and differentially expressed genes (Zeng et al. 2017a, 2017b), but not in the selection of new cultivars. This study hypothesized that combined genotypic and phenotypic data

generated for crested wheatgrass predicts breeding values more precisely than phenotypic data alone and improves selection accuracies in crested wheatgrass breeding.

This study was conducted with the objectives: (1) to apply GBS in combination with the Universal Network Enabled Analysis Kit (UNEAK) and the Haplotag pipelines to identify genome-wide SNP markers and; (2) to assess the informativeness and feasibility of GS for complex traits in diploid and tetraploid crested wheatgrass lines.

5.3. Materials and Methods

5.3.1. *Plant Materials*

The study material comprised ten lines of crested wheatgrass (five cultivars: Fairway, Kirk, AC-Goliath, AC-Parkland and NewKirk, and five breeding lines: S8959E, S9491, S9516, S9542 and S9556) (Table 5.1). These lines were made available from the forage breeding program of the University of Saskatchewan. Field experiments were conducted to evaluate the 10 lines of crested wheatgrass for various agronomic, morphological and nutritive value traits. The trials were established in randomized complete block designs with four replications and with 16 genotypes per line per replication at Saskatoon and Swift Current in 2014. Data were collected for two years (2015 and 2016). Each genotype was spaced 1m within and between rows. AC Parkland, Fairway and S9542 were diploid cultivars, while the other seven lines were tetraploids (Table 5.1). Four plants were randomly selected per replication and young leaf tissues were collected from 160 genotypes (16 randomly selected genotypes for each of the 10 lines in four replication) from Saskatoon and 160 genotypes plus an additional five genotypes (165 genotypes) from Swift Current and stored at -80°C prior to DNA extraction. A total of 325 genotypes from the 10 lines were used for bioinformatics.

Table 5.1 List of the 10 crested wheatgrass (*A. cristatum*) lines used in the study

Lines	Origin	Type	Ploidy
Fairway	Canada	Cultivar	Diploid
Kirk	Canada	Cultivar	Tetraploid
AC-Goliath	Canada	Cultivar	Tetraploid
AC-Parkland	Canada	Cultivar	Diploid
NewKirk	Canada	Cultivar	Tetraploid
S8959E	Siberia/Canada	Breeding line	Tetraploid
S9491	Canada	Breeding line	Tetraploid
S9516	Canada	Breeding line	Tetraploid
S9542	Canada	Breeding line	Diploid
S9556	Canada	Breeding line	Tetraploid

5.3.2. *Morphological traits*

The morphological and agronomic traits evaluated for 160 genotypes in Saskatoon and 165 genotypes in Swift Current included early spring vigor (ESV), days to heading (DTH), plant height (PH), leaf width (LW), clump diameter (CD), tillers per plant (TPP), dry matter yield (DM), regrowth after harvest (RGAH) and fall regrowth (FRG). Descriptions of the measurement of these morphological traits are presented in Table 5.2. Heading days were converted to growing degree days (GDD) as described by Lipak et al. (2014) as follows,

1. The first day in which GDD was recorded was the day after the first five consecutive days where the average temperature was > 0° C.
2. After this day, GDD for a day was calculated as: $[((Adj.Min + Adj.Max))/2] - 0$

   Where, *Adj.Min* is the maximum temperature between the minimum daily temperature and 0° C, and *Adj.Max* is the minimum temperature between the maximum daily temperature and 30° C
3. Cumulative GDD was calculated for each day after the first day in which GDD recording was started. This cumulative GDD was the value used when heading was reached as described in Table 5.2.

Table 5.2 Description of the measurement of morphological and agronomic traits

Traits	Trait description	Year
Early spring vigor	1=least vigorous; 5= most vigorous, scored on first week of May.	2015-2016
Days to heading	50% of stems have 50% emerged panicles.	2015-2016
Plant height (cm)	Height measured from base of the stem to tip of the panicle.	2015-2016
Leaf width (mm)	Widest part of penultimate leaf.	2015-2016
Clump diameter (cm)	Measured on clump after harvest.	2015-2016
Tillers per plant	Number of tillers in each genotype.	2015-2016
Dry matter yield (gm)	Each genotype harvested were dried for 48h at 60° C in a forced air oven and weighed, harvesting done on last week of July.	2015-2016
Regrowth after harvest	1=least vigorous; 5= most vigorous, scored on last week of August.	2015-2016
Fall regrowth	1=least vigorous; 5= most vigorous, scored on first week of October.	2015-2016

5.3.3. *Nutritive value traits*

The genotypes were sub-sampled after dry matter determination during the growing seasons of 2015 and 2016 for forage nutritive value determination. The sub-samples were ground through a 1-mm screen Wiley mill (Thomas-Wiley, Philadelphia, PA). The ground samples were stored in plastic bags prior to determination of crude protein (CP), neutral detergent fiber (NDF) and acid detergent fiber (ADF). Nitrogen concentration was determined by the Dumas combustion method using the Leco CN 628 Dumas analyzer (Leco Corporation, St. Joseph, MI) for all the Saskatoon samples and 2015 Swift Current samples while, the Kjeldahl method was used for samples from Swift Current in the year 2016. Then, CP was calculated as crude protein = nitrogen concentration x 0.625 x 1000. Neutral detergent fiber and ADF concentrations were analyzed using an automated Ankom2000 fiber analyzer (ANKOM Technology Corporation, New York).

5.3.4. *Genotyping-by-Sequencing*

For each of the 325 genotypes, protocols of NucleoSpin® Plant II Kit (Macherey-Nagel, Bethlehem, PA, USA) were used to extract DNA from 0.1 g finely ground tissue and was eluted in a 1.5 mL Eppendorf tube with Elution Buffer. DNA quality was measured by comparing the absorptions at 260 and 280 nm using NanoDrop 8000 (Thermo Fisher Scientific, Waltham, MT, USA). Further quantification of the DNA samples

was through the Quant-iTTM PicoGreen® dsDNA assay kit (Invitrogen, Carisbad, CA, USA) and final dilution to 60 ng/µl with 1× TE buffer was done before sequencing analysis.

A genetic diversity-focused GBS (gd-GBS) protocol was used for the preparation of multiplexed GBS libraries following the protocol developed by Peterson et al (2014) and as described in Baral et al. (2018). Briefly, restriction enzyme combinations *Pst*I and *Msp*I (New England Biolabs, Whitby, ON, Canada) digested 200 ng of purified genomic DNA in each library. On to the 5′ and 3′ ends of the restriction fragments, ligation of enzyme-specific adapters consisting of Illumina index sequence and flow cell annealing complementary sequences by T4 ligase was carried out for all samples. Then, AMPure XP kit (Beckman Coulter, Brea, CA, USA) was used for purification of the ligated fragments. Through PCR amplification, Illumina TruSeq HT multiplexing primers specific to adapters were added following the purification. The amplicon fragments were further quantified, concentrated, and pooled to form 4 subgroups of 12 samples each. Using a Pippin Prep instrument (Sage Science, Beverly, MA, USA), pre-selection of the samples in the subgroups for an insert size range of 100–400 bp were done before pooling the samples into a library. Each pooled library was diluted to 6 pM, and denatured with 5% of sequencing-ready Illumina PhiX Library Control (Illumina, San Diego, CA, USA) that can serve for calibration. Sequencing was completed using an Illumina Hiseq2500 Instrument with paired-ends of 125 bp in length. HiSeq runs generated 672 FASTQ sequence files from 336 genotypes (including randomly selected 11 technical replicates) of 10 lines (one forward and one reverse for each of 336 genotypes).

5.3.5. *Bioinformatics Analysis*

Bioinformatic analysis began with sequence (FASTQ) data cleaning, using Trimmomatic version 0.36 (Bolger et al. 2014) to remove any sequenced-through Illumina adapters, low quality sequences (sliding window of 10 bases, average Phred of 20), and fragments under 64 bases long. As the UNEAK-GBS pipeline (Lu et al. 2013) only considers sequences of 64 bp (after barcode removal) with an intact 5-base *Pst*I residue (TGCAG) at the beginning, each FASTQ file of 125 bp was split with a custom Perl script

fastqHiseqCutandCode-Pst.pl to get the first 64 bases with the *Pst*I residual restriction site. The script also provided an arbitrary barcode sequence (CATCAT) at the start of each sequence fragment, since the UNEAK pipeline expects to de-convolute barcoded sequence reads which are not already separated by sample. The 70-base-long fragments formed, thereafter, were recognized by the UNEAK-GBS pipeline (Lu et al. 2013), and passed into UNEAK.

The fragment set (70 bases long) was analyzed with UNEAK and the Haplotag pipelines (Tinker et al. 2016), resulting in the analysis of a total of 59 bases of genetic sequence. Supplementary Material, Section B, describes the procedures to run UNEAK. Two types of meta data files—a single mergedAll.txt (all tags observed more than 10 times) and a set of individual tagCount files (one per sample) needed for the Haplotag pipeline—were generated from the UNEAK run.

Haplotag was run with the parameters and filtering threshold settings described in the HTinput.txt file and generated a matrix of samples by SNP loci (online Supplementary Material, Section B). A set of tag-level haplotypes ("HTgenos") are first generated by Haplotag, followed by a set of SNP data derived from these haplotypes ("HTSNPgenos"). These two data types are technically redundant, so choosing one of them relies on the implementation and preference of software. In the present study, most (97.5%) haplotypes were found to contain only a single SNP; thus, we decided to analyze the SNP dataset for simplicity and compatibility with downstream analysis software.

The character by Taxa (CbyT) program supplied by N. Tinker was used to generate a filtered SNP file. In brief, Haplotag generated "HTSNPGenos" file, which was run with CbyT. The "minimum presence" value in CbyT was set to 80%, 70%, 60%, and 50% for 20%, 30%, 40%, and 50% missing data, respectively. A SNP-by-sample matrix in the output files was used in further analyses. Additional descriptions of the SNP data matrix and the custom Perl and Shell scripts are available in the Supplementary Material, Section A. Analyses from FASTQ file separation to SNP generation were conducted using Microsoft Windows 7 64-bit OS with an Intel (R) Xeon (R) CPU E5-2623 v3 @ 3.00 GHz (8 threads) and 32 GB RAM.

5.3.6. *GBS data imputation and filtering*

The SNPs marker information for the missing markers at each missing level (20%, 30%, 40% and 50%) were reconstructed using probabilistic PCA, a PCA-based imputation method, using freely available R package "pcaMethods" (Stacklies et al. 2007; Fu 2014). Following the imputation, SNPs were filtered using the technical replicates to get the same SNP information at each locus in the original and the replicate. Later, allele frequency of these SNPs was calculated independently for the 160 genotypes evaluated for nine morphological and three nutritive value traits in Saskatoon, and 159 genotypes evaluated for eight morphological traits and 156 genotypes evaluated for three nutritive value traits in Swift Current. Within each of these subsets, SNPs with minor allele frequency (MAF) < 0.05 were removed prior to their use in genomic selection models. Table 5.3 provides information about the SNPs used for the genomic selection models.

Table 5.3 Counts of SNPs at four levels of missing data and different filtering criteria

Missing level	SNP Counts	SNPs filtered with technical replicates	SNPs with MAF ≥ 0.05 Saskatoon	SNPs with MAF ≥0.05 Swift Current morphological traits	SNPs with MAF ≥0.05 Swift Current quality traits
20%	827	659	286	257	255
30%	3616	2149	1206	1154	1142
40%	14090	8091	5437	5321	5278
50%	46136	21957	17003	16771	16692

Note: MAF is minor allele frequency

5.3.7. *Population structure analysis*

Our study materials consisted of diploid and tetraploid crested wheatgrass lines. Thus, it was important to determine the existence of sub-population structure in the study material and estimate the effect of SNPs arising from the population structure prior to genomic selection. For this, the genetic structure of 325 crested wheatgrass genotypes was examined using a model-based Bayesian method implemented in the program STRUCTURE version 2.2.3 (Pritchard et al. 2000; Falush et al. 2007) where each population subgroup (K = 1–6) was run 10 times, using an admixture model with 10,000 replicates each for burn-in and during the analysis. This analysis was followed by a principal coordinates analysis

(PCoA) using the R routine AveDissR (Yang and Fu 2017; R development core team 2018) to assess genetic distinctness and redundancy, and to assess the genotype associations. Plots of the first two resulting principal coordinates were generated. Based on these results, we hypothesized that the first two principal coordinates (PCoAs) of the imputed markers (filtered with technical replicates but without filtering for MAF < 0.05) at each missing level would sufficiently account for these genetic differences due to ploidy. Hence, we fitted a model for each trait where trait was the response variable and the first two PCoAs from the PCoA of these SNPs were the explanatory variables. The residuals obtained fitting each model were used for genomic selection.

5.3.8. *Phenotypic evaluation*

The 160 genotypes from 10 crested wheatgrass lines were evaluated for nine morphological and three nutritive value traits in Saskatoon, while 159 genotypes were evaluated for eight morphological traits and 156 genotypes were evaluated for three nutritive value traits from Swift Current (the genotypes with incomplete morphological or quality traits in either of the years were not included). Outliers for each of the evaluated traits at each location were examined using studentized deleted residuals (Kutner et al. 2005) from a mixed linear model including year and population as random effects in SAS 9.4 (SAS Institute Inc. 2013). Best linear unbiased predictors (BLUPs) were estimated for each trait in each genotype across year using a mixed linear model fitted in R using the package "lme4" (Bates et al. 2014; R Development Core Team, 2018). Variance component estimates from the model used to obtain the BLUPs were used to estimate broad sense heritability on genotype mean basis ($\widehat{H}$) (Holland et al. 2003; Hung et al. 2012). Then, the Box-cox procedure was implemented to determine the optimal transformation of the traits (Box and Cox 1964).

5.3.9. *Genomic selection*

To access the applicability of genome wide markers generated using gd-GBS application to predict phenotypes of morphological and quality traits in crested wheatgrass, five additive genomic selection models (parametric methods) including BayesA (Meuwissen et al. 2001), BayesB (Meuwissen et

al. 2001), BayesCπ (Habier et al. 2011), Genomic Best Linear Unbiased Prediction (GBLUP) (Habier et al. 2007) and Ridge Regression Best Linear Unbiased Predictor (RRBLUP) (Meuwissen et al. 2001) were chosen. These models differ in their assumptions of the marker effects. BayesA assumes each marker has a distinct variance such that there are many markers with small effects and few markers with moderate effects. BayesB assumes only a portion of the markers explain total variance and most markers explain zero variance (Meuwissen et al. 2001). BayesCπ assumes a common marker effect variance and allows some markers to have no effect (Habier et al. 2011). GBLUP assumes common variance of all markers and uses a genomic relationship matrix. RRBLUP assumes that all markers have common variance with small but non-zero effect and thus shrinks equally for each marker effect (Meuwissen et al. 2001). Based on the assumptions we chose to implement these genomic selection models to estimate the breeding values of traits with unknown genomic architecture in crested wheatgrass. All statistical modeling was done in R 3.5.1 (R Core team, 2018). RRBLUP model was implemented using "rr-BLUP" package (Endelman, 2011) while Bayesian models and GBLUP were implemented using the "BGLR" package (Perez and de los Campos, 2014). The model parameters were considered following the package instructions.

The prediction ability of the models was assessed through ten-fold cross validation. First, the genotypes were randomly partitioned into ten equally sized subgroups. Then, for each cross-validation, nine of the subgroups were used as training set and the remaining one set as prediction set. The prediction set was used to assess the correlation between observed and predicted trait values. This process was repeated ten times, such that each subgroup was assigned as a prediction set exactly once. Prediction ability of each model for each trait was the average Pearson's correlation coefficient across ten folds.

5.4. Results

5.4.1. *Phenotypic variation*

Best linear unbiased predictors exhibited substantial variation in each of the morphological and quality traits. This variation was evident from the difference in minimum to maximum values for the traits (Table 5.4–5.5). In Saskatoon, this difference in minimum and maximum value ranged from 1.12-fold for NDF to 2.86-fold for tillers per plant (Table 5.4). The average broad sense heritability for the morphological traits was 0.59 with range of 0.38 (early spring vigor) to 0.84 (heading days). Likewise, for quality traits, heritability ranged from 0.29 for crude protein to 0.6 for ADF. In Swift Current, difference in minimum and maximum value ranged from 1.09-fold for days to heading to 2.77-fold for dry matter yield (Table 5.5). Heritability among the eight morphological traits ranged from 0.32 (fall regrowth) to 0.74 (plant height) with an average heritability of 0.52. Similarly, heritability among the quality traits ranged from 0.25 (crude protein) to 0.58 (ADF).

5.4.2. *Genotyping-by-sequencing*

The Hiseq run of 336 genotypes (six GBS libraries) from the 10 crested wheatgrass lines (Table 5.1) yielded approximately 888.2 million raw forward (R1) sequence reads. The number of raw forward sequence reads per sample ranged from 677,492 to 5,578,827 with an average of 2,643,717. Combined UNEAK and Haplotag analysis at the 20%, 30%, 40%, and 50% level of missing data generated 827; 3,616; 14,090; and 46,136 SNPs, respectively across the 336 genotypes. In addition, this analysis also generated many metagenomic files associated with the SNP discovery, which are described and accessible in the online Supplementary Materials. The data filtering was done by removing of SNPs differing with technical replicates and again with SNPs having MAF <0.05 resulted in 286, 1,206, 5,437 and 17,003 SNPs at 20%, 30%, 40%, and 50% level of missing data for genomic selection in Saskatoon (Table 5.3). Likewise, 257, 1,154, 5,321 and 16,771 SNPs for Morphological and 255, 1,142, 5,278 and 16,692 SNPs for nutritive

value traits at 20%, 30%, 40%, and 50% level of missing data respectively for genomic selection in Swift Current (Table 5.3).

Table 5.4 Means and range for best linear unbiased predictors (BLUPs) of nine morphological and three quality traits, and estimated heritability on genotype-mean basis for crested wheatgrass genotypes evaluated in two summer environments at Saskatoon

Traits	No. Lines	BLUP mean	BLUP SD	BLUP Range	Heritability (H)
Early spring vigor	160	3.96	0.27	3.21–4.36	0.38
Heading days (GDD)	160	705.88	57.44	572.17–864.88	0.84
Plant height	160	90.03	9.83	48.46–114.57	0.72
Tillers per plant	160	256.5	48.02	140.28–400.93	0.56
Leaf width	160	8.28	0.78	5.89–10.18	0.62
Clump Diameter	160	20.49	1.99	16.40–25.54	0.63
Regrowth score after harvest	160	3.62	0.47	2.46–4.37	0.55
Dry matter yield (Biomass)	160	338.09	53.2	169.47–483.16	0.65
Fall regrowth score	160	3.32	0.44	2.40–4.17	0.51
Acid detergent fiber	160	36.96	1.67	32.47–41.37	0.60
Neutral detergent fiber	160	59.69	1.16	55.94–62.93	0.41
Crude protein	160	3.8	0.21	3.46–4.56	0.29

Table 5.5 Means and range for best linear unbiased predictors (BLUPs) of eight morphological and three quality traits, and estimated heritability on genotype-mean basis for crested wheatgrass genotypes evaluated in two summer environments at Swift Current

Traits	No. Lines	BLUP mean	BLUP SD	BLUP Range	Heritability (H)
Early spring vigor	159	3.84	0.29	2.82–4.34	0.43
Heading days (GDD)	159	716.01	17.52	699.87–760.42	0.38
Plant height	159	83.29	8.78	55.03–102.05	0.74
Leaf width	159	7.91	0.78	6.38–10.55	0.65
Clump Diameter	159	22.72	1.71	18.58–27.21	0.62
Regrowth score after harvest	159	3.89	0.36	2.74–4.43	0.48
Dry matter yield (biomass)	159	508.97	69	279.71–775	0.58
Fall regrowth score	159	3.81	0.16	3.40–4.19	0.32
Acid detergent fiber	156	34.63	1.73	30.61–39.50	0.58
Neutral detergent fiber	156	59.23	1.53	55.10–63.03	0.45
Crude protein	156	7.01	0.3	6.34–7.85	0.25

5.4.3. *Population strucutre*

Population structure estimated for 325 genotypes from 10 crested wheatgrass lines without consideration of prior population information in the STRUCTURE analysis revealed the presence of population structure with two or more clusters (Data not shown). Further investigation of the population structure with principal coordinate analysis (PCoA) revealed that diploid cultivars and breeding lines clustered separately from the tetraploid cultivars and breeding lines with little overlap. This suggests the first two PCoAs of the SNPs were able to infer sufficient level of genetic discrepancies between the two ploidy levels and support our hypothesis of the effectiveness of using the first two PCoAs to mitigate the SNP effect arising from the population structure prior to a genomic selection study.

5.4.4. *Genomic selection potential in crested wheatgrass*

The average prediction ability of the five genomic selection models evaluated for crested wheatgrass varied for each trait at the four SNP densities corresponding to the missing levels (20%, 30%, 40% and 50%) of SNPs markers (Table 5.6–5.7 and Figure 3.2–5.5). However, the prediction ability of the five GS models were similar for most of the traits. The genomic selection models were implemented separately for genotypes evaluated in Saskatoon and Swift Current. In Saskatoon, the average prediction ability of GS models at 50% missing level of SNPs information ranged from low to moderate prediction abilities (0.01 to 0.41) for dry matter yield and yield related morphological traits (DTH, LW, PH, CD, TPP). Likewise, prediction ability of vigor and regrowth scores (ESV, RGAH and FRG) were low, ranging from -0.28 to 0.19. While, for forage quality traits (ADF, NDF and CP) prediction ability ranged from -0.17 to 0.07 (Table 5.6). In Swift Current, the average prediction ability of GS models at 50% missing level of SNPs information ranged from -0.18 to 0.43 for dry matter yield and yield related traits (DTH, LW, PH, CD). Likewise, prediction ability of vigor and regrowth score (ESV, RGAH and FRG) ranged from -0.20 to 0.30; while, for forage quality traits (ADF, NDF and CP) prediction ability ranged from -0.24 to 0.24 (Table 5.7).

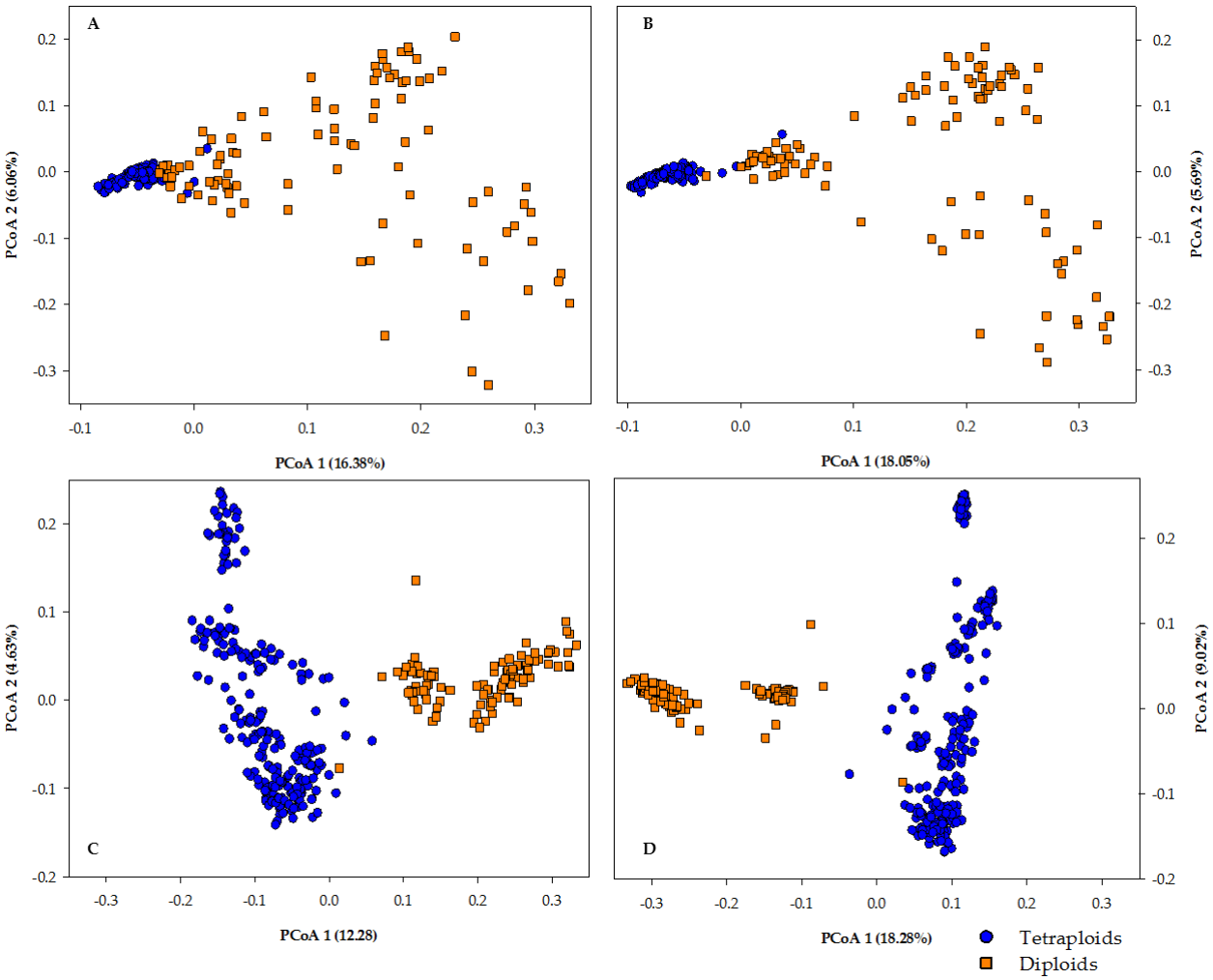

Figure 5.1 Population structure of diploid and tetraploid crested wheatgrass genotypes used for the genomic selection study as explained by PCoA1 and PCoA2 of principal coordinate analysis. (A) SNPs at 20% missing level; (B) SNPs at 30% missing level; (C) SNPs at 40% missing levels and (D) SNPs at 50% missing levels.

Comparison of five additive GS models (parametric methods), BayesA, BayesB, BayesCπ, GBLUP and RRBLUP at SNP density corresponding to 50% missing level for traits evaluated at Saskatoon exhibited similar prediction abilities for DM, DTH, PH, CD, TPP, ESV, RGAH, ADF, NDF and CP. Whereas, for the traits LW and FRG, prediction ability of RRBLUP was lower than the Bayesian models and the GBLUP (Table 5.6). This is evident from average prediction ability of the models and associated

standard errors (Table 5.6). However, prediction ability of RRBLUP was similar to that of three Bayesian models (BayesA, BayesB, BayesCπ) and GBLUP for CD and TPP, while, it was lower for remaining seven morphological traits (DM, DTH, LW, PH, ESV, RGAH, FRG). Likewise, the prediction ability of RRBLUP was comparable to BayesA, BayesB, BayesCπ and GBLUP for CP, while, lower than the remaining four models for ADF and NDF (Table 5.6). Similarly, prediction ability of the five GS models at SNP density corresponding to 50% missing level for traits evaluated at Swift Current were similar for DM, DTH, PH, CD, ESV, RGAH, FRG, ADF and NDF. Whereas, prediction ability of RRBLUP was lower than the Bayesian models and GBLUP for LW and CP (Table 5.7). Overall, the prediction ability of RRBLUP was lower compared to BayesA, BayesB, BayesCπ and GBLUP for DM, DTH, LW, CD, NDF and CP. While, RRBLUP was similar in prediction ability compared to remaining four GS models for PH, ESV, RGAH, FRG and ADF (Table 5.7).

Table 5.6 Prediction accuracies of five genomic selection models at 50% missing level of SNPs information for nine morphological and three forage quality traits with ten-fold cross validation in crested wheatgrass evaluated at Saskatoon

SNPs 50% missing	Bayes A	Bayes B	Bayes Cπ	GBLUP	RRBLUP
Dry matter	0.05 (0.1)	0.05 (0.09)	0.05 (0.09)	0.05 (0.09)	0.01 (0.07)
Days to Heading	0.09 (0.07)	0.11 (0.07)	0.10 (0.09)	0.08 (0.08)	0.01 (0.09)
Leaf width	0.32 (0.07)	0.30 (0.07)	0.30 (0.07)	0.30 (0.07)	0.01 (0.07)
Plant Height	0.26 (0.09)	0.26 (0.09)	0.26 (0.10)	0.26 (0.09)	0.15 (0.08)
Clump Diameter	0.34 (0.06)	0.35 (0.06)	0.34 (0.06)	0.34 (0.07)	0.35 (0.06)
Tillers per plant	0.41 (0.05)	0.41 (0.05)	0.40 (0.05)	0.39 (0.05)	0.40 (0.05)
Spring vigour	-0.05 (0.06)	-0.08 (0.07)	-0.06 (0.06)	-0.08 (0.06)	-0.13 (0.05)
Regrowth score	0.19 (0.07)	0.18 (0.07)	0.18 (0.07)	0.19 (0.07)	0.09 (0.07)
Fall regrowth	0.02 (0.07)	0.01 (0.07)	0.01 (0.06)	-0.01 (0.07)	-0.18 (0.07)
ADF	0.06 (0.07)	0.06 (0.07)	0.07 (0.07)	0.07 (0.07)	-0.01 (0.06)
NDF	0.02 (0.18)	-0.14 (0.19)	-0.03 (0.14)	-0.05 (0.14)	-0.13 (0.13)
Protein	-0.11 (0.07)	-0.17 (0.07)	-0.10 (0.06)	-0.13 (0.08)	-0.14 (0.06)

The values in the parenthesis are standard errors.

Table 5.7 Prediction accuracies of five genomic selection models at 50% missing level of SNPs information for eight morphological and three forage quality traits with ten-fold cross validation in crested wheatgrass evaluated at Swift Current

SNPs 50% missing	Bayes A	Bayes B	Bayes Cπ	GBLUP	RRBLUP
Dry matter	-0.14 (0.08)	-0.14 (0.09)	-0.14 (0.08)	-0.13 (0.09)	-0.18 (0.05)
Days to Heading	0.08 (0.07)	0.07 (0.08)	0.06 (0.08)	0.05 (0.07)	-0.06 (0.07)
Leaf width	0.10 (0.05)	0.13 (0.05)	0.10 (0.05)	0.09 (0.05)	-0.16 (0.07)
Plant Height	0.43 (0.04)	0.42 (0.04)	0.42 (0.04)	0.42 (0.04)	0.42 (0.04)
Clump Diameter	-0.03 (0.07)	-0.01 (0.06)	-0.01 (0.06)	-0.03 (0.07)	-0.06 (0.06)
Spring vigour	-0.18 (0.08)	-0.16 (0.08)	-0.15 (0.08)	-0.2 (0.07)	-0.14 (0.06)
Regrowth score	-0.02 (0.1)	-0.01 (0.1)	-0.05 (0.11)	-0.04 (0.11)	-0.04 (0.09)
Fall regrowth	0.30 (0.09)	0.30 (0.09)	0.30 (0.09)	0.30 (0.09)	0.29 (0.09)
ADF	0.23 (0.06)	0.22 (0.07)	0.23 (0.07)	0.24 (0.07)	0.20 (0.07)
NDF	0.07 (0.09)	0.09 (0.09)	0.08 (0.09)	0.07 (0.09)	0.03 (0.09)
Protein	0.17 (0.06)	0.15 (0.06)	0.16 (0.06)	0.14 (0.05)	-0.24 (0.05)

The values in the parenthesis are standard errors.

Our results demonstrated that, for the genotypes evaluated at Saskatoon, the GS models were similar in their predictive ability for most of the morphological traits (except for NDF at 20% and LW at 50% missing levels) at each level of SNPs density (corresponding to missing SNPs levels). Overall, in Saskatoon, the prediction ability of each of the five genomic selection models for DTH, CD, TPP, ESV, and CP were similar as evident from the standard errors, for the four levels of SNP densities corresponding to missing levels. Likewise, the prediction ability of the models was similar at 30%, 40% and 50% missing levels for ADF, RGAH, FRG, DM, LW and PH, except for the lower prediction ability of RRBLUP for LW and PH at 50% missing level. The prediction ability of the GS models at 20% missing level was lower for DM, FRG and LW except for the prediction ability of RRBLUP for FRG and LW at 50% missing level. While, the prediction ability of the GS models at 20% missing level was higher for ADF but, higher and comparable to that at 30% missing level for RGAH and PH. The prediction ability of GS models for NDF at 30% missing level was lower than that at remaining levels of SNP densities, however, BayesCπ, GBLUP and RRBLUP were comparable to the prediction ability at 40% missing level. Increasing trend in the prediction ability with increase in marker densities was observed for DM, DTH,

LW and FRG. The exception was the prediction ability of RRBLUP for LW, DTH and FRG at SNPs densities with 50% missing level. While, decreasing trend with increasing marker densities was observed for PH and CP. While, for traits TPP, ESV, RGAH, ADF and NDF the prediction ability decreased while increasing the SNP density from 20% to 30% missing level but gradually improved at 40% and 50% missing levels of SNP densities. However, for CD, the prediction ability of the models improved with increasing SNP densities but was inconsistent as the prediction ability decreased at 40% compared to that at 30% and later improved at 50% missing level (Figure 5.2–5.3).

In Swift Current, the prediction ability of each of the GS models was similar at four levels of SNP densities for DTH (except for lower prediction ability of RRBLUP at 50% missing level), PH, RGAH, and FRG. The prediction ability of the GS models for CD and CP was similar at 30%, 40% and 50% missing levels except for the RRBLUP which was similar at all SNP densities for CP and similar in prediction ability at 40% missing level for CD. The prediction ability of the models at 20% missing level was lower than at the remaining SNP densities for CP but, was higher than at the remaining SNP densities for CD. Likewise, for DM and ESV the prediction ability of the models was higher at 30% missing level and lower but similar at 20%, 40% and 50% missing levels. The prediction ability of the GS models for LW, and ADF was higher and similar at 20% and 30% missing levels than at 40 and 50% missing levels. Higher prediction ability of the GS models at 20% and 30% compared to remaining SNP densities was also observed for NDF. Increasing trend of prediction ability with increasing SNP densities was observed for PH, FRG and CP (except for RRBLUP). Decreasing trend of prediction ability of the GS models with increasing SNP densities was observed for LW, CD, RGAH and NDF. The prediction ability of the GS models for DM and ESV improved at 30% missing level but declined with increasing SNP densities above it. However, for DTH and ADF, the prediction ability of the models improved with increasing SNP densities but was inconsistent as the prediction ability decreased at 40% compared to that at 30% and later improved at 50% missing level (Figure 5.4–5.5).

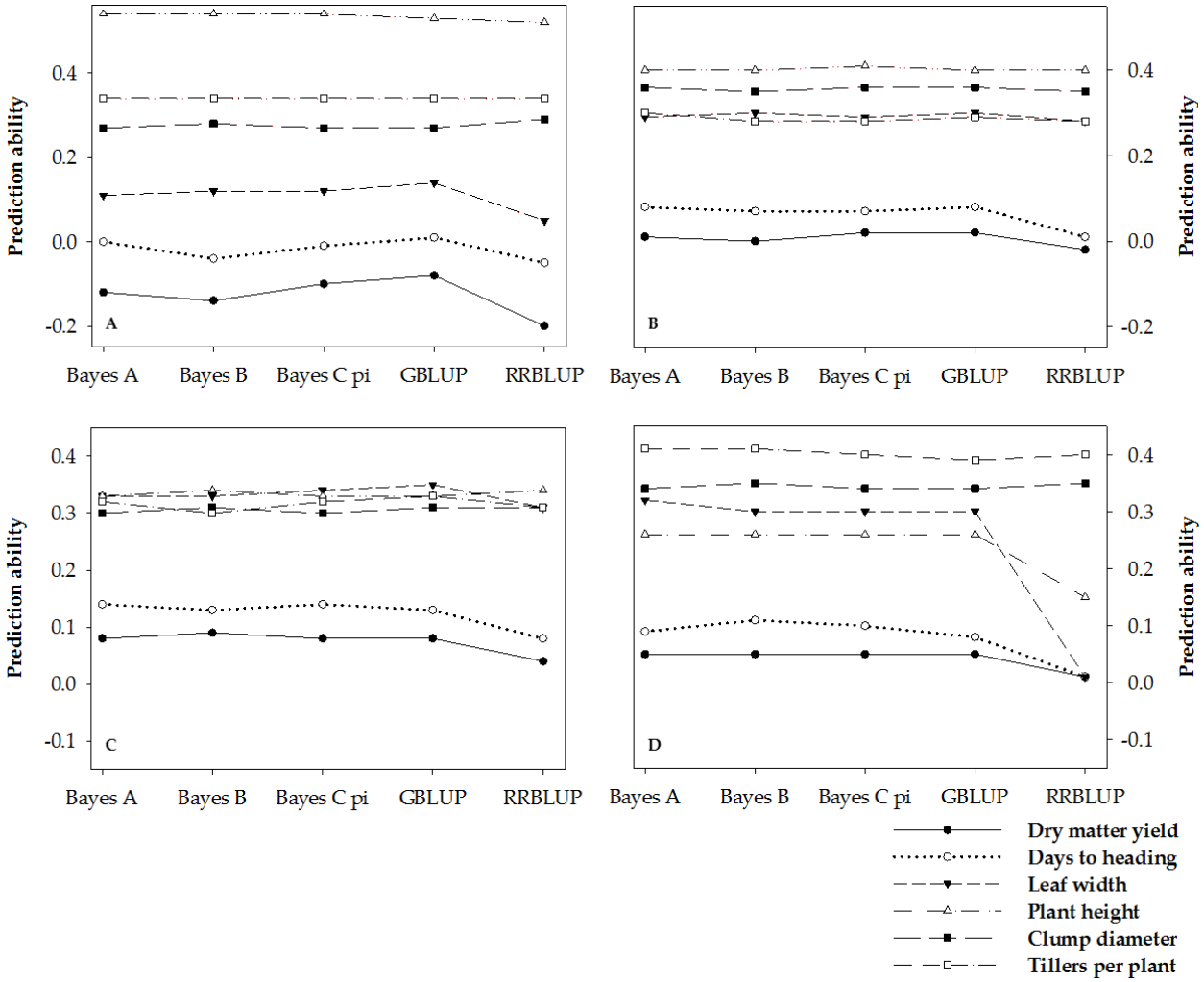

Figure 5.2 Prediction ability of five genomic selection models for dry matter yield, days to heading, leaf width, plant height, clump diameter and tillers per plant with (A) SNPs at 20% missing level; (B) SNPs at 30% missing level; (C) SNPs at 40% missing levels and (D) SNPs at 50% missing levels for genotypes evaluated at Saskatoon.

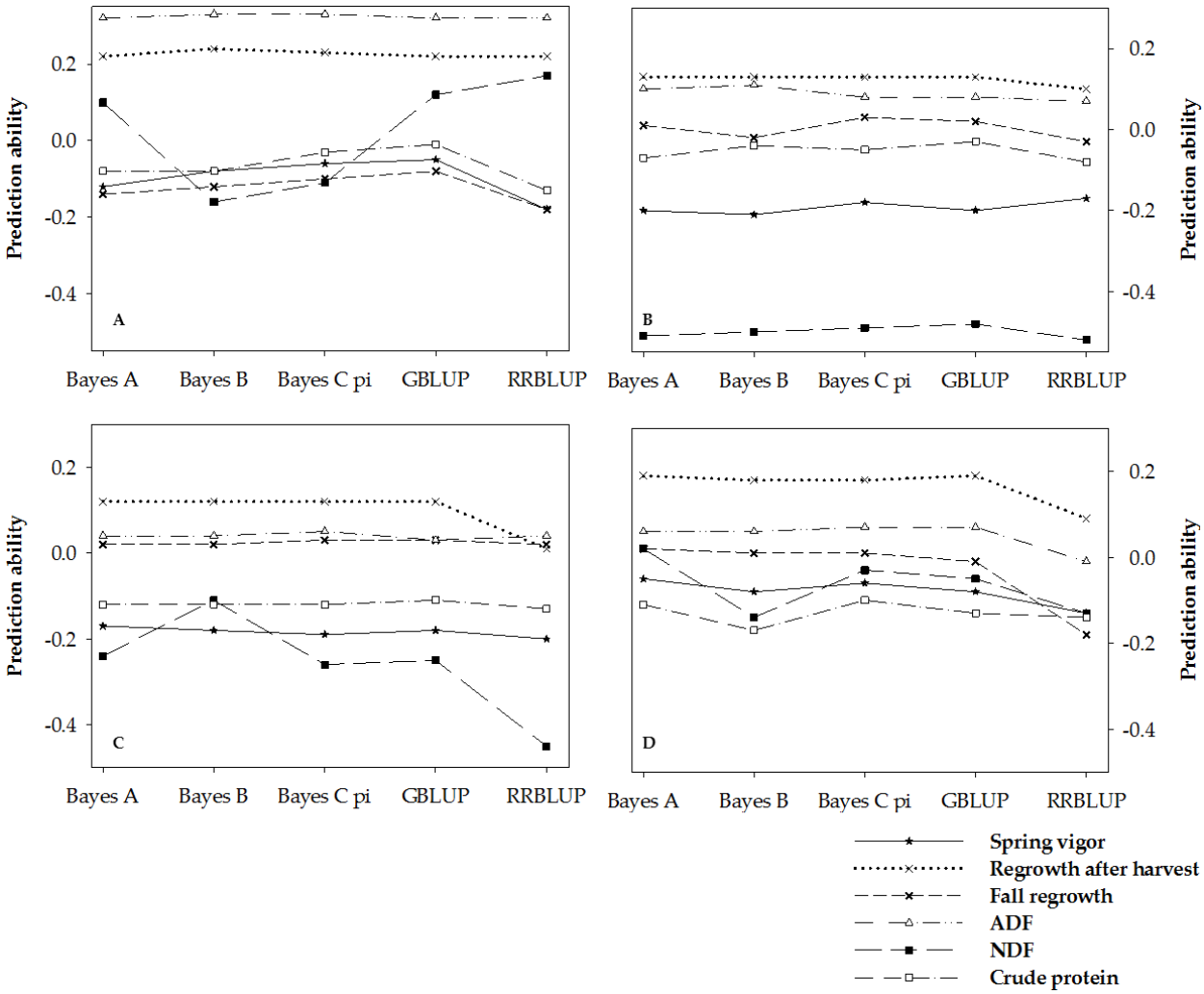

Figure 5.3 Prediction ability of five genomic selection models for early spring vigor, regrowth after harvest, fall regrowth, ADF, NDF and crude protein with (A) SNPs at 20% missing level; (B) SNPs at 30% missing level; (C) SNPs at 40% missing levels and (D) SNPs at 50% missing levels for genotypes evaluated at Saskatoon.

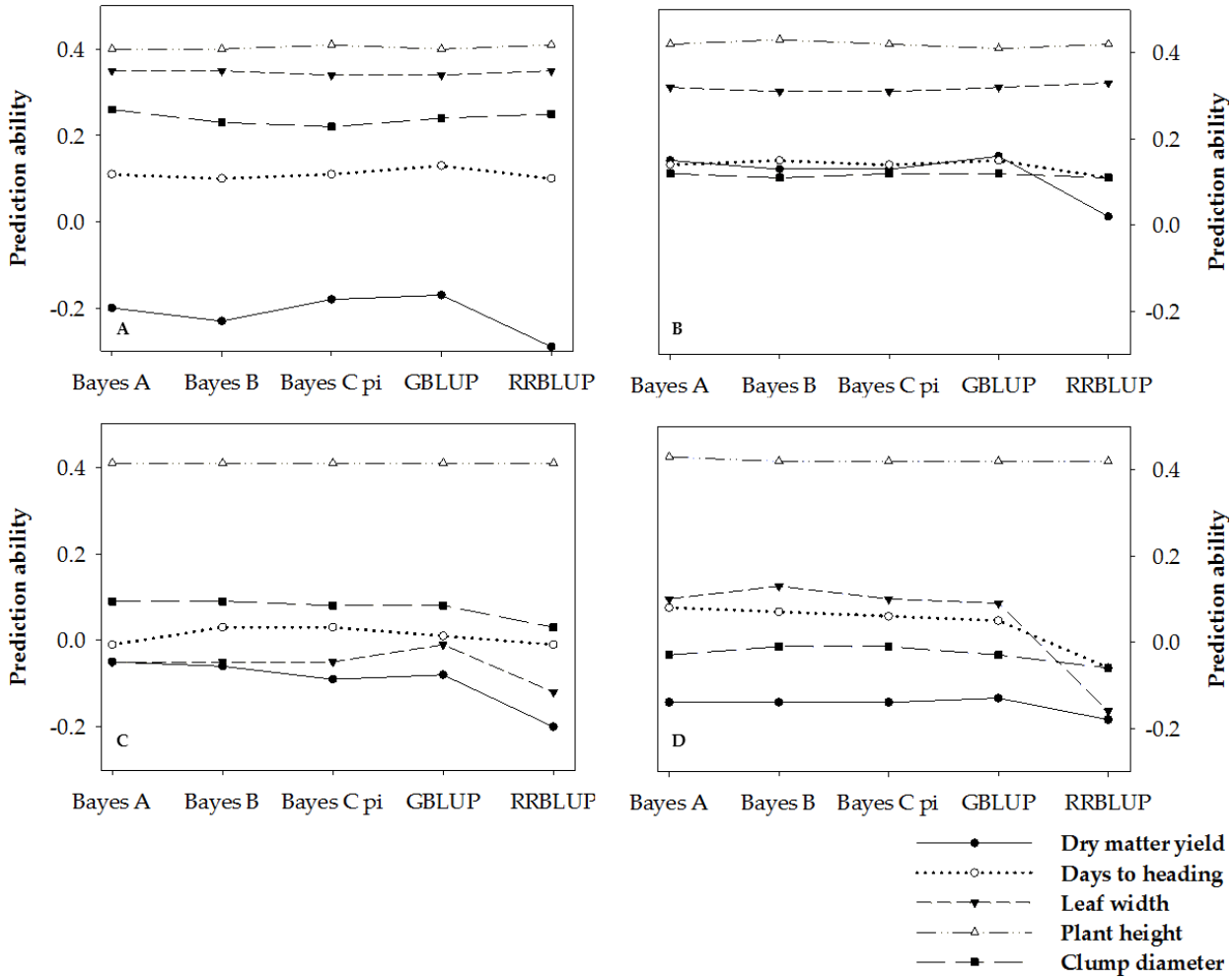

Figure 5.4 Prediction ability of five genomic selection models for dry matter yield, days to heading, leaf width, plant height and clump diameter with(A) SNPs at 20% missing level; (B) SNPs at 30% missing level; (C) SNPs at 40% missing levels and (D) SNPs at 50% missing levels for genotypes evaluated at Swift Current.

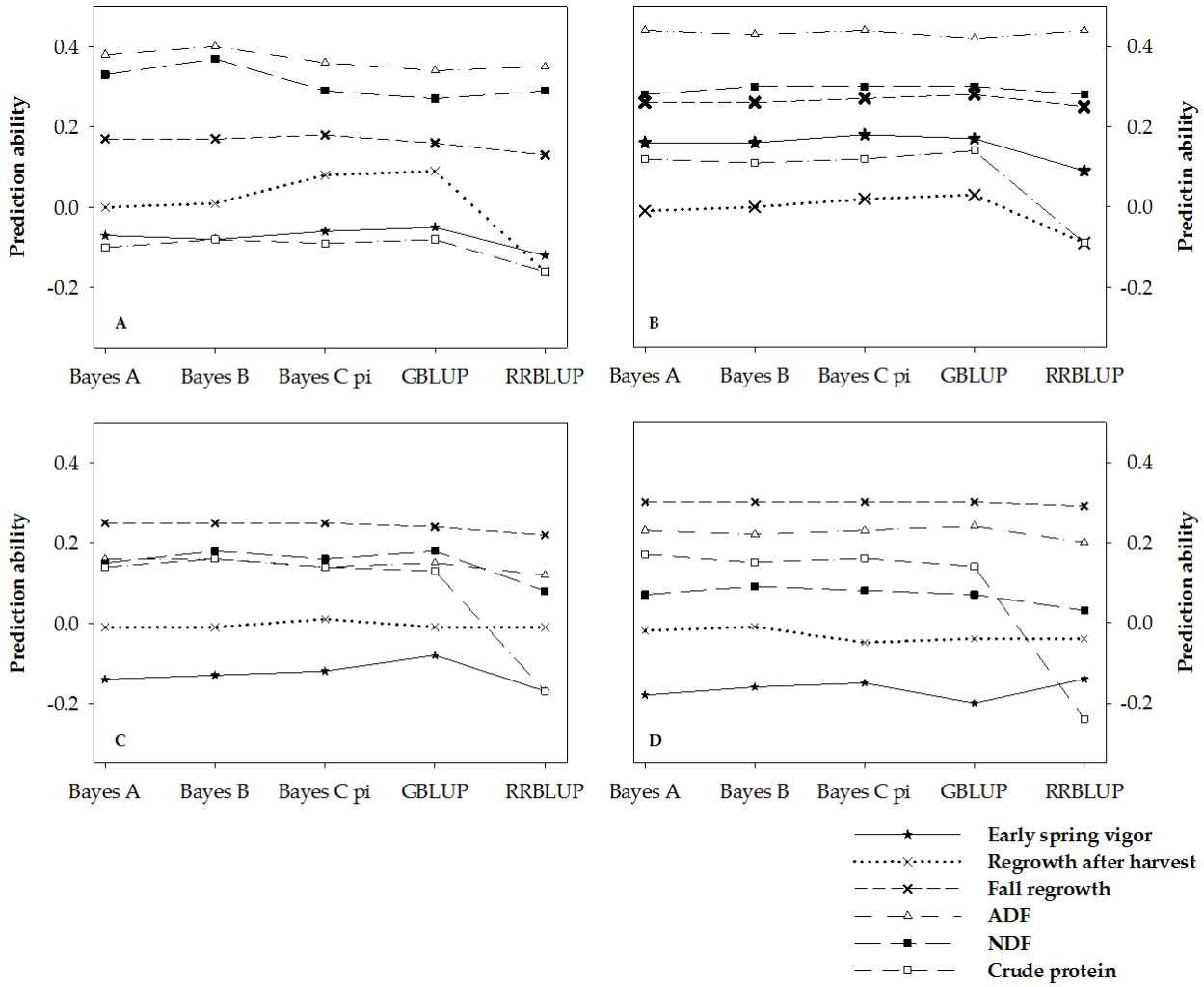

Figure 5.5 Prediction ability of five genomic selection models for early spring vigor, regrowth after harvest, fall regrowth, ADF, NDF and crude protein with (A) SNPs at 20% missing level; (B) SNPs at 30% missing level; (C) SNPs at 40% missing levels and (D) SNPs at 50% missing levels for genotypes evaluated at Swift Current.

5.5. Discussion

This study, for the first time, investigated the prediction ability of genomic selection models in crested wheatgrass breeding utilizing the gd-GBS application for SNP marker generation. This study assessed and compared the prediction ability of the five additive (parametric) GS models. Overall, the GS models were similar in prediction ability for the evaluated traits at each level of SNP densities. However, Bayesian models along with the GBLUP were better performing compared to RRBLUP for the morphological and nutritive value traits in crested wheatgrass. The relationship between the prediction ability of the models and the density of the SNP markers showed differences in prediction ability with increased SNP densities.

Given the long breeding cycles required for varietal development, genomic selection provides an opportunity in crested wheatgrass breeding, enabling estimation of GEBVs from seedlings without requirement of phenotyping which significantly reduces the breeding cycles there by reducing the overall cost and time. The prediction accuracies for the morphological traits and the three forage nutritive value traits from our study are comparable to findings in other perennial forage crops (Lipak et al, 2014; Jia et al 2018). Moderate prediction accuracies (in range of 0.20–0.41) for traits such as ADF, TPP, CD, PH, LW and FRG (Table 5.6 and Table 5.7) indicate the possibility of application of genomic selection in crested wheatgrass to improve the genetic gain per unit of time. Heffner et al (2010) reported genetic gain per year with GS to be greater than MAS for the traits with the prediction accuracies of 0.2 in maize and 0.3 in winter wheat. The present study showed negative prediction abilities ranging from -0.24 to -0.01, for DM, ESV, NDF, CP, CD, RGAH and FRG (Table 5.6 and Table 5.7). Negative prediction accuracies have also been reported from genomic selection studies in maize (Riedelsheimer et al. 2013; Massman et al. 2013), sugar beet (Würschum et al. 2013), perennial ryegrass (Grinberg et al. 2016; Pembleton et al. 2018) and switchgrass (Lipka et al. 2014; Ramstein et al. 2016; Fiedler et al. 2018). These negative prediction accuracies could have resulted from opposite linkage phases between markers and QTL in training and prediction sets as

discussed by Würschum et al. (2013) and Riedelsheimer et al. (2013). Genotypes in the training and validation sets were half-sibs from different families, which could result in opposite linkage phase among the genotypes of training and validation sets resulting in the lower or negative prediction ability as explained in a genomic selection study of maize (Riedelsheimer et al. 2013). A previous simulation study suggested that more accurate predictions are obtained with the inclusion of multiple populations that have marker-QTLs in the same LD phase in the training set (de Roos et al. 2009). Pembleton et al. (2018) reported low to negative prediction accuracies for biomass across seasons and years and stated this could be the result of environmental distinctness, prevalence of genotype by environmental effects (G x E) or even unusual environment between seasons and years.

Our study showed difference in the prediction ability of models for traits evaluated in two different environments. The prediction ability was moderate for yield related traits in Saskatoon, while it was moderate for PH, FRG and ADF in Swift Current (Table 5.7). This difference can be attributed partly to the difference in the genotypes being used in the models and partly to the differences in environmental effects. Saskatoon (moist mixed grassland ecoregion with semi-arid moisture condition and dark brown soil zone) differs in agroclimatic and soil zone from Swift Current (mixed grassland ecoregion with semi-arid condition and brown soil zone) (University of Saskatchewan 2019). Differences in the prediction ability for alfalfa lines evaluated in two different environments have been reported (Annicchiarico et al. 2015). Prediction ability was variable across seasons and years for perennial ryegrass (Pembleton et al. 2018). Such a difference in prediction ability of models evaluated at different locations and years depends on the prevalence and extent of genotype-by-environment interactions (Grinberg et al. 2016).

In crested wheatgrass, the number of genes and their effects on the traits being evaluated are unknown to date. For this reason, we implemented five different additive GS models (parametric models) which differ in their prior assumptions for the marker effects and variance distributions. Regardless of the differences in the prior assumptions, the prediction ability was not, or only modestly, influenced by

the choice of the GS models when considering the standard errors (Table 5.6 and Table 5.7). However, compared to the remaining four GS models, the prediction ability of RRBLUP was inferior for most of the traits except for CD, TPP and CP evaluated at Saskatoon, and PH, FRG and ADF evaluated at Swift Current, where the prediction ability of RRBLUP was similar to the Bayesian and the GBLUP models. The differences were only modest as evident from the standard errors; similar findings have been reported from previous studies (Heffner et al. 2011b; Jarquín et al. 2014; Charmet et al. 2014; Tayeh et al. 2015). In general, the results indicate that the prediction ability is unaffected by the choice of the additive GS models (parametric models) implemented in this study. The difference observed in the prediction ability of models could be attributed to the assumptions about the variance of marker effects. BayesA assigns each marker its unique variance, BayesCπ assigns a proportion of the marker effects to zero and BayesB have both, whereas, the use of a genomic relationship matrix in GBLUP might have resulted in the difference from RRBLUP. The Bayesian models have strong shrinkage of the marker effects for loci with no effect compared to frail shrinkage for loci with large effects. This results in higher prediction abilities of Bayesian models compared to GBLUP and RRBLUP for traits influenced by few QTLs with large effects. However, for the traits affected by large number of small effect QTLs, all QTLs will have some effects and variable shrinkage is irrelevant, leading to similar prediction by the Bayesian, GBLUP and RRBLUP models (Daetwyler et al. 2010; Heffner et al. 2011a). From the similar prediction ability of the Bayesian, GBLUP and RRBLUP models in this study, it could be inferred that the morphological and nutritive value traits in crested wheatgrass (with unknown genetic architecture) could have been affected by a large number of small effect QTLs. Studies from maize and alfalfa have shown similar prediction accuracies of the Bayesian models (Hao et al. 2018; Jia et al. 2018). BayesA and GBLUP were similar in prediction ability for biomass yield in perennial ryegrass (Pembleton et al. 2018). Further, the prediction ability of the GS models depends on the effective population size (N_e) and the trait architecture referred to as the number of QTL (N_{QTL}) affecting the trait and the distribution of their effects. These are the property

of the number of independent chromosomal segments (M_e) in a population which depends on N_e (Daetwyler et al. 2010). The prediction ability of GS models is reported to be influenced by the genetic architecture of the trait and gene interactions. The traits that are under influence of additive effects have been reported to be better predicted by GBLUP and a low prediction could be reasoned for epistatic gene-actions (in *Drosophila melanogaster*) (Ober et al. 2012). A simulation study reported poor prediction ability of parametric models for traits with epistatic gene effects while, a slightly improved prediction ability compared to non-parametric models for traits with additive genetic architecture (Howard et al. 2014). In a recent GS study by Momen et al. (2018), parametric methods were reported to perform better than non-parametric methods under additive gene action, which was opposite under epistatic gene action. Though the models were similar in the prediction ability, low prediction accuracies observed for DM, vigor and regrowth and nutritive value traits could have resulted from the choice of models (parametric models) that were unable to model the underlying genetic architecture of traits related to the gene action involved, which are yet to be determined in crested wheatgrass.

In this study we also attempted to investigate the effect of SNP densities in the prediction ability of genomic selection models with the use of four different SNP densities corresponding to the SNP missing levels. Our results demonstrated different trends of prediction ability for different traits with increasing SNP densities (Figure 5.2–5.5). For the traits evaluated at Saskatoon, an increasing trend of the prediction ability with increasing marker densities was observed for DM, DTH, LW and FRG. A decreasing trend with increasing marker densities was observed for PH and CP. For TPP, ESV, RGAH, ADF and NDF the prediction ability decreased with increasing the SNP density from 20% to 30% missing level but gradually improved at 40% and 50% missing levels of SNP densities. For CD, the prediction ability of the models improved with increasing SNP densities but decreased at 40% compared to that at 30% and later improved at 50% missing level (Figure 5.2–5.3). Likewise, for the traits evaluated at Swift Current, increasing trend of prediction ability with increasing SNP densities was observed for PH, FRG and CP (except for RRBLUP).

Decreasing trend of prediction ability of the GS models with increasing SNP densities was observed for CD, RGAH, NDF and LW, except for the improvement in prediction ability of LW at 50% missing level. The prediction ability of the GS models for DM and ESV improved at 30% missing level but dropped with increasing SNP densities above it. However, for DTH and ADF, the prediction ability of the models improved with increasing SNP densities, but the prediction ability decreased at 40% compared to that at 30% and later improved at 50% missing level (Figure 5.4–5.5). Higher marker densities that are evenly distributed across the genome have been reported to improve the prediction accuracies by increasing the probability of each QTL to be in LD with at least one marker (Goddard 2009; de Roos et al. 2009; Heffner et al. 2011a). In contrast to this assumption, little improvement in prediction ability (increasing trend) with increasing SNP densities in our study could have resulted from uneven genome coverage of the GBS approach, missing markers, and the imputation method implemented to reconstruct the missing markers. Irrespective of the high number of SNPs generated though the GBS, the actual number of independent chromosomal segments (M_e) generated is reported to be generally low (Rabier et al. 2016). This decreases the actual number of markers across the genome affecting the trait controlled by large number of QTLs, consequently limiting the prediction ability of the GS models. Uneven distribution of GBS SNPs across the reference genome of switchgrass have been reported (Fiedler et al. 2018). Fiedler et al (2018), from the same study, reported that prediction ability of GS was not affected by reducing the SNP densities to 3000. Faville et al (2016) in a GS study in perennial ryegrass, utilizing SNP markers generated from GBS, reported the prediction accuracy was similar with about a threefold increase in the marker numbers. Moreover, density of markers required for GS is based on the rate LD decay, which in turn is determined by the product of effective population size (N_e) and the rate of recombination (c) (Gaut and Long 2003). In general, the effective population size of outcrossing species is high, thus the rate of LD decay is high, requiring higher marker densities for higher prediction accuracies of GS (Lin et al. 2014). However, the cultivars and breeding lines used in this study have gone through several cycles of selection and thus have relatively

small number of founders, reducing the N_e which reduces the number of markers required for GS. This could also be the reason for little or no improvement in the prediction ability of the GS models with increasing marker density. Increase in prediction ability for traits with complex genetic architecture, and a decrease in prediction ability for traits with simple genetic architecture with increasing marker density have been reported by Zhang et al. (2019). However, decreasing and increasing trend of the prediction ability with increasing SNP density observed for complex traits in this study could be reasoned for the genetic makeup of the traits and inability of the models to account for gene interactions. Differences in the trends of prediction ability for traits evaluated at Saskatoon and Swift Current could be the result of different genetic interactions in different environments. Difference in the prediction accuracy of oil content evaluated in two different years were reasoned to be influenced by pleiotropic effects of certain genetic factor in specific environments (Werner et al. 2018).

This study attempted to assess SNPs generated using the GBS application in diploid and tetraploid crested wheatgrass plants to predict the GEBVs of agronomic, morphological and nutritive value traits in crested wheatgrass using five additive (parametric) genomic selection models. Overall, this study opens the possibility for application of genomic selection in crested wheatgrass towards accelerated breeding. The GS approach in crested wheatgrass breeding is still in its early stage, but shows promise. Genomic selection has been found to be beneficial in other crops (Sallam et al. 2015; Spindel et al. 2015; Jan et al. 2016), trees (Resende et al. 2012; Ratcliffe et al. 2015; Gamal El-Dien et al. 2015; Isik et al. 2016), animals (Hayes and Goddard 2010; Knol et al. 2016) and forage crops (Lipka et al. 2014; Annicchiarico et al. 2015; Li et al. 2015; Zhang et al. 2016; Grinberg et al. 2016; Ramstein et al. 2016; Jia et al. 2018; Faville et al. 2018; Pembleton et al. 2018; Fiedler et al. 2018). These studies demonstrated genomic selection approaches increase the genetic gain for some traits in a cost-effective way by reducing the length of the selection cycle, and this is expected to be true for the development of high-quality, high-yielding, late maturing crested wheatgrass for extended summer grazing. However, low to moderate prediction

abilities observed in our study suggests that repeating the study may be useful. The repeat would involve refining field experiments for improved phenotyping, increasing the population size, and using genomic selection models that account for genotype-by-environment interactions. This may significantly improve the prediction ability of GS in crested wheatgrass. In addition to this, the genetic architecture of the traits under study influences the prediction ability of the GS models, thus for the traits with unknown genetic architecture in crested wheatgrass, studies are needed for the comparison of parametric models with the non-parametric models which are reported to perform better under epistatic gene effects. We foresee advances in crested wheatgrass research with recent technologies will enable development of high-density markers with a higher level of genome coverage which will in turn provide opportunity for development of a reference genome that will serve as a genomic resource in future to improve crested wheatgrass.

5.6. Conclusions

The genotypes used in this study represents the cultivars and breeding lines from the forage breeding program of the University of Saskatchewan. The genotyping-by-sequencing application generated 827, 3,616, 14,090 and 46,136 SNP markers at four SNP markers densities (20%, 30%, 40% and 50% missing SNPs level), respectively. These markers were useful in investigating population structure of the crested wheatgrass genotypes used in the genomic selection study, and estimated the prediction ability of five additive (parametric) genomic selection models. The GS models showed that moderate prediction ability can be achieved for traits such as LW, PH, CD, TPP, ADF and FRG. Similar prediction ability of the parametric GS models implemented in this study indicate that the choice of the GS model has little or no influence on the prediction ability. Similar prediction results of the Bayesian, GBLUP and RRBLUP models at each level of SNP density in this study leads to the assumption that, the morphological and nutritive value traits in crested wheatgrass (with unknown genetic architecture) were affected by a large number of small effect QTLs. In addition to this, low to moderate prediction ability of

the GS models with increasing SNP density leads to the conclusion that the SNP markers might not have effectively captured the independent chromosomal segments affecting the traits across the genome. The increasing and decreasing trend of the prediction ability observed for the complex traits also indicates the possibility of complex gene interactions such as epistasis could involve in the expression of traits, which were not captured by the parametric GS models implemented in this study. The prediction abilities of the measured characters were low to moderate, similar to findings for these traits in other plant species. These results serve as a foundation for the improvement of GS strategies in crested wheatgrass.

Supplementary Materials: Supplementary materials can be found at https://figshare.com/s/a904cc6d0553aafbe3fa.

6. General Discussion and Conclusions

In this thesis three studies were conducted in an attempt to develop genomic resources for the accelerated breeding of crested wheatgrass, a perennial temperate C_3 forage species, commercially important as a hay and pasture species in western Canada. The work was based on the hypotheses that: 1) GBS application will generate large number of SNP markers in the non-model outcrossing grass species crested wheatgrass; 2) High density genome-wide SNP markers will be useful for studying genetic diversity and population structure within this species. 3) SNP markers generated in crested wheatgrass can provide a higher resolution linkage map for crested wheatgrass than the other genetic markers in F1 mapping population of intraspecific crosses. 4) A combined genotypic and phenotypic data set generated for crested wheatgrass predicts breeding values more precisely than only phenotypic information and improves selection accuracies in forage breeding.

6.1. Genotyping-by-Sequencing for SNP marker discovery

Genome-wide SNP markers required for each of the three projects in the present study were generated utilizing a genetic diversity-focused genotyping-by-sequencing (gd-GBS) protocol (Peterson et al. 2014). The GBS approach based on genome reduction for the construction of multiplexed reduced representation libraries does not require reference genome for SNP marker discovery and genotyping (Elshire et al. 2011; Fu and Peterson 2011; Peterson et al. 2012; Peterson et al. 2014). The genotypes, restriction enzymes used during library preparation, as well as sequencing platforms differed in each of the experiments. The first project, 'Genotyping-by-sequencing enhances genetic diversity analysis of crested wheatgrass [*Agropyron cristatum* (L.) Gaertn.]', utilized two restriction enzymes *Pst*I and *Msp*I (New England Biolabs, Whitby, ON, Canada) for library preparation and Illumina MiSeq Instrument for sequencing. Likewise, the second project, 'Development of a high-density linkage map of crested wheatgrass [*Agropyron cristatum* (L.) Gaertn.] using genotyping-by-sequencing', utilized two different

enzyme combinations *HinfI* and *HpyCH4IV* (New England Biolabs, Whitby, ON, Canada), and *Pst*I and *Msp*I (New England Biolabs, Whitby, ON, Canada) separately for library preparation and sequencing with the Illumina HiSeq2500 instrument. Whereas, the third project, 'Accelerating breeding of crested wheatgrass [*Agropyron cristatum* (L.) Gaertn.] through genotyping-by-sequencing and genomic selection', utilized two restriction enzymes *Pst*I and *Msp*I (New England Biolabs, Whitby, ON, Canada) for library preparation and sequencing with Illumina HiSeq2500 instrument. The differences in the library preparation and sequencing instruments resulted in generation of sequence reads that differ in their length and number. Miseq sequencing generated reads that are 250 bp in length, while, the read length from Hiseq2500 sequencing were 126 bp. As expected, the number of reads generated with Hiseq sequencing were higher than those obtained for Miseq. In the second project we implemented two separate enzyme combinations of restriction enzymes (*HinfI* and *HpyCH4IV*, and *Pst*I and *Msp*I) to increase the number of SNPs generated for linkage mapping. The enzyme combination *HinfI* and *HpyCH4IV* yielded about 1.8 times more SNPs in the Fairway mapping population and about 1.5 times more SNPs in the Parkway mapping population than *Pst*I and *Msp*I enzyme combination, similar to the results reported by Fu et al. (2016). The UNEAK in combination with the Haplotag pipeline was used for SNP generation in the first study and the third study, whereas, ANGSD was used in the second study. The UNEAK and the Haplotag pipeline were able to generate SNP markers in tetraploid and mixture of diploid and tetraploid. The bioinformatic pipeline ANGSD was straight forward and generated SNP information for the diploids used in the second study. The results from these studies indicated the potential of GBS application in generation of genome-wide SNP markers. The SNP markers generated could potentially be utilized for subsequent downstream analysis such as genetic diversity and population structure studies, linkage mapping, and genomic selection.

6.2. Genetic diversity and population structure analysis

Genetic diversity analysis of 192 genotypes from 12 crested wheatgrass lines was conducted with SNP markers generated by GBS application. Three types of diversity analysis performed at individual levels: STRUCTURE analysis, Neighbor joining clustering and the principal coordinate analysis (PCoA) analysis, were able to distinguish the lines Karabalykskij 202 (from Kazakhstan), PGR 16,830 (from Kazakhstan), Vysokij 9 (from Russia) and S8,959E (selected from Vysokij 9) from other lines. This is also supported by the UPGMA dendrogram based on AMOVA. However, the clusters II, III and IV identified by STRUCTURE analysis were found overlapping with PCoA. The results demonstrated that the germplasms could be clustered and the extent of genetic variation present within and among the materials can be identified based on the SNP markers obtained from GBS application. The distinctness of the four lines could be attributed to their origin and lack of inter-pollination with the Canadian breeding lines. Similarly, the distinctness of Canadian breeding lines S9491 and S9514 is related to their synthesis involving lines from Saskatoon, Canada and Utah, USA. Higher within line SNP variation obtained in this study is as expected for outcrossing species (Hamrick and Godt 1989). Further, the proportion of variance residing in the SNPs revealed that higher SNP densities were able to infer within and among line variance more accurately. Similar results were found in the third project, 'Accelerating breeding of crested wheatgrass [*Agropyron cristatum* (L.) Gaertn.] through genotyping-by-sequencing and genomic selection'. The study materials in the first project consisted of tetraploid lines while the third project consisted of diploid and tetraploid lines of crested wheatgrass. The STRUCTURE analysis (Pritchard et al. 2000; Falush et al. 2007) and PCoA analysis using the R routine AveDissR (Yang and Fu 2017; R development core team 2018) distinguished tetraploids from the diploids. These results serve as valuable resources for crested wheatgrass breeding. However, Canadian cultivars and breeding lines share one or more common parents in the genetic background and have gone through several cycles of recurrent selection. This lowers the genetic diversity, which is evident from the higher inbreeding coefficients (F_{st}

values) for these lines. Pembleton et al. (2018) reported that sub-selection of founder genotypes and the cyclic nature of breeding programs lowers the genetic diversity within a breeding program. In such, the distinct lines which have not yet been utilized in intercrossing with Canadian breeding lines and cultivars provide the opportunity to broaden the genetic base of our breeding population. Rauf et al. (2010), in a review, reported increases in genetic diversity by incorporating plant introductions in a program. Reduction in the genetic variance in breeding lines from recurrent selection could be reversed with introduction of genetic resources (Boller and Veteläinen 2010). This provides new avenues for selection and improvement of traits affected by alleles introduced with new genetic resources. However, there is chance of loss of gene combinations affecting traits of interest and adaptation in the advanced breeding materials obtained through recurrent selection. This could be minimized through introduction of materials that show good adaptation to the target environment and simultaneously increasing the introgression of favorable alleles affecting traits of interest (Boller and Veteläinen 2010).

6.3. Genetic linkage mapping in intraspecific F1 mapping population

Sequence information for the P genome of crested wheatgrass is yet unknown due to its large genome size. Recent advancements in NGS and high-throughput genotyping platforms have greatly facilitated the marker discovery and the development of high-density genetic maps even in forage crops having large and complex genomes without prior sequence information such as perennial ryegrass (Velmurugan et al. 2016), intermediate wheatgrass (Kantarski et al. 2017) and napiergrass (Paudel et al. 2018). For crested wheatgrass, this study demonstrates the potential of GBS to capture genome-wide variation and its use in developing genetic maps even in the mapping population created from an intraspecific cross of two closely related parents. The SNPs were distributed in seven linkage groups, providing the evidence of linkage mapping in a segregating intraspecific F_1 mapping population of crested wheatgrass for the first time. The parents used in the present study to generate the mapping population were genetically related which may have impacted the rate of recombination, potentially

reducing the accuracy of the maps generated (Semagn et al. 2006b). In spite of this, the present study found sufficient segregating loci for the development of linkage maps. The maps provide potential for genetic studies of crested wheatgrass and their improvement. Previously developed maps for *Agropyron* were developed from interspecific crosses. The AFLP and RAPD markers used in the previous mapping studies (Yu et al. 2012) are less accurate owing to instability and complexity of the bands, thus are difficult to transfer between different mapping populations and for map integration with genome sequences. Likewise, SALF-seq used for linkage mapping in *Agropyron* (Zhang et al. 2015) were from less conserved non-coding regions, limiting their application. Linkage mapping of *Agropyron* with a wheat 660K SNP array (Zhou et al. 2018) was able to group 913 SNP markers into seven linkage groups, and identifying collinearity and conserved regions between the P genome and wheat genome based on their homeologous relationship. High-density, high quality linkage maps are fundamental to marker-assisted crop improvement. However, linkage maps developed from interspecific crosses provide partial information and are unable to detect events of chromosomal rearrangements within species (Guo et al. 2012). Linkage maps developed from intraspecific crosses, with highly transferable markers such as GBS will facilitate comparative and evolutionary genomic studies, and marker-assisted selection for the improvement of crested wheatgrass. These maps will also serve towards development of physical maps and reference maps in crested wheatgrass.

6.4. Genomic selection in crested wheatgrass for morphological and quality traits

The genomic selection study demonstrated the potential of GBS application (Peterson et al. 2014) to efficiently genotype crested wheatgrass populations (diploid and tetraploid). The genome wide SNP markers were investigated for their application in genomic selection of morphological and nutritive value traits in crested wheatgrass. Five additive genomic selection models (BayesA, BayesB, BayesCπ, GBLUP and RRBLUP) were implemented with ten-fold cross-validation with the objective to identify the most appropriate statistical model to implement GS in crested wheatgrass breeding. The average prediction

ability of the GS models evaluated in Saskatoon ranged from -0.18 to 0.41 (Table 5.6). For traits evaluated at Swift Current the prediction ability of the GS models ranged from -0.24 to 0.43 (Table 5.7). Moderate prediction ability was obtained for some morphological traits, while prediction ability of GS models was low to negative for some traits. The study materials consisted of genotypes of cultivars and breeding lines that have one or more parents in common and share some level of parentage. Thus, the training population in this study can be referred to as 'far related' as reported by Bassi et al. (2016) where sparely represented allelic combinations preclude proper assessment of epistasis by the GS models, reducing the prediction accuracy. On the other hand, the cultivars and breeding lines have been gone through many cycles of recurrent selection for late maturity, yield and nutrient value traits, which reduces the number of alleles segregating for such traits. The prediction ability for those traits using such training population will be reduced (Bassi et al. 2016). The five GS models evaluated were similar in prediction ability across traits except for prediction ability of RRBLUP for traits LW, PH, FRG in Saskatoon and LW and CP in Swift Current. This could be related to the QTL effects underlying a trait. Single trait GS models were similar in prediction ability in absence of large effect QTLs (Haile et al. 2019). Similar prediction ability for BayesA, BayesB and BayesCπ was reported in a GS study in alfalfa (Jia et al. 2018). Four marker densities corresponding to the missing levels (20%, 30%, 40% and 50%) were used to investigate the effect of marker densities on the prediction ability of GS models. The result showed no or little improvement in prediction ability with increasing marker densities for the traits evaluated (Figure 5.2 –5.5) similar to those reported for alfalfa (Annicchiarico et al. 2015). In this study, the size of the training population was constant for all marker densities. This could have resulted in co-linearity among the markers reducing the accuracy at higher marker densities as reported by Muir (2007), thus might have resulted in little improvement in the prediction ability with increasing marker densities.

The observed moderate prediction ability for the morphological traits demonstrates the ability to implement GS to crested wheatgrass breeding program. Whereas, low to negative prediction abilities

obtained could be improved with refining the training population by increasing the size and composition, refining field trials for accurate phenotyping, using GS models that accommodate environmental covariates, and even utilizing multi-trait prediction models.

6.5. Future directions

The results from this study indicated the potential of GBS application in generation of genome-wide SNP markers that could potentially be utilized for subsequent downstream analysis as genetic diversity and population structure study, Linkage mapping and genomic selection. The crested wheatgrass breeding program could benefit from utilizing the diverse breeding materials specific to breeding objectives. The lines of Asian origin could be incorporated into Canadian breeding materials through intercrossing and broaden the genetic base from which selection for late maturity, yield and quality can be made.

Our linkage map created using an intraspecific F_1 mapping population was the first using SNP markers from a GBS application. Further research with the use of more heterozygous parents and more than two mapping populations could be utilized towards developing a higher-density map as achieved in intermediate wheatgrass by Kantarski et al. (2017). Likewise, use of a different GBS protocol involving different combinations of restriction enzymes other than those reported could increase the number of informative SNPs.

Genomic selection in crested wheatgrass for traits (ADF, LW, PH, PD, TPP and FRG,) with moderate prediction ability suggests that GS has potential in crested wheatgrass breeding for certain traits. However, lower to negative prediction abilities for DM, regrowth scores and forage nutritive value traits highlights that more efficient and accurate phenotyping of the traits with refined field trials, improvement of the design and composition of training populations, and improvement of methods implemented for the imputation of the genotypic data may be required to improve selection efficiency using genomic selectin for important agronomic traits.

REFERENCES

Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., and Moreno, R.F. 1991. Complementary DNA sequencing: expressed sequence tags and human genome project. Science **252**: 1651–6.

Adhikari, L., Lindstrom, O.M., Markham, J., and Missaoui, A.M. 2018. Dissecting key adaptation traits in the polyploid perennial *medicago sativa* using GBS-SNP mapping. Front. Plant Sci. **9**: 934. doi:10.3389/fpls.2018.00934.

Al-Hajaj, N., Peterson, G.W., Horbach, C., Al-Shamaa, K., Tinker, N.A., and Fu, Y.-B. 2018. Genotyping-by-sequencing empowered genetic diversity analysis of Jordanian oat wild relative *Avena sterilis*. Genet. Resour. Crop Evol. **65**: 2069–2082. doi:10.1007/s10722-018-0674-x.

Alm, V., Busso, C.S., Ergon, Å., Rudi, H., Larsen, A., Humphreys, M.W., and Rognli, O.A. 2011. QTL analyses and comparative genetic mapping of frost tolerance, winter survival and drought tolerance in meadow fescue (*Festuca pratensis* Huds.). Theor. Appl. Genet. **123**: 369–382. doi:10.1007/s00122-011-1590-z.

Annicchiarico, P., Nazzicari, N., Li, X., Wei, Y., Pecetti, L., and Brummer, E.C. 2015. Accuracy of genomic selection for alfalfa biomass yield in different reference populations. BMC Genomics **16**: 1–13. doi:10.1186/s12864-015-2212-y.

Asay, K.H. 1986. Breeding Strategies in Crested Wheatgrass. Pages 53–57 *in* K.L. Johnson, ed. Crested wheatgrass: its values, problems and myths. Utah State University, Logan.

Asay, K.H., and Jensen, K.B. 1996. Wheatgrasses. Pages 691–724 *in* L.E. Moser, D.R. Buxton, M.D. Casler. eds. Cool-season forage grasses. Agronomy Monograph no. 34, Chap. 22. ASA-CSSASSSA, Madison, WI, USA.

Asay, K.H., Jensen, K.B., Hsiao, C., and Dewey, D.R. 1992. Probable origin of standard crested wheatgrass, *Agropyron desertorum* Fisch Ex Link, Schultes. Can. J. Plant Sci. **72**: 763–772.

Asghari, A., Agayev, Y., and Fathi, S.A.A. 2007. Karyological study of four species of wheat grass (*Agropyron sp*.). Pakistan J. Biol. Sci. **10**: 1093–1097. doi:10.3923/pjbs.2007.1093.1097.

Baral, K., Coulman, B., Biligetu, B., and Fu, Y.-B. 2018. Genotyping-by-Sequencing enhances genetic diversity analysis of crested wheatgrass [*Agropyron cristatum* (L.) Gaertn.]. Int. J. Mol. Sci. **19**: 2587. doi:10.3390/ijms19092587.

Bassi, F.M., Bentley, A.R., Charmet, G., Ortiz, R., and Crossa, J. 2016. Breeding schemes for the implementation of genomic selection in wheat (*Triticum spp*.). Plant Sci. **242**: 23–36. doi:10.1016/j.plantsci.2015.08.021.

Bates, D., Mächler, M., Bolker, B., and Walker, S. 2014. Fitting linear mixed-effects models using lme4. Journal of Statisticval Software **67 (1)**. doi:10.18637/jss.v067.i01.

Baxter, S.W., Davey, J.W., Johnston, J.S., Shelton, A.M., Heckel, D.G., Jiggins, C.D., and Blaxter, M.L. 2011. Linkage mapping and comparative genomics using next-generation rad sequencing of a non-model organism. PLoS One **6**: e19315. doi:10.1371/journal.pone.0019315.

Bekele, W.A., Wight, C.P., Chao, S., Howarth, C.J., and Tinker, N.A. 2018. Haplotype based genotyping-by-sequencing in oat genome research. Plant Biotechnol. J. **16**: 1452–1463. doi:10.1111/pbi.12888.

Biazzi, E., Nazzicari, N., Pecetti, L., Brummer, E.C., Palmonari, A., Tava, A., and Annicchiarico, P. 2017. Genome-wide association mapping and genomic selection for alfalfa (*Medicago sativa*) forage quality traits. PLoS One **12**: 1–17. doi:10.1371/journal.pone.0169234.

Bolger, A.M., Lohse, M., and Usadel, B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics **30**: 2114–2120. doi:10.1093/bioinformatics/btu170.

Boller, B., and Veteläinen, M. 2010. A state of the art of germplasm collections for forage and turf species. Pages 17–28 *in* Christian Huyghe, ed. Sustainable use of genetic diversity in forage and turf breeding. Springer, New York. doi:10.1007/978-90-481-8706-5.

Botstein, D., White, R.L., Skolnick, M., and Davis, R.W. 1980. Construction of a genetic linkage map in

man using restriction fragment length polymorphisms. Am. J. Hum. Genet. **32**: 314–31. [Online] Available: http://www.ncbi.nlm.nih.gov/pubmed/6247908 [2019 Apr. 4].

Bourke, P.M., Voorrips, R.E., Visser, R.G.F., and Maliepaard, C. 2018. Tools for genetic studies in experimental populations of polyploids. Front. Plant Sci. **9**: 513. doi:10.3389/fpls.2018.00513.

Box, G.E.P., and Cox, D.R. 1964. An analysis of transformations. J. R. Stat. Soc. Ser. B **26**: 211–252. [Online] Available: http://www.jstor.org/stable/2984418.

Brown, T.A. 2002. An Introduction to Genomes. Oxford: Wiley-Liss, New York, USA.

Calus, M.P.L., Meuwissen, T.H.E., de Roos, A.P.W., and Veerkamp, R.F. 2008. Accuracy of genomic selection using different methods to define haplotypes. Genetics **178**: 553–61. doi:10.1534/genetics.107.080838.

Cartwright, D.A., Troggio, M., Velasco, R., and Gutin, A. 2007. Genetic mapping in the presence of genotyping errors. Genetics **176**: 2521–7. doi:10.1534/genetics.106.063982.

Casler, M.D. 1997. Breeding for improved forage quality: potentials and problems. Proc. Eighteenth Int. Grassl. Congr.: 8–19. [Online] Available: https://www.internationalgrasslands.org/files/igc/publications/1997/iii-323.pdf.

Casler, M.D., and Brummer, E.C. 2008. Theoretical expected genetic gains for among-and-within-family selection methods in perennial forage crops. Crop Sci. **48**: 890–902. doi:10.2135/cropsci2007.09.0499.

Casler, M.D., and Vogel, K.P. 1999. Accomplishments and impact from breeding for increased forage nutritional value. Crop Sci. **39**: 12–20. doi:10.2135/cropsci1999.0011183X003900010003x.

Charmet, G., Storlie, E., Oury, F.X., Laurent, V., Beghin, D., Chevarin, L., Lapierre, A., Perretant, M.R., Rolland, B., Heumez, E., Duchalais, L., Goudemand, E., Bordes, J., and Robert, O. 2014. Genome-wide prediction of three important traits in bread wheat. Mol. Breed. **34**: 1843–1852. doi:10.1007/s11032-014-0143-y.

Che, Y., Yang, Y., Yang, X., Li, X., and Li, L. 2015. Phylogenetic relationship and diversity among

Agropyron Gaertn. germplasm using SSRs markers. Plant Syst. Evol. **301**: 163–170. doi:10.1007/s00606-014-1062-4.

Che, Y.H., Li, H.J., Yang, Y.P., Yang, X.M., Li, X.Q., and Li, L.H. 2008. On the use of SSR markers for the genetic characterization of the *Agropyron cristatum* (L.) Gaertn. in Northern China. Genet. Resour. Crop Evol. **55**: 389–396. doi:10.1007/s10722-007-9246-1.

Che, Y.H., Yang, Y.P., Yang, X.M., Li, X.Q., and Li, L.H. 2011. Genetic diversity between ex situ and in situ samples of *Agropyron cristatum* (L.) Gaertn. based on simple sequence repeat molecular markers. Crop Pasture Sci. **62**: 639–644. doi:10.1071/CP11065.

Chen, S.Y., Ma, X., Zhang, X.Q., Huang, L.K., and Zhou, J.N. 2013. Genetic diversity and relationships among accessions of five crested wheatgrass species (Poaceae: *Agropyron*) based on gliadin analysis. Genet. Mol. Res. **12**: 5704–5713. doi:10.4238/2013.November.18.19.

Chen, Z., Wang, B., Dong, X., Liu, H., Ren, L., Chen, J., Hauck, A., Song, W., and Lai, J. 2014. An ultra-high density bin-map for rapid QTL mapping for tassel and ear architecture in a large F2 maize population. BMC Genomics **15**: 433. doi:10.1186/1471-2164-15-433.

Collard, B.C.Y., Jahufer, M.Z.Z., Brouwer, J.B., and Pang, E.C.K. 2005. An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. Euphytica **142**: 169–196. doi:10.1007/s10681-005-1681-5.

Collard, B.C.Y., and Mackill, D.J. 2008. Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Philos. Trans. R. Soc. B Biol. Sci. **363**: 557–572. doi:10.1098/rstb.2007.2170.

Combs, E., and Bernardo, R. 2013. Accuracy of genomewide selection for different traits with constant population size, heritability, and number of markers. Plant Genome **6**: 1–7. doi:10.3835/plantgenome2012.11.0030.

Conaghan, P., and Casler, M.D. 2011. A theoretical and practical analysis of the optimum breeding system

for perennial ryegrass. [Online] Available: https://naldc.nal.usda.gov/download/56392/PDF [2019 Mar. 6].

Copete, A., Moreno, R., and Cabrera, A. 2018. Characterization of a world collection of *Agropyron cristatum* accessions. Genet. Resour. Crop Evol. **65**: 1455–1469. doi:10.1007/s10722-018-0630-9.

Coulman, B.E., and Jefferson, P. 2013. Ninety years of perennial forage grass breeding for the Canadian prairie provinces. Pages 290-292 *in* D.L. Michalk, G.D. Millar, W.B. Badgery, and K.M. Broadfoot, eds. Revitalising grasslands to sustain our communities: Proceedings 22nd International Grassland Congress 15–19 September 2013. New South Wales Australia. doi:10.1071/cpv65n6_fo.

Crossa, J., De Los Campos, G., Pérez, P., Gianola, D., Burgueño, J., Araus, J.L., Makumbi, D., Singh, R.P., Dreisigacker, S., Yan, J., Arief, V., Banziger, M., and Braun, H.-J. 2010. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genet. Soc. Am. **186**: 713–724. doi:10.1534/genetics.110.118521.

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de Los Campos, G., Burgueño, J., González-Camacho, J.M., Pérez-Elizalde, S., Beyene, Y., Dreisigacker, S., Singh, R., Zhang, X., Gowda, M., Roorkiwal, M., Rutkoski, J., and Varshney, R.K. 2017. Genomic selection in plant breeding: methods, models, and perspectives. Trends Plant Sci. **22**: 961–975. doi:10.1016/j.tplants.2017.08.011.

Daetwyler, H.D., Pong-Wong, R., Villanueva, B., and Woolliams, J.A. 2010. The impact of genetic architecture on genome-wide evaluation methods. Genet. Soc. Am.: 1021–1031. doi:10.1534/genetics.110.116855.

Daetwyler, H.D., Swan, A.A., van der Werf, J.H., and Hayes, B.J. 2012. Accuracy of pedigree and genomic predictions of carcass and novel meat quality traits in multi-breed sheep data assessed by cross-validation. Genet. Sel. Evol. **44**: 33. doi:10.1186/1297-9686-44-33.

de Los Campos, G., Hickey, J.M., Pong-Wong, R., Daetwyler, H.D., and Calus, M.P.L. 2013. Whole-

genome regression and prediction methods applied to plant and animal breeding. Genetics **193**: 327–345. doi:10.1534/genetics.112.143313.

de Los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., Weigel, K., and Cotes, J.M. 2009. Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics **182**: 375–385. doi:10.1534/genetics.109.101501.

de los Campos, G., Veturi, Y., Vazquez, A.I., Lehermeier, C., and Pérez-Rodríguez, P. 2015. Incorporating genetic heterogeneity in whole-genome regressions using interactions. J. Agric. Biol. Environ. Stat. **20**: 467–490. doi:10.1007/s13253-015-0222-5.

de Roos, A.P.W., Hayes, B.J., and Goddard, M.E. 2009. Reliability of genomic predictions across multiple populations. Genetics **183**: 1545–53. doi:10.1534/genetics.109.104935.

Dewey, D.R. 1974. Reproduction in crested wheatgrass pentaploids. Crop Sci. 14: 867–872.

Dewey, D.R. 1984. The genomic system of classification as a guide to intergeneric hybridization with the perennial triticeae. Pages 209–279 *in* J.P. Gustafson ed. Gene manipulation in plant improvement. Proc. 16th Stadler Genet. Symposium, Columbia, Mo. Plenum, NY, USA.

Dewey, D.R., and Asay, K.H. 1982. Cytogenetic and taxonomic relationships among three diploid crested wheatgrasses. Crop Sci. **22**: 645–650.

Dong, Y.S., Zhou, R.H., Xu, S.J., Li, L.H., Cauderon, Y., and Wang, R.R. 1992. Desirable characteristics in perennial Triticeae collected in China for wheat improvement. Hereditas **116**: 175–178. doi:10.1111/j.1601-5223.1992.tb00224.x.

Elliott, C.R., and Bolton, J.L. 1970. Licensed varieties of cultivated grasses and legumes. Canada Dep. Agr. Publ. 1405.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S., and Mitchell, S.E. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One **6**. doi:10.1371/journal.pone.0019379.

Evanno, G., Regnaut, S., and Goudet, J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. Mol. Ecol. **14**: 2611–2620. doi:10.1111/j.1365-294X.2005.02553.x.

Excoffier, L., and Lischer, H.E.L. 2010. Arlequin suite ver 3.5. 5 a new series of programs to perform population genetics analyses under Linux and Windows. Molecular Ecol. Resour. **10**: 564–567. doi :10.1111/j.1755-0998.2010.02847.x

Falush, D., Stephens, M., and Pritchard, J.K. 2007. Inference of population structure using multilocus genotype data: Dominant markers and null alleles. Mol. Ecol. Notes **7**: 574–578. doi:10.1111/j.1471-8286.2007.01758.x.

Faville, M.J., Ganesh, S., Cao, M., Jahufer, M.Z.Z., Bilton, T.P., Easton, H.S., Ryan, D.L., Trethewey, J.A.K., Rolston, M.P., Griffiths, A.G., Moraga, R., Flay, C., Schmidt, J., Tan, R., and Barrett, B.A. 2018. Predictive ability of genomic selection models in a multi-population perennial ryegrass training set using genotyping-by-sequencing. Theor. Appl. Genet. **131**: 703–720. doi:10.1007/s00122-017-3030-1.

Faville, M.J., Ganesh, S., Moraga, R., Easton, H.S., Jahufer, M.Z.Z., Elshire, R.E., Asp, T., and Barrett, B.A. 2016. Development of Genomic Selection for Perennial Ryegrass. Pages 139–143 *in* Roldán-Ruiz I., Baert J., Reheul D. eds. Breeding in a World of Scarcity. Springer International Publishing, Cham. doi:10.1007/978-3-319-28932-8_21.

Ferreira, A., Flores Da Silva, M., Da Costa E Silva, L., and Damião Cruz, C. 2006. Estimating the effects of population size and type on the accuracy of genetic maps. Genet. Mol. Biol. **29**: 187–192. [Online] Available: www.sbg.org.br [2019 May 7].

Fiedler, J.D., Lanzatella, C., Edmé, S.J., Palmer, N.A., Sarath, G., Mitchell, R., and Tobias, C.M. 2018. Genomic prediction accuracy for switchgrass traits related to bioenergy within differentiated populations. BMC Plant Biol. **18**: 142. doi:10.1186/s12870-018-1360-z.

Forster, J.W., Jones, E.S., Kölliker, R., Drayton, M.C., Dumsday, J., Dupal, M.P., Guthridge, K.M.,

Mahoney, N.L., van Zijll de Jong, E., and Smith, K.F. 2001. Development and implementation of molecular markers for forage crop improvement. Mol. Breed. Forage Crop.: 101–133.

Francia, E., Tacconi, G., Crosatti, C., Barabaschi, D., Bulgarelli, D., Dall'Aglio, E., and Valè, G. 2005. Marker assisted selection in crop plants. Plant Cell. Tissue Organ Cult. **82**: 317–342. doi:10.1007/s11240-005-2387-z.

Fu, Y.-B., and Peterson, G.W. 2011. Genetic diversity analysis with 454 pyrosequencing and genomic reduction confirmed the eastern and western division in the cultivated barley gene pool. Plant Genome J. **4**: 226. doi:10.3835/plantgenome2011.08.0022.

Fu, Y.-B., Peterson, G.W., and Dong, Y. 2016. Increasing genome sampling and improving SNP genotyping for genotyping-by-sequencing with new combinations of restriction enzymes. doi:10.1534/g3.115.025775.

Fu, Y.B., and Yang, M.H. 2017. Genotyping-by-sequencing and its application to oat genomic research. Pages 169–187 *in* Methods in Molecular Biology. doi:10.1007/978-1-4939-6682-0_13.

Gamal El-Dien, O., Ratcliffe, B., Klápště, J., Chen, C., Porth, I., and El-Kassaby, Y.A. 2015. Prediction accuracies for growth and wood attributes of interior spruce in space using genotyping-by-sequencing. BMC Genomics **16**: 370. doi:10.1186/s12864-015-1597-y.

Ganal, M.W., Altmann, T., and Röder, M.S. 2009. SNP identification in crop plants. Curr. Opin. Plant Biol. **12**: 211–217. doi:10.1016/J.PBI.2008.12.009.

Gaut, B.S., and Long, A.D. 2003. The lowdown on linkage disequilibrium. Plant Cell **15**: 1502–1506. doi:10.1105/tpc.150730

Gianola, D., de los Campos, G., Hill, W.G., Manfredi, E., and Fernando, R. 2009. Additive genetic variability and the Bayesian alphabet. Genetics **183**: 347–363. Genetics. doi:10.1534/genetics.109.103952.

Goddard, M. 2009. Genomic selection: prediction of accuracy and maximisation of long term response.

Genetica **136**: 245–257. doi:10.1007/s10709-008-9308-0.

Goonetilleke, S.N., March, T.J., Wirthensohn, M.G., Arús, P., Walker, A.R., and Mather, D.E. 2018. Genotyping by sequencing in almond: SNP discovery, linkage mapping, and marker design. G3 Genes|Genomes|Genetics **8**: 161. doi:10.1534/G3.117.300376.

Gordon, A. 2010. FASTX-toolkit. [Online] Available: http://hannonlab.cshl.edu/fastx_toolkit/index.html. [2017 June 18].

Grinberg, N.F., Lovatt, A., Hegarty, M., Lovatt, A., Skøt, K.P., Kelly, R., Blackmore, T., Thorogood, D., King, R.D., Armstead, I., Powell, W., and Skøt, L. 2016. Implementation of genomic prediction in *Lolium perenne* (L.) breeding populations. Front. Plant Sci. **7**: 133. doi:10.3389/fpls.2016.00133.

Guo, Y., Khanal, S., Tang, S., Bowers, J.E., Heesacker, A.F., Khalilian, N., Nagy, E.D., Zhang, D., Taylor, C.A., Stalker, H.T., Ozias-Akins, P., and Knapp, S.J. 2012. Comparative mapping in intraspecific populations uncovers a high degree of macrosynteny between A- and B-genome diploid species of peanut. BMC Genomics **13**: 608. doi:10.1186/1471-2164-13-608.

Habier, D., Fernando, R.L., and Dekkers, J.C.M. 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics. doi:10.1534/genetics.107.081190.

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

Hackett, C.A., and Broadfoot, L.B. 2003. Effects of genotyping errors, missing values and segregation distortion in molecular marker data on the construction of linkage maps. Heredity (Edinb). **90**: 33–38. doi:10.1038/sj.hdy.6800173.

Haile, T.A., Heidecker, T., Wright, D., Neupane, S., Ramsay, L., Vandenberg, A., and Bett, K.E. 2019. Genomic selection for lentil breeding: empirical evidence. bioRxiv: 608406. Cold Spring Harbor Laboratory. doi:10.1101/608406.

Hamrick, J.L., and Godt, M.J.W. 1989. Allozyme diversity in plant species. Pages 43–63 *in* B.S. Brown,

A.H.D., Clegg, M.T., Kahler, A.L., Weir, ed. Plant population genetics, breeding, and genetic resources. Sinauer Associates Inc., Sunderland, MA, USA. [Online] Available: https://www.cabdirect.org/cabdirect/abstract/19901612624 [2019 Mar. 28].

Hao, Y., Wang, H., Yang, X., Zhang, H., He, C., Li, D., Li, H., Wang, G., Wang, J., and Fu, J. 2018. Genomic prediction using existing historical data contributing to selection in biparental populations: a study of kernel oil in maize. Plant Genome **12**: 1–9. doi:10.3835/plantgenome2018.05.0025.

Hayes, B. 2007. QTL mapping, mas, and genomic selection. [Online] Available: https://www.ans.iastate.edu/files/page/files/notes.pdf [2019 Apr. 22].

Hayes, B., and Goddard, M. 2010. Genome-wide association and genomic selection in animal breeding. Genome **53**: 876–883. doi:10.1139/G10-076.

Hayes, B.J., Cogan, N.O.I., Pembleton, L.W., Goddard, M.E., Wang, J., Spangenberg, G.C., and Forster, J.W. 2013. Prospects for genomic selection in forage plant species. Plant Breed. **132**: 133–143. doi:10.1111/pbr.12037.

He, J., Zhao, X., Laroche, A., Lu, Z.-X., Liu, H., and Li, Z. 2014. Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. Front. Plant Sci. **5**: 484. Frontiers. doi:10.3389/fpls.2014.00484.

Heffner, E.L., Jannink, J.-L., and Sorrells, M.E. 2011a. Genomic selection accuracy using multifamily prediction models in a wheat breeding program. Plant Genome **4**: 65. doi:10.3835/plantgenome2010.12.0029.

Heffner, E.L., Lorenz, A.J., Jannink, J.L., and Sorrells, M.E. 2010. Plant breeding with genomic selection: Gain per unit time and cost. Crop Sci. **50**: 1681–1690. doi:10.2135/cropsci2009.11.0662.

Heffner, E.L., Sorrells, M.E., and Jannink, J.L. 2009. Genomic selection for crop improvement. Crop Sci. **49**: 1–12. doi:10.2135/cropsci2008.08.0512.

Heffner, L.E., Jannink, J.-L., Iwata, H., Souza, E., and Sorrells Mark E. 2011b. Genomic selection accuracy

for grain quality traits in biparental wheat populations. Crop Sci. **51**. doi:10.2135/cropsci2011.05.0253.

Henderson, D.C., and Naeth, M.A. 2005. Multi-scale impacts of crested wheatgrass invasion in mixed-grass prairie. Biol. Invasions **7**: 639–650. doi:10.1007/s10530-004-6669-x.

Heslot, N., Yang, H.-P., Sorrells, M.E., and Jannink, J.-L. 2012. Genomic selection in plant breeding: A comparison of models. Crop Sci. **52**: 146. doi:10.2135/cropsci2011.06.0297.

Holland, J.B., W., N., and C., C. 2003. Estimating and interpreting heritability for plant breeding: An update. Plant Breeding Rev. **22**:109–112 . doi:10.1002/9780470650202.ch2.

Howard, R., Carriquiry, A.L., and Beavis, W.D. 2014. Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. G3: Genes, Genomes, Genetics **4:** 1027–1046. doi:10.1534/g3.114.010298.

Huang, X., Feng, Q., Qian, Q., Zhao, Q., Wang, L., Wang, A., Guan, J., Fan, D., Weng, Q., Huang, T., Dong, G., Sang, T., and Han, B. 2009. High-throughput genotyping by whole-genome resequencing. Genome Res. **19**: 1068–1076. doi:10.1101/gr.089516.108.

Huang, Y.F., Poland, J.A., Wight, C.P., Jackson, E.W., and Tinker, N.A. 2014. Using Genotyping-By-Sequencing (GBS) for genomic discovery in cultivated oat. PLoS One **9**. doi:10.1371/journal.pone.0102448.

Hull, G.J., and Klomp, A.C. 1966. Longevity of crested wheatgrass in the sagebrush-grass type in southern Idaho. J. Range Manag. **19**: 5–11.

Hung, H., Browne, C., Guill, K., Coles, N., Eller, M., Garcia, A., Lepak, N., Melia-Hancock, S., Oropeza-Rosas, M., Salvo, S., Upadyayula, N., Buckler, E.S., Flint-Garcia, S., McMullen, M.D., Rocheford, T.R., and Holland, J.B. 2012. The relationship between parental genetic or phenotypic divergence and progeny variation in the maize nested association mapping population. Heredity (Edinb). **108**: 490–499. doi:10.1038/hdy.2011.103.

Hussain, W., Baenziger, P.S., Belamkar, V., Guttieri, M.J., Venegas, J.P., Easterly, A., Sallam, A., and Poland, J. 2017. Genotyping-by-sequencing derived high-density linkage map and its application to qtl mapping of flag leaf traits in bread wheat. Sci. Rep. **7**: 16394. doi:10.1038/s41598-017-16006-z.

Isik, F., Bartholomé, J., Farjat, A., Chancerel, E., Raffin, A., Sanchez, L., Plomion, C., and Bouffier, L. 2016. Genomic selection in maritime pine. Plant Sci. **242**: 108–119. doi:10.1016/j.plantsci.2015.08.006.

Jan, H.U., Abbadi, A., Lücke, S., Nichols, R.A., and Snowdon, R.J. 2016. Genomic prediction of testcross performance in canola (*Brassica napus*). PLoS One **11**: e0147769. doi:10.1371/journal.pone.0147769.

Jannink, J.-L., Lorenz, A.J., and Iwata, H. 2010. Genomic selection in plant breeding: From theory to practice. Brief. Funct. Genomics **9**: 166–177. doi:10.1093/bfgp/elq001.

Jarquín, D., Kocak, K., Posadas, L., Hyma, K., Jedlicka, J., Graef, G., and Lorenz, A. 2014. Genotyping by sequencing for genomic prediction in a soybean breeding population. BMC Genomics **15**: 740. BioMed Central. doi:10.1186/1471-2164-15-740.

Jia, C., Zhao, H., Wang, Z., Liu, G., Zhao, F., Han, J., and Wang, X. 2018. Genomic prediction for 25 agronomic and quality traits in alfalfa (*Medicago sativa*). Front. Plant Sci. **9**: 1–7. doi:10.3389/fpls.2018.01220.

Kantarski, T., Larson, S., Zhang, X., DeHaan, L., Borevitz, J., Anderson, J., and Poland, J. 2017. Development of the first consensus genetic map of intermediate wheatgrass (*Thinopyrum intermedium*) using genotyping-by-sequencing. Theor. Appl. Genet. **130**: 137–150. doi:10.1007/s00122-016-2799-7.

Kirk, L.E. 1932. Crested wheatgrass. Saskatchewan Agr. Exten. Bull. 54. 24 p.

Knol, E.F., Nielsen, B., and Knap, P.W. 2016. Genomic selection in commercial pig breeding. Anim. Front. **6**: 15–22. doi:10.2527/af.2016-0003.

Knowles, R.P., and Buglass, E. 1982. Crested wheatgrass. Agriculture Canada Publication 1295.

Korneliussen, T.S., Albrechtsen, A., and Nielsen, R. 2014. ANGSD: Analysis of next generation

sequencing data. BMC Bioinformatics **15**: 356. [Online] Available: http://www.biomedcentral.com/1471-2105/15/356 [2019 May 21].

Kumar, S., Stecher, G., and Tamura, K. 2016. MEGA7: Molecular Evolutionary genetics analysis version 7.0 for bigger datasets. Mol. Biol. Evol. **33**: 1870–1874. doi:10.1093/molbev/msw054.

Kutner, M.H., Nachtsheim, C.J., Neter, J., and Li, W. 2005. Applied linear statistical models (5th ed). The McGraw-Hill Companies, Inc., NY, USA.

Lande, R., and Thompson, R. 1990. Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics **124**: 743-756. doi:10.1046/j.1365-2540.1998.00308.x.

Lander, E.S. 1996. The new genomics: Global views of biology. Science **274**: 536–9. doi:10.1126/SCIENCE.274.5287.536.

Li, J.X., Yun, J.F., and Alts, B.D. 2004. Genetic diversity of *Agropyron cristatum* on cytology. Grassland of China, 26: 12-15.

Li, P., Bhattarai, S., Peterson, G., Coulman, B., Schellenberg, M., Biligetu, B., Fu, Y.-B., Li, P., Bhattarai, S., Peterson, G.W., Coulman, B., Schellenberg, M.P., Biligetu, B., and Fu, Y.-B. 2018. Genetic diversity of northern wheatgrass (*Elymus lanceolatus ssp. lanceolatus*) as revealed by genotyping-by-sequencing. Diversity **10**: 23. doi:10.3390/d10020023.

Li, P., Biligetu, B., Coulman, B.E., Schellenberg, M., and Fu, Y.B. 2017. Genotyping-by-sequencing data of 272 crested wheatgrass (*Agropyron cristatum*) genotypes. Data Br. **15**: 401–406. doi:10.1016/j.dib.2017.09.030.

Li, X., Wei, Y., Acharya, A., Hansen, J.L., Crawford, J.L., Viands, D.R., Michaud, R., Claessens, A., and Brummer, E.C. 2015. Genomic prediction of biomass yield in two selection cycles of a tetraploid alfalfa breeding population. Plant Genome **8**: 1–10. doi:10.3835/plantgenome2014.12.0090.

Li, X., Wei, Y., Acharya, A., Jiang, Q., Kang, J., and Brummer, E.C. 2014. A saturated genetic linkage map of autotetraploid alfalfa (*Medicago sativa* L.) developed using genotyping-by-sequencing is highly

syntenous with the *Medicago truncatula* genome. G3 (Bethesda). **4**: 1971–9. doi:10.1534/g3.114.012245.

Lin, Z., Hayes, B.J., and Daetwyler, H.D. 2014. Genomic selection in crops, trees and forages: A review. Crop Pasture Sci. **65**: 1177–1191. doi:10.1071/CP13363.

Lipka, A.E., Lu, F., Cherney, J.H., Buckler, E.S., Casler, M.D., and Costich, D.E. 2014. Accelerating the switchgrass (*Panicum virgatum* L.) breeding cycle using genomic selection approaches. PLoS One **9**: 1–7. doi:10.1371/journal.pone.0112227.

Looman, J., and Heinrichs, D.. 1973. Stability of crested wheatgrass pastures under long-term pasture use. Can. J. Plant Sci. **53**: 501–506.

Lorenz, A.J., Chao, S., Asoro, F.G., Heffner, E.L., Hayashi, T., Iwata, H., Smith, K.P., Sorrells, M.E., and Jannink, J.-L. 2011. Genomic selection in plant breeding: Knowledge and prospects. Adv. Agron. **110**: 77–123. doi:10.1016/B978-0-12-385531-2.00002-5.

Lorenz, R.J. 1986. Introduction and early use of crested wheatgrass in the northern great plains. Pages 9–20 *in*: K.L. Johnson ed. Crested wheatgrass: Its values, problems and myths. Utah State University, Logan, USA. [Online] Available: https://globalrangelands.org/sites/globalrangelands.org/files/dlio/38004/cwgsp-1983-09-20.pdf [2019 Jul. 2].

Lu, F., Lipka, A.E., Glaubitz, J., Elshire, R., Cherney, J.H., Casler, M.D., Buckler, E.S., and Costich, D.E. 2013. Switchgrass genomic diversity, ploidy, and evolution: Novel insights from a network-based SNP discovery protocol. PLoS Genet. **9**. doi:10.1371/journal.pgen.1003215.

Massman, J.M., Gordillo, A., Lorenzana, R.E., and Bernardo, R. 2013. Genomewide predictions from maize single-cross data. Theor. Appl. Genet. **126**: 13–22. doi:10.1007/s00122-012-1955-y.

Mellish, A., Coulman, B., and Ferdinandez, Y. 2002. Genetic relationships among selected crested wheatgrass cultivars and species determined on the basis of AFLP markers. Crop Sci. **42**: 1662–1668. doi:10.2135/cropsci2002.1662.

Metzker, M.L. 2010. Sequencing technologies — the next generation. Nat. Rev. Genet. **11**: 31–46. doi:10.1038/nrg2626.

Meuwissen, T.H.E., Hayes, B.J., and M. E. Goddard 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics **157**: 1819–1829. [Online] Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1461589/pdf/11290733.pdf [2019 Mar. 6].

Momen, M., Mehrgardi, A., Sheikhi, A., Kranis, A., Tusell, L., Morota, G., Rosa, G.J.M., and Gianola, D. 2018. Predictive ability of genome-assisted statistical models under various forms of gene action. Sci. Rep. **8**. doi:10.1038/s41598-018-30089-2.

Mott, I.W., Larson, S.R., Jones, T.A., Robins, J.G., Jensen, K.B., and Peel, M.D. 2011. A molecular genetic linkage map identifying the St and H subgenomes of *Elymus* (Poaceae: Triticeae) wheatgrass. Genome: 819–828. doi:10.1139/G11-045.

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**: 342–355. doi:10.1111/j.1439-0388.2007.00700.x.

Muñoz, P.R., Resende, M.F.R., Gezan, S.A., Resende, M.D.V., De Los Campos, G., Kirst, M., Huber, D., and Peter, G.F. 2014. Unraveling additive from nonadditive effects using genomic relationship matrices. Genetics: **198**: 1759–1768. doi:10.1534/genetics.114.171322.

Ober, U., Ayroles, J.F., Stone, E.A., Richards, S., Zhu, D., Gibbs, R.A., Stricker, C., Gianola, D., Schlather, M., Mackay, T.F.C., and Simianer, H. 2012. Using whole-genome sequence data to predict quantitative trait phenotypes in drosophila melanogaster. PLoS Genet. **8**: e1002685. doi:10.1371/journal.pgen.1002685.

Ochoa, V., Madrid, E., Said, M., Rubiales, D., and Cabrera, A. 2015. Molecular and cytogenetic characterization of a common wheat-*Agropyron cristatum* chromosome translocation conferring resistance to leaf rust. Euphytica **201**: 89–95. doi:10.1007/s10681-014-1190-5.

Paran, I., and Michelmore, R.W. 1993. Development of reliable PCR-based markers linked to downy mildew resistance genes in lettuce. Theor. Appl. Genet. **85**: 985–993. doi:10.1007/BF00215038.

Paterson, A.H., Lander, E.S., Hewitt, J.D., Peterson, S., Lincoln, S.E., and Tanksley, S.D. 1988. Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. Nature **335**: 721–726. [Online] Available: https://www.nature.com/articles/335721a0.pdf [2019 May 7].

Paudel, D., Kannan, B., Yang, X., Harris-Shultz, K., Thudi, M., Varshney, R.K., Altpeter, F., and Wang, J. 2018. Surveying the genome and constructing a high-density genetic map of napiergrass (*Cenchrus purpureus* Schumach). Sci. Rep. **8**: 14419. doi:10.1038/s41598-018-32674-x.

Pembleton, L.W., Inch, C., Baillie, R.C., Drayton, M.C., Thakur, P., Ogaji, Y.O., Spangenberg, G.C., Forster, J.W., Daetwyler, H.D., and Cogan, N.O.I. 2018. Exploitation of data from breeding programs supports rapid implementation of genomic selection for key agronomic traits in perennial ryegrass. Theor. Appl. Genet. **131**: 1891–1902. doi:10.1007/s00122-018-3121-7.

Pérez, P., and de Los Campos, G. 2014. Genome-wide regression and prediction with the BGLR statistical package. Genetics **198**: 483–495. doi:10.1534/genetics.114.164442.

Peterson, B.K., Weber, J.N., Kay, E.H., Fisher, H.S., and Hoekstra, H.E. 2012. Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS One **7** (5). e37135. doi:10.1371/journal.pone.0037135.

Peterson, G.W., Dong, Y., Horbach, C., and Fu, Y.-B. 2014. Genotyping-By-Sequencing for plant genetic diversity analysis: A lab guide for snp genotyping. Diversity **6**: 665–680. doi:10.3390/d6040665.

Piepho, H.P., Möhring, J., Melchinger, A.E., and Büchse, A. 2008. BLUP for phenotypic selection in plant breeding and variety testing. Euphytica **161:** 209–228. doi:10.1007/s10681-007-9449-8.

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

(2): e32253. doi:10.1371/journal.pone.0032253.

Poland, J.A., and Rife, T.W. 2012. Genotyping-by-Sequencing for Plant Breeding and Genetics. Plant Genome. **5 (3)**: 92–102. doi:10.3835/plantgenome2012.05.0005.

Pootakham, W., Jomchai, N., Ruang-Areerate, P., Shearman, J.R., Sonthirod, C., Sangsrakru, D., Tragoonrung, S., and Tangphatsornruang, S. 2015. Genome-wide SNP discovery and identification of QTL associated with agronomic traits in oil palm using genotyping-by-sequencing (GBS). Genomics **105**: 288–95. doi:10.1016/j.ygeno.2015.02.002.

Poudel, H.P., Sanciangco, M.D., Kaeppler, S.M., Buell, C.R., and Casler, M.D. 2019. Quantitative trait loci for freezing tolerance in a lowland x upland switchgrass population. Front. Plant Sci. **10**: 372. doi:10.3389/fpls.2019.00372.

Pritchard, J.K., Stephens, M., and Donnelly, P. 2000. Inference of population structure using multilocus genotype data. Genetics **155**: 945–959. doi:10.1111/j.1471-8286.2007.01758.x.

Rabier, C.-E., Barre, P., Asp, T., Charmet, G., and Mangin, B. 2016. On the accuracy of genomic selection. PLoS One **11**: e0156086. doi:10.1371/journal.pone.0156086.

Ramstein, G.P., Evans, J., Kaeppler, S.M., Mitchell, R.B., Vogel, K.P., Buell, C.R., and Casler, M.D. 2016. Accuracy of genomic prediction in switchgrass (*Panicum virgatum* L.) improved by accounting for linkage disequilibrium. G3 (Bethesda). **6**: 1049–62. doi:10.1534/g3.115.024950.

Ratcliffe, B., El-Dien, O.G., Klápště, J., Porth, I., Chen, C., Jaquish, B., and El-Kassaby, Y.A. 2015. A comparison of genomic selection models across time in interior spruce (*Picea engelmannii* × *glauca*) using unordered SNP imputation methods. Heredity (Edinb). **115**: 547–55. doi:10.1038/hdy.2015.57.

Rauf, S., Silva, J., Asif, A., and Naveed, A. 2010. Consequences of plant breeding on genetic diversity. Int. J. Plant Breed. **4**: 1–21.

Resende, M.F.R., Muñoz, P., Acosta, J.J., Peter, G.F., Davis, J.M., Grattapaglia, D., Resende, M.D.V., and Kirst, M. 2012. Accelerating the domestication of trees using genomic selection: Accuracy of

prediction models across ages and environments. New Phytol. **193**: 617–624. doi:10.1111/j.1469-8137.2011.03895.x.

Resende, R.M.S., Casler, M.D., and Resende, M.D.V. 2013. Selection methods in forage breeding: A quantitative appraisal. Crop Sci. **53**: 1925–1936. doi:10.2135/cropsci2013.03.0143.

Resende, R.M.S., Casler, M.D., and Resende, M.D.V. 2014. Genomic selection in forage breeding: Accuracy and methods. Crop Sci. **54**: 143–156. doi:10.2135/cropsci2013.05.0353.

Riedelsheimer, C., Endelman, J.B., Stange, M., Sorrells, M.E., Jannink, J.-L., and Melchinger, A.E. 2013. Genomic predictability of interconnected biparental maize populations. Genetics **194 (2):** 493–503 doi:10.1534/genetics.113.150227.

Rogers, D.L., and Montalvo, A.M. 2004. Genetically appropriate choices for plant materials to maintain biological diversity. [Online] Available: https://www.fs.usda.gov/Internet/FSE_DOCUMENTS/fsbdev3_039080.pdf [2019 Mar. 28].

Rogler, G.A., and Lorenz, R.L. 1983. Crested wheatgrass-early history in the United States. J. Range Manag. **36**: 91–93.

Said, M., Hřibová, E., Danilova, T. V., Karafiátová, M., Čížková, J., Friebe, B., Doležel, J., Gill, B.S., and Vrána, J. 2018. The *Agropyron cristatum* karyotype, chromosome structure and cross-genome homoeology as revealed by fluorescence in situ hybridization with tandem repeats and wheat single-gene probes. Theor. Appl. Genet. **131**: 2213–2227. doi:10.1007/s00122-018-3148-9.

Salikhov, K., Sacomoto, G., and Kucherov, G. 2014. Using cascading bloom filters to improve the memory usage for de Brujin graphs. Algorithms Mol. Biol. **9**: 2. doi:10.1186/1748-7188-9-2.

Salimath, S.S., Oliveira, A.C. de, Bennetzen, J.L., and Godwin, I.D. 1995. Assessment of genome origins and genetic diversity in the genus *Eleusine* with DNA markers. Genome **38**: 757–763. doi:10.1139/g95-096.

Sallam, A.H., Endelman, J.B., Jannink, J.-L., and Smith, K.P. 2015. Assessing genomic selection prediction

accuracy in a dynamic barley breeding population. Plant Genome **8**: 0. doi:10.3835/plantgenome2014.05.0020.

SAS Institute Inc. 2013. SAS Version 9.4. doi:10.1037/a0034781.

Semagn, K., Bjørnstad, Å., and Ndjiondjop, M. 2006a. An overview of molecular marker methods for plants. African J. Biotechnol. **5**: 2540–2568. [Online] Available: https://www.ajol.info/index.php/ajb/article/view/56080 [2019 Apr. 1].

Semagn, K., Bjørnstad, Å., and Ndjiondjop, M.N. 2006b. Principles, requirements and prospects of genetic mapping in plants. African J. Biotechnol. **5**: 2569–2587. [Online] Available: http://www.academicjournals.org/AJB [2019 Apr. 3].

Sharma, H.C., Gill, B.S., and Uyemoto, J.K. 1984. High levels of resistance in agropyron species to barley yellow dwarf and wheat streak mosaic viruses. J. Phytopathol. **110**: 143–147. doi:10.1111/j.1439-0434.1984.tb03402.x.

Shendure, J., and Ji, H. 2008. Next-generation DNA sequencing. Nat. Biotechnol. **26**: 1135–1145. doi:10.1038/nbt1486.

Shinozuka, H., Cogan, N.O., Spangenberg, G.C., and Forster, J.W. 2012. Quantitative Trait Locus (QTL) meta-analysis and comparative genomics for candidate gene prediction in perennial ryegrass (*Lolium perenne* L.). BMC Genet. **13**: 101. doi:10.1186/1471-2156-13-101.

Singh, B.D., and Singh, A.K. 2015. Marker-assisted plant breeding: Principles and practices. Springer India. doi:10.1007/978-81-322-2316-0.

Smoliak, S., Johnston, A., and Lodge, R.W. 1981. Managing crested wheatgrass in pastures. Information Services, Agriculture Canada, K1A 0C7. [Online] Available: https://archive.org/stream/managementofcres00smol/managementofcres00smol_djvu.txt

Spindel, J., Begum, H., Akdemir, D., Virk, P., Collard, B., Redoña, E., Atlin, G., Jannink, J.-L., and 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**: e1004982. doi:10.1371/journal.pgen.1004982.

Stapley, J., Reger, J., Feulner, P.G.D., Smadja, C., Galindo, J., Ekblom, R., Bennison, C., Ball, A.D., Beckerman, A.P., and Slate, J. 2010. Adaptation genomics: the next generation. Trends Ecol. Evol. **25**: 705–712. doi:10.1016/j.tree.2010.09.002.

Statistics Canada 2016. Crops - Hay and field crops. [Online] Available: https://www150.statcan.gc.ca/n1/pub/95-634-x/2017001/article/54904-eng.htm [2019 Jun. 30].

Su, C., Wang, W., Gong, S., Zuo, J., Li, S., and Xu, S. 2017. High density linkage map construction and mapping of yield trait QTLs in maize (*zea mays*) using the genotyping-by-sequencing (GBS) technology. Front. Plant Sci. **8**: 706. doi:10.3389/fpls.2017.00706.

Talukder, S.K., and Saha, M.C. 2017. Toward genomics-based breeding in C3 cool-season perennial grasses. Front. Plant Sci. **8**: 1317. doi:10.3389/fpls.2017.01317.

Tandoh, S. 2019. Characterization of crested wheatgrass germplasms for plant maturity and associated physiological and morphological traits. Master's thesis, University of Saskatchewan, Saskatoon, SK.

Tayeh, N., Klein, A., Le Paslier, M.-C., Jacquin, F., Houtin, H., Rond, C., Chabert-Martinello, M., Magnin-Robert, J.-B., Marget, P., Aubert, G., and Burstin, J. 2015. Genomic prediction in pea: Effect of marker density and training population size and composition on prediction accuracy. Front. Plant Sci. **6**: 941. doi:10.3389/fpls.2015.00941.

Tester, M., and Langridge, P. 2010. Breeding technologies to increase crop production in a changing world. Science **327**: 818–22. doi:10.1126/science.1183700.

Tinker, N.A., Bekele, W.A., and Hattori, J. 2016. Haplotag: Software for haplotype-based genotyping-by-sequencing analysis. G3 **6**: 857–863. doi:10.1534/g3.115.024596.

University of Saskatcehwan 2019. Ecoregions of Saskatchewan. [Online] Available:

http://www.usask.ca/biology/rareplants_sk/root/htm/en/researcher/4_ecoreg.php [2019 Jun. 28].

Van Ooijen, J.W. 2006. JoinMap ® 4, Software for the calculation of genetic linkage maps in experimental populations. Kyazma BV, Wageningen.

Van Ooijen, J.W. 2011. Multipoint maximum likelihood mapping in a full-sib family of an outbreeding species. Genet. Res., Camb **93**: 343–349. doi:10.1017/S0016672311000279.

Vaness, B.M., and Wilson, S.D. 2007. Impact and management of crested wheatgrass (*Agropyron cristatum*) in the northern Great Plains. Can. J. Plant Sci. **87**: 1023–1028. doi:10.4141/CJPS07120.

Varshney, R.K., Roorkiwal, M., and Sorrells, M.E. 2017. Genomic selection for crop improvement: New molecular breeding strategies for crop improvement. Springer International Publishing, Cham, Switzerland. doi:10.1007/978-3-319-63170-7.

Velmurugan, J., Mollison, E., Barth, S., Marshall, D., Milne, L., Creevey, C.J., Lynch, B., Meally, H., McCabe, M., and Milbourne, D. 2016. An ultra-high density genetic linkage map of perennial ryegrass (*Lolium perenne*) using genotyping by sequencing (GBS) based on a reference shotgun genome assembly. Ann. Bot. **118**: 71–87. doi:10.1093/aob/mcw081.

Vining, K.J., Salinas, N., Tennessen, J.A., Zurn, J.D., Sargent, D.J., Hancock, J., and Bassil, N. V. 2017. Genotyping-by-sequencing enables linkage mapping in three octoploid cultivated strawberry families. PeerJ **5**: e3731. doi:10.7717/peerj.3731.

Vitezica, Z.G., Aguilar, I., Misztal, I., and Legarra, A. 2011. Bias in genomic predictions for populations under selection. Genet. Res., Camb. **93**: 357–366. doi:10.1017/S001667231100022X.

Vogel, K.P., and Lamb, J.F.S. 2007. Forage breeding. Pages 1–808 *in*: K.J. Moore, R.F. Barnes, C.J. Nelson, and M. Collins eds. Forages Volume II: The science of grassland agriculture, 6th edition. Wiley-Blackwell.

Vogel, K.P., and Pedersen, J.F. 1993. Breeding systems for cross-pollinated perennial grasses. Plant Breed. Rev. **11**: 251–274. [Online] Available: http://digitalcommons.unl.edu/agronomyfacpub/948 [2019

Apr. 1].

Voorrips, R.E. 2002. MapChart: Software for the graphical presentation of linkage maps and QTLs. J. Hered. **93**: 77–78. doi:10.1093/jhered/93.1.77.

Vos, P., Hogers, R., Bleeker, M., Reijans, M., Lee, T. van de, Hornes, M., Friters, A., Pot, J., Paleman, J., Kuiper, M., and Zabeau, M. 1995. AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res. **23**: 4407–4414. doi:10.1093/nar/23.21.4407.

Wang, L., Wan, Z.Y., Bai, B., Huang, S.Q., Chua, E., Lee, M., Pang, H.Y., Wen, Y.F., Liu, P., Liu, F., Sun, F., Lin, G., Ye, B.Q., and Yue, G.H. 2015. Construction of a high-density linkage map and fine mapping of QTL for growth in Asian seabass. Sci. Rep. **5**: 16358. doi:10.1038/srep16358.

Weir, B.S., and Hill, W.G. 2002. Estimating F-statistics. Annu. Rev. Genet **36**: 721–50. doi:10.1146/annurev.genet.36.

Werner, C.R., Voss-Fels, K.P., Miller, C.N., Qian, W., Hua, W., Guan, C.-Y., Snowdon, R.J., and Qian, L. 2018. Effective genomic selection in a narrow-genepool crop with low-density markers: Asian rapeseed as an example. Plant Genome **11**: 1–12. doi:10.3835/plantgenome2017.09.0084.

White, T.L., Adams, W.T., and Neale, D.B. (eds.) 2007. Forest Genetics - Google Books. CABI Publishing, Cambridge, MA. [Online] Available: https://books.google.ca/books?hl=en&lr=&id=_MKWuexx52YC&oi=fnd&pg=PR7&dq=White,+T.+L.,+Adams,+W.+T.+and+Neale,+D.+B.+(2007)+Forest+genetics.&ots=uvchtgMc2h&sig=6GQDByzv-4jLBVrQ2OQCAaoEapM#v=onepage&q=White%2C T. L.%2C Adams%2C W. T. and Neal [2019 Apr. 15].

Whittaker, J.C., Thompson, R., and Denham, M.C. 2000. Marker-assisted selection using ridge regression. Genet. Res. **75**: 249–252.

Wilkins, P.W., and Humphreys, M.O. 2003.March. Progress in breeding perennial forage grasses for temperate agriculture. The Journal of Agricultural Science **140**(2): 129-150.

doi:10.1017/S0021859603003058.

Williams, J.G.K., Kubelik, A.R., Livak, K.J., Rafalski, J.A., and Tingey, S. V. 1990. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Res. **18**: 6531–6535. Narnia. doi:10.1093/nar/18.22.6531.

Wu, J., Yang, X., Wang, H., Li, H., Li, L., Li, X., and Liu, W. 2006. The introgression of chromosome 6P specifying for increased numbers of florets and kernels from Agropyron cristatum into wheat. Theor. Appl. Genet. **114**: 13–20. doi:10.1007/s00122-006-0405-0.

Würschum, T., Reif, J.C., Kraft, T., Janssen, G., and Zhao, Y. 2013. Genomic selection in sugar beet breeding populations. BMC Genet. **14**: 85. doi:10.1186/1471-2156-14-85.

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., and 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**: 2133–2149. doi:10.1007/s00122-016-2762-7.

Yang, M.-H., and Fu, Y.-B. 2017. AveDissR: An R function for assessing genetic distinctness and genetic redundancy. Appl. Plant Sci. **5**: 1700018. doi:10.3732/apps.1700018.

Yu, X., Li, X., Ma, Y., Yu, Z., and Li, Z. 2012. A genetic linkage map of crested wheatgrass based on AFLP and RAPD markers. Genome **55**: 327–335. doi:https://doi.org/10.1139/g2012-014.

Zeng, F., Biligetu, B., Coulman, B., Schellenberg, M.P., and Fu, Y. 2017a. RNA-Seq analysis of plant maturity in crested wheatgrass (*Agropyron cristatum*). Genes **8**: 291. doi:10.3390/genes8110291.

Zeng, F., Biligetu, B., Coulman, B., Schellenberg, M.P., and Fu, Y.B. 2017b. RNA-Seq analysis of gene expression for floral development in crested wheatgrass (*Agropyron cristatum* L.). PLoS One **12**: 1–21. doi:10.1371/journal.pone.0177417.

Zhang, H., Yin, L., Wang, M., Yuan, X., and Liu, X. 2019. Factors affecting the accuracy of genomic selection for agricultural economic traits in maize, cattle, and pig populations. Front. Genet. **10**: 189.

doi:10.3389/fgene.2019.00189.

Zhang, J., Liu, W., Han, H., Song, L., Bai, L., Gao, Z., Zhang, Y., Yang, X., Li, X., Gao, A., and Li, L. 2015. De novo transcriptome sequencing of *Agropyron cristatum* to identify available gene resources for the enhancement of wheat. Genomics **106**: 129–136. doi:10.1016/j.ygeno.2015.04.003.

Zhang, X., Sallam, A., Gao, L., Kantarski, T., Poland, J., DeHaan, L.R., Wyse, D.L., and Anderson, J.A. 2016. Establishment and optimization of genomic selection to accelerate the domestication and improvement of intermediate wheatgrass. Plant Genome **9**: 1–18. doi:10.3835/plantgenome2015.07.0059.

Zhang, Y., Zhang, J., Huang, L., Gao, A., Zhang, J., Yang, X., Liu, W., Li, X., and Li, L. 2015b. A high-density genetic map for P genome of *Agropyron* Gaertn. based on specific-locus amplified fragment sequencing (SLAF-seq). Planta **242**: 1335–1347. doi:10.1007/s00425-015-2372-7.

Zhao, X., Huang, L., Zhang, X., Wang, J., Yan, D., Li, J., Tang, L., Li, X., and Shi, T. 2016. Construction of high-density genetic linkage map and identification of flowering-time QTLs in orchardgrass using SSRs and SLAF-seq. Sci Rep **6**: 29345. doi:10.1038/srep29345.

Zhou, S., Zhang, J., Che, Y., Liu, W., Lu, Y., Yang, X., Li, X., Jia, J., Liu, X., and Li, L. 2018. Construction of *Agropyron* Gaertn. genetic linkage maps using a wheat 660K SNP array reveals a homoeologous relationship with the wheat genome. Plant Biotechnol. J. **16**: 818–827. doi:10.1111/pbi.12831.

Zlatnik, E. 1999. Agropyron cristatum. In: Fire Effects Information System, [Online]. U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station, Fire Sciences Laboratory (Producer). Available: https://www.fs.fed.us/database/feis/plants/graminoid/agrcri/all.html [2019 Mar. 28].

APPENDICES

Appendix A

The online Supplementary materials in research component 1 comprise of information about how the supplementary materials can be accessed and the contents in them. In brief, the supplementary material is divided into two sections: Section A consists of list of supplementary materials (three files and five zip folders) and they are available online (DOI://10.6084/m9.figshare.7001414). Section B provides detailed procedure for analyzing FASTQ files using UNEAK and HAPLOTAG to generate tag-level SNP data.

Appendix B

The online Supplementary materials in research component 3 comprise of information about how the supplementary materials can be accessed and the contents in them. In brief, the supplementary material is divided into two sections: Section A consists of list of supplementary materials (five files and five zip folders) and they are available online (https://figshare.com/s/a904cc6d0553aafbe3fa). Section B provides detailed procedure for analyzing FASTQ files using UNEAK and HAPLOTAG to generate tag-level SNP data.