Genome-wide association and genomic selection in *Brassica napus* L.

By

Jia Sun

A Thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Department of Plant Science

University of Manitoba

Winnipeg, Manitoba

ACKNOWLEDGEMENTS

Throughout my PhD I have received a great deal of support and assistance.

I would first like to thank my supervisor, Dr. Rob Duncan, whose expertise was invaluable in formulating the research questions and methodology. I want to extend my sincere acknowledgement to members of my committee: Dr. Anita Brûlé-Babel and Dr. Francis Zvomuya, for their guidance throughout my studies. I would like to thank Dr. Genyi Li for his help and kind suggestions. I would also like to express my sincere gratitude to Dr. Mike Domaratzki, with whom I had many valuable discussions on data analysis and the interpretations of the results.

To all current and previous members of the Brassica breeding program for their support. To all supporting staff at the Plant Science Department for their support and help during my PhD. To my friends in the Plant Science Department for going through the joyful and stressful times with me.

I would like to thank NSERC, DL Seeds, Nutrien, Bunge Canada, and all other scholarship providers for the financial support to complete this research.

To Fei, thank you for being patient with my fuzziness and grumpiness during my writing, and taking in my negative thoughts and make them positive. Thank you for believing in me when I didn't. Thank you for all the silly jokes. And to Mochi, our cat, who came to me at a very stressful time during my PhD. Thank you for generously keeping me companied during my writing and keeping me calm with your endless purring.

谢谢许飞，在我写作期间包容我鼓励我，谢谢你愿意帮我消化负面情绪。谢谢你在我不相信自己的时候依然坚定的相信我。谢谢我们的小猫咪 Mochi，在我压力极大的时候来到我的身边，在我写作期间给我无私的陪伴，谢谢你的呼噜声让我保持稳定的情绪。

To my parents who have always been there for me and believed in me all the way since day one. Thank you for all the encouragement throughout my PhD and your unconditional love and support throughout my life that made me who I am today.

感谢我的爸爸妈妈，你们是我生命里最坚强的后盾。谢谢你们在我读博期间对我的鼓励，谢谢你们给予我的所有的无条件的爱和支持，让我得以成为今天的我。

DEDICATION

Dedicated to my parents Yongxia and Wancang, for everything they have taught me.

谨以此文献给我的妈妈爸爸，孙永霞和孙万仓。谢谢你们教给我的一切。

TABLE OF CONTENTS

List of Tables

List of Figures

ABSTRACT

Linkage and association mapping are two of the most common ways to identify genes or quantitative trait loci (QTL) associated with quantitative traits. The identified genes or QTL can then be used in marker-assisted selection (MAS) with various breeding methodologies. Although MAS has gained great success in improving traits controlled by fewer genes or QTL with large effects, its application is limited in improving traits controlled by many loci with small effects. Genomic selection (GS), as a variant of MAS, was proposed to address this issue by utilizing markers along the whole genome instead of only focusing on the major-effect markers. Currently, the application of GS in *Brassica napus* L. breeding is at a preliminary stage. Therefore, this research was conducted to explore the potential of applying GS in *B. napus* breeding. Chapter 3 demonstrated the application of genome-wide association mapping (GWAS) in *B. napus* on important agronomic and seed quality traits. In total, 141 significant MTAs were detected. Thirty candidate genes had been previously identified in *B. napus* associated with abiotic stress responses and pathogen infection. Chapter 4 investigated the factors that could affect GS prediction accuracies in hybrid *B. napus* including training population (TP) size and composition, marker density and the choice of GS model. The prediction accuracy significantly improved by combining 91 parents and 345 hybrids in the TP, indicating the composition and size of the TP are crucial to GS performance. Higher marker density did not necessarily increase the prediction accuracy, which was possibly due to the high relatedness among the individuals in the TP and test population. Chapter 5 explored the application of GWAS-guided GS and the prediction accuracies were compared across different traits, marker sets and GS models. Compared to conventional GS, GWAS-guided GS showed improvements in prediction accuracy, yet the improvements were not consistent across traits, models or marker sets. In addition, Bayesian models required significantly

longer computational time than penalized approaches (rrBLUP and GBLUP). Taken together, the work presented here demonstrated the potential and impact of GS in assisting and optimizing hybrid *B. napus* breeding programs.

FOREWORD

This thesis has been written in the manuscript style and follows the format guidelines outlined by the Faculty of Graduate Studies at the University of Manitoba. This thesis includes a general introduction, literature review, three manuscripts, a general discussion and future work recommendations. The machine learning section in the literature review (2.5.1.4.2) was written based on a draft completed by Dr. Mike Domaratzki (Department of Computer Science, Western University). The first manuscript "Genome-wide association study (GWAS) of agronomic and seed quality traits in *Brassica napus* L." will be submitted to the journal Scientific Reports or another suitable journal. The second manuscript "Genomic selection of agronomic and seed quality traits in hybrid *Brassica napus* L. based on parametric and machine learning methods" will be submitted to the journal Theoretical and Applied Genetics or another suitable journal or another suitable journal. The machine learning section was collaborative research completed and mostly written by Dr. Mike Domaratzki. The third manuscript "GWAS-guided GS of agronomic and seed quality traits in *Brassica napus* L." will be submitted to the journal Frontiers in Plant Science or another suitable journal.

1. GENERAL INTRODUCTION

Brassica crops are among the earliest crops known to humans (Rakow 2004). Canola (*Brassica napus* L.), following soybean [Glycine *max* (L.) Merr.] and oil palm (*Elaeis guineensis* Jacq.), ranks as the third-largest oilseed crop produced in the world (FAO 2021). Canada has the largest production of canola in the world, followed by the European Union and China (USDA 2020). As the world's leading canola exporter, Canada exports about 90% of its canola to over 50 countries worldwide (Canola Council of Canada 2017). Canola's contribution to Canadian economic output has increased by 35% in the last decade, reaching $29.9 billion annually (LMC International 2020). By 2025, the market demand for Canadian canola production will need to exceed 2,914 kg ha^{-1} (52 bushels/acre) to achieve the 26 Mt production goal (Canola Council of Canada 2014). Therefore, canola breeders have always sought to improve yield and its related traits.

A major contributor to canola yield increases during 2000-2013 was the improvement of canola genetics, particularly the transition from open-pollinated to hybrid cultivars and herbicide-resistant cultivars (Morrison et al. 2016). One of the major challenges in hybrid canola breeding concerns the identification of ideal parental combinations that can lead to larger heterosis in the offspring, contributing to better agronomic performance and superior seed quality (Starmer et al. 1998). Even though numerous previous studies have been undertaken to investigate heterosis in order to improve hybrid performance, the development of heterotic pools in canola has not progressed due to the limited diversity of this new crop species (Bus et al. 2011; Habibur et al. 2015; Jan et al. 2016).

Ideal canola cultivars need to have high seed yield, ease of cultivation for growers (disease resistance, lodging tolerance, shattering tolerance, appropriate maturity, among other agronomic traits) and the appropriate nutritional value for consumers (seed quality traits). Linkage mapping

[quantitative trait loci (QTL) mapping] is one of the most common methods for identifying QTL that are associated with complex traits. Linkage mapping has been routinely implemented in canola breeding for investigating or improving the complex traits mentioned above (Chen et al. 2010; Chen et al. 2007; Fu et al. 2015; Huang et al. 2016; Mei et al. 2009; Nesi et al. 2008; Zhao et al. 2006). However, most of the reported studies focused on the identification of the QTL and only a few reported applying the identified QTL in breeding new genotypes through marker-assisted selection (MAS).

A newer tool, genome-wide association study (GWAS), has emerged in identifying marker-trait associations (MTAs) (Huang and Han 2014), which could assist breeders in understanding the genetic structure underlying complex but important economic traits in canola (Honsdorf et al. 2010). Numerous GWAS studies have been conducted to investigate traits related to seed yield and seed quality (Kittipol et al. 2019; Korber et al. 2016; Li et al. 2016a; Li et al. 2014b; Liu et al. 2016a; Schiessl et al. 2015; Sun et al. 2016a; Sun et al. 2016b; Wang et al. 2018a; Wei et al. 2019a; Wu et al. 2016b; Xiao et al. 2019; Zheng et al. 2017). However, the results of GWAS often differed depending on the difference in population size and structure, as well as the target population composition.

With the availability of low-cost genomic information, a variety of molecular breeding methods are now widely used, including marker-assisted selection and whole genome prediction/selection (GS) (Bernardo 2016). Though considered as a variant of MAS, GS does not focus on major-effect QTL. Instead, GS considers all markers that have small effects on the target traits (Heffner et al. 2009). Many economically important traits are controlled by multiple QTL with small effect, which makes GS advantageous when compared to QTL mapping since genomic selection considers markers along the whole genome, regardless of the amount contributed to a particular

trait (Desta and Ortiz 2014; Goddard and Hayes 2007). Genomic selection has been more commonly applied in maize and soybean breeding (Bernardo 2016), but its use in canola breeding has been somewhat limited. Genomic selection has also been combined with GWAS research to improve the prediction accuracy, where the results from GWAS were implemented into GS as fixed effects (Bian and Holland 2017; Fiedler et al. 2017; Tsai et al. 2020). Overall, GS has shown potential as a plant breeding tool.

The accuracy of GS can be influenced by a number of factors, including training population size and composition, marker density, trait heritability, and model performance (Tan et al. 2017; Zhang et al. 2019a). The challenge breeders encounter when choosing a GS model is that there is no "one size fits all" solution; thus, there is no model that works equally well for all crops, or even the same crop with different traits (Lorenz et al. 2011). These factors need to be carefully considered depending upon the specific purpose when applying GS in breeding.

Genomic selection has shown potential in canola breeding, where a few studies have been reported for investigating traits such as time to flowering (Li et al. 2015a), plant height (Würschum et al. 2014), grain yield and seed glucosinolate content (Jan et al. 2016) and blackleg [*Leptosphaeria maculans* (Desm.) Ces. & de Not.] resistance (Fikere et al. 2018). However, none of these studies examined the seed yield, plant height, seed protein content, seed oil content and seed glucosinolate content at the same time and none of them applied GWAS-guided GS in canola breeding.

Based on genotypic data obtained from high-throughput genotyping (*Brassica* 60K array) and unbalanced phenotypic data for *B. napus* from multi-year, multi-location field trials, three experiments were designed to explore the potential of GWAS in examining the marker-trait associations and that of GS in predicting progeny performance. The objectives of the first experiment, genome-wide association study (GWAS) of agronomic and seed quality traits in

rapeseed (*B. napus* L.), were to examine the structure of the populations of interest (one consisting of parental genotypes and the second consisting of parental genotypes and hybrids), and then to identify the significant MTAs and potential genes associated with the selected agronomic and seed quality traits. This was accomplished using different GWAS models and comparing the performance of the selected GWAS models. The objectives of the second experiment, genomic selection and performance prediction in hybrid *B. napus* L., were to evaluate the effects of training population (TP) and marker density on prediction accuracy, to evaluate the prediction accuracy of GS on selected traits (seed yield, plant height, seed protein content, seed oil content and seed glucosinolate content) and to compare the performance of different GS models. The objectives of the third experiment, GWAS-guided GS of agronomic and seed quality traits in *B. napus* L., were to evaluate and compare the performance of GWAS-guided GS through different models and marker sets on the traits mentioned above. The results from this research will offer valuable information on utilizing unbalanced phenotypic data in GWAS and GS. The discovery of significant MTAs and genes will provide insight into improving agronomic performance and seed quality in *B. napus* using GWAS and GS.

2. LITERATURE REVIEW

2.1 Background

2.1.1 Origin of *Brassica napus* L.

Brassica crops are among the most ancient crops known to humans (Rakow 2004). The family Brassicaceae is also known as the mustard family, containing over 3,700 species in 338 genera (Hayward 2011). The "Triangle of U" (Figure 2.1) shows the relationships amongst six economically important *Brassica* species across the world (Rakow 2004). Three allotetraploid species, *B. juncea* L., *B. napus* L., and *B. carinata* A. Braun, are derived from interspecific crosses of the three diploid species *B. rapa* L. (syn. *campestris*), *B. nigra* L., and *B. oleracea* L. (Morinaga 1934; U 1935). Canola (*Brassica napus* L.), also known as oil rapeseed, contains both the AA and CC genomes (AACC, 2n = 38) (Falk 2009; Hayward 2011; Morrison et al. 2016). About 7,500 years ago, hybridization events occurred between *B. rapa* (genome AA, 2n = 20) and *B. oleracea* (genome CC, 2n = 18), and the subsequent allopolyploid led to the formation of *B. napus* (Chalhoub et al. 2014; Falk 2009; Hayward 2011; Wang et al. 2011), which is also confirmed by recent genomic and cytological analysis (Allender and King 2010; Chalhoub et al. 2014; Snowdon et al. 2002). Southern Europe is believed to be the centre of origin of *B. napus*, which was then introduced into Asia in the 1700s (Daun 2011). Currently, *B. napus* is well adapted to growth in a wide range of areas including Canada, the United States, China, Japan, India, Western and Eastern Europe, Australia, Argentina, Chile, South Africa, Egypt and Iran (Daun 2011).

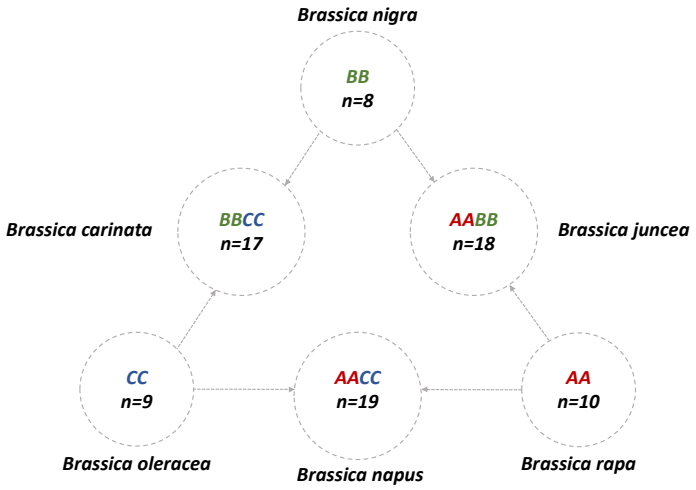


Figure 2.1 Triangle of U showing the relationship among six major cultivated *Brassica* species. The six genomes are denoted as AA (*Brassica rapa*, 2n=20), BB (*Brassica nigra*, 2n=16), CC (*Brassica oleracea*, 2n=18), AABB (*Brassica juncea*, 2n=36), AACC (*Brassica napus*, 2n= 38), and BBCC (*Brassica carinata*, 2n=34). Figure modified from Snowdon (2007).

2.1.2 History and development of canola

Since the end of 19th century, petroleum replaced *B. napus* as the major source for lamp oil (Snowdon et al. 2007). Oilseed rape was first introduced to Canada in 1936 by a Polish immigrant farmer Fred Solvonik, who lived in Shellbrook, Saskatchewan and received *Brassica campestris* (also known as *B. rapa*) seeds from Poland (Bell 1982). During World War II, the Canadian government expanded oilseed rape production in Canada (Morrison et al. 2016). In Western Canada, rapeseed oil was in high demand mainly for steam and marine engine service (Morrison et al. 2016). Due to their relatively high yield, good quality and adaption to the Canadian Prairies, *B. napus* and *B. rapa* quickly became a common oilseed crop in Canada (Daun 2011).

Double-low *B. napus* (low erucic acid and low glucosinolate levels) is a relatively new crop developed in the 1970s by Canadian breeders Drs. Keith Downey (Agriculture and Agri-food Canada Saskatoon) and Baldur R. Stefansson (University of Manitoba) (Stefansson and Kondra 1975). There were two breakthroughs leading to the development of canola. The first was the discovery of an erucic acid-free genotype derived from a German cultivar "Liho" in the early 1960s (Snowdon et al. 2007; Stefansson et al. 1961). This discovery was of great importance in studying the genetic inheritance of the erucic acid content in the seed oil of *B. napus*. Researchers found that low erucic acid content was controlled by two genes with little or no dominance (Downey and Harvey 1963; Harvey and Downey 1964; Stefansson and Hougen 1964). The second breakthrough was the discovery of a low-glucosinolate Polish canola cultivar Bronowski in 1969 (Kondra and Stefansson 1970). Using backcross techniques, these two desired traits (low erucic acid content and low glucosinolate content) were successfully transferred from Liho and Bronowski to Target and Turret (Stefansson and Kondra 1975). After generations of selection, the first double-low

"canola" cultivar was derived and registered as "Tower" in 1974 (Morrison et al. 2016; Stefansson and Kondra 1975).

Today, *B. napus* production provides raw materials for a more diverse range of end products, including livestock feed, biofuel, biodegradable plastics, industrial lubricants, as well as edible oils for human consumption (Jan et al. 2016; Snowdon et al. 2007). Specifically, rapeseed cultivars produce oils with 45% or higher erucic acid (C22:1), which were often used as lamp oil, lubricating oil, and in plastics manufacturing (Raymer 2002). Compared with rapeseed cultivars, canola seed contains less than 2% erucic acid and less than 30 μmol g^{-1} of aliphatic glucosinolate (Eskin and McDonald 1991). These two characteristics make canola oil readily usable for human consumption (Raymer 2002). The byproduct of *B. napus*, seed meal (known as canola meal), has been commonly used in poultry and dairy cow diets (Huhtanen et al. 2011).

2.1.3 Growth and development of *Brassica napus*

Three ecotypes of *B. napus* (winter, semi-winter and spring types) are currently grown worldwide (Zou et al. 2019). Grown predominantly in Europe due to the mild climates, winter-type cultivars are biannual and require vernalization (Watts 2013; Wei et al. 2017). Derived from winter-type cultivars that were adapted to the local climate after being introduced to China, semi-winter type cultivars are also biannual, but require moderate vernalization (Wang et al. 2011; Wei et al. 2017). Spring cultivars, that do not require vernalization are grown in Northern Europe, Australia, and Canada (Wang et al. 2011; Watts 2013).

There are two major stages in the growth and development of a *B. napus* plant: the vegetative stage and reproductive stage (Harper and Berkenkamp 1975). Lancashire et al. (1991) proposed a universal scale (known as the BBCH scale) which characterized the growth and development stages of common crops and weeds. Based on this two-decimal-coded scale, *B. napus* has ten

general growth stages (stages 0 to 9). Within each general stage, more detailed stages are coded with two-digit numbers (Table 2.1). For example, stages 10-19 are the leaf development stages that describe the changes starting from the unfolding of the cotyledons to the development of more than 9 leaves on the main stem. Stages 60-69 describe the flowering stages starting from the beginning of flowering to the end of flowering. Maturity of canola plants varies depending upon the genotype, location, the growing season and the seeding date. The growth and development of *B. napus* is also measured in growing degree days (GDDs), which is an equation used to calculate the accumulated heat over a particular period of time (Derscheid and Lytle 1977). In the Canadian Prairies, *B. napus* needs about 1,500 GDDs to reach the stage ready for swathing, which corresponds to growth stage 8.4 in the BBCH scale (Canola Council of Canada 2013). Typically canola plants need 90 to 120 days after seeding to reach physiological maturity in the western prairies (Canola Council of Canada 2021a; Koscielny 2018).

2.1.4 Factors that affect growth and development of *Brassica napus*

Temperature critically impacts the growth and development of canola (Nuttall et al. 1992). As a cool-season crop (Karamanos et al. 2002; Koscielny et al. 2018; Morrison and Stewart 2002), canola performs the best between 12 and 20 °C (Assefa et al. 2018; Morrison et al. 2016). High temperature during the day (>25°C) and warm night temperatures can adversely affect the development by resulting in male and female sterility, leading to a negative effect on canola yield and quality (Harker et al. 2012; Polowick and Sawhney 1988). Kutcher et al. (2010) showed that a 75.5 kg ha^{-1} yield loss in canola could occur corresponding to every 1°C increase, based on an analysis of canola yield data collected over 34 years in Saskatchewan. Koscielny et al. (2020) also demonstrated that heat stress caused a 14.6% to 18.2% seed yield reduction in canola in all testing locations.

Table 2.1 A summarized biologische bundesanstalt, bundessortenamt and chemical industry (BBCH) scale of *Brassica napus*[1].

Stage	BBCH code	Description	
		Beginning of the stage	End of the stage
0: Germination	00-09	Dry seed	Emergence
1: Leaf development	10-19	Cotyledons completely unfold	≥ 9 leaves
2: Formation of side shoots	20-29	No side shoots	≥ 9 side shoots
3: Stem elongation	30-39	No internodes	≥ 9 visibly extended internodes
4: Development of harvestable vegetative plant parts	40-49	N/A	NA
5: Inflorescence emergence	50-59	Flower buds present, yet still enclosed by leaves	First petals visible, flower buds still closed
6: Flowering	60-69	First flower open	End of flowering
7: Development of fruit	70-79	10% of pods reach final size	Nearly all pods reach final size
8: Ripening	80-89	Seed green, filling pod cavity	Nearly all pods ripe, seeds black and hard
9: Senescence	90-99	Plants dead and dry	Harvested product

[1] Information retrieved from Lancashire et al. (1991).

Another environmental factor affecting *B. napus* growth and development is precipitation. Nuttall et al. (1992) found an increase in *B. napus* grain yield as the total precipitation increased. Harker et al. (2015) examined the effects of crop rotation on *B. napus* yield and observed a higher yield in cooler locations with sufficient, consistent precipitation. Meng et al. (2017) confirmed the positive effects of precipitation on *B. napus* yield based on their analysis of the 1987-2010 period in Saskatchewan. More specifically, a 10% increase in precipitation that occurred during October and April increased seed yield by 0.7% (Meng et al. 2017).

Appropriate field management practices are also crucial in increasing productivity of *B. napus* and maintaining the quality of the final product (Sokólski et al. 2020). Assefa et al. (2018) thoroughly reviewed the management factors that limited *B. napus* yield potential in North America and found that different growing regions had specific managing practices in terms of seeding rate, planting depth, nutrient requirements, crop rotation and tillage. All these specific requirements need to be coordinated to optimize the performance of *B. napus,* which in return, lead to higher yield (Morrison et al. 2016; Sidlauskas and Bernotas 2003). Zheng et al. (2020) also stated that crop rotation was utilized as an effective method in controlling soil-borne diseases in *B. napus* and played a significant role in affecting yield of *B. napus*. As stated by Assefa et al. (2018), to fulfill the current yield gap of 25% to 50% between the actual and potential seed yield, researchers must learn to manage the negative impacts caused by limited resources and a changing climate.

2.1.5 Current canola production in the world and Canada

Canola is currently the third-largest oilseed crop in the world, only after soybean [Glycine *max* (L.) Merr] and oil palm (*Elaeis guineensis* Jacq.) (FAO 2021). In 2018/19, 72.41 Mt of canola/rapeseed were produced worldwide (USDA 2020). Canada is the largest canola producer globally, followed by the European Union and China (USDA 2020). Within Canada, canola is mostly grown in the

western prairies, while there is some production in southern Ontario and Quebec (Canola Council of Canada 2016). *Brassica napus,* also known as Argentine canola, is the most commonly grown canola species in Canada (Canola Council of Canada 2021b). In 2019, Canadian farmers harvested 18.65 Mt of canola from a harvested area of 8.32 million hectares, where Saskatchewan led canola production, followed by Alberta, Manitoba, British Columbia and Ontario (Figure 2.2) (Statistics Canada 2019).

As the largest global exporter of canola, Canada exports about 90% of its canola/rapeseed to more than 50 countries worldwide, including the United States, China, Mexico, Japan, India and the European Union (Canola Council of Canada 2017). In 2018/19, 747.7 tonnes of canola/rapeseed seeds were exported from Canada (Figure 2.3).

2.1.6 *Brassica napus* genome

In 2014, the genome of a European winter oilseed cultivar "Darmor-bzh" (*B. napus*) was assembled (Chalhoub et al. 2014). The assembled A_n and C_n subgenomes are 314.2 Mb and 525.8 Mb, respectively (Chalhoub et al. 2014). Numerous homeologous exchanges between the two subgenomes vary in size (Chalhoub et al. 2014). In 2019, Lu et al. (2019) revealed that the A subgenome was possibly originated from the progenitor of European turnip, while the C subgenome was possibly originated from the common progenitor of cauliflower, broccoli, and Chinese kale.

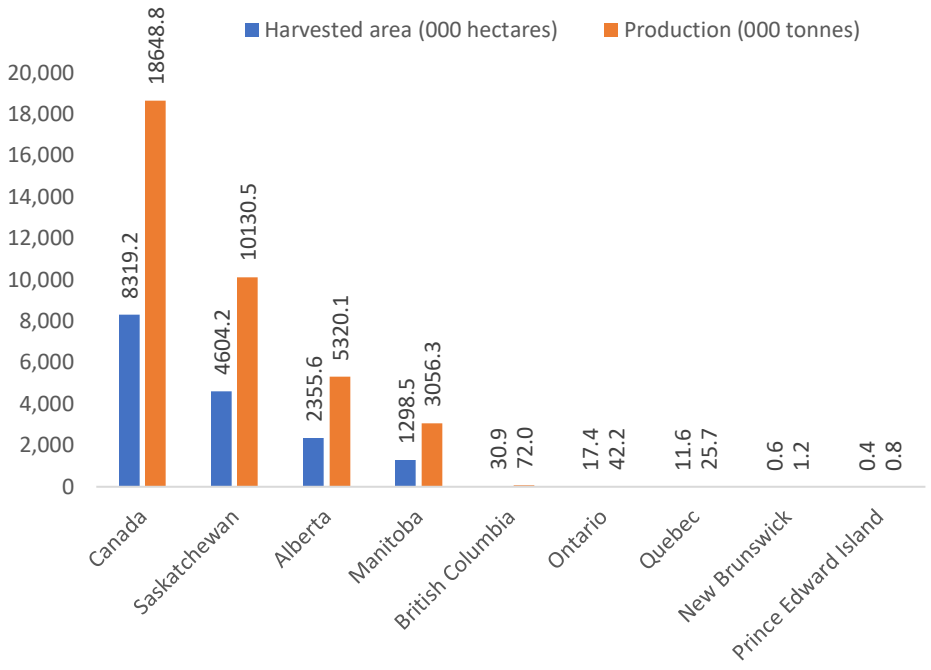


Figure 2.2 Canola/rapeseed harvested area and production in Canada by province in 2019. Data retrieved from Statistics Canada (2019).

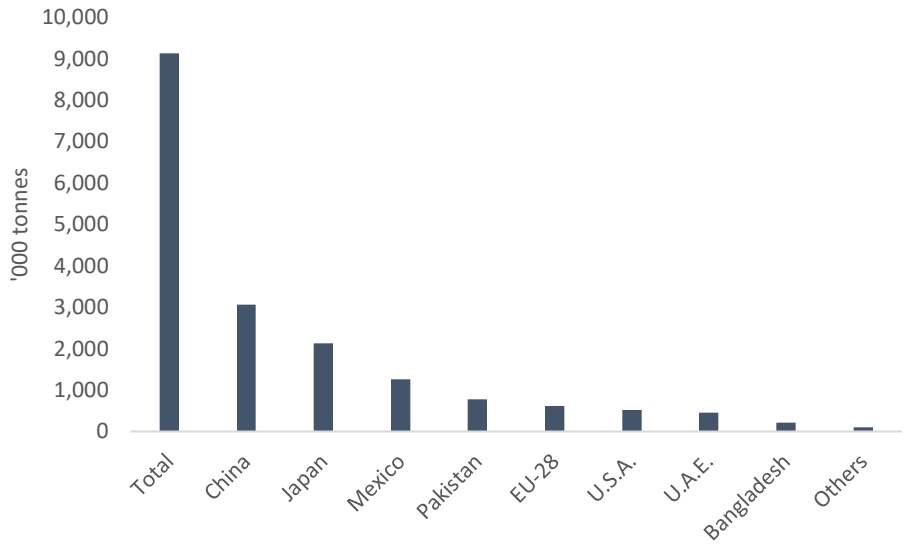


Figure 2.3 Canadian canola/rapeseed seeds exports by country in 2018-2019. Data retrieved from Canola Council of Canada (2019).

Recently, two pan-genomes were constructed for *B. napus*: one based on eight accessions and consisted of winter, semi-winter and spring types of *B. napus* (Song et al. 2020) and another based on 12 different rapeseed genotypes (NRGene 2021). Tettelin et al. (2005) first came up with the concept of pan-genome based on a bacterial species *Streptococcus agalactiae*. The term "pan-genome" was proposed to capture the diversity of a particular species by including multiple accessions from this species and characterizing the consensus genes shared across different accessions of the same species and the genes only presented in some of the accessions (Bayer et al. 2020). The development of a pan-genome addressed the issue that the diversity of a species cannot be sufficiently demonstrated by only one particular genome (Song et al. 2020). With pan-genomes, it is possible to compare multiple individuals of the same species at the genome level and achieve a more thorough understanding of the traits of interest (Bayer et al. 2020).

2.2 Plant breeding and selection

2.2.1 Challenges in conventional breeding

With an increasing global population, a changing climate, decreasing resources and more diverse consumer preferences, breeders are targeting higher yield, more durable disease resistance and abiotic stress tolerance for most crops (Collard and Mackill 2008; Fess et al. 2011). In contrast with the increasing global population and a rising demand for food resources, farmland worldwide has been decreasing for the past 40 years. These contrasting trends increase the difficulty of positively impacting total crop production moving forward (Fess et al. 2011).

Canola is in high demand as the world population grows. During 2000-2013, Canadian land sown to canola increased from 4.9 to 8.3 million hectares, while canola seed yield increased from 1.33 to 2.02 t ha^{-1} (Canola council of Canada 2018; Morrison et al. 2016). By 2025, the market demand for Canadian production will need to exceed 26 Mt. The main factor contributing to canola yield

gain during 2000-2013 was the improvement of canola genetics (i.e. the conversion from open-pollinated cultivars to hybrid and herbicide-tolerant cultivars) (Morrison et al. 2016). Agronomic practices or field management contributed to 13% of the yield gain (Morrison et al. 2016). This means that exploring and improving canola genetics is a key contributor to reach the ultimate yield goal of 2.9 t ha^{-1}, and a production goal of 26 Mt by 2025 (Canola Council of Canada 2014).

Over the last century, conventional breeding has been continuously improving yield in most major crops (Ahmar et al. 2020; Fedoroff 2010; Hickey et al. 2019; Prohens 2011). However, despite the 2.2% increase between 1950 and 1990 globally, the annual increase in the world grain yield was only 1.3% from 1990 to 2011 (Brown 2012). "Upper yield plateau" is a term used to describe the situation in some cereal-producing areas where yield has not increased for some time after a consistent linear increase (Grassini et al. 2013). Rice (*Oryza sativa* L.) and wheat (*Triticum aestivum* L.) yield increases reached a plateau or a potential decline during 1986-1999 in Asia (Pingali and Heisey 2001). In East Asia, where there is 33% of the world's rice production, a yield plateau has been observed (Brown 2012). Statistically significant upper yield plateaus were also observed in wheat production in Northwestern Europe, including the UK, France, Germany, the Netherlands and Denmark, based on a study that analyzed the crop production in 36 countries (Grassini et al. 2013).

Improving the genetics of a crop is the primary goal of a breeding program. The improvement is often quantified as "genetic gain" (ΔG), which is the genetic improvement of a trait in a population over a breeding cycle (Rutkoski 2019). Previous improvement on crop yield has mainly depended on selection of observed phenotypic variation, also known as the phenotypic selection (Fu et al. 2017). Phenotypic selection is effective for selecting traits that are controlled by major alleles with larger effects (Prohens 2011). As a result, the favourable alleles that contributed to quantitative

traits, yet only had minor effects, were harder to capture in the selection process. This reality leads to a significant limitation in conventional breeding (Prohens 2011). Developing new cultivars solely with conventional breeding can also be time-consuming and labour-intensive (Ashkani et al. 2015). This process can take decades depending on the crop (Wieczorek and Wright 2012). Due to these limitations, it is evident that the adoption of newer breeding tools/methodologies is critical to advancing genetic gain.

The gap between the current crop yield increase and potential future demand emphasizes the need for plant breeders to accelerate crop genetic gain (Voss-Fels et al. 2019). Modern technologies that can contribute to accelerating genetic gain can include marker-assisted selection (MAS) and genomic selection (GS), which are reviewed in sections 2.2.2 and 2.5, respectively. These tools have become promising supporting approaches for breeders to continue the increase in crop yield (Collard and Mackill 2008; Huang et al. 1997; Ortiz 1998; Ruttan 1999; Voss-Fels et al. 2019).

2.2.2 Marker-assisted selection

Individuals or species can be distinguished by genetic markers, and these markers can be morphological, biochemical or molecular (Collard et al. 2005). Marker-assisted selection, which is the application of molecular makers in plant breeding, has been applied in plant breeding since the 1990s (Collard and Mackill 2008). Marker-assisted selection uses molecular markers to select desirable genotypes via indirect selection (Ashraf et al. 2012; Desta and Ortiz 2014).

The basic concepts of MAS have been thoroughly reviewed (Collard et al. 2005; Francia et al. 2005; Ribaut and Hoisington 1998; Xu and Crouch 2008). Briefly, there are four steps in MAS. The very first step in the marker development process is to develop a suitable population for linkage mapping which labels the positions of the markers and aligns them based on their relative genetic distances (Collard and Mackill 2008; Collard et al. 2005). Commonly used population

types in constructing a linkage map include F_2, doubled-haploid (DH), backcross (BC) and recombinant inbred line (RIL) populations (Collard and Mackill 2008). The occurrence of genetic recombination events during meiosis leads to segregation in genes or markers (Collard et al. 2005). In the second step, a linkage map that illustrates the relative arrangement of markers is constructed based on the analysis of segregation (Collard et al. 2005). In the third step, the Quantitative Trait Loci (QTL), which are the genomic regions associated with a trait of interest, can be identified from the linkage map by examining the difference between phenotypic means of different genotypic groups (Langridge and Chalmers 2005; Sleper and Poehlman 2006). The next step is to validate the detected QTL across different conditions and genetic backgrounds to confirm that they are reliable in predicting the phenotypes (Collard and Mackill 2008). The validated markers can then be used in MAS, assisting breeders in selecting individuals with the desired traits of interest (Nadeem et al. 2017).

Different types of molecular markers have been used in MAS in plant breeding. These markers include restriction fragment length polymorphism (RFLP) (Botstein et al. 1980), microsatellites or simple sequence repeat (SSR) (Jeffreys et al. 1985; Litt and Luty 1989), random amplified polymorphic DNA (RAPD) (Williams et al. 1990), sequence characterized amplified region (SCAR) (Paran and Michelmore 1993), amplified fragment length polymorphism (AFLP) (Vos et al. 1995), single nucleotide polymorphism (SNP) (Wang et al. 1998) and diversity arrays technology (DarT) (Jaccoud et al. 2001). Today, most high-throughput detection platforms utilize SNPs, the variation that occurs on a single nucleotide at a certain point in the genome (Celton et al. 2010). Mammadov et al. (2012) reviewed the impacts of SNP markers on plant breeding. Essentially, SNP markers are excellent to distinguish individuals within a population and have

become popular due to their genome-wide abundance (John 2012). Details on SNP arrays are discussed in section 2.3.2.

2.2.2.1 Advantages in marker-assisted selection

Marker-assisted selection has brought many opportunities and challenges to breeding and has been practised as a standard tool in breeding programs for some time (Jiang 2015). Compared with phenotypic selection, MAS is more efficient for transferring target genomic regions for traits of interest in a precise manner (Wijerathna 2015). Selection can be made at a very early stage of plant development (Jiang 2015), for example at the seedling stage (Collard and Mackill 2008) or even directly from the seed, prior to planting (Xu and Crouch 2008). Thus, the total number of plants phenotypically evaluated can be reduced, and the cost of greenhouse or field research would decrease accordingly. With the advancements in genotyping technologies and innovative gene-editing technologies, MAS will continue to be a useful breeding tool in future plant breeding (Cobb et al. 2019).

2.2.2.2 Limitations in marker-assisted selection

Although there are many advantages to utilizing MAS in plant breeding, there are several factors that limit additional genetic gains using MAS. Examples of the limitations include: (a) limited proportion of genetic variance that can be explained by markers; (b) limited accuracy of estimated QTL effects; (c) limited ability in shortening the breeding cycle; and (d) limited accuracy of estimated breeding value (EBV) (Goddard and Hayes 2007). A breeder must consider these limitations when utilizing MAS. For example, one has to be cautious of the reliability and accuracy of QTL mapping, or the interaction between QTL and the environment (Cobb et al. 2019; Collard and Mackill 2008). As an indirect selection method, MAS relies on the association between

phenotypic variation and the markers, thus the existence and significance of QTL might vary depending upon the environments the field data were collected from, as well as the population that they were identified from (Cobb et al. 2019). In addition, minor gene effects were often difficult to detect using MAS and association genetics (Desta and Ortiz 2014). Compared to complex traits controlled by many genes, MAS is more effective in selecting qualitative traits or traits controlled by a few major genes (Fu et al. 2017; Jiang 2013; Snowdon and Friedt 2004).

2.2.2.3 The success of marker-assisted selection in some major crops

Marker-assisted selection has been widely used in different breeding programs in various crops. For example, MAS has been effective at improving seed oil and seed protein content, which have been two of the most important traits being studied in soybean (Leite et al. 2016; Patil et al. 2018; Zhang et al. 2016; Zhang et al. 2019b; Zhang et al. 2018b; Zhang et al. 2015b). Marker-assisted selection was also used in maize (*Zea mays* L.) breeding. In maize, molecular markers have been frequently used to investigate and develop high-quality protein maize cultivars (Gibbon and Larkins 2005; Hossain et al. 2018; Krishna et al. 2017). In hybrid maize breeding programs, MAS was utilized to evaluate heterosis (Collard and Mackill 2008). Marker-assisted selection was also used to evaluate and improve yield performance in maize (Abdulmalik et al. 2017; Beyene et al. 2016; Ribaut and Ragot 2007) and has been shown to be more effective than pedigree breeding (Beyene et al. 2016). In wheat, MAS has gained success in developing drought-resistant cultivars (Fleury et al. 2010; Tuberosa and Salvi 2006). The development of disease-resistant wheat cultivars for powdery mildew (*Blumeria graminis* DC. Speer f. sp. *tritici*) resistance (Li et al. 2018a; Ma et al. 2018), stripe rust (*Puccinia striiformis* Westend.) resistance (Miedaner and Korzun 2012) and *Fusarium* head blight (*Fusarium graminearum* Schwabe *sensu lato;* FHB) resistance was enabled by MAS (Anderson 2007; Clark et al. 2016; Liu et al. 2019). In Canada,

nine wheat cultivars developed through MAS have been released, including rust, FHB, wheat midge [*Sitodiplosis mosellana* (Gehin)] and wheat stem sawfly resistant cultivars (Randhawa et al. 2013). In other crops such as barley (*Hordeum vulgare* L.), MAS has been applied in breeding disease resistant cultivars (Miedaner and Korzun 2012). In rice, cooking qualities and abiotic stress tolerance have also been successfully improved by MAS as reviewed by Phing Lau et al. (2016). In upland cotton (*Gossypium hirsutum* L.), MAS has been used in improving fibre quality (Ijaz et al. 2019). Regardless of the disadvantages of MAS, numerous achievements have been obtained through its application, and previously reviewed (Boopathi 2013; Francia et al. 2005; Wijerathna 2015; Xu and Crouch 2008).

2.2.2.4 Marker-assisted selection in *Brassica napus*

The development of DNA marker technologies has enabled the generation of high-density molecular maps, QTL and associated markers, followed by MAS within *Brassica* crop species (Snowdon and Friedt 2004). Numerous studies have been conducted in identifying QTL on different traits in *B. napus* in the past, however, only a few reported cases have applied MAS in improving the genetics of new *B. napus* cultivars. For example, MAS has been successfully applied in improving seed quality traits such as linolenic acid, oleic acid, oil content (Cheung et al. 1998; Jourdren et al. 1996; Rakow et al. 1999; Somers et al. 1999; Spasibionek et al. 2020) and erucic acid content in *B. napus* (Rahman et al. 2008). Yellow seed coat colour was investigated by utilizing MAS in *B. napus* (Liu et al. 2005; Rakow et al. 1999; Somers et al. 2001). Marker-assisted selection has also been applied in selecting for traits related to disease resistance such as white rust resistance [*Albugo candida* (Pers) Kunze] (Cheung et al. 1998; Jourdren et al. 1996; Somers et al. 1999) and clubroot (*Plasmodiophora brassicae* Woronin) resistance (Hirani et al. 2016; Rahman et al. 2014). In addition, self-incompatibility alleles can be screened in early developmental stages

in *B. napus* using genes as markers (Žaludová et al. 2013). Although there have been abundant QTL identified on many traits, the application of MAS primarily focused on traits controlled by fewer genes.

2.3 High-throughput genotyping

High-throughput genotyping is a type of genotyping that can process hundreds to thousands of individuals using hundreds to thousands of markers simultaneously, which quickly became a popular approach to identify SNPs because of the high efficiency and cost effectiveness (Edwards et al. 2013; Singh and Singh 2015). The most commonly used genotyping methods have been reviewed by Scheben et al. (2017). To date, genotyping-by-sequencing (GBS) and SNP arrays are the most widely utilized platforms (You et al. 2018b).

2.3.1 Genotype-by-sequencing and single nucleotide polymorphism arrays

Genotyping-by-sequencing is a highly multiplexed high-throughput genotyping method that utilizes restriction enzymes (RE) to sequence genome subsets and can be modified to apply to any species (Elshire et al. 2011). It is both time and cost-effective since the process of genomic DNA digestion and adaptor ligation all take place within one well, which avoids errors that could occur in transferring samples between wells or plates (Elshire et al. 2011).

A DNA microarray is a solid surface where a collection of nucleic acids is attached to and typically used to measure the relative concentrations of the nucleic acid species in solution through hybridization (Bumgarner 2013). A SNP array is a type of DNA microarray used in detecting polymorphisms (LaFramboise 2009), which has also been widely used in high-throughput genotyping. (You et al. 2018b). The probes on the array were designed to hybridize with the target DNA fragments, which will then produce signals that can be measured and determine the specific alleles of the SNPs (LaFramboise 2009). SNPs scored using this method are reliable and have been

widely applied in plant breeding in different crops (Elbasyoni et al. 2018). For example, in cereal crops such as wheat and oat, various SNP array platforms have been made available such as the Illumina Wheat 9K iSelect SNP array (Cavanagh et al. 2013), the Illumina Wheat 90K iSelect SNP array (Wang et al. 2014), the Axiom® Wheat 660K SNP array (Sun et al. 2020) and the Illumina Oat 6K array etc. All of these platforms provide great opportunities in shortening the breeding cycle and improving economically important traits in crops (Rasheed et al. 2017).

2.3.2 *Brassica napus* Illumina Infinium™ SNP Array

Genotyping *B. napus* can be challenging since *B. napus* has a complex genome due to the historical genome duplications that occurred in its ancestors of *B. rapa* (AA) and *B. oleracea* (CC), as well as the homeologous exchanges between the A- and C- subgenomes of *B. napus* (Chalhoub et al. 2014; Fu et al. 2016). Developed for allotetraploid *B. napus*, the *Brassica napus* Illumina Infinium™ SNP array (60K SNP Chip) became commercially available in 2013 (Mason et al. 2017). It is a high-density SNP array containing 52,157 markers (Mason et al. 2017). This brought new possibilities for genomic selection, as it is a new cost-efficient method that offers high-density, high-throughput whole genome screening for polymorphism in *B. napus* populations (Liu et al. 2013; Snowdon and Iniguez Luy 2012). The initial processing of the genotypic data generated from the 60K SNP array could be completed in GenomeStudio®, software developed by Illumina® for the preliminary data screening, filtration, and cluster adjustment (Illumina Inc 2016).

Since its release, the *B. napus* 60K SNP genotyping array has been applied in numerous studies that covered a wide range of research topics (Mason et al. 2017). The 60K SNP was used in investigating seed quality traits such as oleic acid content (Yao et al. 2020b), oil content (Liu et al. 2016a), erucic acid, stearic acid and glucosinolate content (Zou et al. 2016), as well as disease resistance related traits such as *Sclerotinia sclerotiorum* (Lib.) de Bary stem rot resistance (Wei et

al. 2016) and clubroot resistance caused by *Plasmodiophora Brassicae* Woronin (Fredua-Agyeman et al. 2020). It has also been used in genomic prediction in *B. napus* such as performance prediction of *B. napus* hybrids (Knoch et al. 2021; Werner et al. 2018a), evaluation of the effect of low marker density on prediction accuracy (Werner et al. 2018b) and population effect on prediction accuracy in cross-validation (Werner et al. 2020).

2.4 Genome-wide association study

Genome-wide association study (GWAS) examines the relationship between the traits of interest and the associated genes, QTL or SNPs through various statistical comparisons (Scherer and Christensen 2016). It was initially applied in human genetic studies, mainly focusing on identifying the association between SNP markers and diseases (Klein et al. 2005; Rafalski 2010; Visscher et al. 2012). Currently, GWAS has been widely accepted and utilized in human genetic research, especially for common diseases, which have shown great impact in determining the molecular mechanisms and genetic basis (Huang and Han 2014). Following the establishment in human genetic research of GWAS, the application of GWAS has been extended into non-human species and has become a new tool for animal and plant breeders to improve the genetics of breeding materials (Scherer and Christensen 2016).

The population used clearly impacts the final results of GWAS (Gupta et al. 2014) and various types of plant populations can be used. For example, historical germplasm, bi-parental mapping populations and breeding populations are frequently used (Gupta et al. 2014). Considering that phenotypic data is regularly collected from breeding populations, implementing GWAS based on breeding populations is relatively cost-effective (Gupta et al. 2014). Multi-parental populations are also used in GWAS such as multi-parental advanced generation intercrosses (MAGIC) (Cavanagh et al. 2008) and nested association mapping (NAM) populations (Yu et al. 2008). To construct a

typical MAGIC population, usually four, eight or 16 parents (generation 0) are used to produce F_1 hybrids (generation I) and subsequent crosses are made between the F_1 hybrids to produce generation II. Crosses are then made between generation II individuals that do not share common ancestors followed by selfing until the individuals reach a desired inbred level (Cavanagh et al. 2008). To construct a NAM population, crosses are performed between one parental genotype and several founder parental genotypes instead of performing crosses among the selected parental genotypes, which therefore makes the NAM population desirable for joint linkage association mapping (Gupta et al. 2014).

Genome-wide association study has also been combined with other techniques in finding candidate genes and looking into genetic architecture of a particular trait. These techniques include bulk segregant analysis (BSA) (Gyawali et al. 2019), linkage mapping (Deng et al. 2017; Li et al. 2016d; Li et al. 2015b; Liu et al. 2020a; Wang et al. 2018b), QTL mapping (Ju et al. 2017; Peiffer et al. 2014; Zhao et al. 2018b), and genomic prediction/selection (Bian and Holland 2017; Fiedler et al. 2017; Tsai et al. 2020). With the support from other techniques, GWAS is a more powerful tool in dissecting genetic architecture of traits.

2.4.1 Advantages and limitations of genome-wide association study

Compared to studies that focus solely on QTL identification, GWAS has shown numerous advantages. Genome-wide association study can be directly performed on breeding populations or a collection of germplasm, which is time-efficient as it does not require a population development process (Cortes et al. 2021; Gupta et al. 2014). The resolution from GWAS was often higher than QTL studies using a bi-parental population (Huang and Han 2014). Bi-parental crosses limit the diversity of genetic variations that can be evaluated in the target population (Korte and Farlow 2013). Genome-wide association study has become a great tool in examining complex traits and

examining the genetic variation involved in crop plants (Huang and Han 2014). In fact, there were more advantages in applying GWAS in plants than in human studies (Rafalski 2010). For example, constructing a sizeable population consisting of homozygous individuals and testing its performance on different traits is more feasible in plants compared to humans (Rafalski 2010). These advantages demonstrate why genome-wide association studies can be so efficient and effective in understanding the genetics of a trait.

Genome-wide association studies also have several limitations. False negatives are quite common when the trait of interest is easily affected by environmental effects (Brachi et al. 2011). In most GWAS studies, markers with minor allele frequencies (MAF) under 5% or 10% are removed from association analysis, due to the fact that rare variants can cause an artificial increase in association score estimation, which makes it challenging to identify SNPs that have true associations with the phenotypic variations (Miyagawa et al. 2008). The power of GWAS decreases when the trait of interest is controlled by many alleles with small effects (Korte and Farlow 2013). Researchers also have to be cautious when constructing a GWAS population, since population structure can also affect the power of GWAS or any genome-wide studies (Asoro et al. 2011; Liu and Yan 2019). The stratification underlying the population structure will cause linkage disequilibrium (LD) even though the involved loci are not physically linked, which then may cause increases in the false positive rate and possible inflations in the test results in typical GWAS that use single locus models (Segura et al. 2012). In other words, it may make it difficult to decide whether the identified associations are true or false (Atwell et al. 2010). This also leads to the difficulty in detecting epistasis in a population that consists of unrelated individuals (Liu and Yan 2019). Several statistical methods have been proposed to address this issue including a multi-locus mixed-model (MLMM) (Segura et al. 2012), efficient mixed-model association (EMMA) (Kang et al. 2008),

empirical Bayesian method (EB) (Wang et al. 2016b) and fast-empirical Bayesian linear model (FAST-EB-LMM) (Chang et al. 2019).

2.4.2 Application of genome-wide association study in major crops

Using GWAS in soybean, numerous significant SNPs/genes were associated with yield-related traits such as time to flowering and maturity (Zatybekov et al. 2017), seed weight (Zatybekov et al. 2017; Zhao et al. 2019b), pod dehiscence (Hu et al. 2019), internode number, seed yield per plant, plant height (Assefa et al. 2019; Zatybekov et al. 2017) and seeds per plant (Zatybekov et al. 2017). Genomic regions significantly associated with seed quality traits on seed protein and oil content were also identified (Hwang et al. 2014). Significant quantitative trait nucleotides (QTNs) were identified in abiotic stress-tolerant characteristics such as seed-flooding tolerance (Yu et al. 2019). Pest resistant QTL for soybean aphids (*Aphis glycines* Matsumura) and soybean cyst nematodes (*Heterodera glycines* Ichinohe) have also been identified by GWAS (Neupane et al. 2019).

In maize, GWAS has been widely accepted (Xiao et al. 2017). Using GWAS, various significant SNPs/genes were associated with agronomic traits such as plant height (Gyawali et al. 2019; Peiffer et al. 2014; Zhang et al. 2019c), flowering traits (days to tassel, days to silk, days to anthesis and anthesis-silking interval) in a multiple hybrid population (Wang et al. 2017a) and husk traits (number of layers, length, width, and thickness) from a diverse panel that consisted of 508 genotypes (Cui et al. 2016). Marker-trait associations (MTA) were also identified with abiotic stress resistant traits such as drought tolerance at the seedling stage (Wang et al. 2016c) and cold tolerance (Revilla et al. 2016). In a study that focused on forage quality traits (acid detergent fiber, neutral detergent fiber and in vitro dry matter digestibility) in mature stalks, SNPs were identified from a diverse population consisting of 369 inbred genotypes, with each SNP accounting for 4.2%

to 6.2% of the phenotypic variation (Wang et al. 2016a). The abundant discoveries derived from GWAS studies mentioned above have assisted maize breeders in improving the agronomic traits as well as abiotic resistance in maize cultivars.

In wheat, GWAS has been well-established and applied. Significantly associated SNPs have been detected for numerous traits including heading date (Ogbonnaya et al. 2017; Zhang et al. 2018a), spike length (Ogbonnaya et al. 2017), flowering date (Zhang et al. 2018a), plant height (Li et al. 2019a), thousand kernel weight (Liu et al. 2017c; Ogbonnaya et al. 2017; Sun et al. 2017), grain yield (Ogbonnaya et al. 2017) and pre-harvest sprouting resistance (Lin et al. 2016; Lin et al. 2017). Significant MTA were also identified with disease-resistant traits such as powdery mildew resistance (Liu et al. 2017a), seedling leaf rust resistance (Li et al. 2016b) and FHB resistance (Wang et al. 2017c). In terms of seed quality of winter wheat, Lin et al. (2016) identified four QTL that were significantly associated with grain colour, while Tsai et al. (2020) identified two SNPs associated with seed moisture and one SNP associated with seed starch content. These studies increased the understanding of wheat genetics through GWAS and the identified SNPs could be useful in MAS in wheat breeding.

Aside from the major crops mentioned above, GWAS has also been applied in barley (Bellucci et al. 2017; Jabbari et al. 2018; Tsai et al. 2020), rice (Pantaliao et al. 2016; Yang et al. 2018; Yano et al. 2016), oat (Carlson et al. 2019; Newell et al. 2011), flax (He et al. 2018; Soto-Cerda et al. 2018; Xie et al. 2019; You et al. 2018a), and pulse crops such as common bean (*Phaseolus vulgaris* L.) (Hoyos-Villegas et al. 2017; Nascimento et al. 2018), chickpea (*Cicer arietinum* L.) (Bajaj et al. 2016; Basu et al. 2018; Upadhyaya et al. 2015), and lentil (Khazaei et al. 2017), as well as in many other crops.

2.4.3 Application of genome-wide association study in *Brassica napus*

Similar to other major crops, GWAS has been well established in *B. napus*, allowing scientists to have a better understanding of the genetic structure of numerous traits in *B. napus*. It was used to detect SNPs/QTL/candidate genes that were significantly associated with yield-related traits such as flowering time (Korber et al. 2016; Raman et al. 2019; Schiessl et al. 2015; Xu et al. 2016; Zhou et al. 2018), maturity time (Zhou et al. 2018), plant height (Korber et al. 2016; Li et al. 2016a; Schiessl et al. 2015; Sun et al. 2016a; Zheng et al. 2017), stem strength (Li et al. 2018b), primary branch number (He et al. 2017; Li et al. 2016a), number of seeds per pod, number of pods per branch, number of pods per plant, number of pods on main inflorescence, branch yield, main inflorescence yield (Lu et al. 2017), harvest index (Lu et al. 2016), seed weight (Li et al. 2014b; Lu et al. 2017), seed yield (Korber et al. 2016; Schiessl et al. 2015) and lodging coefficient (Li et al. 2018b).

Marker-trait association has also been identified in seed quality traits such as seed coat colour (Wang et al. 2017b), erucic acid content (Korber et al. 2016; Li et al. 2014b; Wang et al. 2018a), fatty acid composition (Gacek et al. 2016; Qu et al. 2017; Xue et al. 2018), oleic acid content (Zhao et al. 2019a), seed oil content (Korber et al. 2016; Li et al. 2014b; Liu et al. 2016a; Sun et al. 2016b; Wang et al. 2018a; Wu et al. 2016b; Xiao et al. 2019), glucosinolate content (Kittipol et al. 2019; Korber et al. 2016; Li et al. 2014b; Wang et al. 2018a; Wei et al. 2019a) and seed acid detergent lignin and hull content (Wang et al. 2015a).

Significant associations have also been identified in physiological characteristics such as hypocotyl elongation (Luo et al. 2017), seed germination and vigour (Hatzig et al. 2015), root system architecture traits (primary root length, shoot dry weight, root dry weight, total root length, lateral root density, lateral root length, lateral root number and mean lateral root length) (Wang et

al. 2017d), calcium accumulation (Alcock et al. 2017; Chen et al. 2018) and magnesium accumulation (Alcock et al. 2017). For abiotic stress tolerant characteristics, MTAs have also been identified in salt tolerance (Wan et al. 2017; Yong et al. 2015). In addition, MTAs have been identified for disease resistant characteristics such as *Sclerotinia* stem rot resistance (Wei et al. 2016; Wu et al. 2016a), clubroot resistance (Li et al. 2016c) and blackleg resistance (Raman et al. 2016).

2.5 Genomic selection

It is commonly recognized that many traits are often controlled by numerous markers that have small effects on a trait, which is difficult to depict by QTL mapping or MAS (Cobb et al. 2019; de Los Campos et al. 2013). Genomic selection (GS) is considered as a variant of MAS which assumes that "at least one marker is in linkage disequilibrium (LD) with the locus/loci" that control(s) the trait of interest instead of focusing on major marker effects or novel genes as in QTL mapping or MAS (Desta and Ortiz 2014; Goddard and Hayes 2007). Genomic selection attempts to examine and evaluate all gene/marker effects along the entire genome for traits of interest for each genotype (Heffner et al. 2011a; Newell and Jannink 2014).

Genomic selection was initially used for improving the rate of genetic gain in animal breeding (Meuwissen et al. 2001). In GS, a training population (TP, also known as reference population) is often divided into a training set and a validation set (Desta and Ortiz 2014). All individuals in the TP have been phenotyped and genotyped and are used to train a model (Heffner et al. 2011a; Jannink et al. 2010). This is called the "cross validation" process (Equation 2.1).

$$y = \mu + \sum_k \chi_k \beta_k + e \qquad [2.1] \text{ (Desta and Ortiz 2014)}$$

where y is the vector of phenotype of the given trait, μ is the population mean of the phenotype, k stands for the locus, χ_k is the allelic status at locus k, and β_k is marker effect at locus k, and e is the residual effects ranging from $0 \sim \sigma_e^2$, where σ_e^2 is the residual variance.

Based on the genotypic data of the untested test population (or the candidate /breeding population), this model is then used to evaluate and predict the phenotypic performance of each individual in the test population (Isidro et al. 2015; Mangin et al. 2017) (Equation 2.2).

$$GEBV = x_{new}\widehat{\beta_k} \qquad [2.2] \text{ (Desta and Ortiz 2014)}$$

where x_{new} represents a matrix consisting of the allelic status of individuals in a test population, and $\widehat{\beta_k}$ is the regression coefficient of β_k.

Then the genomic estimated breeding value (GEBV) of a genotype can be calculated by summing all its SNP effects (Su et al. 2010). The identified highest GEBVs can then be used as the selection criteria without phenotypic data by breeders (Desta and Ortiz 2014; Heffner et al. 2011a; Jannink et al. 2010; Thavamanikumar et al. 2015). Correlation between GEBVs and empirically EBV is then computed for prediction accuracy of the applied model (Desta and Ortiz 2014) (Equation 2.3).

$$r_A = \sqrt{\frac{h^2}{h^2 + \frac{M_e}{N_p}}} \qquad [2.3] \text{ (Desta and Ortiz 2014)}$$

where h^2 stands for the narrow sense heritability,

M_e stands for the number of independent chromosome blocks,

N_p stands for the number of individuals in the TP.

2.5.1 Factors that affect prediction accuracy of genomic selection

When breeders try to choose a model for GS, unfortunately there is no such thing as "one size fits all" (Lorenz et al. 2011). Numerous factors, such as size and components of the TP, marker density, trait heritability and model performances can all influence prediction accuracy (Tan et al. 2017;

Zhang et al. 2019a). In fact, there is no perfect GS model that can be applied to all species of crops, or even on the same species for different traits (Lorenz et al. 2011). Therefore, these related (but not limited to) factors should be adjusted according to specific purposes.

2.5.1.1 Training population

2.5.1.1.1 Size of the training population and relatedness within training population

Prediction accuracy in GS is largely affected by the size of the TP (Robertsen et al. 2019). Although other factors need to be considered, prediction accuracy increases when TP size increases (Desta and Ortiz 2014). In structured populations, population size is a crucial factor when estimating genomic heritability and evaluating genomic prediction (Guo et al. 2014). Theoretically, when constructing a TP, it should include a wide range of genotypes, which offer diverse genetic background information to allow more accurate predictions in the next steps (Calus 2010). However, this is neither feasible, nor practical, which makes it reasonable to construct a TP that is genetically closely related to the test population to improve prediction accuracy (Calus 2010). Thus, the design and structure of the TP becomes another important factor to consider.

2.5.1.1.2 Genetic structure of training population

An appropriate design of the TP is crucial in a successful GS project. Size of the TP and the relatedness between training set and validation set could greatly impact prediction accuracy (Lozada et al. 2019). When the test population is closely related to the TP, GS produces more precise GEBV values (Calus 2010; Clark and van der Werf 2013). This is because the relatedness between TP and test population can help capture the genetic relationship within the test population (Clark et al. 2011). Schulz-Streeck et al. (2012) found that combining different population groups or genotypes clusters improved prediction accuracy in maize.

Genomic selection accuracy can also be affected by population structure in a stratified population (de Los Campos et al. 2015; Guo et al. 2014). Genomic selection assumes that population structure is consistent amongst the training set, the validation set, and the candidate set, which makes it possible to predict the performance of the test population based on genotypic and phenotypic information from the training population.

2.5.1.2 Markers

Marker density is among the major factors that affect prediction accuracy. Generally, higher marker density is preferred, since lower marker density results in lower GS prediction accuracy (Moser et al. 2009). Assuming at least one marker is in LD with loci that contribute to the trait of interest is the fundamental concept of GS. Thus, higher marker density would ensure the association between markers and QTL/genes, and therefore produce higher prediction accuracy (Desta and Ortiz 2014). It was reported that in winter wheat, when the marker number decreased from 1158 to 192, prediction accuracy decreased by 10% (Heffner et al. 2011a). However, "the more the better" is not always the case for marker density in GS (Combs and Bernardo 2013). Although prediction accuracy increased with higher marker density, gains in the genome-wide prediction accuracy stayed at the same level when a relatively high marker density was reached (Combs and Bernardo 2013). In addition, the issue of overfitting might arise when GS models were fitted with a large number of markers, meaning that non-genetic effects are ascribed to markers (Hickey et al. 2014).

The ideal marker density actually relies on the LD span of the species of interest and the population size (Desta and Ortiz 2014). For example, compared with wheat and barley whose LD span is longer, maize would prefer a higher marker density since its LD span is shorter (Desta and Ortiz 2014). When the relatedness between the training population and the test population is high, fewer

markers are needed due to the common LD blocks (Hickey et al. 2014). Therefore when using a diverse TP to predict the performance of a poorly related test population the increase in prediction accuracy corresponding to greater marker density was higher (Norman et al. 2018). The ideal marker density needed in GS is also related to the trait of interest since the response to greater marker density varied depending on the particular trait being investigated (Norman et al. 2018).

Different types of markers could be used in GS (Heffner et al. 2011a; Solberg et al. 2008). Heffner et al. (2011a) used SSR and DarT markers to predict grain quality traits in wheat. Low density SCAR markers were used in GS with good performance in predicting seed weight in soybean (Shu et al. 2013). SSR markers were also used in genomic selection in oil palm (*Elaeis guineensis* Jacq.) (Cros et al. 2015). Elbasyoni et al. (2018) found that GBS-scored SNPs performed similar or better compared to array-scored SNPs in terms of GS prediction accuracy. Marker density is also related with the type of markers used in GS. For example, Solberg et al. (2008) found that compared to SSR markers, two to three times greater SNP marker density was needed to obtain a comparable accuracy of GS.

2.5.1.3 Trait heritability

It is expected that GS prediction accuracy is often higher on traits with high heritability compared to traits with low heritability (Moser et al. 2009). However, it is also stated that $h^2 \times n$ (where h^2 is the heritability of the trait and n is the TP size) is the most important factor that affects the prediction accuracy instead of h^2 or n individually (Combs and Bernardo 2013). Thus, a sufficient TP size can compensate for the impact of low trait heritability to produce an accurate prediction model (Combs and Bernardo 2013; Solberg et al. 2008).

2.5.1.4 Model performances

In GS, a statistical model is used to predict the breeding values of the individuals in the test population whose phenotypic performance is unknown (Bassi et al. 2016). Whole-genome regression (WGR) models in general are categorized into parametric and non-parametric regressions (Desta and Ortiz 2014) (Table 2.2). Specifically, a parametric model characterizes the parameters of a probability distribution, while a non-parametric model focuses more on the shape of the function (Roehrig 1988).

Different models based on varying assumptions have been applied to estimate marker effects in GS (Desta and Ortiz 2014; Lorenz et al. 2011). As summarized by Desta and Ortiz (2014) (Table 2.3) the most commonly used models include ridge regression best linear unbiased prediction (rrBLUP) (Endelman 2011), Bayesian shrinkage regression methods (BayesA, BayesB, BayesCπ) (Calus et al. 2008; Meuwissen et al. 2001; ter Braak et al. 2005; Xu 2003) and reproducing kernel hilbert spaces regression (RKHS regression) (Gianola et al. 2006).

2.5.1.4.1 Parametric models

The major factor that differentiates these prediction models is their assumptions on how much genetic variance could be explained by the individual loci (Clark et al. 2011; Tan et al. 2017). GBLUP and rrBLUP are the basic and most widely applied models, and are considered equivalent in general (Habier et al. 2007). Both of them assume all loci contribute equally to the genetic variance (Clark et al. 2011; Tan et al. 2017; Whittaker et al. 2000; Würschum et al. 2014). GBLUP applies a genomic relationship matrix (GRM) to estimate the additive effects (Tan et al. 2017), while rrBLUP relies on the estimated marker effects (Endelman 2011). In rrBLUP, the penalization is equal to all markers (Resende et al. 2012). One of the main differences between them is that the size of genetic effects matrix in GBLUP is $n \times m$, where n is the number of individuals in the

Table 2.2 Classification of whole-genome regression models[1].

Parametric regressions	Penalized approach	Ridge regression best linear unbiased prediction (rrBLUP)
		Lease absolute shrinkage and selector operator (LASSO)
		Elastic net (EN)
		Support vector regression (SVR)
		Neural networks (NN)
		Reproducing kernels Hilbert spaces regression (RKHS)
		Genomic best linear unbiased prediction (GBLUP)
	Bayesian approach	Genomic best linear unbiased prediction (GBLUP)
		Bayesian ridge regressions (BRR)
		Bayesian LASSO (BL)
		BayesA
		BayesB
		BayesC
		Neural networks (NN)
		Reproducing kernels Hilbert spaces regression (RKHS)
Non-parametric regressions	Support vector regression (SVR)[2]	
	Random forest (RF)[2]	
	Neural networks (NN)[2]	
	Reproducing kernels Hilbert spaces regression (RKHS)[2]	

[1] This table is adapted from Figure 2 of "*Genomic selection: genome-wide prediction in plant improvement*" (Desta and Ortiz 2014)

[2] Machine-learning methods

Table 2.3 Main features of genome-wide prediction models[1]

Model	Features
rrBLUP	Assumes that all markers have equal variances with small but non-zero effect. Applies homogeneous shrinkage of predictors towards zero but allows for markers to have uneven effects. Computed from a realized-relation matrix based on markers. Some QTL are in LD to marker loci, whereas others are not.
LASSO	Combines both shrinkage and variable selection methods. rrBLUP does not use variable selection but outsmarts LASSO when there is multi-colinearity between the predictors.
EN	Double regularization using ℓ_1 and ℓ_2 penalty norms combines the merited features of these norms to confront the challenge of high-dimensional data.
BRR	Induces homogeneous shrinkage of all marker effects towards zero and yields a Gaussian distribution of marker effects. Similar to RR-BLUP, there is a problem of QTL linkages to the marker loci.
BL	Applies to both shrinkage and variable selection. Has an exponential prior on marker variances resulting in a double exponential (DE) distribution The DE distribution has a higher mass density at zero and heavier prior tails compared with a Gaussian distribution
BayesA	Utilizes an inverse chi-square (χ^2) on marker variances yielding a scaled *t*-distribution for marker effects Similar to BL and in contrast to BRR, it shrinks tiny marker effects towards zero and larger values survive Has a higher peak of mass density zero compared with the DE distribution
BayesB	Similar to BayesA, uses an inverse χ^2 resulting in a scaled t-distribution Unlike BayesA, utilizes both shrinkage and variable selection methods When $\pi = 0$, then it is similar to BayesA
BayesC	Applies both shrinkage and variable selection methods Characterized by a Gaussian distribution BayesB and BayesC consist of point of mass at zero in their slab priors
BayesCπ	A modified variant of BayesB Used to alleviate the shortcomings of BayesA and BayesB Unlike BayesB, π is not fixed, but estimated from the data
RKHS	Based on genetic distance and a kernel function with a smoothing parameter to regulate the distribution of QTL effects Effective for detecting nonadditive gene effects
RF	Uses the regression model rooted in bootstrapping sample observations Takes the average of all tree nodes to find the best prediction model Captures the interactions between markers

[1] This table is adapted from Table 1 of "*Genomic selection: genome-wide prediction in plant improvement*" (Desta and Ortiz 2014)

population, while that in rrBLUP is $m \times m$, where m is the number of markers, which is often significantly larger than n (Clark and van der Werf 2013).

Each method of Bayesian models has its own prior distribution of the genetic variance explained by the loci, and they assume that markers could have different effects across the loci (Perez and de los Campos 2014; Thavamanikumar et al. 2015). BayesA ($\pi = 0$) assumes the prior specification of marker effects follow a scaled-t distribution (Perez and de los Campos 2014) and allows some of the markers to be treated as zero since they do not contribute to the phenotype (Habier et al. 2011a). BayesB method ($\pi > 0$) allows the probability that some markers might not have any effects at all and thus can be excluded from the model, meaning that genetic variances are only found on a few loci (Meuwissen et al. 2001). BayesC was developed based on BayesB, which utilizes a common effect variance in substitution for the variance specific to a certain loci in BayesB (Colombani et al. 2013). Unlike the previous Bayesian models, π is treated as unknown in BayesCπ method (Colombani et al. 2013). Bayesian LASSO (BL) assumes a double exponential (DE) prior specification, while Bayesian Ridge Regression (BRR) assumes a similar prior specification with Gaussian or normal distribution, with similar-level shrunk effects (Perez and de los Campos 2014). Although these Bayesian models have different assumptions, it was reported that their performance was very similar with each other (Colombani et al. 2013; Resende et al. 2012).

It has been reported that GS model prediction accuracy can be improved if a non-linear model is used to evaluate non-additive genetic effects (Desta and Ortiz 2014). The RKHS regression model linearly combines the basic feature of reproducing kernel, meaning that it is an additive genetic model combined with a kernel function to generate a matrix that can be used in a linear model (de Los Campos et al. 2009; Gianola et al. 2006; Tan et al. 2017). As a nonparametric method, RKHS

was proposed to cope with problems such as “dimensionality, multicollinearity, and the inability to deal effectively with epistasis” that exists in current popular models (eg.: rrBLUP, BayesA, BayesB) (Sun et al. 2012).

There are models that could include G × E effects into GS prediction as well, and they were found to be more accurate than those that do not consider G × E effects (Acosta-Pech et al. 2017; Crossa et al. 2014; Cuevas et al. 2017; Montesinos-Lopez et al. 2019). As a result, when utilizing models that account for G × E effects, breeders would have a better idea about the stability of the genotypes and thus be able to select the genotypes with the best performance within a certain site or across sites (Roorkiwal et al. 2018). Heslot et al. (2014) proposed a model that integrated weather data which improved prediction accuracy using a large historical winter wheat breeding dataset. In hybrid maize breeding, Acosta-Pech et al. (2017) found that the prediction performance was better when the model included interaction of general and specific combining ability with environments. Gapare et al. (2018) demonstrated that GS prediction accuracy of the model that included marker-by-environment effect performed better than the single-location and the across-location models for both fibre length and fibre strength.

Some models can capture epistasis, which would improve prediction accuracy. For example, extended genomic best linear unbiased prediction (EG-BLUP) was found to improve prediction accuracy in self-pollinated crops (Jiang and Reif 2015). Another example is the RKHS regression, which could take epistasis or dominance into consideration (de Los Campos et al. 2009; Gianola et al. 2006). He et al. (2016) found that models that considered both additive and epistasis effects improved prediction accuracy compared to models that only considered additive effects in a commercial winter wheat population.

2.5.1.4.2 Non-parametric models

In addition to the parametric models described above, non-parametric models such as machine learning (ML) methods have also been applied in GS in plant breeding (van Dijk et al. 2021). Machine learning is a subdivision of artificial intelligence that develops algorithms based on training data and uses them to perform specific tasks (van Dijk et al. 2021). In GS, ML algorithms are more flexible in managing complicated associations (Montesinos-Lopez et al. 2021) since they are expected to capture relationships between markers and phenotypes differently from the linear models for GS (Heslot et al. 2012).

Support-vector regression (SVR) is a machine learning algorithm that considers data instances in the training set as points in a high-dimensional vectors space (that is, the vector space has a dimension equal to the number of features for each instance) (Drucker et al. 1997). With SVR, a hyperplane is constructed to perform regression, but with a different optimization compared to traditional regression such as least-squares or Lasso regression: the quadratic optimization seeks to minimize coefficients associated with the slope of the hyperplane (Drucker et al. 1997). The formulation of the SVR ensures that the solution is sparse, in that it depends on few instances in the training set, which reduces overfitting and improves efficiency.

Two tree-based ensemble methods have also been used in GS: extreme gradient boosting (XGBoost) (Chen and Guestrin 2016; Friedman 2001) and random forests (RF) (Breiman 2001). Both methods rely on decision trees, which are flowcharts based on binary decisions for a single feature. Building a decision tree from a training set is accomplished by growing a tree from the root to the leaves by finding decisions for new internal nodes that split the training set into two subsets that are as uniform as possible in their output values (Myles et al. 2004). Each tree is a set of decision rules for characterizing instances – the rules are based on values of single features

(Myles et al. 2004). The predicted value from each tree is averaged to achieve a final prediction from the entire ensemble (Chen and Guestrin 2016). Random forests are an ensemble ML method where large collections of trees are used to obtain a prediction (Breiman 2001). The number of trees in a RF is a hyperparameter, as is the maximum depth of the trees. Typically, the number of trees in a RF is in the range of 10s to 100s, while the depth is typically less than 50 (Breiman 2001). To obtain a prediction from a RF, each tree is used to obtain an output prediction, and the results for all trees are averaged to obtain a final prediction (Holliday et al. 2012). XGBoost shares many aspects with RFs – both are ensemble techniques consisting of a collection of decision trees (Friedman 2001). The primary difference is that while the trees are constructed independently for RFs, in XGBoost, each tree is constructed using information from all previously constructed trees. As a relatively newer approach in plant breeding, GS still has its limitations (Desta and Ortiz 2014; Jonas and de Koning 2013). Although a great amount of GS models have been proposed for plant breeding, so far none of them can outperform all the others in all circumstances, due to different genetic architectures and the specific traits of interest (Lorenz et al. 2011).

2.5.2 Advantages and limitations of genomic selection

With the help of GS, breeders can make selections with predicted performance of a breeding population instead of the observed performance (Combs and Bernardo 2013), which theoretically shortens the breeding cycle and reduces the cost in phenotyping potential candidates (Desta and Ortiz 2014). One of the greatest advantages of GS is that it has the potential to accurately predict GEBVs without repeated phenotyping over multiple generations, which should cut down the cost of phenotyping as well as generation interval (Habier et al. 2007; Jannink et al. 2010). Genomic selection is also a promising tool for integrating historical data from a breeding program to aid with current research (Heslot et al. 2014). Annicchiarico et al. (2015) found that GS for biomass

yield was more effective than conventional selection. Genomic selection had a higher prediction accuracy than MAS in a wheat population that included multiple families (Heffner et al. 2011b). Genomic selection also outperformed in MAS predicting seed weight in soybean (Zhang et al. 2016). However, Beyene et al. (2019) also compared the efficiency of GS versus phenotypic selection in hybrid maize and concluded that GS did not necessarily perform better than phenotypic selection, although it did reduce cost by 32%, which was one of the greatest advantages in incorporating GS into maize breeding.

A major limitation when implementing GS into breeding programs was the limited number of markers available as well as the very high cost of genotyping (Goddard and Hayes 2007). This has now become less of an issue due to the fast development and improvement of newer sequencing technologies, which has led to a dramatic reduction in the cost of genotyping (Rasheed et al. 2017). Another limitation is the loss of genetic diversity in the GS process. In some conventional GS approaches, short-term gain is achieved at the cost of the long term potential, due to the reduction in genetic diversity (Moeinizade et al. 2019). A new approach called the optimal population value (OPV) was developed to address this issue (Goiffon et al. 2017). However, this method could be time-consuming since its GEBV calculation is based on the best individuals derived from the unlimited generation of the best group of the population (Goiffon et al. 2017). Future GS studies also need to address the issues of evaluating marker effects more accurately, examining the effectiveness of GS in populations with variable LD structures, as well as optimizing the scheme in resource allocation (Moeinizade et al. 2019).

2.5.3 Current application of genomic selection in plant breeding

Over the last decade, following the success of its application in animal breeding, GS has become a useful tool in selecting complex traits in crop breeding (Heffner et al. 2011a). Genomic selection

has been evaluated for its potential performance in various major crops such as soybean (Shu et al. 2013), wheat (Crossa et al. 2010; Fleury et al. 2010; Heffner et al. 2011a; Heffner et al. 2011b), maize (Crossa et al. 2010; Lorenzana and Bernardo 2009; Riedelsheimer et al. 2012), barley (Lorenzana and Bernardo 2009), rice (Grenier et al. 2015; Spindel et al. 2015), and rapeseed (Jan et al. 2016; Snowdon and Iniguez Luy 2012; Würschum et al. 2014).

In wheat, GS has been applied extensively in predicting yield related traits such as heading date (Elbasyoni et al. 2018; Lozada et al. 2019; Sarinelli et al. 2019; Zhao et al. 2014), flowering time (Bentley et al. 2014; Watson et al. 2019), maturity time (Elbasyoni et al. 2018), plant height (Bentley et al. 2014; Elbasyoni et al. 2018; Lozada et al. 2019; Sarinelli et al. 2019; Watson et al. 2019; Zhao et al. 2014) and grain yield (Bentley et al. 2014; Elbasyoni et al. 2018; Hoffstetter et al. 2016; Lozada et al. 2019; Michel et al. 2019). It has also been used in predicting seed quality traits such as protein content (Michel et al. 2019), processing quality and end use quality (Battenfield et al. 2016; Michel et al. 2019). GS has also been applied in breeding for disease resistance in wheat such as rust resistance (Daetwyler et al. 2014; Ornella et al. 2017), FHB severity (Schulthess et al. 2018), FHB resistance (Hoffstetter et al. 2016) and powdery mildew resistance (Sarinelli et al. 2019). In addition, GS has been utilized in hybrid wheat breeding (Longin et al. 2015).

In maize, GS has been utilized in predicting flowering time, ear height and grain yield (Guo et al. 2019) root length (Pace et al. 2015), drought resistance (Vivek et al. 2017), drought tolerance (Dias et al. 2018), and witchweed (*Striga hermonthica* (Del.) Benth.) resistance (Badu-Apraku et al. 2019). Genomic selection was also intensively studied in predicting hybrid performance in maize (Andorf et al. 2019; Guo et al. 2019; Technow et al. 2012; Technow et al. 2014). For example, Acosta-Pech et al. (2017) examined the effect of adding a GxE term in to the prediction model and

found that the prediction accuracy increased by 16.73%, 12.30% and 21.74 for silage yield, starch content and dry matter content, respectively.

Genomic selection has been utilized for numerous traits in *B. napus*. Würschum et al. (2014) described that the rrBLUP method had the highest prediction accuracy in predicting plant height but the lowest prediction accuracy in glucosinolate content and grain yield in a DH winter rapeseed population. Li et al. (2015a) reported using GS in flowering time prediction in *B. napus* and obtained relatively high prediction accuracy using different models. Zou et al. (2016) reported applying genomic prediction in predicting seed quality traits with high prediction accuracies in *B. napus*, and found that the choice of prediction models did not impact prediction accuracies significantly. Werner et al. (2018b) reported that high prediction accuracies were achievable through utilizing low density markers, ranging from hundreds to a few thousand, and the results were comparable to that through high density arrays. Fikere et al. (2018) reported that previously known QTL information on blackleg resistance only accounted for less than 30% of its genetic variance, and a big chunk of the genetic variation remained unknown, which could be characterized by genomic selection.

Genomic selection is a relatively new approach to estimate hybrid performance in rapeseed breeding. Only a limited number of studies have been reported. Jan et al (2016) reported that genomic selection was applied in a testcross study in order to select the best parental combination for hybrids. Liu et al. (2017b) found that the performance of the hybrids in an immortalized F_2 population was determined by additive, dominance and epistatic effects together. Similar with Zou et al (2016), Wener et al. (2018a) also stated that compared to the choice of models, prediction accuracy of genomic prediction relied more on the nature of the traits of interest. In terms of

genotyping platform, the *Brassica* 60K genotyping array was reported to be of great value in predicting hybrid performance in *B. napus* (Werner et al. 2018a).

With the increasing availability of genomic sequencing data as well as the decreasing cost for high-throughput genotyping (Edwards et al. 2013), genotypic information is no longer among the major limitations in implementing GS into *B. napus* breeding. In fact, one of the major challenges for *B. napus* breeders is how to relate the huge amount of genomic information to the corresponding phenotypic data (Snowdon and Iniguez Luy 2012). In addition, more comprehensive statistical models are needed to identify favourable cross combinations to give better hybrid performance predictions. Another tool that can be implemented in GS in *B. napus* breeding is high-throughput phenotyping that has been utilized in wheat breeding (Rutkoski et al. 2016), which would offer more accurate phenotypic data in the analysis and therefore help improve the prediction accuracy.

The objectives of the following three experiments were as follows: (1) the first experiment, genome-wide association study (GWAS) of agronomic and seed quality traits in rapeseed (*B. napus* L.), examined the effects of population structure, population size, population composition, marker density and choice of models on GWAS and GS; (2) the second experiment GS and performance prediction in hybrid *B. napus* L., assessed the potential of GS in predicting the agronomic performance and seed quality traits of hybrid canola using various conventional GS models; (3) the final experiment, GWAS-guided GS of agronomic and seed quality traits in *B. napus* L., compared the performance of conventional GS models and the GS + *de novo* GWAS models.

3. GENOME-WIDE ASSOCIATION STUDY OF AGRONOMIC AND SEED QUALITY TRAITS IN *Brassica napus* L.

3.1 Abstract

Canola (*Brassica napus* L.) is currently the third-largest oilseed crop in the world, after soybean [Glycine *max* (L.) Merr.] and palm (*Elaeis guineensis* Jacq.). In canola, the majority of agronomic traits are controlled quantitatively, and the environment greatly impacts the phenotypic performance of these traits. To improve canola production performance, it is critical to understand the underlying genetics of canola traits related with yield and seed quality. In this research, five traits were investigated: seed yield, plant height, seed protein content, seed oil content and seed glucosinolate content. A parental population and a combined population (parents and hybrids) were examined in this research. Phenotypic data of the parental genotypes were collected from five site-years across southern Manitoba, while hybrid data were collected from 43 site-years across western Canada. Population structure analysis revealed that the genetic background of both the parental population and the combined population had a relatively low level of diversity. Two marker sets that contained 26,651 and 16,855 single nucleotide polymorphisms (SNPs), respectively, were used in genome-wide association study analysis with six models. The multiple locus mixed linear model (MLMM), fixed and random model circulating probability unification (FarmCPU) and compressed mixed linear model (CMLM) performed better than three mixed linear models that considered population structure (MLM+K, MLM+K+PCA and MLM+K+Q). In total, 141 significant marker-trait associations (MTAs) were identified based on the two marker sets, and 222 genes were predicted and annotated under the Brassicales order. Thirty genes were

identified in this research that had been previously identified in *B. napus* associated with abiotic stress response, disease resistance and glucosinolate synthesis.

3.2 Introduction

Canola (*Brassica napus* L.) is an economically important crop (Paterson et al. 2001) and is one of the most important sources of plant-based edible oil (Cartea et al. 2019). Currently, 13.3% to 16.0% of the global vegetable seed oil for human consumption and industrial products is provided by canola production (Delourme et al. 2006; Wang et al. 2018a). Canola contribution to the Canadian economy has increased by 35% to $29.9 billion per year in the past ten years (LMC International 2020). As a result, improving canola yield and yield-related traits is a major breeding and production goal. By 2025, the market demand for Canadian canola production will need to exceed 2,914 kg ha^{-1} (52 bushels/acre) to reach the 26 Mt production goal (Canola Council of Canada 2014). Canola breeding efforts have focused on yield-related agronomic traits (seed yield, plant height, disease resistance, lodging resistance and shattering resistance) and seed quality traits (oil content, fatty acid profile, glucosinolate content and more recently seed protein content). Plant height (HT) is an important trait in canola as it is significantly correlated with lodging and seed yield (YLD), and serves as an important selection criterion in canola breeding (Ivanovska et al. 2007; Wu and Ma 2016). As an oilseed crop, seed oil content (SOC) confers the most important economic value in *B. napus* and has been extensively studied (Delourme et al. 2006; Jiang et al. 2014; Sun et al. 2016b; Tang et al. 2021; Vigeolas et al. 2007; Wang et al. 2010; Zou et al. 2010). Numerous studies have also been conducted on protein content (SPC) (Chao et al. 2017; Jolivet et al. 2009; Nesi et al. 2008), since canola protein contains a balanced amino acid profile, and has been suggested as an alternative protein for human consumption (Poisson et al. 2019). Glucosinolate content is also an important trait in canola as low seed glucosinolate content (GSL)

is essential for human consumption. High leaf glucosinolate content helps resistance against pests and diseases (Liu et al. 2020b), therefore manipulating GSL has always been important in canola breeding (Li et al. 2014a; Qu et al. 2015; Zhao and Meng 2003). It is vital to understand the genetic basis of these important yield-related and seed quality traits in order to produce high-yielding cultivars with improved quality.

Genome-wide association study (GWAS) examines the marker-trait associations (MTAs) through statistical analysis on the whole genome of an organism (Lekklar et al. 2019). As an effective method to analyze agronomic traits (Li et al. 2014a), GWAS has been widely used in numerous crops such as soybean [Glycine *max* (L.) Merr.] (Hwang et al. 2014; Yu et al. 2019; Zatybekov et al. 2017), maize (*Zea mays* L.) (Chen et al. 2015; Gyawali et al. 2019; Wang et al. 2016c) and wheat (*Triticum aestivum* L.) (Ogbonnaya et al. 2017; Wang et al. 2017c) to identify genes or markers that are associated with complex traits. Compared with linkage mapping [quantitative trait loci (QTL) mapping] that only offers information on the trait of interest relative to a specific population that is genetically related, GWAS can be applied to more diverse populations and therefore, covers a broader genetic background (Gupta et al. 2014). As a result, GWAS allows for a higher mapping resolution using single nucleotide polymorphism (SNP) markers in comparison to linkage mapping (Lekklar et al. 2019). Another advantage of GWAS is that it is usually considered more time-efficient and cost-effective as it is does not require the development of a specific population (Cortes et al. 2021). However, the power of GWAS can be significantly affected by population structure (Asoro et al. 2011; Liu and Yan 2019). Due to the stratification underlying the population structure, linkage disequilibrium (LD) exists even though the loci involved are not physically linked. This can cause increased false positive rates and inflation in test results for GWAS using single locus models (Segura et al. 2012). This problem has been

addressed by several statistical methods including multi-locus mixed-model (MLMM) (Segura et al. 2012), fixed and random model circulating probability unification (FarmCPU) (Liu et al. 2016b), and efficient mixed-model association (EMMA) (Kang et al. 2008). Another effective way to account for population structure is to implement a kinship matrix that recognizes the relationships between individuals (Zhang et al. 2010).

Genome-wide association studies have been used to identify SNPs significantly associated with numerous traits in canola such as YLD, HT and seed quality traits such as SPC, SOC, and GSL (Kittipol et al. 2019; Korber et al. 2016; Li et al. 2016a; Li et al. 2014b; Liu et al. 2016a; Schiessl et al. 2015; Sun et al. 2016a; Sun et al. 2016b; Wang et al. 2018a; Wei et al. 2019a; Wu et al. 2016b; Xiao et al. 2019; Zheng et al. 2017). However, these studies rarely used unbalanced data from hybrid breeding experiments with spring-type canola, in combination with the effects of SNP marker density on GWAS analysis. Thus, the objectives of the current research were to evaluate the effects of population composition and marker density on the accuracy of GWAS based on different single-locus and multi-locus models, as well as to identify the markers associated with YLD, HT, SOC, SPC and GSL in spring-type *B. napus*. Based on the objectives we hypothesize that the population composition, different marker density and choice of GWAS models have impacts on the identification of MTAs related with YLD, HT, SOC, SPC and GSL in spring-type *B. napus*.

3.3 Materials and methods

3.3.1 Plant materials and phenotypic data

Ninety-two parents including 31 females ("B-line") and 61 males ("R-line") within the *ogu* Institute for Agricultural Research (INRA) (now National Research Institute for Agriculture, Food and Environment, INREA) cytoplasmic male sterility (CMS) pollination control system (Ogura

1968) and 362 F_1 hybrid genotypes (developed from the parents listed above) were evaluated under the field conditions in this research. The parents were phenotyped in five site-years across southern Manitoba, Canada in locations with different soil types (Table 3.1). Each site-year contained three sub-experiments. In each sub-experiment one third of the parental genotypes and control genotypes were tested using a randomized complete block design (RCBD) with three replicates (blocks). Among all five site-years, Glenlea 2016 used 3 m double-nursery rows with a row spacing of 0.40 m, and the remaining site-years used six-row plots of 6 m by 1.6 m , with 0.20 m row spacing. Edge® Granular Herbicide (5%) (Gowman Canada, Winnipeg) was incorporated into the soil at 30.9 kg ha^{-1} in the fall before each field experiment took place. All seeds were treated with HELIX XTra® (Syngenta Canada Inc., Calgary) at 15 ml kg^{-1} seed before seeding. Fertilizer application rates for a yield goal of 2.5 t ha^{-1} were calculated based on the nutrient recommandations from the Canola Council of Canada (2020) (Table 3.2). Soil tests were conducted in the fall to determine appropriate fertilizer application rates, which varied depending upon the year, but averaged 110 kg N, 44 kg P, 13 kg S and 18 kg K per hectare. The fertilizer was broadcast applied (46-0-0, 43%), (11-52-0, 31%), (20-0-0-24, 19%), (0-0-60, 7%). During the growing season, insecticides were applied five days after emergence followed by another application one week after the first application. Decis® (Bayer CropScience, Leverkusen, Germany) was applied at 0.2 L ha^{-1} when it was under 25 °C for control of flea beetles. When the temperature was higher than 25 °C, Matador® (Syngenta Canada Inc., Calgary) was applied at 0.08 L ha^{-1}. After seeding, herbicides were applied at the 2-4 leaf stages (BBCH 12-14). A mixture of Poast® Ultra (BASF, Ludwigshafen, Germany), Muster® (DuPont Canada, Mississauga, ON), and Lontrel™ 360

Table 3.1 Five site-years of field experiments that included 92 *Brassica napus* L. parental genotypes.

Site-year	Geographic coordinates	Plot size	Soil type description
Glenlea 2016	49.6° N, 97.1° W	3 m double-row	Rego black chernozem (typic hapludert) of scantenbury series[1]
Carman 2017	49.5° N, 98.0° W	5 m x 2 m	Orthic blacks developed on sandy, coarse loamy, fine loamy and clayey sediments[2]
Portage la Prairie 2017	50.0° N, 98.3° W	5 m x 2 m	Chernozemic with occurrence of regosolic and gleysolic[3]
Glenlea 2018	49.6° N, 97.1° W	5 m x 2 m	Rego black chernozem (typic hapludert) of scantenbury series
Portage la Prairie 2018	50.0° N, 98.3° W	5 m x 2 m	Chernozemic with occurrence of regosolic and gleysolic

[1] Information retrieved from Xu et al. (2013)
[2] Information retrieved from Mills and Haluschak (1993)
[3] Information retrieved from Michalyna and Smith (1972)

Table 3.2 Canola nutrient requirements of N, P, S, K for a target yield of 2.5 t ha^{-1}. Calculation was based on recommendations from Canola Council of Canada (2020).

Soil nutrient	Recommended rate for one kg seed yield (kg)	Recommended rate for the target yield (kg ha^{-1})
Nitrogen (N)	0.88 to 0.13	100.80 to 151.20
Phosphate (P)	0.06 to 0.07	63.00 to 75.60
Sulphur (S)	0.02 to 0.04	25.20 to 40.30
Potassium (K)	0.10 to 0.11	115.92 to 126.00

(Dow AgroSciences, Indianapolis, IN) were applied during BBCH 12-16 at 0.67 L ha^{-1}, 4.9 g ha^{-1} and 0.67 L ha^{-1}, respectively, for grassy weed and broadleaf weed control. Fungicide PROLINE® (Bayer CropScience, Leverkusen, Germany) was applied during flowering time (BBCH 61-65) at 0.37 L ha^{-1} for management of Sclerotinia stem rot caused by *Sclerotinia sclerotiorum* (Lib.) de Bary. When necessary, Pounce® 384EC Insecticide was applied at 0.18 L ha^{-1} for control of flea beetles prior to swathing (BBCH 83, 30% of pods ripe, seeds black and hard).

Plant height (HT) was measured in cm from the centre of each plot at growth stage BBCH 80-83. Swathing took place at growth stage BBCH 87, when the majority of seeds had reached physiological maturity, with approximately 30 to 35% average seed moisture. Harvest was completed by combining the plants at growth stage BBCH 99 where plants were dry (8 - 10% seed moisture). After harvest, seeds from each plot were dried and cleaned manually. Seed yield (YLD) was recorded after cleaning and converted to kg ha^{-1} for consistency in data processing. Seed quality traits, including seed protein content (SPC), seed oil content (SOC), and seed glucosinolate content (GSL) were phenotyped using six grams of cleaned seed at 0% moisture using a FOSS NIR System (Model 6500, Foss NIR Systems Inc., Maryland, USA) at the Canadian Grain Commission certified seed quality lab at the University of Manitoba. The NIR instrument was calibrated and verified based on the description by DeClercq et al. (1998). The measurements followed the protocol described by Elahi et al. (2016).

The same phenotypic data for hybrid genotypes were collected during 2014-2018 from 19 locations across Alberta, Saskatchewan and Manitoba, totalling 43 different site-years (See Table S3.1 in the Appendix for details). Fertilizer, herbicide, insecticide, and fungicide were applied as described above. Briefly, 279 out of 362 hybrid genotypes were replicated five to nine times.

Thirty-four genotypes were replicated ten to 15 times. Forty-nine genotypes were replicated more than 15 times, among which genotype 13OH76 had 120 replicates.

PROC UNIVARIATE in SAS® Studio V. 3.8 was used to examine the distribution of the residuals. Data with large deviation (studentized residuals larger than 5) were identified as outliers and were removed from the analysis. The parent and the hybrid phenotypic data were combined to compute the best linear unbiased prediction (BLUP) in SAS® Studio V. 3.8 using the Proc Mixed statement. Since not every genotype was replicated in the same site-year, the site-year effect was nested within the genotype effect. Furthermore, since the block effect was nested within the site-year effect, it was accounted for as nested within genotype-by-site-year effect. The genotype, site-year nested within genotype and block within site-year-by-genotype were all modeled as random effects. The operational model to compute the BLUP value of each genotype is shown in [3.1]:

$$y = \mu + g_i + s_{ij} + r_{jk} + e_{ijk} \qquad [3.1]$$

where: μ = overall population mean of a specific trait;

g_i = effect of the i^{th} genotype;

s_{ij} = effect of the j^{th} site-year nested within the i^{th} genotype;

r_{jk} = effect of the k^{th} block nested within the j^{th} site-year;

e_{ijk}= residual

3.3.2 Genotypic data

An average of 0.05 g leaf tissue of each genotype was sampled from the first two true leaves of one plant (BBCH 11-12) grown in the greenhouse facility at the Department of Plant Science, University of Manitoba (day temperature 25 °C; night temperature 22 °C; relative humidity 40 – 50%; light cycle 16 h light, 8 h dark). Tissue samples were stored at -80 °C until DNA extraction could occur. Genomic DNA was extracted following a standard CTAB protocol (Porebski et al.

1997), with modifications that excluded the use of polyvinylpyrrolidone and 2-mercaptoethanol, and the replacement of octanol with phenol. The quantification of isolated DNA was completed using a NanoDrop^{TM} 2000 Spectrophotometer (Thermo Scientific, MA, USA) as per manufacturer protocols. Isolated DNA samples were adjusted to 50 ng/µL for genotyping.

Isolated DNA of all 454 genotypes were sent to and genotyped at Agriculture and Agri-Food Canada (AAFC) Saskatoon (Dr. Isobel Parkin's lab) using the *Brassica* 60K Illumina Infinium SNP array (Illumina Inc., CA, USA) (Clarke et al. 2016). After the completion of genotyping, raw intensity data files and a custom cluster file were obtained from AAFC for data analysis (Clarke et al. 2016). All markers were mapped to the *B. napus "Darmor-bzh"* reference genome (Chalhoub et al. 2014) obtained from AAFC. In GenomeStudio 2.0 software V. 2.0.4 (Illumina Inc., CA, USA), all markers with more than 5% missing data and a Gentrain score of zero were excluded. All markers without chromosome number and position were removed. Genotypes with more than 20% missing data were removed for marker quality control. In the end, 436 genotypes and 26,651 markers were used in the GWAS (hereinafter referred to as MS-1) analyses (see a complete list of genotypes in the Appendix Table S3.2). A second set of markers was created by applying one more filter in GenomeStudio 2.0 to remove markers with a minor allele frequency (MAF) less than 0.05, which contained 16,855 SNP markers (hereinafter referred to as MS-2).

3.3.3 Linkage disequilibrium evaluation

The squared correlation (r^2) between markers was calculated in PLINK V. 1.90b6.16 (Chang et al. 2015). In PLINK, the flags applied were "--r2 --ld-window-r2 0 --ld-window 999999 --ld-window-kb 71850". Linkage disequilibrium (LD) was evaluated between SNPs within a window of 71,850 kb which was the length of the longest chromosome based on the *Darmor-bzh* reference genome (Bayer et al. 2017). Markers were binned by 10 kb across the genome and for each bin the mean

r^2 was calculated. The mean r^2 values were then ordered from the smallest to the largest. The LD decay distance was assessed where the r^2 value dropped below 0.2. The results were visualized in RStudio V. 1.3.1073 (RStudio Team 2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/) using R code adapted from Biostars (https://www.biostars.org/p/300381/). Linkage disequilibrium decay was plotted by A- and C-sub-genomes, the whole genome as well as each chromosome.

3.3.4 Population structure

A LD-pruned subset of markers were created using PLINK 1.9 (Chang et al. 2015). By applying a 50-5-0.2 filter, the whole genome was scanned at an r^2 threshold of 0.2 with a 50-marker-window and a scanning interval of five markers during the scanning. Following the scanning, a subset of 3,205 independent markers was extracted for population structure analysis (hereinafter referred to as "LD-pruned markers" in this chapter).

STRUCTURE V. 2.3.4 (Pritchard et al. 2000) and principal component analysis (PCA) were applied in population structure analysis using the LD-pruned markers. In STRUCTURE, Bayesian clustering was performed and an admixture model was applied with 100,000 Markov Chain Monte Carlo (MCMC) iterations and burn-in period both set as 100,000. The value range for k, which was the number of potential subpopulations, was set from one to ten. The STRUCTURE results were then uploaded to Structure Harvester Web V. 0.6.94 (July 2014) (Earl 2012) to determine and visualize the optimal k value. The STRUCTURE analysis was repeated by setting the k value as 2 to extract the Q matrix of the population.

Principal component analysis (PCA) was conducted in RStudio V. 1.3.1073. Missing data in the marker dataset was imputed using the "A.mat" function from R package "rrBLUP" (Endelman 2011) and were treated as the mean value of the specific marker across the population. R built-in

function "prcomp" was used in conducting PCA using the LD-pruned marker set as well as phenotypic data of all five traits, including HT, YLD, SPC, SOC and GSL. The results were visualized using the R package "ggplot2" (Wickham 2016) based on the category of the genotype ("R-line", "B-line" or "Hybrid").

3.3.5 Identification of marker-trait associations

Two sets of markers (MS-1 and MS-2) were used in GWAS analysis. Six GWAS models were applied to all five traits (YLD, HT, SPC, SOC and GSL). In TASSEL5 (Bradbury et al. 2007), three mixed linear models (MLM) models were used (Zhang et al. 2010) including MLM+K, MLM+K+Q and MLM+K+PCA. The difference between these three models was the implementation of the population structure. K represented the kinship matrix, which was calculated in TASSEL using the default Centered IBS algorithm. Q represented the population structure extracted from the previous STRUCTURE analysis based on Bayesian clustering, and PCA represented the principal components from marker based PCA. In GAPIT V. 3 (Wang and Zhang 2020), three models were used including compression mixed linear model (CMLM) (Zhang et al. 2010), multi-locus mixed linear model (MLMM) (Segura et al. 2012), and fixed and random model circulating probability unification (FarmCPU) (Liu et al. 2016b).

To compare the performance of the six GWAS models, Quantile-quantile (Q-Q) plots were created using the R package "dplyr" V. 1.0.6 by plotting the observed *p* values plotted against the expected *p* values. The root mean square error (RMSE) of each model was computed in RStudio to evaluate the accuracy. Smaller RMSE indicated better accuracy. The average deviation of the observed p values was plotted against expected *p* values to visualize model performances.

A Bonferroni-corrected threshold was used to identify MTAs. The computation of the Bonferroni-corrected threshold is shown in [2]. GWAS results were then plotted as Manhattan plots using the R package "CMplot" (Yin et al. 2020).

$$Bonferroni - corrected\ p = \frac{\alpha}{n} \qquad [2]$$

where: α = the original p value, in this study 0.05 was used;

n = the total number of markers, in this study n either equaled 26,651 or 16,855.

3.3.6 Identification of candidate genes

Haplotype blocks were identified using PLINK V.1.9 with a window of 10 Mb using both MS-1 and MS-2 based on both the parental and the combined populations. If a significant MTA was found within a haplotype block corresponding to the population and marker set where the MTA was identified from, the flanking markers of this block were used to extract genes from the "*Brassica napus* Genome Browser" (Chalhoub et al. 2014). If a significant MTA was not found in any haplotype block corresponding to the population and marker set where the MTA was identified from, the flanking markers of this marker were used for the extraction. The extracted information was organized as .fasta files and loaded into Omicsbox V. 1.4.11 (BioBam, Valencia, Spain) for functional annotation using the default annotation pipeline (Gotz et al. 2008). All sequences were BLASTed against the *Arabidopsis* (3701) and *B. napus* (3708) query databases. A list of GO ID was created based on the identified genes for enrichment analysis through Fisher's exact analysis (Gotz et al. 2008). The top GO terms were visualized using the R package "ggplot2" (Wickham 2016).

3.4 Results

3.4.1 Phenotypic variations

Large variation was observed in the raw data for parental genotypes as well as the hybrids. The YLD of parental genotypes ranged between 53.0 and 3086.2 kg ha^{-1} with a mean of 1118.0 kg ha^{-1} and a median of 1066.0 kg ha^{-1}), while that of the hybrid genotypes ranged between 686.4 and 5312.9 kg ha^{-1} with a mean of 2304.8 kg ha^{-1} and a median of 2122.8 kg ha^{-1}. The HT ranged from 50.0 to 150.0 cm in the parental genotypes with a mean of 92.8 cm and a median of 93 cm, and 65.0 to 172.0 cm in hybrid genotypes with a mean of 113.1 cm and a median of 110.0 cm. The SPC ranged 22.5 to 37.7% in the parental genotypes with a mean of 30.0% and a median of 29.9%, while in the hybrids it ranged between 16.5 to 31.4% with a mean of 25.7% and a median of 26.0%. The SOC of parental genotypes ranged between 32.7 and 52.3% with a mean of 43.5% and a median of 43.7%, while that of the hybrid genotypes ranged between 40.6 and 57.1% with a mean of 48.3% and a median of 48.0%. Lastly, GSL ranged 3.1 to 50.5 μmol g^{-1} in the parental genotypes with a mean of 18.6 μmol g^{-1} and a median of 17. μmol g^{-1} and 0.19 to 24.7 μmol g^{-1} in the hybrid genotypes with a mean of 9.63 μmol g^{-1} and a median of 10.0 μmol g^{-1}. See a more detailed summary of the phenotype data of the parental genotypes in separate site-years in the Appendix (Table S3.3).

The best linear unbiased predictions (BLUPs) corrected the unbalance in the raw data by taking the site-year effect and the unequal replications into account (Table 3.3). The YLD of the combined population ranged between 1276.9 and 2813.2 kg ha^{-1} with a mean of 2161.0 kg ha^{-1}. Height varied from 82.2 to 122.6 cm with a mean of 106.9 cm. Seed protein content ranged from 23.6 to 29.7% with a mean of 26.1% while the SOC of the combined population ranged from 43.4 to 53.0% with

Table 3.3 A summary of phenotype best linear unbiased predictions (BLUPs) of 30 B-lines, 60 R-lines and 345 hybrid genotypes derived from the 91 parental genotypes in the combined *Brassica napus* L. population. The computation of BLUPs were based on field experiments across Canadian Prairies conducted in 2014-2018.

Trait	Min	Max	Mean	SD	CV (%)
YLD[1] (kg ha^{-1})	1276.86	2813.16	2160.96	235.57	10.90
HT[2] (cm)	82.21	122.55	106.91	6.45	6.04
SPC[3] (%)	23.59	29.72	26.05	0.99	3.82
SOC[4] (%)	43.41	53.03	48.97	1.51	3.08
GSL[5] (μmol g^{-1})	7.73	37.32	16.65	4.76	28.57

[1] Yield.
[2] Plant height.
[3] Seed protein content.
[4] Seed oil content.
[5] Seed glucosinolate content.

a mean of 49.0%. And lastly, GSL ranged from 7.7 to 37.3 μmol g^{-1} with a mean of 16.7 μmol g^{-1}.

All traits were significantly correlated with each other (Figure 3.1). HT and SOC had significant positive correlations with YLD, while SPC and GSL had significant negative correlations with YLD ($p < 0.001$). SOC and SPC had a significant negative correlation ($r = -0.900$, $p < 0.001$). SOC was also significantly negatively correlated with GSL ($r = -0.525$, $p < 0.001$).

Principal component analysis (PCA) was conducted based on the LD-pruned markers of all genotypes (Figure 3.2). PC1, PC2 and PC3 explained 93.51% of the phenotypic variation. While there were clear clusters for B- and R-lines in (A) PC1 vs. PC2 and (C) PC1 vs PC3, they overlapped in (B) PC2 vs. PC3 and could not be separated from each other. Although hybrid genotypes tended to cluster together, it was difficult to separate the hybrids from their parental genotypes.

3.4.2 Marker density

Marker densities of each chromosome, A- and C- subgenomes and the whole genome were calculated by dividing the distance by the total number of markers and expressing the marker density as kb/marker (Clarke et al. 2016). Marker density varied among the three marker sets. Based on MS-1, the whole genome had dense coverage (Figure 3.3.A). Marker density ranged from 10 kb/marker on chromosome A7 to 49 kb/marker on chromosome C9 (Table 3.4). Compared to the C-subgenome (28 kb/marker), the A-subgenome had a higher marker density (20 kb/marker) on average, consistent with Clarke et al. (2016). For the whole genome, marker density was 24 kb/marker.

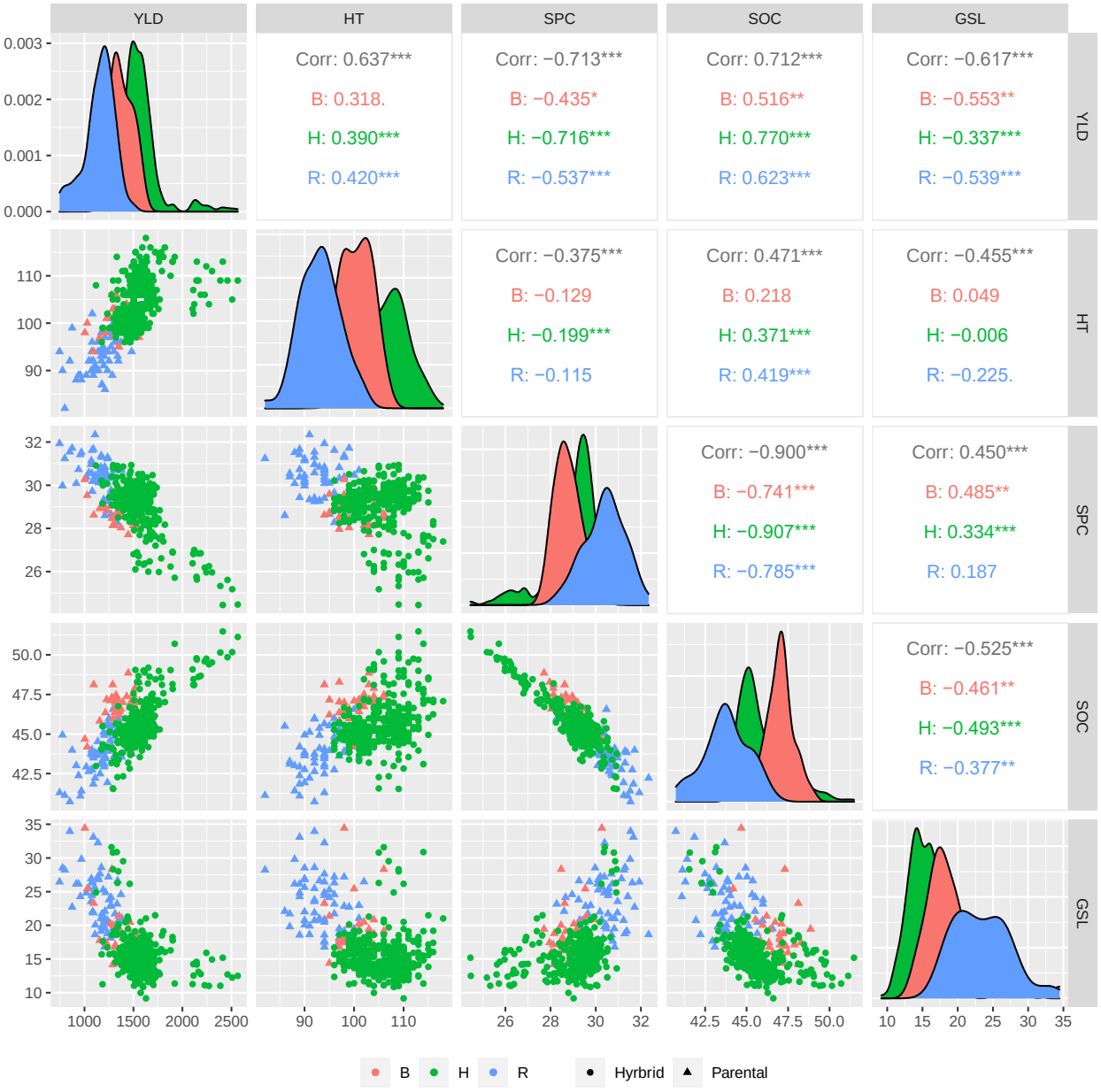


Figure 3.1 Correlation matrix of the traits based on the phenotype best linear unbiased predictions (BLUPs) from the combined *Brassica napus* L. population including all the parental and hybrid genotypes. The computation of BLUPs were based on field experiments across Alberta, Saskatchewan and Manitoba conducted in 2014-2018. The 31 B-lines, 60 R-lines and 345 hybrids are represented by red, blue and green, respectively. The upper half of the panel shows the correlations among the traits. The level of significance is noted by asterisks. The diagonal shows the distribution of the phenotype BLUPs of all traits. The lower half of the panel shows the scatterplot of the traits and each data point represents the BLUP for a genotype. Abbreviations: YLD: seed yield; HT: plant height; SPC: seed protein; SOC: seed oil content; GSL: seed glucosinolate content.

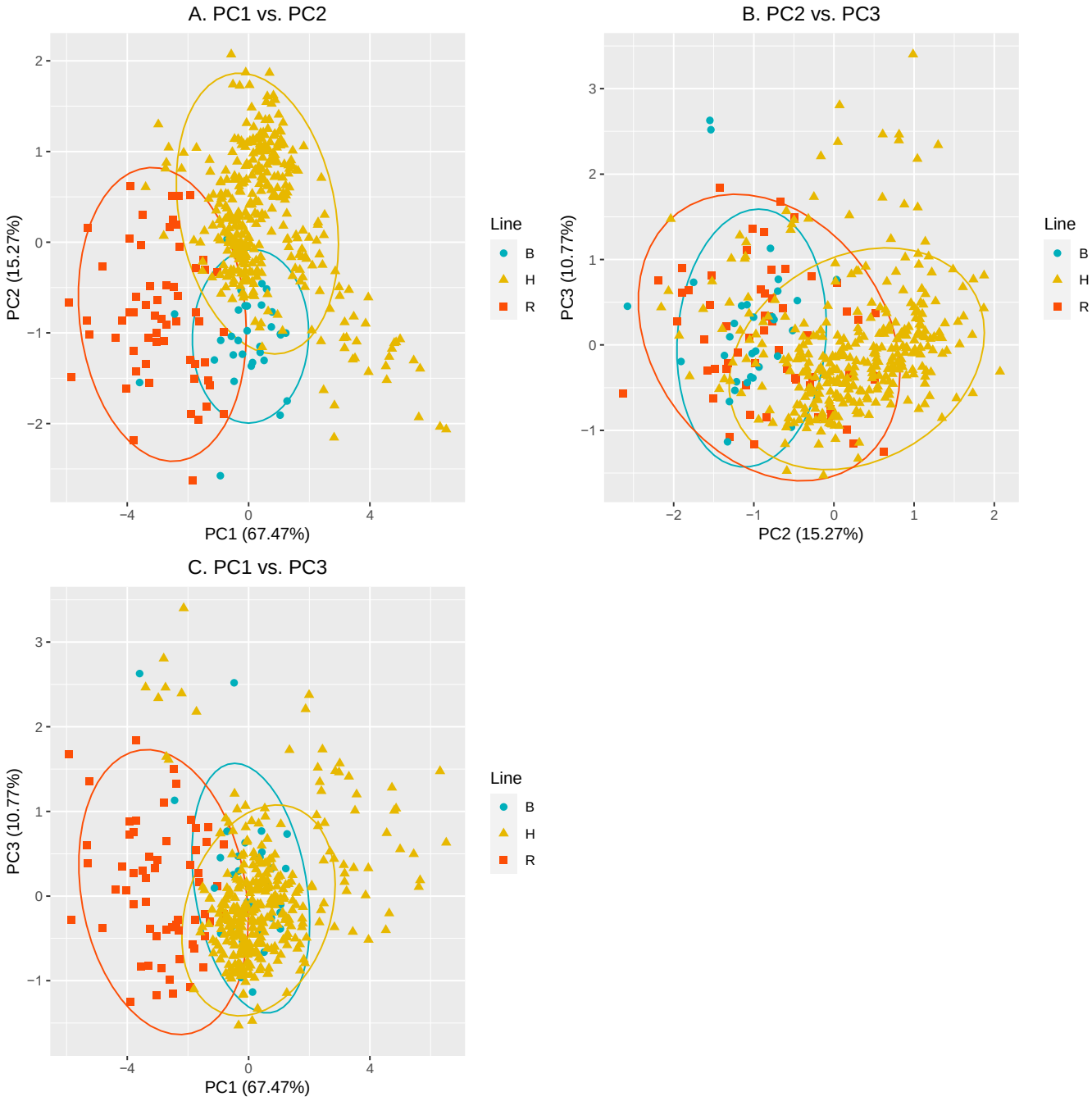
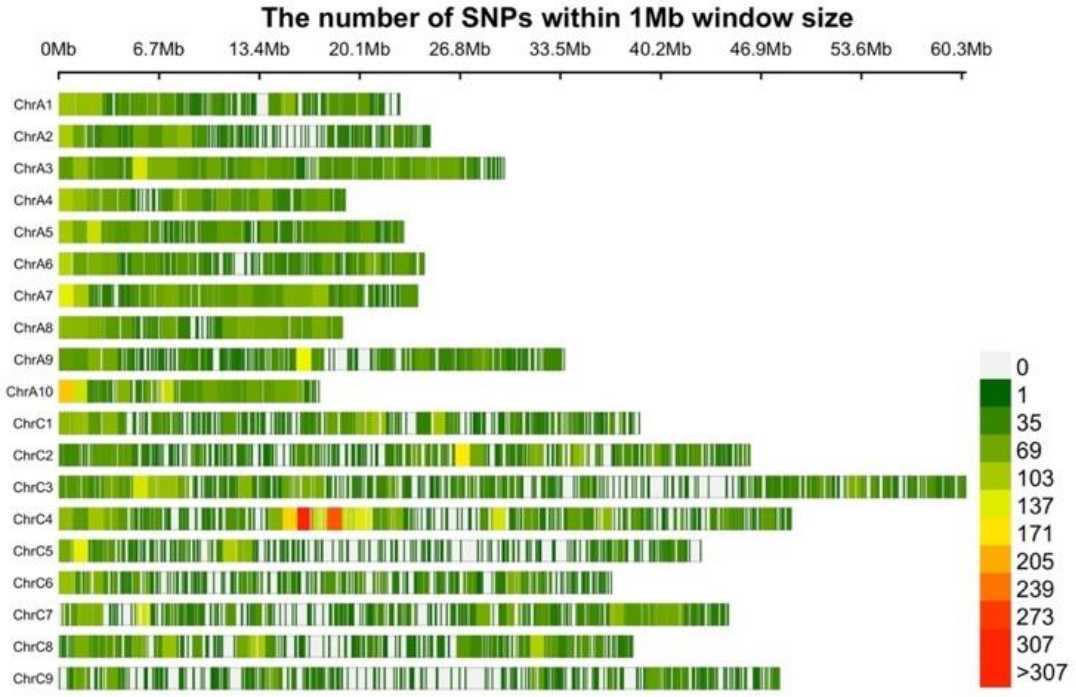
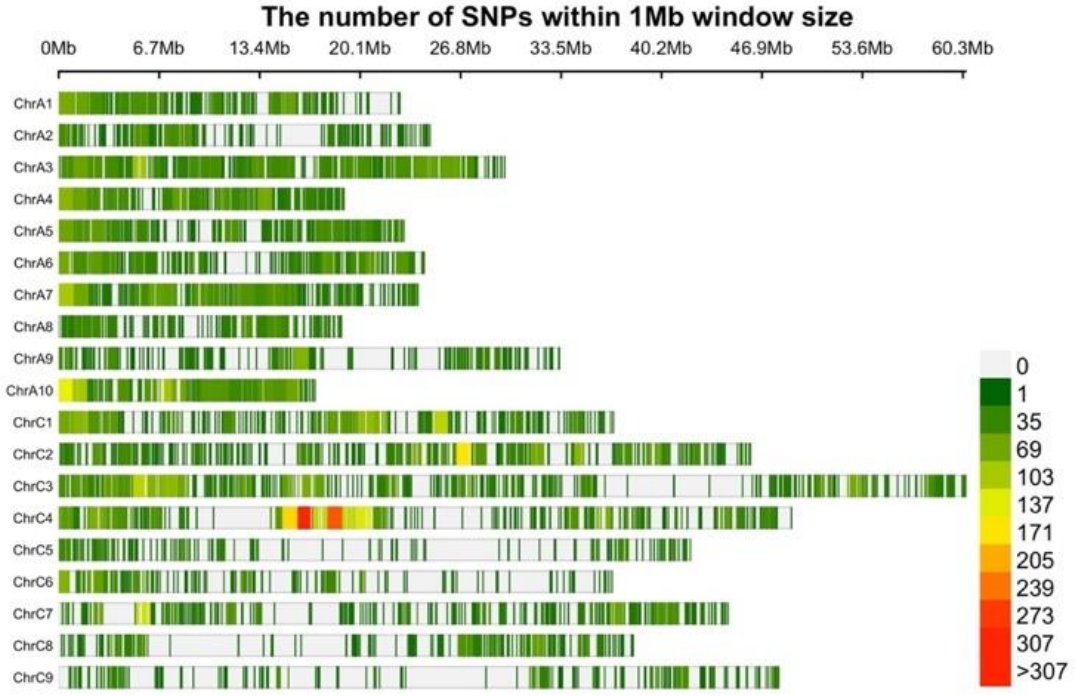


Figure 3.2 Principal component analysis (PCA) showing the subpopulation structure of a *Brassica napus* L. population consisting of 30 B-lines (blue circles), 61 R-lines (red squares) and 345 hybrids (yellow triangles) based on the phenotype best linear unbiased predictions (BLUPs). Three PCs are shown: (A) PC1 VS. PC2; (B) PC2 VS. PC3; (C) PC1 VS. PC3.

A.


The number of SNPs within 1Mb window size
0Mb
6.7Mb
13.4Mb
20.1Mb
26.8Mb
33.5Mb
40.2Mb
46.9Mb
53.6Mb
60.3Mb
ChrA1
ChrA2
ChrA3
ChrA4
ChrA5
ChrA6
ChrA7
ChrA8
ChrA9
ChrA10
ChrC1
ChrC2
ChrC3
ChrC4
ChrC5
ChrC6
ChrC7
ChrC8
ChrC9
0
1
35
69
103
137
171
205
239
273
307
>307

B.


The number of SNPs within 1Mb window size
0Mb
6.7Mb
13.4Mb
20.1Mb
26.8Mb
33.5Mb
40.2Mb
46.9Mb
53.6Mb
60.3Mb
ChrA1
ChrA2
ChrA3
ChrA4
ChrA5
ChrA6
ChrA7
ChrA8
ChrA9
ChrA10
ChrC1
ChrC2
ChrC3
ChrC4
ChrC5
ChrC6
ChrC7
ChrC8
ChrC9
0
1
35
69
103
137
171
205
239
273
307
>307

C.

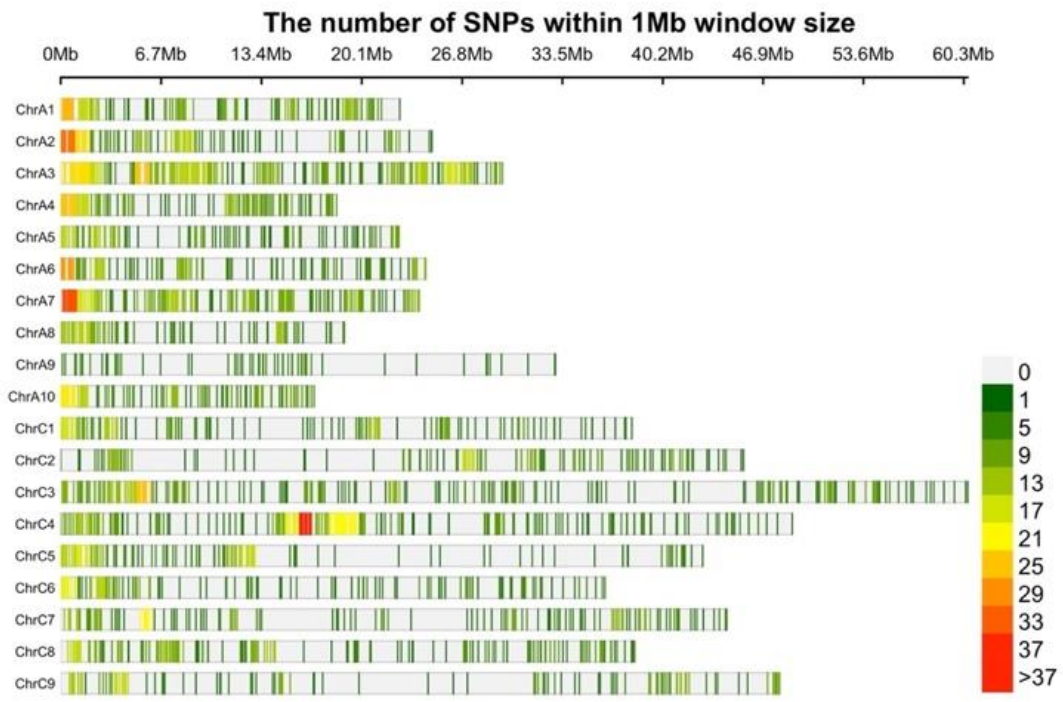


Figure 3.3 Marker density distribution in a *Brassica napus* L. population consisting of 31 B-lines, 60 R-lines and the 345 hybrids. Colour scale represents number of SNP markers per megabase. (A) Marker distribution based on MS-1 (26,651 SNP markers); (B) Marker distribution based on MS-2 (16,855 SNP markers); (C) Marker distribution based on LD-pruned markers (3,205 SNP markers).

Table 3.4 Marker density on each chromosome, subgenome and the whole genome based on a *Brassica napus* L. population consisting of 31 B-lines, 60 R-lines and the 345 hybrids, calculated using MS-1 (26,651 SNP markers).

Chromosome	Distance (kb)	Total number of markers	Mean marker density (kb/marker)
Whole Genome	642,535	26,651	24
A-subgenome	237,707	12,182	20
C-subgenome	404,827	14,469	28
A1	22,744	1,077	21
A2	24,768	905	27
A3	29,728	1,680	18
A4	19,104	418	46
A5	23,015	1,236	19
A6	24,381	1,157	21
A7	23,912	2,462	10
A8	18,931	1,086	17
A9	33,757	835	40
A10	17,366	1,242	14
C1	38,752	1,528	25
C2	46,131	1,711	27
C3	60,522	2,183	28
C4	48,891	2,896	17
C5	42,857	992	43
C6	36,888	1,179	31
C7	44,462	1,582	28
C8	38,243	1,419	27
C9	48,080	979	49

Marker density based on MS-2 was generally lower than using MS-1 (Figure 3.3.B), which was expected since the total number of markers dropped by 9,796. Marker density ranged from 17 kb/marker on chromosome A10 to 103 kb/marker on chromosome C5 (Table 3.5). On average, marker densities of the A-, C- subgenomes and the whole genome were 31 kb/marker, 43 kb/marker and 38 kb/marker, respectively.

Compared with MS 1 and 2, LD pruned markers had even lower marker densities, which resulted from a drastically reduced number of markers (Figure 3.3.C). Using LD pruned markers, mean marker densities of the A- and C- subgenomes and the whole genome were 138 kb, 239 kb and 188 kb, respectively (Table 3.6). It was consistent with the results from MS-1 and MS-2 that the A-subgenome had relatively higher marker densities than the C-subgenome.

3.4.3 Population structure

The results from STUCTURE were visualized and revealed peaks at k=2 (i.e. two subdivisions) for both the parental population and the combined population (Figure 3.4). This indicated that both populations can be grouped into two clusters.

PCA analysis was then performed to evaluate the population structure (Figure 3.5). The first 15 principal components explained 24.61% of the total genotypic variation while PC1, PC2 and PC3 explained 8.20% combined. There were no clear clusters formed based on the type of individuals (i.e. B-lines, R-lines or hybrids) although Figure 3.5.C showed that the majority of the parental genotypes gathered in the upper portion of the figure. The hybrid genotypes were evenly distributed among the parental genotypes. This indicated that there was no clear stratification in the target population.

Table 3.5 Marker density on each chromosome, subgenome and the whole genome based on a *Brassica napus* L. population consisting of 31 B-lines, 60 R-lines and the 345 hybrids, calculated using MS-2 (16,855 SNP markers).

Chromosome	Distance (kb)	Total number of markers	Mean marker density (kb/marker)
Whole Genome	637,850	168,55	38
A-subgenome	236,761	7521	31
C-subgenome	401,089	9,334	43
A1	22,744	688	33
A2	24,727	446	55
A3	29,728	1,202	25
A4	19,008	805	24
A5	22,986	698	33
A6	24,349	764	32
A7	23,911	884	27
A8	18,848	468	40
A9	33,376	557	60
A10	17,083	1,009	17
C1	36,957	1,134	33
C2	46,025	1,371	34
C3	60,447	1,707	35
C4	48,859	2,082	23
C5	42,113	410	103
C6	36,888	582	63
C7	44,386	963	46
C8	38,105	564	68
C9	47,310	521	91

Table 3.6 Marker density on each chromosome, subgenome and the whole genome based on a *Brassica napus* L. population consisting of 31 B-lines, 60 R-lines and the 345 hybrids, calculated using LD-pruned markers (3,205 SNPs).

Chromosome	Distance (kb)	Total number of markers	Mean marker density (kb/marker)
Whole Genome	636,256	3376	188
A-subgenome	234,682	1,699	138
C-subgenome	401,574	1,677	239
A1	22,631	167	136
A2	24,768	157	158
A3	29,490	309	95
A4	18,381	309	59
A5	22,544	122	185
A6	24,354	137	178
A7	23,796	218	109
A8	18,921	97	195
A9	32,903	58	567
A10	16,892	125	135
C1	38,121	153	249
C2	45,581	160	285
C3	60,434	276	219
C4	48,842	290	168
C5	42,857	142	302
C6	36,258	147	247
C7	44,219	181	244
C8	37,841	191	198
C9	47,422	137	346

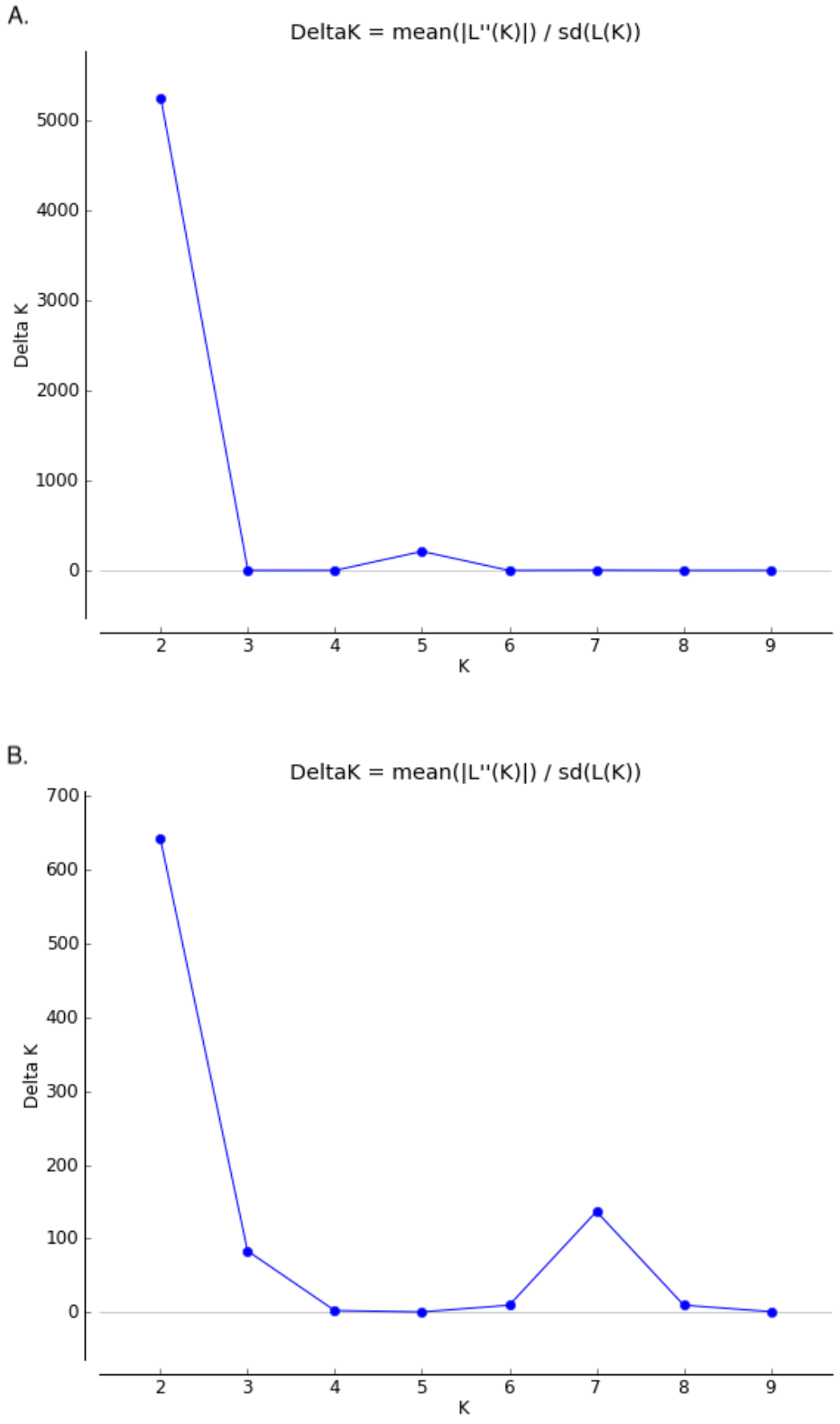


Figure 3.4 Population structure of two *Brassica napus* L. populations. (A) subpopulation of the parental genotypes consisting of 31 B-lines and 60 R-lines. (B) represents the combined population consisting of the parental population and 345 hybrid genotypes.

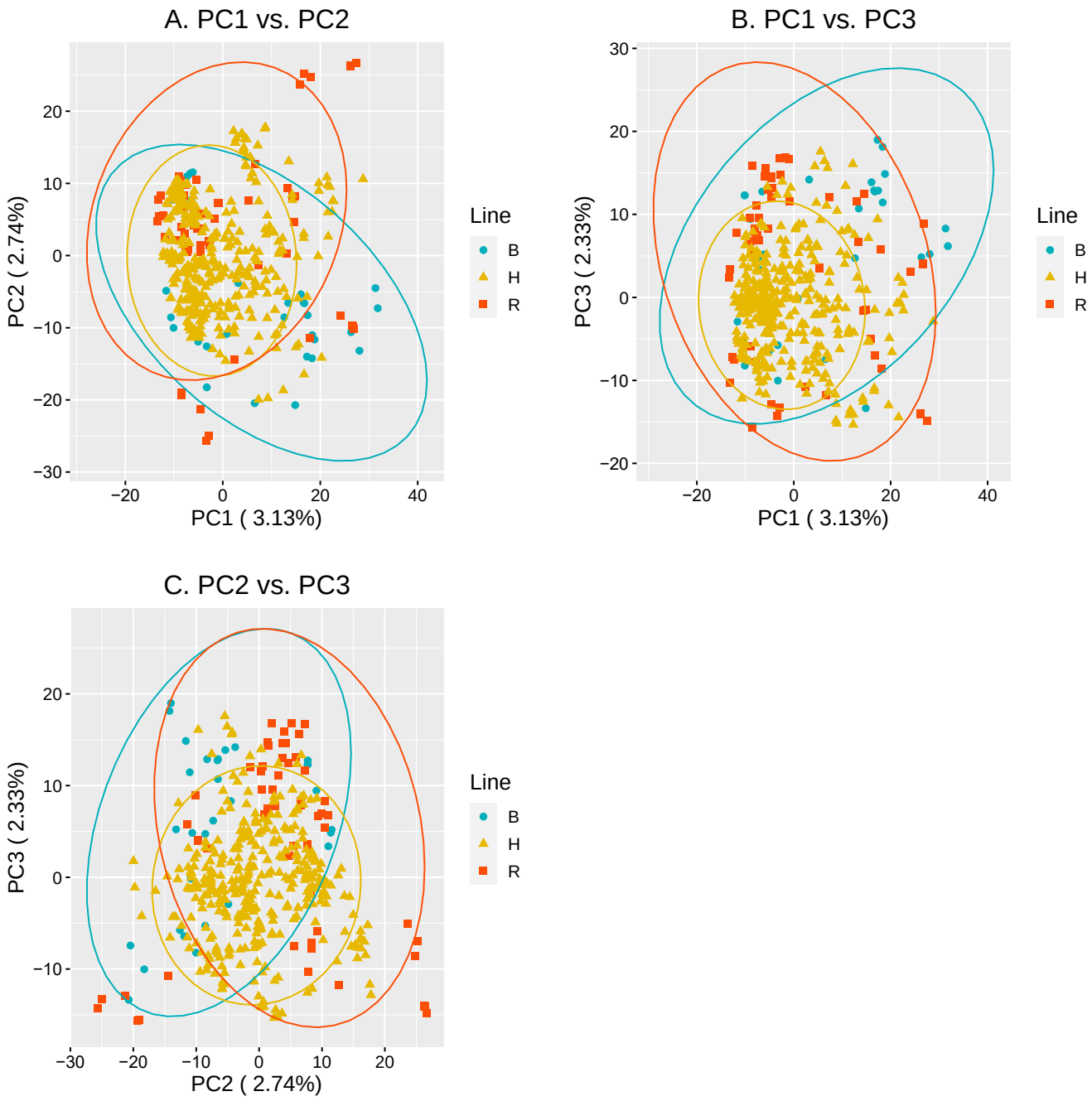


Figure 3.5 Principal component analysis (PCA) on the *B. napus* population consisting of 31 B-lines, 60 R-lines and 345 hybrids based on the LD-pruned markers (3,205 SNPs). A represents PC1 vs. PC2; B represents PC1 vs. PC3 and C represents PC2 vs. PC3.

3.4.4 Linkage disequilibrium

The threshold for linkage disequilibrium (LD) decay was set as $r^2 = 0.2$. Based on MS-1, the whole genome decayed at ~4.0 Mb, and the A- and C-subgenome decayed at ~400.0 kb and 5.6 Mb, respectively (Figure 3.6A). Based on MS-2 the whole genome decayed at ~5.2 Mb, and the A- and C-subgenome decayed at ~600.0 kb and 6.0 Mb, respectively (Figure 3.6B). The LD decay generally occurred faster on chromosomes 1 to 10 (the A-subgenome) compared to chromosomes 11 to 19 (the C-subgenome) (Figure 3.7). Based on MS-1, the intrachromosomal LD decay within A- and C- subgenomes varied between 350 kb ~ 1.1 Mb and 250 kb ~ 7.2 Mb, respectively (Figure 3.7.A). Based on MS-2, the intrachromosomal LD decay within A- and C-subgenomes varied between 250 kb ~ 2.1 Mb and 450 kb ~ 9.0 Mb, respectively (Figure 3.7.B).

3.4.5 Model comparison

3.4.5.1 Parental population

Six GWAS models used to identify significant MTAs in this study were CMLM, FarmCPU, MLMM, MLM+K, MLM+K+Q, MLM+PCA+Q. For the parental population, the six models performed similar (Figure 3.8 and Figure 3.9) with two groups based on the RMSE values (Table 3.7). The first group consisted of the three mixed linear models: MLM+K, MLM+K+Q and MLM+K+PCA, and the second group consisted of MLMM, FarmCPU and CMLM. The second group generally had smaller RMSE values, which indicated that this group was more accurate than the first group. Root mean square errors of the models varied between 0.03 (MLM+K on GSL) and 0.32 (MLM+K+PCA on SPC) (Table 3.7), indicating the fitness of the GWAS models differed depending on the trait of interest. In the Q-Q plots for each trait, the x-axis represented the negative logarithms of the *p* values from the six GWAS models and the y axis represented their expected value under the null hypothesis, which assumed no association between SNP markers with the trait.

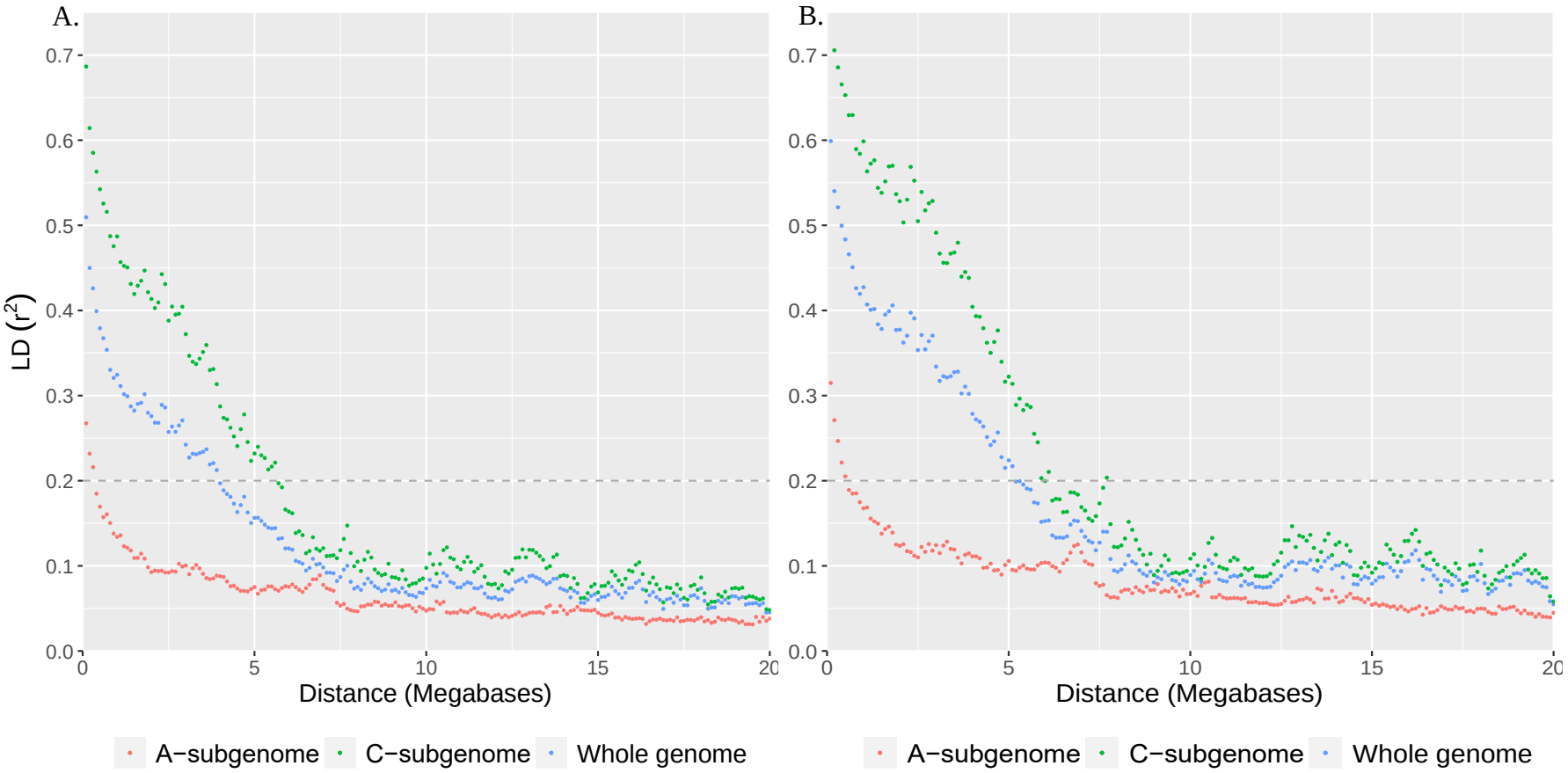


Figure 3.6 Linkage disequilibrium (LD) decay of the whole genome (blue line), A-subgenome (red line) and C-subgenome (green line) evaluated in a *Brassica napus* L. population consisting of 31 B-lines, 60 R-lines and 345 hybrids based on MS-1 that contained 26,651 SNP markers (A) and MS-2 that contained 16,855 SNP markers (B).

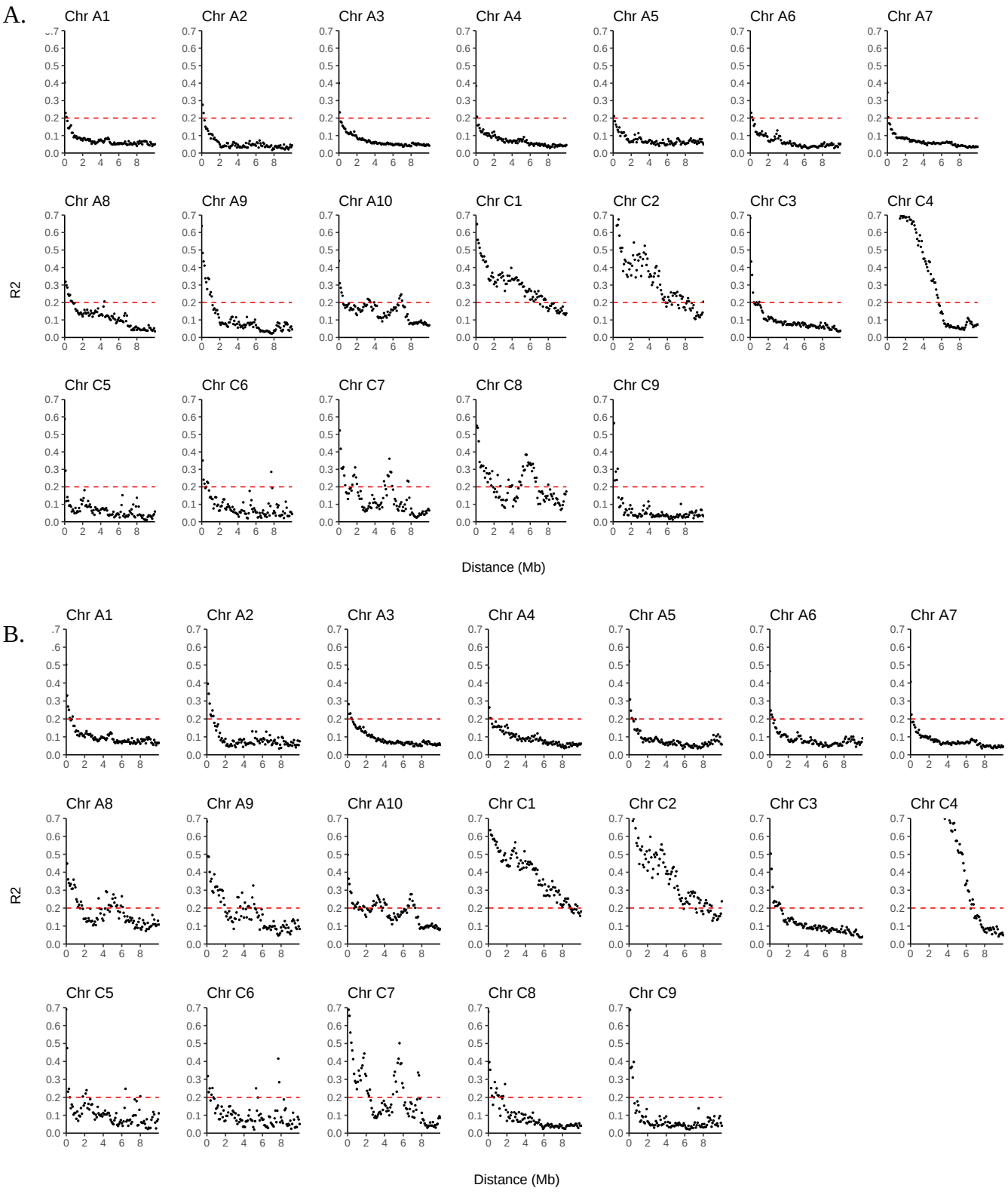
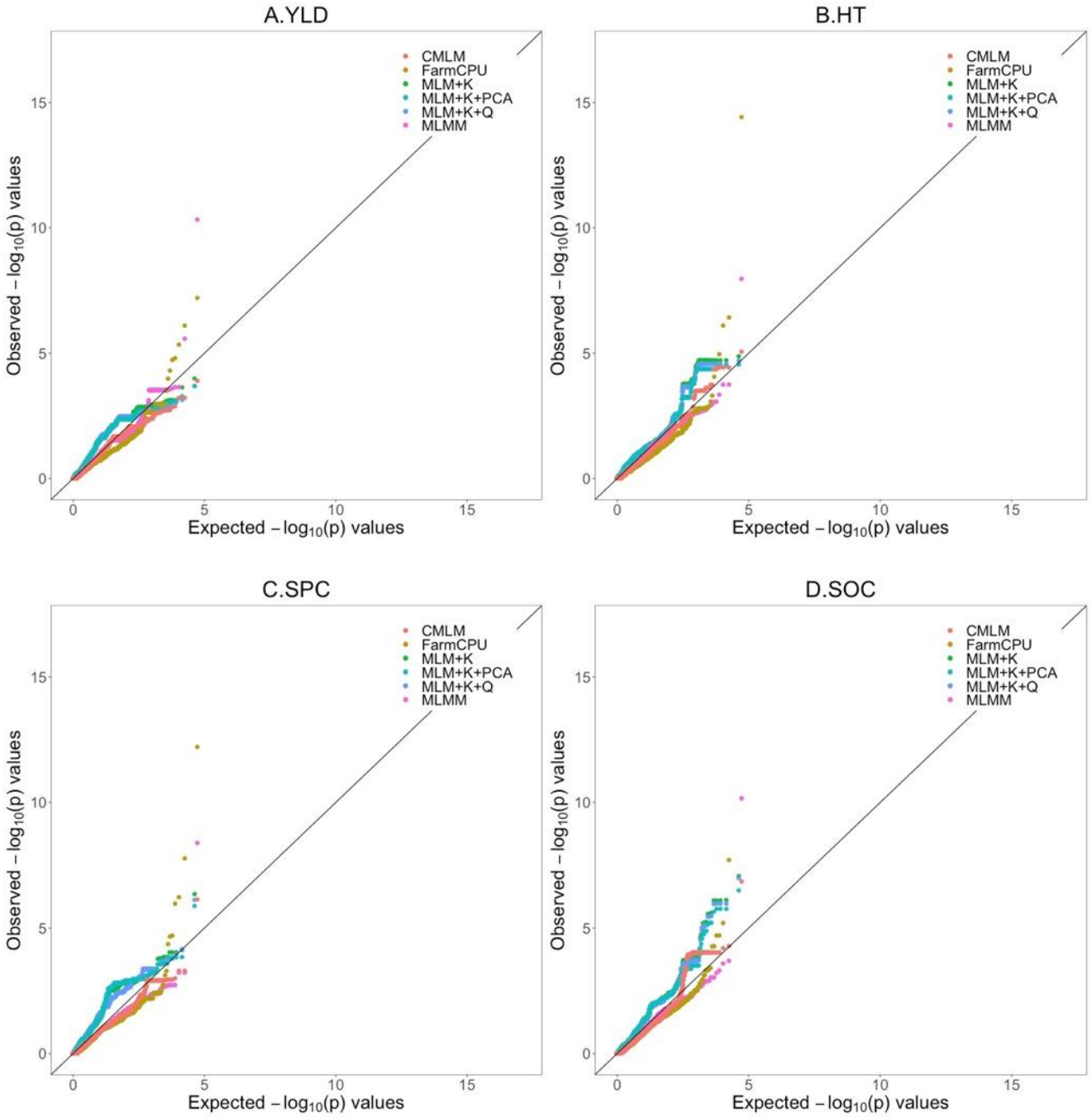


Figure 3.7 Linkage disequilibrium decay plots of all 19 chromosomes (chromosomes A1 to C9) in a *Brassica napus* L. population consisting of 31 B-lines, 60 R-lines and 345 hybrids based on MS-1 (26,651 SNPs) (A) and MS-2 (16,855 SNPs) (B).


A.YLD
B.HT
C.SPC
D.SOC
CMLM
FarmCPU
MLM+K
MLM+K+PCA
MLM+K+Q
MLMM
Observed – log10(p) values
Expected – log10(p) values
0
5
10
15

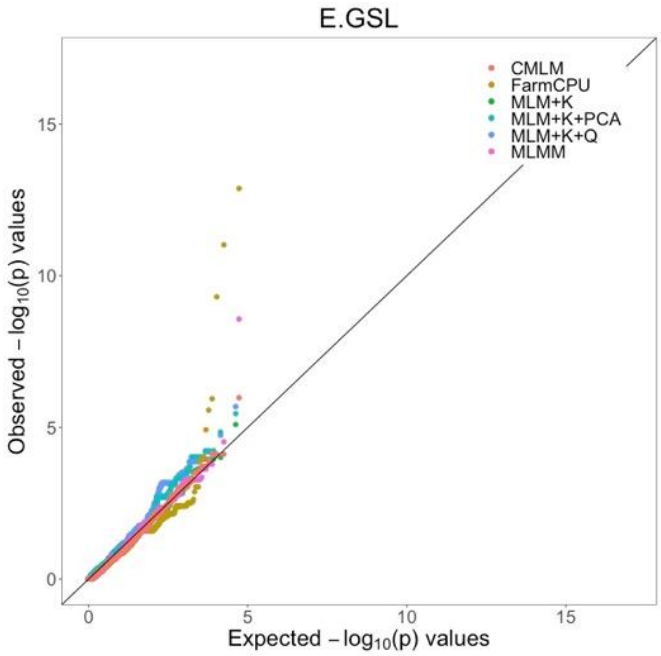
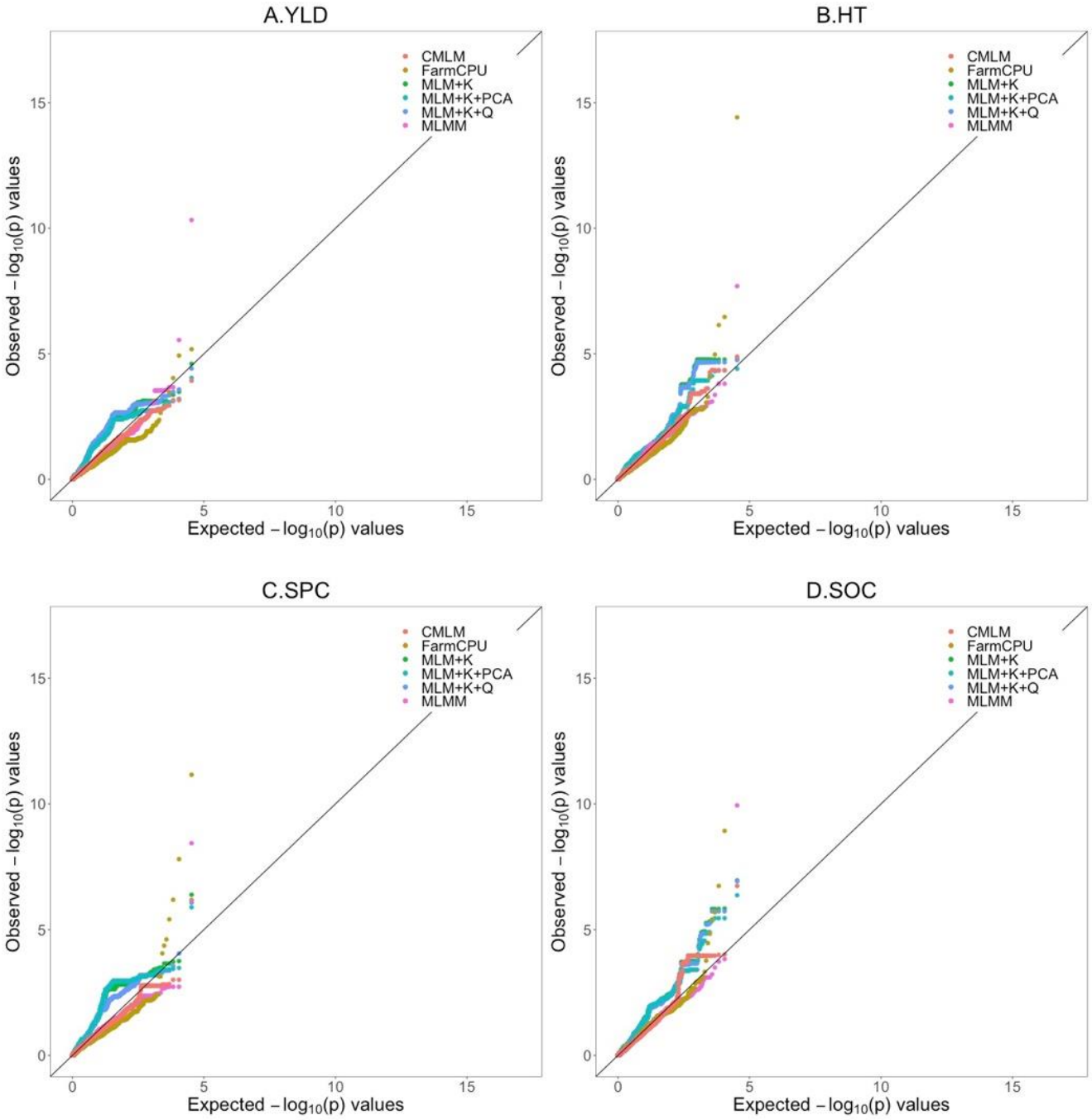


Figure 3.8 Model performance comparison among CMLM, FarmCPU, MLMM, MLM+K, MLM+K+Q, MLM+PCA+Q for the *Brassica napus* L. parental population based on MS-1 (26,651 markers). Observed p values were plotted against the expected p values in the quantile-quantile plots on all five traits (A) YLD (seed yield); (B) HT (plant height); (C) SPC (seed protein content); (D) SOC (seed oil content); (E) GSL (seed glucosinolate content). Abbreviations of models: CMLM: compression mixed linear model; FarmCPU: fixed and random model circulating probability unification; MLM: mixed linear model; K: kinship matrix; Q: population structure matrix based on Bayesian clustering; PCA: principal component analysis; and MLMM: multi-locus mixed linear model.


A.YLD
B.HT
C.SPC
D.SOC
CMLM
FarmCPU
MLM+K
MLM+K+PCA
MLM+K+Q
MLMM
Observed – log10(p) values
Expected – log10(p) values
0
5
10
15

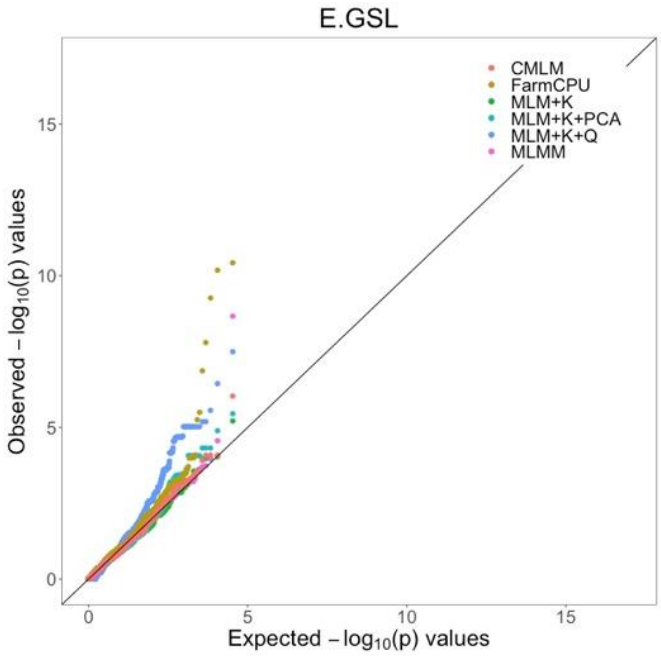


Figure 3.9 Model performance comparison among CMLM, FarmCPU, MLMM, MLM+K, MLM+K+Q, MLM+PCA+Q for the *Brassica napus* L. parental population based on MS-2 (16,855 markers). Observed p values were plotted against the expected p values in the quantile-quantile plots on all five traits (A) YLD (seed yield); (B) HT (plant height); (C) SPC (seed protein content); (D) SOC (seed oil content); (E) GSL (seed glucosinolate content). Abbreviations of models: CMLM: compression mixed linear model; FarmCPU: fixed and random model circulating probability unification; MLM: mixed linear model; K: kinship matrix; Q: population structure matrix based on Bayesian clustering; PCA: principal component analysis; and MLMM: multi-locus mixed linear model.

Table 3.7 Root mean square error (RMSE) values of GWAS models applied on the parental *Brassica napus* L. population consisting of 31 B-lines and 60 R lines. Traits evaluated included seed yield (YLD), plant height (HT), seed protein content (SPC), seed oil content (SOC) and seed glucosinolate content (GSL).

Markers	Model	YLD	HT	SPC	SOC	GSL
MS-1 (26,651)	MLM+K[1]	0.24	0.13	0.29	0.19	0.03
	MLM+K+Q[2]	0.23	0.12	0.28	0.17	0.08
	MLM+K+PCA[3]	0.22	0.15	0.32	0.18	0.05
	MLMM[4]	0.10	0.11	0.13	0.13	0.11
	FarmCPU[5]	0.17	0.21	0.18	0.18	0.14
	CMLM[6]	0.11	0.12	0.12	0.14	0.10
MS-2 (16,855)	MLM+K	0.30	0.14	0.37	0.23	0.04
	MLM+K+Q	0.34	0.14	0.29	0.21	0.30
	MLM+K+PCA	0.25	0.16	0.41	0.23	0.06
	MLMM	0.09	0.05	0.06	0.06	0.05
	FarmCPU	0.18	0.15	0.16	0.13	0.11
	CMLM	0.06	0.05	0.06	0.09	0.04

[1] mixed linear model considering kinship.
[2] mixed linear model considering kinship and subpopulation structure via Bayesian clustering.
[3] mixed linear model considering kinship and subpopulation structure via principal component analysis.
[4] multi-locus mixed linear model.
[5] fixed and random model circulating probability unification.
[6] compression mixed linear model.

For the parental population (91 genotypes) the performance of these six models based on MS-1 were similar to each other especially in the case of glucosinolate content (Figure 3.8).

This was the same for the parental genotypes based on MS-2 as well. Again, there was variation in the RMSE values of the models using MS-2, where the lowest RMSE was identified on GSL content based on MLM+K model (0.04) and the highest RMSE on SPC based on MLM+K+PCA model (0.41) (Table 3.7). All six models performed similar for each trait (Figure 3.9) with slight differences among traits. This indicated that different traits had dissimilar responses to the choice of a particular GWAS model, which would result in variation in the accuracy. All models were included in the process of MTA identification based on the parental population.

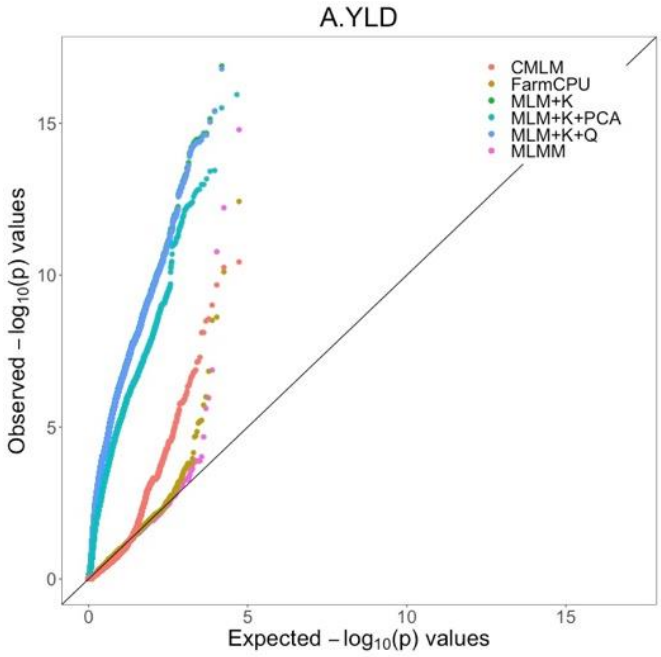
3.4.5.2 Combined population

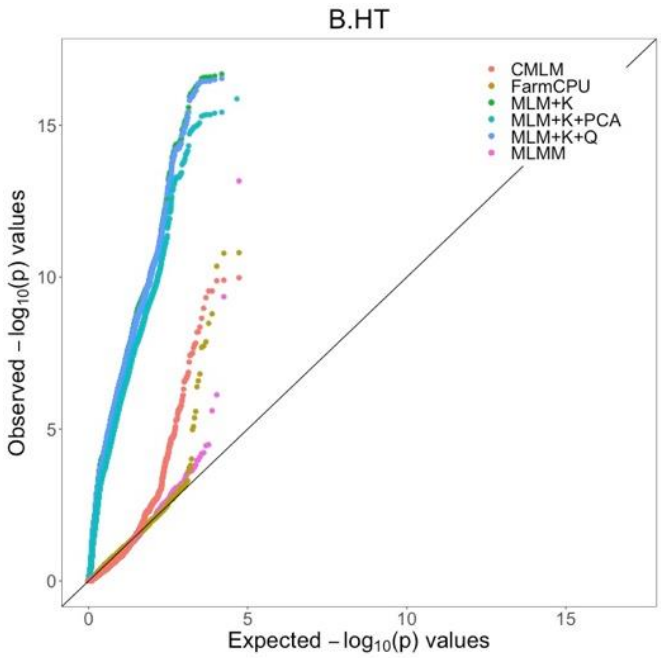
In terms of model performance based on all 436 genotypes, there were clear differences based on the choice of models. The RMSE values were larger than that of the parental population (Table 3.8), which suggested that there were larger differences between observed and predicted values in the combined population compared to the training population (Nakano et al. 2020). The lowest RMSE value based on MS-1 was 0.07 (SPC based on MLMM) and the highest was 3.55 (HT based on MLM+K). The lowest RMSE value based on MS-2 was 0.03 (SOC based on CLMM) and the highest was 4.53 (HT based on MLM+K+Q). Larger deviations were observed in the Q-Q plots for all traits with both MS-1 and 2 (Figure 3.10 and Figure 3.11). Overall, CMLM, FarmCPU and MLMM performed better than the MLM models for all five traits. Therefore, only three models (CMLM. FarmCPU, MLMM) were applied in the identification of significant MTAs to avoid false positives.

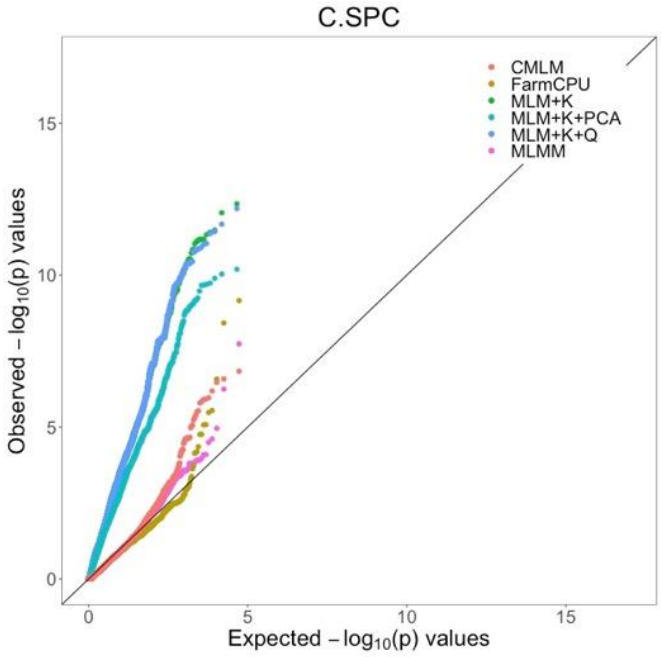
Table 3.8 Root mean square error (RMSE) values of six GWAS models applied on the combined population of *Brassica napus* L. (436 genotypes). Traits evaluated included seed yield (YLD), plant height (HT), seed protein contend (SPC), seed oil content (SOC) and seed glucosinolate content (GSL).

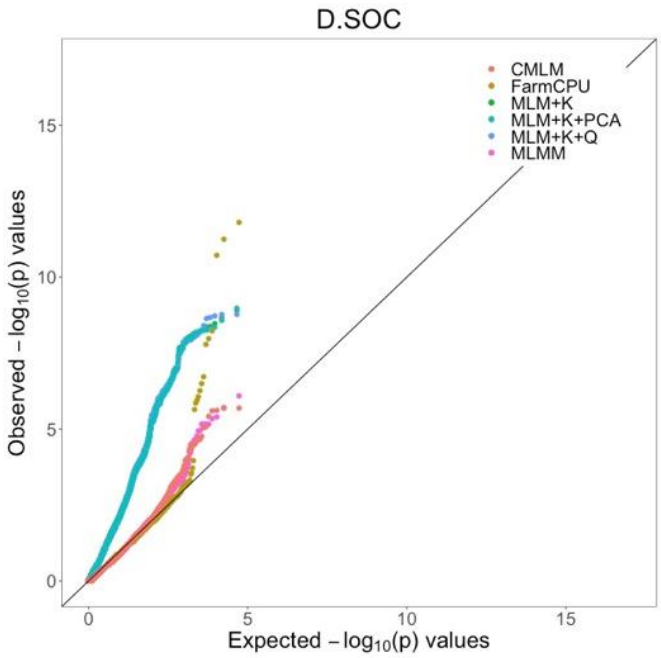
Markers	Model	YLD	HT	SPC	SOC	GSL
MS-1 (26,651)	MLM+K[1]	3.44	3.55	0.81	1.53	2.72
	MLM+K+Q[2]	3.41	3.52	0.81	1.52	2.67
	MLM+K+PCA[3]	2.60	3.09	0.80	1.14	0.64
	MLMM[4]	0.11	0.11	0.07	0.08	0.17
	FarmCPU[5]	0.09	0.11	0.11	0.09	0.22
	CMLM[6]	0.25	0.23	0.08	0.11	0.22
MS-2 (16,855)	MLM+K	4.17	4.52	1.66	0.86	3.78
	MLM+K+Q	4.14	4.53	1.68	0.86	3.77
	MLM+K+PCA	3.31	4.00	1.58	0.77	3.43
	MLMM	0.12	0.10	0.04	0.09	0.21
	FarmCPU	0.14	0.15	0.11	0.12	0.49
	CMLM	0.20	0.23	0.07	0.03	0.20

[1] mixed linear model considering kinship.
[2] mixed linear model considering kinship and subpopulation structure via Bayesian clustering.
[3] mixed linear model considering kinship and subpopulation structure via principal component analysis.
[4] multi-locus mixed linear model.
[5] fixed and random model circulating probability unification.
[6] compression mixed linear model.


A.YLD
CMLM
FarmCPU
MLM+K
MLM+K+PCA
MLM+K+Q
MLMM
Observed – log10(p) values
Expected – log10(p) values
0
5
10
15


B.HT
CMLM
FarmCPU
MLM+K
MLM+K+PCA
MLM+K+Q
MLMM
Observed – log10(p) values
Expected – log10(p) values
0
5
10
15


C.SPC
CMLM
FarmCPU
MLM+K
MLM+K+PCA
MLM+K+Q
MLMM
Observed – log10(p) values
Expected – log10(p) values
0
5
10
15


D.SOC
CMLM
FarmCPU
MLM+K
MLM+K+PCA
MLM+K+Q
MLMM
Observed – log10(p) values
Expected – log10(p) values
0
5
10
15

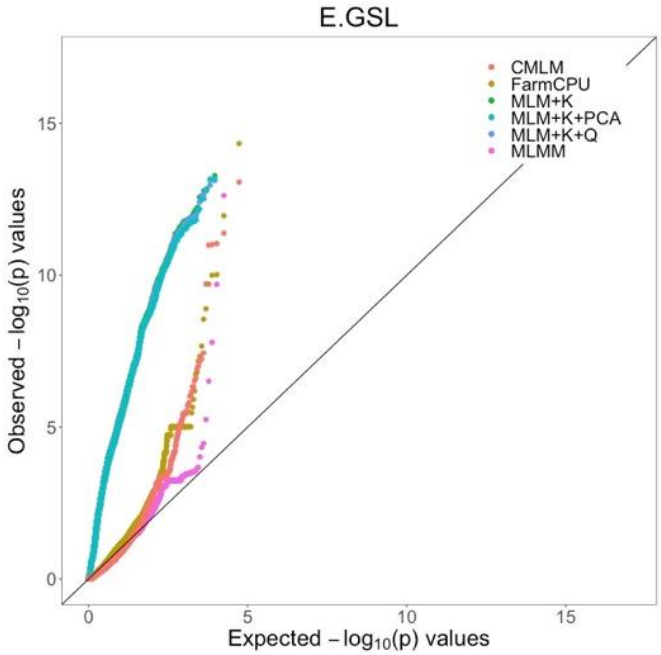
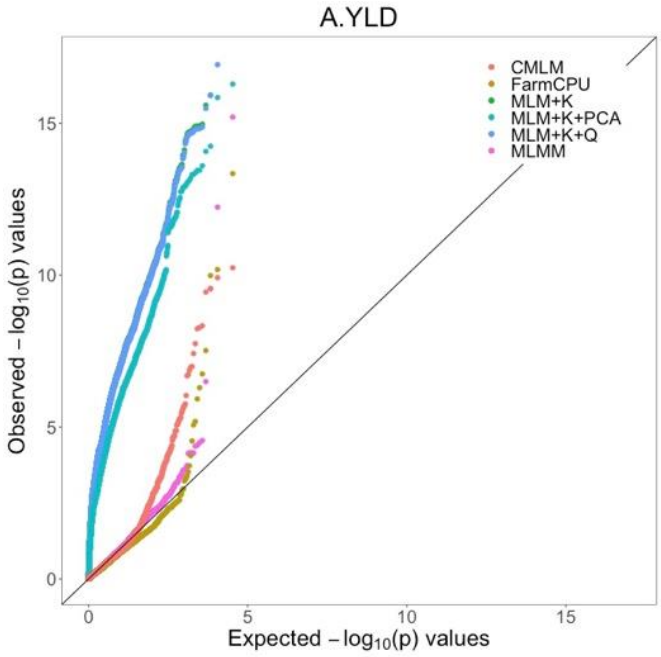
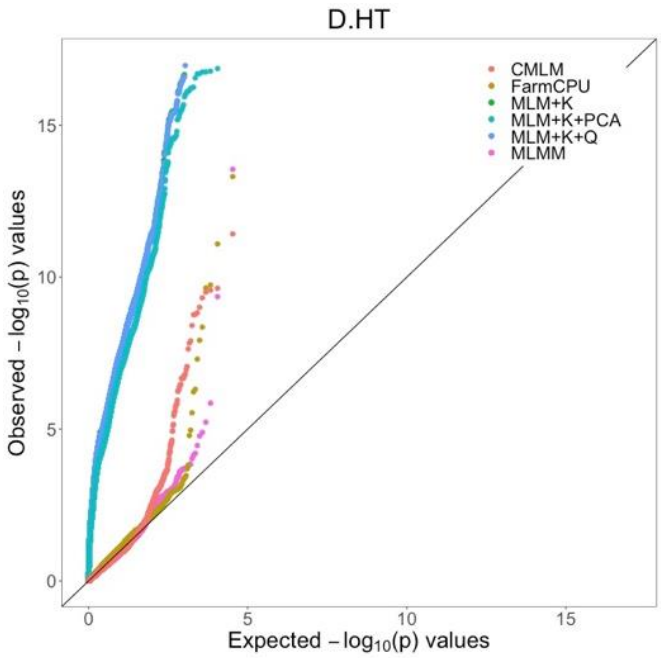
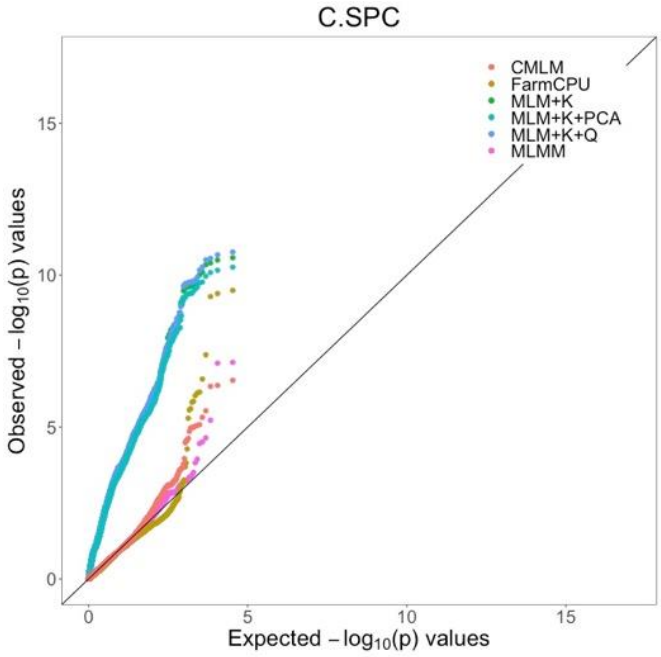
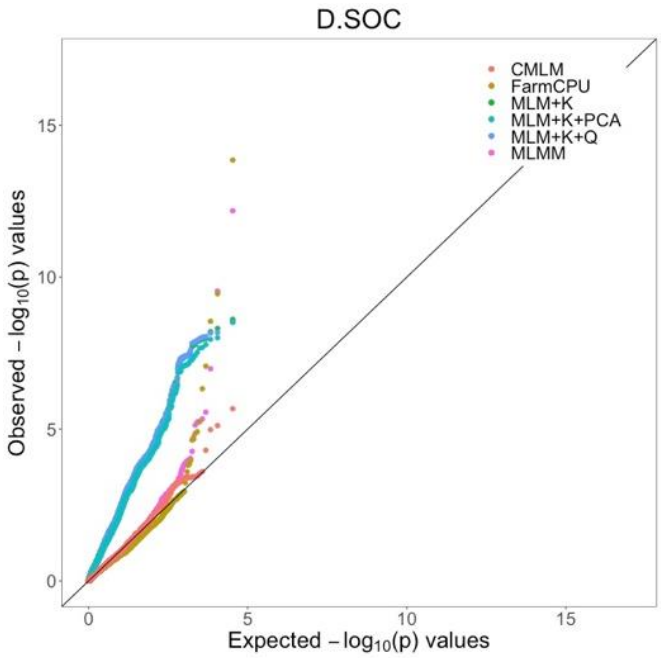


Figure 3.10 Model performance comparison among CMLM, FarmCPU, MLMM, MLM+K, MLM+K+Q, MLM+PCA+Q for the *Brassica napus* L. combined population based on MS-1 (26,651 markers). Observed p values were plotted against the expected p values in the quantile-quantile (Q-Q) plots on all five traits (A) YLD (seed yield); (B) HT (plant height); (C) SPC (seed protein content); (D) SOC (seed oil content); (E) GSL (seed glucosinolate content). Abbreviations: CMLM: compression mixed linear model; FarmCPU: fixed and random model circulating probability unification; MLM: mixed linear model; K: kinship matrix; Q: population structure matrix based on Bayesian clustering; PCA: principal component analysis; and MLMM: multi-locus mixed linear model.


A.YLD
CMLM
FarmCPU
MLM+K
MLM+K+PCA
MLM+K+Q
MLMM
Observed – log10(p) values
Expected – log10(p) values
0
5
10
15


D.HT
CMLM
FarmCPU
MLM+K
MLM+K+PCA
MLM+K+Q
MLMM
Observed – log10(p) values
Expected – log10(p) values
0
5
10
15


C.SPC
CMLM
FarmCPU
MLM+K
MLM+K+PCA
MLM+K+Q
MLMM
Observed – log10(p) values
Expected – log10(p) values
0
5
10
15


D.SOC
CMLM
FarmCPU
MLM+K
MLM+K+PCA
MLM+K+Q
MLMM
Observed – log10(p) values
Expected – log10(p) values
0
5
10
15

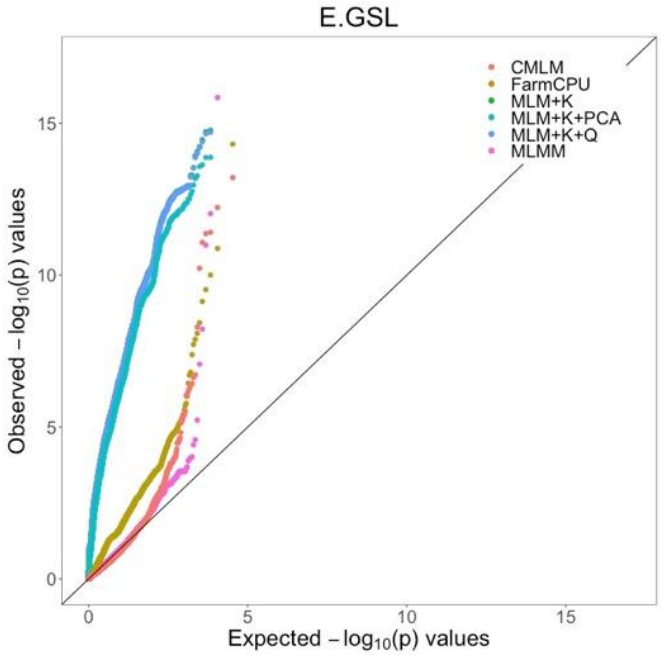


Figure 3.11 Model performance comparison among CMLM, FarmCPU, MLMM, MLM+K, MLM+K+Q, MLM+PCA+Q for the *Brassica napus* L. combined population based on MS-2 (16,855 markers). Observed p values were plotted against the expected p values in the quantile-quantile (Q-Q) plots on all five traits (A) YLD (seed yield); (B) HT (plant height); (C) SPC (seed protein content); (D) SOC (seed oil content); (E) GSL (seed glucosinolate content). Abbreviations of models: CMLM: compression mixed linear model; FarmCPU: fixed and random model circulating probability unification; MLM: mixed linear model; K: kinship matrix; Q: population structure matrix based on Bayesian clustering; PCA: principal component analysis; and MLMM: multi-locus mixed linear model.

3.4.6 Marker-trait association identification

After combing results from all traits there were 141 significant MTAs identified from the parental populations and the combined population, based on two sets of markers. Most of the significant MTAs were unique as per the population or the marker set, while some of the significant MTAs were shared between populations or marker sets (Figure 3.12). Based on the parental population, there were 13 consensus significant MTAs shared between MS-1 and MS-2. Based on the combined population, there were 17 consensus markers identified when comparing results from MS-1 and MS-2. The full lists of significant SNP markers identified from the parental and hybrid population based on two marker sets and their corresponding traits can be found in the Appendix (Tables S3.4-S.3.7).

Five common SNPs were identified as significant across populations as well as marker sets (Table 3.9). Interestingly, they were all significantly associated with SOC in the parental population, while in the combined population they were all significantly associated with GSL. Some of the significant MTAs identified were pleiotropic, meaning that they were found to be associated with multiple traits. More specifically, one pleiotropic and two pleiotropic SNPs were identified from the parental population based on MS-1 and MS-2, respectively. In the combined population, 33 and 12 pleiotropic SNPs were identified based on MS-1 and MS-2, respectively.

As mentioned earlier, different traits from the same population based on the same markers responded differently to the models. There was variation in the number of significant MTAs detected by the same model on the same trait when comparing across the populations and marker sets (Table 3.10).

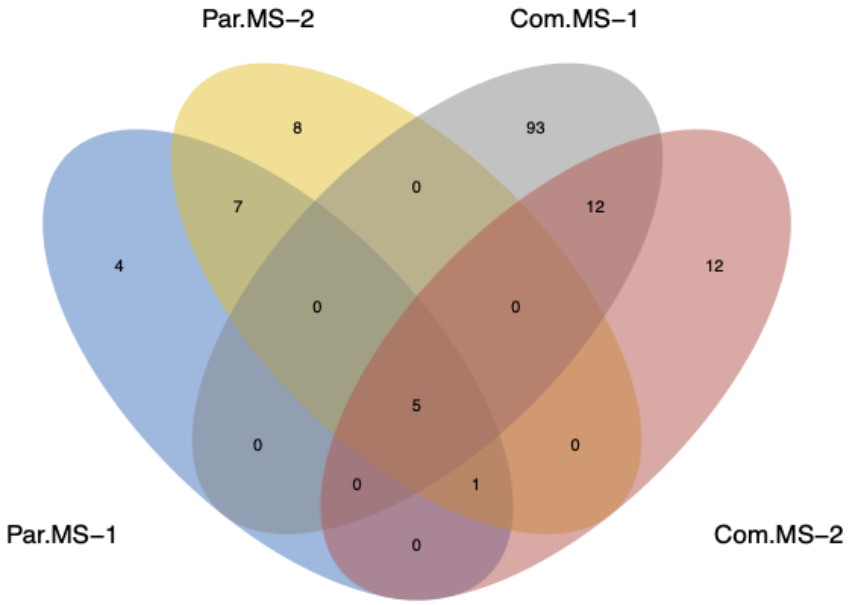


Figure 3.12 A Venn diagram demonstrating the number of significant marker-trait associations (MTAs) detected across the parental and combined populations of *Brassica napus* L. and marker sets 1 and 2 that contained 26,651 and 16,855 SNPs. Abbreviations: Par: parental population; Com: combined population; MS-1: marker set 1; MS-2: marker set 2.

Table 3.9 Significant MTAs commonly shared between the *Brassica napus* L. parental and combined population based on both MS-1 (26,651 SNP markers) and MS-2 (16,855 SNP markers). The parental population consisted of 31 B-lines and 60 R-lines, while the combined population consisted of the parental population and 345 hybrids.

SNP	Chromo-some	Parental Population		Combined population	
		MS-1	MS-2	MS-1	MS-2
Bn-scaff_15712_13-p63324	A2	SOC[1]	SOC	YLD[2], HT[3], GSL[4]	YLD, GSL
Bn-A09-p264743	A9	YLD ,HT, SPC[5], SOC	YLD, HT, SPC, SOC	HT, GSL	GSL
Bn-scaff_18514_1-p28001	C2	SOC	SOC, GSL	GSL	GSL
Bn-scaff_15712_13-p38138	C2	SOC	SOC	GSL	GSL
Bn-scaff_15712_13-p43168	C2	SOC	SOC	YLD, GSL	YLD, GSL

[1] Seed oil content.
[2] Seed yield.
[3] Plant height.
[4] Seed glucosinolate content.
[5] Seed protein content.

Table 3.10 Number of significant SNPs identified by six models based on MS-1 (26,651 SNP markers) and MS-2 (16,855 SNP markers) based on two *Brassica napus* L. populations. The parental population consisted of 31 B-lines and 60 R-lines, while combined population consisted of the parental population and 345 hybrids. Traits evaluated included seed yield (YLD), plant height (HT), seed protein content (SPC), seed oil content (SOC) and seed glucosinolate content (GSL).

Population	Marker	Model	YLD	HT	SPC	SOC	GSL	Unique SNPs
Parental	MS-1	MLM+K[1]	0	0	1	5	0	-
		MLM+K+Q[2]	0	0	1	5	0	
		MLM+K+PCA[3]	0	0	1	4	0	
		MLMM[4]	1	1	1	1	1	
		FarmCPU[5]	2	3	4	2	4	
		CMLM[6]	0	0	1	1	1	
		Unique SNPs	2	3	4	6	4	16
	MS-2	MLM+K	0	0	1	5	0	-
		MLM+K+Q	0	0	1	5	3	
		MLM+K+PCA	0	0	1	1	0	
		MLMM	2	3	1	1	1	
		FarmCPU	0	1	3	4	5	
		CMLM	0	0	1	1	1	
		Unique SNPs	2	3	3	8	8	20
Combined	MS-1	MLMM	5	3	2	1	5	-
		FarmCPU	6	11	3	12	14	
		CMLM	37	35	9	0	19	
		Unique SNPs	44	44	12	13	34	110
	MS-2	MLMM	4	3	2	4	15	-
		FarmCPU	7	10	12	5	17	
		CMLM	19	34	4	1	15	
		Unique SNPs	24	42	14	10	29	86

[1] mixed linear model considering kinship.
[2] mixed linear model considering kinship and subpopulation structure via Bayesian clustering.
[3] mixed linear model considering kinship and subpopulation structure via principal component analysis.
[4] multi-locus mixed linear model.
[5] fixed and random model circulating probability unification.
[6] compression mixed linear model.

3.4.6.1 Seed yield

Results from across the populations and marker sets were pooled together for further interpretation. There were 47 significant MTAs identified for YLD in the association analysis, located across the entire genome along every chromosome except C7. There were three SNP peaks located on chromosomes A2, A8 and C2 that were detected by CMLM from the combined population. The analysis based on the MS-1 detected the most abundant MTAs (Figure 3.13). All Manhattan plots for yield based on MS-2 can be found in the Appendix (Figure S3.1).

In total there were 773 candidate genes identified based on the regions identified by the significant MTAs. Regarding biological processes, the top three GO terms included nucleic acid metabolic processes (GO:0090304), gene expression (GO:0010467) and macromolecule biosynthetic processes (GO:0009059) (Figure 3.14). The top three GO terms for cellular components were intracellular membrane-bounded organelle (GO:0043231), chloroplast (GO:0009507) and cell-cell junction (GO:0005911). For molecular function, the top three GO terms were metal ion binding (GO:0046872), purine nucleotide binding (GO:0017076) and purine ribonucleotide binding (GO:0032555). Seventeen GO terms were enriched in biological processes, molecular function and cellular components. Out of the 773 candidate genes, 56 were predicted genes belonged to the Brassicales order and were related to different aspects of growth and development stages in plants. See the full list of the genes identified under Brassicales in the Appendix (Table S3.8). Seven of the 56 predicted genes were previously identified in *B. napus* (Table 3.11), which are c-repeat Binding Factor 5 (*CBF5*), cytokinin dehydrogenase 3 (*CKX3*), calcium-dependent protein kinase 18 and 23 (*CPK18* and *CPK 23*), sucrose transporters 2 (*SUT2*), *WRKY41* and *WRKY72*.

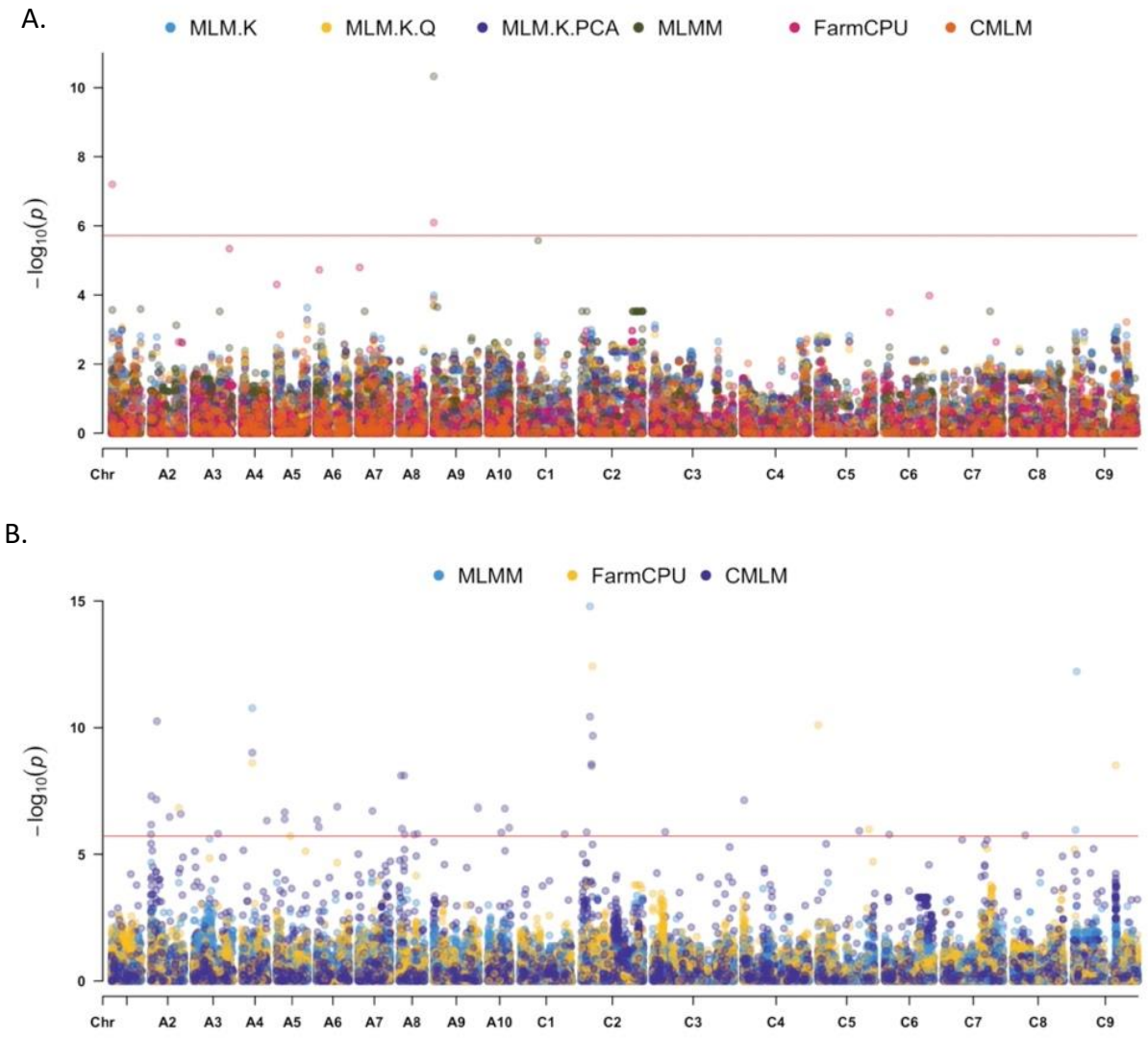


Figure 3.13 Manhattan plots of seed yield (YLD) based on the MS-1 (26,651 SNP markers). Bonferroni-corrected significance threshold was shown as the red horizontal line at $-\log_{10}(0.05/26651) = 5.73$. (A) Results from was the *Brassica napus* L. parental population based on six models including mixed linear models considering kinship (MLM+K), mixed linear models considering subpopulation structure via Bayesian clustering (MLM+K+Q), mixed linear models considering subpopulation structure via principal component analysis (MLM+K+PCA), multi-loci mixed model (MLMM), Fixed and random model circulating probability unification (FarmCPU) and compression mixed linear model (CMLM). (B) Results from the *Brassica napus* L. combined population based on three models MLMM, FarmCPU and CMLM.

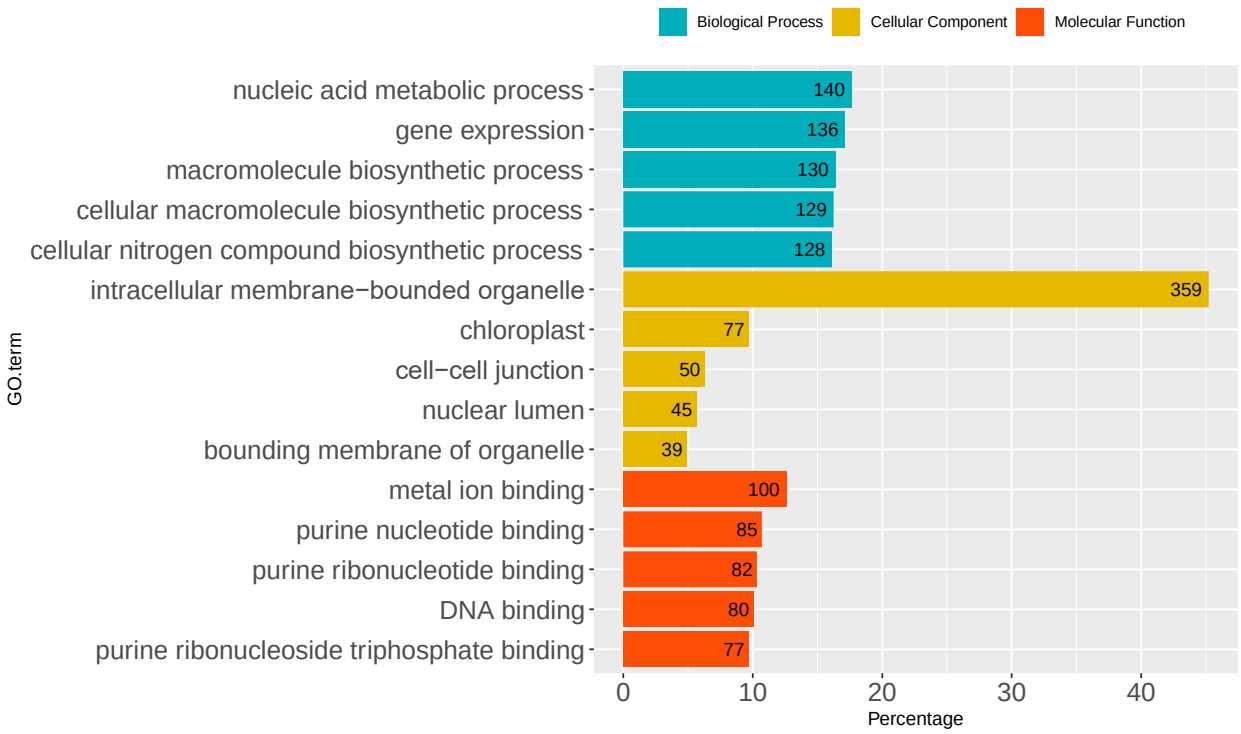


Figure 3.14 The distribution of top five gene ontology (GO) terms associated with YLD (seed yield) identified from pooled results from GWAS conducted on a parental and a hybrid population of *Brassica napus* L. based on two sets of makers that contained 26,651 and 16,855 SNP markers with six GWAS models. The number on the end of each bar represents the number of mapped sequences. The x- axis represents the proportion of sequences mapped to the GO terms out of the total number of sequences, representing the abundance of the GO term. The y-axis represents the name of the GO term.

Table 3.11 Predicted genes that were previously identified and described in *Brassica napus*.

Trait	Gene name	Description	Query ID	Reference
YLD[1]	C-repeat Binding Factor 5 (*CBF5*)	related with the development of essential cold tolerance	BnaA02g21770D	Savitch et al. (2005)
	Cytokinin dehydrogenase 3 (*CKX3*)	involved in pod development and stress response	BnaA02g08420D	Liu et al. (2018).
	Calcium-dependent protein kinase 3 (*CPK18*)	gene family involved in stress-related signal transduction pathways	BnaA08g04270D	Zhang et al. (2014a).
	Calcium-dependent protein kinase (*CPK23*)	gene family involved in stress-related signal transduction pathways	BnaA02g21060D	Zhang et al. (2014a).
	Sucrose transporters 2 (*SUT2*)	sucrose transporter, ameliorating impacts from drought stress	BnaC02g34840D	La et al. (2019)
	WRKY41	regulation of anthocyanin biosynthesis	BnaA02g21850D	Duan et al. (2018)
	WRKY72	response to cold stress and *Sclerotinia sclerotiorum* inoculation	BnaA02g02500D	Li et al. (2019b)
HT[2]	Calcium-dependent protein kinase 18 (*CPK18*)	gene family involved in stress-related signal transduction pathways	BnaA08g04270D	Zhang et al. (2014a)
	Calcium-dependent protein kinase 24 (*CPK24*)	gene family involved in stress-related signal transduction pathways	BnaC04g13100D	Zhang et al. (2014a)
	Fatty acid desaturase 3 (*FAD3*)	control of the oleic acid (C18:1) and linolenic acid (C18:3) contents	BnaC04g14820D	Yang et al. (2012)
SPC[3]	Cell wall invertase (*CWINV*)	potentially associated with providing hexoses in the process of anther and ovary development and petal expansion	BnaC01g37450D	Song et al. (2015)
	Fatty acid desaturase 7 (*FAD7*)	synthesis of linolenic acid (C18:3) from linoleic in the plastids	BnaC09g18650D	Dar et al. (2017)
	Glutathione transferases F2 (*GSTF2*)	important gene related to resistance to blackleg disease caused by *Leptosphaeria maculans*	BnaA09g00850D, BnaA09g00860D	Wei et al. (2019b)
	Light harvesting complex gene 3 (*LHCA3*)	involved in photosynthesis or light absorption	BnaA09g13710D	Marmagne et al. (2010)

	Sodium hydrogen exchanger (*NHX1*)	antioxidant defense gene	BnaA09g03600D	Zhao et al. (2018a)
	The pyrabactin resistance 1-like (*PYL*)	important candidates for improving tolerance to abiotic stresses.	BnaC09g19620D	Di et al. (2018)
	Somatic embryogenesis receptor-like kinases 1 (*SERK1*)	involved in the process of microspore embryogenesis induction, development, and plantlet regeneration.	BnaC06g32810D	Ahmadi et al. (2016)
	SHOOT MERISTEMLESS (*STM*)	promoting seed SOC production and desirable alterations of fatty acid and GSL levels	BnaA09g13310D	Elhiti et al. (2012)
	Vacuolar protein sorting 34 (*VPS34*)	potentially needed in division and development of flower organs and germinated seeds	BnaA09g14140D	Das et al. (2005)
	WRKY13	cadmium accumulation and sensitivity	BnaC07g47230D	Sheng et al. (2019)
SOC[4]	C-repeat Binding Factor 5 (*CBF5*)	enhancement of energy conversion efficiency under low temperature	BnaA07g17200D	Huang et al. (2020)
	Calcineurin B-like proteins (*CBL*)	response to different abiotic or hormone signaling	BnaA07g17150D	Zhang et al. (2014b)
	Calcineurin B-like proteins (*CBL2*)	gene families respond to different abiotic or hormone signaling	BnaC02g12710D	Zhang et al. (2014b)
	DICER-LIKE 4 (*DCL4*)	response to verticillium wilt (*Verticillium dahliae*)	BnaC09g37430D	Shen et al. (2014)
	EF-Tu receptor (*EFR*)	response to low temperature stress	BnaC09g37350D	Luo et al. (2019)
	Fatty acid desaturase 7 (*FAD7*)	synthesis of linolenic acid (C18:3) from linoleic in the plastids	BnaC09g18650D	Dar et al. (2017)
	pyrabactin resistance 1-like 1 (*PYL1*)	might be important candidates for improving tolerance to abiotic stresses	BnaC09g19620D	Di et al. (2018)
GSL[5]	Calcineurin B-like proteins (*CBL2*)	gene families respond to different abiotic or hormone signaling	BnaC02g12710D	Zhang et al. (2014b)
	Calcium-dependent protein kinase 2 (*CPK2*)	gene family involved in stress-related signal transduction pathways	BnaC03g36720D, BnaC03g36730D	Zhang et al. (2014a)

	Calcium-dependent protein kinase 20 (*CPK20*)	gene family involved in stress-related signal transduction pathways	BnaC03g21760D	Zhang et al. (2014a)
	Cytochrome P450 family 83 subfamily A polypeptide 1 (*CYP83A1*)	GSL synthesis genes; highly responsive to *S. sclerotiorum* and *B. cinerea* infection	BnaA04g06630D	Zhang et al. (2015a)
	Fatty acid desaturase 7 (*FAD7*)	synthesis of linolenic acid (C18:3) from linoleic in the plastids	BnaC03g37090D	Dar et al. (2017)
	RNA-dependent RNA polymerase 1 (*RDR1*)	may contribute to *B. napus* defense against *S. sclerotiorum*	BnaC05g10980D	Cao et al. (2016)
	WRKY69	response to jasmonic acid and *S. sclerotiorum* induction	BnaC08g29410D	Yao et al. (2020a)

[1] Yield.
[2] Plant height.
[3] Seed protein content.
[4] Seed oil content.
[5] Seed glucosinolate content.

3.4.6.2 Plant height

In the combined population, there were 47 significant MTAs with plant height identified above the threshold 5.73 and they were distributed on 17 chromosomes excluding chromosomes A1 and A3 (Figure 3.15). Based on the parental population there was one SNP peak on chromosome A9 (Figure 3.15.A), which was also identified based on the combined population. SNP peaks were detected on chromosomes A2, A5, A6, A8, A10 and C2 by CMLM, FarmCPU and CMLM (Figure 3.15.B). All Manhattan plots for HT based on MS-2 can be found in the Appendix (Figure S3.2). Figure 3.16 shows the distribution of the top five GO terms associated with HT. The top three GO terms in biological process were cellular protein metabolic process (GO:0044267), nucleic acid metabolic process (GO:0090304) and macromolecule biosynthetic process (GO:0009059). The top three GO terms in the cellular component were intracellular membrane-bounded organelle (GO:0043231), chloroplast (GO:0009507) and cell-cell junction (GO:0005911). The top three GO terms in molecular function were metal ion binding (GO:0046872), purine nucleotide binding (GO:0017076) and purine ribonucleotide binding (GO:0032555). One GO term (GO:0008234) for molecular function was enriched in cysteine-type peptidase activity. In total, there were 1,236 candidate genes identified that were associated with plant height. Out of the 1,236 candidate genes, 78 predicted genes belonged to the Brassicales order. The full list of the predicted genes can be found in the Appendix (Table S3.8). Among the 78 predicted genes, three were previously described in *B. napus* (Table 3.11) which are fatty acid desaturase 3 (*FAD3*), C*PK18* and *CPK24* (calcium-dependent protein kinase (*CPK*) gene family).

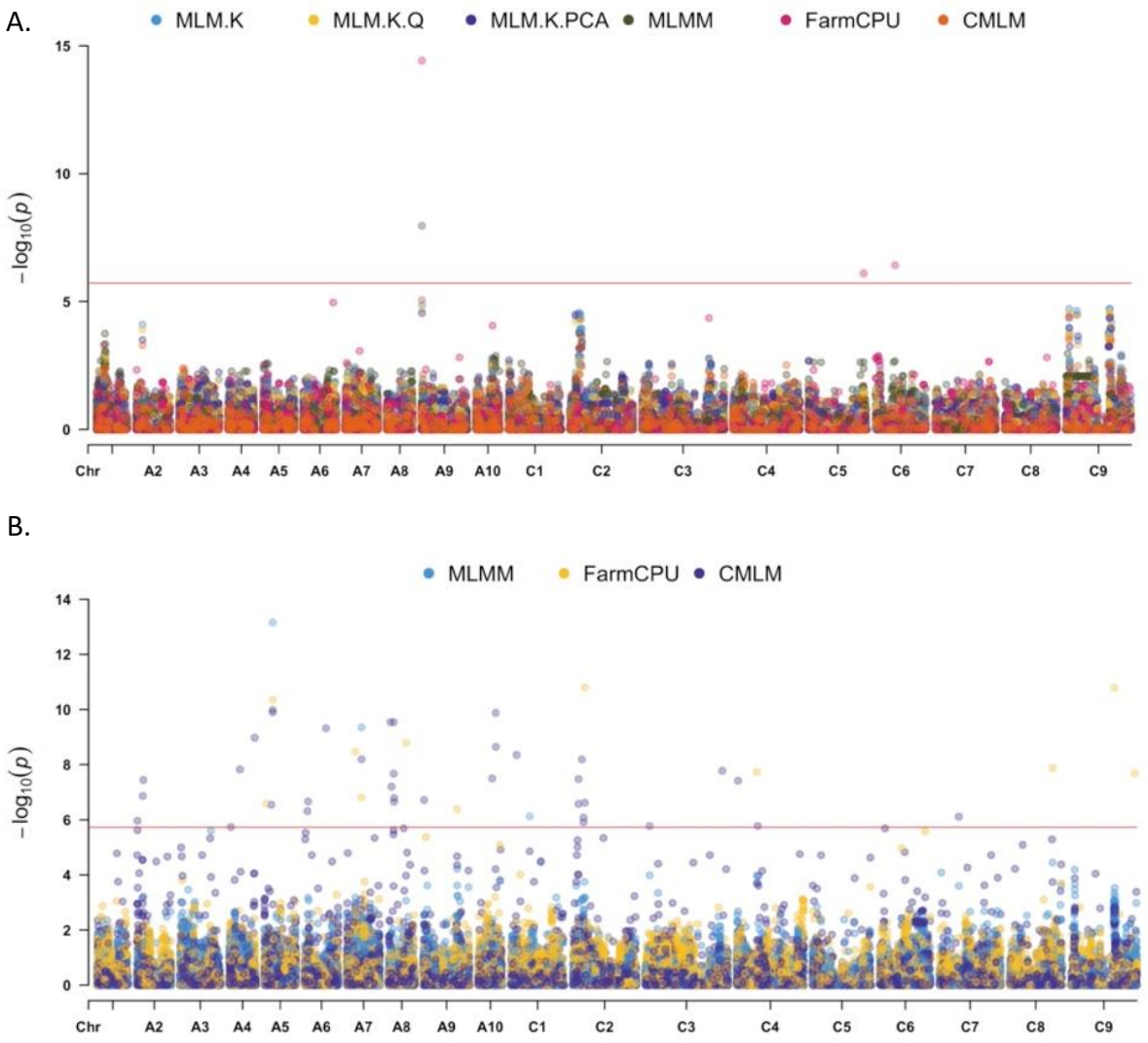


Figure 3.15 Manhattan plots of plant height (HT) based on the MS-1 (26,651 SNP markers). Bonferroni-corrected significance threshold was shown as the red horizontal line at $-\log_{10}$ (0.05/26651) =5.73. (A) Results from was the *Brassica napus* L. parental population based on six models including mixed linear models considering kinship (MLM+K), mixed linear models considering subpopulation structure via Bayesian clustering (MLM+K+Q), mixed linear models considering subpopulation structure via principal component analysis (MLM+K+PCA), multi-loci mixed model (MLMM), Fixed and random model circulating probability unification (FarmCPU) and compression mixed linear model (CMLM). (B) Results from the *Brassica napus* L. combined population based on three models MLMM, FarmCPU and CMLM.

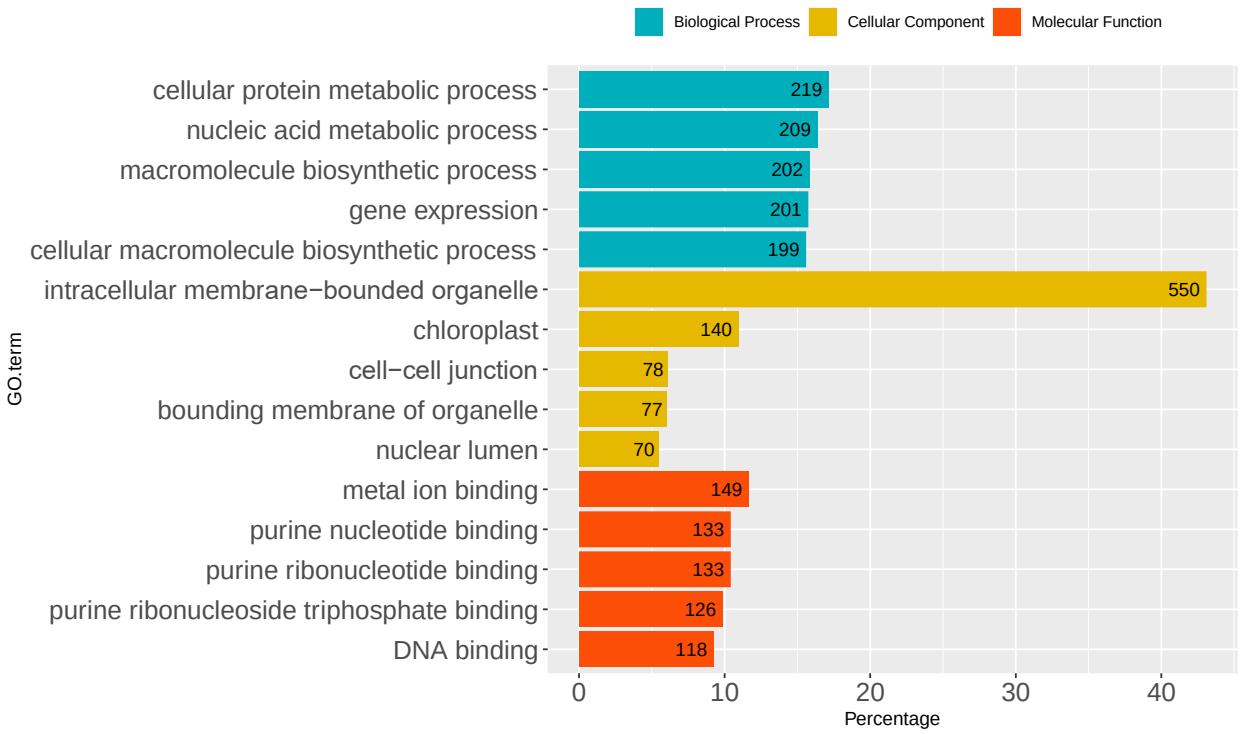


Figure 3.16 The distribution of top five gene ontology (GO) terms associated with HT (plant height) identified from pooled results from GWAS conducted on a parental and a hybrid population of *Brassica napus* L. based on two sets of makers that contained 26,651 and 16,855 SNP markers with six GWAS models. The number on the end of each bar represents the number of mapped sequences. The x- axis represents the proportion of sequences mapped to the GO terms out of the total number of sequences, representing the abundance of the GO term. The y-axis represents the name of the GO term.

3.4.6.3 Seed protein content

There were 12 MTAs detected for SPC above the threshold 5.73. Nine of them were located on the A-subgenome (chromosomes A2, A3, A5, A8 and A9) and three of them were located on the C-subgenome (chromosomes C3, C5 and C9) (Figure 3.17). There was one SNP peak identified from the parental population on chromosome A9 by all models expect MLM+K+Q. However, this peak wasn't identified in the combined population. All Manhattan plots for SPC based on MS-2 can be found in the Appendix (Figure S3.3).

In total there were 972 candidate genes identified based in the regions identified by the significant MTAs (see the full list of the genes identified under Brassicales in Table S3.6). The top three GO terms in biological process were gene expression (GO:0010467), nucleic acid metabolic process (GO:0090304) and macromolecule biosynthetic process (GO:0009059) (Figure 3.18). The top three GO terms in cellular components were intracellular membrane-bounded organelle (GO:0043231), chloroplast (GO:0009507) and cell-cell junction (GO:0005911). The top three GO terms in molecular function were purine nucleotide binding (GO:0017076), purine ribonucleotide binding (GO:0032555) and metal ion binding (GO:0046872). Thirteen GO terms were significantly enriched in biological process, molecular function, and cell component. Among the 54 genes predicted under the Brassicales order, ten were previously identified in *B. napus* (Table 3.11): cell wall invertase (*CWINV*) gene, fatty acid desaturase 7 (*FAD7*), glutathione transferases F2 (*GSTF2*) and light harvesting complex gene 3 (*LHCA3*), Sodium hydrogen exchanger (*NHX1*), pyrabactin resistance 1-like (*PYL*), somatic embryogenesis receptor-like kinases 1 (*SERK1*), SHOOT MERISTEMLESS (*STM*), vacuolar protein sorting 34 (*VPS34*) and *WRKY13*.

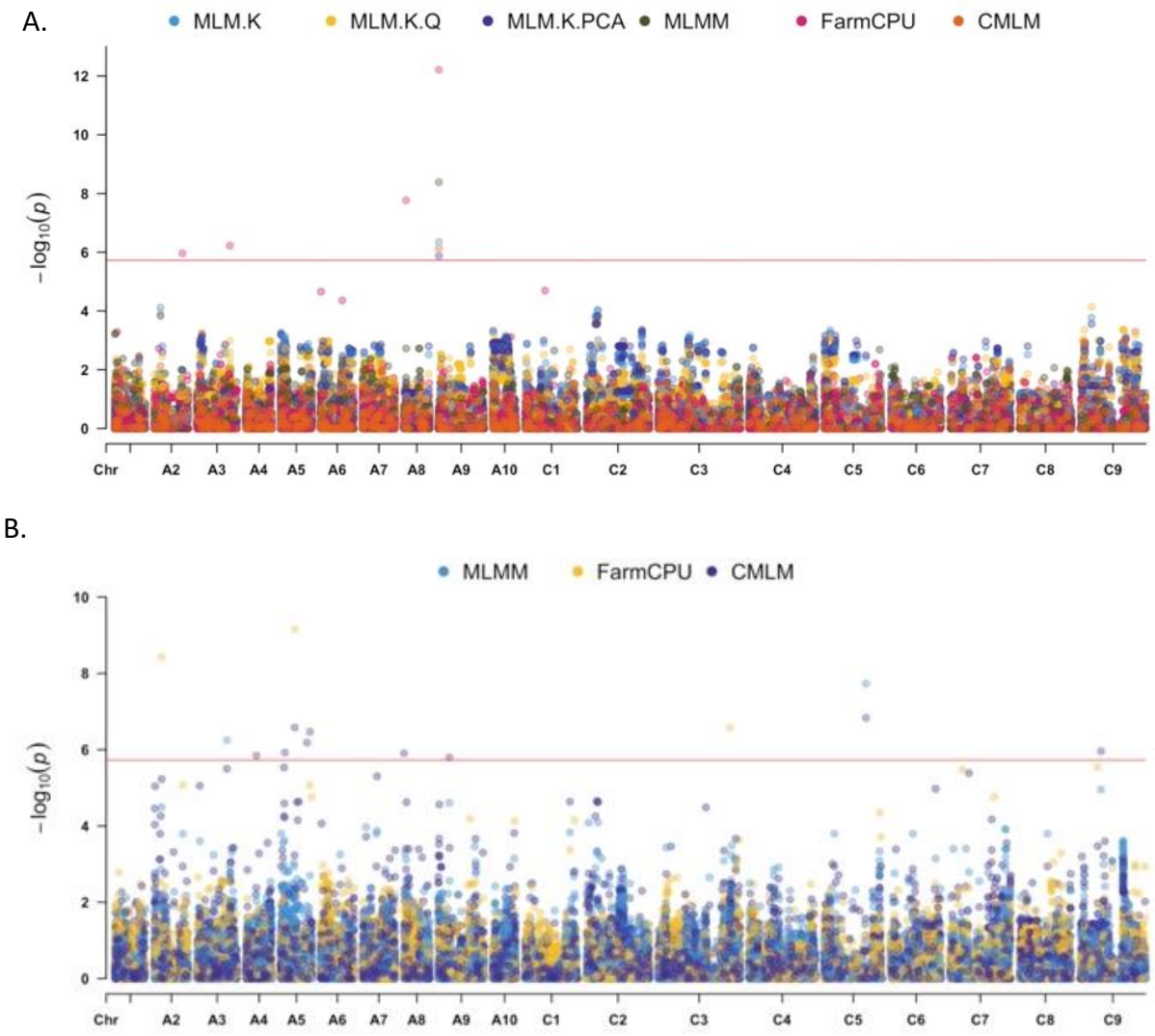


Figure 3.17 Manhattan plots of seed protein content (SPC) based on the MS-1 (26,651 SNP markers). Bonferroni-corrected significance threshold was shown as the red horizontal line at -$\log_{10}$ (0.05/26651) =5.73. (A) Results from was the *Brassica napus* L. parental population based on six models including mixed linear models considering kinship (MLM+K), mixed linear models considering subpopulation structure via Bayesian clustering (MLM+K+Q), mixed linear models considering subpopulation structure via principal component analysis (MLM+K+PCA), multi-loci mixed model (MLMM), Fixed and random model circulating probability unification (FarmCPU) and compression mixed linear model (CMLM). (B) Results from the *Brassica napus* L. combined population based on three models MLMM, FarmCPU and CMLM.

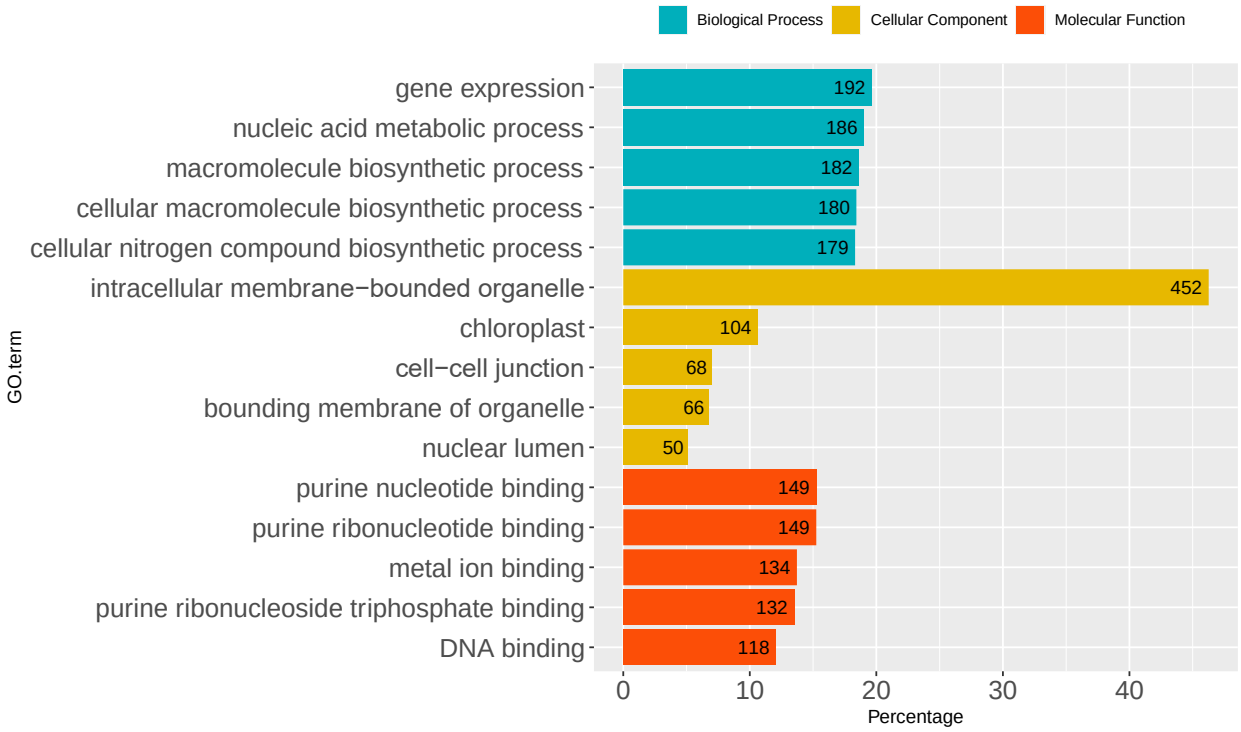


Figure 3.18 The distribution of top five gene ontology (GO) terms associated with SPC (seed protein content) identified from pooled results from GWAS conducted on a parental and a hybrid population of *Brassica napus* L. based on two sets of makers that contained 26,651 and 16,855 SNP markers with six GWAS models. The number on the end of each bar represents the number of mapped sequences. The x- axis represents the proportion of sequences mapped to the GO terms out of the total number of sequences, representing the abundance of the GO term. The y-axis represents the name of the GO term.

3.4.6.4 Seed oil content

For SOC there were 36 MTAs detected from 13 different chromosomes (A2, A5, A7, A9, C2, C3, C5 and C9) (Figure 3.19). FarmCPU and MLMM detected 12 and one significant SNP, respectively. Based on the parental population, there were three SNP peaks identified which located on chromosomes A2, A9 and C2 (Figure 3.19.A). The distribution of significant MTAs was more dispersed on multiple chromosomes compared to the parental population. All Manhattan plots for SOC based on MS-2 can be found in the Appendix (Figure S3.4).

There were 873 candidate genes identified, but no GO term was significantly enriched. See the full list of the genes identified under Brassicales in the Appendix (Table S3.8). The top three GO terms in biological process were nucleic acid metabolic process (GO:0090304), gene expression (GO:0010467) and macromolecule biosynthetic process (GO:0009059) (Figure 3.20). The top three GO terms in cellular component were intracellular membrane-bounded organelle (GO:0043231), chloroplast (GO:0009507) and bounding membrane of organelle (GO:0098588). The top three GO terms in molecular function were purine ribonucleotide binding (GO:0032555), purine nucleotide binding (GO:0017076) and metal ion binding (GO:0046872). Under the Brassicales order there were 799 candidate genes and seven of them were previously characterised in *B. napus* including c-repeat binding factor 5 (*CBF5*), calcineurin B-like proteins (*CBL*), calcineurin B-like proteins 2 (*CBL2*), DICER-LIKE 4 (*DCL4*), EF-Tu receptor (*EFR*), fatty acid desaturase 7 (*FAD7*) and the pyrabactin resistance 1-like 1 (*PYL1*) (Table 3.11).

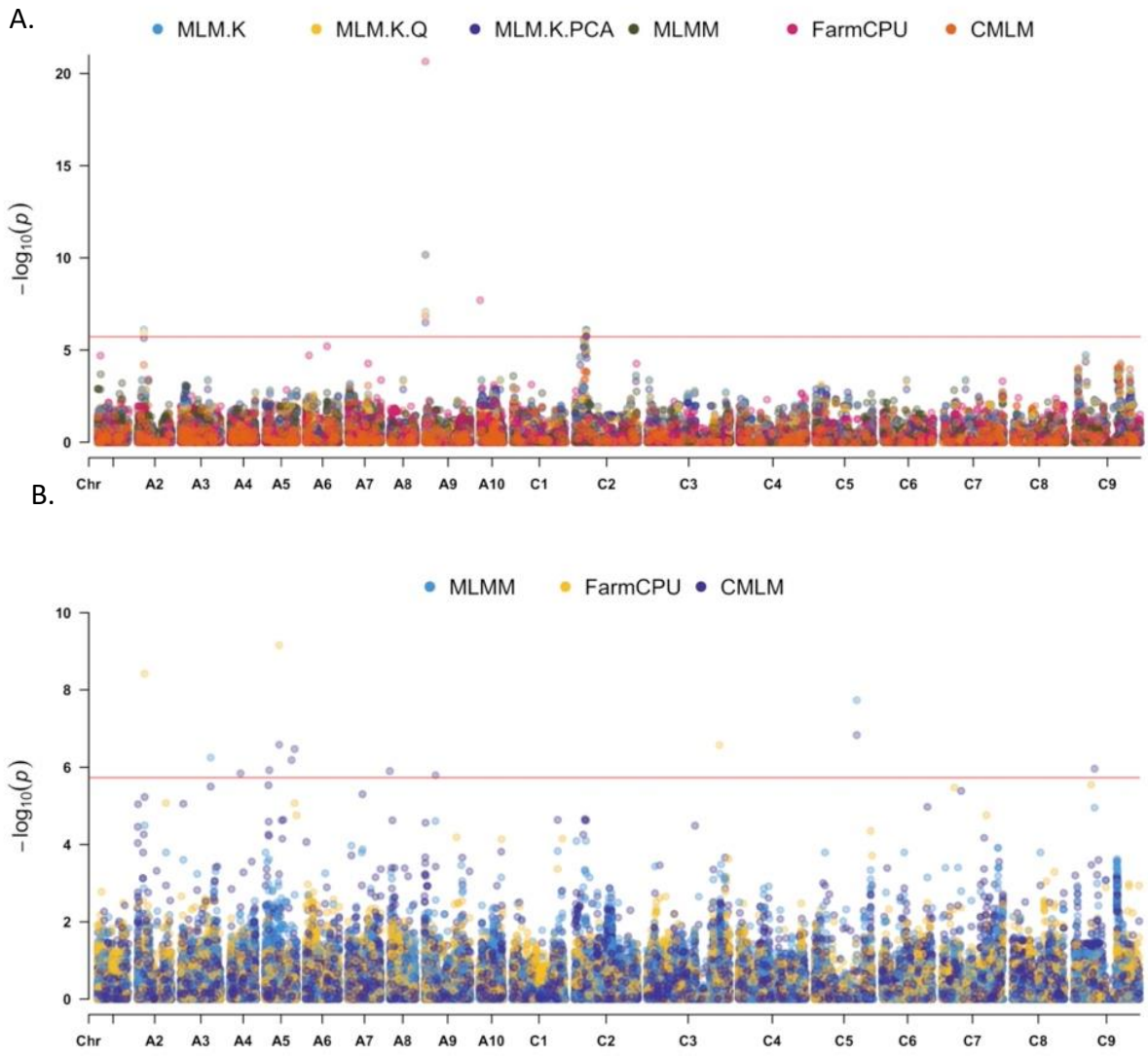


Figure 3.19 Manhattan plots of seed oil content (SOC) based on the MS-1 (26,651 SNP markers). Bonferroni-corrected significance threshold was shown as the red horizontal line at -$\log_{10}$ (0.05/26651) =5.73. (A) Results from was the *Brassica napus* L. parental population based on six models including mixed linear models considering kinship (MLM+K), mixed linear models considering subpopulation structure via Bayesian clustering (MLM+K+Q), mixed linear models considering subpopulation structure via principal component analysis (MLM+K+PCA), multi-loci mixed model (MLMM), Fixed and random model circulating probability unification (FarmCPU) and compression mixed linear model (CMLM). (B) Results from the *Brassica napus* L. combined population based on three models MLMM, FarmCPU and CMLM.

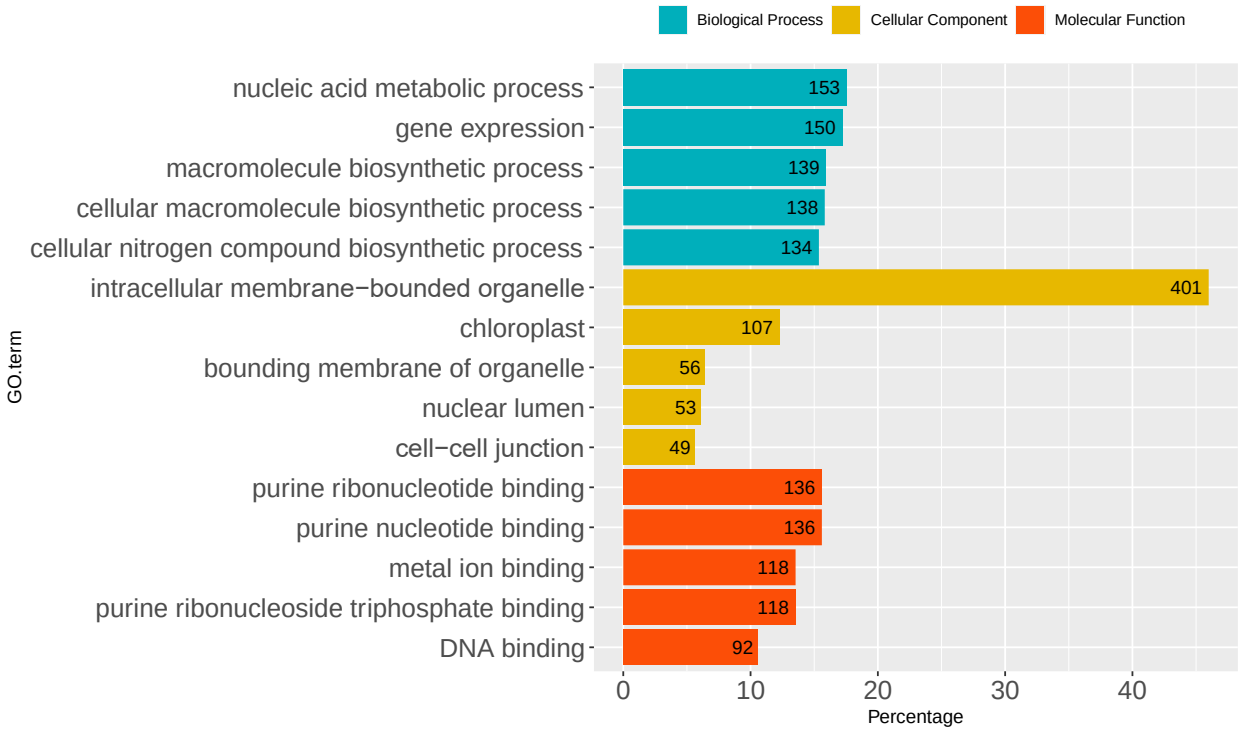


Figure 3.20 The distribution of top five gene ontology (GO) terms associated with SOC (seed oil content) identified from pooled results from GWAS conducted on a parental and a hybrid population of *Brassica napus* L. based on two sets of makers that contained 26,651 and 16,855 SNP markers with six GWAS models. The number on the end of each bar represents the number of mapped sequences. The x- axis represents the proportion of sequences mapped to the GO terms out of the total number of sequences, representing the abundance of the GO term. The y-axis represents the name of the GO term.

3.4.6.5 Seed glucosinolate content

Thirty-four significant MTAs were detected for GSL (Figure 3.21). CMLM, MLMM and FarmCPU detected 19, 5, and 14 significant MTAs, respectively. SNP peaks were observed on chromosomes C2 and C5. Based on the parental population, there was a peak of SNPs on chromosome A5 detected by multiple models (MLM+K+Q, MLMM, FarmCPU and CMLM). However, this peak was not detected based on the combined population. All Manhattan plots for GSL based on MS-2 can be found in the Appendix (Figure S3.5).

There were 815 candidate genes identified, but no GO term was significantly enriched. The full list of the genes identified under Brassicales in the Appendix (Table S3.8). The top three GO terms were nucleic acid metabolic process (GO:0090304), cellular protein metabolic process (GO:0044267) and gene expression (GO:0010467) (Figure 3.22). The top three GO terms in cellular component were intracellular membrane-bounded organelle (GO:0043231), chloroplast (GO:0009507) and bounding membrane of organelle (GO:0098588). The top three GO terms in molecular function were purine ribonucleotide binding (GO:0032555), purine nucleotide binding (GO:0017076) and metal ion binding (GO:0046872). 560 candidate genes were predicted under the Brassicale order and seven of them were previously identified in *B. napus* including calcineurin B-like proteins (*CBL2*), calcium-dependent protein kinase 2 (*CPK2*), calcium-dependent protein kinase 20 (*CPK20*), cytochrome P450 family 83 subfamily A polypeptide 1 (*CYP83A1*), fatty acid desaturase 7 (*FAD7*), RNA-dependent RNA polymerase 1 (*RDR1*) and *WRKY69* (Table 3.11).

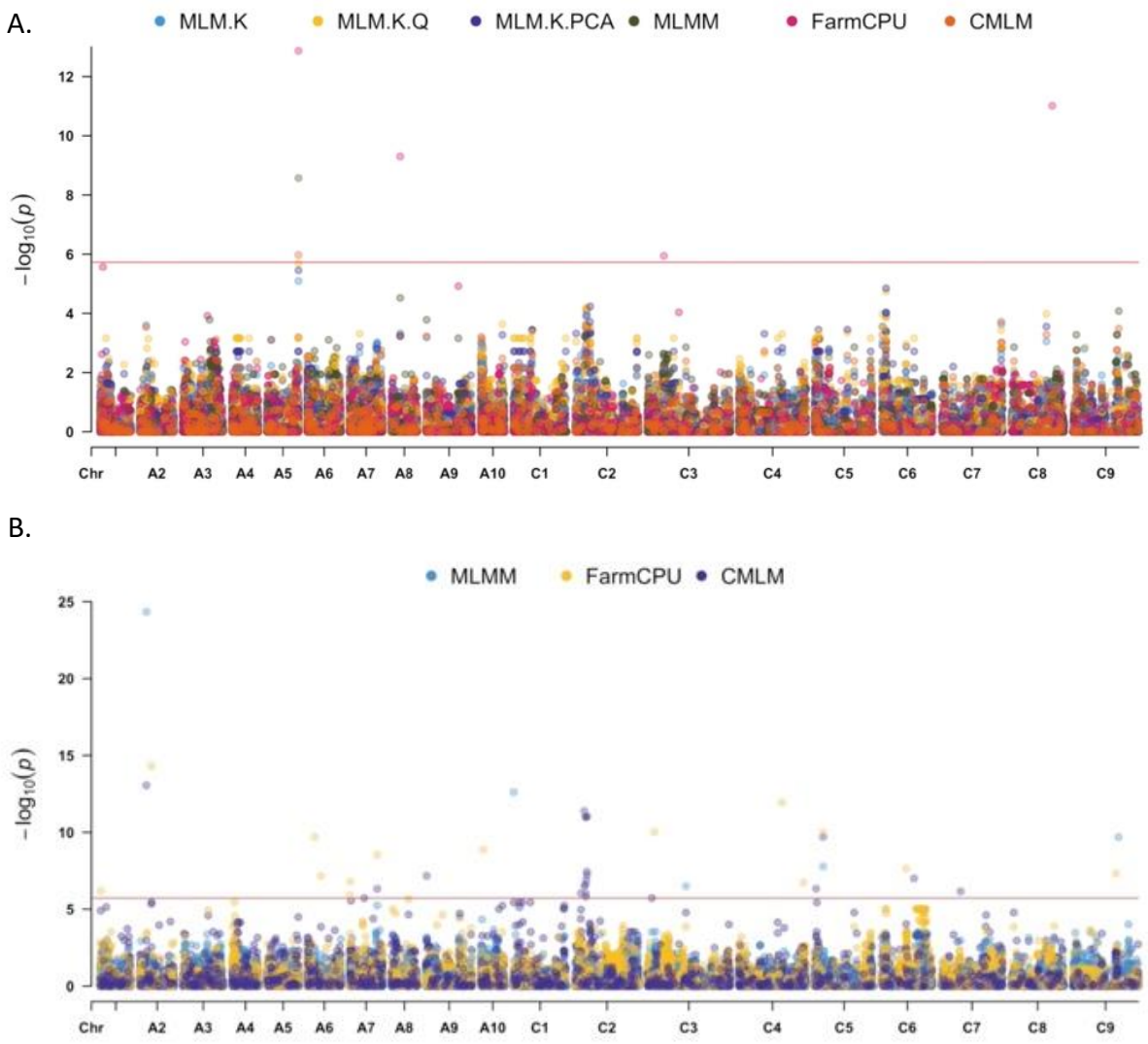


Figure 3.21 Manhattan plots showing seed glucosinolate content (GSL) based on the MS-1 (26,651 SNP markers). Bonferroni-corrected significance threshold was shown as the red horizontal line at $-\log_{10}(0.05/26651) = 5.73$. (A) Results from was the *Brassica napus* L. parental population based on six models including mixed linear models considering kinship (MLM+K), mixed linear models considering subpopulation structure via Bayesian clustering (MLM+K+Q), mixed linear models considering subpopulation structure via principal component analysis (MLM+K+PCA), multi-loci mixed model (MLMM), Fixed and random model circulating probability unification (FarmCPU) and compression mixed linear model (CMLM). (B) Results from the *Brassica napus* L. combined population based on three models MLMM, FarmCPU and CMLM.

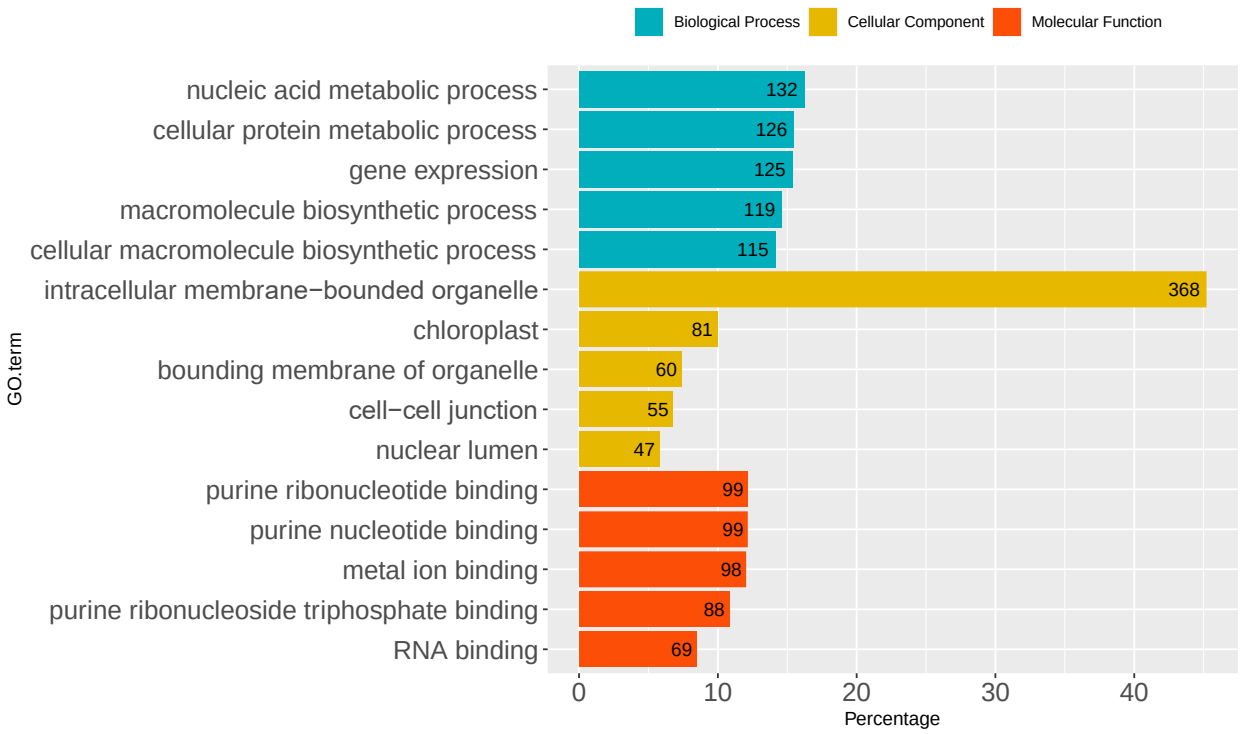


Figure 3.22 The distribution of top five gene ontology (GO) terms associated with GSL (seed glucosinolate content) identified from pooled results from GWAS conducted on a parental and a hybrid population of *Brassica napus* L. based on two sets of makers that contained 26,651 and 16,855 SNP markers with six GWAS models. The number on the end of each bar represents the number of mapped sequences. The x- axis represents the proportion of sequences mapped to the GO terms out of the total number of sequences, representing the abundance of the GO term. The y-axis represents the name of the GO term.

3.5 Discussion

Genome-wide association mapping is often considered as a complementary investigation method to linkage or QTL mapping in identifying candidate QTL or genes that control a certain trait (Korte and Farlow 2013). The detection of a true MTA by GWAS (i.e. the power of GWAS) depends on the phenotypic variance of the population that can be explained by the SNP(s) applied (Korte and Farlow 2013). Therefore, the target population size and structure as well as marker density could all affect the power of GWAS (Ibrahim et al. 2020). To improve the accuracy of GWAS, understanding the structure of the target population is crucial (Li et al. 2014a). This research examined the effects of population size as well as population composition on GWAS. As stated by Sebastiani et al. (2009), a small population is often not ideal for GWAS studies. There is a sample size threshold above which the rate of locus discovery increases since increased population size and marker density could improve the power of GWAS based on empirical evidence (Alseekh et al. 2021). Therefore, a larger population with relatively more diversity is preferred when conducting GWAS as it provides more statistical power.

In this research, we included hybrids in the analysis for two reasons. First, by adding the hybrid genotypes (F_1) in the combined population, the target population size increased and offered greater power in detecting possible MTAs. For example, using MS-1, the total number of significant MTAs detected increased from 16 to 110 when comparing the results from the parental population with the combined population (Table 3.10). Using MS-2, the number of MTAs increased from 20 to 64. Another reason to include the hybrid genotypes in this research was GWAS conducted on an inbred population could not be applied in evaluating hybrid performance. Research using inbreds is unable to disclose what lies hidden in hybrids, nor does it reveal the variables that contribute to hybrid performance (Wang et al. 2017a). Instead, a multiple hybrid population

comprised of a large number of hybrids derived from a certain mating design (Diallel, North Carolina Designs, triple test crosses or simplified triple testcrosses) is more appropriate and powerful in GWAS for hybrid crops compared with bi-/multi-parental populations or natural populations (Wang et al. 2017a). Therefore, the addition of hybrid genotypes provided greater insight into the performance of the population over multiple environments (Zhang et al. 2019c). This is crucial to the Canadian canola industry since hybrid cultivars account for more than 95% of all cultivars grown in the Canadian Prairies(Morrison et al. 2016).

Marker density is another factor that could affect GWAS accuracy (Ibrahim et al. 2020). In GWAS, the design of the marker panel should take several factors into account such as genome size, the LD extent and the traits of interest (Ballesta et al. 2020). Higher marker density is required for plants with a larger genome size and fast LD decay (Ballesta et al. 2020; Cui et al. 2020). Maize, for example, requires 0.5–1.0 million or more markers to perform successful GWAS (Yan et al. 2011). In this study, MS-1 contained 9,796 more markers than MS-2, which provided increased coverage along the genome (Figure 3.3). Therefore, MS-1 was able to detect more significant MTAs in the combined population. However, in the parental population, MS-2 performed better than MS-1. This might be due to a reduction in the SNP effects associated with QTL as the marker density increases (Chang et al. 2018), especially considering that the parental population had a smaller population size. Another possible explanation is that often higher marker density is needed for populations with shorter LD decay so that QTL associated with a certain trait is in linkage with one or more markers (Kainer et al. 2019). Muller et al. (2017) found there was no difference in predictive ability when using subsets of SNP markers (~5000 to 10,000 SNPs) instead of full sets of markers (3,787 and 19,506 SNPs) in two breeding populations of *Eucalyptus*. Thus, it is likely that MS-2 that contained 16,855 SNP markers already provided sufficient coverage for the parental

population and therefore reduced potential noise that could affect the power to detect significant MTAs.

In this research, MLMM, FarmCPU and CMLM performed better in both the parental populations and the combined populations. This is due to the difference amongst the computational methods of these models. Even though MLM models manage the *p*-value inflation effectively, it also generates false negatives, which could cause a reduction in the power to identify of the true associations (Zhang et al. 2010). Therefore, CMLM was proposed to address this issue in which individuals were clustered into groups and their genetic values are fit in the model as random effects (Kaler et al. 2019). The biggest advantage of CMLM is that it simplified the iteration process and is more computational efficient than MLM that considers the relatedness within the population (Zhang et al. 2010). CMLM groups similar individuals together and utilizes a reduced kinship in the analysis, which was found to have more statistical power as well (Zhang et al. 2010). MLMM and FarmCPU are models based on multiple loci, which, compared with single-locus models, have better control on issues that arise from existing population structure in the target populations (Wang and Zhang 2020). Kaler et al. (2019) compared the performance of eight GWAS models including analysis of variance (ANOVA), general linear model (GLM), MLM, CMLM, enriched compressed MLM (ECMLM), Settlement of MLM Under Progressively Exclusive Relationship (SUPER), MLMM and FarmCPU. The authors found that complex models (MLM, CMLM and ECMLM) were able to control false positives effectively, but increased false negatives, while the multi-locus model FarmCPU controlled both false positives and false negatives effectively. The current research found that the three MLM models were able to control both false positives and false negatives well in the parental population, as the Q-Q plots showed a straight line close to 1:1 line with a sharp deviated tail for all traits (Figure 3.8 and Figure 3.9).

However, in the combined population the MLM models deviated from the 1:1 line in the Q-Q plots. In contrast, MLMM, FarmCPU and CMLM performed well in controlling both the false positives and negatives. However, despite the fact that MLMM, FarmCPU and CMLM performed better than the other three models in this study, and that the populations we used in this study did not show significant population structure, the deviations observed in the Q-Q plots still revealed stratifications that were not accounted for by the models (Ehret 2010).

Understanding the genetics of complex traits is crucial in improving traits of interest in plant breeding. In this study, a total of 29 different predicted genes for critical agronomic or seed quality traits were identified that had been previously identified in *B. napus*. In general, genes predicted based on the identified MTAs in this research are involved in abiotic stress responses (such as cold, heat and drought) and pathogen infection. In this research, several predicted genes associated with seed yield were previously identified (Table 3.11). For example, *CKX3* was found to play a role in enhancing yield in the model plant *Arabidopsis thaliana* (L.) Heynh., as well as crop plants such as wheat and chickpea (*Cicer arietinum* L.) (Bartrina et al. 2011; Chen et al. 2020). *WRKY72* was involved in responding to cold stress and *S. sclerotiorum*, the causal agent of Sclerotinia stem rot (Khan et al. 2020). *CBF5* also responds to cold stress (Savitch et al. 2005). *SUT2* responds to heat stress, a crucial gene in improving *B. napus* yield since fertility can be adversely affected by high temperature (Harker et al. 2012; Polowick and Sawhney 1988). In terms of predicted genes associated with seed quality traits, predicted genes from SOC were found to be involved in response to low temperature (*CBF5*) and abiotic or hormone signaling (*CBL* and *CBL2*). Previous studies have demonstrated that temperature can affect the performance and quality traits of crops (Odukoya et al. 2019). More specifically, in canola, the composition of fatty acids can be altered under low temperature such that the content of highly unsaturated fatty acids decreases and the

content of oleic acid increases (Canvin 1965). *CYP83A1* was previously identified as an important gene in GSL synthesis. Besides, *CYP83A1* was found to be an important enzyme in glucosinolate biosynthesis in *Brassica oleracea* var. *acephala* (Cuong et al. 2019). The characteristics listed above are important factors associated with traits related to yield and seed quality in canola, which may require further investigation to gain a better understanding of the roles they play in the development of the crop.

3.6 Conclusion

The size and structure of the target population can significantly affect the performance of GWAS. Marker density and the choice of models can also impact GWAS; therefore, models that appropriately fit the trait and data need to be selected. In this research, genes associated with abiotic stress response, disease resistance and glucosinolate synthesis were predicted, which offers valuable information for future research in improving the genetics of these traits in canola.

4. GENOMIC SELECTION OF AGRONOMIC AND SEED QUALITY TRAITS IN HYBRID *Brassica napus* L. BASED ON PARAMETRIC AND MACHINE LEARNING METHODS

4.1 Abstract

Genomic selection (GS) has become a useful tool in plant breeding for its advantages in shortening the breeding cycle and improving cost efficiencies due to its potential to reduce the number of field experiments. As an important commodity in Canada, canola (*Brassica napus* L.) contributes $29.9 billion to Canadian economy annually. In this study, various factors that affect the prediction accuracy of important agronomic and seed quality traits on hybrid *Brassica napus* were examined based on a mixed population consisting of 91 parental genotypes and 345 F_1 hybrids derived from the parental genotypes. Five traits were studied: seed yield (YLD), plant height (HT), seed protein content (SPC), seed oil content (SOC) and seed glucosinolates content (GSL). Based on rrBLUP, we found that the prediction of hybrid performance using the 91 parental genotypes ($N_{TP} = 91$) produced a prediction accuracy that varied between 1% to 2% and 23 to 24% on seed YLD and SPC, respectively. Meanwhile a mixed training population (TP) ($N_{TP} = 91$) consisting of both parental and hybrid genotypes performed significantly better (28% and 45% for YLD and SPC, respectively). A greater prediction accuracy was shown as the mixed TP size increased ($N_{TP} = 262$) (38% and 58% for YLD and SPC, respectively). Marker density also impacted the prediction accuracy. Three sets of markers (26,651, 16,855, and 3,205 markers) were also compared in this research. Interestingly, we found that the set that had the highest marker density (26,651 SNPs) did not produce the highest prediction accuracy based on rrBLUP. The marker set that contained 16,855 performed the best amongst the three marker sets. In addition, we compared the model

performance of rrBLUP (ridge regression best linear unbiased prediction), GBLUP (Genomic best linear unbiased prediction), Bayesian A (BayesA), Bayesian B (BayesB), Bayesian C (BayesC) and Bayesian Ridge Regression regression method (BRR). We found that even though considered equivalent to GBLUP mathematically, rrBLUP had the poorest performance among all models across different traits. As one of the most used GS models, GBLUP performed quite similar to the Bayesian models. Although the prediction accuracy based on GBLUP was slightly lower than the Bayesian models on YLD and HT, it had equal performance with BayesA on SPC, SOC and GSL, while its computational time was significantly shorter than BayesA. Lastly, we compared the performance of three machine learning (ML) algorithms including support vector regression (SVR), Extreme Gradient Boosting (XGBoost) and random forest (RF). All these ML methods showed strong robustness in predicting the five traits, with the lowest prediction accuracy produced on YLD (69% to 72%) and the highest on GSL (84% to 87%). Taken together, this research offers valuable information on implementing GS in hybrid breeding of *B. napus*.

4.2 Introduction

Canola (*Brassica napus* L.), together with soybean [Glycine *max* (L.) Merr.] and oil palm (*Elaeis guineensis* Jacq.), are currently the three largest oilseed crops in the world(FAO 2021). Today, canola production provides raw materials for a wide range of end products, including livestock feed, biofuel, biodegradable plastics, industrial lubricants, as well as edible oils for human consumption (Jan et al. 2016; Snowdon et al. 2007). As the largest global producer and exporter of canola, Canada exports about 90% of its canola to more than 50 countries worldwide (Canola Council of Canada 2017; USDA 2020).

The improvement of canola genetics has been a major contributing factor to yield increases during 2000-2013. Two main improvements during this time included the conversion from open-

pollinated cultivars to hybrid and herbicide-tolerant cultivars (Morrison et al. 2016). In the development of high-performance hybrids, a critical consideration is how to identify the best parental combinations with the potential to create superior agronomic performance and outstanding seed quality (Starmer et al. 1998).

Genomic selection (GS) was first proposed by Meuwissen et al. (2001) in the early 2000s to improve the efficiency in animal breeding. Genomic selection is considered a variant of MAS, but instead of focusing on major-effect QTL, GS utilizes all markers on the whole genome and assumes one or more markers are in linkage disequilibrium (LD) with the loci that control the trait of interest (Desta and Ortiz 2014). Recently, plant breeders have adopted it as a tool to improve efficiency in plant breeding in various crops such as wheat (*Triticum aestivum* L.) (Elbasyoni et al. 2018; Lozada et al. 2019; Sarinelli et al. 2019; Zhao et al. 2014), maize (*Zea mays* L.) (Dias et al. 2018; Guo et al. 2019; Pace et al. 2015; Vivek et al. 2017), soybean (Shu et al. 2013), barley (*Hordeum vulgare* L.) (Lorenzana and Bernardo 2009) and rice (*Oryza sativa* L.) (Grenier et al. 2015; Spindel et al. 2015). Genomic selection has also being applied to canola breeding (Jan et al. 2016; Snowdon and Iniguez Luy 2012; Würschum et al. 2014) on various traits such as flowering time (Li et al. 2015a), plant height (Würschum et al. 2014), grain yield and seed glucosinolate content (Jan et al. 2016), and blackleg [*Leptosphaeria maculans* (Desm.) Ces. & de Not.] resistance (Fikere et al. 2018).

Genomic selection is a relatively new approach to estimate hybrid performance in canola breeding. Only a limited number of studies have been reported. Based on 950 F_1 testcross hybrids, Jan et al. (2016) examined the testcross performance through genomic prediction and obtained moderate to high prediction accuracy in seed glucosinolate content (61%) and seed oil content (81%). The authors suggested that the moderate to high prediction accuracy estimated based on the additive

effects indicated the low heterotic levels in their population. Effects of training population (TP) size were also examined, and the authors found that the prediction accuracy plateaued when the size of TP accounted for 70% to 80% of the whole population (Jan et al. 2016). Liu et al. (2017b) found that the performance of the hybrids in an immortalized F_2 population was determined by a combination of additive, dominance and epistatic effects. Knoch et al. (2021) compared the prediction accuracy using parental omics data to predict hybrids and found that using transcriptomic data instead of genetic marker data could improve the accuracy in *B. napus*. However, it is still unclear how GS can be applied to obtain reliable prediction accuracies in hybrid *B. napus*, considering that many factors can affect the prediction accuracy.

In this study, a mixed population consisting of 91 parental genotypes and 345 hybrid genotypes derived from the 91 parental genotypes were used as the training/validation population. The effects of different training set (TP) and validation set (VP) were examined based on the rrBLUP model. Genomic selection was performed with three different marker sets representing different marker densities (26,651 SNPs, 16,855 SNPs and 3,205 SNPs). Two penalized approaches (rrBLUP, GBLUP), four Bayesian approaches (BayesA, BayesB, BayesC and Bayes Ridge Regression) and three machine leaning (ML) approaches (Support-vector regression, Extreme Gradient Boosting and Random Forests) were used in the prediction of the target traits including seed yield (YLD), plant height (HT), seed protein content (SPC), seed oil content (SOC) and seed glucosinolate content (GSL) to compare the model performance. We hypothesize that for all the traits of interest, the TP size and composition will affect GS prediction accuracy on the as well as marker density, the choice of GS models, and the nature of the target trait.

4.3 Materials and methods

4.3.1 Phenotypic data and genotypic data

The details of the phenotypic data collection and curation were described thoroughly in Chapter 3 under 3.3.1. Briefly, the training and validation populations used the "combined population" which consisted of the 91 parental genotypes and 345 hybrid genotypes (see Table S3.2 in Appendix for details). Phenotypic data were collected on seed yield (YLD), plant height (HT), seed protein content (SPC), seed oil content (SOC) and seed glucosinolates content (GSL).

The parental genotypes were tested under field conditions in RCBD experiments with three replications per site-year for five site-years in southern Manitoba (Glenlea 2016, Carman 2017, Portage 2017, Glenlea 2018, Portage 2018). The hybrid genotypes were tested in RCBD across 19 locations in Western Canada, totalling 43 site-years. However, unlike the parental genotypes, these hybrid genotypes did not have an equal number of replicates due to the nature of selection within a breeding program where only the favoured genotypes were selected for further field experiments. This led to an unbalanced phenotypic data set in the hybrids. To correct the unevenness, a best linear unbiased prediction (BLUP) value was calculated for each genotype, which was then used as the phenotype input for this chapter (see 3.3.1 for details).

DNA samples of all genotypes were extracted following a modified standard CTAB protocol (Porebski et al. 1997) that eliminated polyvinylpyrrolidone and 2-mercaptoethanol and replaced octanol with phenol. Genotyping took place at Agriculture and Agri-Food Canada (AAFC) Saskatoon (Dr. Isobel Parkin's lab) using the *Brassica* 60K Illumina Infinium SNP array (Illumina Inc., CA, USA). All three sets of markers described in sections 3.3.2 (MS-1 and MS-2) and 3.3.4 (the LD pruned markers) were used in this study.

4.3.2 Effect of training and validation population

The effect of training and validation population was completed based on the ridge regression best linear unbiased prediction (rrBLUP) (Endelman 2011). To estimate the effects of training population on prediction accuracy on the five traits of interest, two different approaches were used. In the first approach, all 91 parental genotypes were used as the TP (N_{TP} = 91) to predict the performance of the hybrids. In this approach, three methods to subsample the validation set (VP) were used including: 1) 69 random hybrid genotypes (20% of all hybrids) as the VP; 2) 207 random hybrid genotypes (60% of all hybrid genotypes) as the VP; and 3) all 345 hybrid genotypes as the VP. In the second approach, two methods were applied in selecting individuals as the TP. First, 91 random individuals were selected from the entire population (parents and hybrids) as the TP (N_{TP} = 91), which accounted for approximately 20% of the entire population. Secondly, a random subset of 262 individuals were selected as the TP (N_{TP} = 262), which also contained a mix of parental genotypes as well as hybrid genotypes and accounted for approximately 60% of the entire population.

4.3.3 Genomic selection with different marker density

In the previous chapter, a pruned set of markers that contained 3,205 SNPs (hereinafter referred to as MS-3 in this chapter) were used to analyze the population structure, while two sets of markers (MS-1 contained 26,651 SNPs and MS-2 contained 16,855 SNPs) were used in conducting the association analysis (see 3.3.2 for details). To estimate the effects of marker density in GS, all three sets of markers were used in this research. The model used in evaluating marker density effect was rrBLUP, as it is one of the most basic models in GS. When performing GS, all marker sets were imputed using the "A.mat" function in R package rrBLUP V. 4.6.1 (Endelman 2011), which replaced the missing data of a particular marker with its mean value across the population.

4.3.4 Parametric regression models

Two penalized approach GS models were used in this research including ridge regression best linear unbiased prediction (rrBLUP) (Endelman 2011) and genomic best linear unbiased prediction (GBLUP) (Clark and van der Werf 2013). The performance of four Bayesian models including Bayesian A (BayesA), Bayesian B (BayesB) (Meuwissen et al. 2001), Bayesian C (BayesC) (Habier et al. 2011b), Bayesian Ridge Regression (BRR) (Perez and de los Campos 2014) were compared.

The rrBLUP model was fit using the ridge regression and other kernels for genomic selection (rrBLUP) package V. 4.6.1. The rrBLUP package was developed mainly for performing GS based on mixed models (Endelman 2011). The core function of this package is "mixed solve", which solves single-variance mixed models (i.e. only one variance in the model except the error term) (Endelman 2011). The GBLUP and Bayesian models were fit using the R package Bayesian generalized linear regression (BGLR) V. 1.0.8 (Perez and de los Campos 2014). The default settings of BGLR were applied, i.e. the degrees of freedom was set as 5 and the scaled parameter was solved to coordinate with the partition of the phenotype variance (Perez and de los Campos 2014). The assumptions of the models were described in detail in section 2.5.1.4.

4.3.4.1 Cross validation

The cross-validation process (CV) was based on a subset of the population to validate another subset of the population (Haile 2018). The marker effect was computed based on a TP. The marker effect matrix obtained was then used in estimating the genomic predictions for the corresponding validation set (VP). Specifically, the marker effect matrix of the TP was multiplied with the marker matrix of the VP, and the product was the genomic estimated breeding values (GEBVs) of the genotypes in the VP. The Pearson's correlation values (*r*) were then computed between the GEBVs

and the observed phenotype of a trait, which was the prediction accuracy of the model. For the rrBLUP model, the TP sets used were described in detail in section 4.2.2. In rrBLUP, this process was iterated 500 times and the prediction accuracy was calculated as the grand mean of the 500 iterations. For the Bayesian models, the TP was 262 randomly sampled genotypes (60% of the entire population), and the VP represented the rest of the population. All Bayesian models were iterated 12,000 times and the burn-in was set as 5,000, meaning that the results from first 5,000 samples were discarded.

4.3.5 Non-parametric regression algorithms

Three ML algorithms were also considered for each trait: two tree-based ensemble ML methods [Extreme Gradient Boosting (XGBoost) (Chen and Guestrin 2016; Friedman 2001) and Random Forests (RF) (Breiman 2001)] as well as support-vector regression (SVR) using the MS-2 set of markers. As with the rrBLUP and GBLUP methods, 60% of the data was used for training and 40% for evaluation. The split of data into training and test sets was randomly performed 500 times. Support-vector regression is a ML algorithm that considers data instances in the training set as points in a high-dimensional vectors space (that is, the vector space has a dimension equal to the number of features for each instance) (Drucker et al. 1997).

All training and evaluation were done using Python 3.5 with the Scikit-learn package for RFs (SKL) (Pedregosa et al. 2011) and the XGBoost library (XGB) (Chen and Guestrin 2016). Missing training data were imputed using the most frequent value for that marker, then all data were encoded with a one-hot encoding, which was observed to improve prediction accuracy. For each training set, a grid search on hyperparameters with CV was performed (5-fold CV). That is, for each fold of the CV, all combinations of hyperparameters were considered and a model was trained

for each combination. The performance of all choices of hyperparameters was calculated for each fold.

The hyperparameters for the grid search of the three ML algorithms are listed in Table 4.1, – thus, on each fold, 6 x 5 x 4= 120 different XGBoost models were trained and evaluated. For CV, the folds were scored using Pearson's Correlation coefficient between the observed and predicted trait values. The best set of hyperparameters was chosen over all folds and the model was evaluated on the training data using these hyperparameters. The mean and standard deviation of the Pearson's Correlation coefficient on the testing set over the 500 iterations were reported.

4.4 Results

4.4.1 Genomic selection with different training population

The prediction accuracy tended to be very low when the TP only included parental genotypes, regardless of the size of the validation population, or the trait being predicted (Figure 4.1). When using the parental genotypes to predict a small subset of the hybrids (20%, 69 genotypes), the prediction accuracy for YLD was as low as 2%, which dropped to even lower (1%) when the VP size increased to 207 or 345, which accounted for 60% and 100% of the hybrid genotypes, respectively. The prediction accuracy was higher for height and the seed quality traits (SPC, SOC and GSL). The highest prediction accuracy was observed for GSL which varied between 32% to 33%.

When the TP consisted of a mix of parental and hybrid genotypes, the prediction accuracy increased significantly. Overall, a larger TP population size resulted in a higher prediction accuracy for YLD, when the TP of the same size with the parental genotypes was sampled randomly from the entire population (N_{TP} = 91), the prediction accuracy increased to 28%. It increased even higher to 38% when the randomly sampled TP size increased to 262 (N_{TP} = 262).

Table 4.1 Hyperparameters for grid search using three machine learning algorithms.

Machine learning algorithm	Hyperparameter	Range of values
Extreme Gradient Boosting (XGBoost)	Number of Trees	10, 50, 100, 200, 300, 400
	Maximum Depth of Tree	2, 5, 7, 10, 15
	Learning Rate	10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}
Random Forest (RF)	Number of Trees	10, 50, 100, 200, 300, 400
	Maximum Depth of Tree	2, 5, 7, 10, 15
Support vector regression (SVR)	RBF[1] Kernel - gamma	2^n for n = -17, -14, -11, -8, -5, -2, 1
	RBF Kernel - C	2^n for n = -5, -1, 3, 8, 12, 16
	Linear Kernel – C	1, 10, 100, 1000

[1] Radial basis function.

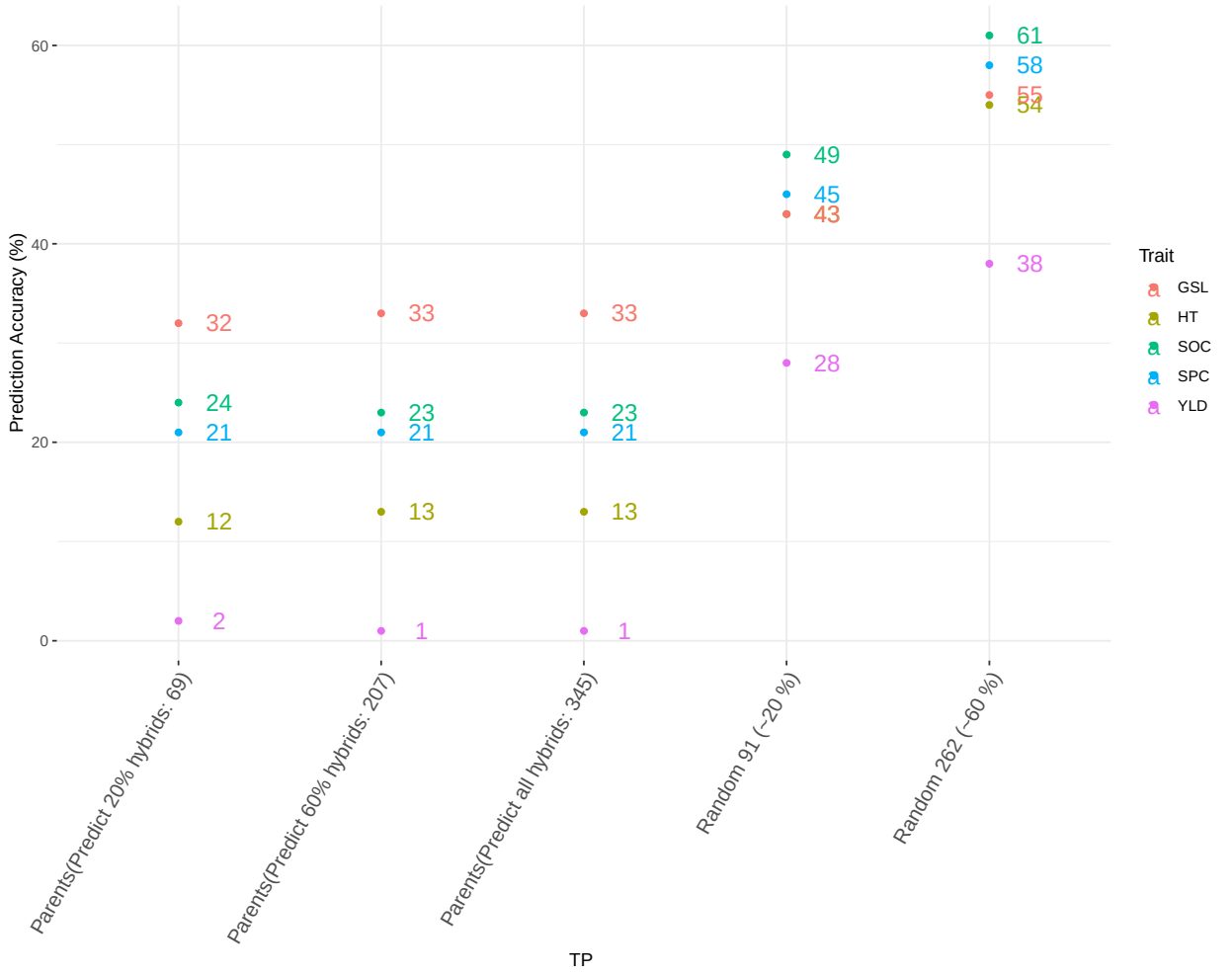


Figure 4.1 Prediction accuracy (Pearson's Correlation (%) between predicted and actual values) by rrBLUP (ridge regression best linear unbiased prediction) based on MS-1 that contained 26,651 SNP markers based on a combined population of *Brassica napus* L. consisting of 31 B-lines, 60 R-lines and 345 hybrids. The x-axis represents the different TP and VP types. Traits evaluated included YLD (seed yield), HT (plant height), SPC (seed protein content), SOC (seed oil content) and GSL (seed glucosinolate content). From left to right: using all parental genotypes to predict the performance of a subset of the hybrid genotypes (20% of hybrids); using all parental genotypes to predict the performance of a subset of the hybrid genotypes (60% of hybrids); using all parental genotypes to predict the performance of all hybrid genotypes; using randomly sampled 91 genotypes across the entire population, which accounted for about 20% of the population, to predict all hybrid genotypes; using randomly sampled 262 genotypes across the entire population, which accounted for about 60% of the population, to predict all hybrid genotypes. The y-axis represents the prediction accuracy in percentage. Traits are denoted by different colours.

This trend was consistent for HT and the seed quality traits (Figure 4.1). For example, the prediction accuracy of HT increased to 43% (overlapped with GSL in Figure 4.1) when based on a mixed TP of 91 genotypes, 31% higher than using a TP consisted of 91 parents. Prediction accuracy of SPC and SOC also increased to 45% and 49%, respectively. When the mixed TP increased to 262 randomly sampled genotypes, the prediction on HT increased to 54%. The prediction accuracy also improved on SPC, SOC and GSL when increasing the size of the mixed TP to 262 genotypes, reaching 58%, 61% and 55%, respectively.

4.4.2 Marker density affected prediction accuracy

Prediction accuracy varied as the marker density changed. As described in Chapter 3 section 3.4.2, marker density of MS-1, MS-2, and MS-3 were 24 kb/marker, 38 kb/marker and 188 kb/marker on the whole genome, respectively. This indicated that among the three marker sets, MS-1 had the best genome coverage while MS-3 had lower coverage (see details in section 3.4.2). Since the predictions were similar based solely on the parental genotypes regardless of the VP sizes as described in section 4.4.2, the group that was based on parental genotypes to predict all hybrid genotypes were chosen to present in the evaluation of the marker density effect.

The prediction accuracy varied with the rrBLUP model and different marker sets, although the difference varied depending on the trait (Figure 4.2). Based on a TP of 91 parents, MS-1, MS-2 and MS-3 produced similar prediction accuracy for YLD (0% to 6%), HT (12 to 14%), SPC (21% to 23%) and SOC (23% to 27%), but performed differently for GSL. Based on MS-2 the GSL prediction accuracy (51%) was significantly higher compared to MS-1 (33%) and MS-3 (31%).

The same trend was identified based on a mixed TP consisting of 91 or 262 randomly sampled individuals. Based on a TP consisting of 91 randomly sampled individuals MS-1, MS-2 and MS-3 produced similar prediction accuracy for YLD (28% to 30%), HT (43% to 44%), SPC (44% to

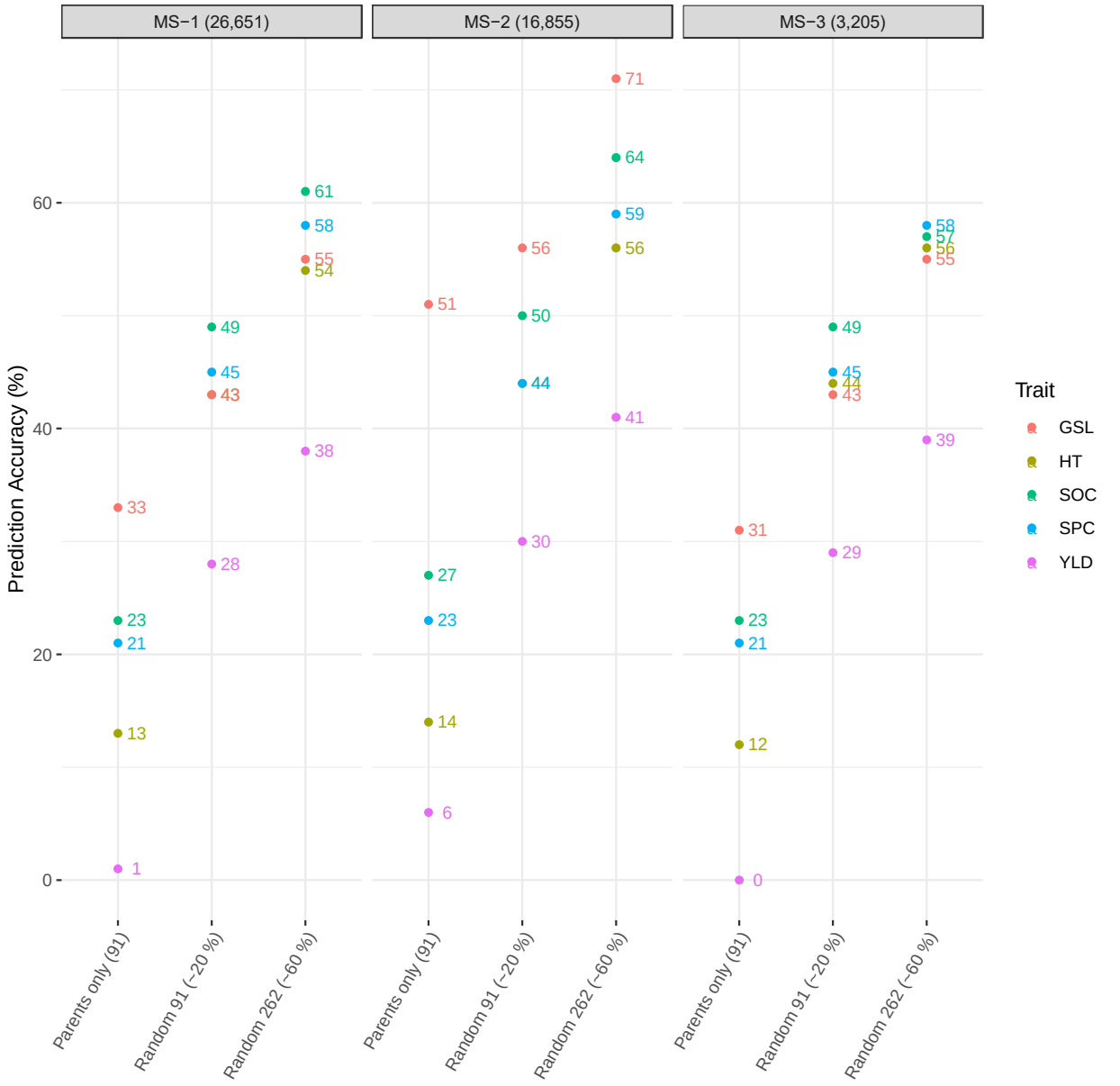


Figure 4.2 Prediction accuracy (Pearson's Correlation (%) between predicted and actual values) using rrBLUP (ridge regression best linear unbiased prediction) based on all three marker sets and a combined population of *Brassica napus* L. consisting of 31 B-lines, 60 R-lines and 345 hybrids. MS-1, MS-2 and MS3 contained 26,651, 16855 and 3,205 SNP markers, respectively. Traits evaluated included YLD (seed yield), HT (plant height), SPC (seed protein content), SOC (seed oil content) and GSL (seed glucosinolate content). Each panel represents prediction accuracy from a marker set. The x-axis represents the different training set, and the y-axis represents the prediction accuracy in percentage. Traits are denoted with different colours.

45%) and SOC (49% to 50%) but performed differently for GSL. The highest prediction for GSL (56%) was produced with MS-2, which was higher compared to MS-1 (43%) and MS-2 (43%). Based on the TP that consisted of 262 randomly sampled individuals MS-1, MS-2 and MS-3 produced similar prediction accuracy for YLD (38% to 41%), HT (54% to 56%), SPC (58% to 59%) and SOC (57% to 64%) but performed differently for GSL. Marker set-2 produced the highest prediction for GSL (71%) which was higher compared to MS-1 (55%) and MS-2 (55%).

4.4.3 Model performance comparison

Since MS-2 performed the best using the TP consisting of 262 randomly selected genotypes (60% of the entire population), the comparison of model effects was performed based on these two attributes (N_{TP}=262 and MS-2). Overall, moderate to high predictions were obtained based on the Bayesian models, GBLUP and the ML algorithms, which all performed better compared to rrBLUP (Figure 4.3).

4.4.3.1 Parametric regressions

Within the Bayesian models, BayeB produced a prediction accuracy of 76% for YLD, which performed slightly better than the BayesA (73%), BayesC (69%) or BRR (69%) (Figure 4.3). BayesB also produced the highest prediction accuracy for HT (88%), which was slightly better than BayesA (87%), Bayes C (84%) and BRR (84%). For the seed quality traits SPC, SOC and GSL, the Bayesian models performed quite similar with each other. Among all traits the biggest difference in prediction accuracy due to the choice of models was observed in YLD. The prediction accuracy difference on seed quality was very minimal for SPC, SOC and GSL.

GBLUP was more efficient in terms of computation time while producing similar prediction accuracy with the Bayesian models. For YLD, even though lower than BayesB by 8%, GBLUP's

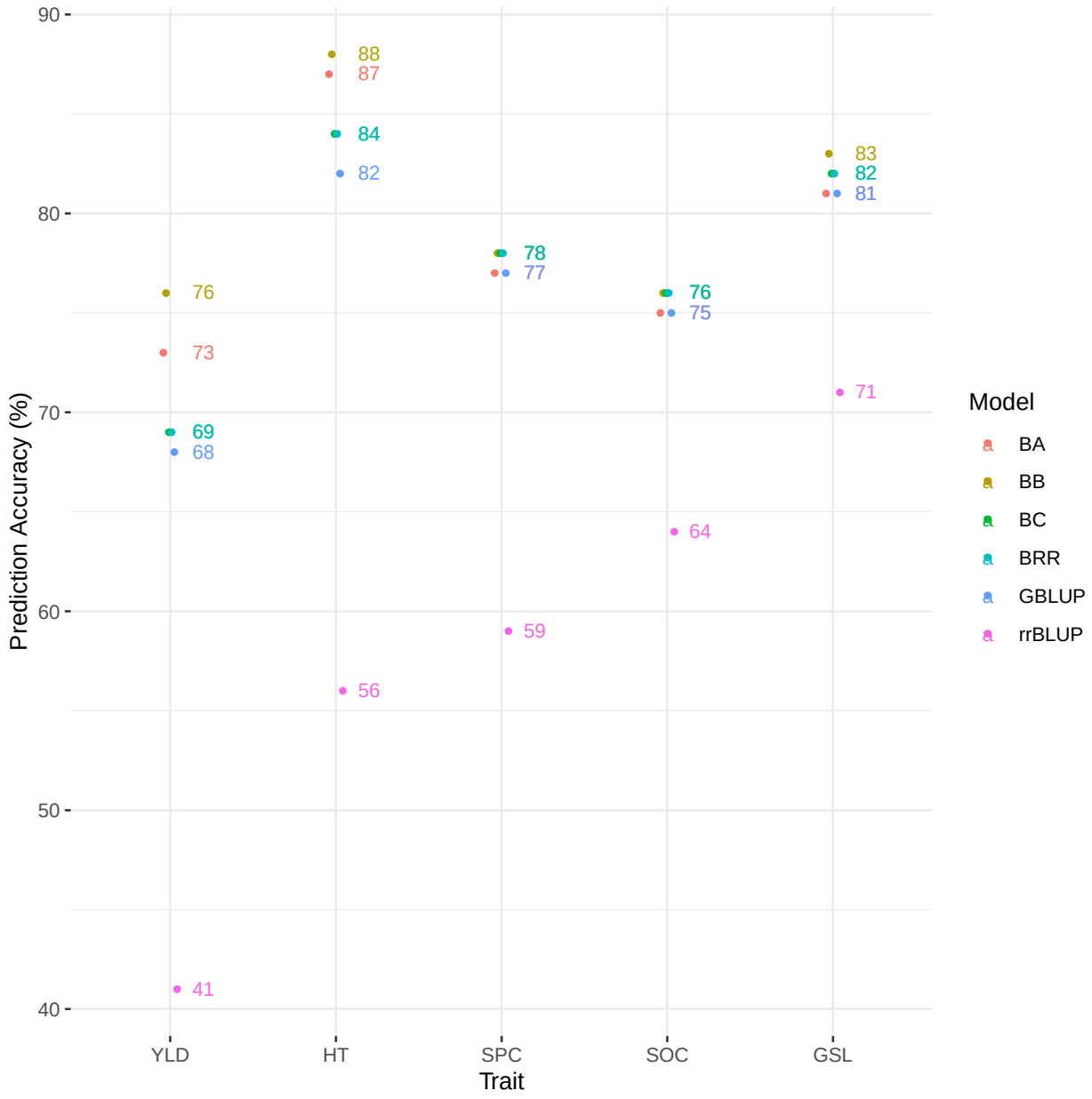


Figure 4.3 Prediction accuracy (Pearson's Correlation (%) between predicted and actual values) comparison based on BayesA, BayesB, BayesC, BRR, GBLUP and rrBLUP using MS-2 using 262 randomly sampled individuals as the TP based on a combined population of *Brassica napus* L. consisting of 31 B-lines, 60 R-lines and 345 hybrids. Traits evaluated included YLD (seed yield), HT (plant height), SPC (seed protein content), SOC (seed oil content) and GSL (seed glucosinolate content). The x-axis represents the traits. The y-axis represents the prediction accuracy. Abbreviations: BRR: Bayesian Ridge Regression; rrBLUP: (ridge regression best linear unbiased prediction); GBLUP: genomic best linear unbiased prediction.

prediction accuracy (68%) was only 1% lower than BayesC and BRR. For HT, GBLUP had a slightly lower prediction accuracy (82%) compared to the Bayesian models (84% to 88%).

For SPC, SOC and GSL, GBLUP had the same prediction accuracy with BayesA. Using the rrBLUP model, YLD had the lowest prediction accuracy (41%), while HT, SPC, SOC and GSL obtained moderate prediction accuracy which varied between 56% to 71%, depending on the trait.

4.4.3.2 Non-parametric regressions

Example scatter plots of a random run of the grid search for each of the three ML models are shown in Figure 4.4. For SVR (Figure 4.4.A), this figure shows the performance of one of the 500 train-test iterations. In particular, the behaviour of the optimal model, given by the hyperparameters chosen by the 5-fold CV, is shown on both the training and test set. In this example, the Pearson's Correlation coefficient on the test set was 73%, while on the training set, it is 85%. The difference between training and test performance was similar for the other two models (Figure 4.4.B and Figure 4.4 .C) and this comparison between ML methods was similar for all traits (not shown in the results). While the model shows somewhat higher performance on the training set, an indication of possible overfitting was also observed in the scatter plot where the test set (black dots) were not identified in the tail of the training set (red dots) (Figure 4.4.C).

The overall performance of the model (i.e. the correlation between the predicted and actual values) on the test sets and the competitiveness with the other methods demonstrated that the results were suitable for general use (Table 4.2). Among all non-parametric methods BayesB had the best prediction accuracy overall and therefore, was selected for comparison with the ML methods. With the ML models, the YLD was the most difficult to predict, as YLD had the lowest prediction accuracy among all traits (69% to 72%), which was slightly lower than BayesB (76%). For HT, SPC and SOC, ML algorithms also performed well and produced slightly lower

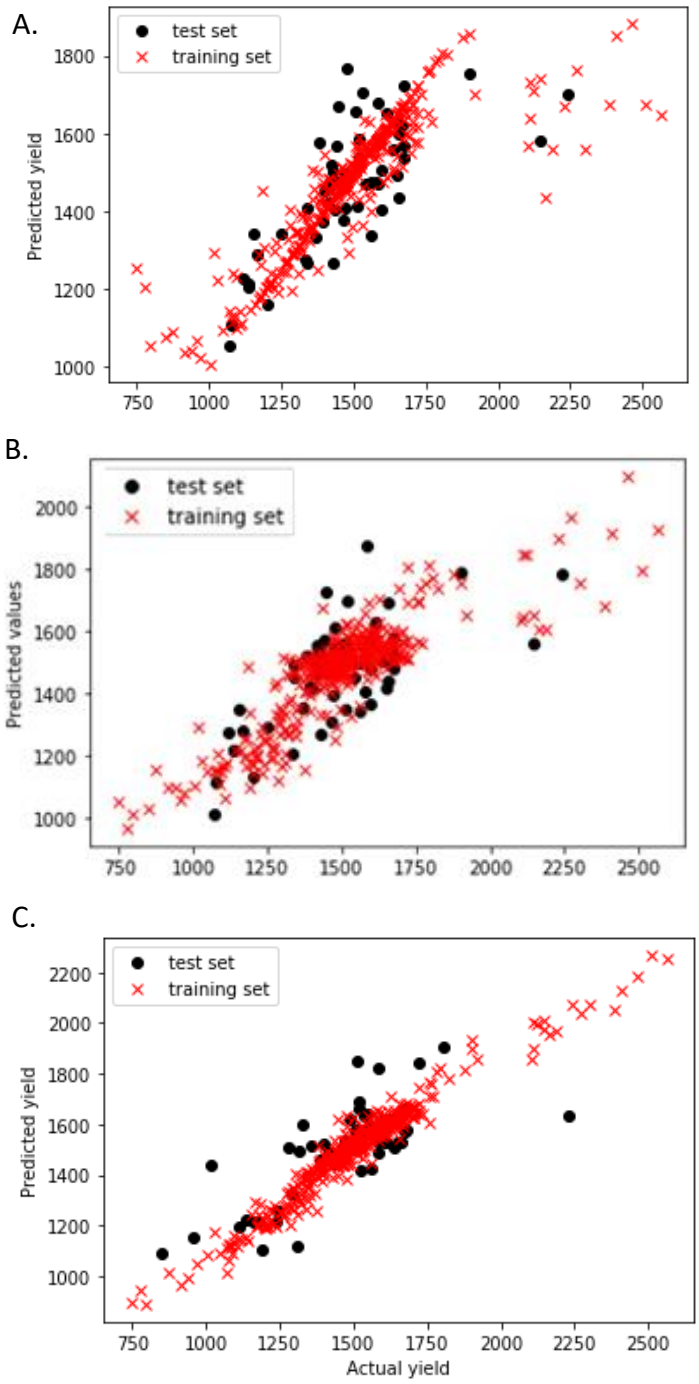


Figure 4.4 Example scatter plots of a random run (one of the 500 train-test iterations) of the grid search for (A) Support-vector machines (gamma = 2^{-14}, C=2048); (B) Extreme gradient boosting scatter (PCC test 0.688), and (C) Random forests scatter (PCC test 0.744) based on seed yield of a *Brassica napus* L. population consisting of 31 B-line, 60 R lines and 345 hybrids. Black dots represent model performance based on the test set (VP) and the red X's represent model performance based on the training set (TP). The x-axis represents the observed yield and the y-axis represents the predicted yield.

Table 4.2 Mean Pearson's Correlation between predicted and actual values over 500 iterations of three machine learning algorithms: Support vector regression (SVR), Extreme Gradient Boosting (XGBoost) and Random Forests (RF) and BayesianB (BayesB) (the best performing parametric model in this research) based on a *Brassica napus* L. population consisting of 31 B-lines, 60 R-lines and the 345 hybrids. Five traits evaluated included seed yield (YLD), plant height (HT), seed protein content (SPC), seed oil content (SOC) and seed glucosinolate content (GSL).

	YLD	HT	SPC	SOC	GSL
XGBoost	69%	81%	73%	73%	84%
RF	70%	82%	71%	72%	84%
SVR	72%	81%	74%	73%	87%
BayesB	76%	88%	78%	76%	83%

predictions (81% to 82%, 71% to 74%, and 72% to 73% for HT, SPC and SOC, respectively) than BayesB (88%, 78% and 76%). All ML algorithms had higher prediction accuracy than BayesB on GSL. XGBoost and RF outperformed BayesB by 1% and SVR outperformed it by 4%. Standard deviations were also calculated for the ML methods and found to be minimal (varied between 0.04 to 0.08 for 500 iterations).

4.5 Discussion

The composition of the TP is a key factor for the success of GS (Voss-Fels et al. 2019). Results from this study showed that both the composition of the population and the size of the population had effects on the prediction accuracy of GS. When using parents only to predict the performance of the hybrid genotypes, the prediction accuracy was quite low. The size of the VP did not have significant impacts on the prediction accuracy. When the TP only contained parents, the prediction accuracy of the same trait was quite similar. Daetwyler et al. (2013) and Asoro et al. (2011) demonstrated that the prediction accuracy tended to be higher when the TP was closely related to the genotypes being predicted. Norman et al. (2018) suggested implementing a TP highly related to the individuals being predicted to increase the genetic response. This research revealed that the prediction accuracy increased with the addition of the hybrids into the TP. A randomly sampled TP (parents and hybrids) had a higher predictive ability (for all traits), which indicated that prediction of hybrid genotypes cannot solely rely on the parental genotypes. Similarly, Liang et al. (2018) found in pearl millet [*Cenchrus americanus* (L.) Morrone.] hybrids that the addition of inbred parents (i.e. computing the BLUPs of the population after adding inbred parents to hybrids) instead of accounting for impacts of heterosis actually reduced the prediction accuracy, especially for traits with high heterosis. By implementing phenotype BLUP of parental and hybrid genotypes into GS, this negative effect caused by simply combining inbred and hybrid data could be cancelled

and therefore increase the prediction accuracy for some traits compared to performing genomic prediction based on hybrid genotypes alone (Liang et al. 2018). In terms of TP size effects, the prediction accuracy increased as a randomly sampled TP increased in size (91 to 262 genotypes). This indicated that a larger TP produced a higher prediction accuracy and is consistent with previous studies (Asoro et al. 2011; Norman et al. 2018; Tayeh et al. 2015; Xu et al. 2018).

A difference in prediction accuracy among traits was also observed in this study. Overall, the lowest prediction was observed for YLD, which was expected due to the very complex nature of the trait. The prediction accuracy of 0% for yield based on MS-3 indicated a negative impact of low marker density on prediction accuracy for a complex trait. Norman et al. (2018) suggested that more variation in prediction accuracy was identified in grain yield when the prediction was based on a smaller TP. Differences in the prediction accuracy can be affected by the genetic complexity of the trait. Therefore, a larger TP is needed to provide more allelic observations for the prediction of small effect QTL on more complex traits (Gilmour 2007). However, Maulana et al. (2021) found that the variation in the prediction accuracy responding to the TP size change was very minimal when investigating a complex trait, and suggested that the relatedness between TP and VP is more important in achieving higher prediction accuracies. In this research, we observed that compared with the seed quality traits, YLD had large variation responding to a TP size increase, TP composition change, and marker density, as well as different models. Therefore, we suggest the complexity of the target trait has a large impact on the prediction accuracy.

Overall, the impact of marker density was not as great as the TP effect, which is consistent with results from a study on hybrid rice (Xu et al. 2018). Although MS-1 had the highest maker density, it did not have the best prediction accuracy among the three marker sets. Hickey et al. (2014) described this situation as the model overfitting by a large number of markers, where markers

accounted for non-genetic effects. This can cause a decrease in the prediction accuracy when the prediction was performed on data sets that do not have common non-genetic effects (Jannink et al. 2010). In particular, higher marker densities are more advantageous when a large TP is utilized, especially when prediction is performed among unrelated individuals (Meuwissen 2009). Authors of a previous study conducted in a wheat population found that the prediction accuracy increased with an increased marker density when the TP and VP were more distinct (Norman et al. 2018). Therefore, marker density required in GS is indeed associated with the relatedness among individuals in the target population. In this research we found that MS-2 performed well compared with a larger number of markers (MS-1), which indicated that lower marker density is needed when there is high relatedness between the TP and VP. As stated by Meuwissen (2009), the required marker density is lower when the individuals being predicted are the progenies of the individuals from the training set. Hickey et al. (2014) had the same findings based on a simulated maize population, suggesting that lower marker density is needed when the TP and VP share highly related genotypes. Some studies also demonstrated that increasing the phenotypic data had larger impacts on GS prediction accuracy than increasing marker density (Lorenz et al. 2011; VanRaden et al. 2009). This suggests that even though MS-2 (16,855) performed the best in this research, the predictive ability could still be improved if the size of the TP increased. The results from this study also indicated that a subset of markers evenly distributed along the genome could be sufficient to perform GS. This has benefits in reducing the cost of genotyping as well as the computation time. In this research, most of the models performed quite similar except for rrBLUP. GBLUP was found to be as good as more complicated models such as Bayesian models. Knoch et al. (2021) examined prediction accuracy of hybrid performance based on parental empirical data and large omics datasets using GBLUP and obtained moderate to high prediction accuracies on SOC, SPC and

GSL, which is consistent with our results. Even though GBLUP was described as mathematically equivalent with rrBLUP (Habier et al. 2007) there are still some differences between them. More specifically, they compute GEBVs through different approaches. In rrBLUP, all markers are assumed to have equal variances while the marker effects shrink to zero, and GEBVs are estimated based on marker effects (Endelman 2011). In GBLUP, the markers are used to compute a genomic relationship matrix (GRM), which is then used to estimate the breeding value of an individual (Clark and van der Werf 2013), meaning that the computation of GEBVs does not depend on the marker effect estimation (Tan et al. 2017). The comparison on prediction accuracy between GBLUP and rrBLUP was not consistent in the past. For example, studies found rrBLUP could outperform GBLUP (Wang et al. 2015b; Wang et al. 2015c); whereas in other studies, the performance of GBLUP was quite similar with rrBLUP (Bhering et al. 2015; Gilmour 2007; Habier et al. 2007; Tan et al. 2017). In the current research, GBLUP performed significantly better than rrBLUP for all traits. In fact, rrBLUP had the lowest prediction accuracy among all six models evaluated. We assume this might be caused by the difference in the algorithms as well as the assumptions of the variances of these models. More specifically, since marker effects estimated are based on a $m \times m$ matrix in rrBLUP, where m represents the number of markers, it potentially introduces too much noise, negatively affecting the prediction accuracy. It is commonly known that rrBLUP can only capture additive effects in the model, which leads to the fact that the heterosis in the hybrids cannot be well captured. Therefore, we assume the deficiency observed in this research in rrBLUP compared with other models was possibly due to rrBLUP not being capable of characterizing the actual QTL effects.

In a simulation study that compared rrBLUP, GBLUP, BayesA, BayesB, BayesC π and BayesLASSO, Wang et al. (2015c) found that BayesB produced the highest prediction accuracy

when the target trait had a lower heritability (0.3 and 0.5), or when the trait was impacted by a small number of QTL (20). In another simulation study that compared GBLUP and BayesB based on different population sizes and numbers of QTL, Daetwyler et al. (2013) found that BayesB performed better than GBLUP when the number of QTL was low. However, as the number of QTL increased, GBLUP outperformed BayesB. These findings are consistent with Wang et al. (2015b), based on simulations that BayesB outperformed GBLUP in some cases where a large number of markers were applied and the marker effects were assumed to follow a non-normal distribution. In our research, BayesB slightly outperformed other models on YLD, HT and GSL, while obtaining the same prediction accuracy with BayesC and BRR on SPC and SOC. Even though BayesB could produce higher prediction accuracy in some situations (e.g. the trait being controlled by moderate to large-effect QTL), it is not computationally efficient compared to GBLUP due to the algorithm of BayesB (Metropolis Hastings algorithm) (Wang et al. 2015b). BayesA uses Gibbs sampling which often needs less computation time, but is still slow when large numbers of markers and genotypes are applied (Wang et al. 2015b). Although the difference in prediction accuracy among the models was minimal, our results align with BayesB having the highest prediction accuracy on YLD. The similarity in prediction accuracy between GBLUP and the Bayesian models indicated the complexity in the traits being predicted or the differences need larger data sets for accurate prediction (Daetwyler et al. 2013). Ali et al. (2020) investigated six traits and seven GS models in winter wheat and found that no GS models consistently outperformed others. Similarly, Daetwyler et al. (2013) also stated no single GS model could be used as the benchmark for genomic prediction. Therefore, they recommended comparing methods where all loci are treated as equal contributors to the target trait with a variable selection model where some loci contribute more to the target trait (eg. BayesB).

In this research we also applied three commonly used ML algorithms, which produced medium to high prediction accuracy for all five traits of interest. As non-parametric models, ML algorithms are expected to capture different relationships between markers and phenotypes compared to the linear models for GS (Heslot et al. 2012). Therefore, ML algorithms are more flexible in managing complicated associations (Montesinos-Lopez et al. 2021). Despite the minimal differences observed between the different algorithms, our results showed consistency between BLUP and the ML models, as they performed relatively similarly on all traits. Compared to the other four traits, YLD had lowest prediction accuracy based on the ML algorithms which might be due to the nature of the trait. Except for the genetic nature, other factors could affect seed yield such as agronomic practice and environmental effects, as well as the interaction among these factors (Parmley et al. 2019). In our research, on both the training and test set, the SVR model had reduced ability to make predictions accurately for YLD at the top range of performance. This was observed relatively consistently over all iterations. However, the overall selection of promising candidates (i.e., those with high YLD) would not be significantly compromised. All three ML algorisms outperformed the parametric regressions for GSL. This could be related with the high heritability of GSL in *B. napus* (Kittipol et al. 2019). Deep learning (DL) is a subset of ML procedures (Abdollahi-Arpanahi et al. 2020). Montesinos-Lopez et al. (2021) reviewed the applications of DL in GS for animal and plant breeding and concluded that high quality data of the TP and a large TP are essential for DL approaches. In our research, the ML methods performed well and produced high prediction accuracy with low standard deviation. However, similar to Montesinos-Lopez et al. (2021), we did not identify significant improvement on the prediction power based on ML algorithms compared to conventional GS methods.

4.6 Conclusion

Genomic selection is a promising tool in predicting the performance of *B. napus* hybrids. Breeders may decrease the number of hybrids that require development and field evaluation and as a result perform selection more efficiently based on the results of GS. Even though multiple factors could affect prediction accuracy simultaneously, we found that the magnitude of their impacts did vary. In this research, TP composition and size has significantly higher impacts on prediction accuracy than marker density. Model performance, on the other hand, did not affect prediction accuracy greatly as most of them had very similar prediction accuracy. However, it is still recommended that the breeders try applying various models at the same time on different traits of interest. Importantly, the prediction of hybrid performance needs to include parental and hybrid data. Machine learning algorithms performed very similar to the conventional GS models and produced high prediction accuracy in this research, but would require larger TP with high data quality to obtain better performance. When taking all these factors mentioned above (nature of trait, marker density, TP effect, prediction model) into account, moderate to high prediction accuracy can be produced for the important traits we investigated based on models that only consider additive effects.

5. GENOME-WIDE ASSOCIATION STUDY – GUIDED GENOMIC SELECTION OF AGRONOMIC AND SEED QUALITY TRAITS IN *Brassica napus* L.

5.1 Abstract

The use of genomic selection (GS) in plant breeding has gained popularity because it offers several advantages over conventional phenotypic selection such as reduced breeding cycles and cost effectiveness through reducing required field experiments. Genomic selection has been applied in improving seed yield as well as seed quality traits in canola. However, none of the currently available models have performed consistently across populations or traits in canola. We examined the prediction accuracy of genome-wide association study (GWAS) – guided genomic selection (GS). First proposed in 2016, this method uses significant SNPs identified from the training set to fit GS models as fixed effects and the results are then validated using a cross-validation (CV) technique. FarmCPU (Fixed and random model circulating probability unification) was applied to identify the significant SNPs that were fit in the GS models as fixed effects. Six parametric GS models including BayesianA (BayesA), BayesianB (BayesB), BayesianC (BayesC), Bayesian Ridge Regression (BRR), genomic best linear unbiased prediction (GBLUP) and ridge regression best linear unbiased prediction (rrBLUP) were applied in evaluating the prediction accuracies for five traits including seed yield (YLD), plant height (HT), seed protein content (SPC), seed oil content (SOC) and seed glucosinolates content (GSL). Two marker sets with different marker densities were used in the analysis. The results revealed that the prediction accuracies of GWAS-guided GS varied based on the choice of GS models and were higher than conventional GS methods in several situations. Conventional GBLUP and rrBLUP were robust and had stable performance across different traits based on marker set 1 (MS-1, 26,651 SNPs). However, based

on MS-2, Farm-CPU guided Bayesian models had better performance compared to conventional Bayesian models across all traits. Marker density did not have significant impacts on the prediction accuracy of conventional GS but did have impacts on that of the FarmCPU-guided GS models. In general, prediction accuracy increased based on MS-2 compared to MS-1 for FarmCPU-guided GS approaches for all traits except SOC. Marker density also greatly affected the computational efficiency. Compared with FarmCPU-guided Bayesian approaches based on MS-2, conventional Bayesian models took an extra 4.6 to 6.4 h to complete based on MS-1. Overall, GWAS-guided GS is a promising tool, but will need to be performed with caution as its prediction accuracy is impacted by multiple factors.

5.2 Introduction

In the early 2000s, genomic selection (GS) was first proposed as a method to increase the efficiency of animal breeding (Meuwissen et al. 2001). Although considered as a variant of marker assisted selection (MAS), GS does not focus on major-effect quantitative trait loci (QTL). Genomic selection utilizes all markers across the whole genome and assumes the loci that control the target trait are in linkage disequilibrium (LD) with one or more markers (Desta and Ortiz 2014). Since the first applications in animal breeding, GS has gained great success and has been embraced by plant breeders as a new method to improve plant breeding efficiency for a wide range of crops including canola (Jan et al. 2016; Snowdon and Iniguez Luy 2012; Würschum et al. 2014). However, none of the existing GS models can serve as a benchmark in all situations due to variations in multiple aspects (e.g., population history and genome structure) (Daetwyler et al. 2013).

Several approaches have been proposed to combine genome-wide association study (GWAS) and GS to improve the prediction accuracy (Bian and Holland 2017; Fiedler et al. 2017; Spindel et al.

2015; Tsai et al. 2020; Zhang et al. 2014c). Most of these studies attempted to integrate GWAS results in GS models as fixed effects. For example, Zhang et al. (2014c) proposed a GS model called "BLUP|GA" that applied previous knowledge of significant loci detected by GWAS. However, this method could be problematic considering that the GWAS results are often affected by the population structure (Ibrahim et al. 2020). This indicates that previous GWAS results might not reflect the true genetic structure of the traits in the population of interest (Spindel et al. 2016). Bian and Holland (2017) proposed new main and nested-effect GWAS models based on a multi-parental NAM population in maize and combined it with GBLUP for genomic prediction on simulated traits. Genomic selection was also conducted simultaneously on advanced breeding lines with GWAS in barley (*Hordeum vulgare* L.) and wheat (*Triticum aestivum* L.) to study the genetic structure of the traits of interest and examine the prediction accuracy (Tsai et al. 2020). A major issue remained in this type of research regarding how to avoid bias and validate that the identified significant SNPs are truly contributing to the traits of interest. More specifically, breeders need to remain cautious regarding the significant markers implemented in the GS model, and whether they could provide comprehensive background information about the genetic architecture of a certain trait. Therefore, Spindel et al. (2016) proposed a new approach that incorporated GS and *de novo* GWAS to improve the prediction accuracy. Instead of utilizing the existing knowledge of QTL obtained from previous GWAS research, their "GS + *de novo* GWAS" method uses significant SNPs identified from the training set to fit into the GS models as fixed effects and the results were validated using a cross-validation (CV) technique. This approach, in theory, should perform more efficiently than the GS + historical GWAS method or conventional GS as it utilizes significant SNPs that are obtained from the population being investigated (Spindel et al. 2016).

Genomic selection, as a newer tool in canola breeding, has been applied in predicting flowering time (Li et al. 2015a), plant height (Würschum et al. 2014), grain yield and seed glucosinolate content (Jan et al. 2016), and blackleg [*Leptosphaeria maculans* (Desm.) Ces. & de Not.] resistance (Fikere et al. 2018). However, to our knowledge, there has been no reported GWAS-guided GS research on canola based on the "GS + *de novo* GWAS" method. Therefore, we performed GWAS-guided GS to examine its effectiveness in improving the prediction accuracy based on a population of parents and their hybrids. The GS + *de novo* GWAS approach proposed by Spindel et al. (2016) was used in this research. One GWAS model (FarmCPU) and six GS models (four Bayesian models, GBLUP and rrBLUP) were used to evaluate the prediction accuracy.

5.3 Materials and methods

5.3.1 Phenotypic data and genotypic data

An in-depth description of the collection and curation of phenotypic data was provided in chapter 3 in section 3.3.1. Briefly, the training and validation population in this chapter consisted of the "combined population" which included 91 parental genotypes (31 B-lines and 60 R-lines) and 345 hybrid genotypes (see Table S3.2 in Appendix for details). The following phenotypic data were collected: seed yield (YLD), plant height (HT), seed protein content (SPC), seed oil content (SOC) and seed glucosinolates content (GSL) as described in section 3.3.1.

The parental genotypes were tested under field conditions in multiple RCBD experiments with three replications per site-year for five site-years in southern Manitoba (Glenlea 2016, Carman 2017, Portage 2017, Glenlea 2018, Portage 2018). The hybrid genotypes were tested in 19 locations in Western Canada, totalling 43 site-years. Due to the manner the hybrid selection was conducted, only selected genotypes were advanced for further field evaluation, leading to hybrid

genotypes with unequal numbers of replications. Consequently, there was unevenness in the phenotype data. Thus, a BLUP (best linear unbiased prediction) value was computed for each genotype to standardize the data for each trait (Piepho et al. 2008; Robinson 1991). The BLUP values served as the phenotypic input of this section (see section 3.3.1 for further details).

5.3.2 Genotypic data

Genotyping of the population took place at Agriculture and Agri-Food Canada (AAFC) Saskatoon (Dr. Isobel Parkin's Lab) using the *Brassica* 60K Illumina Infinium SNP array (Illumina Inc., CA, USA). The genotypic data were exported from GenomeStudio 2.0 software version 2.0.4 (Illumina Inc., CA, USA). Two marker sets, MS-1 with 26,651 SNPs and MS-2 with 16,855 SNPs were used in this chapter as the genotypic data. They were then converted to hapmap format in Trait Analysis by Association Evolution and Linkage (Tassel) 5 version 20200110 (Bradbury et al. 2007) and imported to Intelligent Prediction and Association Tool (iPat) version 1.0 (15.0.1) (Chen and Zhang 2018) for data conversion. Two output files were created after the conversion. One converted file, called the dat file (.dat), included the genotype names and their corresponding marker information in a numeric format. Homozygous genotypes were converted to 0s and 2s while heterozygous genotypes were converted to 1s. Missing data was formatted as NA. The other converted dataset was called the map file (.map), containing the name of the SNP markers as well as their chromosome numbers and positions in base pairs (bp).

5.3.3 Genome-wide association-guided genomic selection based on rrBLUP and GBLUP

Methods described here were modified from a previous study on combining GWAS in GS by implementing significant markers identified from GWAS as the fixed effects in GS (Spindel et al. 2016). The phenotype data (BLUPs), genotype file (.dat) and SNP marker information file (.map) were loaded into iPat for the GWAS analysis. Since iPat could not process a large marker data set in FarmCPU-guided GS analysis based on rrBLUP and GBLUP, GWAS and GS analysis were conducted separately for GWAS-rrBLUP and GWAS-GBLUP based on MS-1 in this research, as suggested by the author of iPat (Dr. Chunpeng James Chen, University of California, Davis).

FarmCPU is a multi-locus GWAS model that provides control in both false positives and false negatives (Liu et al. 2016b). Based on the results from Chapter 3 (Figure 3.10 and Figure 3.11), FarmCPU controlled both false positives and false negatives better than other GWAS models since it "followed closely to the 1:1 line with a sharp upward deviated tail" (Kaler et al. 2019). This was consistently observed based on both MS-1 and MS-2 in the combined population (see details under section 3.5.2). Therefore, FarmCPU was selected to identify significant MTAs in this research. Genome Association and Prediction Integrated Tool (GAPIT) 3 (Wang and Zhang 2020) (previously used in chapter 3) was implemented to iPat, which included fixed and random model circulating probability unification (FarmCPU) (Liu et al. 2016b). The number of principal components (PCs) included in the GWAS analysis was set as 3 which was the default setting of the program. After the completion of the GWAS analysis, three SNP markers with the lowest false discovery rate (FDR) adjusted *p* values of each trait from each GWAS model were extracted from the original marker data. The genotypic data of these SNPs were then converted to a numeric format as described in section 5.3.2 and this was saved as the covariate input file (.cov), which was later fit in the GS models as fixed effects. The genotypic data without the extracted SNPs was

saved as the genotype input for each trait in GS. Thus, each trait has its own set of genotypic data covariates for the GS process.

The phenotypic data, genotypic data, map file and the covariates were then loaded into iPat for the GS analysis. For MS-2, iPat was able to process the marker data set and the input data were directly loaded into iPat for analysis. For both rrBLUP and GBLUP, the validation on accuracy option was selected and the fold number was set as 5 for CV, with 100 iterations which was the highest number of iterations available. Computation time was recorded for each analysis for efficiency comparison between models.

5.3.4 Genome-wide association-guided genomic selection based on Bayesian models

In this section, the input data were the same as section 5.3.3 and both MS-1 and MS-2 were used. The "BGLR" package (Perez and de los Campos 2014) was implemented in iPat and BayesA, BayesB, BayesC and Bayes Ridge Regression (BRR) were used in the GS step. The parameter for number of iterations (nIter) was set as 10,000 and the burn-in period was set as 3,000. For the CV process, CV fold number was set as 5, and the iteration number was set as 10. Computation time was recorded for each analysis.

5.3.5 Conventional genomic selection

Instead of directly comparing results from this chapter and Chapter 4, conventional GS was conducted in iPat since iPat utilizes a different CV technique from Chapter 4. In this technique, the population was divided into five subsets of approximately equal size. In each fold, four of these subsets were combined to form the training set, with the remaining subset representing the validation set. This process was repeated until the five subsets rotated, and each was used as the validation set once. Pearson's correlation coefficients (r) were calculated between the predicted genomic estimated breeding values (GEBVs) and the observed phenotype of the validation set.

The mean of the Pearson's correlation coefficients from each fold was reported as the final prediction accuracy of a particular trait based on a specific model or marker set.

The input data were the same as section 5.3.3 and both MS-1 and MS-2 were used. For rrBLUP and GBLUP, the CV fold number and iteration number were set as 5 and 100, respectively. For Bayesian models, the same settings as described in section 5.3.4 were used. Computation times were recorded for all models to compare the efficiency of different models.

5.4 Results

5.4.1 Genome-wide association-guided genomic selection based on 26,651 SNPs

Prediction accuracies on YLD were relatively low overall and had large variation in FarmCPU-guided GS (Figure 3.1). The highest prediction accuracy was 37% (FarmCPU-guided rrBLUP) and the lowest was 0 (FarmCPU-guided BayesA and FarmCPU-guided BRR). Conventional GS had more uniform performance across different GS models, which varied between 22% to 29%. For HT, there was larger variation in the prediction accuracy which varied between 4% to 45% based on FarmCPU-guided GS, with the lowest produced by FarmCPU-guided BayesA and the highest produced by FarmCPU-guided rrBLUP. Based on the conventional GS models the prediction accuracy was more uniform, which varied between 40% to 46% for HT. Similarly, based on FarmCPU-guided GS models prediction accuracy on SPC also had large variation (2% to 48%). FarmCPU-guided Bayesian models had relatively higher prediction accuracy (47% to 48%) while FarmCPU-guided GBLUP and FarmCPU-guided rrBLUP had lower prediction accuracy (2% and 30%).

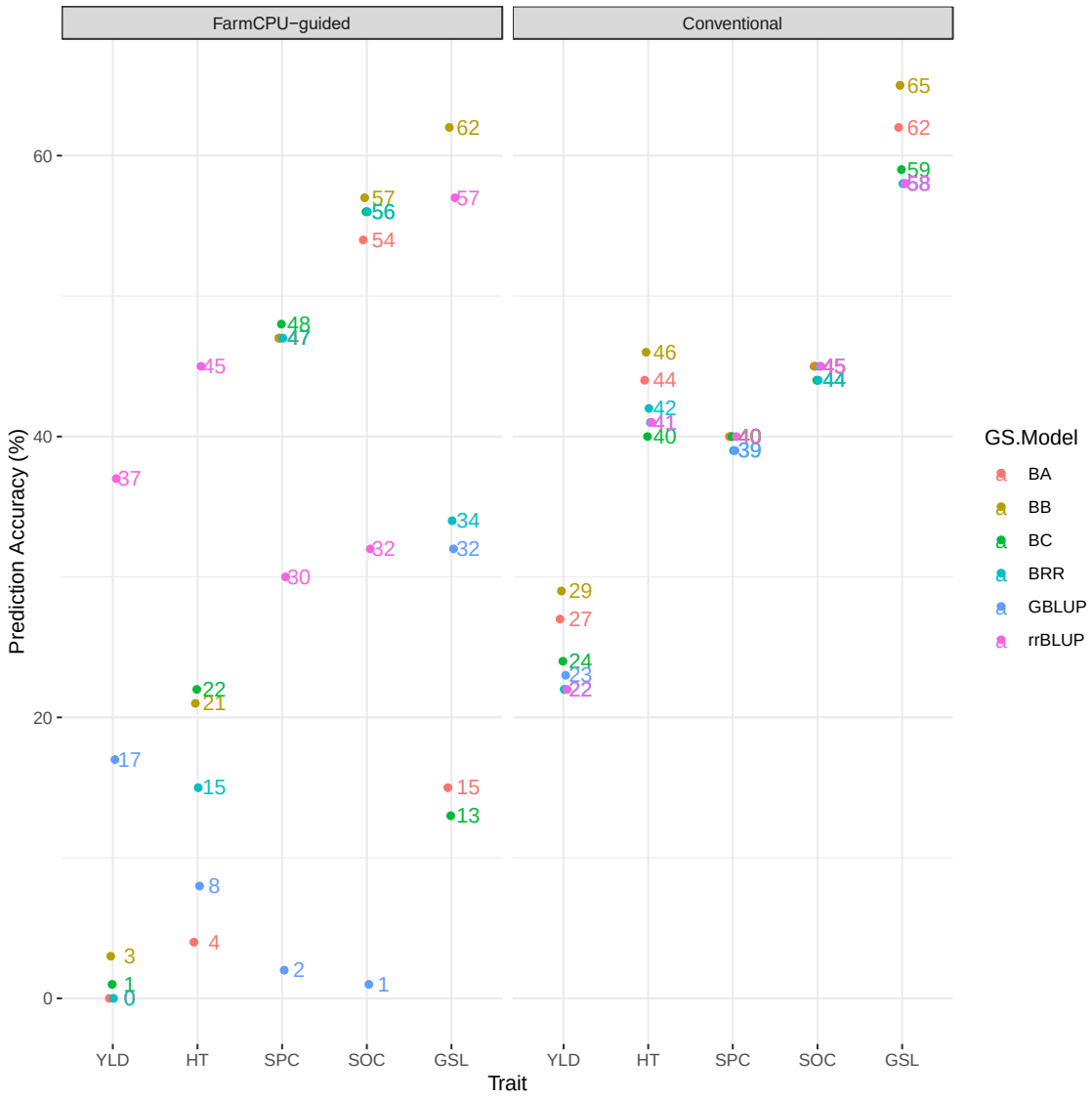


Figure 5.1 Prediction accuracy (Pearson's Correlation (%) between predicted and actual values) based on FarmCPU-guided GS and conventional GS using five-fold cross-validation technique with MS-1 (26,651 SNP markers) based on a combined population of *Brassica napus* L. consisting of 31 B-lines, 60 R-lines and 345 hybrids. Genomic selection models applied included BayesA, BayesB, BayesC, BRR, GBLUP and rrBLUP. The x-axis represents the traits evaluated: YLD (seed yield), HT (plant height), SPC (seed protein content), SOC (seed oil content) and GSL (seed glucosinolate content). The y-axis represents the prediction accuracy. Abbreviations: GS: genomic selection; FarmCPU: Fixed and random model circulating probability unification; BRR: Bayesian ridge regression; rrBLUP: ridge regression best linear unbiased prediction; GBLUP: genomic best linear unbiased prediction.

Conventional GS had small variation, but performed relatively poor on predicting SPC, which varied between 39% to 40%. FarmCPU-guided Bayesian models on SOC had higher prediction accuracies (54% to 57%) compared with conventional GS models (44% to 45%), whereas FarmCPU-guided GBLUP and FarmCPU-guided rrBLUP only had prediction accuracies of 1% and 32%. The prediction accuracies on GSL based on conventional GS models were relatively stable with prediction accuracies between 58% to 65%. FarmCPU-guided GS had similar prediction accuracy based on BayesB and rrBLUP but had significant lower prediction accuracy of GSL based on other GS models (13% to 34%). When comparing FarmCPU-guided GS and conventional GS horizontally, FarmCPU did not improve the prediction accuracy consistently based on MS-1 (Table 5.1). Conventional GS showed strong robustness and mostly performed better than FarmCPU except for SPC and SOC based on Bayesian models.

5.4.2 Genome-wide association-guided genomic selection based on 16,855 SNPs

Overall, FarmCPU-guided GS had better prediction accuracy when using MS-2 (16,855 SNPs) (Figure 5.2). For conventional GS, the difference in prediction accuracy (Pearson's Correlation (%) values between predicted and actual values) was minimal when comparing results based on the two marker sets (Figure 3.1 and Figure 5.2).

For FarmCPU-guided GS models (YLD), the prediction accuracy varied between 41% to 43%, while FarmCPU-guided GBLUP had a relatively low prediction accuracy (26%). Compared with FarmCPU-guided GS models, conventional GS had lower prediction accuracy on YLD, which varied between 19% to 25%, with the lowest produced from BayesC, GBLUP and rrBLUP and the highest produced from BayesB. For HT, FarmCPU-guided Bayesian and FarmCPU-guided rrBLUP prediction accuracies varied between 46% to 57%. FarmCPU-guided GBLUP had a lower

Table 5.1 Prediction accuracy (Pearson's Correlation (%) values between predicted and actual values) difference between FarmCPU-guided GS and conventional GS based on MS-1 (26,651 SNPs). "Difference" was calculated by deducting prediction accuracy of conventional GS models from that of FarmCPU-guided GS models of the same trait. All values were in percentages.

Trait	GS model	FarmCPU-guided GS	Conventional GS	Difference
YLD	BA	0	27	-27
	BB	3	29	-26
	BC	1	24	-23
	BRR	0	22	-22
	GBLUP	17	23	-6
	rrBLUP	37	22	15
HT	BA	4	44	-40
	BB	21	46	-25
	BC	22	40	-18
	BRR	15	42	-27
	GBLUP	8	41	-33
	rrBLUP	45	41	4
SPC	BA	47	40	7
	BB	47	40	7
	BC	48	40	8
	BRR	47	39	8
	GBLUP	2	39	-37
	rrBLUP	30	40	-10
SOC	BA	54	45	9
	BB	57	45	12
	BC	56	44	12
	BRR	56	44	12
	GBLUP	1	45	-44
	rrBLUP	32	45	-13
GSL	BA	15	62	-47
	BB	62	65	-3
	BC	13	59	-46
	BRR	34	58	-24
	GBLUP	32	58	-26
	rrBLUP	57	58	-1

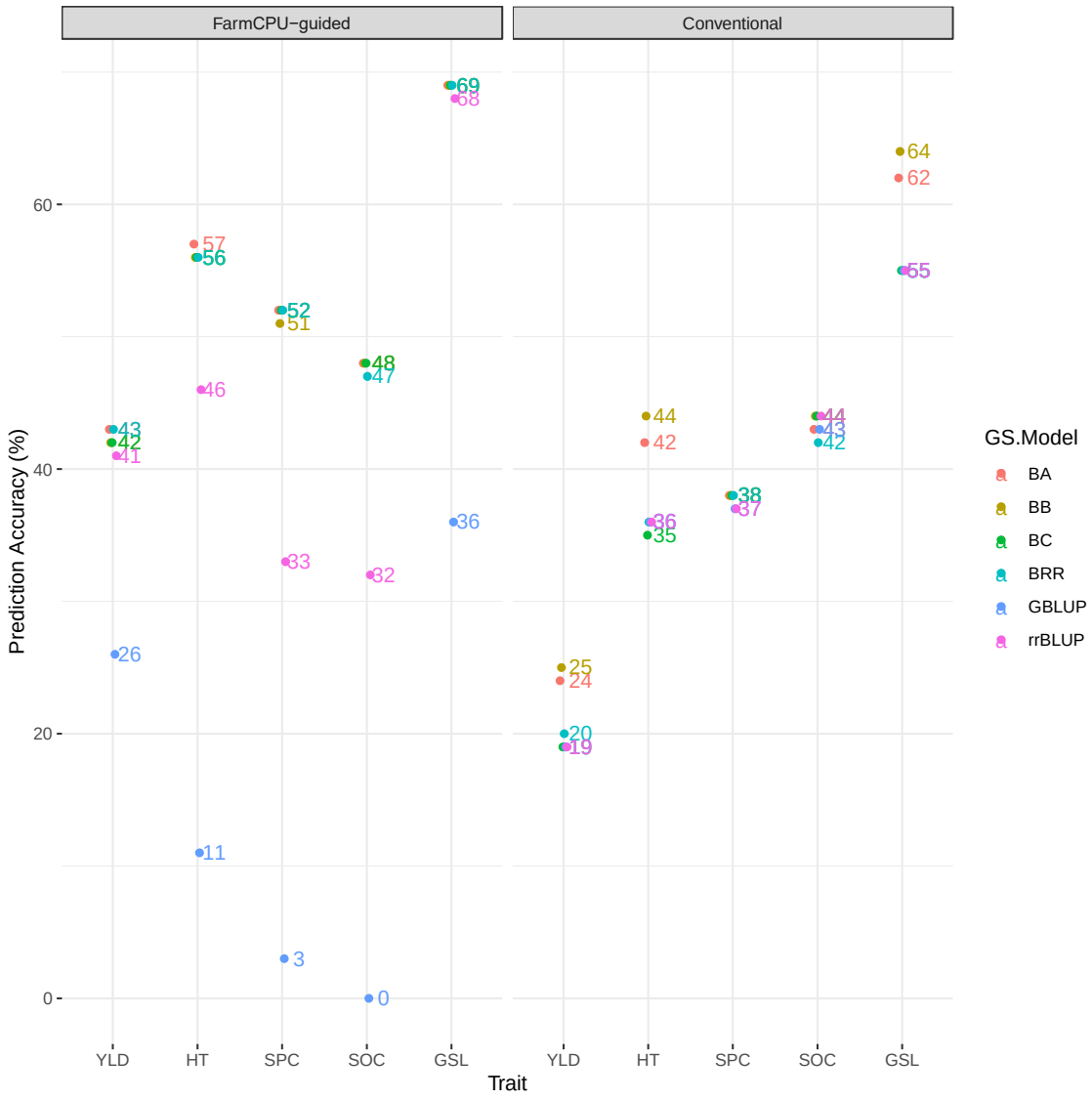


Figure 5.2 Prediction accuracy (Pearson's Correlation (%) values between predicted and actual values) comparison based on FarmCPU-guided GS and conventional GS using five-fold cross-validation technique with MS-2 (16,855 SNP markers) based on a combined population of *Brassica napus* L consisting of 31 B-lines, 60 R-lines and 345 hybrids. Genomic selection models applied include BayesA, BayesB, BayesC, BRR, GBLUP and rrBLUP. The x-axis represents the traits evaluated: YLD (seed yield), HT (plant height), SPC (seed protein content), SOC (seed oil content) and GSL (seed glucosinolate content). The y-axis represents the prediction accuracy. Abbreviations: GS: genomic selection; FarmCPU: Fixed and random model circulating probability unification; BRR: Bayesian ridge regression; rrBLUP: ridge regression best linear unbiased prediction; GBLUP: genomic best linear unbiased prediction.

prediction accuracy (11%). Conventional GS performed relatively uniformly, with prediction accuracies between 35% to 44% with the lowest produced from Bayes C and highest from BayesB. For SPC, FarmCPU-guided Bayesian models also had higher prediction accuracy (51% to 52%) while FarmCPU-guided GBLUP and FarmCPU-guided rrBLUP had lower prediction accuracy (3% and 33%). Similar with HT, conventional GS had lower prediction accuracy (37% to 38%) for SPC. For SOC, Farm-CPU guided Bayesian models had prediction accuracy varied between 47% to 48% while FarmCPU-guided GBLUP and FarmCPU-guided rrBLUP had lower prediction accuracy (0% and 32%). Conventional GS models had similar prediction accuracy which varied between 42% to 44%. Compared with other traits, GSL had very high prediction accuracy based on FarmCPU-guided Bayesian models and FarmCPU-guided rrBLUP (68% to 69%) while a lower prediction accuracy (36%) was obtained by FarmCPU-guided GBLUP. Conventional GS also obtained moderate-high prediction accuracy of 55% to 64% for GSL, with the lowest produced based on BayesC, BRR, GBLUP and rrBLUP and the highest produced based on BayesB.

When comparing FarmCPU-guided GS and conventional GS, FarmCPU consistently improved prediction accuracy based on MS-2 (Table 5.2). More specifically, FarmCPU-guided Bayesian models performed better than conventional GS on all traits. FarmCPU-guided GBLUP and FarmCPU-guided rrBLUP underperformed conventional GBLUP and rrBLUP for YLD, but outperformed for the other four traits.

When comparing the performance of MS-1 and MS-2, fewer markers resulted in an increase in the prediction accuracy of FarmCPU-guided models, except for SOC. For example, prediction accuracy on YLD based on MS-1 varied between 0% to 3% based on FarmCPU-guided Bayesian models, while that of YLD based on MS-2 varied between 42% to 43%. This trend was also observed based on rrBLUP and GBLUP, where the prediction accuracy of FarmCPU-guided

Table 5.2 Prediction accuracy (Pearson's Correlation (%) values between predicted and actual values) difference between FarmCPU-guided GS and conventional GS based on MS-2 (16,855 SNPs). "Difference" was calculated by deducting prediction accuracy of conventional GS models from that of FarmCPU-guided GS models of the same trait. All values were in percentages.

Trait	GS Model	FarmCPU-guided GS	Conventional GS	Difference
YLD	BA	43	24	19
	BB	42	25	17
	BC	42	19	23
	BRR	43	20	23
	GBLUP	26	19	7
	rrBLUP	41	19	22
HT	BA	57	42	15
	BB	56	44	12
	BC	56	35	21
	BRR	56	36	20
	GBLUP	11	36	-25
	rrBLUP	46	36	10
SPC	BA	52	38	14
	BB	51	38	13
	BC	52	38	14
	BRR	52	38	14
	GBLUP	3	37	-34
	rrBLUP	33	37	-4
SOC	BA	48	43	5
	BB	48	44	4
	BC	48	44	4
	BRR	47	42	5
	GBLUP	0	43	-43
	rrBLUP	32	44	-12
GSL	BA	69	62	7
	BB	69	64	5
	BC	69	55	14
	BRR	69	55	14
	GBLUP	36	55	-19
	rrBLUP	68	55	13

GBLUP and FarmCPU-guided rrBLUP based on MS-1 were 17% and 37%, respectively, which increased to 26% and 41% based on MS-2, respectively.

5.4.3 Computational efficiency

With a larger marker set (MS-1), the computation time was significantly higher than MS-2 (Table 5.3). For instance, FarmCPU-guided BayesA based on MS-1 both took 14.8 h, while the conventional BayesA took 16.4 h. In comparison, FarmCPU-guided BayesA based on MS-2 took 9.3 h, while the conventional BayesA took 14.2 h. This was consistent across other Bayesian models, GBLUP and rrBLUP. When comparing the computation time difference based on the same model but different marker sets, larger differences were observed in the FarmCPU-guided GS models than the conventional GS approaches.

Regardless of the choice of marker set, GBLUP and rrBLUP were more efficient than the Bayesian models in terms of computation time. Between GBLUP and rrBLUP, GBLUP was even faster than rrBLUP. Computation time based on FarmCPU-guided GS varied based on the choice of GS models. In general, the difference between these two approaches were not significant.

5.5 Discussion

In the initial GS + *de novo* GWAS research, the authors found that GS combined with the *de novo* GWAS (i.e. GWAS-guided GS) performed better than the standard rrBLUP approach in every case (Spindel et al. 2016). In our research, the performance of GWAS-guided rrBLUP using MS-1 exceeded conventional rrBLUP on YLD and HT by 15% and 4%, respectively (Figure 3.1). Meanwhile, FarmCPU-guided rrBLUP had a lower prediction accuracy on SPC (30%), SOC (32%) and GSL (57%) than conventional rrBLUP (40%, 45% and 58%, respectively) based on MS-1.

Table 5.3 Computation time of conventional GS models and FarmCPU-guided models. The *Brassica napus* L. population consisted of 31 B- lines, 60 R-lines and 345 hybrids. The unit of computation time is hour. Genomic selection models applied include BayesA, BayesB, BayesC, BRR, rrBLUP and GBLUP. Traits evaluated included YLD (seed yield), HT (plant height), SPC (seed protein content), SOC (seed oil content) and GSL (seed glucosinolate content).

Marker Set	GS[1] Model	FarmCPU-guided[2]	Conventional GS
MS[3]-1	BA[4]	14.8	16.4
	BB[5]	17.0	18.5
	BC[6]	15.1	14.9
	BRR[7]	12.8	20.1
	GBLUP[8]	0.8	0.4
	rrBLUP[9]	6.0	3.7
MS-2	BA	9.3	14.2
	BB	10.6	15.6
	BC	9.4	9.6
	BRR	8.2	6.4
	GBLUP	0.2	0.2
	rrBLUP	2.4	3.8

[1] Genomic selection.
[2] FarmCPU: Fixed and random model circulating probability unification.
[3] MS: marker set.
[4] BA: BayesA.
[5] BB: BayesB.
[6] BC: BayesC.
[7] BRR: Bayesian ridge regression.
[8] GBLUP: genomic best linear unbiased prediction.
[9] rrBLUP: ridge regression best linear unbiased prediction.

Based on MS-2, FarmCPU-guided rrBLUP exceeded conventional rrBLUP on YLD, HT and GSL by 22%, 10% and 13%, respectively (Figure 5.2). These results suggested that the prediction performance of GWAS-guided GS was better for some traits when compared with conventional GS. The performance was also impacted by marker density and choice of model (Tan et al. 2017; Zhang et al. 2019a).

As expected, different GS models produced different prediction accuracies even on the same trait. This is due to the fact that the models have different pre-assumptions for the distribution of the variances (Perez and de los Campos 2014). Interestingly, in this research FarmCPU-guided rrBLUP showed improvements in prediction accuracies for YLD and HT, yet, adversely affected the prediction accuracies on all three seed quality traits. This unexpected result suggested that the SNPs fit in the GS models as fixed effects might not link to the true QTL that positively contribute to the trait of interest (Spindel et al. 2016). GWAS-guided Bayesian models in general performed similar to each other on the same trait with some minor variation, which is consistent with results from the previous chapter (section 4.3.3). These similar prediction accuracies again pointed to the complexity of the traits being predicted and the need to analyze data sets of larger size to identify the potential differences (Daetwyler et al. 2013).

There is large variation when comparing prediction accuracies among different traits. For instance, YLD exhibited the largest variation when prediction accuracies were compared across different models and marker sets. As a highly complex trait, YLD can be severely affected by the environment, thus, yield can be difficult to characterize even if the marker set applied had uniform coverage along the genome. Based on FarmCPU-guided GS models, YLD had low to medium prediction accuracy regardless of marker sets or models which varied between 0% to 37% or 26% to 43% based on MS-1 and MS-2, respectively. As Zhang et al. (2014c) stated, genetic architecture

can affect the performance of whole genome prediction (GS) as traits with higher heritability are controlled by fewer loci tend to produce higher prediction accuracies in genomic prediction. Bernardo (2014) also found that fitting major genes as fixed effects increases the prediction accuracy in GS when a quantitative trait is mainly controlled by one to three genes, with each gene accounting for 10% of the genetic variance.

Differences in prediction accuracy were also observed when comparing results from Chapter 4. For YLD, the prediction accuracy was as high as 76% based on MS-2 with a training population size of 262, which decreased to 42% in this research based on conventional BayesB with MS-2. This was possibly caused by the different cross-validation (CV) techniques applied. Previously, to simulate the issues that may be encountered in the prediction process, different CV approaches have been created for GS (Crossa et al. 2017). Haile (2018) identified variation in prediction accuracies due to different CV techniques and recommended choosing a CV technique that mimics the real prediction problem. Similar variation was also observed in other traits such as SPC and SOC based on MS-2. Taken together, the difference observed in prediction accuracy caused by the application of different CV indicated that the CV method used in Chapter 4 was more suitable for this research.

From a cost-efficiency perspective, it is crucial to understand the optimal marker density applied in GS studies, as a lower marker density would be more cost-effective if similar prediction accuracies can be reached compared to higher marker densities. The optimal marker density depends on the nature of the species and traits being investigated as well as the genotyping platform (Kriaridou et al. 2020). In theory, higher marker density would produce higher prediction accuracy since more markers are assumed to be in LD with the loci that control the traits of interest (Desta and Ortiz 2014; Heffner et al. 2009). However, in this research, we found that more markers did

not necessarily improve the prediction accuracy in FarmCPU-guided GS, or in conventional GS. This was consistent with the previous chapter, where the prediction accuracies based on the larger marker set (MS-1, 26,651 SNPs) were lower than the smaller marker set (MS-2, 16,855 SNPs). This possibly resulted from the high relatedness amongst the individuals in the population, which is among the factors that could affect the optimal marker density needed in GS studies (Meuwissen 2009).

Computation efficiency is another factor to consider when conducting GS studies. In general, GWAS-guided GS in this research required similar computation time with conventional GS. FarmCPU-guided Bayesian models required significantly longer than FarmCPU-guided GBLUP and FarmCPU-guided rrBLUP, which is consistent with previous studies (Wang et al. 2015b; Zhang et al. 2014c).

5.6 Conclusion

In conclusion, GWAS-guided GS showed some improvements compared to conventional GS approaches. Using an optimized marker set (MS-2), GWAS-guided GS showed improvements in all traits using Bayesian models. These improvements were model specific and impacted by the marker density. To further develop our understanding GWAS-guided GS prediction accuracy, it is recommended to apply GWAS-guided GS on additional *Brassica napus* populations, but it is clear that prediction performance will differ depending upon the trait of interest.

6 GENERAL DISCUSSION

As an economically important crop, canola (*Brassica napus* L.) is a crucial source of plant-based edible oils (Cartea et al. 2019; Paterson et al. 2001). Canada is the largest canola producer globally (USDA 2020), and canola's annual economic contribution to the Canadian economy has increased 35% in the past ten years to $29.9 billion CAD (LMC International 2020). Today, *B. napus* is also grown to provide raw materials for a wide range of end products such as livestock feed, biofuel, biodegradable plastics and industrial lubricants (Jan et al. 2016; Snowdon et al. 2007). As the global population grows, environmental change and resource shortages increase, and consumer preferences change, breeders are targeting higher yields, stronger disease resistance, and greater abiotic stress tolerance for most crops (Collard and Mackill 2008; Fess et al. 2011). As a result, improving canola yield and yield-related traits will continue to be a major breeding goal for canola breeders. To meet market demand by 2025, Canada's canola production must reach 2,914 kg ha^{-1} to meet the 26 Mt production goal (Canola Council of Canada 2014).

To efficiently improve canola traits related with yield and seed quality, it is crucial to understand the genetic architecture of the traits of interest and the genetic variation in the target population. In Chapter 3, a genome-wide association study (GWAS) was conducted to identify marker-trait associations (MTAs) related to seed yield, plant height, seed protein content, seed oil content and seed glucosinolates content. In this research we examined different factors that could affect the power of GWAS, for example population size and composition. We found that GWAS based on the parental genotypes performed (tested across five site-years) poorly, but with the addition of hybrids (tested across 43 site-years) to the parental genotypes, the performance of GWAS on hybrids was enhanced by not only increasing the population size, but also accounting for the multi-environment effect. This finding was consistent with a previous GWAS study conducted on a

hybrid maize (*Zea mays* L.) population where the authors stated GWAS results from the inbred population cannot be directly applied in understanding the genetic background of a hybrid population (Zhang et al. 2019c). We also compared the marker density effects on GWAS and found that the higher marker density did not necessarily perform better than the lower marker density in this research. In a previous study conducted on *Eucalyptus*, the authors found that the predictive ability based on subsets of SNP markers did not differ from that based on full SNP marker sets. Often, a higher marker density is needed when the linkage disequilibrium decays fast (Kainer et al. 2019), while in our population linkage disequilibrium decayed relatively slow (~ 4.0 Mb based on MS-1 and 5.3 Mb based on MS-2). In terms of model performance, consistent with what was previously characterised by Kaler et al. (2019), complex models such as MLMM, FarmCPU and CMLM performed well in controlling false positives in GWAS. However, there indeed were stratifications that were not controlled well by the complex models since there were deviations observed in the Q-Q plots (Ehret 2010). This research provided a foundation in understanding the population effect of our target populations and the effect of marker density for Chapters 4 and 5.

In addition to understanding the genetics of the traits of interest, appropriate selection models are required. In the Canadian Prairies, more than 95% of all cultivars grown are hybrid cultivars (Morrison et al. 2016). Thus, it is important to identify efficient methodologies for selecting parental combinations that produce the best hybrids. In Chapter 4, we preformed genomic selection (GS) and evaluated the factors that affect the prediction accuracies such as training population, marker density and model performance. We found that prediction accuracies increased when hybrid genotypes were included in the training population, which was consistent with a previous study where the authors stated that simply adding parental genotypes (i.e computing BLUPs based

on combined inbreds and hybirds instead of computing their BLUPs separately before combining) reduced the prediction accuracy in hybrid pearl millet [*Cenchrus americanus* (L.) Morrone.], especially for traits with high heterosis (Liang et al. 2018). We also found that with a larger training population, the prediction accuracies tend to be higher. This is consistent with previous research since a larger training population can offer a wider range of allelic observations for the prediction of small effect QTL on complex quantitative traits (Asoro et al. 2011; Gilmour 2007; Norman et al. 2018; Tayeh et al. 2015; Xu et al. 2018). Similar with the findings from Chapter 3 on marker density, the highest marker density did not produce the highest prediction accuracy. This is possibly caused by the high relatedness amongst the individuals in the population. Previous research has indicated that the required marker density is lower when the relatedness is high (Meuwissen 2009). In terms of variation in the prediction accuracies of different traits, we found that compared with seed yield and plant height, the three seed quality traits tended to have higher prediction accuracies, which is consistent with Knoch et al. (2021). Model performance also affected prediction accuracy in this research. Bayesian models performed quite similar to each other, which indicated a larger population may be needed to reveal the differences in the model performances (Daetwyler et al. 2013). In summary of Chapter 4, many factors affect the prediction accuracy of GS (training population size and composition, marker density and model choice). Therefore, one has to consider all of these factors when developing a GS methodology.

Even though GS selection is regarded as a promising tool, breeders are continuing to improve the prediction accuracy. In an effort to improve the performance of GS, results from GWAS were integrated to GS models as fixed effects in previous studies (Bian and Holland 2017; Fiedler et al. 2017; Spindel et al. 2015; Tsai et al. 2020; Zhang et al. 2014c). In Chapter 5 we followed the approach proposed by Spindel et al. (2016) where a GS + *de novo* GWAS was performed to

examine if the prediction accuracies were improved compared to conventional GS approaches. By comparing the results from Chapter 4 and Chapter 5 we found that different cross validation techniques could affect prediction accuracy, and proper validation techniques need to be applied when performing GS (Haile 2018). In the initial GS + *de novo* GWAS research, Spindel et al. (2016) found that the prediction accuracy based on GS + *de novo* GWAS was consistently higher than that of the conventional rrBLUP. In contrast, the improvements in the performance of GS + *de novo* GWAS compared to conventional rrBLUP were not consistent across traits in our research. Similarly, in a GS study on wheat, the authors found no significant difference on prediction accuracy between GS + *de novo* GWAS and conventional rrBLUP (Haile 2018). We also compared the computation efficiency of different models and identified that GBLUP and rrBLUP (both GWAS-guided and conventional) had the shortest computation duration compared with the Bayesian models. In conclusion, GWAS-guided GS could improve the prediction accuracy compared to conventional GS, but may need more empirical studies to verify its power, since its performance may vary depending on the trait, marker set and the GS model.

Collectively, this research used GWAS in identifying significant SNPs associated with important agronomic and seed quality traits and demonstrated the application of conventional and GWAS-guided GS in hybrid canola development. The methods and results provide valuable information required for implementing GS into canola breeding programs, ultimately advancing the canola industry.

7 FUTURE RESEARCH RECOMMENDATIONS

In this research, we found that genome-wide association study (GWAS) and genomic selection (GS) have strong potential in improving canola breeding efficiencies; however, we suggest the following to improve GS in future breeding efforts.

The research in Chapter 3 would benefit from including a greater number of globally-collected accessions, increasing the genetic diversity of the population, which will also increase the statistical power in detecting significant SNPs associated with the trait of interest. At the same time, traits of interest can be expanded to include flowering, maturity, other seed quality traits and thousand seed weight to identify significant markers.

Regarding Chapters 4 and 5, the performance of GS can be improved by increasing the size of the population, particularly on traits with a lower heritability. Increasing population size will also provide more information to differentiate the performance of different models or markers densities. This will aid in optimizing the performance of GS in its practical application. In addition, high-throughput phenotyping can be implemented in collecting phenotypic data from the field experiments, which could offer more accurate phenotypic data and avoid human error during the phenotyping process. Moreover, covariates can be added to GS models including, but not limited to, annual precipitation, average growing season temperatures of the site-year, soil texture and pest/disease severity. Additional research could also focus on exploring more machine learning methods such multilayer perceptrons (MLPs), recurrent neural networks (RNN) and convolutional neural networks (CNN). Collectively, the above recommendations will facilitate the establishment of GS in the hybrid canola breeding industry in the future.

8. REFERENCE MATTER

8.1 Literature cited

Abdollahi-Arpanahi R, Gianola D, Penagaricano F (2020) Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. Genet Sel Evol 52:12

Abdulmalik RO, Menkir A, Meseka SK, Unachukwu N, Ado SG, Olarewaju JD, Aba DA, Hearne S, Crossa J, Gedil M (2017) Genetic gains in grain yield of a maize population improved through marker assisted recurrent selection under stress and non-stress conditions in west Africa. Front Plant Sci 8:841

Acosta-Pech R, Crossa J, de Los Campos G, Teyssedre S, Claustres B, Perez-Elizalde S, Perez-Rodriguez P (2017) Genomic models with genotype x environment interaction for predicting hybrid performance: an application in maize hybrids. Theor Appl Genet 130:1431-1440

Ahmadi B, Masoomi-Aladizgeh F, Shariatpanahi ME, Azadi P, Keshavarz-Alizadeh M (2016) Molecular characterization and expression analysis of *SERK1* and *SERK2* in *Brassica napus* L.: implication for microspore embryogenesis and plant regeneration. Plant Cell Rep 35:185-193

Ahmar S, Gill RA, Jung KH, Faheem A, Qasim MU, Mubeen M, Zhou W (2020) Conventional and molecular techniques from simple breeding to speed breeding in crop plants: recent advances and future outlook. Int J Mol Sci 21:2590

Alcock TD, Havlickova L, He Z, Bancroft I, White PJ, Broadley MR, Graham NS (2017) Identification of candidate genes for calcium and magnesium accumulation in *Brassica napus* L. by association genetics. Front Plant Sci 8:1968

Ali M, Zhang Y, Rasheed A, Wang J, Zhang L (2020) Genomic prediction for grain yield and yield-related traits in chinese winter wheat. Int J Mol Sci 21:1342

Allender CJ, King GJ (2010) Origins of the amphiploid species *Brassica napus* L. investigated by chloroplast and nuclear molecular markers. BMC Plant Biol 10:1-9

Alseekh S, Kostova D, Bulut M, Fernie AR (2021) Genome-wide association studies: assessing trait characteristics in model and crop plants. Cell Mol Life Sci 78:5743–5754

Anderson JA (2007) Marker-assisted selection for *Fusarium* head blight resistance in wheat. Int J Food Microbiol 119:51-53

Andorf C, Beavis WD, Hufford M, Smith S, Suza WP, Wang K, Woodhouse M, Yu J, Lubberstedt T (2019) Technological advances in maize breeding: past, present and future. Theor Appl Genet 132:817-849

Annicchiarico P, Nazzicari N, Li X, Wei Y, Pecetti L, Brummer EC (2015) Accuracy of genomic selection for alfalfa biomass yield in different reference populations. BMC Genomics 16:1020

Ashkani S, Rafii MY, Shabanimofrad M, Miah G, Sahebi M, Azizi P, Tanweer FA, Akhtar MS, Nasehi A (2015) Molecular breeding strategy and challenges towards improvement of blast disease resistance in rice crop. Front Plant Sci 6:886

Ashraf M, Akram NA, Mehboob-ur-Rahman, Foolad MR (2012) Marker-Assisted Selection in Plant Breeding for Salinity Tolerance. In: Shabala S, Cuin AT (eds) Plant Salt Tolerance: Methods and Protocols. Humana Press, Totowa, NJ, pp 305-333

Asoro FG, Newell MA, Beavis W, Scott M, Jannink J (2011) Accuracy and training population design for genomic selection on quantitative traits in elite North American oats. Plant Genome 4:132-144

Assefa T, Otyama PI, Brown AV, Kalberer SR, Kulkarni RS, Cannon SB (2019) Genome-wide associations and epistatic interactions for internode number, plant height, seed weight and seed yield in soybean. BMC Genomics 20:527

Assefa Y, Prasad PVV, Foster C, Wright Y, Young S, Bradley P, Stamm M, Ciampitti IA (2018) Major management factors determining spring and winter canola yield in North America. Crop Sci 58:1-16

Atwell S, Huang YS, Vilhjalmsson BJ, Willems G, Horton M, Li Y, Meng D, Platt A, Tarone AM, Hu TT, Jiang R, Muliyati NW, Zhang X, Amer MA, Baxter I, Brachi B, Chory J, Dean C, Debieu M, de Meaux J, Ecker JR, Faure N, Kniskern JM, Jones JD, Michael T, Nemri A, Roux F, Salt DE, Tang C, Todesco M, Traw MB, Weigel D, Marjoram P, Borevitz JO, Bergelson J, Nordborg M (2010) Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. Nature 465:627-631

Badu-Apraku B, Talabi AO, Fakorede MAB, Fasanmade Y, Gedil M, Magorokosho C, Asiedu R (2019) Yield gains and associated changes in an early yellow bi-parental maize population following genomic selection for Striga resistance and drought tolerance. BMC Plant Biol 19:129

Bajaj D, Upadhyaya HD, Das S, Kumar V, Gowda CL, Sharma S, Tyagi AK, Parida SK (2016) Identification of candidate genes for dissecting complex branch number trait in chickpea. Plant Sci 245:61-70

Ballesta P, Bush D, Silva FF, Mora F (2020) Genomic predictions using low-density snp markers, pedigree and gwas information: a case study with the non-model species *Eucalyptus cladocalyx*. Plants (Basel) 9:99

Bartrina I, Otto E, Strnad M, Werner T, Schmulling T (2011) Cytokinin regulates the activity of reproductive meristems, flower organ size, ovule formation, and thus seed yield in *Arabidopsis thaliana*. Plant Cell 23:69-80

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

Basu U, Srivastava R, Bajaj D, Thakro V, Daware A, Malik N, Upadhyaya HD, Parida SK (2018) Genome-wide generation and genotyping of informative SNPs to scan molecular signatures for seed yield in chickpea. Sci Rep 8:13240

Battenfield SD, Guzman C, Gaynor RC, Singh RP, Pena RJ, Dreisigacker S, Fritz AK, Poland JA (2016) Genomic selection for processing and end-use quality traits in the CIMMYT spring bread wheat breeding program. Plant Genome 9:1-12

Bayer PE, Golicz AA, Scheben A, Batley J, Edwards D (2020) Plant pan-genomes are the new reference. Nat Plants 6:914-920

Bayer PE, Hurgobin B, Golicz AA, Chan CK, Yuan Y, Lee H, Renton M, Meng J, Li R, Long Y, Zou J, Bancroft I, Chalhoub B, King GJ, Batley J, Edwards D (2017) Assembly and comparison of two closely related *Brassica napus* genomes. Plant Biotechnol J 15:1602-1610

Bell JM (1982) From rapeseed to canola: a brief history of research for superior meal and edible oil. Poultry Science 61:613-622

Bellucci A, Tondelli A, Fangel JU, Torp AM, Xu X, Willats WG, Flavell A, Cattivelli L, Rasmussen SK (2017) Genome-wide association mapping in winter barley for grain yield and culm cell wall polymer content using the high-throughput CoMPP technique. PLoS One 12:e0173313

Bentley AR, Scutari M, Gosman N, Faure S, Bedford F, Howell P, Cockram J, Rose GA, Barber T, Irigoyen J, Horsnell R, Pumfrey C, Winnie E, Schacht J, Beauchene K, Praud S, Greenland A, Balding D, Mackay IJ (2014) Applying association mapping and genomic selection to the dissection of key traits in elite European wheat. Theor Appl Genet 127:2619-2633

Bernardo R (2014) Genomewide selection when major genes are known. Crop Sci 54:68-75

Bernardo R (2016) Bandwagons I, too, have known. Theor Appl Genet 129:2323-2332

Beyene Y, Gowda M, Olsen M, Robbins KR, Perez-Rodriguez P, Alvarado G, Dreher K, Gao SY, Mugo S, Prasanna BM, Crossa J (2019) Empirical comparison of tropical maize hybrids selected through genomic and phenotypic selections. Front Plant Sci 10:1502

Beyene Y, Semagn K, Mugo S, Prasanna BM, Tarekegne A, Gakunga J, Sehabiague P, Meisel B, Oikeh SO, Olsen M, Crossa J (2016) Performance and grain yield stability of maize populations developed using marker-assisted recurrent selection and pedigree selection procedures. Euphytica 208:285-297

Bhering L, Junqueira V, Peixoto L, Cruz C, Laviola B (2015) Comparison of methods used to identify superior individuals in genomic selection in plant breeding. Genet Mol Res 14:10888-10896

Bian Y, Holland JB (2017) Enhancing genomic prediction with genome-wide association studies in multiparental maize populations. Heredity (Edinb) 118:585-593

Boopathi NM (2013) Marker-Assisted Selection. Genetic Mapping and Marker Assisted Selection: Basics, Practice and Benefits. Springer India, India, pp 173-186

Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32:314-331

Brachi B, Morris GP, Borevitz JO (2011) Genome-wide association studies in plants: the missing heritability is in the field. Genome Biol 12:232

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633-2635

Breiman L (2001) Random Forests. Machine Learning 45:5-32

Brown LR (2012) Grain yields starting to plateau. Full Planet, Empty Plates: The New Geopolitics of Food Scarcity, 1 edn. Earth Policy Institute, New York, pp 72-82

Bumgarner R (2013) Overview of DNA microarrays: types, applications, and their future. Curr Protoc Mol Biol 101:22.21.21-22.21.11

Bus A, Korber N, Snowdon RJ, Stich B (2011) Patterns of molecular variation in a species-wide germplasm set of *Brassica napus*. Theor Appl Genet 123:1413-1423

Calus MP (2010) Genomic breeding value prediction: methods and procedures. Animal 4:157-164

Calus MP, Meuwissen TH, de Roos AP, Veerkamp RF (2008) Accuracy of genomic selection using different methods to define haplotypes. Genetics 178:553-561

Canola Council of Canada (2013) GDDs to date.

Canola Council of Canada (2014) Canada's canola industry sets bold new target for 2025.

Canola Council of Canada (2016) Canadian Canola Harvested Acreage.

Canola Council of Canada (2017) Markets of Canadian canola.

Canola council of Canada (2018) Canadian canola yield (tonnes/acre).

Canola council of Canada (2019) Canadian canola seed exports.

Canola Council of Canada (2020) How much fertilizer does canola need?

Canola Council of Canada (2021a) Growth stages of the canola plant.

Canola Council of Canada (2021b) History of canola seed development.

Canvin DT (1965) The effect of temperature on the oil content and fatty acid composition of the oils from several oil seed crops. Can J Bot 43:63-69

Cao JY, Xu YP, Li W, Li SS, Rahman H, Cai XZ (2016) Genome-Wide identification of dicer-like, argonaute, and rna-dependent rna polymerase gene families in *Brassica Species* and functional analyses of their *Arabidopsis* homologs in resistance to *Sclerotinia sclerotiorum*. Front Plant Sci 7:1614

Carlson MO, Montilla-Bascon G, Hoekenga OA, Tinker NA, Poland J, Baseggio M, Sorrells ME, Jannink JL, Gore MA, Yeats TH (2019) Multivariate genome-wide association analyses reveal the genetic basis of seed fatty acid composition in oat (*Avena sativa* L.). G3 (Bethesda) 9:2963-2975

Cartea E, De Haro-Bailon A, Padilla G, Obregon-Cano S, Del Rio-Celestino M, Ordas A (2019) Seed oil quality of *Brassica napus* and *Brassica rapa* germplasm from Northwestern Spain. Foods 8:292

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

Cavanagh CR, Chao S, Wang S, Huang BE, Stephen S, Kiani S, Forrest K, Saintenac C, Brown-Guedira GL, Akhunova A, See D, Bai G, Pumphrey M, Tomar L, Wong D, Kong S, Reynolds M, da Silva ML, Bockelman H, Talbert L, Anderson JA, Dreisigacker S, Baenziger S, Carter A, Korzun V, Morrell PL, Dubcovsky J, Morell MK, Sorrells ME, Hayden MJ, Akhunov E (2013) Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. Proc Natl Acad Sci U S A 110:8057-8062

Celton JM, Christoffels A, Sargent DJ, Xu X, Rees DJ (2010) Genome-wide SNP identification by high-throughput sequencing and selective mapping allows sequence assembly positioning using a framework genetic linkage map. BMC Biol 8:155

Chalhoub B, Denoeud F, Liu S, Parkin IA, Tang H, Wang X, Chiquet J, Belcram H, Tong C, Samans B, Correa M, Da Silva C, Just J, Falentin C, Koh CS, Le Clainche I, Bernard M, Bento P, Noel B, Labadie K, Alberti A, Charles M, Arnaud D, Guo H, Daviaud C, Alamery S, Jabbari K, Zhao M, Edger PP, Chelaifa H, Tack D, Lassalle G, Mestiri I, Schnel N, Le Paslier MC, Fan G, Renault V, Bayer PE, Golicz AA, Manoli S, Lee TH, Thi VH, Chalabi S, Hu Q, Fan C, Tollenaere R, Lu Y, Battail C, Shen J, Sidebottom CH, Wang X, Canaguier A, Chauveau A, Berard A, Deniot G, Guan M, Liu Z, Sun F, Lim YP, Lyons E, Town CD, Bancroft I, Wang X, Meng J, Ma J, Pires JC, King GJ, Brunel D, Delourme R, Renard M, Aury JM, Adams KL, Batley J, Snowdon RJ, Tost J, Edwards D, Zhou Y, Hua W, Sharpe AG, Paterson AH, Guan C, Wincker P (2014) Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. Science 345:950-953

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4:7

Chang LY, Toghiani S, Ling A, Aggrey SE, Rekaya R (2018) High density marker panels, SNPs prioritizing and accuracy of genomic selection. BMC Genet 19:4

Chang T, Wei J, Liang M, An B, Wang X, Zhu B, Xu L, Zhang L, Gao X, Chen Y, Li J, Gao H (2019) A fast and powerful empirical Bayes method for genome-wide association studies. Animals (Basel) 9

Chao H, Wang H, Wang X, Guo L, Gu J, Zhao W, Li B, Chen D, Raboanatahiry N, Li M (2017) Genetic dissection of seed oil and protein content and identification of networks associated with oil content in *Brassica napus*. Sci Rep 7:46295

Chen CJ, Zhang Z (2018) iPat: intelligent prediction and association tool for genomic research. Bioinformatics 34:1925-1927

Chen G, Geng J, Rahman M, Liu X, Tu J, Fu T, Li G, McVetty PBE, Tahir M (2010) Identification of QTL for oil content, seed yield, and flowering time in oilseed rape (*Brassica napus*). Euphytica 175:161-174

Chen G, Wang X, Hao J, Yan J, Ding J (2015) Genome-wide association implicates candidate genes conferring resistance to maize rough dwarf disease in maize. PLoS One 10:e0142001

Chen L, Wan H, Qian J, Guo J, Sun C, Wen J, Yi B, Ma C, Tu J, Song L, Fu T, Shen J (2018) Genome-wide association study of cadmium accumulation at the seedling stage in rapeseed (*Brassica napus* L.). Front Plant Sci 9:375

Chen L, Zhao J, Song J, Jameson PE (2020) Cytokinin dehydrogenase: a genetic target for yield improvement in wheat. Plant Biotechnol J 18:614-630

Chen T, Guestrin C (2016) XGBoost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp 785-794

Chen W, Zhang Y, Liu X, Chen B, Tu J, Fu T (2007) Detection of QTL for six yield-related traits in oilseed rape (*Brassica napus*) using DH and immortalized F2 populations. Theor Appl Genet 115:849-858

Cheung WY, Gugel RK, Landry BS (1998) Identification of RFLP markers linked to the white rust resistance gene (*Acr*) in mustard (*Brassica juncea* (L.) Czern. and Coss.). Genome 41:626-628

Clark AJ, Sarti-Dvorjak D, Brown-Guedira G, Dong Y, Baik BK, Van Sanford DA (2016) Identifying rare FHB-resistant segregants in intransigent backcross and F2 winter wheat populations. Front Microbiol 7:277

Clark SA, Hickey JM, van der Werf JH (2011) Different models of genetic variation and their effect on genomic evaluation. Genet Sel Evol 43:18

Clark SA, van der Werf J (2013) Genomic best linear unbiased prediction (gBLUP) for the estimation of genomic breeding values. In: Gondro C, Werf Jvd, Hayes B (eds) Genome-Wide Association Studies and Genomic Prediction, pp 321-330

Clarke WE, Higgins EE, Plieske J, Wieseke R, Sidebottom C, Khedikar Y, Batley J, Edwards D, Meng J, Li R, Lawley CT, Pauquet J, Laga B, Cheung W, Iniguez-Luy F, Dyrszka E, Rae S, Stich B, Snowdon RJ, Sharpe AG, Ganal MW, Parkin IA (2016) A high-density SNP genotyping array for *Brassica napus* and its ancestral diploid species based on optimised selection of single-locus markers in the allotetraploid genome. Theor Appl Genet 129:1887-1899

Cobb JN, Biswas PS, Platten JD (2019) Back to the future: revisiting MAS as a tool for modern plant breeding. Theor Appl Genet 132:647-667

Collard BC, Mackill DJ (2008) Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Philosophical Transactions of the Royal Society B 363:557-572

Collard BCY, Jahufer MZZ, Brouwer JB, Pang ECK (2005) An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. Euphytica 142:169-196

Colombani C, Legarra A, Fritz S, Guillaume F, Croiseau P, Ducrocq V, Robert-Granie C (2013) Application of Bayesian least absolute shrinkage and selection operator (LASSO) and BayesCpi methods for genomic selection in French Holstein and Montbeliarde breeds. J Dairy Sci 96:575-591

Combs E, Bernardo R (2013) Accuracy of genomewide selection for different traits with constant population size, heritability, and number of markers. Plant Genome 6:1-7

Cortes LT, Zhang Z, Yu J (2021) Status and prospects of genome-wide association studies in plants. Plant Genome 14:e20077

Cros D, Denis M, Sanchez L, Cochard B, Flori A, Durand-Gasselin T, Nouy B, Omore A, Pomies V, Riou V, Suryana E, Bouvet JM (2015) Genomic selection prediction accuracy in a perennial crop: case study of oil palm (*Elaeis guineensis* Jacq.). Theor Appl Genet 128:397-410

Crossa J, Campos Gde L, Perez P, Gianola D, Burgueno J, Araus JL, Makumbi D, Singh RP, Dreisigacker S, Yan J, Arief V, Banziger M, Braun HJ (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186:713-724

Crossa J, Perez P, Hickey J, Burgueno J, Ornella L, Ceron-Rojas J, Zhang X, Dreisigacker S, Babu R, Li Y, Bonnett D, Mathews K (2014) Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity (Edinb) 112:48-60

Crossa J, Pérez-Rodríguez P, Cuevas J, Montesinos-López O, Jarquín D, de los Campos G, Burgueño J, González-Camacho JM, Pérez-Elizalde S, Beyene Y, Dreisigacker S, Singh R, Zhang X, Gowda M, Roorkiwal M, Rutkoski J, Varshney RK (2017) Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. Trends Plant Sci 22:961-975

Cuevas J, Crossa J, Montesinos-Lopez OA, Burgueno J, Perez-Rodriguez P, de Los Campos G (2017) Bayesian genomic prediction with genotype x environment interaction kernel models. G3 (Bethesda) 7:41-53

Cui Z, Dong H, Zhang A, Ruan Y, Jiang S, He Y, Zhang Z (2020) Denser markers and advanced statistical method identified more genetic loci associated with husk traits in maize. Sci Rep 10:8165

Cui Z, Luo J, Qi C, Ruan Y, Li J, Zhang A, Yang X, He Y (2016) Genome-wide association study (GWAS) reveals the genetic architecture of four husk traits in maize. BMC Genomics 17:946

Cuong DM, Park SU, Park CH, Kim NS, Bong SJ, Lee SY (2019) Comparative analysis of glucosinolate production in hairy roots of green and red kale (*Brassica oleracea* var. *acephala*). Prep Biochem Biotechnol 49:775-782

Daetwyler HD, Bansal UK, Bariana HS, Hayden MJ, Hayes BJ (2014) Genomic prediction for rust resistance in diverse wheat landraces. Theor Appl Genet 127:1795-1803

Daetwyler HD, Calus MP, Pong-Wong R, de Los Campos G, Hickey JM (2013) Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. Genetics 193:347-365

Dar AA, Choudhury AR, Kancharla PK, Arumugam N (2017) The *FAD2* Gene in Plants: Occurrence, Regulation, and Role. Front Plant Sci 8:1789

Das S, Hussain A, Bock C, Keller WA, Georges F (2005) Cloning of *Brassica napus* phospholipase C2 (BnPLC2), phosphatidylinositol 3-kinase (BnVPS34) and phosphatidylinositol synthase1 (BnPtdIns S1)--comparative analysis of the effect of abiotic stresses on the expression of phosphatidylinositol signal transduction-related genes in *B. napus*. Planta 220:777-784

Daun JK (2011) Origin, distribution, and production. In: Daun JK, Eskin NAM, Hickling D (eds) Canola: chemistry, production, processing, and utilization. AOCS Press, Champaign, pp 1-27

de Los Campos G, Gianola D, Rosa GJ (2009) Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. J Anim Sci 87:1883-1887

de Los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MP (2013) Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193:327-345

de Los Campos G, Sorensen D, Gianola D (2015) Genomic heritability: what is it? PLoS Genet 11:e1005048

DeClercq DR, Daun JK, Tipples K (1998) Quality of western Canadian canola. Canadian Grain Commission

Delourme R, Falentin C, Huteau V, Clouet V, Horvais R, Gandon B, Specel S, Hanneton L, Dheu JE, Deschamps M, Margale E, Vincourt P, Renard M (2006) Genetic control of oil content in oilseed rape (*Brassica napus* L.). Theor Appl Genet 113:1331-1345

Deng M, Li D, Luo J, Xiao Y, Liu H, Pan Q, Zhang X, Jin M, Zhao M, Yan J (2017) The genetic architecture of amino acids dissection by association and linkage analysis in maize. Plant Biotechnol J 15:1250-1263

Derscheid LA, Lytle WF (1977) Growing degree days (GDD)

Desta ZA, Ortiz R (2014) Genomic selection: genome-wide prediction in plant improvement. Trends Plant Sci 19:592-601

Di F, Jian H, Wang T, Chen X, Ding Y, Du H, Lu K, Li J, Liu L (2018) Genome-Wide analysis of the *PYL* gene family and identification of *PYL* genes that respond to abiotic stress in *Brassica napus*. Genes (Basel) 9:156

Dias K, Gezan SA, Guimaraes CT, Nazarian A, da Costa ESL, Parentoni SN, de Oliveira Guimaraes PE, de Oliveira Anoni C, Padua JMV, de Oliveira Pinto M, Noda RW, Ribeiro CAG, de Magalhaes JV, Garcia AAF, de Souza JC, Guimaraes LJM, Pastina MM (2018) Improving accuracies of genomic predictions for drought tolerance in maize by joint modeling of additive and dominance effects in multi-environment trials. Heredity (Edinb) 121:24-37

Downey RK, Harvey BL (1963) Methods of breeding for oil quality in rape. Can J Plant Sci 43:271-275

Drucker H, Burges C, Kaufman L, Smola A, Vapnik V (1997) Support vector regression machines. Adv Neural Inf Process Syst 9:155-161

Duan S, Wang J, Gao C, Jin C, Li D, Peng D, Du G, Li Y, Chen M (2018) Functional characterization of a heterologously expressed *Brassica napus WRKY41-1* transcription factor in regulating anthocyanin biosynthesis in *Arabidopsis thaliana*. Plant Sci 268:47-53

Earl DA (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation Genetics Resources 4:359-361

Edwards D, Batley J, Snowdon RJ (2013) Accessing complex crop genomes with next-generation sequencing. Theor Appl Genet 126:1-11

Ehret GB (2010) Genome-wide association studies: contribution of genomics to understanding blood pressure and essential hypertension. Curr Hypertens Rep 12:17-25

Elahi N, Duncan RW, Stasolla C (2016) Modification of oil and glucosinolate content in canola seeds with altered expression of *Brassica napus LEAFY COTYLEDON1*. Plant Physiol Biochem 100:52-63

Elbasyoni IS, Lorenz AJ, Guttieri M, Frels K, Baenziger PS, Poland J, Akhunov E (2018) A comparison between genotyping-by-sequencing and array-based scoring of SNPs for genomic prediction accuracy in winter wheat. Plant Sci 270:123-130

Elhiti M, Yang C, Chan A, Durnin DC, Belmonte MF, Ayele BT, Tahir M, Stasolla C (2012) Altered seed oil and glucosinolate levels in transgenic plants overexpressing the *Brassica napus SHOOTMERISTEMLESS* gene. J Exp Bot 63:4447-4461

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A Robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLOS ONE 6:e19379

Endelman JB (2011) Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome 4:250-255

Eskin NM, McDonald B (1991) Canola oil. Nutrition Bulletin 16:138-146

Falk K (2009) Heterosis assessment for agronomic and seed quality traits in hybrid canola (*Brassica napus* L.). Plant Sci. University of Manitoba, Winnipeg, p 132

FAO (2021) Crops and livestock products. In: Food and Agriculture Organization of the United Nations (FAO) (ed)

Fedoroff NV (2010) The past, present and future of crop genetic modification. N Biotechnol 27:461-465

Fess TL, Kotcon JB, Benedito VA (2011) Crop breeding for low input agriculture: a sustainable response to feed a growing world population. Sustainability 3:1742

Fiedler JD, Salsman E, Liu Y, Michalak de Jimenez M, Hegstad JB, Chen B, Manthey FA, Chao S, Xu S, Elias EM, Li X (2017) Genome-wide association and prediction of grain and semolina quality traits in durum wheat breeding populations. Plant Genome 10:1-12

Fikere M, Barbulescu DM, Malmberg MM, Shi F, Koh JCO, Slater AT, MacLeod IM, Bowman PJ, Salisbury PA, Spangenberg GC, Cogan NOI, Daetwyler HD (2018) Genomic prediction using prior quantitative trait loci information reveals a large reservoir of underutilised blackleg resistance in diverse canola (*Brassica napus* L.) lines. Plant Genome 11:1-16

Fleury D, Jefferies S, Kuchel H, Langridge P (2010) Genetic and genomic tools to improve drought tolerance in wheat. J Exp Bot 61:3211-3222

Francia E, Tacconi G, Crosatti C, Barabaschi D, Bulgarelli D, Dall'Aglio E, Valè G (2005) Marker assisted selection in crop plants. Plant Cell Tiss Org Cult 82:317-342

Fredua-Agyeman R, Yu Z, Hwang S-F, Strelkov SE (2020) Genome-wide mapping of loci associated with resistance to clubroot in *Brassica napus* ssp. *napobrassica* (rutabaga) accessions from Nordic countries. Front Plant Sci 11:742

Friedman JH (2001) Greedy function approximation: A gradient boosting machine. Ann Stat 29:1189-1232, 1144

Fu D, Mason AS, Xiao M, Yan H (2016) Effects of genome structure variation, homeologous genes and repetitive DNA on polyploid crop research in the age of genomics. Plant Sci 242:37-46

Fu Y, Lu K, Qian L, Mei J, Wei D, Peng X, Xu X, Li J, Frauen M, Dreyer F, Snowdon RJ, Qian W (2015) Development of genic cleavage markers in association with seed glucosinolate content in canola. Theor Appl Genet 128:1029-1037

Fu YB, Yang MH, Zeng F, Biligetu B (2017) Searching for an accurate marker-based prediction of an individual quantitative trait in molecular plant breeding. Front Plant Sci 8:1182

Gacek K, Bayer PE, Bartkowiak-Broda I, Szala L, Bocianowski J, Edwards D, Batley J (2016) Genome-wide association study of genetic control of seed fatty acid biosynthesis in *Brassica napus*. Front Plant Sci 7:2062

Gapare W, Liu S, Conaty W, Zhu QH, Gillespie V, Llewellyn D, Stiller W, Wilson I (2018) Historical datasets support genomic selection models for the prediction of cotton fiber quality phenotypes across multiple environments. G3 (Bethesda) 8:1721-1732

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

Gibbon BC, Larkins BA (2005) Molecular genetic approaches to developing quality protein maize. Trends Genet 21:227-233

Gilmour AR (2007) Mixed model regression mapping for QTL detection in experimental crosses. Comput Stat Data Anal 51:3749-3764

Goddard ME, Hayes BJ (2007) Genomic selection. J Anim Breed Genet 124:323-330

Goiffon M, Kusmec A, Wang L, Hu G, Schnable PS (2017) Improving Response in Genomic Selection with a Population-Based Selection Strategy: Optimal Population Value Selection. Genetics 206:1675-1682

Gotz S, Garcia-Gomez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talon M, Dopazo J, Conesa A (2008) High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res 36:3420-3435

Grassini P, Eskridge KM, Cassman KG (2013) Distinguishing between yield advances and yield plateaus in historical crop production trends. Nat Commun 4:2918

Grenier C, Cao TV, Ospina Y, Quintero C, Chatel MH, Tohme J, Courtois B, Ahmadi N (2015) Accuracy of genomic selection in a rice synthetic population developed for recurrent selection breeding. PLoS One 10:e0136594

Guo T, Yu X, Li X, Zhang H, Zhu C, Flint-Garcia S, McMullen MD, Holland JB, Szalma SJ, Wisser RJ, Yu J (2019) Optimal designs for genomic selection in hybrid crops. Mol Plant 12:390-401

Guo Z, Tucker DM, Basten CJ, Gandhi H, Ersoz E, Guo B, Xu Z, Wang D, Gay G (2014) The impact of population structure on genomic prediction in stratified populations. Theor Appl Genet 127:749-762

Gupta PK, Kulwal PL, Jaiswal V (2014) Association mapping in crop plants: opportunities and challenges. Adv Genet 85:109-147

Gyawali A, Shrestha V, Guill KE, Flint-Garcia S, Beissinger TM (2019) Single-plant GWAS coupled with bulk segregant analysis allows rapid identification and corroboration of plant-height candidate SNPs. BMC Plant Biol 19:412

Habibur R, A. BR, Ginette S-S (2015) Broadening genetic diversity in *Brassica napus* canola: Development of canola-quality spring *B. napus* from *B. napus*×*B. oleracea* var. *alboglabra* interspecific crosses. Can J Plant Sci 95:29-41

Habier D, Fernando RL, Dekkers JC (2007) The impact of genetic relationship information on genome-assisted breeding values. Genetics 177:2389-2397

Habier D, Fernando RL, Kizilkaya K, Garrick DJ (2011a) Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics 12:186

Habier D, Fernando RL, Kizilkaya K, Garrick DJ (2011b) Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics 12:186

Haile TA (2018) Genomic selection, quantitative trait loci and genome-wide association mapping for spring bread wheat (*Triticum aestivum* L.) improvement. University of Saskatchewan

Harker KN, O'Donovan JT, Turkington TK, Blackshaw RE, Lupwayi NZ, Smith EG, Klein-Gebbinck H, Dosdall LM, Hall LM, Willenborg CJ (2012) High-yield no-till canola production on the Canadian Prairies. Can J Plant Sci 92:221-233

Harker KN, O'Donovan JT, Turkington TK, Blackshaw RE, Lupwayi NZ, Smith EG, Johnson EN, Gan Y, Kutcher HR, Dosdall LM, Peng G (2015) Canola rotation frequency impacts canola yield and associated pest species. Can J Plant Sci 95:9-20

Harper FR, Berkenkamp B (1975) Revised growth-stage key for *Brassica campestris* and *B. napus*. Can J Plant Sci 55:657-658

Harvey BL, Downey RK (1964) The inheritance of erucic acid content in rapeseed (*Brassica napus*). Can J Plant Sci 44:104-111

Hatzig SV, Frisch M, Breuer F, Nesi N, Ducournau S, Wagner MH, Leckband G, Abbadi A, Snowdon RJ (2015) Genome-wide association mapping unravels the genetic control of seed germination and vigor in *Brassica napus*. Front Plant Sci 6:221

Hayward A (2011) Introduction: Oilseed Brassicas. Genetics, Genomics and Breeding of Oilseed Brassicas. Science Publishers, pp 1-13

He L, Xiao J, Rashid KY, Yao Z, Li P, Jia G, Wang X, Cloutier S, You FM (2018) Genome-wide association studies for pasmo resistance in flax (*Linum usitatissimum* L.). Front Plant Sci 9:1982

He S, Schulthess AW, Mirdita V, Zhao Y, Korzun V, Bothe R, Ebmeyer E, Reif JC, Jiang Y (2016) Genomic selection in a commercial winter wheat population. Theor Appl Genet 129:641-651

He Y, Wu D, Wei D, Fu Y, Cui Y, Dong H, Tan C, Qian W (2017) GWAS, QTL mapping and gene expression analyses in *Brassica napus* reveal genetic control of branching morphogenesis. Sci Rep 7:15971

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

Heffner EL, Jannink JL, Sorrells ME (2011b) Genomic selection accuracy using multifamily prediction models in a wheat breeding program. Plant Genome 4:65-75

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

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

Heslot N, Yang HP, Sorrells ME, Jannink JL (2012) Genomic selection in plant breeding: a comparison of models. Crop Sci 52:146-160

Hickey JM, Dreisigacker S, Crossa J, Hearne S, Babu R, Prasanna BM, Grondona M, Zambelli A, Windhausen VS, Mathews K, Gorjanc G (2014) Evaluation of genomic selection training population designs and genotyping strategies in plant breeding programs using simulation. Crop Sci 54:1476-1488

Hickey LT, Hafeez AN, Robinson H, Jackson SA, Leal-Bertioli SCM, Tester M, Gao C, Godwin ID, Hayes BJ, Wulff BBH (2019) Breeding crops to feed 10 billion. Nat Biotechnol 37:744-754

Hirani AH, Gao F, Liu J, Fu G, Wu C, Yuan Y, Li W, Hou J, Duncan R, Li G (2016) Transferring clubroot resistance from Chinese cabbage (*Brassica rapa*) to canola (*B. napus*). Can J Plant Pathol 38:82-90

Hoffstetter A, Cabrera A, Huang M, Sneller C (2016) Optimizing training population data and validation of genomic selection for economic traits in soft winter wheat. G3 (Bethesda) 6:2919-2928

Holliday JA, Wang T, Aitken S (2012) Predicting adaptive phenotypes from multilocus genotypes in sitka spruce (*Picea sitchensis*) using random forest. G3 (Bethesda) 2:1085-1093

Honsdorf N, Becker HC, Ecke W (2010) Association mapping for phenological, morphological, and quality traits in canola quality winter rapeseed (*Brassica napus* L.). Genome 53:899-907

Hossain F, Muthusamy V, Pandey N, Vishwakarma AK, Baveja A, Zunjare RU, Thirunavukkarasu N, Saha S, Manjaiah KMM, Prasanna BM, Gupta HS (2018) Marker-assisted introgression of *opaque2* allele for rapid conversion of elite hybrids into quality protein maize. J Genet 97:287-298

Hoyos-Villegas V, Song Q, Kelly JD (2017) Genome-wide association analysis for drought tolerance and associated traits in common bean. Plant Genome 10:1-17

Hu D, Kan G, Hu W, Li Y, Hao D, Li X, Yang H, Yang Z, He X, Huang F, Yu D (2019) Identification of loci and candidate genes responsible for pod dehiscence in soybean via genome-wide association analysis across multiple environments. Front Plant Sci 10:811

Huang N, Angeles ER, Domingo J, Magpantay G, Singh S, Zhang G, Kumaravadivel N, Bennett J, Khush GS (1997) Pyramiding of bacterial blight resistance genes in rice: marker-assisted selection using RFLP and PCR. Theor Appl Genet 95:313-320

Huang X, Han B (2014) Natural variations and genome-wide association studies in crop plants. Annu Rev Plant Biol 65:531-551

Huang X-Q, Huang T, Hou G-Z, Li L, Hou Y, Lu Y-H (2016) Identification of QTLs for seed quality traits in rapeseed (*Brassica napus* L.) using recombinant inbred lines (RILs). Euphytica 210:1-16

Huang Y, Hussain MA, Luo D, Xu H, Zeng C, Havlickova L, Bancroft I, Tian Z, Zhang X, Cheng Y, Zou X, Lu G, Lv Y (2020) A *Brassica napus* reductase gene dissected by associative transcriptomics enhances plant adaption to freezing stress. Front Plant Sci 11:971

Huhtanen P, Hetta M, Swensson C (2011) Evaluation of canola meal as a protein supplement for dairy cows: A review and a meta-analysis. Can J Anim Sci 91:529-543

Hwang EY, Song Q, Jia G, Specht JE, Hyten DL, Costa J, Cregan PB (2014) A genome-wide association study of seed protein and oil content in soybean. BMC Genomics 15:1

Ibrahim AK, Zhang L, Niyitanga S, Afzal MZ, Xu Y, Zhang L, Zhang L, Qi J (2020) Principles and approaches of association mapping in plant breeding. Trop Plant Biol 13:212-224

Ijaz B, Zhao N, Kong J, Hua J (2019) Fiber quality improvement in upland cotton (*Gossypium hirsutum* L.): quantitative trait loci mapping and marker assisted selection application. Front Plant Sci 10:1585

Illumina Inc (2016) GenomeStudio® Genotyping Module v2.0 Software Guide. Illumina Proprietary

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

Ivanovska S, Stojkovski C, Dimov Z, Marjanović-Jeromela A, Jankulovska M, Jankuloski L (2007) Interrelationship between yield and yield related traits of spring canola (*Brassica napus* L.) genotypes. Genetika 39:325-332

Jabbari M, Fakheri BA, Aghnoum R, Mahdi Nezhad N, Ataei R (2018) GWAS analysis in spring barley (*Hordeum vulgare* L.) for morphological traits exposed to drought. PLoS One 13:e0204952

Jaccoud D, Peng K, Feinstein D, Kilian A (2001) Diversity arrays: a solid state technology for sequence information independent genotyping. Nucleic Acids Res 29:E25

Jan HU, Abbadi A, Lucke S, Nichols RA, Snowdon RJ (2016) Genomic prediction of testcross performance in canola (*Brassica napus*). PLoS One 11:e0147769

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

Jeffreys AJ, Wilson V, Thein SL (1985) Hypervariable 'minisatellite' regions in human DNA. Nature 314:67-73

Jiang C, Shi J, Li R, Long Y, Wang H, Li D, Zhao J, Meng J (2014) Quantitative trait loci that control the oil content variation of rapeseed (*Brassica napus* L.). Theor Appl Genet 127:957-968

Jiang G-L (2013) Molecular markers and marker-assisted breeding in plants. In: Andersen SB (ed) Plant Breeding from Laboratories to Fields. In Tech, pp 45-83

Jiang G-L (2015) Molecular marker-assisted breeding: a plant breeder's review. In: Al-Khayri JM, Jain SM, Johnson DV (eds) Advances in Plant Breeding Strategies: Breeding, Biotechnology and Molecular Tools. Springer International Publishing, Cham, pp 431-472

Jiang Y, Reif JC (2015) Modeling epistasis in genomic selection. Genetics 201:759-768

John MB (2012) Single nucleotide polymorphisms and applications. Advanced Topics in Forensic DNA Typing: Methodology. Academic Press, Walthan, MA, pp 347-369

Jolivet P, Boulard C, Bellamy A, Larre C, Barre M, Rogniaux H, d'Andréa S, Chardot T, Nesi N (2009) Protein composition of oil bodies from mature *Brassica napus* seeds. Proteomics 9:3268-3284

Jonas E, de Koning DJ (2013) Does genomic selection have a future in plant breeding? Trends Biotechnol 31:497-504

Jourdren C, Barret P, Brunel D, Delourme R, Renard M (1996) Specific molecular marker of the genes controlling linolenic acid content in rapeseed. Theor Appl Genet 93:512-518

Ju M, Zhou Z, Mu C, Zhang X, Gao J, Liang Y, Chen J, Wu Y, Li X, Wang S, Wen J, Yang L, Wu J (2017) Dissecting the genetic architecture of *Fusarium verticillioides* seed rot resistance in maize by combining QTL mapping and genome-wide association analysis. Sci Rep 7:46446

Kainer D, Padovan A, Degenhardt J, Krause S, Mondal P, Foley WJ, Külheim C (2019) High marker density GWAS provides novel insights into the genomic architecture of terpene oil yield in *Eucalyptus*. New Phytol 223:1489-1504

Kaler AS, Gillman JD, Beissinger T, Purcell LC (2019) Comparing different statistical models and multiple testing corrections for association mapping in soybean and maize. Front Plant Sci 10:1794

Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E (2008) Efficient control of population structure in model organism association mapping. Genetics 178:1709-1723

Karamanos RE, Harapiak J, Flore NA (2002) Fall and early spring seeding of canola (*Brassica napus* L.) using different methods of seeding and phosphorus placement. Can J Plant Sci 82:21-26

Khan MA, Cowling W, Banga SS, You MP, Tyagi V, Bharti B, Barbetti MJ (2020) Patterns of inheritance for cotyledon resistance against *Sclerotinia sclerotiorum* in *Brassica napus*. Euphytica 216:79

Khazaei H, Podder R, Caron CT, Kundu SS, Diapari M, Vandenberg A, Bett KE (2017) Marker-trait association analysis of iron and zinc concentration in lentil (*Lens culinaris* Medik.) seeds. Plant Genome 10:1-8

Kittipol V, He Z, Wang L, Doheny-Adams T, Langer S, Bancroft I (2019) Genetic architecture of glucosinolate variation in *Brassica napus*. J Plant Physiol 240:152988

Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J (2005) Complement factor H polymorphism in age-related macular degeneration. Science 308:385-389

Knoch D, Werner CR, Meyer RC, Riewe D, Abbadi A, Lucke S, Snowdon RJ, Altmann T (2021) Multi-omics-based prediction of hybrid performance in canola. Theor Appl Genet 134:1147-1165

Kondra Z, Stefansson B (1970) Inheritance of the major glucosinolates of rapeseed (B*rassica napus*) meal. Can J Plant Sci 50:643-647

Korber N, Bus A, Li J, Parkin IA, Wittkop B, Snowdon RJ, Stich B (2016) Agronomic and seed quality traits dissected by genome-wide association mapping in *Brassica napus*. Front Plant Sci 7:386

Korte A, Farlow A (2013) The advantages and limitations of trait analysis with GWAS: a review. Plant Methods 9:29

Koscielny CB (2018) Analysis of thermotolerance in *Brassica napus* L. Department of Plant Science. University of Manitoba, Winnipeg

Koscielny CB, Gardner SW, Technow F, Duncan RW (2020) Linkage mapping and whole-genome predictions in canola (*Brassica napus*) subjected to differing temperature treatments. Crop Pasture Sci 71:229-238

Koscielny CB, Hazebroek J, Duncan RW (2018) Phenotypic and metabolic variation among spring *Brassica napus* genotypes during heat stress. Crop Pasture Sci 69:284-295

Kriaridou C, Tsairidou S, Houston RD, Robledo D (2020) Genomic prediction using low density marker panels in aquaculture: performance across species, traits, and genotyping platforms. Front Genet 11:124

Krishna MSR, Sokka Reddy S, Satyanarayana SDV (2017) Marker-assisted breeding for introgression of opaque-2 allele into elite maize inbred line BML-7. 3 Biotech 7:165

Kutcher HR, Warland JS, Brandt SA (2010) Temperature and precipitation effects on canola yields in Saskatchewan, Canada. Agric For Meteorol 150:161-165

La VH, Lee B-R, Islam MT, Park S-H, Lee H, Bae D-W, Kim T-H (2019) Antagonistic shifting from abscisic acid- to salicylic acid-mediated sucrose accumulation contributes to drought tolerance in *Brassica napus*. Environ Exp Bot 162:38-47

LaFramboise T (2009) Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. Nucleic Acids Res 37:4181-4193

Lancashire PD, Bleiholder H, Boom TVD, LangelÜDdeke P, Stauss R, Weber E, Witzenberger A (1991) A uniform decimal code for growth stages of crops and weeds. Ann Appl Biol 119:561-601

Langridge P, Chalmers K (2005) The principle: identification and application of molecular markers. In: H. L, G. W (eds) Molecular Marker Systems in Plant Breeding and Crop Improvement. Springer, Berlin Heidelberg, pp 3-22

Leite DC, Pinheiro JB, Campos JB, Di Mauro AO, Uneda-Trevisoli SH (2016) QTL mapping of soybean oil content for marker-assisted selection in plant breeding program. Genet Mol Res 15:1-11

Lekklar C, Pongpanich M, Suriya-Arunroj D, Chinpongpanich A, Tsai H, Comai L, Chadchawan S, Buaboocha T (2019) Genome-wide association study for salinity tolerance at the flowering stage in a panel of rice accessions from Thailand. BMC Genomics 20:76

Li CX, Xu WG, Guo R, Zhang JZ, Qi XL, Hu L, Zhao MZ (2018a) Molecular marker assisted breeding and genome composition analysis of Zhengmai 7698, an elite winter wheat cultivar. Sci Rep 8:322

Li F, Chen B, Xu K, Gao G, Yan G, Qiao J, Li J, Li H, Li L, Xiao X, Zhang T, Nishio T, Wu X (2016a) A genome-wide association study of plant height and primary branch number in rapeseed (*Brassica napus*). Plant Sci 242:169-177

Li F, Chen B, Xu K, Wu J, Song W, Bancroft I, Harper AL, Trick M, Liu S, Gao G, Wang N, Yan G, Qiao J, Li J, Li H, Xiao X, Zhang T, Wu X (2014a) Genome-wide association study dissects

the genetic architecture of seed weight and seed quality in rapeseed (*Brassica napus* L.). DNA Res 21:355-367

Li F, Chen B, Xu K, Wu J, Song W, Bancroft I, Harper AL, Trick M, Liu S, Gao G, Wang N, Yan G, Qiao J, Li J, Li H, Xiao X, Zhang T, Wu X (2014b) Genome-wide association study dissects the genetic architecture of seed weight and seed quality in rapeseed (*Brassica napus* L.). DNA Res 21:355-367

Li G, Xu X, Bai G, Carver BF, Hunger R, Bonman JM, Kolmer J, Dong H (2016b) Genome-wide association mapping reveals novel QTL for seedling leaf rust resistance in a worldwide collection of winter wheat. Plant Genome 9:1-12

Li H, Cheng X, Zhang L, Hu J, Zhang F, Chen B, Xu K, Gao G, Li H, Li L, Huang Q, Li Z, Yan G, Wu X (2018b) An integration of genome-wide association study and gene co-expression network analysis identifies candidate genes of stem lodging-related traits in *Brassica napus*. Front Plant Sci 9:796

Li L, Long Y, Zhang L, Dalton-Morgan J, Batley J, Yu L, Meng J, Li M (2015a) Genome wide analysis of flowering time trait in multiple environments via high-throughput genotyping technique in *Brassica napus* L. PLoS One 10:e0119425

Li L, Luo Y, Chen B, Xu K, Zhang F, Li H, Huang Q, Xiao X, Zhang T, Hu J, Li F, Wu X (2016c) A genome-wide association study reveals new loci for resistance to clubroot disease in *Brassica napus*. Front Plant Sci 7:1483

Li L, Peng Z, Mao X, Wang J, Chang X, Reynolds M, Jing R (2019a) Genome-wide association study reveals genomic regions controlling root and shoot traits at late growth stages in wheat. Ann Bot 124:993-1006

Li Q, LI C, Wang C, Sun Y, Jiang Y, Yang B (2019b) Gene cloning, expression analysis and identification of interacting proteins of transcription factor *WRKY72* in oilseed rape (*Brassica napus*). J Agric Biotechnol 27:761-772

Li X, Zhou Z, Ding J, Wu Y, Zhou B, Wang R, Ma J, Wang S, Zhang X, Xia Z, Chen J, Wu J (2016d) Combined linkage and association mapping reveals QTL and candidate genes for plant and ear height in maize. Front Plant Sci 7:833

Li YX, Wu X, Jaqueth J, Zhang D, Cui D, Li C, Hu G, Dong H, Song YC, Shi YS, Wang T, Li B, Li Y (2015b) The identification of two head smut resistance-related QTL in maize by the joint approach of linkage mapping and association analysis. PLoS One 10:e0145549

Liang Z, Gupta SK, Yeh C-T, Zhang Y, Ngu DW, Kumar R, Patil HT, Mungra KD, Yadav DV, Rathore A, Srivastava RK, Gupta R, Yang J, Varshney RK, Schnable PS, Schnable JC (2018) Phenotypic data from inbred parents can improve genomic prediction in pearl millet hybrids. G3 (Bethesda, Md) 8:2513-2522

Lin M, Zhang D, Liu S, Zhang G, Yu J, Fritz AK, Bai G (2016) Genome-wide association analysis on pre-harvest sprouting resistance and grain color in U.S. winter wheat. BMC Genomics 17:794

Lin Y, Liu S, Liu Y, Liu Y, Chen G, Xu J, Deng M, Jiang Q, Wei Y, Lu Y, Zheng Y (2017) Genome-wide association study of pre-harvest sprouting resistance in Chinese wheat founder parents. Genet Mol Biol 40:620-629

Litt M, Luty JA (1989) A hypervariable microsatellite revealed by *in vitro* amplification of a dinucleotide repeat within the cardiac muscle actin gene. Am J Hum Genet 44:397-401

Liu HJ, Yan J (2019) Crop genome-wide association study: a harvest of biological relevance. Plant J 97:8-18

Liu L, Qu C, Wittkop B, Yi B, Xiao Y, He Y, Snowdon RJ, Li J (2013) A high-density SNP map for accurate mapping of seed fibre QTL in *Brassica napus* L. PLoS One 8:e83052

Liu M, Tan X, Yang Y, Liu P, Zhang X, Zhang Y, Wang L, Hu Y, Ma L, Li Z, Zhang Y, Zou C, Lin H, Gao S, Lee M, Lubberstedt T, Pan G, Shen Y (2020a) Analysis of the genetic architecture of maize kernel size traits by combined linkage and association mapping. Plant Biotechnol J 18:207-221

Liu N, Bai G, Lin M, Xu X, Zheng W (2017a) Genome-wide association analysis of powdery mildew resistance in U.S. winter wheat. Sci Rep 7:11743

Liu P, Zhang C, Ma J-Q, Zhang L-Y, Yang B, Tang X-Y, Huang L, Zhou X-T, Lu K, Li J-N (2018) Genome-wide identification and expression profiling of cytokinin oxidase/dehydrogenase (*CKX*) genes reveal likely roles in pod development and stress responses in oilseed rape (*Brassica napus* L.). Genes (Basel) 9:168

Liu P, Zhao Y, Liu G, Wang M, Hu D, Hu J, Meng J, Reif JC, Zou J (2017b) Hybrid performance of an immortalized F2 rapeseed population is driven by additive, dominance, and epistatic effects. Front Plant Sci 8:815

Liu S, Fan C, Li J, Cai G, Yang Q, Wu J, Yi X, Zhang C, Zhou Y (2016a) A genome-wide association study reveals novel elite allelic variations in seed oil content of *Brassica napus*. Theor Appl Genet 129:1203-1215

Liu S, Huang H, Yi X, Zhang Y, Yang Q, Zhang C, Fan C, Zhou Y (2020b) Dissection of genetic architecture for glucosinolate accumulations in leaves and seeds of *Brassica napus* by genome-wide association study. Plant Biotechnol J 18:1472-1484

Liu X, Huang M, Fan B, Buckler ES, Zhang Z (2016b) Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. PLoS Genet 12:e1005767

Liu Y, Lin Y, Gao S, Li Z, Ma J, Deng M, Chen G, Wei Y, Zheng Y (2017c) A genome-wide association study of 23 agronomic traits in Chinese wheat landraces. Plant J 91:861-873

Liu Y, Salsman E, Fiedler JD, Hegstad JB, Green A, Mergoum M, Zhong S, Li X (2019) Genetic mapping and prediction analysis of FHB resistance in a hard red spring wheat breeding population. Front Plant Sci 10:1007

Liu Z-W, Fu T-D, Tu J-X, Chen B-y (2005) Inheritance of seed colour and identification of RAPD and AFLP markers linked to the seed colour gene in rapeseed (*Brassica napus* L.). Theor Appl Genet 110:303-310

LMC International (2020) The economic impact of canola on the Canadian cconomy: 2020 Update. Canola Council of Canada

Longin CF, Mi X, Wurschum T (2015) Genomic selection in wheat: optimum allocation of test resources and comparison of breeding strategies for line and hybrid breeding. Theor Appl Genet 128:1297-1306

Lorenz AJ, Chao S, Asoro FG, Heffner EL, Hayashi T, Iwata H, Smith KP, Sorrells ME, Jannink J-L (2011) Genomic selection in plant breeding: knowledge and prospects. Adv Agron 110:77

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

Lozada DN, Mason RE, Sarinelli JM, Brown-Guedira G (2019) Accuracy of genomic selection for grain yield and agronomic traits in soft red winter wheat. BMC Genet 20:82

Lu K, Peng L, Zhang C, Lu J, Yang B, Xiao Z, Liang Y, Xu X, Qu C, Zhang K, Liu L, Zhu Q, Fu M, Yuan X, Li J (2017) Genome-wide association and transcriptome analyses reveal candidate genes underlying yield-determining traits in *Brassica napus*. Front Plant Sci 8:206

Lu K, Wei L, Li X, Wang Y, Wu J, Liu M, Zhang C, Chen Z, Xiao Z, Jian H, Cheng F, Zhang K, Du H, Cheng X, Qu C, Qian W, Liu L, Wang R, Zou Q, Ying J, Xu X, Mei J, Liang Y, Chai YR, Tang Z, Wan H, Ni Y, He Y, Lin N, Fan Y, Sun W, Li NN, Zhou G, Zheng H, Wang X, Paterson AH, Li J (2019) Whole-genome resequencing reveals *Brassica napus* origin and genetic loci involved in its improvement. Nat Commun 10:1154

Lu K, Xiao Z, Jian H, Peng L, Qu C, Fu M, He B, Tie L, Liang Y, Xu X, Li J (2016) A combination of genome-wide association and transcriptome analysis reveals candidate genes controlling harvest index-related traits in *Brassica napus*. Sci Rep 6:36452

Luo T, Xian M, Zhang C, Zhang C, Hu L, Xu Z (2019) Associating transcriptional regulation for rapid germination of rapeseed (*Brassica napus* L.) under low temperature stress through weighted gene co-expression network analysis. Sci Rep 9:55

Luo X, Xue Z, Ma C, Hu K, Zeng Z, Dou S, Tu J, Shen J, Yi B, Fu T (2017) Joint genome-wide association and transcriptome sequencing reveals a complex polygenic network underlying hypocotyl elongation in rapeseed (*Brassica napus* L.). Sci Rep 7:41561

Ma P, Xu H, Xu Y, Song L, Liang S, Sheng Y, Han G, Zhang X, An D (2018) Characterization of a powdery mildew resistance gene in wheat breeding line 10V-2 and its application in marker-assisted selection. Plant Dis 102:925-931

Mammadov J, Aggarwal R, Buyyarapu R, Kumpatla S (2012) SNP markers and their impact on plant breeding. Int J Plant Genomics 2012:728398

Mangin B, Bonnafous F, Blanchet N, Boniface MC, Bret-Mestries E, Carrere S, Cottret L, Legrand L, Marage G, Pegot-Espagnet P, Munos S, Pouilly N, Vear F, Vincourt P, Langlade NB (2017) Genomic prediction of sunflower hybrids oil content. Front Plant Sci 8:1633

Marmagne A, Brabant P, Thiellement H, Alix K (2010) Analysis of gene expression in resynthesized *Brassica napus* allotetraploids: transcriptional changes do not explain differential protein regulation. New Phytol 186:216-227

Mason AS, Higgins EE, Snowdon RJ, Batley J, Stein A, Werner C, Parkin IA (2017) A user guide to the *Brassica* 60K Illumina Infinium SNP genotyping array. Theor Appl Genet 130:621-633

Maulana F, Kim K-S, Anderson JD, Sorrells ME, Butler TJ, Liu S, Baenziger PS, Byrne PF, Ma X-F (2021) Genomic selection of forage agronomic traits in winter wheat. Crop Sci 61:410-421

Mei DS, Wang HZ, Hu Q, Li YD, Xu YS, Li YC (2009) QTL analysis on plant height and flowering time in *Brassica napus*. Plant Breed 128:458-465

Meng T, Carew R, Florkowski WJ, Klepacka AM (2017) Analyzing temperature and precipitation influences on yield distributions of canola and spring wheat in Saskatchewan. J Appl Meteorol Climatol 56:897-913

Meuwissen TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819-1829

Meuwissen THE (2009) Accuracy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping. Genet Sel Evol 41:35

Michalyna W, Smith RE (1972) Soils of the Portage la Prairie area. Soils Report

Michel S, Loschenberger F, Ametz C, Pachler B, Sparry E, Burstmayr H (2019) Combining grain yield, protein content and protein quality by multi-trait genomic selection in bread wheat. Theor Appl Genet 132:2767-2780

Miedaner T, Korzun V (2012) Marker-assisted selection for disease resistance in wheat and barley breeding. Phytopathology 102:560-566

Mills GF, Haluschak P (1993) Soils of the Carman research station. Canada-Manitoba Soil Survey. Agriculture Canada,Manitoba Department of Agriculture,Department of Soil Science, University of Manitoba

Miyagawa T, Nishida N, Ohashi J, Kimura R, Fujimoto A, Kawashima M, Koike A, Sasaki T, Tanii H, Otowa T, Momose Y, Nakahara Y, Gotoh J, Okazaki Y, Tsuji S, Tokunaga K (2008) Appropriate data cleaning methods for genome-wide association study. J Hum Genet 53:886-893

Moeinizade S, Hu G, Wang L, Schnable PS (2019) Optimizing selection and mating in genomic selection with a look-ahead approach: an operations research framework. G3 (Bethesda) 9:2123-2133

Montesinos-Lopez OA, Montesinos-Lopez A, Perez-Rodriguez P, Barron-Lopez JA, Martini JWR, Fajardo-Flores SB, Gaytan-Lugo LS, Santana-Mancilla PC, Crossa J (2021) A review of deep learning applications for genomic selection. BMC Genomics 22:19

Montesinos-Lopez OA, Montesinos-Lopez A, Tuberosa R, Maccaferri M, Sciara G, Ammar K, Crossa J (2019) Multi-trait, multi-environment genomic prediction of durum wheat with genomic best linear unbiased predictor and deep learning methods. Front Plant Sci 10:1311

Morinaga T (1934) Interspecific hybridization in *Brassica*. VI. The cytology of F1 hybrids of *B. juncea* and *B. nigra*. Cytologia 6:62-67

Morrison MJ, Harker KN, Blackshaw RE, Holzapfel CJ, O'Donovan JT (2016) Canola yield improvement on the Canadian Prairies from 2000 to 2013. Crop Pasture Sci 67:245-252

Morrison MJ, Stewart DW (2002) Heat stress during flowering in summer *Brassica*. Crop Sci 42:797-803

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

Muller BSF, Neves LG, de Almeida Filho JE, Resende MFR, Jr., Munoz PR, Dos Santos PET, Filho EP, Kirst M, Grattapaglia D (2017) Genomic prediction in contrast to a genome-wide association study in explaining heritable variation of complex growth traits in breeding populations of *Eucalyptus*. BMC Genomics 18:524

Myles AJ, Feudale RN, Liu Y, Woody NA, Brown SD (2004) An introduction to decision tree modeling. J Chemometrics 18:275-285

Nadeem MA, Nawaz MA, Shahid MQ, Doğan Y, Comertpay G, Yıldız M, Hatipoğlu R, Ahmad F, Alsaleh A, Labhane N, Özkan H, Chung G, Baloch FS (2017) DNA molecular markers in plant breeding: current status and recent advancements in genomic selection and genome editing. Biotechnol Biotechnol Equip 32:261-285

Nakano Y, Kusunoki K, Hoekenga OA, Tanaka K, Iuchi S, Sakata Y, Kobayashi M, Yamamoto YY, Koyama H, Kobayashi Y (2020) Genome-wide association study and genomic prediction elucidate the distinct genetic architecture of aluminum and proton tolerance in *Arabidopsis thaliana*. Front Plant Sci 11:405

Nascimento M, Nascimento ACC, Silva FFE, Barili LD, Vale NMD, Carneiro JE, Cruz CD, Carneiro PCS, Serao NVL (2018) Quantile regression for genome-wide association study of flowering time-related traits in common bean. PLoS One 13:e0190303

Nesi N, Delourme R, Brégeon M, Falentin C, Renard M (2008) Genetic and molecular approaches to improve nutritional value of *Brassica napus* L. seed. C R Biol 331:763-771

Neupane S, Purintun JM, Mathew FM, Varenhorst AJ, Nepal MP (2019) Molecular basis of soybean resistance to soybean aphids and soybean cyst nematodes. Plants (Basel) 8:374

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

Newell MA, Jannink JL (2014) Genomic selection in plant breeding. Methods Mol Biol 1145:117-130

Norman A, Taylor J, Edwards J, Kuchel H (2018) Optimising genomic selection in wheat: effect of marker density, population size and population structure on prediction accuracy. G3 (Bethesda) 8:2889-2899

NRGene (2021) Canola/rapeseed pan-genome consortium results reveal broad genetic diversity of the crop. NRGene

Nuttall WF, Moulin AP, Townley-Smith LJ (1992) Yield response of canola to nitrogen, phosphorus, precipitation, and temperature. Agron J 84:765-768

Odukoya J, Lambert R, Sakrabani R (2019) Understanding the impacts of crude oil and its induced abiotic stresses on agrifood production: A review. Horticulturae 5:47

Ogbonnaya FC, Rasheed A, Okechukwu EC, Jighly A, Makdis F, Wuletaw T, Hagras A, Uguru MI, Agbo CU (2017) Genome-wide association study for agronomic and physiological traits in spring wheat evaluated in a range of heat prone environments. Theor Appl Genet 130:1819-1835

Ogura H (1968) Studies on the new male sterility in Japanese radish, with special references on the utilization of this sterility towards the practical raising of hybrid seeds. Mem Fac Agric Kagoshima Univ 6:40-75

Ornella L, Gonzalez-Camacho JM, Dreisigacker S, Crossa J (2017) Applications of genomic selection in breeding wheat for rust resistance. Methods Mol Biol 1659:173-182

Ortiz R (1998) Critical role of plant biotechnology for the genetic improvement of food crops: perspectives for the next millennium. Electron J Biotechnol 1:16-17

Pace J, Yu X, Lubberstedt T (2015) Genomic prediction of seedling root length in maize (*Zea mays* L.). Plant J 83:903-912

Pantaliao GF, Narciso M, Guimaraes C, Castro A, Colombari JM, Breseghello F, Rodrigues L, Vianello RP, Borba TO, Brondani C (2016) Genome wide association study (GWAS) for grain yield in rice cultivated under water deficit. Genetica 144:651-664

Paran I, Michelmore RW (1993) Development of reliable PCR-based markers linked to downy mildew resistance genes in lettuce. Theor Appl Genet 85:985-993

Parmley KA, Higgins RH, Ganapathysubramanian B, Sarkar S, Singh AK (2019) Machine learning approach for prescriptive plant breeding. Sci Rep 9:17132

Paterson AH, Lan TH, Amasino R, Osborn TC, Quiros C (2001) *Brassica* genomics: a complement to, and early beneficiary of, the *Arabidopsis* sequence. Genome Biol 2:1-4

Patil G, Vuong TD, Kale S, Valliyodan B, Deshmukh R, Zhu C, Wu X, Bai Y, Yungbluth D, Lu F, Kumpatla S, Shannon JG, Varshney RK, Nguyen HT (2018) Dissecting genomic hotspots underlying seed protein, oil, and sucrose content in an interspecific mapping population of soybean using high-density linkage mapping. Plant Biotechnol J 16:1939-1953

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V (2011) Scikit-learn: Machine learning in Python. J Mach Learn Res 12:2825-2830

Peiffer JA, Romay MC, Gore MA, Flint-Garcia SA, Zhang Z, Millard MJ, Gardner CA, McMullen MD, Holland JB, Bradbury PJ, Buckler ES (2014) The genetic architecture of maize height. Genetics 196:1337-1356

Perez P, de los Campos G (2014) Genome-wide regression and prediction with the BGLR statistical package. Genetics 198:483-495

Phing Lau WC, Latif MA, M YR, Ismail MR, Puteh A (2016) Advances to improve the eating and cooking qualities of rice by marker-assisted breeding. Crit Rev Biotechnol 36:87-98

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

Pingali PL, Heisey PW (2001) Cereal-crop productivity in developing countries: past trends and future prospects. In: Alston JM, Pardey PG, Taylor MJ (eds) Agricultural science policy: Changing global agendas. The Johns Hopkins University Press, pp 56-82

Poisson E, Trouverie J, Brunel-Muguet S, Akmouche Y, Pontet C, Pinochet X, Avice JC (2019) Seed yield components and seed quality of oilseed rape are impacted by sulfur fertilization and its interactions with nitrogen fertilization. Front Plant Sci 10:458

Polowick PL, Sawhney VK (1988) High temperature induced male and female sterility in canola (*Brassica napus* L.). Ann Bot 62:83-86

Porebski S, Bailey LG, Baum BR (1997) Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. Plant Mol Biol Report 15:8-15

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945-959

Prohens J (2011) Plant breeding: a success story to be continued thanks to the advances in genomics. Front Plant Sci 2:51

Qu C, Jia L, Fu F, Zhao H, Lu K, Wei L, Xu X, Liang Y, Li S, Wang R, Li J (2017) Genome-wide association mapping and Identification of candidate genes for fatty acid composition in *Brassica napus* L. using SNP markers. BMC Genomics 18:232

Qu C-M, Li S-M, Duan X-J, Fan J-H, Jia L-D, Zhao H-Y, Lu K, Li J-N, Xu X-F, Wang R (2015) Identification of candidate genes for seed glucosinolate content using association mapping in *Brassica napus* L. Genes (Basel) 6:1215-1229

Rafalski JA (2010) Association genetics in crop improvement. Curr Opin Plant Biol 13:174-180

Rahman H, Peng G, Yu F, Falk KC, Kulkarni M, Selvaraj G (2014) Genetics and breeding for clubroot resistance in Canadian spring canola (*Brassica napus* L.). Can J Plant Pathol 36:122-134

Rahman M, Sun Z, McVetty PB, Li G (2008) High throughput genome-specific and gene-specific molecular markers for erucic acid genes in *Brassica napus* (L.) for marker-assisted selection in plant breeding. Theor Appl Genet 117:895-904

Rakow G (2004) Species Origin and Economic Importance of Brassica. In: Pua E-C, Douglas CJ (eds) Brassica. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 3-11

Rakow G, Raney J, Relf-Eckstein J (1999) Agronomic performance and seed quality of a new source of yellow seeded *Brassica napus*. 10th International Rapeseed Congress, Canberra, Australia

Raman H, Raman R, Coombes N, Song J, Diffey S, Kilian A, Lindbeck K, Barbulescu DM, Batley J, Edwards D, Salisbury PA, Marcroft S (2016) Genome-wide association study identifies new loci for resistance to *Leptosphaeria maculans* in canola. Front Plant Sci 7:1513

Raman H, Raman R, Qiu Y, Yadav AS, Sureshkumar S, Borg L, Rohan M, Wheeler D, Owen O, Menz I, Balasubramanian S (2019) GWAS hints at pleiotropic roles for *FLOWERING LOCUS T* in flowering time and yield-related traits in canola. BMC Genomics 20:636

Randhawa HS, Asif M, Pozniak C, Clarke JM, Graf RJ, Fox SL, Humphreys DG, Knox RE, DePauw RM, Singh AK, Cuthbert RD, Hucl P, Spaner D, Gupta P (2013) Application of molecular markers to wheat breeding in Canada. Plant Breed 132:458-471

Rasheed A, Hao Y, Xia X, Khan A, Xu Y, Varshney RK, He Z (2017) Crop breeding chips and genotyping platforms: progress, challenges, and perspectives. Mol Plant 10:1047-1064

Raymer PL (2002) Canola: an emerging oilseed crop. In: Janick J, Whipkey A (eds) Trends in New Crops and New Uses. ASHS Press, Alexandria, Virginia, pp 122-126

Resende MF, Jr., Munoz P, Resende MD, Garrick DJ, Fernando RL, Davis JM, Jokela EJ, Martin TA, Peter GF, Kirst M (2012) Accuracy of genomic selection methods in a standard data set of loblolly pine (*Pinus taeda* L.). Genetics 190:1503-1510

Revilla P, Rodriguez VM, Ordas A, Rincent R, Charcosset A, Giauffret C, Melchinger AE, Schon CC, Bauer E, Altmann T, Brunel D, Moreno-Gonzalez J, Campo L, Ouzunova M, Alvarez A, Ruiz

de Galarreta JI, Laborde J, Malvar RA (2016) Association mapping for cold tolerance in two large maize inbred panels. BMC Plant Biol 16:127

Ribaut J-M, Hoisington D (1998) Marker-assisted selection: new tools and strategies. Trends Plant Sci 3:236-239

Ribaut JM, Ragot M (2007) Marker-assisted selection to improve drought adaptation in maize: the backcross approach, perspectives, limitations, and alternatives. J Exp Bot 58:351-360

Riedelsheimer C, Czedik-Eysenberg A, Grieder C, Lisec J, Technow F, Sulpice R, Altmann T, Stitt M, Willmitzer L, Melchinger AE (2012) Genomic and metabolic prediction of complex heterotic traits in hybrid maize. Nat Genet 44:217-220

Robertsen C, Hjortshøj R, Janss L (2019) Genomic selection in cereal breeding. Agronomy 9:95

Robinson GK (1991) That BLUP is a good thing: the estimation of random effects. Statistical science 6:15-32

Roehrig CS (1988) Conditions for identification in nonparametric and parametric models. Econometrica 56:433-447

Roorkiwal M, Jarquin D, Singh MK, Gaur PM, Bharadwaj C, Rathore A, Howard R, Srinivasan S, Jain A, Garg V, Kale S, Chitikineni A, Tripathi S, Jones E, Robbins KR, Crossa J, Varshney RK (2018) Genomic-enabled prediction models using multi-environment trials to estimate the effect of genotype x environment interaction on prediction accuracy in chickpea. Sci Rep 8:11701

RStudio Team (2020) RStudio: Integrated Development for R. 1.3.1073 edn. RStudio, PBC, Boston, MA

Rutkoski J, Poland J, Mondal S, Autrique E, Perez LG, Crossa J, Reynolds M, Singh R (2016) Canopy temperature and vegetation indices from high-throughput phenotyping improve accuracy of pedigree and genomic selection for grain yield in wheat. G3 (Bethesda) 6:2799-2808

Rutkoski JE (2019) A practical guide to genetic gain. In: Sparks DL (ed) Adv Agron. Academic Press, pp 217-249

Ruttan VW (1999) The transition to agricultural sustainability. Proceedings of the National Academy of Sciences of the USA 96:5960-5967

Sarinelli JM, Murphy JP, Tyagi P, Holland JB, Johnson JW, Mergoum M, Mason RE, Babar A, Harrison S, Sutton R, Griffey CA, Brown-Guedira G (2019) Training population selection and use of fixed effects to optimize genomic predictions in a historical USA winter wheat panel. Theor Appl Genet 132:1247-1261

Savitch LV, Allard G, Seki M, Robert LS, Tinker NA, Huner NP, Shinozaki K, Singh J (2005) The effect of overexpression of two *Brassica CBF/DREB1*-like transcription factors on photosynthetic capacity and freezing tolerance in *Brassica napus*. Plant Cell Physiol 46:1525-1539

Scheben A, Batley J, Edwards D (2017) Genotyping-by-sequencing approaches to characterize crop genomes: choosing the right tool for the right application. Plant Biotechnol J 15:149-161

Scherer A, Christensen GB (2016) Concepts and relevance of genome-wide association studies. Sci Prog 99:59-67

Schiessl S, Iniguez-Luy F, Qian W, Snowdon RJ (2015) Diverse regulatory factors associate with flowering time and yield responses in winter-type *Brassica napus*. BMC Genomics 16:737

Schulthess AW, Zhao Y, Longin CFH, Reif JC (2018) Advantages and limitations of multiple-trait genomic prediction for *Fusarium* head blight severity in hybrid wheat (*Triticum aestivum* L.). Theor Appl Genet 131:685-701

Schulz-Streeck T, Ogutu JO, Karaman Z, Knaak C, Piepho HP (2012) Genomic selection using multiple populations. Crop Sci 52:2453-2461

Sebastiani P, Timofeev N, Dworkis DA, Perls TT, Steinberg MH (2009) Genome-wide association studies and the genetic dissection of complex traits. Am J Hematol 84:504-515

Segura V, Vilhjalmsson BJ, Platt A, Korte A, Seren U, Long Q, Nordborg M (2012) An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat Genet 44:825-830

Shen D, Suhrkamp I, Wang Y, Liu S, Menkhaus J, Verreet JA, Fan L, Cai D (2014) Identification and characterization of microRNAs in oilseed rape (*Brassica napus*) responsive to infection with the pathogenic fungus *Verticillium longisporum* using *Brassica* AA (*Brassica rapa*) and CC (*Brassica oleracea*) as reference genomes. New Phytol 204:577-594

Sheng Y, Yan X, Huang Y, Han Y, Zhang C, Ren Y, Fan T, Xiao F, Liu Y, Cao S (2019) The *WRKY* transcription factor, *WRKY13*, activates *PDR8* expression to positively regulate cadmium tolerance in *Arabidopsis*. Plant Cell Environ 42:891-903

Shu YJ, Yu DS, Wang D, Bai X, Zhu YM, Guo CH (2013) Genomic selection of seed weight based on low-density SCAR markers in soybean. Genet Mol Res 12:2178-2188

Sidlauskas G, Bernotas S (2003) Some factors affecting seed yield of spring oilseed rape (*Brassica napus* L.). Agron Res 1:229-243

Singh BD, Singh AK (2015) High-throughput SNP genotyping. Marker-Assisted Plant Breeding: Principles and Practices. Springer, India, New Delhi, pp 367-400

Sleper DA, Poehlman JM (2006) Biotechnology and Plant Breeding. Breeding Field Crops, 5 edn. Blackwell Publishing, Ames, Iowa, pp 115-134

Snowdon RJ (2007) Cytogenetics and genome analysis in *Brassica* crops. Chromosome Res 15:85-95

Snowdon RJ, Friedrich T, Friedt W, Köhler W (2002) Identifying the chromosomes of the A- and C-genome diploid Brassica species *B. rapa* (syn. *campestris*) and *B. oleracea* in their amphidiploid *B. napus*. Theor Appl Genet 104:533-538

Snowdon RJ, Friedt W (2004) Molecular markers in Brassica oilseed breeding: current status and future possibilities. Plant Breed 123:1-8

Snowdon RJ, Iniguez Luy FL (2012) Potential to improve oilseed rape and canola breeding in the genomics era. Plant Breed 131:351-360

Snowdon RJ, Lühs W, Friedt W (2007) Oilseed rape. In: Kole C (ed) Genome Mapping and Molecular Breeding in Plants. Springer-Verlag Berlin Heidelberg, pp 55-114

Sokólski M, Jankowski KJ, Załuski D, Szatkowski A (2020) Productivity, Energy and Economic Balance in the Production of Different Cultivars of Winter Oilseed Rape. A Case Study in North-Eastern Poland. Agronomy 10:508

Solberg TR, Sonesson AK, Woolliams JA, Meuwissen TH (2008) Genomic selection using different marker types and densities. J Anim Sci 86:2447-2454

Somers D, Rakow G, Raney P, Prabhu V, Séguin-Swartz G, Rimmer R, Gugel R, Lydiate D, Sharpe A (1999) Developing marker-assisted breeding for quality and disease resistance traits in *Brassica* oilseeds. 10th International Rapeseed Congress, Canberra, Australia

Somers DJ, Rakow G, Prabhu VK, Friesen KR (2001) Identification of a major gene and RAPD markers for yellow seed coat colour in *Brassica napus*. Genome 44:1077-1082

Song J, Jiang L, Jameson PE (2015) Expression patterns of *Brassica napus* genes implicate IPT, CKX, sucrose transporter, cell wall invertase, and amino acid permease gene family members in leaf, flower, silique, and seed development. J Exp Bot 66:5067-5082

Song JM, Guan Z, Hu J, Guo C, Yang Z, Wang S, Liu D, Wang B, Lu S, Zhou R, Xie WZ, Cheng Y, Zhang Y, Liu K, Yang QY, Chen LL, Guo L (2020) Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of Brassica napus. Nat Plants 6:34-45

Soto-Cerda BJ, Cloutier S, Quian R, Gajardo HA, Olivos M, You FM (2018) Genome-wide association analysis of mucilage and hull content in flax (*Linum usitatissimum* L.) seeds. Int J Mol Sci 19:2870

Spasibionek S, Mikolajczyk K, Cwiek-Kupczynska H, Pietka T, Krotka K, Matuszczak M, Nowakowska J, Michalski K, Bartkowiak-Broda I (2020) Marker assisted selection of new high oleic and low linolenic winter oilseed rape (*Brassica napus* L.) inbred lines revealing good agricultural value. PLoS One 15:e0233959

Spindel J, Begum H, Akdemir D, Collard B, Redoña E, Jannink J, McCouch S (2016) Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. Heredity 116:395-408

Spindel J, Begum H, Akdemir D, Virk P, Collard B, Redona E, Atlin G, Jannink JL, McCouch SR (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

Starmer KP, Brown J, Davis JB (1998) Heterosis in spring canola hybrids grown in northern Idaho. Crop Sci 38:376-380

Statistics Canada (2019) Table 32-10-0359-01 Estimated areas, yield, production, average farm price and total farm value of principal field crops, in metric and imperial units.

Stefansson BR, Hougen FW (1964) Selection of rape plants (*Brassica napus*) with seed oil practically free from erucic acid. Can J Plant Sci 44:359-364

Stefansson BR, Hougen FW, Downey RK (1961) Note on the isolation of rape plants with seed oil free from erucic acid. Can J Plant Sci 41:218-219

Stefansson BR, Kondra ZP (1975) Tower summer rape. Can J Plant Sci 55:343-344

Su G, Guldbrandtsen B, Gregersen VR, Lund MS (2010) Preliminary investigation on reliability of genomic estimated breeding values in the Danish Holstein population. J Dairy Sci 93:1175-1183

Sun C, Dong Z, Zhao L, Ren Y, Zhang N, Chen F (2020) The Wheat 660K SNP array demonstrates great potential for marker-assisted selection in polyploid wheat. Plant Biotechnol J 18:1354-1360

Sun C, Wang B, Yan L, Hu K, Liu S, Zhou Y, Guan C, Zhang Z, Li J, Zhang J, Chen S, Wen J, Ma C, Tu J, Shen J, Fu T, Yi B (2016a) Genome-wide association study provides insight into the genetic control of plant height in rapeseed (*Brassica napus* L.). Front Plant Sci 7:1102

Sun C, Zhang F, Yan X, Zhang X, Dong Z, Cui D, Chen F (2017) Genome-wide association study for 13 agronomic traits reveals distribution of superior alleles in bread wheat from the Yellow and Huai Valley of China. Plant Biotechnol J 15:953-969

Sun F, Liu J, Hua W, Sun X, Wang X, Wang H (2016b) Identification of stable QTLs for seed oil content by combined linkage and association mapping in *Brassica napus*. Plant Sci 252:388-399

Sun X, Ma P, Mumm RH (2012) Nonparametric method for genomics-based prediction of performance of quantitative traits involving epistasis in plant breeding. PLoS One 7:e50604

Tan B, Grattapaglia D, Martins GS, Ferreira KZ, Sundberg B, Ingvarsson PK (2017) Evaluating the accuracy of genomic prediction of growth and wood traits in two Eucalyptus species and their F1 hybrids. BMC Plant Biol 17:110

Tang S, Zhao H, Lu S, Yu L, Zhang G, Zhang Y, Yang QY, Zhou Y, Wang X, Ma W, Xie W, Guo L (2021) Genome- and transcriptome-wide association studies provide insights into the genetic basis of natural variation of seed oil content in *Brassica napus*. Mol Plant 14:470-487

Tayeh N, Klein A, Le Paslier MC, Jacquin F, Houtin H, Rond C, Chabert-Martinello M, Magnin-Robert JB, Marget P, Aubert G, 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

Technow F, Riedelsheimer C, Schrag TA, Melchinger AE (2012) Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects. Theor Appl Genet 125:1181-1194

Technow F, Schrag TA, Schipprack W, Bauer E, Simianer H, Melchinger AE (2014) Genome properties and prospects of genomic prediction of hybrid performance in a breeding program of maize. Genetics 197:1343-1355

ter Braak CJ, Boer MP, Bink MC (2005) Extending Xu's Bayesian model for estimating polygenic effects using markers of the entire genome. Genetics 170:1435-1438

Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, Deboy RT, Davidsen TM, Mora M, Scarselli M, Margarit y Ros I, Peterson JD, Hauser CR, Sundaram JP, Nelson WC, Madupu R, Brinkac LM, Dodson RJ, Rosovitz MJ, Sullivan SA, Daugherty SC, Haft DH, Selengut J, Gwinn ML, Zhou L, Zafar N, Khouri H, Radune D, Dimitrov G, Watkins K, O'Connor KJ, Smith S, Utterback TR, White O, Rubens CE, Grandi G, Madoff LC, Kasper DL, Telford JL, Wessels MR, Rappuoli R, Fraser CM (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". Proc Natl Acad Sci U S A 102:13950-13955

Thavamanikumar S, Dolferus R, Thumma BR (2015) Comparison of genomic selection models to predict flowering time and spike grain number in two hexaploid wheat doubled haploid populations. G3 (Bethesda) 5:1991-1998

Tsai HY, Janss LL, Andersen JR, Orabi J, Jensen JD, Jahoor A, Jensen J (2020) Genomic prediction and GWAS of yield, quality and disease-related traits in spring barley and winter wheat. Sci Rep 10:3347

Tuberosa R, Salvi S (2006) Genomics-based approaches to improve drought tolerance of crops. Trends Plant Sci 11:405-412

U N (1935) Genome analysis in *Brassica* with special reference to the experimental formation of *B. napus* and peculiar mode of fertilization. The Journal of Japanese Botany 7:389-452

Upadhyaya HD, Bajaj D, Das S, Saxena MS, Badoni S, Kumar V, Tripathi S, Gowda CL, Sharma S, Tyagi AK, Parida SK (2015) A genome-scale integrated approach aids in genetic dissection of complex flowering time trait in chickpea. Plant Mol Biol 89:403-420

USDA (2020) World agricultural production. Circular Series

van Dijk ADJ, Kootstra G, Kruijer W, de Ridder D (2021) Machine learning in plant science and plant breeding. iScience 24:101890

VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor JF, Schenkel FS (2009) Invited review: reliability of genomic predictions for North American Holstein bulls. J Dairy Sci 92:16-24

Vigeolas H, Waldeck P, Zank T, Geigenberger P (2007) Increasing seed oil content in oil-seed rape (*Brassica napus* L.) by over-expression of a yeast glycerol-3-phosphate dehydrogenase under the control of a seed-specific promoter. Plant Biotechnol J 5:431-441

Visscher PM, Brown MA, McCarthy MI, Yang J (2012) Five years of GWAS discovery. Am J Hum Genet 90:7-24

Vivek BS, Krishna GK, Vengadessan V, Babu R, Zaidi PH, Kha LQ, Mandal SS, Grudloyma P, Takalkar S, Krothapalli K, Singh IS, Ocampo ETM, Xingming F, Burgueno J, Azrai M, Singh RP, Crossa J (2017) Use of genomic estimated breeding values results in rapid genetic gains for drought tolerance in maize. Plant Genome 10:1-8

Vos P, Hogers R, Bleeker M, Reijans M, van de Lee T, Hornes M, Frijters A, Pot J, Peleman J, Kuiper M, et al. (1995) AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res 23:4407-4414

Voss-Fels KP, Cooper M, Hayes BJ (2019) Accelerating crop genetic gains with genomic selection. Theor Appl Genet 132:669-686

Wan H, Chen L, Guo J, Li Q, Wen J, Yi B, Ma C, Tu J, Fu T, Shen J (2017) Genome-wide association study reveals the genetic architecture underlying salt tolerance-related traits in rapeseed (*Brassica napus* L.). Front Plant Sci 8:593

Wang B, Wu Z, Li Z, Zhang Q, Hu J, Xiao Y, Cai D, Wu J, King GJ, Li H, Liu K (2018a) Dissection of the genetic architecture of three seed-quality traits and consequences for breeding in *Brassica napus*. Plant Biotechnol J 16:1336-1348

Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J, Kruglyak L, Stein L, Hsie L, Topaloglou T, Hubbell E, Robinson E, Mittmann M, Morris MS, Shen N, Kilburn D, Rioux J, Nusbaum C, Rozen S, Hudson TJ, Lipshutz R, Chee M, Lander ES (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 280:1077-1082

Wang H, Li K, Hu X, Liu Z, Wu Y, Huang C (2016a) Genome-wide association analysis of forage quality in maize mature stalk. BMC Plant Biol 16:227

Wang H, Xu C, Liu X, Guo Z, Xu X, Wang S, Xie C, Li WX, Zou C, Xu Y (2017a) Development of a multiple-hybrid population for genome-wide association studies: theoretical consideration and genetic mapping of flowering traits in maize. Sci Rep 7:40239

Wang H, Xu S, Fan Y, Liu N, Zhan W, Liu H, Xiao Y, Li K, Pan Q, Li W, Deng M, Liu J, Jin M, Yang X, Li J, Li Q, Yan J (2018b) Beyond pathways: genetic dissection of tocopherol content in maize kernels by combining linkage and association analyses. Plant Biotechnol J 16:1464-1475

Wang J, Jian H, Wei L, Qu C, Xu X, Lu K, Qian W, Li J, Li M, Liu L (2015a) Genome-wide analysis of seed acid detergent lignin (ADL) and hull content in rapeseed (*Brassica napus* L.). PLoS One 10:e0145045

Wang J, Xian X, Xu X, Qu C, Lu K, Li J, Liu L (2017b) Genome-wide association mapping of seed coat color in *Brassica napus*. J Agric Food Chem 65:5229-5237

Wang J, Zhang Z (2020) GAPIT Version 3: boosting power and accuracy for genomic association and prediction. bioRxiv:2020.2011.2029.403170

Wang N, Qian W, Suppanz I, Wei L, Mao B, Long Y, Meng J, Müller AE, Jung C (2011) Flowering time variation in oilseed rape (*Brassica napus* L.) is associated with allelic variation in the *FRIGIDA* homologue *BnaA.FRI.a*. J Exp Bot 62:5641-5658

Wang Q, Wei J, Pan Y, Xu S (2016b) An efficient empirical Bayes method for genomewide association studies. J Anim Breed Genet 133:253-263

Wang R, Chen J, Anderson JA, Zhang J, Zhao W, Wheeler J, Klassen N, See DR, Dong Y (2017c) Genome-wide association mapping of *Fusarium* head blight resistance in spring wheat lines developed in the Pacific Northwest and CIMMYT. Phytopathology 107:1486-1495

Wang T, Chen YP, Goddard ME, Meuwissen TH, Kemper KE, Hayes BJ (2015b) A computationally efficient algorithm for genomic prediction using a Bayesian model. Genet Sel Evol 47:34

Wang X, Chen Y, Thomas CL, Ding G, Xu P, Shi D, Grandke F, Jin K, Cai H, Xu F, Yi B, Broadley MR, Shi L (2017d) Genetic variants associated with the root system architecture of oilseed rape (*Brassica napus* L.) under contrasting phosphate supply. DNA Res 24:407-417

Wang X, Liu G, Yang Q, Hua W, Liu J, Wang H (2010) Genetic analysis on oil content in rapeseed (*Brassica napus* L.). Euphytica 173:17-24

Wang X, Wang H, Liu S, Ferjani A, Li J, Yan J, Yang X, Qin F (2016c) Genetic variation in *ZmVPP1* contributes to drought tolerance in maize seedlings. Nat Genet 48:1233-1241

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

Wang Y, Cheng X, Shan Q, Zhang Y, Liu J, Gao C, Qiu JL (2014) Simultaneous editing of three homoeoalleles in hexaploid bread wheat confers heritable resistance to powdery mildew. Nat Biotechnol 32:947-951

Watson A, Hickey LT, Christopher J, Rutkoski J, Poland J, Hayes BJ (2019) Multivariate genomic selection and potential of rapid indirect selection with speed breeding in spring wheat. Crop Sci 59:1945-1959

Watts R (2013) Development of linkage map of *Brassica juncea* using molecular markers and detection of quantitative trait loci for oil content, seed protein and fatty acids. Plant Sci. University of Manitoba, Winnipeg, p 123

Wei D, Cui Y, He Y, Xiong Q, Qian L, Tong C, Lu G, Ding Y, Li J, Jung C, Qian W (2017) A genome-wide survey with different rapeseed ecotypes uncovers footprints of domestication and breeding. J Exp Bot 68:4791-4801

Wei D, Cui Y, Mei J, Qian L, Lu K, Wang ZM, Li J, Tang Q, Qian W (2019a) Genome-wide identification of loci affecting seed glucosinolate contents in *Brassica napus* L. J Integr Plant Biol 61:611-623

Wei L, Jian H, Lu K, Filardo F, Yin N, Liu L, Qu C, Li W, Du H, Li J (2016) Genome-wide association analysis and differential expression analysis of resistance to *Sclerotinia* stem rot in *Brassica napus*. Plant Biotechnol J 14:1368-1380

Wei L, Zhu Y, Liu R, Zhang A, Zhu M, Xu W, Lin A, Lu K, Li J (2019b) Genome wide identification and comparative analysis of glutathione transferases (GST) family genes in *Brassica napus*. Sci Rep 9:9196

Werner CR, Gaynor RC, Gorjanc G, Hickey JM, Kox T, Abbadi A, Leckband G, Snowdon RJ, Stahl A (2020) How population structure impacts genomic selection accuracy in cross-validation: implications for practical breeding. Front Plant Sci 11:592977

Werner CR, Qian L, Voss-Fels KP, Abbadi A, Leckband G, Frisch M, Snowdon RJ (2018a) Genome-wide regression models considering general and specific combining ability predict hybrid performance in oilseed rape with similar accuracy regardless of trait architecture. Theor Appl Genet 131:299-317

Werner CR, Voss-Fels KP, Miller CN, Qian W, Hua W, Guan CY, Snowdon RJ, Qian L (2018b) Effective genomic selection in a narrow-genepool crop with low-density markers: Asian rapeseed as an example. Plant Genome 11:1-14

Whittaker JC, Thompson R, Denham MC (2000) Marker-assisted selection using ridge regression. Genet Res 75:249-252

Wickham H (2016) ggplot2: Elegant Graphics for Data Analysis., 2 edn. Springer-Verlag New York

Wieczorek AM, Wright MG (2012) History of agricultural biotechnology: How crop development has evolved. Nat Edu 3:9

Wijerathna YMAM (2015) Marker assisted selection: biotechnology tool for rice molecular breeding. Adv Crop Sci Technol 3:187

Williams JG, Kubelik AR, Livak KJ, Rafalski JA, Tingey SV (1990) DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Res 18:6531-6535

Wu J, Zhao Q, Liu S, Shahid M, Lan L, Cai G, Zhang C, Fan C, Wang Y, Zhou Y (2016a) Genome-wide association study identifies new loci for resistance to *Sclerotinia* stem rot in *Brassica napus*. Front Plant Sci 7:1418

Wu W, Ma BL (2016) A new method for assessing plant lodging and the impact of management options on lodging in canola crop production. Sci Rep 6:31890

Wu Z, Wang B, Chen X, Wu J, King GJ, Xiao Y, Liu K (2016b) Evaluation of linkage disequilibrium pattern and association study on seed oil content in *Brassica napus* using ddRAD sequencing. PLoS One 11:e0146383

Würschum T, Abel S, Zhao Y, Léon J (2014) Potential of genomic selection in rapeseed (*Brassica napus* L.) breeding. Plant Breed 133:45-51

Xiao Y, Liu H, Wu L, Warburton M, Yan J (2017) Genome-wide association studies in maize: praise and stargaze. Mol Plant 10:359-374

Xiao Z, Zhang C, Tang F, Yang B, Zhang L, Liu J, Huo Q, Wang S, Li S, Wei L, Du H, Qu C, Lu K, Li J, Li N (2019) Identification of candidate genes controlling oil content by combination of genome-wide association and transcriptome analysis in the oilseed crop *Brassica napus*. Biotechnol Biofuels 12:216

Xie D, Dai Z, Yang Z, Tang Q, Deng C, Xu Y, Wang J, Chen J, Zhao D, Zhang S, Zhang S, Su J (2019) Combined genome-wide association analysis and transcriptome sequencing to identify candidate genes for flax seed fatty acid metabolism. Plant Sci 286:98-107

Xu L, Hu K, Zhang Z, Guan C, Chen S, Hua W, Li J, Wen J, Yi B, Shen J, Ma C, Tu J, Fu T (2016) Genome-wide association study reveals the genetic architecture of flowering time in rapeseed (*Brassica napus* L.). DNA Res 23:43-52

Xu N, Wilson HF, Saiers JE, Entz M (2013) Effects of crop rotation and management system on water-extractable organic matter concentration, structure, and bioavailability in a chernozemic agricultural soil. J Environ Qual 42:179-190

Xu S (2003) Estimating polygenic effects using markers of the entire genome. Genetics 163:789-801

Xu Y, Crouch JH (2008) Marker-assisted selection in plant breeding: from publications to practice. Crop Sci 48:391-407

Xu Y, Wang X, Ding X, Zheng X, Yang Z, Xu C, Hu Z (2018) Genomic selection of agronomic traits in hybrid rice using an NCII population. Rice (N Y) 11:32

Xue Y, Chen B, Wang R, Win AN, Li J, Chai Y (2018) Genome-wide survey and characterization of fatty acid desaturase gene family in *Brassica napus* and its parental species. Appl Biochem Biotechnol 184:582-598

Yan J, Warburton M, Crouch J (2011) Association mapping for enhancing maize (*Zea mays* L.) genetic improvement. Crop Sci 51:433-449

Yang M, Lu K, Zhao FJ, Xie W, Ramakrishna P, Wang G, Du Q, Liang L, Sun C, Zhao H, Zhang Z, Liu Z, Tian J, Huang XY, Wang W, Dong H, Hu J, Ming L, Xing Y, Wang G, Xiao J, Salt DE, Lian X (2018) Genome-Wide Association Studies Reveal the Genetic Basis of Ionomic Variation in Rice. Plant Cell 30:2720-2740

Yang Q, Fan C, Guo Z, Qin J, Wu J, Li Q, Fu T, Zhou Y (2012) Identification of *FAD2* and *FAD3* genes in *Brassica napus* genome and development of allele-specific markers for high oleic and low linolenic acid contents. Theor Appl Genet 125:715-729

Yano K, Yamamoto E, Aya K, Takeuchi H, Lo PC, Hu L, Yamasaki M, Yoshida S, Kitano H, Hirano K, Matsuoka M (2016) Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. Nat Genet 48:927-934

Yao L, Cui X, Liang W, Gao S, Zhao P, Chen Q, Yan J, Li C, Jiang Y, Yang B (2020a) cDNA cloning and expression characterization of WRKY69 gene in oilseed rape (*Brassica napus*). J Agric Biotechnol 28:191-200

Yao M, Guan M, Zhang Z, Zhang Q, Cui Y, Chen H, Liu W, Jan HU, Voss-Fels KP, Werner CR, He X, Liu Z, Guan C, Snowdon RJ, Hua W, Qian L (2020b) GWAS and co-expression network combination uncovers multigenes with close linkage effects on the oleic acid content accumulation in *Brassica napus*. BMC Genomics 21:320

Yin L, Zhang H, Tang Z, Xu J, Yin D, Zhang Z, Yuan X, Zhu M, Zhao S, Li X, Liu X (2020) rMVP: A memory-efficient, visualization-enhanced, and parallel-accelerated tool for genome-wide association study. bioRxiv:2020.2008.2020.258491

Yong HY, Wang C, Bancroft I, Li F, Wu X, Kitashiba H, Nishio T (2015) Identification of a gene controlling variation in the salt tolerance of rapeseed (*Brassica napus* L.). Planta 242:313-326

You FM, Xiao J, Li P, Yao Z, Jia G, He L, Kumar S, Soto-Cerda B, Duguid SD, Booker HM, Rashid KY, Cloutier S (2018a) Genome-wide association study and selection signatures detect genomic regions associated with seed yield and oil quality in flax. Int J Mol Sci 19:2303

You Q, Yang X, Peng Z, Xu L, Wang J (2018b) Development and Applications of a High Throughput Genotyping Tool for Polyploid Crops: Single Nucleotide Polymorphism (SNP) Array. Front Plant Sci 9:104

Yu J, Holland JB, McMullen MD, Buckler ES (2008) Genetic design and statistical power of nested association mapping in maize. Genetics 178:539-551

Yu Z, Chang F, Lv W, Sharmin RA, Wang Z, Kong J, Bhat JA, Zhao T (2019) Identification of QTN and candidate gene for seed-flooding tolerance in soybean [*Glycine max* (L.) Merr.] using genome-wide association study (GWAS). Genes (Basel) 10:957

Žaludová J, Havlíčková L, Jozová E, Kučera V, Vyvadilová M, Klíma M, Čurn V (2013) Marker assisted selection as a tool for detection of *Brassica napus* plants carrying self-incompatibility alleles, in hybrid breeding programs. Rom Agric Res 30:13-22

Zatybekov A, Abugalieva S, Didorenko S, Gerasimova Y, Sidorik I, Anuarbek S, Turuspekov Y (2017) GWAS of agronomic traits in soybean collection included in breeding pool in Kazakhstan. BMC Plant Biol 17:179

Zhang H, Liu W-Z, Zhang Y, Deng M, Niu F, Yang B, Wang X, Wang B, Liang W, Deyholos MK, Jiang Y-Q (2014a) Identification, expression and interaction analyses of calcium-dependent protein kinase (CPK) genes in canola (*Brassica napus* L.). BMC Genomics 15:211

Zhang H, Yang B, Liu WZ, Li H, Wang L, Wang B, Deng M, Liang W, Deyholos MK, Jiang YQ (2014b) Identification and characterization of CBL and CIPK gene families in canola (*Brassica napus* L.). BMC Plant Biol 14:8

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

Zhang J, Song Q, Cregan PB, Jiang GL (2016) Genome-wide association study, genomic prediction and marker-assisted selection for seed weight in soybean (*Glycine max*). Theor Appl Genet 129:117-130

Zhang T, Wu T, Wang L, Jiang B, Zhen C, Yuan S, Hou W, Wu C, Han T, Sun S (2019b) A combined linkage and GWAS analysis identifies QTLs linked to soybean seed protein and oil content. Int J Mol Sci 20:5915

Zhang X, Chen J, Yan Y, Yan X, Shi C, Zhao L, Chen F (2018a) Genome-wide association study of heading and flowering dates and construction of its prediction equation in Chinese common wheat. Theor Appl Genet 131:2271-2285

Zhang Y, Huai D, Yang Q, Cheng Y, Ma M, Kliebenstein DJ, Zhou Y (2015a) Overexpression of three glucosinolate biosynthesis genes in *Brassica napus* identifies enhanced resistance to *Sclerotinia sclerotiorum* and *Botrytis cinerea*. PLoS One 10:e0140491

Zhang Y, Li W, Lin Y, Zhang L, Wang C, Xu R (2018b) Construction of a high-density genetic map and mapping of QTLs for soybean (*Glycine max*) agronomic and seed quality traits by specific length amplified fragment sequencing. BMC Genomics 19:641

Zhang Y, Wan J, He L, Lan H, Li L (2019c) Genome-wide association analysis of plant height using the maize F1 population. Plants (Basel) 8:432

Zhang YH, Liu MF, He JB, Wang YF, Xing GN, Li Y, Yang SP, Zhao TJ, Gai JY (2015b) Marker-assisted breeding for transgressive seed protein content in soybean [*Glycine max* (L.) Merr]. Theor Appl Genet 128:1061-1072

Zhang Z, Ersoz E, Lai CQ, Todhunter RJ, Tiwari HK, Gore MA, Bradbury PJ, Yu J, Arnett DK, Ordovas JM, Buckler ES (2010) Mixed linear model approach adapted for genome-wide association studies. Nat Genet 42:355-360

Zhang Z, Ober U, Erbe M, Zhang H, Gao N, He J, Li J, Simianer H (2014c) Improving the accuracy of whole genome prediction for complex traits using the results of genome wide association studies. PLOS ONE 9:e93017

Zhao G, Zhao Y, Yu X, Kiprotich F, Han H, Guan R, Wang R, Shen W (2018a) Nitric oxide is required for melatonin-enhanced tolerance against salinity stress in rapeseed (*Brassica napus* L.) seedlings. Int J Mol Sci 19:1912

Zhao J, Becker HC, Zhang D, Zhang Y, Ecke W (2006) Conditional QTL mapping of oil content in rapeseed with respect to protein content and traits related to plant development and grain yield. Theor Appl Genet 113:33-38

Zhao J, Meng J (2003) Detection of loci controlling seed glucosinolate content and their association with Sclerotinia resistance in *Brassica napus*. Plant Breed 122:19-23

Zhao Q, Wu J, Cai G, Yang Q, Shahid M, Fan C, Zhang C, Zhou Y (2019a) A novel quantitative trait locus on chromosome A9 controlling oleic acid content in *Brassica napus*. Plant Biotechnol J 17:2313-2324

Zhao X, Dong H, Chang H, Zhao J, Teng W, Qiu L, Li W, Han Y (2019b) Genome wide association mapping and candidate gene analysis for hundred seed weight in soybean [*Glycine max* (L.) Merrill]. BMC Genomics 20:648

Zhao X, Luo L, Cao Y, Liu Y, Li Y, Wu W, Lan Y, Jiang Y, Gao S, Zhang Z, Shen Y, Pan G, Lin H (2018b) Genome-wide association analysis and QTL mapping reveal the genetic control of cadmium accumulation in maize leaf. BMC Genomics 19:91

Zhao Y, Mette MF, Gowda M, Longin CF, Reif JC (2014) Bridging the gap between marker-assisted and genomic selection of heading time and plant height in hybrid wheat. Heredity (Edinb) 112:638-645

Zheng M, Peng C, Liu H, Tang M, Yang H, Li X, Liu J, Sun X, Wang X, Xu J, Hua W, Wang H (2017) Genome-wide association study reveals candidate genes for control of plant height, branch initiation height and branch number in rapeseed (*Brassica napus* L.). Front Plant Sci 8:1246

Zheng X, Koopmann B, Ulber B, von Tiedemann A (2020) A global survey on diseases and pests in oilseed rape—current challenges and innovative strategies of control. Frontiers in Agronomy 2:590908

Zhou Q, Han D, Mason AS, Zhou C, Zheng W, Li Y, Wu C, Fu D, Huang Y (2018) Earliness traits in rapeseed (*Brassica napus*): SNP loci and candidate genes identified by genome-wide association analysis. DNA Res 25:229-244

Zou J, Jiang C, Cao Z, Li R, Long Y, Chen S, Meng J (2010) Association mapping of seed oil content in *Brassica napus* and comparison with quantitative trait loci identified from linkage mapping. Genome 53:908-916

Zou J, Mao L, Qiu J, Wang M, Jia L, Wu D, He Z, Chen M, Shen Y, Shen E, Huang Y, Li R, Hu D, Shi L, Wang K, Zhu Q, Ye C, Bancroft I, King GJ, Meng J, Fan L (2019) Genome-wide selection footprints and deleterious variations in young Asian allotetraploid rapeseed. Plant Biotechnol J 17:1998-2010

Zou J, Zhao Y, Liu P, Shi L, Wang X, Wang M, Meng J, Reif JC (2016) Seed quality traits can be predicted with high accuracy in *Brassica napus* using genomic data. PLoS One 11:e0166624

8.2 Appendices

8.2.1 List of abbreviations

AFLP	Amplified fragment length polymorphism
ANOVA	Analysis of variance
BayesA	Bayesian A regression method
BayesB	Bayesian B regression method
BayesC	Bayesian C regression method
BC	Backcross
BGLR	Bayesian Generalized Linear Regression
BLUP	Best linear unbiased prediction
bp	Base pairs
BRR	Bayesian Ridge Regression method
C.V.	Coefficient of variation
CMLM	Compression mixed linear model
CMS	Cytoplasmic male sterility
CTAB	Cetyl methylammonium bromide
CV	Cross validation
DarT	Diversity arrays technology
DE	Double exponential
DH	Doubled haploid
EB	Empirical Bayesian
EG-BLUP	Extended genomic best linear unbiased prediction
EMMA	Efficient mixed-model association
FarmCPU	Fixed and random model circulating probability unification
FAST-EB-LMM	Fast-empirical Bayesian linear model
Gapit	Genome Association and Prediction Integrated Tool
GBLUP	Genomic best linear unbiased prediction
GBS	Genotyping-by-sequencing
GDD	Growing degree days
GEBV	Genomic estimated breeding value

GO	Gene ontology
GRM	Genomic relationship matrix
GS	Genomic selection
GSL	Glucosinolate
GWAS	Genome-wide association study
HT	Height
INRA	Institut National de la Recherche Agronomique
iPat	Intelligent prediction and association tool
kb	Kilo base pairs
LD	Linkage disequilibrium
MAF	Minor allele frequency
MAS	Marker-assisted selection
Mb	Mega base pairs
ML	Machine learning
MLM	Mixed linear model
MLM+K	Mixed linear models considering kinship
MLM+K+PCA	Mixed linear models considering kinship and subpopulation structure via principal component analysis
MLM+K+Q	Mixed linear models considering kinship and subpopulation structure via Bayesian clustering
MLMM	Multi-locus mixed model
MS	Marker set
Mt	Million tonnes
MTA	Marker-trait association
PCA	Principal component analysis
QTL	Quantitative trait loci
Quantile-Quantile	Q-Q
RAPD	Random amplified polymorphic DNA
RBF	Radial basis function
RCBD	Randomized complete block design
RF	Random forest

RFLP	Restriction fragment length polymorphism
RIL	Recombinant inbred line
RMSE	Root mean square error
rrBLUP	Ridge regression best linear unbiased prediction
SCAR	Sequence characterized amplified region
SNP	Single nucleotide polymorphism
SOC	Seed oil content
SPC	Seed protein content
SSR	Simple sequence repeat, or microsatellites
SVR	Support vector regression
TASSEL	Trait analysis by association, evolution and linkage
TP	Training population/set
VP	Validation population/set
WGR	Whole-genome regression
XGB/XGBoost	Extreme gradient boosting
YLD	Yield

8.2.2 Supplemental tables and figures from Chapter 3

Table S3.1 Summary of locations and number of 362 *Brassica napus* L. hybrid genotypes tested across Alberta, Saskatchewan and Manitoba during 2014-2018.

Site-year #	Year	Site	Number of genotypes evaluated
1	2014	Bison (Winnipeg), MB	34
2	2014	Carman, MB	37
3	2015	Arboretum (Winnipeg), MB	31
4	2015	Bison (Winnipeg), MB	182
5	2015	Carman, MB	182
6	2015	Portage la Prairie, MB	182
7	2016	Arboretum (Winnipeg), MB	35
8	2016	Bison (Winnipeg), MB	198
9	2016	Carman, MB	236
10	2016	Holland, MB	3
11	2016	Killam, AB	3
12	2016	North Battleford, SK	3
13	2016	Pense, SK	3
14	2016	Portage la Prairie, MB	233
15	2016	Rosebank, MB	3
16	2016	Rosetown, SK	3
17	2016	Saint Albert, AB	3
18	2016	Thornhill, MB	3
19	2016	Wawanesa, MB	3
20	2016	Yellow Grass, SK	3
21	2017	Bison (Winnipeg), MB	35
22	2017	Carman, MB	45
23	2017	Holland, MB	10
24	2017	Killam, AB	10
25	2017	Lake Lenore, SK	10
26	2017	Marquis, SK	10
27	2017	Portage la Prairie, MB	35
28	2017	Rosebank, MB	10
29	2017	Rosetown, SK	10
30	2017	Saint Albert, AB	10
31	2017	Thornhill, MB	10
32	2017	Vanscoy, SK	10
33	2017	Watrous, SK	10
34	2018	Carman, MB	15
35	2018	Carstairs, AB	9

36	2018	Killam, AB	1
37	2018	Lake Lenore, SK	9
38	2018	Marquis, SK	9
39	2018	Portage la Prairie, MB	15
40	2018	Rosebank, MB	9
41	2018	Saint Albert, AB	10
42	2018	Vanscoy, SK	10
43	2018	Watrous, SK	10

Table S3.2 List of 436 *Brassica napus* accessions used in genome-wideassociation analysis. There were 61 restorer lines (R), 30 maintainer lines (B) and 345 hybrids (H), totalling 436 accessions.

Genotype #	Name	Line	Paternal Parent	Maternal Parent
1	11DH91	R	RedRiver1997	NR06-24122
2	11DH92	R	RedRiver1997	NR06-24122
3	11DH97	R	RedRiver1997	NR06-24122
4	11DH108	R	RedRiver1997	NR06-24122
5	11DH109	R	RedRiver1997	NR06-24122
6	11DH114	R	RedRiver1997	NR06-24122
7	11DH122	R	RedRiver1997	NR06-24122
8	11DH137	R	RedRiver1997	NR06-24122
9	11DH144	R	RedRiver1997	NR06-24122
10	11DH148	R	RedRiver1997	NR06-24122
11	11DH149	R	RedRiver1997	NR06-24122
12	11DH162	R	RedRiver1997	NR06-24122
13	12DH384	R	Castor	NR07-29768
14	12DH430	R	RedRiver1997	4434
15	12DH478	R	RedRiver1997	71-45
16	12DH915	R	RedRiver1861	NR07-29768
17	12DH949	R	RedRiver1861	NR07-29768
18	14DH1	R	RedRiver1861	NE06-20351
19	14DH3	R	RedRiver1861	NE06-20351
20	14DH4	R	RedRiver1861	NE06-20351
21	14DH5	R	RedRiver1861	NE06-20351
22	14DH7	R	RedRiver1861	NE06-20351
23	14DH9	R	LLHR1074	NE06-20351
24	14DH31	R	LLHR1074	NE06-20351
25	14DH33	R	LLHR1074	NE06-20351
26	14DH35	R	LLHR1074	NE06-20351
27	14DH36	R	LLHR1074	NE06-20351
28	14DH52	R	LLHR1112	NE06-20351
29	14DH53	R	LLHR1112	NE06-20351
30	14DH54	R	LLHR1112	NE06-20351
31	14DH66	R	RedRiver1861	NE06-20351
32	14DH89	R	LLHR1074	NE06-21498
33	14DH90	R	LLHR1074	NE06-21498
34	14DH94	R	LLHR1074	NE06-21498
35	14DH101	R	LLHR1079	NE06-21498
36	14QL370	R	Castor	NR07-29768
37	14QL375	R	Castor	NR07-29768

38	14QL397	R	Castor	NR07-29768
39	14QL434	R	Castor	NR07-29768
40	14QL531	R	Industry	FU27_4Mill03__NE06-20351
41	14QL544	R	Industry	FU27_4Mill03__NE06-20351
42	14QL545	R	Industry	FU27_4Mill03__NE06-20351
43	14QL645	R	Savery	FU27_4Mill03__NE06-20351
44	14QL647	R	Savery	FU27_4Mill03__NE06-20351
45	14R8181	R	RedRiver1861	NR06-24122
46	14R8351	R	RedRiver1861	NR07-29768
47	14R8752-1-1	R	RedRiver1861	4414RR
48	14R8762-1-1	R	RedRiver1861	4414RR
49	ZSDH8135	R	ZSDH2602	FU27_5M03__NE06-20351
50	ZSDH8136	R	ZSDH2602	FU27_5M03__NE06-20351
51	ZSDH8153	R	ZSDH2602	FU27_5M03__NE06-20351
52	ZSDH8173	R	ZSDH2602	FU27_5M03__NE06-20351
53	ZSDH8175	R	ZSDH2631	FU27_5M03__NE06-20380
54	ZSDH8193	R	ZSDH2634	FU27_5M03__NE06-20383
55	ZSDH8196	R	ZSDH2602	FU27_5M03__NE06-20351
56	ZSDH8205	R	ZSDH2602	FU27_5M03__NE06-20351
57	ZSDH8230	R	ZSDH2650	FU27_5M03__NE06-20399
58	ZSDH8238	R	ZSDH2602	FU27_5M03__NE06-20351
59	ZSDH8488	R	ZSDH2602	FU27_5M03__NE06-20351
60	ZSDH8511	R	ZSDH2673	FU27_5M03__NE06-20422
61	Industry	B	Industry	Industry
62	Savery	B	Savery	Savery
63	Mill03	B	Mercury	Cyclone
64	RedRiver1826	B	UM1-73	Castor
65	RedRiver1852	B	UM1-73	Castor

66	RedRiver1861	B	SPBucky	Castor_UM1-73
67	RedRiver1997	B	SPBucky	Mill03_LG3333207-143
68	RRHR204	B	35-25	Mill03
69	RRHR404	B	Kelsey	Mill03
70	RRHR503	B	UM1-73	Castor
71	RRHR5815	B	UM1-73	Castor
72	RRHR9707	B	33-95	RedRiver1826
73	08C702	B	SPBucky	Castor_UM1-73
74	08C712	B	33-95	Mill03_LG3333
75	08C847	B	PR6336	RedRiver1826
76	ZSDH5225	B	Savery	Savery
77	ZSDH5825	B	04R2026_Mill03	ZSDH582514-24_SHEAR69
78	ZSDH6550	B	04R2026_Mill03	ZSDH582514-24_SHEAR69
79	13728	B	RedRiver1826	Canterra1768S
80	13729	B	RedRiver1826	Canterra1768S
81	13738	B	RedRiver1997	30412-B6RR
82	13742	B	RedRiver1997	30412-B6RR
83	13755	B	ZSDH2602	RedRiver1826
84	13774	B	ZSDH2602	RedRiver1852
85	13776	B	ZSDH2602	RedRiver1852
86	13778	B	ZSDH2602	RedRiver1852
87	13785	B	ZSDH2602	RedRiver1852
88	13786	B	ZSDH2602	RedRiver1852
89	13810	B	30216-C7RR	RedRiver1997
90	13831	B	30220-D8RR	RedRiver1997
91	13851	B	30408-C7RR	RedRiver1997
92	12OH1	H	08C702	Industry
93	12OH2	H	08C702	Savery
94	12OH7	H	RedRiver1852	Industry
95	12OH14	H	RRHR503	Mill03
96	13OH55	H	11DH92	Savery
97	13OH56	H	11DH97	Savery
98	13OH57	H	11DH108	Savery
99	13OH58	H	11DH109	Savery
100	13OH59	H	11DH114	Savery
101	13OH60	H	11DH122	Savery
102	13OH61	H	11DH137	Savery
103	13OH62	H	11DH144	Savery

104	13OH63	H	11DH148	Savery
105	13OH64	H	11DH149	Savery
106	13OH68	H	12DH915	Industry
107	13OH69	H	12DH915	Savery
108	13OH71	H	12DH949	Industry
109	13OH72	H	12DH949	Savery
110	13OH75	H	11DH137	RedRiver1826
111	13OH76	H	11DH137	RedRiver1861
112	13OH77	H	11DH137	RedRiver1861
113	13OH78	H	11DH137	RRHR204
114	13OH79	H	11DH137	RRHR404
115	13OH80	H	11DH137	RRHR5815
116	13OH81	H	11DH137	08C702
117	13OH82	H	11DH137	08C712
118	13OH83	H	11DH137	08C847
119	13OH84	H	11DH137	RRHR9707
120	13OH85	H	12DH915	RedRiver1997
121	13OH86	H	12DH915	RedRiver1997
122	13OH87	H	12DH915	RRHR404
123	13OH88	H	12DH915	08C712
124	13OH89	H	12DH949	RedRiver1997
125	13OH90	H	12DH949	RedRiver1997
126	13OH91	H	12DH949	RRHR404
127	13OH92	H	12DH949	08C712
128	13OH93	H	RRHR503	Industry
129	13OH100	H	RedRiver1826	Savery
130	14OH101	H	14DH101	RedRiver1826
131	14OH102	H	14DH101	RedRiver1861
132	14OH103	H	14DH101	RRHR503
133	14OH104	H	14DH101	RRHR404
134	14OH105	H	14DH101	RedRiver1826
135	14OH106	H	14DH101	RedRiver1826
136	14OH107	H	14DH101	RedRiver1826
137	14OH108	H	14DH101	RedRiver1852
138	14OH109	H	14DH101	RedRiver1852
139	14OH112	H	14DH101	RedRiver1861
140	14OH113	H	14DH101	RRHR204
141	14OH115	H	14DH101	08C702
142	14OH116	H	14DH101	08C712
143	14OH117	H	14DH101	08C847

144	14OH118	H	14DH101	RRHR9707
145	14OH119	H	14DH3	RedRiver1826
146	14OH120	H	14DH3	RRHR503
147	14OH121	H	14DH3	RRHR404
148	14OH124	H	14DH5	RRHR404
149	14OH125	H	14DH31	RedRiver1826
150	14OH126	H	14DH31	RedRiver1861
151	14OH127	H	14DH31	RRHR503
152	14OH128	H	14DH31	RRHR404
153	14OH130	H	14DH52	RedRiver1861
154	14OH131	H	14DH52	RRHR503
155	14OH132	H	14DH52	RRHR404
156	14OH133	H	14DH53	RedRiver1826
157	14OH135	H	14DH53	RRHR503
158	14OH136	H	14DH53	RRHR404
159	14OH137	H	14DH54	RedRiver1826
160	14OH139	H	14DH54	RRHR503
161	14OH140	H	14DH54	RRHR404
162	14OH141	H	14DH90	RedRiver1826
163	14OH142	H	14DH90	RedRiver1861
164	14OH143	H	14DH90	RRHR503
165	14OH144	H	14DH90	RRHR404
166	14OH145	H	14QL370	RedRiver1826
167	14OH146	H	14QL370	RedRiver1861
168	14OH147	H	14QL370	RRHR503
169	14OH148	H	14QL370	RRHR404
170	14OH149	H	14QL375	RedRiver1826
171	14OH150	H	14QL375	RedRiver1861
172	14OH151	H	14QL375	RRHR503
173	14OH152	H	14QL375	RRHR404
174	14OH153	H	14QL397	RedRiver1826
175	14OH154	H	14QL397	RedRiver1861
176	14OH155	H	14QL397	RRHR503
177	14OH156	H	14QL397	RRHR404
178	14OH157	H	14QL434	RedRiver1826
179	14OH158	H	14QL434	RedRiver1861
180	14OH159	H	14QL434	RRHR503
181	14OH160	H	14QL434	RRHR404
182	14OH161	H	14QL531	RedRiver1826
183	14OH162	H	14QL531	RedRiver1861

184	14OH163	H	14QL531	RRHR503
185	14OH164	H	14QL531	RRHR404
186	14OH165	H	14QL545	RedRiver1826
187	14OH166	H	14QL545	RedRiver1861
188	14OH168	H	14QL545	RRHR404
189	14OH169	H	14QL645	RedRiver1826
190	14OH170	H	14QL645	RedRiver1861
191	14OH171	H	14QL645	RRHR503
192	14OH172	H	14QL645	RRHR404
193	14OH173	H	14QL647	RedRiver1826
194	14OH174	H	14QL647	RedRiver1861
195	14OH175	H	14QL647	RRHR503
196	14OH176	H	14QL647	RRHR404
197	14OH177	H	12DH378	RedRiver1826
198	14OH178	H	12DH378	RedRiver1861
199	14OH179	H	12DH378	RRHR503
200	14OH180	H	12DH378	RRHR404
201	14OH181	H	14DH4	RedRiver1826
202	14OH182	H	14DH4	RRHR503
203	14OH183	H	14DH4	RRHR404
204	14OH184	H	14DH4	RedRiver1826
205	14OH185	H	14DH4	RedRiver1826
206	14OH186	H	14DH4	RedRiver1826
207	14OH187	H	14DH4	RedRiver1852
208	14OH188	H	14DH4	RedRiver1852
209	14OH189	H	14DH4	RedRiver1997
210	14OH190	H	14DH4	RedRiver1997
211	14OH191	H	14DH4	RRHR204
212	14OH192	H	14DH4	RRHR5815
213	14OH193	H	14DH4	08C702
214	14OH194	H	14DH4	08C712
215	14OH195	H	14DH4	08C847
216	14OH196	H	14DH4	RRHR9707
217	14OH197	H	14DH1	RedRiver1826
218	14OH198	H	14DH1	RRHR503
219	14OH199	H	14DH1	RRHR404
220	14OH200	H	14DH7	RedRiver1826
221	14OH201	H	14DH7	RRHR503
222	14OH202	H	14DH7	RRHR404
223	14OH203	H	14DH9	RedRiver1826

224	14OH204	H	14DH9	RRHR503
225	14OH205	H	14DH9	RRHR404
226	14OH206	H	14DH66	RedRiver1826
227	14OH207	H	14DH66	RRHR503
228	14OH208	H	14DH66	RRHR404
229	14OH209	H	12DH384	RedRiver1826
230	14OH210	H	12DH384	RedRiver1861
231	14OH211	H	12DH384	RRHR503
232	14OH212	H	12DH384	RRHR404
233	14OH213	H	12DH430	RedRiver1826
234	14OH214	H	12DH430	RedRiver1861
235	14OH215	H	12DH430	RRHR503
236	14OH216	H	12DH430	RRHR404
237	14OH217	H	12DH478	RedRiver1826
238	14OH218	H	12DH478	RedRiver1861
239	14OH219	H	12DH478	RRHR503
240	14OH220	H	12DH478	RRHR404
241	14OH221	H	12DH915	RedRiver1826
242	14OH222	H	12DH915	RRHR503
243	14OH223	H	12DH949	RedRiver1826
244	14OH224	H	12DH949	RRHR503
245	14OH225	H	14DH4	Mill03
246	14OH226	H	14DH4	Savery_ZSDH5225
247	14OH227	H	14DH4	Savery_ZSDH5825
248	14OH228	H	14DH4	Savery_ZSDH6550
249	14OH229	H	14DH1	Mill03
250	14OH230	H	14DH1	Savery_ZSDH5225
251	14OH231	H	14DH1	Savery_ZSDH5825
252	14OH232	H	14DH1	Savery_ZSDH6550
253	14OH233	H	14DH7	Mill03
254	14OH234	H	14DH7	Savery_ZSDH5225
255	14OH235	H	14DH7	Savery_ZSDH5825
256	14OH236	H	14DH7	Savery_ZSDH6550
257	14OH237	H	14DH9	Mill03
258	14OH238	H	14DH9	Savery_ZSDH5225
259	14OH239	H	14DH9	Savery_ZSDH5825
260	14OH243	H	14DH66	Savery_ZSDH5225
261	14OH246	H	14DH66	Savery_ZSDH5825
262	14OH247	H	12DH384	Savery_ZSDH5225
263	14OH250	H	12DH430	Savery_ZSDH5225

264	14OH167	H	14QL545	RRHR503
265	14OH254	H	12DH478	Savery_ZSDH5225
266	14OH255	H	12DH478	Savery_ZSDH5825
267	14OH257	H	12DH915	Mill03
268	14OH258	H	12DH915	Savery_ZSDH5225
269	14OH259	H	12DH915	Savery_ZSDH5825
270	14OH260	H	12DH915	Savery_ZSDH6550
271	14OH261	H	12DH949	Mill03
272	14OH262	H	12DH949	Savery_ZSDH5225
273	14OH263	H	12DH949	Savery_ZSDH5825
274	15OH265	H	14DH3	RRHR5815
275	15OH266	H	14DH3	08C712
276	15OH267	H	14DH3	08C847
277	15OH268	H	14DH3	RRHR9707
278	15OH269	H	14DH5	RRHR5815
279	15OH271	H	14DH5	08C847
280	15OH272	H	14DH5	08C847
281	15OH273	H	14DH31	RedRiver1861
282	15OH274	H	14DH31	RRHR5815
283	15OH275	H	14DH31	08C702
284	15OH276	H	14DH31	08C712
285	15OH277	H	14DH31	08C847
286	15OH278	H	14DH31	RRHR9707
287	15OH279	H	14DH52	RedRiver1861
288	15OH280	H	14DH52	RRHR5815
289	15OH281	H	14DH52	08C702
290	15OH282	H	14DH52	08C712
291	15OH283	H	14DH52	08C847
292	15OH284	H	14DH52	RRHR9707
293	15OH285	H	14DH53	RedRiver1861
294	15OH286	H	14DH53	RRHR5815
295	15OH287	H	14DH53	08C702
296	15OH288	H	14DH53	08C712
297	15OH290	H	14DH53	RRHR9707
298	15OH291	H	14DH54	RedRiver1861
299	15OH292	H	14DH54	RRHR5815
300	15OH293	H	14DH54	08C702
301	15OH294	H	14DH54	08C712
302	15OH296	H	14DH54	RRHR9707
303	15OH297	H	14QL545	RedRiver1861

304	15OH298	H	14QL545	RRHR5815
305	15OH299	H	14QL545	08C702
306	15OH300	H	14QL545	08C712
307	15OH301	H	14QL545	08C847
308	15OH302	H	14QL545	RRHR9707
309	15OH303	H	11DH137	Mill03
310	15OH305	H	11DH137	Savery_ZSDH5825
311	15OH306	H	11DH137	Savery_ZSDH6550
312	15OH307	H	12DH478	RedRiver1861
313	15OH308	H	12DH478	RRHR5815
314	15OH309	H	12DH478	08C702
315	15OH310	H	12DH478	08C712
316	15OH311	H	12DH478	08C847
317	15OH312	H	12DH478	RRHR9707
318	15OH314	H	12DH915	RRHR9707
319	15OH316	H	12DH949	RRHR9707
320	15OH317	H	14DH1	RRHR5815
321	15OH318	H	14DH1	08C712
322	15OH319	H	14DH1	08C847
323	15OH320	H	14DH1	RRHR9707
324	15OH321	H	14DH7	RRHR5815
325	15OH322	H	14DH7	08C712
326	15OH323	H	14DH7	08C847
327	15OH324	H	14DH7	RRHR9707
328	15OH325	H	14DH9	RRHR5815
329	15OH326	H	14DH9	08C712
330	15OH327	H	14DH9	08C847
331	15OH328	H	14DH9	RRHR9707
332	15OH329	H	14DH66	RRHR5815
333	15OH330	H	14DH66	08C712
334	15OH331	H	14DH66	08C847
335	15OH332	H	14DH66	RRHR9707
336	15OH333	H	14QL544	RedRiver1861
337	15OH334	H	14QL544	RedRiver1861
338	15OH335	H	14QL544	RRHR503
339	15OH336	H	14QL544	RRHR5815
340	15OH337	H	14QL544	08C702
341	15OH338	H	14QL544	08C712
342	15OH339	H	14QL544	08C847
343	15OH340	H	14QL544	RRHR9707

344	15OH342	H	14QL547	RedRiver1861
345	15OH343	H	14QL547	RRHR503
346	15OH344	H	14QL547	RRHR5815
347	15OH345	H	14QL547	08C702
348	15OH346	H	14QL547	08C712
349	15OH347	H	14QL547	08C847
350	15OH348	H	14QL547	RRHR9707
351	15OH349	H	14DH33	RedRiver1861
352	15OH350	H	14DH33	RedRiver1861
353	15OH351	H	14DH33	RRHR503
354	15OH352	H	14DH33	RRHR5815
355	15OH353	H	14DH33	08C702
356	15OH354	H	14DH33	08C712
357	15OH355	H	14DH33	08C847
358	15OH356	H	14DH33	RRHR9707
359	15OH357	H	14DH35	RedRiver1861
360	15OH358	H	14DH35	RedRiver1861
361	15OH359	H	14DH35	RRHR503
362	15OH360	H	14DH35	RRHR5815
363	15OH361	H	14DH35	08C702
364	15OH362	H	14DH35	08C712
365	15OH363	H	14DH35	08C847
366	15OH364	H	14DH35	RRHR9707
367	15OH365	H	14DH36	RedRiver1861
368	15OH366	H	14DH36	RedRiver1861
369	15OH367	H	14DH36	RRHR503
370	15OH368	H	14DH36	RRHR5815
371	15OH369	H	14DH36	08C702
372	15OH370	H	14DH36	08C712
373	15OH371	H	14DH36	08C847
374	15OH372	H	14DH36	RRHR9707
375	15OH373	H	14DH89	RedRiver1861
376	15OH374	H	14DH89	RedRiver1861
377	15OH375	H	14DH89	RRHR503
378	15OH376	H	14DH89	RRHR5815
379	15OH377	H	14DH89	08C702
380	15OH378	H	14DH89	08C712
381	15OH379	H	14DH89	08C847
382	15OH380	H	14DH89	RRHR9707
383	15OH381	H	14DH94	RedRiver1861

384	15OH382	H	14DH94	RedRiver1861
385	15OH383	H	14DH94	RRHR503
386	15OH384	H	14DH94	RRHR5815
387	15OH385	H	14DH94	08C702
388	15OH386	H	14DH94	08C712
389	15OH387	H	14DH94	08C847
390	15OH388	H	14DH94	RRHR9707
391	15OH389	H	14R8181	Mill03
392	15OH391	H	14R8181	Savery
393	15OH393	H	14R8181	Savery_ZSDH5825
394	15OH394	H	14R8181	Savery_ZSDH6550
395	15OH395	H	14R8181	RRHR503
396	15OH396	H	14R8181	08C712
397	15OH397	H	14R8181	08C847
398	15OH399	H	14R8351	Mill03
399	15OH401	H	14R8351	Savery
400	15OH403	H	14R8351	Savery_ZSDH5825
401	15OH404	H	14R8351	Savery_ZSDH6550
402	15OH405	H	14R8351	RRHR503
403	15OH406	H	14R8351	08C712
404	15OH407	H	14R8351	08C847
405	15OH408	H	14R8351	RRHR9707
406	15OH409	H	14R8712	Mill03
407	15OH410	H	14R8712	Industry
408	15OH411	H	14R8712	Savery
409	15OH413	H	14R8712	Savery_ZSDH5825
410	15OH414	H	14R8712	Savery_ZSDH6550
411	15OH415	H	14R8712	RRHR503
412	15OH416	H	14R8712	08C712
413	15OH417	H	14R8712	08C847
414	15OH418	H	14R8712	RRHR9707
415	15OH419	H	14R8793	Mill03
416	15OH421	H	14R8793	Savery
417	15OH423	H	14R8793	Savery_ZSDH5825
418	15OH424	H	14R8793	Savery_ZSDH6550
419	15OH425	H	14R8793	RRHR503
420	15OH426	H	14R8793	08C712
421	15OH427	H	14R8793	08C847
422	15OH428	H	14R8793	RRHR9707
423	15OH429	H	14R8762-1-1	Mill03

424	15OH431	H	14R8762-1-1	Savery
425	15OH433	H	14R8762-1-1	Savery_ZSDH5825
426	15OH434	H	14R8762-1-1	Savery_ZSDH6550
427	15OH435	H	14R8762-1-1	RRHR503
428	15OH436	H	14R8762-1-1	08C712
429	15OH437	H	14R8762-1-1	08C847
430	15OH438	H	14R8762-1-1	RRHR9707
431	15OH439	H	14R8752-1-1	Mill03
432	15OH441	H	14R8752-1-1	Savery
433	15OH443	H	14R8752-1-1	Savery_ZSDH5825
434	15OH444	H	14R8752-1-1	Savery_ZSDH6550
435	15OH446	H	14R8752-1-1	08C712
436	15OH448	H	14R8752-1-1	RRHR9707

Table S3.3 A summary of raw phenotype data of the *Brassica napus* L. parental genotypes. Data were collected from five site-years across southern Manitoba: Glenlea 2016, Carman 2017, Portage la Prairie 2017, Glenlea 2018 and Portage la Prairie 2018.

Site-year	Trait	Min	Max	SD	Mean	C.V.[1] (%)
Portage 2018	YLD[2]	121.0	2370.0	401.9	1062.2	37.8
	HT[3]	50.0	113.0	11.3	87.0	13.0
	SPC[4]	25.9	34.6	1.7	30.6	5.7
	SOC[5]	34.8	49.9	2.8	43.3	6.6
	GSL[6]	8.9	50.5	6.2	19.1	32.6
Glenlea 2018	YLD	53.0	1610.0	289.0	686.8	42.1
	HT	52.0	93.0	8.2	70.5	11.6
	SPC	24.2	35.6	2.2	31.4	7.1
	SOC	36.1	49.9	3.3	42.6	7.6
	GSL	8.7	47.2	6.7	20.9	31.9
Portage 2017	YLD	108.0	2704.0	445.3	1640.0	27.2
	HT	78.0	150.0	14.8	113.7	13.0
	SPC	22.5	36.8	2.2	27.8	7.5
	SOC	36.6	52.3	2.4	45.2	5.4
	GSL	3.1	34.1	5.7	16.5	34.7
Carman 2017	YLD	77.0	2356.0	405.9	1115.3	36.4
	HT	57.0	133.0	13.7	94.6	14.5
	SPC	24.1	32.4	1.8	28.5	6.3
	SOC	35.3	51.8	2.7	44.4	612
	GSL	6.0	37.0	5.4	18.3	29.3
Glenlea 2016	YLD	100.2	3086.2	528.8	1095.3	48.3
	HT	57.5	120.0	10.8	98.7	11.0
	SPC	25.9	37.7	2.1	31.5	6.72
	SOC	32.7	48.7	3.1	41.7	7.4
	GSL	5.3	46.3	5.9	18.2	32.2

[1]C.V. values represent the population variation of a certain trait ($CV_{population} = \frac{SD_{population}}{mean_{population}}$).

[2] Seed yield (kg ha^{-1}).

[3] Plant height (cm).

[4] Seed protein content (%).

[5] Seed oil content (%).

[6] Seed glucosinolates content (μmol g^{-1}).

Table S3.4 Significant MTAs identified based on the *Brassica napus* L. parental population based on MS-1 (26,651 SNP markers) on all five traits: seed yield (YLD), plant height (HT), seed protein content (SPC), seed oil content (SOC) and seed glucosinolate content (GSL).

Trait	SNP	Model	Chromosome	Position (bp)
YLD	Bn-Scaffold000253-p27839	FarmCPU	A1	826,013
	Bn-A09-p264743	FarmCPU, MLMM	A9	429,423
HT	Bn-A09-p264743	FarmCPU	A9	429,423
	Bn-scaff_17441_3-p31296	FarmCPU	C5	40,328,601
	Bn-scaff_16510_1-p242507	FarmCPU	C6	14,081,295
	Bn-A09-p264743	MLMM	A9	429,423
SPC	Bn-A09-p264743	CMLM	A9	429,423
	Bn-scaff_15712_5-p649917	FarmCPU	A2	20,655,454
	Bn-A03-p25949676	FarmCPU	A3	24,308,492
	Bn-A08-p13303525	FarmCPU	A8	1,664,920
	Bn-A09-p264743	FarmCPU, MLM+K, MLM+K+PCA, MLM+K+Q, MLMM	A9	429,423
SOC	Bn-A09-p264743	FarmCPU, CMLM	A9	429,423
	Bn-A09-p264743	CMLM	A9	429,423
	Bn-A10-p2535072	FarmCPU	A10	13,530
	Bn-scaff_15712_13-p63324	MLM+K	A2	4,500,718
	Bn-A09-p264743	MLM+K	A9	429,423
	Bn-scaff_18514_1-p28001	MLM+K	C2	7,925,571
	Bn-scaff_15712_13-p38138	MLM+K	C2	8,302,329
	Bn-scaff_15712_13-p43168	MLM+K	C2	8,307,658
	Bn-A09-p264743	MLM+K+PCA	A9	429,423
	Bn-scaff_18514_1-p28001	MLM+K+PCA	C2	7,925,571
	Bn-scaff_15712_13-p38138	MLM+K+PCA	C2	8,302,329
	Bn-scaff_15712_13-p43168	MLM+K+PCA	C2	8,307,658
	Bn-scaff_15712_13-p63324	MLM+K+Q	A2	4,500,718
	Bn-A09-p264743	MLM+K+Q	A9	429,423
	Bn-scaff_18514_1-p28001	MLM+K+Q	C2	7,925,571
	Bn-scaff_15712_13-p38138	MLM+K+Q	C2	8,302,329

	Bn-scaff_15712_13-p43168	MLM+K+Q	C2	8,307,658
	Bn-A09-p264743	MLMM	A9	429,423
GSL	Bn-A05-p23873413	CMLM, FarmCPU	A5	22,850,229
	Bn-A08-p7496720	FarmCPU	A8	6,503,838
	Bn-scaff_16002_1-p2298646	FarmCPU	C3	12,092,758
	Bn-scaff_16361_1-p2115007	FarmCPU	C8	29,655,788
	Bn-A05-p23873413	MLMM	A5	22,850,229

Abbreviations: SNP: single nucleotide polymorphism; CMLM: compression mixed linear model; FarmCPU: fixed and random model circulating probability unification; MLM: mixed linear model; K: kinship matrix; Q: population structure matrix based on Bayesian clustering; PCA: principal component analysis; MLMM: multi-locus mixed linear model.

Table S3.5 Significant MTAs identified based on the *Brassica napus* L. combined population based on MS-1 (26,651 SNP markers) on all five traits: seed yield (YLD), plant height (HT), seed protein content (SPC), seed oil content (SOC) and seed glucosinolate content (GSL).

Trait	SNP	Model	Chromosome	Position (bp)
YLD	Bn-A02-p6751482	CMLM	A2	188,202
	Bn-A02-p7351109	CMLM	A2	270,573
	Bn-A02-p2851231	CMLM	A2	393,886
	Bn-A02-p7004091	CMLM	A2	4,037,640
	Bn-scaff_15712_13-p63324	CMLM	A2	4,500,718
	Bn-A02-p16666866	CMLM	A2	13,863,707
	Bn-A02-p23836232	CMLM	A2	22,068,630
	Bn-A03-p19501529	CMLM	A3	18,476,014
	Bn-A04-p6389136	CMLM	A4	7,593,915
	Bn-A05-p2143241	CMLM	A4	18,442,758
	Bn-A05-p6405021	CMLM	A5	5,956,466
	Bn-A05-p6405336	CMLM	A5	5,956,781
	Bn-A02-p27262588	CMLM	A6	487,133
	Bn-A06-p1874196	CMLM	A6	1,839,761
	Bn-A06-p19419215	CMLM	A6	15,432,566
	Bn-A07-p9115370	CMLM	A7	10,581,719
	Bn-A08-p10147092	CMLM	A8	1,382,860
	Bn-A08-p2585836	CMLM	A8	2,022,032
	Bn-A08-p4146148	CMLM	A8	3,547,924
	Bn-A08-p4537897	CMLM	A8	3,956,547
	Bn-A08-p13214314	CMLM	A8	10,959,702
	Bn-A08-p15945454	CMLM	A8	13,407,212
	Bn-A09-p35656352	CMLM	A9	32,788,001
	Bn-A10-p8409315	CMLM	A10	9,898,825
	Bn-A10-p11268601	CMLM	A10	12,562,289
	Bn-A09-p5267535	CMLM	A10	15,656,938
	Bn-A10-p17125825	CMLM	C1	32,941,300
	Bn-scaff_15714_1-p346291	CMLM	C2	4,012,095
	Bn-scaff_16269_1-p192431	CMLM	C2	6,421,917
	Bn-scaff_18374_1-p23965	CMLM	C2	7,565,729
	Bn-scaff_16804_4-p112136	CMLM	C2	7,610,450
	Bn-scaff_15712_5-p1021105	CMLM	C2	8,512,187
	Bn-scaff_21312_1-p895326	CMLM	C3	9,539,709
	Bn-scaff_16534_1-p259614	CMLM	C4	854,745
	Bn-scaff_15695_2-p413042	CMLM	C5	30,679,838
	Bn-scaff_16485_1-p630261	CMLM	C6	3,538,459

	Bn-A08-p7355372	CMLM	C8	9,556,938
	Bn-A02-p22443535	FarmCPU	A2	20,723,010
	Bn-A04-p6389136	FarmCPU	A4	7,593,915
	Bn-scaff_15712_13-p43168	FarmCPU	C2	8,307,658
	Bn-A10-p4553957	FarmCPU	C5	369,884
	Bn-scaff_23408_1-p96976	FarmCPU	C5	37,805,713
	Bn-scaff_20619_1-p169094	FarmCPU	C9	31,677,194
	Bn-A04-p6389136	MLMM	A4	7,593,915
	Bn-A09-p35656352	MLMM	A9	32,788,001
	Bn-scaff_16269_1-p192431	MLMM	C2	6,421,917
	Bn-scaff_17048_1-p243085	MLMM	C9	1,985,336
	Bn-scaff_16246_2-p3090	MLMM	C9	2,856,998
HT	Bn-A02-p6751482	CMLM	A2	188,202
	Bn-A02-p7004091	CMLM	A2	4,037,640
	Bn-scaff_15712_13-p63324	CMLM	A2	4,500,718
	Bn-A04-p14023419	CMLM	A4	1,021,817
	Bn-A04-p6389136	CMLM	A4	7,593,915
	Bn-A05-p2143241	CMLM	A4	18,442,758
	Bn-A05-p5058143	CMLM	A5	4,877,416
	Bn-A05-p6405021	CMLM	A5	5,956,466
	Bn-A05-p6405336	CMLM	A5	5,956,781
	Bn-A06-p1874196	CMLM	A6	1,839,761
	Bn-A06-p2365869	CMLM	A6	2,381,259
	Bn-A06-p19419215	CMLM	A6	15,432,566
	Bn-A07-p9115370	CMLM	A7	10,581,719
	Bn-A08-p10147092	CMLM	A8	1,382,860
	Bn-A08-p2585836	CMLM	A8	2,022,032
	Bn-A08-p4146148	CMLM	A8	3,547,924
	Bn-A08-p4207231	CMLM	A8	3,626,293
	Bn-A08-p4452888	CMLM	A8	3,865,227
	Bn-A08-p4537897	CMLM	A8	3,956,547
	Bn-A09-p264743	CMLM	A9	429,423
	Bn-A10-p8409315	CMLM	A10	9,898,825
	Bn-A10-p11268601	CMLM	A10	12,562,289
	Bn-A10-p11376164	CMLM	A10	12,679,162
	Bn-scaff_16553_1-p34303	CMLM	C1	4,026,215
	Bn-scaff_15714_1-p357596	CMLM	C2	4,003,578
	Bn-scaff_15714_1-p346291	CMLM	C2	4,012,095
	Bn-scaff_16269_1-p192431	CMLM	C2	6,421,917
	Bn-scaff_18374_1-p23965	CMLM	C2	7,565,729

	Bn-scaff_16804_4-p112136	CMLM	C2	7,610,450
	Bn-scaff_15712_5-p1021105	CMLM	C2	8,512,187
	Bn-scaff_18936_1-p744477	CMLM	C3	3,325,084
	Bn-scaff_16996_1-p230771	CMLM	C3	56,423,745
	Bn-scaff_16534_1-p259614	CMLM	C4	854,745
	Bn-scaff_19253_1-p285404	CMLM	C4	15,292,281
	Bn-scaff_17580_1-p281057	CMLM	C7	13,974,728
	Bn-A05-p1107888	FarmCPU	A5	1,243,846
	Bn-A05-p6405336	FarmCPU	A5	5,956,781
	Bn-A07-p4097388	FarmCPU	A7	6,014,828
	Bn-A07-p9115370	FarmCPU	A7	10,581,719
	Bn-A10-p10847605	FarmCPU	A8	12,590,859
	Bn-A09-p26532777	FarmCPU	A9	24,583,494
	Bn-scaff_15712_5-p1021105	FarmCPU	C2	8,512,187
	Bn-scaff_18505_1-p288572	FarmCPU	C4	14,546,495
	Bn-scaff_16197_1-p2958698	FarmCPU	C8	31,316,266
	Bn-scaff_20619_1-p167640	FarmCPU	C9	31,678,632
	Bn-scaff_17750_1-p1839429	FarmCPU	C9	46,646,099
	Bn-A05-p6405336	MLMM	A5	5,956,781
	Bn-A07-p9115370	MLMM	A7	10,581,719
	Bn-scaff_19614_1-p53484	MLMM	C1	13,527,850
SPC	Bn-A04-p6389136	CMLM	A4	7,593,915
	Bn-A02-p26975795	CMLM	A5	3,004,035
	Bn-A05-p11689852	CMLM	A5	10,221,712
	Bn-A05-p21081145	CMLM	A5	19,232,253
	Bn-A02-p23280472	CMLM	A5	21,444,314
	Bn-A05-p11702049	CMLM	A8	88,915
	Bn-A09-p8635608	CMLM	A9	8,062,605
	Bn-scaff_15695_2-p413042	CMLM	C5	30,679,838
	Bn-scaff_17801_1-p220808	CMLM	C9	15,322,388
	Bn-A02-p8169424	FarmCPU	A2	5,160,109
	Bn-A05-p11689852	FarmCPU	A5	10,221,712
	Bn-scaff_26642_1-p55504	FarmCPU	C3	52,941,890
	Bn-scaff_16069_1-p1204477	MLMM	A3	22,019,044
	Bn-scaff_15695_2-p413042	MLMM	C5	30,679,838
SOC	Bn-A02-p3121742	FarmCPU	A2	605,811
	Bn-A02-p23280472	FarmCPU	A5	21,444,314
	Bn-A07-p22398500	FarmCPU	A7	23,796,382
	Bn-A01-p8821722	FarmCPU	A9	1,770,199
	Bn-scaff_27039_1-p405965	FarmCPU	C2	1,452,154

	Bn-scaff_20942_1-p432654	FarmCPU	C2	10,800,822
	Bn-scaff_23761_1-p736623	FarmCPU	C3	6,000,302
	Bn-scaff_18602_1-p263207	FarmCPU	C3	51,660,331
	Bn-scaffold5411-p214	FarmCPU	C5	4,957,352
	Bn-scaff_20219_1-p200426	FarmCPU	C5	40,827,867
	Bn-scaff_17801_1-p220808	FarmCPU	C9	15,322,388
	Bn-scaff_19899_1-p356624	FarmCPU	C9	40,719,329
	Bn-scaff_15695_2-p413042	MLMM	C5	30,679,838
GSL	Bn-scaff_15712_13-p63324	CMLM	A2	4,500,718
	Bn-scaff_23957_1-p127219	CMLM	A7	20,022,579
	Bn-A09-p264743	CMLM	A9	429,423
	Bn-scaff_15714_1-p346291	CMLM	C2	4,012,095
	Bn-scaff_16269_1-p192431	CMLM	C2	6,421,917
	Bn-scaff_16269_1-p57060	CMLM	C2	6,574,199
	Bn-scaff_16804_1-p635989	CMLM	C2	7,408,306
	Bn-scaff_18374_1-p23965	CMLM	C2	7,565,729
	Bn-scaff_16804_4-p107537	CMLM	C2	7,605,785
	Bn-scaff_16804_4-p111688	CMLM	C2	7,610,005
	Bn-scaff_16804_4-p112136	CMLM	C2	7,610,450
	Bn-scaff_18514_1-p28001	CMLM	C2	7,925,571
	Bn-scaff_15712_13-p38138	CMLM	C2	8,302,329
	Bn-scaff_15712_13-p43168	CMLM	C2	8,307,658
	Bn-scaff_15712_5-p1021105	CMLM	C2	8,512,187
	Bn-scaff_16414_1-p863592	CMLM	C5	1,091,262
	Bn-scaff_22728_1-p947436	CMLM	C5	6,146,830
	Bn-scaff_15746_1-p176373	CMLM	C6	23,292,130
	Bn-scaff_17580_1-p281057	CMLM	C7	13,974,728
	Bn-A01-p5567077	FarmCPU	A1	322,699
	Bn-A02-p11248742	FarmCPU	A2	8,255,799
	Bn-A06-p5741557	FarmCPU	A6	5,182,655
	Bn-A06-p10297153	FarmCPU	A6	9,686,244
	Bn-A07-p1450961	FarmCPU	A7	210,238
	Bn-A07-p132150	FarmCPU	A7	323,068
	Bn-scaff_23957_1-p127219	FarmCPU	A7	20,022,579
	Bn-A10-p2311797	FarmCPU	A10	1,540,475
	Bn-scaff_21778_1-p262139	FarmCPU	C3	5,054,038
	Bn-scaff_16394_1-p73139	FarmCPU	C4	31,472,468
	Bn-scaff_20270_1-p1013122	FarmCPU	C4	47,080,450
	Bn-scaff_22728_1-p947436	FarmCPU	C5	6,146,830
	Bn-scaff_15818_2-p1146207	FarmCPU	C6	17,529,142

Bn-scaff_20619_1-p167640	FarmCPU	C9	31,678,632
Bn-scaff_15712_13-p63324	MLMM	A2	4,500,718
Bn-scaff_17731_1-p166950	MLMM	C1	114,921
Bn-scaff_16092_1-p494537	MLMM	C3	28,173,069
Bn-scaff_22728_1-p947436	MLMM	C5	6,146,830
Bn-scaff_17910_1-p132433	MLMM	C9	33,581,091

Abbreviations: SNP: single nucleotide polymorphism; CMLM: compression mixed linear model; FarmCPU: fixed and random model circulating probability unification; MLMM: multi-locus mixed linear model.

Table S3.6 Significant MTAs identified based on the *Brassica napus* L. parental population using MS-2 (16,855 SNP markers) on all five traits: seed yield (YLD), plant height (HT), seed protein content (SPC), seed oil content (SOC) and seed glucosinolate content (GSL).

Trait	SNP	Model	Chromosome	Position
YLD	Bn-A09-p264743	MLMM	A9	429,423
	Bn-scaff_15712_5-p332728	MLMM	C1	13,638,163
HT	Bn-A09-p264743	FarmCPU, MLMM	A9	429,423
	Bn-scaff_17441_3-p31296	MLMM	C5	40,328,601
	Bn-scaff_16510_1-p88138	MLMM	C6	1,437,302
SPC	Bn-A09-p264743	CLMM	A9	429,423
	Bn-A08-p13303525	FarmCPU	A8	1,664,920
	Bn-A09-p264743	FarmCPU	A9	429,423
	Bn-scaff_15712_2-p767471	FarmCPU	C2	38,977,697
	Bn-A09-p264743	MLM+K, MLM+L+PCA, MLM+L+Q, MLMM	A9	429,423
SOC	Bn-A09-p264743	CMLM	A9	429,423
	Bn-A10-p2535072	FarmCPU	A10	13,530
	Bn-A06-p2268393	FarmCPU	A6	2,274,725
	Bn-scaff_18206_3-p536950	FarmCPU	A7	14,459,805
	Bn-A09-p264743	FarmCPU	A9	429,423
	Bn-scaff_15712_13-p63324	MLM+K	A2	4,500,718
	Bn-A09-p264743	MLM+K	A9	429,423
	Bn-scaff_18514_1-p28001	MLM+K	C2	7,925,571
	Bn-scaff_15712_13-p38138	MLM+K	C2	8,302,329
	Bn-scaff_15712_13-p43168	MLM+K	C2	8,307,658
	Bn-A09-p264743	MLM+K+PCA	A9	429,423
	Bn-scaff_15712_13-p63324	MLM+K+Q	A2	4,500,718
	Bn-A09-p264743	MLM+K+Q	A9	429,423
	Bn-scaff_18514_1-p28001	MLM+K+Q	C2	7,925,571
	Bn-scaff_15712_13-p38138	MLM+K+Q	C2	8,302,329
	Bn-scaff_15712_13-p43168	MLM+K+Q	C2	8,307,658

	Bn-A09-p264743	MLMM	A9	429,423
GSL	Bn-A05-p23873413	CMLM	A5	22,850,229
	Bn-A05-p23873413	FarmCPU	A5	22,850,229
	Bn-A08-p7496720	FarmCPU	A8	6,503,838
	Bn-scaff_18514_1-p28001	FarmCPU	C2	7,925,571
	Bn-scaff_16002_1-p2298646	FarmCPU	C3	12,092,758
	Bn-scaff_16361_1-p2115007	FarmCPU	C8	29,655,788
GSL	Bn-A07-p1227140	MLM+K+Q	A7	861,713
	Bn-scaff_15838_5-p886564	MLM+K+Q	C1	3,719,679
GSL	Bn-scaff_19523_1-p28969	MLM+K+Q	C3	18,526,881
GSL	Bn-A05-p23873413	MLMM	A5	22,850,229

Abbreviations: SNP: single nucleotide polymorphism; CMLM: compression mixed linear model; FarmCPU: fixed and random model circulating probability unification; MLM: mixed linear model; K: kinship matrix; Q: population structure matrix based on Bayesian clustering; PCA: principal component analysis; MLMM: multi-locus mixed linear model.

Table S3.7 Significant MTAs identified based on the *Brassica napus* L. combined population based on MS-2 (16,855 SNP markers) on all five traits: seed yield (YLD), plant height (HT), seed protein content (SPC), seed oil content (SOC) and seed glucosinolate content (GSL).

Trait	Marker	Model	Chromosome	Position (bp)
YLD	Bn-A02-p6766615	CMLM	A2	204,182
	Bn-A02-p7351109	CMLM	A2	270,573
	Bn-A02-p3658139	CMLM	A2	1,093,648
	Bn-A02-p7004091	CMLM	A2	4,037,640
	Bn-scaff_15712_13-p63324	CMLM	A2	4,500,718
	Bn-A02-p16666866	CMLM	A2	13,863,707
	Bn-A02-p23836232	CMLM	A2	22,068,630
	Bn-A02-p26154951	CMLM	A2	23,778,815
	Bn-A04-p6389136	CMLM	A4	7,593,915
	Bn-A05-p6405336	CMLM	A5	5,956,781
	Bn-A07-p9115370	CMLM	A7	10,581,719
	Bn-A08-p4207231	CMLM	A8	3,626,293
	Bn-A10-p8409315	CMLM	A10	9,898,825
	Bn-scaff_16269_1-p192431	CMLM	C2	6,421,917
	Bn-scaff_18374_1-p23965	CMLM	C2	7,565,729
	Bn-scaff_16804_4-p112136	CMLM	C2	7,610,450
	Bn-scaff_15712_5-p1021105	CMLM	C2	8,512,187
	Bn-scaff_16534_1-p259614	CMLM	C4	854,745
	Bn-scaff_18520_1-p622295	CMLM	C7	32,478,815
	Bn-A04-p6389136	FarmCPU	A4	7,593,915
	Bn-A10-p16164252	FarmCPU	A10	15,400,988
	Bn-scaff_15712_13-p43168	FarmCPU	C2	8,307,658
	Bn-scaff_20270_1-p1324392	FarmCPU	C5	41,751,856
	Bn-scaff_18520_1-p622295	FarmCPU	C7	32,478,815
	Bn-scaff_17048_1-p243085	FarmCPU	C9	1,985,336
	Bn-scaff_16246_2-p3090	FarmCPU	C9	2,856,998
	Bn-A04-p6389136	MLMM	A4	7,593,915
	Bn-scaff_16269_1-p192431	MLMM	C2	6,421,917
	Bn-scaff_17048_1-p243085	MLMM	C9	1,985,336
	Bn-scaff_16246_2-p3090	MLMM	C9	2,856,998
HT	Bn-scaff_15911_1-p546022	CMLM	A1	14,317,565
	Bn-A02-p6766615	CMLM	A2	204,182
	Bn-A02-p7004091	CMLM	A2	4,037,640
	Bn-scaff_15712_13-p63324	CMLM	A2	4,500,718
	Bn-A02-p16666866	CMLM	A2	13,863,707
	Bn-A02-p23836232	CMLM	A2	22,068,630

Bn-A03-p9605319	CMLM	A3	787,447
Bn-A04-p6389136	CMLM	A4	7,593,915
Bn-A05-p5058143	CMLM	A5	4,877,416
Bn-A05-p6405336	CMLM	A5	5,956,781
Bn-A06-p3206017	CMLM	A6	196,789
Bn-A06-p2365869	CMLM	A6	2,381,259
Bn-A07-p9115370	CMLM	A7	10,581,719
Bn-scaff_23957_1-p127219	CMLM	A7	20,022,579
Bn-A08-p4145883	CMLM	A8	3,547,660
Bn-A08-p4146394	CMLM	A8	3,548,169
Bn-A08-p4146828	CMLM	A8	3,550,256
Bn-A08-p4147963	CMLM	A8	3,551,391
Bn-A08-p4207231	CMLM	A8	3,626,293
Bn-A09-p264743	CMLM	A9	429,423
Bn-A10-p8409315	CMLM	A10	9,898,825
Bn-A10-p15437444	CMLM	A10	16,055,360
Bn-scaff_16553_1-p34303	CMLM	C1	4,026,215
Bn-scaff_16269_1-p192431	CMLM	C2	6,421,917
Bn-scaff_18374_1-p23965	CMLM	C2	7,565,729
Bn-scaff_16804_4-p112136	CMLM	C2	7,610,450
Bn-scaff_15712_5-p1021105	CMLM	C2	8,512,187
Bn-scaff_18675_1-p421921	CMLM	C2	22,122,071
Bn-scaff_18936_1-p744477	CMLM	C3	3,325,084
Bn-scaff_16996_1-p230771	CMLM	C3	56,423,745
Bn-scaff_16534_1-p259614	CMLM	C4	854,745
Bn-scaff_19253_1-p285404	CMLM	C4	15,292,281
Bn-scaff_16888_1-p1803454	CMLM	C4	45,940,032
Bn-scaff_20376_1-p218414	CMLM	C5	42,120,912
Bn-scaff_15712_13-p63324	FarmCPU	A2	4,500,718
Bn-A05-p6405336	FarmCPU	A5	5,956,781
Bn-A06-p2940743	FarmCPU	A6	2,840,416
Bn-A06-p10297153	FarmCPU	A6	9,686,244
Bn-A07-p4097388	FarmCPU	A7	6,014,828
Bn-A10-p13016246	FarmCPU	A10	13,051,612
Bn-scaff_20376_1-p218414	FarmCPU	C5	42,120,912
Bn-scaff_20619_1-p167640	FarmCPU	C9	31,678,632
Bn-scaff_19899_1-p283181	FarmCPU	C9	40,792,238
Bn-scaff_17750_1-p1839429	FarmCPU	C9	46,646,099
Bn-A03-p23996628	MLMM	A3	22,640,491
Bn-A05-p6405336	MLMM	A5	5,956,781

	Bn-A07-p9115370	MLMM	A7	10,581,719
SPC	Bn-A04-p6389136	CMLM	A4	7,593,915
	Bn-A07-p9115370	CMLM	A7	10,581,719
	Bn-A09-p8635608	CMLM	A9	8,062,605
	Bn-scaff_17801_1-p220808	CMLM	C9	15,322,388
	Bn-A07-p9115370	FarmCPU	A7	10,581,719
	Bn-A09-p135892	FarmCPU	A9	539,846
	Bn-A01-p8821722	FarmCPU	A9	1,770,199
	Bn-A09-p8635608	FarmCPU	A9	8,062,605
	Bn-scaff_15936_1-p245327	FarmCPU	C1	36,380,585
	Bn-scaff_26642_1-p13971	FarmCPU	C3	52,895,415
	Bn-A05-p23290863	FarmCPU	C5	41,789,253
	Bn-scaff_16397_1-p114405	FarmCPU	C6	32,795,141
	Bn-scaff_16110_1-p436278	FarmCPU	C7	44,461,558
	Bn-scaff_17048_1-p243085	FarmCPU	C9	1,985,336
	Bn-scaff_16246_2-p3090	FarmCPU	C9	2,856,998
	Bn-scaff_17750_1-p1810873	FarmCPU	C9	46,663,402
	Bn-A04-p6389136	MLMM	A4	7,593,915
	Bn-A09-p8635608	MLMM	A9	8,062,605
SOC	Bn-A02-p8169424	CMLM	A2	5,160,109
	Bn-A02-p3121742	FarmCPU	A2	605,811
	Bn-scaff_18602_1-p263207	FarmCPU	C3	51,660,331
	Bn-scaff_20219_1-p200426	FarmCPU	C5	40,827,867
	Bn-scaff_17801_1-p220808	FarmCPU	C9	15,322,388
	Bn-scaff_19899_1-p356624	FarmCPU	C9	40,719,329
	Bn-scaff_18374_1-p23965	MLMM	C2	7,565,729
	Bn-scaff_15818_2-p1146207	MLMM	C6	17,529,142
	Bn-scaff_17048_1-p243085	MLMM	C9	1,985,336
	Bn-scaff_20619_1-p112877	MLMM	C9	31,738,963
GSL	Bn-scaff_15712_13-p63324	CMLM	A2	4,500,718
	Bn-A07-p9115370	CMLM	A7	10,581,719
	Bn-scaff_23957_1-p127219	CMLM	A7	20,022,579
	Bn-A09-p264743	CMLM	A9	429,423
	Bn-scaff_16269_1-p192431	CMLM	C2	6,421,917
	Bn-scaff_16269_1-p57060	CMLM	C2	6,574,199
	Bn-scaff_16804_1-p635989	CMLM	C2	7,408,306
	Bn-scaff_18374_1-p23965	CMLM	C2	7,565,729
	Bn-scaff_16804_4-p112136	CMLM	C2	7,610,450
	Bn-scaff_18514_1-p28001	CMLM	C2	7,925,571
	Bn-scaff_15712_13-p38138	CMLM	C2	8,302,329

Bn-scaff_15712_13-p43168	CMLM	C2	8,307,658
Bn-scaff_15712_5-p1021105	CMLM	C2	8,512,187
Bn-scaff_18936_1-p744477	CMLM	C3	3,325,084
Bn-scaff_22728_1-p947436	CMLM	C5	6,146,830
Bn-A01-p5567077	FarmCPU	A1	322,699
Bn-scaff_26139_1-p322324	FarmCPU	A4	5,291,782
Bn-scaff_26787_1-p6892	FarmCPU	A4	17,975,686
Bn-A05-p671910	FarmCPU	A5	793,277
Bn-A06-p18438509	FarmCPU	A6	19,808,118
Bn-scaff_23957_1-p127219	FarmCPU	A7	20,022,579
Bn-A08-p19575209	FarmCPU	A8	2,027,576
Bn-A10-p2311797	FarmCPU	A10	1,540,475
Bn-scaff_15838_1-p2212925	FarmCPU	C1	2,576,689
Bn-A02-p10719296	FarmCPU	C2	13,849,225
Bn-scaff_18936_1-p744477	FarmCPU	C3	3,325,084
Bn-scaff_23761_1-p738056	FarmCPU	C3	6,013,746
Bn-scaff_17298_1-p35170	FarmCPU	C3	22,345,273
Bn-scaff_16888_1-p1803454	FarmCPU	C4	45,940,032
Bn-scaff_22728_1-p947436	FarmCPU	C5	6,146,830
Bn-scaff_15892_1-p368466	FarmCPU	C6	26,266,184
Bn-scaff_16394_1-p83655	FarmCPU	C7	24,655,094
Bn-scaff_15712_13-p63324	MLMM	A2	4,500,718
Bn-A07-p9115370	MLMM	A7	10,581,719
Bn-scaff_23957_1-p127219	MLMM	A7	20,022,579
Bn-A09-p264743	MLMM	A9	429,423
Bn-scaff_16269_1-p192431	MLMM	C2	6,421,917
Bn-scaff_16269_1-p57060	MLMM	C2	6,574,199
Bn-scaff_16804_1-p635989	MLMM	C2	7,408,306
Bn-scaff_18374_1-p23965	MLMM	C2	7,565,729
Bn-scaff_16804_4-p112136	MLMM	C2	7,610,450
Bn-scaff_18514_1-p28001	MLMM	C2	7,925,571
Bn-scaff_15712_13-p38138	MLMM	C2	8,302,329
Bn-scaff_15712_13-p43168	MLMM	C2	8,307,658
Bn-scaff_15712_5-p1021105	MLMM	C2	8,512,187
Bn-scaff_18936_1-p744477	MLMM	C3	3,325,084
Bn-scaff_22728_1-p947436	MLMM	C5	6,146,830

Abbreviations: SNP: single nucleotide polymorphism; CMLM: compression mixed linear model; FarmCPU: fixed and random model circulating probability unification; MLMM: multi-locus mixed linear model.

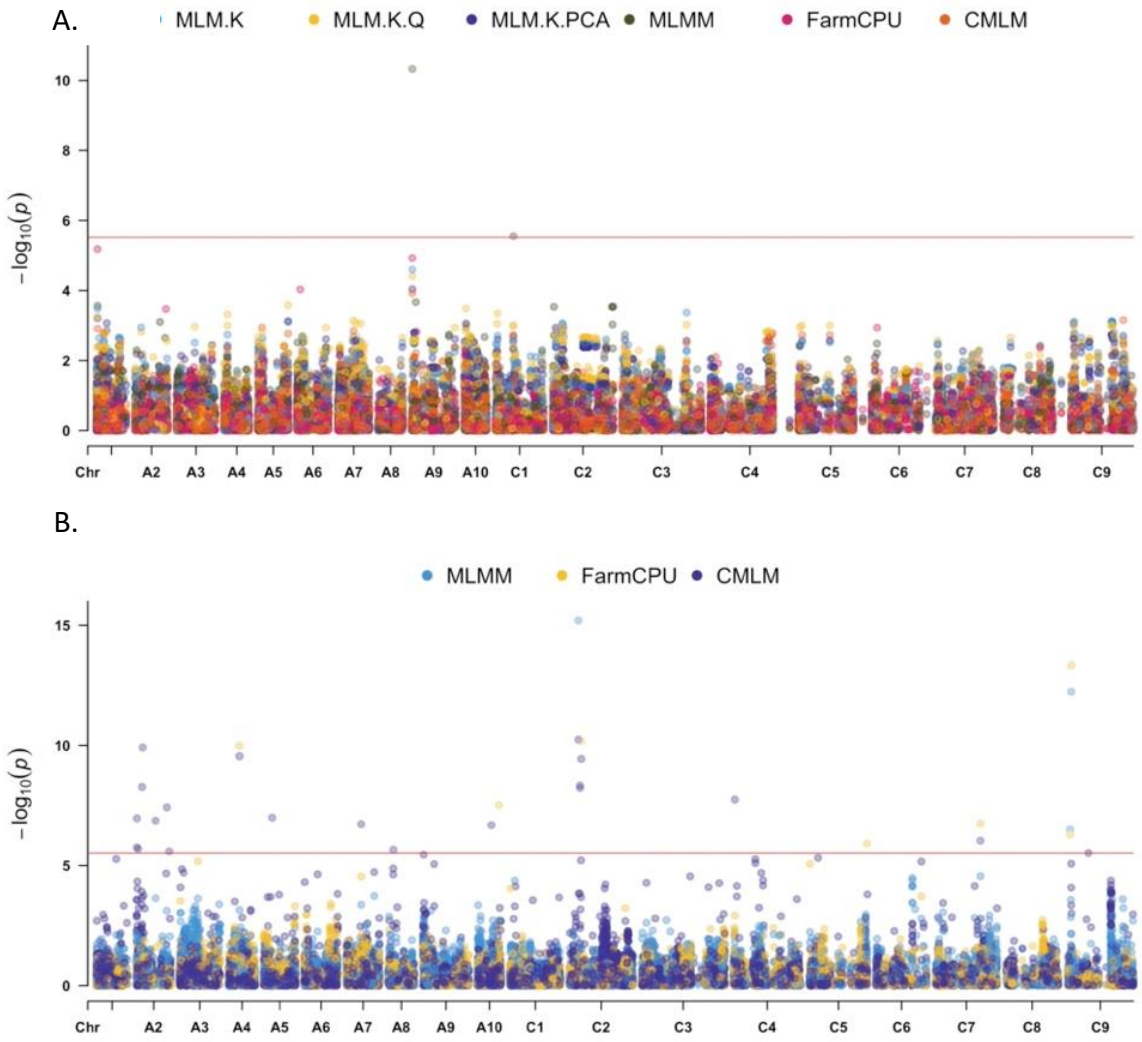


Figure S3.1 Manhattan plots showing seed yield (YLD) based on the MS-2 (16,855 SNP markers). Bonferroni-corrected significance threshold was shown as the red horizontal line at $-\log_{10}(0.05/16855) = 5.53$. (A) Results from the *Brassica napus* L. parental population based on six models including mixed linear models considering kinship (MLM+K), mixed linear models considering subpopulation structure via Bayesian clustering (MLM+K+Q), mixed linear models considering subpopulation structure via principal component analysis (MLM+K+PCA), multi-loci mixed model (MLMM), Fixed and random model circulating probability unification (FarmCPU) and compression mixed linear model (CMLM). (B) Results from the *Brassica napus* L. combined population based on three models MLMM, FarmCPU and CMLM.

Table S3.8 Candidate genes predicted under the Brassicales order based on combined significant MTAs identified from the parental population and combined population of *Brassica napus* L. Traits evaluated are YLD (seed yield), HT (plant height), seed protein content (SPC), seed oil content (SOC) and GSL (seed glucosinloate content).

Trait	Query ID	Gene Name	EggNOG Description	E-Value	Bit-Score
YLD	BnaC02g36370D	AFB2	auxin signaling f-box	0	1171.8
YLD	BnaA02g33110D	AGD1	ADP-ribosylation factor GTPase-activating protein	0	1565.4
YLD	BnaA02g21070D	ATPC	ATP synthase gamma chain 1	2.50E-171	608.6
YLD	BnaA02g20870D	BSL1	Serine threonine-protein phosphatase	0	1780.8
YLD	BnaC02g34690D	CAND1	Cullin-associated	0	2268.4
YLD	BnaA02g21770D	CBF5	H ACA ribonucleoprotein complex subunit	5.60E-241	840.1
YLD	BnaA02g30320D	CHS2	Chalcone and stilbene synthases, N-terminal domain	2.20E-226	791.2
YLD	BnaA02g30340D	CHS2	Chalcone and stilbene synthases, N-terminal domain	1.60E-28	132.1
YLD	BnaC01g19610D	CIP7	COP1-interacting protein 7	0	1690.6
YLD	BnaC01g19600D	CIP7	COP1-interacting protein 7	4.70E-29	134.4
YLD	BnaA02g08420D	CKX3	Cytokinin dehydrogenase 1, FAD and cytokinin binding	2.70E-304	1050.4
YLD	BnaA02g00750D	CM2	chorismate mutase	1.70E-137	495.4

YLD	BnaC02g13300D	CNX5	Plays a central role in 2-thiolation of mcm(5)S(2)U at tRNA wobble positions of cytosolic tRNA(Lys), tRNA(Glu) and tRNA(Gln). Also essential during biosynthesis of the molybdenum cofactor. Acts by mediating the C-terminal thiocarboxylation of sulfur carriers URM1 and MOCS2A. Its N-terminus first activates URM1 and MOCS2A as acyl-adenylates (-COAMP), then the persulfide sulfur on the catalytic cysteine is transferred to URM1 and MOCS2A to form thiocarboxylation (-COSH) of their C-terminus. The reaction probably involves hydrogen sulfide that is generated from the persulfide intermediate and that acts as nucleophile towards URM1 and MOCS2A. Subsequently, a transient disulfide bond is formed. Does not use thiosulfate as sulfur donor	5.80E-258	896.3
YLD	BnaA08g04270D	CPK18	calcium-dependent protein kinase	7.20E-17	92.4
YLD	BnaA02g21060D	CPK23	Calcium-dependent protein kinase	2.20E-298	1030.8
YLD	BnaA08g04030D	CYCA3-1	Belongs to the cyclin family	8.70E-201	706.1
YLD	BnaA08g04010D	CYCA3-1	Belongs to the cyclin family	4.50E-197	693.7
YLD	BnaC02g36490D	DRB3	dsRNA-binding protein	8.00E-189	666.4
YLD	BnaC02g36690D	DREB2A	Dehydration-responsive element-binding protein	6.00E-34	149.8

YLD	BnaC02g36700D	DREB2A	Dehydration-responsive element-binding protein	2.30E-93	348.2
YLD	BnaA02g33430D	EB1	Microtubule-associated protein RP EB family member	7.30E-147	526.6
YLD	BnaA02g21720D	EIF3A	RNA-binding component of the eukaryotic translation initiation factor 3 (eIF-3) complex, which is involved in protein synthesis of a specialized repertoire of mRNAs and, together with other initiation factors, stimulates binding of mRNA and methionyl-tRNAi to the 40S ribosome. The eIF-3 complex specifically targets and initiates translation of a subset of mRNAs involved in cell proliferation	0	1452.2
YLD	BnaC02g11150D	EIF4G	translation initiation factor	0	1469.9
YLD	BnaC07g26600D	FKBP15-2	peptidyl-prolyl cis-trans isomerase	7.00E-86	323.2
YLD	BnaC02g13190D	FTSZ1-1	Cell division protein FtsZ homolog 1	6.40E-227	793.1
YLD	BnaA02g21470D	GAE5	4-epimerase 5	8.90E-245	852.4
YLD	BnaC02g36240D	GAPA	Belongs to the glyceraldehyde-3-phosphate dehydrogenase family	3.00E-212	744.2
YLD	BnaC09g04970D	HCF136	Photosystem II stability assembly factor	2.00E-18	97.8
YLD	BnaC09g04960D	HCF136	Photosystem II stability assembly factor	2.90E-49	201.4

YLD	BnaA02g21580D	HMGS	This enzyme condenses acetyl-CoA with acetoacetyl-CoA to form HMG-CoA, which is the substrate for HMG-CoA reductase	1.40E-267	928.3
YLD	BnaC02g35320D	HSP90C	Hsp90 protein	2.9E-311	1073.9
YLD	BnaA02g33000D	HUA2	enhancer of ag-4	1.30E-111	409.1
YLD	BnaA02g33010D	HUA2	enhancer of ag-4	0	1650.6
YLD	BnaC02g34780D	IQD29	IQ-domain	6.30E-246	856.7
YLD	BnaC02g35130D	Lhcb6-1	The light-harvesting complex (LHC) functions as a light receptor, it captures and delivers excitation energy to photosystems with which it is closely associated	0.0018	49.7
YLD	BnaA08g04050D	MAF1	WPP domain-containing protein	1.20E-75	289.3
YLD	BnaA02g02390D	MAN2A2	alpha-mannosidase	0	2393.6
YLD	BnaA02g33410D	MYB31	RNA polymerase II transcription regulator recruiting activity	7.50E-141	506.5
YLD	BnaC02g34810D	MYB88	RNA polymerase II transcription regulator recruiting activity	3.00E-241	840.9

YLD	BnaA02g33590D	NBP35	Component of the cytosolic iron-sulfur (Fe-S) protein assembly (CIA) machinery. Required for maturation of extramitochondrial Fe-S proteins. Functions as Fe-S scaffold, mediating the de novo assembly of an Fe-S cluster and its transfer to target apoproteins. Essential for embryo development	1.40E-13	82.8
YLD	BnaC02g35480D	NIK2	Belongs to the protein kinase superfamily. Ser Thr protein kinase family	2.90E-25	121.3
YLD	BnaC02g35750D	NIMIN-2	Nim1-interacting 2	6.20E-34	149.8
YLD	BnaA08g03990D	PAF2	The proteasome is a multicatalytic proteinase complex which is characterized by its ability to cleave peptides with Arg, Phe, Tyr, Leu, and Glu adjacent to the leaving group at neutral or slightly basic pH	2.10E-135	488.4
YLD	BnaA02g33180D	PER1	Per1-like family	6.10E-204	716.5
YLD	BnaC02g10920D	PHOT2	FMN binding blue light photoreceptor kinase protein serine threonine kinase	0	1666.4
YLD	BnaA02g08340D	PIN2	May act as a component of the auxin efflux carrier	2.80E-303	1047.3
YLD	BnaC01g19630D	PORA	Protochlorophyllide reductase	4.90E-213	746.9
YLD	BnaC02g13600D	PRPL24	Belongs to the universal ribosomal protein uL24 family	1.00E-102	379.4

YLD	BnaC09g04920D	PYR4	Belongs to the dihydroorotate dehydrogenase family. Type 2 subfamily	3.20E-264	917.1
YLD	BnaC02g13240D	RAN1	GTP-binding protein involved in nucleocytoplasmic transport. Required for the import of protein into the nucleus and also for RNA export. Involved in chromatin condensation and control of cell cycle	5.00E-115	420.6
YLD	BnaC02g35650D	RAV1	AP2 ERF and B3 domain-containing transcription factor	4.60E-188	663.7
YLD	BnaC02g12070D	RPL31	60s ribosomal protein	5.40E-59	233.4
YLD	BnaC02g12080D	RPS30	Belongs to the eukaryotic ribosomal protein eS30 family	5.70E-31	140.2
YLD	BnaC02g34840D	SUT2	PUCC protein	1.30E-298	1031.6
YLD	BnaC02g36550D	TCP20	Transcription factor	6.50E-11	73.9
YLD	BnaC02g36570D	TCP20	Transcription factor	1.30E-09	68.9
YLD	BnaC02g13170D	TOP1	DNA Topoisomerase I (eukaryota)	0	1242.6
YLD	BnaA02g33480D	UBC3	ubiquitin-conjugating enzyme	1.20E-84	318.9
YLD	BnaA02g00690D	UBP22	ubiquitin carboxyl-terminal hydrolase	0	1094.3
YLD	BnaA08g04100D	WOX4	Homeodomain	3.80E-134	484.2
YLD	BnaC02g34990D	WRKY3	Transcription factor	1.60E-205	722.2
YLD	BnaA02g21850D	WRKY41	DNA binding domain	3.60E-185	654.1
YLD	BnaA02g02500D	WRKY72	transcription factor	2.20E-136	491.9
HT	BnaA03g45080D	APR1	reductase 1	2.70E-263	914.1
HT	BnaC04g19270D	APR1	reductase 2	1.50E-261	908.3

HT	BnaA06g03040D	ABF1	DNA binding protein binding transcription activator transcription factor	6.10E-161	573.5
HT	BnaC04g18780D	ACA1	This magnesium-dependent enzyme catalyzes the hydrolysis of ATP coupled with the transport of calcium	3.80E-32	143.7
HT	BnaC06g11800D	APC7	Tetratricopeptide repeat	4.60E-73	280.4
HT	BnaC04g18310D	APE2	Triose phosphate phosphate translocator	6.30E-107	393.7
HT	BnaA10g16550D	APY2	nucleoside-diphosphatase activity	5.70E-269	932.9
HT	BnaC04g15900D	ARF10	Auxin response factors (ARFs) are transcriptional factors that bind specifically to the DNA sequence 5'-TGTCTC-3' found in the auxin-responsive promoter elements (AuxREs)	0	1351.3
HT	BnaC04g13500D	ATG9	Involved in autophagy and cytoplasm to vacuole transport (Cvt) vesicle formation. Plays a key role in the organization of the preautophagosomal structure phagophore assembly site (PAS), the nucleating site for formation of the sequestering vesicle	2.30E-153	548.1
HT	BnaC09g47340D	AtpB	Produces ATP from ADP in the presence of a proton gradient across the membrane	0.00E+00	1067.4

HT	BnaC09g47330D	AtpB	Produces ATP from ADP in the presence of a proton gradient across the membrane	7.10E-253	879.4
HT	BnaC03g66580D	BAK1	Belongs to the protein kinase superfamily. Ser Thr protein kinase family	7.20E-140	503.8
HT	BnaC04g16300D	BBX30	zinc finger	6.00E-69	266.5
HT	BnaC03g07000D	BCCP2	first, biotin carboxylase catalyzes the carboxylation of the carrier protein and then the transcarboxylase transfers the carboxyl group to form malonyl-CoA	2.20E-97	362.1
HT	BnaC04g15750D	BGAL8	beta-galactosidase	0	1733.8
HT	BnaA03g44890D	BGLU47	Belongs to the glycosyl hydrolase 1 family	5.80E-304	1049.3
HT	BnaA03g44900D	BGLU47	Belongs to the glycosyl hydrolase 1 family	2.60E-109	401.4
HT	BnaA03g44910D	BGLU47	Belongs to the glycosyl hydrolase 1 family	3.10E-172	610.9
HT	BnaA08g15040D	BTI1	Reticulon-like protein	5.00E-124	450.7
HT	BnaC04g14970D	CDC6	cell division control	1.50E-264	918.3
HT	BnaA04g25810D	CHL-CPN10	Belongs to the GroES chaperonin family	8.50E-72	276.2
HT	BnaC04g14410D	CIPK11	Non-specific serine threonine protein kinase	1.00E-256	892.1
HT	BnaA06g03950D	CIPK17	CBL-interacting protein kinase	1.40E-237	828.6
HT	BnaA06g03930D	CIPK17	CBL-interacting protein kinase	2.30E-232	811.2
HT	BnaC04g17180D	CMK	GHMP kinases N terminal domain	1.50E-219	768.5

HT	BnaC02g13300D	CNX5	Plays a central role in 2-thiolation of mcm(5)S(2)U at tRNA wobble positions of cytosolic tRNA(Lys), tRNA(Glu) and tRNA(Gln). Also essential during biosynthesis of the molybdenum cofactor. Acts by mediating the C-terminal thiocarboxylation of sulfur carriers URM1 and MOCS2A. Its N-terminus first activates URM1 and MOCS2A as acyl-adenylates (-COAMP), then the persulfide sulfur on the catalytic cysteine is transferred to URM1 and MOCS2A to form thiocarboxylation (-COSH) of their C-terminus. The reaction probably involves hydrogen sulfide that is generated from the persulfide intermediate and that acts as nucleophile towards URM1 and MOCS2A. Subsequently, a transient disulfide bond is formed. Does not use thiosulfate as sulfur donor	5.80E-258	896.3
HT	BnaC04g13690D	COQ3	Belongs to the class I-like SAM-binding methyltransferase superfamily. UbiG COQ3 family	8.50E-21	105.9
HT	BnaA08g04270D	CPK18	calcium-dependent protein kinase	7.20E-17	92.4

HT	BnaC04g13100D	CPK24	ATP binding calcium ion binding calmodulin-dependent protein kinase kinase protein kinase protein serine threonine kinase	9.70E-273	945.7
HT	BnaC04g16310D	CPN60A	Belongs to the chaperonin (HSP60) family	1.10E-252	879
HT	BnaA08g04010D	CYCA3-1	Belongs to the cyclin family	4.50E-197	693.7
HT	BnaA08g04030D	CYCA3-1	Belongs to the cyclin family	8.70E-201	706.1
HT	BnaC04g15310D	CYP707A2	Belongs to the cytochrome P450 family	1.50E-277	961.4
HT	BnaC04g14620D	DRP4A	Belongs to the TRAFAC class dynamin-like GTPase superfamily. Dynamin Fzo YdjA family	9.60E-215	753.1
HT	BnaC04g18250D	EF1Bgamma2	elongation factor	5.20E-94	350.9
HT	BnaC06g11860D	EGY1	zinc metalloprotease EGY1, chloroplastic	6.10E-54	216.5
HT	BnaC06g11850D	EGY1	zinc metalloprotease EGY1, chloroplastic	0	1094.7
HT	BnaC06g11880D	eIF(iso)4E	Initiation factor	2.10E-100	371.7
HT	BnaC02g11150D	EIF4G	translation initiation factor	0	1469.9
HT	BnaC03g66460D	EX1	Domain of unknown function (DUF3506)	0	1135.9
HT	BnaC04g15200D	EXPA6	Rare lipoprotein A (RlpA)-like double-psi beta-barrel	2.60E-96	358.2
HT	BnaC04g14590D	FABD	carrier protein transacylase	8.70E-199	699.5
HT	BnaC04g14820D	FAD3	omega-3 fatty acid desaturase	3.60E-229	800.4
HT	BnaC02g13190D	FTSZ1-1	Cell division protein FtsZ homolog 1	6.40E-227	793.1
HT	BnaA05g08850D	HAK11	Potassium transporter	0	1464.1

HT	BnaC04g14350D	HSI2	High-level expression of sugar-inducible gene 2	0	1520
HT	BnaA03g45240D	IleS	Belongs to the class-I aminoacyl-tRNA synthetase family	0	1616.7
HT	BnaA03g45250D	IleS	Belongs to the class-I aminoacyl-tRNA synthetase family	1.30E-89	336.3
HT	BnaA08g04050D	MAF1	WPP domain-containing protein	1.20E-75	289.3
HT	BnaA03g44930D	MSRB2	methionine sulfoxide reductase	1.20E-109	402.5
HT	BnaC01g19550D	NAC4	(NAC) domain-containing protein	9.20E-172	609.4
HT	BnaC09g29150D	ndhH	NDH shuttles electrons from NAD(P)H plastoquinone, via FMN and iron-sulfur (Fe-S) centers, to quinones in the photosynthetic chain and possibly in a chloroplast respiratory chain. The immediate electron acceptor for the enzyme in this species is believed to be plastoquinone. Couples the redox reaction to proton translocation, and thus conserves the redox energy in a proton gradient	1.10E-71	275.8

HT	BnaC09g29160D	ndhH	NDH shuttles electrons from NAD(P)H plastoquinone, via FMN and iron-sulfur (Fe-S) centers, to quinones in the photosynthetic chain and possibly in a chloroplast respiratory chain. The immediate electron acceptor for the enzyme in this species is believed to be plastoquinone. Couples the redox reaction to proton translocation, and thus conserves the redox energy in a proton gradient	3.30E-140	504.2
HT	BnaC04g16320D	NPC1	Niemann-Pick C1 protein-like	0	2439.8
HT	BnaA08g03990D	PAF2	The proteasome is a multicatalytic proteinase complex which is characterized by its ability to cleave peptides with Arg, Phe, Tyr, Leu, and Glu adjacent to the leaving group at neutral or slightly basic pH	2.10E-135	488.4
HT	BnaC04g17530D	PAF2	The proteasome is a multicatalytic proteinase complex which is characterized by its ability to cleave peptides with Arg, Phe, Tyr, Leu, and Glu adjacent to the leaving group at neutral or slightly basic pH	3.10E-150	537.7

HT	BnaC04g17110D	PAG1	The proteasome is a multicatalytic proteinase complex which is characterized by its ability to cleave peptides with Arg, Phe, Tyr, Leu, and Glu adjacent to the leaving group at neutral or slightly basic pH	2.10E-137	495
HT	BnaC04g13660D	PARP1	poly ADP-ribose polymerase	0	1830.8
HT	BnaC04g18160D	PBG1	The proteasome is a multicatalytic proteinase complex which is characterized by its ability to cleave peptides with Arg, Phe, Tyr, Leu, and Glu adjacent to the leaving group at neutral or slightly basic pH	6.80E-141	506.5
HT	BnaC02g10920D	PHOT2	FMN binding blue light photoreceptor kinase protein serine threonine kinase	0	1666.4
HT	BnaC06g11890D	PHYC	Regulatory photoreceptor which exists in two forms that are reversibly interconvertible by light the Pr form that absorbs maximally in the red region of the spectrum and the Pfr form that absorbs maximally in the far-red region	0	2194.1

HT	BnaA08g15050D	PPT1	Catalyzes the prenylation of para-hydroxybenzoate (PHB) with an all-trans polyprenyl group. Mediates the second step in the final reaction sequence of coenzyme Q (CoQ) biosynthesis, which is the condensation of the polyisoprenoid side chain with PHB, generating the first membrane-bound Q intermediate	4.70E-120	437.6
HT	BnaC02g13600D	PRPL24	Belongs to the universal ribosomal protein uL24 family	1.00E-102	379.4
HT	BnaC04g15570D	psaA	PsaA and PsaB bind P700, the primary electron donor of photosystem I (PSI), as well as the electron acceptors A0, A1 and FX. PSI is a plastocyanin-ferredoxin oxidoreductase, converting photonic excitation into a charge separation, which transfers an electron from the donor P700 chlorophyll pair to the spectroscopically characterized acceptors A0, A1, FX, FA and FB in turn. Oxidized P700 is reduced on the lumenal side of the thylakoid membrane by plastocyanin	0	1550.4

HT	BnaC04g15560D	psaB	PsaA and PsaB bind P700, the primary electron donor of photosystem I (PSI), as well as the electron acceptors A0, A1 and FX. PSI is a plastocyanin-ferredoxin oxidoreductase, converting photonic excitation into a charge separation, which transfers an electron from the donor P700 chlorophyll pair to the spectroscopically characterized acceptors A0, A1, FX, FA and FB in turn. Oxidized P700 is reduced on the lumenal side of the thylakoid membrane by plastocyanin	0	1100.1
HT	BnaC02g13240D	RAN1	GTP-binding protein involved in nucleocytoplasmic transport. Required for the import of protein into the nucleus and also for RNA export. Involved in chromatin condensation and control of cell cycle	5.00E-115	420.6
HT	BnaC04g15270D	RBL1	Rhomboid family	1.00E-215	755.7

HT	BnaA03g44840D	RPB2	DNA-dependent RNA polymerase catalyzes the transcription of DNA into RNA using the four ribonucleoside triphosphates as substrates	0	2344.3
HT	BnaC02g12070D	RPL31	60s ribosomal protein	5.40E-59	233.4
HT	BnaC04g13530D	RPS25	ribosomal protein	2.30E-09	68.2
HT	BnaC04g13040D	RPS3	40S ribosomal protein	1.30E-132	479.2
HT	BnaC02g12080D	RPS30	Belongs to the eukaryotic ribosomal protein eS30 family	5.70E-31	140.2
HT	BnaA02g00470D	RPS6	Belongs to the eukaryotic ribosomal protein eS6 family	7.00E-133	479.9
HT	BnaC04g17690D	RPT4A	Belongs to the AAA ATPase family	3.40E-222	777.3
HT	BnaA03g45360D	SCD1	DENN domain and WD repeat-containing protein SCD1	4.10E-118	431
HT	BnaC09g29060D	TAP46	TAP42-like family	2.20E-200	704.9
HT	BnaC02g13170D	TOP1	DNA Topoisomerase I (eukaryota)	0	1242.6
HT	BnaA03g44620D	TPP2	Tripeptidyl-peptidase	0	2507.6
HT	BnaC09g29430D	TRY	transcription regulator recruiting activity	3.20E-50	204.1
HT	BnaA03g44640D	TUB9	Tubulin is the major constituent of microtubules. It binds two moles of GTP, one at an exchangeable site on the beta chain and one at a non-exchangeable site on the alpha chain	3.10E-256	890.6
HT	BnaC04g14640D	UBA1	Belongs to the ubiquitin-activating E1 family	0	2107.8

HT	BnaC09g28960D	UGT84B2	Belongs to the UDP-glycosyltransferase family	4.30E-28	131.3
HT	BnaC04g14870D	VLN1	Villin headpiece domain	0	1788.9
HT	BnaA08g04100D	WOX4	Homeodomain	3.80E-134	484.2
HT	BnaC04g13870D	WRKY21	Transcription factor	1.20E-180	639
HT	BnaC04g19430D	WRKY6	Transcription factor	1.30E-231	808.9
HT	BnaC09g29130D	ycf1	BEST Arabidopsis thaliana protein match is Ycf1 protein (TAIR	0	1855.1
HT	BnaC09g29140D	ycf1	BEST Arabidopsis thaliana protein match is Ycf1 protein (TAIR	7.50E-152	543.5
HT	BnaC09g29210D	ycf1	BEST Arabidopsis thaliana protein match is Ycf1 protein (TAIR	5.40E-184	650.2
HT	BnaC04g14740D	ZIP6	Transporter	1.60E-175	622.1
SPC	BnaA09g13300D	ACO1	Belongs to the iron ascorbate-dependent oxidoreductase family	1.10E-186	659.1
SPC	BnaC03g62970D	AG	Floral homeotic protein	1.10E-125	456.1
SPC	BnaC03g64810D	BGLU3	Belongs to the glycosyl hydrolase 1 family	1.20E-100	372.5
SPC	BnaC03g64790D	BGLU5	Belongs to the glycosyl hydrolase 1 family	1.60E-94	352.8
SPC	BnaC09g20030D	BI-1	Belongs to the BI1 family	5.00E-123	447.2
SPC	BnaC07g47240D	BRI1	Belongs to the protein kinase superfamily. Ser Thr protein kinase family	0	1262.7
SPC	BnaC09g19210D	CUL4	Belongs to the cullin family	0	1461.4
SPC	BnaC01g37450D	CWINV	Belongs to the glycosyl hydrolase 32 family	0	1212.2
SPC	BnaC03g63590D	CYCT1	Belongs to the cyclin family	9.90E-286	988.8
SPC	BnaC09g18860D	CYP707A1	Belongs to the cytochrome P450 family	2.10E-271	941

SPC	BnaC03g63040D	ERD3	AT4G19120 (E 0.0) ERD3 ERD3 (early-responsive to dehydration 3)	0	1253
SPC	BnaC09g18650D	FAD7	omega-3 fatty acid desaturase	1.90E-174	618.2
SPC	BnaA09g13190D	FMOGS-OX1	flavin-containing monooxygenase	8.30E-273	945.7
SPC	BnaA03g47400D	GA20ox1	Belongs to the iron ascorbate-dependent oxidoreductase family	3.20E-208	730.7
SPC	BnaA09g14180D	GolS4	Belongs to the glycosyltransferase 8 family	3.20E-197	694.1
SPC	BnaC01g37200D	GOX1	(S)-2-hydroxy-acid oxidase	5.50E-195	686.8
SPC	BnaA09g03700D	GSH2	glutathione synthetase	7.50E-266	922.5
SPC	BnaA09g00850D	GSTF2	Belongs to the GST superfamily	7.50E-58	229.6
SPC	BnaA09g00860D	GSTF2	Belongs to the GST superfamily	1.90E-14	84.7
SPC	BnaC09g04960D	HCF136	Photosystem II stability assembly factor	2.90E-49	201.4
SPC	BnaC09g04970D	HCF136	Photosystem II stability assembly factor	2.00E-18	97.8
SPC	BnaC09g19280D	KAS1	Belongs to the beta-ketoacyl-ACP synthases family	1.10E-269	935.3
SPC	BnaC09g19900D	LBA1	Regulator of nonsense transcripts 1 homolog	0	1402.1
SPC	BnaA09g13710D	LHCA3	The light-harvesting complex (LHC) functions as a light receptor, it captures and delivers excitation energy to photosystems with which it is closely associated	1.70E-153	548.5
SPC	BnaC09g20020D	LIL	Lil3 protein	5.70E-135	486.9
SPC	BnaA09g13650D	LIS	S-( )-linalool synthase	6.80E-106	390.2
SPC	BnaA09g13640D	LIS	S-( )-linalool synthase	3.90E-62	244.2

SPC	BnaC05g31340D	NFU4	NIFU-like protein	5.80E-152	543.5
SPC	BnaA09g03600D	NHX1	Sodium hydrogen exchanger	7.00E-300	1035.8
SPC	BnaA09g03590D	NOP5	Nucleolar protein	9.70E-249	865.9
SPC	BnaA08g00080D	NOP56	Nucleolar protein	3.00E-247	860.9
SPC	BnaA05g05760D	P5CS	P5CS plays a key role in proline biosynthesis, leading to osmoregulation in plants	0	1382.9
SPC	BnaC03g64690D	PER1	Per1-like family	1.30E-52	212.2
SPC	BnaC03g64700D	PER1	Per1-like family	4.10E-64	250.8
SPC	BnaA09g04240D	PWD1	Phosphoglucan, water dikinase	0	2219.5
SPC	BnaC09g19620D	PYL1	Polyketide cyclase / dehydrase and lipid transport	6.60E-99	366.7
SPC	BnaC09g04920D	PYR4	Belongs to the dihydroorotate dehydrogenase family. Type 2 subfamily	3.20E-264	917.1
SPC	BnaC03g62330D	RHA3B	Ring-H2 finger	2.30E-91	341.7
SPC	BnaC03g62560D	RPB1	DNA-dependent RNA polymerase catalyzes the transcription of DNA into RNA using the four ribonucleoside triphosphates as substrates	1.10E-06	59.3
SPC	BnaC03g64530D	RPB2	DNA-dependent RNA polymerase catalyzes the transcription of DNA into RNA using the four ribonucleoside triphosphates as substrates	0	1143.3

SPC	BnaC03g64540D	RPB2	DNA-dependent RNA polymerase catalyzes the transcription of DNA into RNA using the four ribonucleoside triphosphates as substrates	2.20E-117	428.3
SPC	BnaC03g63980D	RPB2	DNA-dependent RNA polymerase catalyzes the transcription of DNA into RNA using the four ribonucleoside triphosphates as substrates	0	2248.8
SPC	BnaC05g46250D	RPL22	60s ribosomal protein	3.50E-61	240.7
SPC	BnaA09g03840D	RPS21	Belongs to the eukaryotic ribosomal protein eS21 family	2.30E-40	171
SPC	BnaC07g47340D	RPS25	40s ribosomal protein	2.10E-36	158.3
SPC	BnaC05g46270D	RPT5A	Belongs to the AAA ATPase family	1.80E-237	828.2
SPC	BnaC06g32810D	SERK1	Belongs to the protein kinase superfamily. Ser Thr protein kinase family	1.40E-294	1018.5
SPC	BnaA09g13310D	STM	ELK	4.40E-187	660.6
SPC	BnaC03g63800D	TAF11	tbp-associated factor	4.90E-39	167.5

SPC	BnaA02g27980D	THI1	Involved in biosynthesis of the thiamine precursor thiazole. Catalyzes the conversion of NAD and glycine to adenosine diphosphate 5-(2-hydroxyethyl)-4-methylthiazole-2-carboxylic acid (ADT), an adenylated thiazole intermediate. The reaction includes an iron-dependent sulfide transfer from a conserved cysteine residue of the protein to a thiazole intermediate. The enzyme can only undergo a single turnover, which suggests it is a suicide enzyme. May have additional roles in adaptation to various stress conditions and in DNA damage tolerance	2.80E-154	551.2
SPC	BnaA08g00200D	TIM50	Mitochondrial import inner membrane translocase subunit	8.40E-183	646.4
SPC	BnaA09g13040D	TIM9	Belongs to the small Tim family	1.00E-44	185.7
SPC	BnaA08g02120D	UBC20	protein modification by small protein conjugation	3.00E-63	247.7
SPC	BnaC03g64120D	VPS28	Component of the ESCRT-I complex (endosomal sorting complex required for transport I), a regulator of vesicular trafficking process	3.30E-104	384.4

SPC	BnaA09g14140D	VPS34	Belongs to the PI3 PI4-kinase family	0	1629
SPC	BnaC07g47230D	WRKY13	transcription factor	3.10E-145	521.2
SPC	BnaA09g13370D	WRKY6	Transcription factor	1.10E-290	1005.4
SPC	BnaC09g19320D	WRKY8	transcription factor	8.40E-139	500
SPC	BnaC03g64360D	ZAC	ADP-ribosylation factor GTPase-activating protein	1.30E-164	585.9
SOC	BnaC06g14830D	AFC1	ATP binding kinase protein kinase protein serine threonine kinase protein tyrosine kinase	0	1515.4
SOC	BnaC03g62970D	AG	Floral homeotic protein	1.10E-125	456.1
SOC	BnaA07g16990D	AREB3	ABSCISIC ACID-INSENSITIVE 5-like protein 2	4.90E-113	414.1
SOC	BnaA07g17010D	AREB3	ABSCISIC ACID-INSENSITIVE 5-like protein 2	1.50E-44	186
SOC	BnaC02g12420D	ARP8	cytoskeleton organization	2.80E-268	930.6
SOC	BnaA02g01640D	ATP5	ATP synthase	1.10E-124	452.6
SOC	BnaA07g17230D	BG3	Belongs to the glycosyl hydrolase 17 family	2.60E-191	674.5
SOC	BnaC03g64810D	BGLU3	Belongs to the glycosyl hydrolase 1 family	1.20E-100	372.5
SOC	BnaC03g64790D	BGLU5	Belongs to the glycosyl hydrolase 1 family	1.60E-94	352.8
SOC	BnaC09g20030D	BI-1	Belongs to the BI1 family	5.00E-123	447.2
SOC	BnaA07g17200D	CBF5	H ACA ribonucleoprotein complex subunit	1.00E-245	855.9
SOC	BnaA07g17150D	CBL	cystathionine beta-lyase	2.60E-26	124.8
SOC	BnaC02g12710D	CBL2	Calcineurin B-like protein 2	3.70E-125	454.1
SOC	BnaA07g17040D	CRD1	magnesium-protoporphyrin IX monomethyl ester (oxidative) cyclase	7.50E-225	786.2
SOC	BnaC09g19210D	CUL4	Belongs to the cullin family	0	1461.4

SOC	BnaC03g63590D	CYCT1	Belongs to the cyclin family	9.90E-286	988.8
SOC	BnaC09g18860D	CYP707A1	Belongs to the cytochrome P450 family	2.10E-271	941
SOC	BnaC09g37430D	DCL4	Belongs to the helicase family. Dicer subfamily	0	3107.4
SOC	BnaA02g01720D	DCP2	hydrolase m7G(5')pppN diphosphatase mRNA binding protein homodimerization	3.10E-173	614.4
SOC	BnaC09g37350D	EFR	Belongs to the protein kinase superfamily. Ser Thr protein kinase family	4.40E-224	784.6
SOC	BnaC03g63040D	ERD3	AT4G19120 (E 0.0) ERD3 ERD3 (early-responsive to dehydration 3)	0	1253
SOC	BnaC02g12340D	EXPA14	Rare lipoprotein A (RlpA)-like double-psi beta-barrel	1.30E-147	528.9
SOC	BnaC09g18650D	FAD7	omega-3 fatty acid desaturase	1.90E-174	618.2
SOC	BnaC02g13190D	FTSZ1-1	Cell division protein FtsZ homolog 1	6.40E-227	793.1
SOC	BnaC02g12750D	HSP81-3	heat shock protein	6.70E-272	943
SOC	BnaC02g12680D	HSP81-3	heat shock protein	9.60E-37	159.5
SOC	BnaA02g01650D	IQD11	IQ-domain	3.60E-184	651
SOC	BnaC09g19280D	KAS1	Belongs to the beta-ketoacyl-ACP synthases family	1.10E-269	935.3
SOC	BnaC09g19900D	LBA1	Regulator of nonsense transcripts 1 homolog	0	1402.1
SOC	BnaC09g20020D	LIL	Lil3 protein	5.70E-135	486.9
SOC	BnaA02g01700D	LKHA4	peptidase M1 family protein	0	1206.4

SOC	BnaC09g29160D	ndhH	NDH shuttles electrons from NAD(P)H plastoquinone, via FMN and iron-sulfur (Fe-S) centers, to quinones in the photosynthetic chain and possibly in a chloroplast respiratory chain. The immediate electron acceptor for the enzyme in this species is believed to be plastoquinone. Couples the redox reaction to proton translocation, and thus conserves the redox energy in a proton gradient	3.30E-140	504.2
SOC	BnaC09g29150D	ndhH	NDH shuttles electrons from NAD(P)H plastoquinone, via FMN and iron-sulfur (Fe-S) centers, to quinones in the photosynthetic chain and possibly in a chloroplast respiratory chain. The immediate electron acceptor for the enzyme in this species is believed to be plastoquinone. Couples the redox reaction to proton translocation, and thus conserves the redox energy in a proton gradient	1.10E-71	275.8
SOC	BnaC05g44530D	NS2	asparagine-tRNA ligase	0	1132.9

SOC	BnaA07g17160D	ORC5	Origin recognition complex (ORC) subunit 5 C-terminus	2.90E-290	1003.8
SOC	BnaC03g64700D	PER1	Per1-like family	4.10E-64	250.8
SOC	BnaC03g64690D	PER1	Per1-like family	1.30E-52	212.2
SOC	BnaA02g01880D	PPH	Alpha/beta hydrolase family	8.90E-269	932.6
SOC	BnaC09g19620D	PYL1	Polyketide cyclase / dehydrase and lipid transport	6.60E-99	366.7
SOC	BnaA06g03620D	RACK1A	Guanine nucleotide-binding protein subunit beta-like protein	1.70E-133	482.3
SOC	BnaC02g13240D	RAN1	GTP-binding protein involved in nucleocytoplasmic transport. Required for the import of protein into the nucleus and also for RNA export. Involved in chromatin condensation and control of cell cycle	5.00E-115	420.6
SOC	BnaC09g37280D	RBX1	RING-box protein	9.20E-59	232.6
SOC	BnaC03g62330D	RHA3B	Ring-H2 finger	2.30E-91	341.7
SOC	BnaC03g62560D	RPB1	DNA-dependent RNA polymerase catalyzes the transcription of DNA into RNA using the four ribonucleoside triphosphates as substrates	1.10E-06	59.3
SOC	BnaC03g64530D	RPB2	DNA-dependent RNA polymerase catalyzes the transcription of DNA into RNA using the four ribonucleoside triphosphates as substrates	0	1143.3

SOC	BnaC03g63980D	RPB2	DNA-dependent RNA polymerase catalyzes the transcription of DNA into RNA using the four ribonucleoside triphosphates as substrates	0	2248.8
SOC	BnaC03g64540D	RPB2	DNA-dependent RNA polymerase catalyzes the transcription of DNA into RNA using the four ribonucleoside triphosphates as substrates	2.20E-117	428.3
SOC	BnaC02g12070D	RPL31	60s ribosomal protein	5.40E-59	233.4
SOC	BnaC02g12080D	RPS30	Belongs to the eukaryotic ribosomal protein eS30 family	5.70E-31	140.2
SOC	BnaA07g17050D	SIP2-1	Belongs to the MIP aquaporin (TC 1.A.8) family	2.00E-129	468.4
SOC	BnaC09g37040D	SUS1	Sucrose-cleaving enzyme that provides UDP-glucose and fructose for various metabolic pathways	0	1638.2
SOC	BnaC03g63800D	TAF11	tbp-associated factor	4.90E-39	167.5
SOC	BnaC09g29060D	TAP46	TAP42-like family	2.20E-200	704.9
SOC	BnaA06g03690D	TOC64-1	Translocon at the outer membrane of chloroplasts 64-III	2.10E-300	1037.7
SOC	BnaC02g13170D	TOP1	DNA Topoisomerase I (eukaryota)	0	1242.6
SOC	BnaC09g29430D	TRY	transcription regulator recruiting activity	3.20E-50	204.1
SOC	BnaC09g28960D	UGT84B2	Belongs to the UDP-glycosyltransferase family	4.30E-28	131.3

SOC	BnaC03g64120D	VPS28	Component of the ESCRT-I complex (endosomal sorting complex required for transport I), a regulator of vesicular trafficking process	3.30E-104	384.4
SOC	BnaC09g19320D	WRKY8	transcription factor	8.40E-139	500
SOC	BnaC09g29130D	ycf1	BEST Arabidopsis thaliana protein match is Ycf1 protein (TAIR	0	1855.1
SOC	BnaC09g29210D	ycf1	BEST Arabidopsis thaliana protein match is Ycf1 protein (TAIR	5.40E-184	650.2
SOC	BnaC09g29140D	ycf1	BEST Arabidopsis thaliana protein match is Ycf1 protein (TAIR	7.50E-152	543.5
SOC	BnaC03g64360D	ZAC	ADP-ribosylation factor GTPase-activating protein	1.30E-164	585.9
GSL	BnaC08g29410D	WRKY69	transcription factor	3.70E-84	317.8
GSL	BnaC05g10200D	WRKY4	Transcription factor	3.40E-218	764.2
GSL	BnaA08g06570D	UPF3	ATUPF3, UPF3 Smg-4 UPF3 family protein	6.30E-113	414.1
GSL	BnaA08g06560D	UPF3	ATUPF3, UPF3 Smg-4 UPF3 family protein	3.60E-258	897.1
GSL	BnaC03g21990D	UPF2	Up-frameshift suppressor 2	0	2032.7
GSL	BnaC02g13170D	TOP1	DNA Topoisomerase I (eukaryota)	0	1242.6
GSL	BnaC05g10850D	SOS1	sodium hydrogen	0	1388.2
GSL	BnaC03g36650D	SIR1	Sulfite reductase	0	1297.3
GSL	BnaC03g36500D	SEC8	exocyst complex	0	1897.9
GSL	BnaC05g10920D	SDS	Belongs to the cyclin family	3.40E-308	1063.5
GSL	BnaC06g24140D	SDP1	Transcription factor	3.70E-227	794.3

GSL	BnaC06g24350D	RPSa	Required for the assembly and or stability of the 40S ribosomal subunit. Required for the processing of the 20S rRNA- precursor to mature 18S rRNA in a late step of the maturation of 40S ribosomal subunits	1.60E-165	588.6
GSL	BnaC02g12080D	RPS30	Belongs to the eukaryotic ribosomal protein eS30 family	5.70E-31	140.2
GSL	BnaC02g12070D	RPL31	60s ribosomal protein	5.40E-59	233.4
GSL	BnaC05g10550D	RPL10	Ribosomal protein L16p/L10e	1.20E-126	459.1
GSL	BnaC05g10980D	RDR1	Probably involved in the RNA silencing pathway and required for the generation of small interfering RNAs (siRNAs)	0	2184.5
GSL	BnaC03g36360D	RAPTOR1	Regulatory-associated protein of TOR	0	2610.9
GSL	BnaC02g13240D	RAN1	GTP-binding protein involved in nucleocytoplasmic transport. Required for the import of protein into the nucleus and also for RNA export. Involved in chromatin condensation and control of cell cycle	5.00E-115	420.6
GSL	BnaC07g18200D	RAN1	Copper-transporting ATPase	0	1847.4
GSL	BnaC02g13600D	PRPL24	Belongs to the universal ribosomal protein uL24 family	1.00E-102	379.4

GSL	BnaC02g10920D	PHOT2	FMN binding blue light photoreceptor kinase protein serine threonine kinase	0	1666.4
GSL	BnaA06g28960D	NOP5	Nucleolar protein	1.00E-250	872.5
GSL	BnaC03g12430D	NADP-MDH	lactate/malate dehydrogenase, alpha/beta C-terminal domain	1.70E-251	874.8
GSL	BnaC06g24580D	MAK16	Belongs to the MAK16 family	1.70E-109	402.5
GSL	BnaC02g12680D	HSP81-3	heat shock protein	9.60E-37	159.5
GSL	BnaC02g12750D	HSP81-3	heat shock protein	6.70E-272	943
GSL	BnaA08g06530D	HLIP	One-helix protein	6.40E-69	266.9
GSL	BnaC03g36740D	GYRA	DNA gyrase subunit A	0	1750.7
GSL	BnaC02g13190D	FTSZ1-1	Cell division protein FtsZ homolog 1	6.40E-227	793.1
GSL	BnaC03g36510D	FLD	flowering locus	0	1677.5
GSL	BnaC03g37090D	FAD7	omega-3 fatty acid desaturase	4.90E-267	926.4
GSL	BnaC02g12340D	EXPA14	Rare lipoprotein A (RlpA)-like double-psi beta-barrel	1.30E-147	528.9
GSL	BnaC03g30750D	EOL1	Tetratricopeptide repeats	0	1604.7
GSL	BnaC02g11150D	EIF4G	translation initiation factor	0	1469.9
GSL	BnaC03g37030D	DREB2A	Dehydration-responsive element-binding protein	5.60E-47	194.5
GSL	BnaA04g06630D	CYP83A1	cytochrome P450	1.50E-291	1008.1
GSL	BnaA07g27740D	CRC	YABBY protein	5.70E-68	263.8
GSL	BnaC03g21760D	CPK20	calcium-dependent protein kinase	0	1088.6
GSL	BnaC03g36720D	CPK2	calmodulin-domain protein kinase cdpk isoform 2	1.50E-258	898.7
GSL	BnaC03g36730D	CPK2	calmodulin-domain protein kinase cdpk isoform 2	2.40E-181	641.3

GSL	BnaC04g46710D	COI1	coi1 coi1 (coronatine insensitive 1)	0	1198.3
GSL	BnaC02g13300D	CNX5	Plays a central role in 2-thiolation of mcm(5)S(2)U at tRNA wobble positions of cytosolic tRNA(Lys), tRNA(Glu) and tRNA(Gln). Also essential during biosynthesis of the molybdenum cofactor. Acts by mediating the C-terminal thiocarboxylation of sulfur carriers URM1 and MOCS2A. Its N-terminus first activates URM1 and MOCS2A as acyl-adenylates (-COAMP), then the persulfide sulfur on the catalytic cysteine is transferred to URM1 and MOCS2A to form thiocarboxylation (-COSH) of their C-terminus. The reaction probably involves hydrogen sulfide that is generated from the persulfide intermediate and that acts as nucleophile towards URM1 and MOCS2A. Subsequently, a transient disulfide bond is formed. Does not use thiosulfate as sulfur donor	5.80E-258	896.3
GSL	BnaC03g36530D	CBP3	serine carboxypeptidase-like 49	3.40E-291	1006.9
GSL	BnaC02g12710D	CBL2	Calcineurin B-like protein 2	3.70E-125	454.1
GSL	BnaC03g22180D	BIK1	belongs to the protein kinase superfamily	5.90E-227	793.1

GSL	BnaC03g07000D	BCCP2	first, biotin carboxylase catalyzes the carboxylation of the carrier protein and then the transcarboxylase transfers the carboxyl group to form malonyl-CoA	2.20E-97	362.1
GSL	BnaC05g10870D	BBR/BPC1	basic pentacysteine	8.20E-256	889.4
GSL	BnaC03g36760D	ASD1	Alpha-L-arabinofuranosidase	0	1370.5
GSL	BnaC02g12420D	ARP8	cytoskeleton organization	2.80E-268	930.6
GSL	BnaA05g01310D	ARF1	Auxin response factors (ARFs) are transcriptional factors that bind specifically to the DNA sequence 5'-TGTCTC-3' found in the auxin-responsive promoter elements (AuxREs)	0	1151
GSL	BnaA07g27710D	AP1	transcription factor that promotes early floral meristem identity in synergy with APETALA1, FRUITFULL and LEAFY. Is required subsequently for the transition of an inflorescence meristem into a floral meristem. Seems to be partially redundant to the function of APETALA1	6.50E-126	456.8
GSL	BnaC08g29530D	AG	Agamous-like MADS-box protein	4.60E-67	260.4
GSL	BnaC08g29520D	AG	Agamous-like MADS-box protein	1.00E-66	259.2
GSL	BnaA07g27660D	ACR4	ACT domain	6.10E-252	876.3
GSL	BnaC03g30470D	ABP1	Auxin binding protein	3.50E-111	407.5

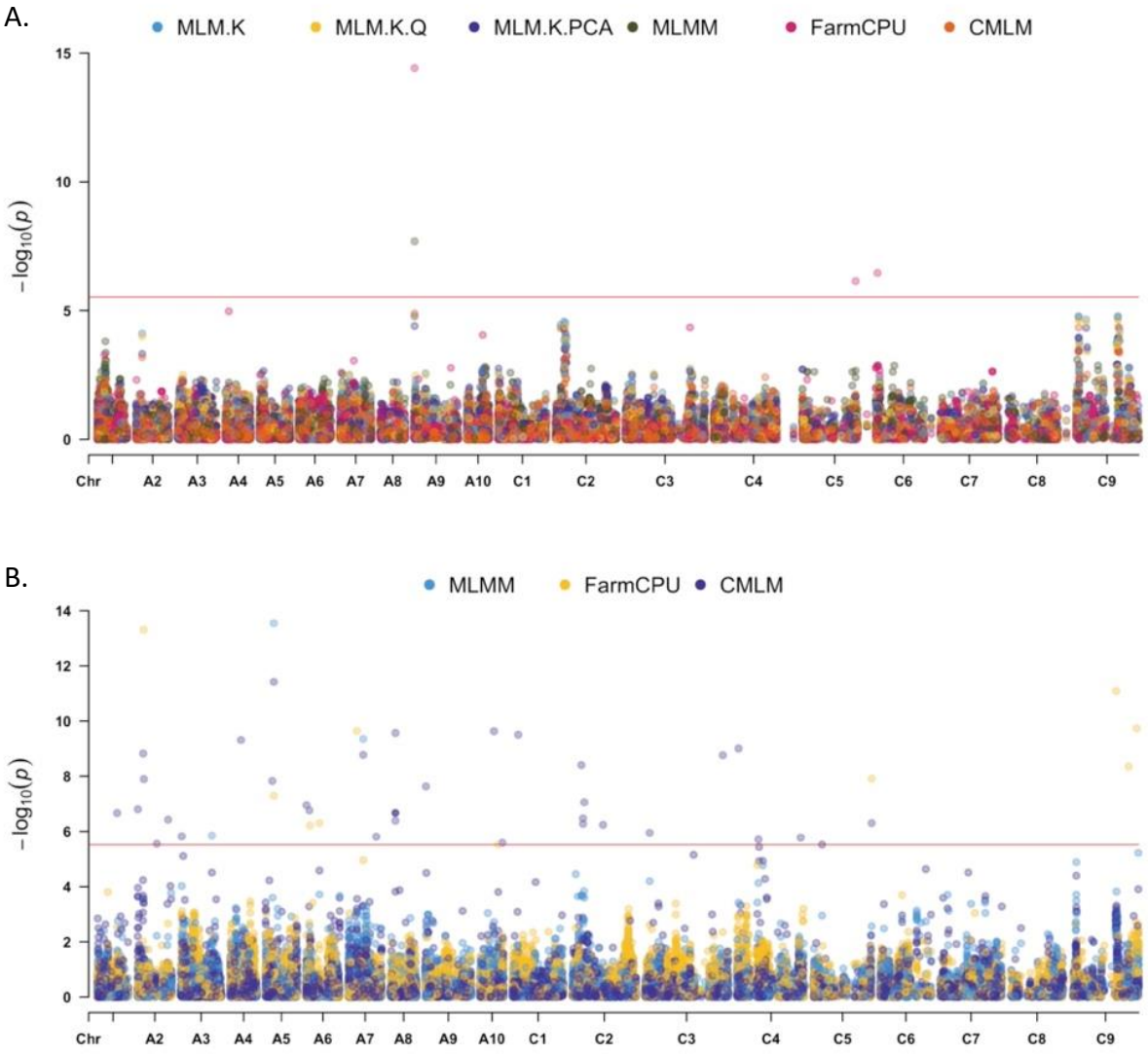


Figure S3.2 Manhattan plots showing plant height (HT) based on the MS-2 (16,855 SNP markers). Bonferroni-corrected significance threshold was shown as the red horizontal line at -log10 (0.05/16855) =5.53. (A) Results from the Brassica napus L. parental population based on six models including mixed linear models considering kinship (MLM+K), mixed linear models considering subpopulation structure via Bayesian clustering (MLM+K+Q), mixed linear models considering subpopulation structure via principal component analysis (MLM+K+PCA), multi-loci mixed model (MLMM), Fixed and random model circulating probability unification (FarmCPU) and compression mixed linear model (CMLM). (B) Results from the *Brassica napus* L. combined population based on three models MLMM, FarmCPU and CMLM.

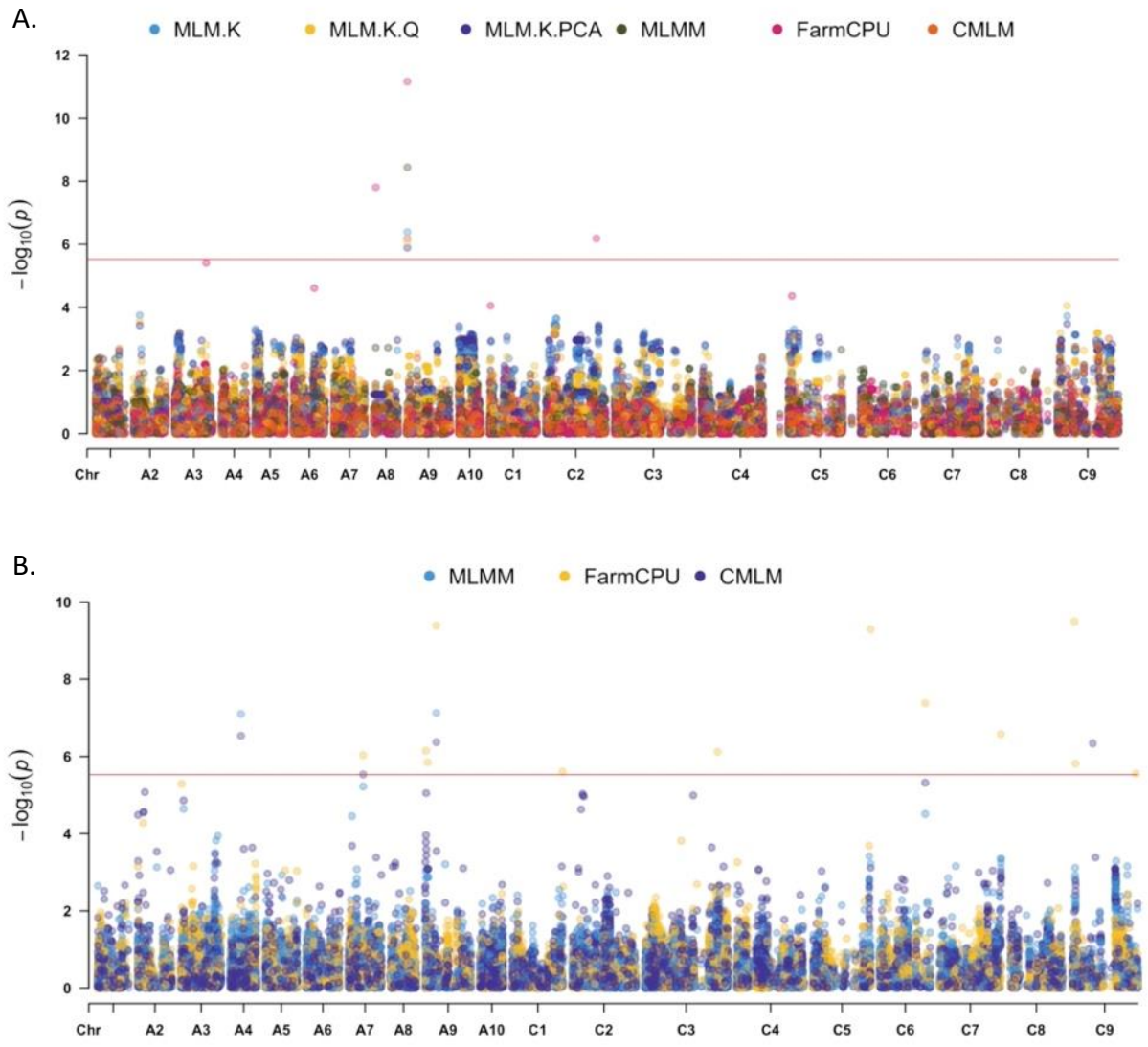


Figure S3.3 Manhattan plots showing seed protein content (SPC) based on the MS-2 (16,855 SNP markers). Bonferroni-corrected significance threshold was shown as the red horizontal line at -$\log_{10}$(0.05/16855) =5.53. (A) Results from the *Brassica napus* L. parental population based on six models including mixed linear models considering kinship (MLM+K), mixed linear models considering subpopulation structure via Bayesian clustering (MLM+K+Q), mixed linear models considering subpopulation structure via principal component analysis (MLM+K+PCA), multi-loci mixed model (MLMM), Fixed and random model circulating probability unification (FarmCPU) and compression mixed linear model (CMLM). (B) Results from the *Brassica napus* L. combined population based on three models MLMM, FarmCPU and CMLM.

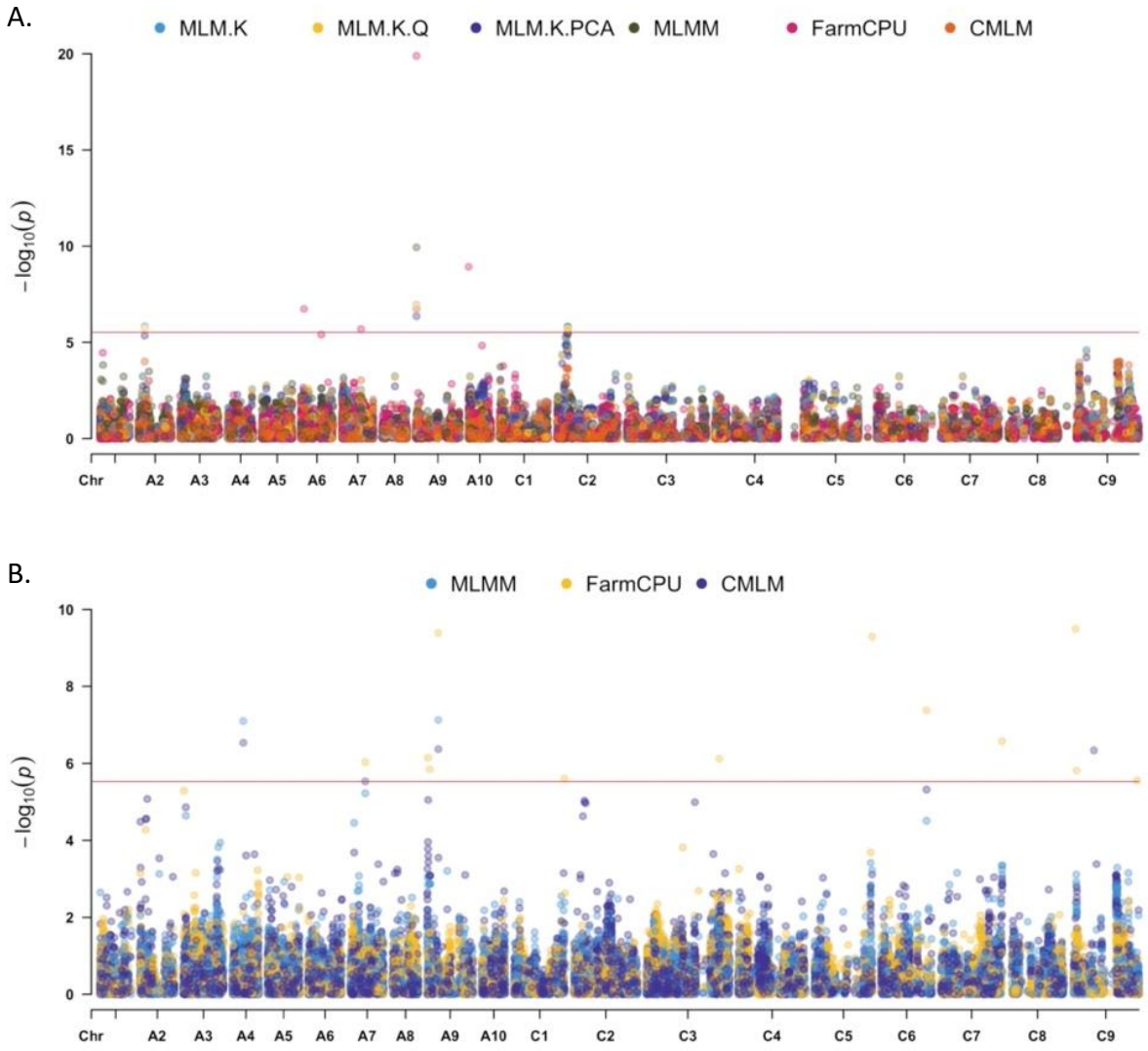


Figure S3.4 Manhattan plots showing seed oil content (SOC) based on the MS-2 (16,855 SNP markers). Bonferroni-corrected significance threshold was shown as the red horizontal line at $-\log_{10}(0.05/16855) = 5.53$. (A) Results from the *Brassica napus* L. parental population based on six models including mixed linear models considering kinship (MLM+K), mixed linear models considering subpopulation structure via Bayesian clustering (MLM+K+Q), mixed linear models considering subpopulation structure via principal component analysis (MLM+K+PCA), multi-loci mixed model (MLMM), Fixed and random model circulating probability unification (FarmCPU) and compression mixed linear model (CMLM). (B) Results from the *Brassica napus* L. combined population based on three models MLMM, FarmCPU and CMLM.

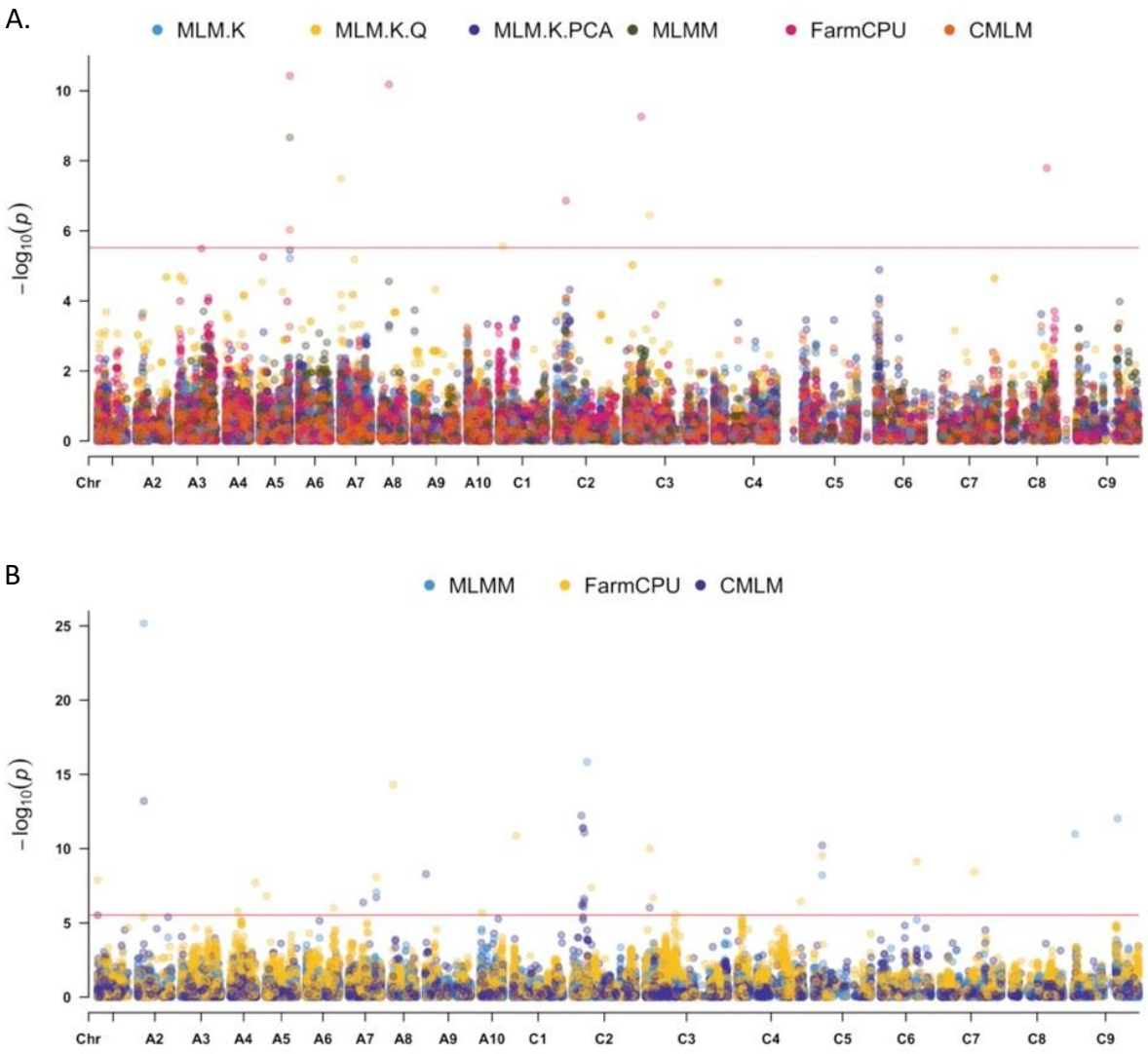


Figure S3.5 Manhattan plots showing seed glucosinolate content (GSL) based on the MS-2 (16,855 SNP markers). Bonferroni-corrected significance threshold was shown as the red horizontal line at $-\log_{10}(0.05/16855) = 5.53$. (A) Results from the *Brassica napus* L. parental population based on six models including mixed linear models considering kinship (MLM+K), mixed linear models considering subpopulation structure via Bayesian clustering (MLM+K+Q), mixed linear models considering subpopulation structure via principal component analysis (MLM+K+PCA), multi-loci mixed model (MLMM), Fixed and random model circulating probability unification (FarmCPU) and compression mixed linear model (CMLM). (B) Results from the *Brassica napus* L. combined population based on three models MLMM, FarmCPU and CMLM.