Examination Commission

Chairperson: Prof. Dr. Gesine Lühken

Supervisor: Prof. Dr. Rod Snowdon

Co-supervisor: Prof. Dr. Richard Nichols

Examiner: Prof. Dr. Matthias Frisch

Examiner: Prof. Dr. Karl-Heinz Kogel

Date of Defense: April 20th, 2016

Dedication

This thesis work is dedicated to the following people:

- To my loving parents (*Ammi and Jee*), my brothers (*Asim Sheikh and Maaz Sheikh*), sisters (*Sarah Sheikh and Sumaira Sheikh*), all other members of my family and all my teachers for their unceasing good wishes and prayers during my stay abroad.
- To my wife *Azra Noor* for her endless support during our stay in Germany and our sweet daughter *Rimsha Habib Sheikh* for being a source of joy to us.
- To all those unsung heroes in the world who work hard day and night for the benefit and welfare of humanity.

Table of Contents

List of figures

List of tables

Chapter 1: General Introduction

1.1 Plant breeding in the modern era

Conventional plant breeding has been practiced for hundreds of years by mankind benefiting from the available genetic diversity found in nature and is still being vehemently pursued today with modern tools and techniques. Darwin writings (published in 1859) and Mendel's discovery of the laws of inheritance laid the scientific foundations of crop plant breeding (Borlaug, 1983). Pedigree based breeding approach to estimate the individual's genetic merit has been used by plant breeders since early 1900 (Henderson, 1975; Koebner and Summers, 2003; Piepho et al. 2007; Piepho, 2009). This trend has been extremely beneficial with the integration of modern agricultural practices in the provision of a sustainable supply of novel cultivars and resulted in a tremendous increase in the yield of majority of crops (Moose and Mumm, 2008; Prohens, 2011). The famous 'Green Revolution' is a living example of the recent history where new cultivars were developed through selective breeding approaches that could give higher yields and thus, exponentially elevating the amount of food production worldwide to serve the global society (Jain, 2010). Since the launching of the 'Green Revolution' in the past half century, the world annual yield of cereals, coarse grains, roots and tubers, pulses and oil crops has increased from 1.8 billion tonnes to 4.6 billion tonnes (FAO, 2011). Despite all the miracles and benefits of the classical breeding methodology, there are however, some challenges and limitations attached to such approach. The predominant limitations include the following: the large amount of time required to declare a new variety, arduous and expensive phenotyping, and lengthy breeding cycles (Fehr, 1987; Resende et al. 2012). Furthermore, genetic evaluations based on conventional pedigree based selection failed to account for individual alleles and thus, potentially ignore the 'Mendelian segregation' that explains half of the genetic variability under (Fisher, 1918) 'infinitesimal additive model' and in the absence of inbreeding (Avendano et al. 2005, van Raden, 2008). In pursuing the development of this

important area of agricultural science in the 20th century, plant breeders' started selecting phenotypically superior plants and crossed these plants to get rather new and improved varieties of various crops with desirable traits, better adapted to environmental fluctuations. Thus, they potentially started taking advantage of cross-breeding or hybridisation and introduced 'modern' plant breeding. This 'modern' plant breeding was also manifested through mutation breeding, plant tissue culture and other novel selection approaches.

More recently,thanks to the relatively inexpensive genomic technologies, plant breeding entered into a new modern era of genome-based molecular breeding. There is a great opportunity for plant breeders and researchers to make use of the available diverse genome sequence information for designing novel varieties, better resistant to biotic and abiotic stresses (Snowdon and Iñiguez-Luy, 2012). The active use of molecular markers in practical plant breeding started in the 1980s (Stuber et al. 1980; Stuber et al. 1982; Tanksley et al. 1988). Initially, these markers were used to detect quantitative trait loci (QTL) that led to the discovery of few genes responsible for variation in the trait of interest; typically a small number of loci were detected for each trait (Soller and Plotkin-Hazan 1977; Soller 1978; Paterson et al. 1988, Lander and Botstein, 1989). This practice is still being pursued by various labs around the globe. But traditional standard QTL mapping is limited by both mapping resolution and capturing the genetic variation (Pasam et al. 2012; Cai et al. 2014). The unprecedented availability of cost-effective molecular markers in huge quantity due to the unparalleled advancement in high-throughput technologies make it easier to detect and exploit DNA polymorphism in various plant species. These DNA polymorphisms play an important role in association mapping where molecular markers are associated with trait phenotype, and is a promising approach for the dissection of complex polygenic traits (Snowdon and Iñiguez-Luy, 2012; Li et al. 2014). Single nucleotide polymorphisms (SNPs) are widely used in genomic analyses. SNP markers are considered

as the ‘choice of markers’ due to their abundance, stability and genome-wide occurrence. Hayward et al. (2012) found that SNP markers can be used effectively to investigate genetic diversity in crops with ancient genome duplications, i.e. *Brassica napus*.

1.2 *Brassica napus* L. (Canola/Oilseed rape/Rapeseed)

Ancient civilisations in Asia and Mediterranean have been reported to have grown *Brassica* vegetables and oilseeds. In India, their uses, especially Brassica *rapa* have recorded as early as 4000 BC and later on some 2000 years ago, it spread into other Asian territories, for example, China and Japan. In Europe, it has been primarily cultivated for its uses as oil for lamps since the 13th century and then, during World War II, the production of canola or rapeseed increased when the oil was used especially for margarine production (Snowdon et al. 2006). From a commercial view point, *Brassica napus* was first grown in Canada in 1942 to be used as lubricant for warships and subsequently, in Australia, their cultivation on commercial scale has started in 1969. Before the introduction of modern *Brassica napus*, it was considered unsuitable for use as food for either humans or other animals in the western world (The Biology of *Brassica napus*, 2008).

The *Brassica* genus belongs to *Brassicaceae* family (which was formerly known as *Cruciferae*) and consists of approximately 100 species, including *Brassica napus* L. *Brassica napus* is divided into two subspecies, Swedes (*B.napus* ssp. *napo brassica*) and (*B.napus* ssp. *napus*) that includes winter, spring oilseed, fodder and vegetable rape (Snowdon et al. 2006) . *Brassica napus* L., (AACC, 2n=38) which is also commonly known as oilseed rape/rapeseed or canola, ranks among the second largest oilseed crop after soybean in the world. Winter-sown canola is also used as a sustainable source of biodiesel in Europe (Snowdon and Friedt, 2012). Germany is the second largest producer of canola within Europe after France with a total production of 4.8 million tons in 2012. In 2012,

canola was cultivated on 1.306 million hectares in Germany, where the total arable land was 12 million hectares (FAO, 2013).

This allopolyploid was naturally formed only about ~7500 years ago from spontaneous inter-specific hybridisations between cabbage (*Brassica oleracea* CC, 2n=18) and turnip rape (*Brassica rapa*; AA, 2n=20) (Chalhoub et al. 2014) (Figure 1.1). Strong 'conscious' selection by breeders for important quality traits resulted modern double-low (00) varieties with low levels of seed glucosinolate (<30 µmol/g seed) and erucic acid (<2 %). These modern breeding materials of rapeseed however, have developed a narrow gene pool in response to continuous selection by breeders for some quality traits (Hasan et al. 2006; Bus et al. 2011).

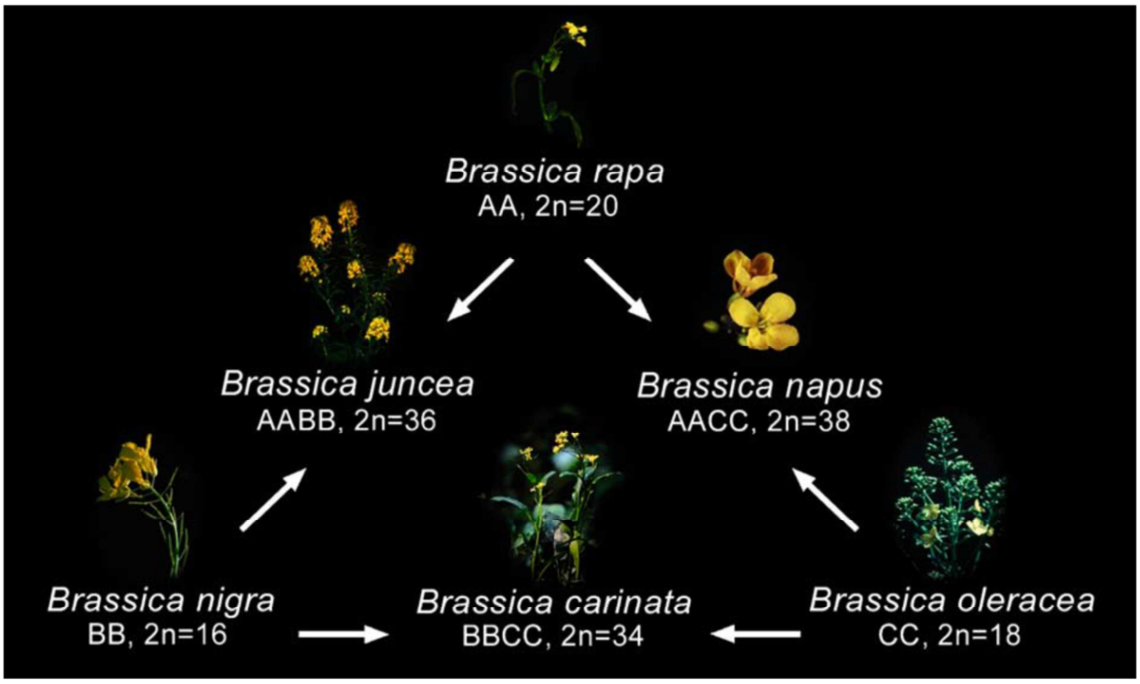


Figure 1.1: The *Brassica* triangle. *Brassica napus* (2n= 38) arose from spontaneous interspecific hybridisation between its two diploid progenitors. Source: Snowdon, (2007).

According to Cowling (2007), the effective population size in Australian spring-type canola is just N_e =11, that indicates the overall shallow genetic diversity of *Brassica napus*. Low

genetic diversity often leads to high level of biotic and abiotic susceptibility and weak response to selection (Kebede et al. 2010; Falconer and Mackay, 1996). The narrow genetic pool in *Brassica napus* poses a great challenge for breeders to not only boost up the genetic diversity to withstand environmental fluctuations, but at the same time improved productivity and seed quality (Snowdon and Iñiguez-Luy, 2012). Introgression of diverse germplasm into elite materials can minimise genetic losses (Haussmann et al. 2004). China is the biggest producer of rapeseed, where attempts were made to elevate the genetic diversity in *Brassica napus* through marker-assisted introgressions from the related species and thus, to increase heterosis (Zou et al. 2010, Mei et al. 2011, Xiao et al. 2012). Development of new heterotic pools in *Brassica napus*, particularly through marker-assisted introgressions of novel germplasm from the diploid progenitors or other exotic gene pools, is an astute strategy to overcome this problem (Qian et al. 2005, Basunanda et al. 2007, Zou et al. 2010, and Girke et al. 2012).

Molecular genetic approaches are coupled with the classical genetic analysis to advance our understanding of the inheritance of various qualitative and quantitative traits, mapping and estimation of the causative genes and their effects and for the identification of DNA-based markers associated with traits of agronomic importance in modern *Brassica* breeding programmes (Snowdon and Friedt, 2004). Therefore, the use of molecular markers, especially in marker-assisted selection (MAS) breeding and other advanced molecular breeding plays a key role.

SNP markers are abundantly found across the genome and can be used in association and genome-wide prediction in *Brassica napus* due to their genome-wide distribution, availability and simplicity in scoring their genotype (Snowdon and Iñiguez-Luy, 2012).

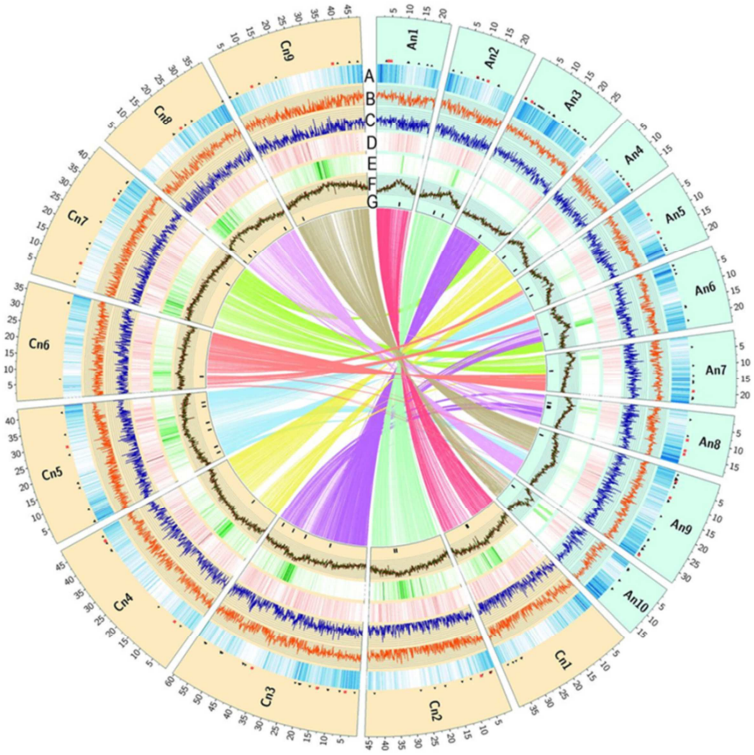


Figure 1.2: The reference genome of the *B. napus* 'Darmor-bzh 'cultivar. Subgenome C has 9 chromosomes and subgenome A has 10. Circular tracks represent (A) gene density (B and C) Transcription states in (B) leaves and (C) roots, (D) DNA transposon density. (E) Retrotransposon density. (F) CpG methylation in leaves (green) and roots (brown); both curves are overlapping. (G) Centromeric repeats. Homeologous relationship between chromosomes of subgenome A and C shown with coloured connecting lines coloured according to the Cn chromosomes. Source: Chalhoub et al. (2014).

Genome-based prediction is a highly promising method to integrate novel, unadapted genetic variation effectively in canola breeding (Snowdon and Iñiguez-Luy, 2012). Recently genomic prediction has been demonstrated for estimation of testcross performance in different crop species, for example, in maize (Albrecht et al. 2011, Zhao et al. 2012), sugar beet (Hofheinz et al. 2012) and rye (Wang et al. 2014).

The recent publication of the reference genome of *Brassica napus* cv. Darmor-bzh in Science (Chalhoub et al. 2014) (Figure 1.2) and the availability of 60k SNP Infinium

consortium array (Illumina Inc., San Diego, CA; USA) opens up the opportunities for further genomic research in exploring the genetic framework, comparative genomics and better understanding the genomic landscape and development of the allopolyploid *Brassica napus* on a population-scale.

1.3 Genome-wide association studies (GWAS) and its role in molecular plant breeding

Genome-wide association studies (GWAS) have emerged as a powerful approach in genetics for the dissection of complex traits. It is the statistical correlation of molecular markers with phenotypic variations in a natural population (Nordborg and Weigel, 2008; Mandel et al. 2013; Li et al. 2014). Association mapping also known as linkage disequilibrium mapping (LD; non-random association of alleles between loci), is based on the exploitation of the historical recombination events occurred in the ancestral population during the evolutionary or cladistics phase (Pasam et al. 2012). GWAS is a well-established approach in the context of human genetics and has successfully unraveled many positive associations between common variants and complex diseases, but is still challenged by the 'missing heritability' problem where perhaps thousands of individuals and millions of molecular markers would be required to detect majority of the QTLs. In plant breeding, however, the problem of 'missing heritability' appears to be less severe because common genetic variants explain the major proportion of the phenotypic variations (Brachi et al. 2011). In tracking the genetic underpinnings of complex traits, various studies of GWAS were carried out in different plant species, for example, in *Arabidopsis thaliana*, rice, maize etc. (Atwell et al. 2010, Zhao et al. 2011; Tian et al. 2011; Huang et al. 2012). These studies explained much greater proportion of phenotypic variance than that explained in human GWAS and hence, GWAS approaches in plant breeding are more successful (Brachi et al. 2011).

The downside in genome-wide association studies (GWAS) is the 'spurious or false' associations between genetic markers and the trait of interest. It has already been diagnosed that cryptic population structure is one of the main causes of fake causal relations (Li, 1969; Lander and Schork, 1994). Prithard et al. (2000) inferred population structure based on a Bayesian clustering approach (STRUCTURE). They assumed a model with *K* populations where individuals were assigned to different populations on the basis of their genotypes and at the same time estimating the allele frequencies of the population. Patterson et al. (2006) introduced a new technique to examine population structure in genetic data through the use of principal component analysis (Cavalli Sforza and Feldman, 2003) that determines the statistically significant 'axes of variation'. However, all these various approaches have had limited success in dealing with this issue effectively (Pritchard et al. 2000; Price et al. 2006; Yu et al. 2006). Variation at the DNA level can provide enough information about the underlying population structure apart from the conventional approaches of pedigree or phenotypic records (Varshney et al. 2005). This knowledge of population structure can play an important role in organising the efficient exploitation of germplasm in crop breeding (Bus et al. 2011). GWAS has the potential to be used directly in plant breeding programmes (Jannink et al. 2010).

There are some ancestrally conserved regions in the genome with low recombination rate that are inherited in a 'block-like' fashion and are termed as 'haplotype-blocks'. These haplotype blocks have potential applications in genome-wide association mappings (Jeffreys et al. 2001; Stumpf, 2004). Haplotype diversity has been used to exploit genetic diversity in different crops, i.e. rice and maize (Cho et al. 2008; Zhang et al. 2013).

Heterosis or hybrid vigour is a complex phenomenon and is manifested with the improved performance of a hybrid compared to its inbred parental lines (Shull, 1908). Hybrid breeding has been instrumental in the exploitation of heterosis by plant breeders for improved

agronomic traits, especially for seed yield gain and yield stability (Duvick, 1999). Rapeseed has well-defined pollination control systems, for example, cytoplasmic male sterility system (CMS), genic male sterility system (GMS), etc. and can be used for the efficient production of hybrid seed (Buzza, 1995; Renard et al. 1997). A QTL based approach was also used in rapeseed to identify chromosomal regions in rapeseed contributing to heterosis for yield and its components (Radoev et al. 2008), but this approach has its own limitations, i.e. the difficulty to generate segregating populations in some species, low genetic diversity and limited recombination, etc. Thus, GWAS can be used effectively to avoid such pitfalls.

In hybrid breeding system, male inbred lines are crossed with genetically distant female lines or ‘testers’ and general combining ability (GCA) is estimated. Information on GCA plays an important role in a breeder’s decision making to identify a viable hybrid (Beck et al. 1990). GCA information has been used recently in various GWAS and genome-wide prediction studies (Riedelsheimer et al. 2012; Riedelsheimer et al. 2013; Reif et al. 2012; Zhao et al. 2013). Qian et al. (2007) reported that, in canola hybrid breeding, additive gene effects are the main contributors to heterosis, rather than specifc combining ability (SCA). GCA accounts for additive gene effects.

1.4 Genomic selection (GS) in the context of plant breeding

The seminal paper of Meuwissen et al. (2001) paved the way for genomic selection (GS) in the context of animal breeding. This trend is also gaining impetus in plant breeding for the prediction of unphenotyped materials (Heffner et al. 2009; Jannink, 2010). GS is a major step forward in marker assisted selection (MAS), where instead of identifying individual genes with larger effects, the effects of high density genetic markers anchored across the whole-genome of an organism are simultaneously estimated (Meuwissen et al. 2011; Kumar et al. 2012; Zhao et al. 2011 and Asoro et al. 2011). In GS, the genotype of an organism is employed to predict the unknown phenotype, as opposed to conventional plant

and animal breeding approach where only phenotypic data on the individuals, their close relatives and their progenies would be used for the prediction of the genotype of the next generation (Meuwissen et al. 2001). Thus, primarily in GS a ‘black box‘approach is followed where no prior knowledge regarding the effects of molecular markers is required (Jonas and de-Koning, 2013). Through GS, rapid genetic gain due to early pre-selection and high prediction accuracy (defined as the correlation between the true observed trait value and predicted trait value) can be obtained (Resende et al. 2012; Meuwissen et al. 2001). Unlike linkage and association mapping, GS estimates breeding values on the basis of dense molecular markers spanning over the whole genome, rather than just estimating the effects of individual genes (Jannink et al. 2010). A large number of selected candidates can be computationally evaluated without any field trials and thus, dramatically reducing the breeding cycle (Lorenz et al. 2011). Goddard and Hayes (2007) further argued that because GS has the ability to effectively capture the realised relationship matrix (RRM) compared to the average relationship matrix, it predicts breeding values more accurately than selection based on only pedigree and phenotype based performance. This ‘RRM’ is based on the assumption that a very large number of genes throughout the genome have small effects from a common distribution on the trait in question, linking to the ‘infinitesimal model’. Various simulations and empirical studies confirm that GS is an optimum way to increase genetic gain and speed up the breeding cycle, as compared to QTL derived markers and phenotypic selection (Heslot et al. 2012). Various experimental works indicate that in the coming time, GS will play an essential role to improve crop breeding (Cabrera-Bosquet et al. 2012).

GS utilizes training and validation or breeding populations (Asoro et al. 2011, Zhao et al. 2011, and Heffner et al. 2009) (Figure 1.3). A training population includes both phenotypic and genomic data, whereas the validation population is represented only by genomic data

and the effects for the genetic markers estimated in the training population (Heffner et al. 2009; Goddard and Hayes, 2007).

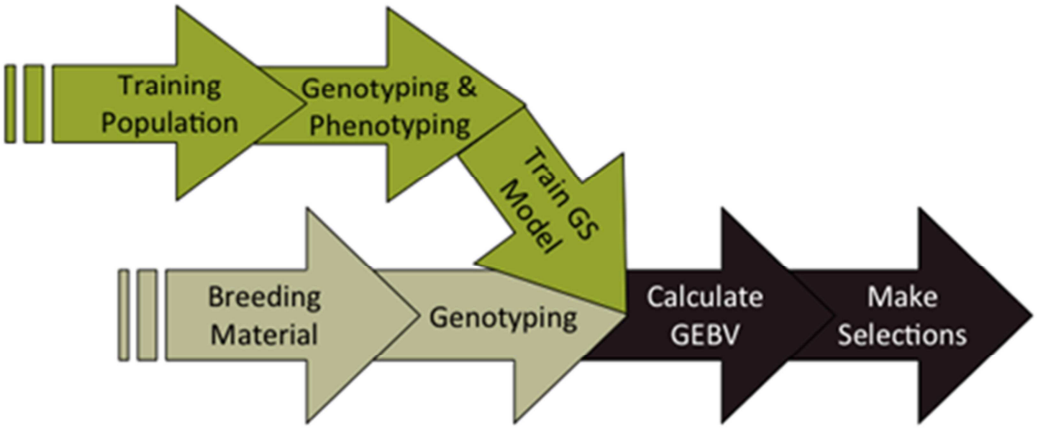


Figure 1.3: General diagram of genomic selection (GS) process. It starts with training population (genotypes & phenotypes) and finally genomic estimated breeding values based selection. Note: This is an example of single model training which is repeated as new phenotype and marker data accumulate. Source: Heffner et al. (2009).

Thus, genomic estimated breeding values (GEBVs) are calculated for non-phenotyped individuals using only genomic data from the training population (Meuwissen et al. 2001). As the generations increase between the training population and validation population, the accuracy of prediction decreases in both traditional pedigrees based Best Linear Unbiased Prediction (BLUP) and genomic BLUP (Meuwissen et al. 2001; Sonessen and Meuwissen, 2009).

The structure of the training population has an important role in affecting persistency of accuracies. If the individuals in the training population are closely related, this means larger parts of the chromosomes near the QTLs will be shared among them and ultimately enabling the more distant markers to explain QTL variation within the training population. As the recombination rates are higher between these distant markers and the QTLs, so the predictive value will go down quickly compared to the markers lying in the close vicinity of

the QTLs (Sved, 1971; and Bastiaansen et al. 2012). Various studies have reported higher reliabilities of genomic predictions where individuals in the validation population were closely related to the training population (Meuwissen, 2009; Habier et al. 2010; Makowsky et al. 2011).

The efficiency of GS is, however, susceptible to various factors, including heritability (h^2) and genetic architecture of the underlying trait, size and structure of the population (Meuwissen et al. 2001; Calus and Veerkamp, 2007; Calus et al. 2008; Goddard, 2009; Solberg et al. 2009; and Coster et al. 2010; and Daetwyler et al. 2010).

The reliability of genome-wide predictions depends on the existence of LDbetween specific alleles of SNPs and QTLs (Meuwissen et al. 2001). The stronger the LD, the higher the accuracy of GS expected (Calus et al. 2008; Solberg et al. 2008). But over generations, LD between QTL and SNP tend to decrease, therefore, re-estimation of SNPs effects become important for the reliability of GS in more recent generations (Muir, 2007). As the distance between the two markers on a chromosome increases, so does the decay of LD intensity. The extent of LD depends on the number of effective population size (N_e) in a population (Sved, 1971). In case of lower N_e, the level of kinship among individuals is high, resulting stronger LD (Falconer and Mackay, 1996).

Increase in LD may be due to inbreeding, small population size, population subdivision, admixture, low recombination rate and balancing selection, etc. while a decrease in LD may be as a result of outcrossing, high recombination rate and high mutation rate and drift (Flint-Garcia, 2003, Gupta et al. 2005). Therefore, evolutionary changes, mating, population size, rate of recombination and selection sweeps have a direct bearing on the rate and the trend of LD (Gaut and Long, 2003). For example, in a small population size, drift operates rapidly leading to a stronger LD. This stronger LD, however, can be balanced by recombination at

equilibrium, causing decay in LD. Therefore, the appropriate choice of marker densityhas a close relationship with the effective population size (Lorenz et al. 2011).

The commonly used LD measures are the D (Lewontin and Kojima, 1960), D' Lewontin (1964) and r^2 (Hill and Robertson (1968) and Franklin and Lewontin (1970). However, r^2 ('the square of the correlation coefficient between the two indicator variables; one representing the presence or absence of a particular allele at the first locus and the other representing the presence or absence of a particular allele at the second locus'), is the most relevant to GS. In GS, every gene affecting the trait of interest is in LD with at least one marker (Hayes et al. 2013; Kumar et al. 2012). LD can be used as an indication to determine the optimum marker density required for the higher accuracy of GEBV. In the case of high linkage, LD will persist with time, but in the absence of linkage, LD decays faster (Mackay and Powell, 2007).

1.5 Genomic selection as predictive strategy

Plant breeders are interested in evaluating inbred lines not on their *per se* performance, but by the optimum performance of the hybrids it produces. Inbred lines are crossed with genetically distant testers to estimate the general combining ability (GCA). Thus, for the accurate GCA prediction of complex traits, information on the parental lines are used (Riedelsheimer et al. 2012). Therefore, selection of the best combination of parental lines to produce optimum F1 hybrids in crops like rapeseed and maize leads to 'predictive' breeding programmes based on genome-wide molecular markers (Edwards et al. 2013). The general aim of the plant breeders is to introduce new varieties that are cost-effective, sustainable and render higher yields with nutritional qualities. The efficient use of genome-wide SNP markers can help breeders enhance better productivity. But the real challenge for both plant and animal breeders of today is, to effectively correlate these genomic data to traits of economic importance by utilising novel algorithms and computational tools (Crossa et al.

2010; and Snowdon and Iñiguez-Luy, 2012). GS takes care of the inbreeding in a breeding population because genome-wide molecular markers increase the individual information available. On the other hand, in classical breeding, information on the relatives is used that leads to co-selection of the shared genes and thus, increasing inbreeding (Meuwissen et al. 2013). The germplasm in hybrid breeding is comprised of different genetically distant heterotic pools. These genetic distances can be assessed with molecular markers along with a relatively new technique of gene expression profiling known as transcriptome analysis. Frisch et al. (2010) showed that the prediction of hybrid performance can be determined in a more precise way with the addition of transcriptome data.

Developing performance prediction statistical models based on multivariate analysis of genome-wide molecular markers are of utmost importance to harness the effects of genome-wide markers. Nevertheless, selecting an appropriate model to gain enough prediction accuracy can both be tricky and challenging (Pérez et al. 2010). In GS, to achieve maximum genomic estimated breeding values (GEVB) accuracy, the statistical models should have the ability to address many marker effects simultaneously with limited phenotypes (Heffner et al. 2011). In breeding value estimation, the addition of polygenic effects in the model helps in capturing the effects of small quantitative trait loci which are normally unaccounted for by markers having big effects (Goddard and Hayes, 2007). According to Zhao et al. (2012), inclusion of epistasis in the genomic model could also enhance the prediction accuracy within populations. Different genomic statistical models have been reported in various GS studies on plants and animals. The main purpose of using these proposed models in GS is to minimise the ‘curse of dimensionality’ due to marker effects because the number of molecular markers (*p*) is always much larger than the number of observations or phenotypes (*n*). This leads to a situation of lack of degrees of freedom, therefore, estimation of marker effects through multiple regressions by ordinary

least square (OLS) methods is not possible. In such a scenario, penalised regression methods for the marker effects are introduced, i.e. ridge regression best linear unbiased estimates (RR-BLUP) and Bayesian approaches for the correct estimation of marker effects (Pérez et al. 2010).

There is a distinct difference between RR-BLUP and Bayesian approaches based on their prior assumptions. In RR-BLUP, the prior assumption is based on the equal marker variance across the whole-genome and this approximates Fisher's famous 'infinitesimal model'. Bayesian methods are sufficiently flexible to allow a different variance for each marker effect, for example, in the case of Bayes A and for Bayes B, where marker effects are drawn from a distribution with different variances (Meuwissen et al. 2001). In the RR-BLUP approach, the genetic effects are assumed to be drawn from a normal distribution, while in Bayes B, the effects are drawn from the Student's t-distribution with lower and wider tails than the normal distribution. This enables it to capture the effects from extreme genetic values (Pungpapong et al. 2012). Bayes A models variances as an inverse chi squared distribution; the conjugate prior to the variance of a normal distribution (Meuwissen et al. 2001).

Using real plant datasets, Crossa et al. (2010) studied wheat (*Triticum aestivum*) and maize (Zea mays) and evaluated Bayesian LASSO (least absolute shrinkage and selection operator), BLUP and RKHS (reproducing kernel Hilbert space) methods for GS. They reported that the model where molecular data was used along with pedigree data performed far better than the model where only pedigree data was used, in terms of higher prediction ability for selected economic traits under multi-environmental conditions. Hofheinz et al. (2012) reported in sugar beet (*Beta vulgaris* L.), that the prediction accuracy achieved through cross validation in one breeding cycle may not be used as a positive indicator to predict accuracy in the subsequent breeding cycle.

For polygenic traits that follow closely the pattern of an ‘infinitesimal model’, the use of RR-BLUP methods is favourable. In contrast, the use of the Bayesian approach is appropriate wherever traits controlled by a single or a few genes of larger effects, for example, some monogenic resistance traits in plants (Zhang et al. 2009, Albrecht et al. 2011). For the effective implementation of GS, different R packages (R Development core team, 2014) have also been introduced and developed, i.e. rrBlupMethod6, rrblup, BLR (Bayesian Linear Regression) etc. (Piepho et al. 2012; Endelman, 2011; and de los Compos et al. 2010). In a study, Heslot et al. (2012) compared eleven GS models and assessed the accuracy of their predictive ability by using empirical datasets from wheat (*Triticum aestivum* L.), barley (*Hordeum vulgare* L.), *Arabidopsis thaliana* (L.), and maize (*Zea mays* L.). They concluded that, there is no single ’all purpose’ GS model or models combination to be used as a master model in plant breeding. However, the selection of the statistical model depends heavily on the genetic architecture of the trait, population structure and molecular marker density (Zhao et al. 2012).

GS is still not widely adopted in plant breeding as an alternative strategy; this may be due to insufficient understanding and the challenges of GS to be used in practical plant breeding (Nakaya and Isobe, 2012). However, the availability of sequence information in different crops will bring a paradigm shift in the breeding of these crops in the near future. Trends towards GS in important crops like Wheat (*Triticum aestivum* L.), Maize (*Zea mays* L.) and Barley (*Hordeum vulgare* L.) etc. has already started with higher prediction accuracies compared to phenotypic selection or pedigree based selections (Lorenza and Bernardo, 2009; de los Compos et al. 2009; Crossa et al. 2010; Heffner et al. 2011, Crossa et al. 2013).

In comparison to classical hybrid crops like maize, in which genetically distinct heterotic pools have been established over many decades of hybrid breeding, there are no such

clear heterotic pools available within canola germplasm. Development of new heterotic pools within adapted germplasm types, particularly through marker-assisted introgression of novel germplasm from the diploid progenitors or other exotic gene pools, is an important strategy to overcome this problem (Qian et al. 2005; Basunanda et al. 2007; Girke et al. 2012; Snowdon et al. 2015). In a hybrid breeding programme, efficient selection of the most promising combinations between male and female parental lines is a vital step to avoid expensive field testing of poor performing hybrids (Reif et al. 2013). This becomes particularly important in crops like canola where the absence of distinct genetic pools prohibits an *ab initio* assumption of heterotic potential between any two potential hybrid parents. Various studies have reported methods for optimum exploitation of heterosis in crop breeding using both morphological and molecular marker data (Schrag et al. 2006; Riedelsheimer et al. 2012). Piepho (2009) described how the performance of untested hybrids can also be predicted effectively using genomic selection methodology.

Technical difficulties associated with the development of male-sterile lines in canola generally lead breeders to choose relatively small panels of maternal lines. On the other hand, some of the most widely used male-sterility systems have the benefit that all known *B. napus* accessions are restorers, so that testcross performance with available maternal lines is an important selection criterion for breeding of pollinators.

Even in the absence of heterotic pools, genomic prediction of testcross performance is a highly promising method in canola breeding to select promising germplasm for advancement into male-sterile maternal lines or fertility restorers (Snowdon and Iñiguez-Luy, 2012). Recently genomic prediction has been demonstrated for estimation of testcross performance in various crops, for example in maize (Albrecht et al. 2011; Zhao et al. 2012; Windhausen et al. 2012), sugar beet (Hofheinz et al. 2012) and rye (Wang et al. 2014).

1.6 Thesis scope and aims

Genome-based modern predictive breeding approaches in the context of spring-type or spring-grown hybrid *Brassica napus* L. will be described in this thesis. The main objectives of this thesis were threefold. First, I explored the population structure and linkage disequilibrium decay across both A- and C- subgenomes of spring-type canola based on molecular marker data. Then, I aimed to elucidate the genetic underpinnings behind hybrid performance in spring-type canola through genome-wide association study (GWAS) and haplotype analysis by using both DNA and general combining ability (GCA) information. Lastly, genomic prediction of complex agronomic traits related to seed yield and seed quality traits to estimate testcross performance in hybrid canola was conducted. My aims in particular were:

(1) To Explore population structure and the extent of LD decay across both A- and C-subgenomes in spring-type canola
(2) The use of the GWAS approach to trace genomic regions involved in hybrid performance bearing candidate genes of pleiotropic effects.
(3) To investigate the effectiveness of genomic prediction of F1 testcross performance in *Brassica napus.*
(4) To Examine the effect of training population sample size on the prediction accuracy
(5) To describe the overall potential of genomic selection in the context of canola hybrid breeding.

Chapter 2: Material and methods

2.1 Genetic material

The experimental material used in this thesis comprised a diverse population of spring-type *Brassica napus* with double-low seed quality (low erucic acid, low glucosinolate content) from a commercial canola breeding programme. The material was carrying introgressions from the diploid progenitors of *B. napus*. Two representative male sterile female testers from a pool of testers carrying the Male Sterility Lembke (MSL) sterility system (*NPZ* Lembke, Hohenlieth, Germany) were crossed with a total of 475 pollinators to generate seed from 950 F1 hybrids.

2.2 Phenotype data

The 950 testcrosses were evaluated at four different locations across Denmark, Germany, Poland and Estonia during the 2012 growing season through our commercial partner *NPZ* Lembke, Hohenlieth, Germany. Un-replicated trials were performed in each of the four environments within Europe by *NPZ* Lembke for various traits of commercial importance.

2.3 Best linear unbiased estimators (BLUEs)

Best linear unbiased estimator (BLUE) values for each trait using their respective phenotype values were generated. The restricted maximum likelihood (REML) method was used to estimate variance components assuming a random effect model. BLUE values were estimated for each trait, treating genotype as a random effect with locations as fixed effect. The Pearson's correlation coefficient (*r*) was calculated between all the seven traits.

All calculations were performed using the statistical software package SPSS Statistics for Windows Version 22.0 (IBM Corp., Armonk, NY, USA).

2.4 Broad sense heritability (H^2)

Broad sense heritability which is the ratio of genotypic to phenotypic variance was calculated for each trait following the method reported in Bekele et al. (2014).

$$H^2(\%) = \left[\sigma^2_g / (\sigma^2_g + \sigma^2_\varepsilon / n) \right] \times 100$$

where σ^2_g is the genotypic variance, σ^2_ε is the estimated error variance, and n is the number of locations. Estimates of error variance were divided by the number of locations.

2.5 Genotype data

Each of the 475 pollinator lines and the two testers were genotyped using the *Brassica* 60k SNP Infinium consortium array (Illumina Inc., San Diego, CA; USA). Spring-type canola in the form of seed was obtained from our commercial partner *NPZ* Lembke, Hohenlieth, Germany. The seed was grown in the greenhouse of the Department of Plant Breeding, JLU Giessen. Upon germination, young leaves from each line were collected carefully. Genomic DNA was extracted from young leaf samples collected 20 days after sowing, shock frozen in liquid nitrogen and stored at -20°C until further processing. The DNA extractions were performed using a BioSprint 96 magnetic bead nucleic acid extraction robot (Qiagen, Hilden, Germany) according to the manufacturer's instructions. After fluorometric quantification of DNA concentrations using a Qubit 2.0 fluorometer (Life Technologies, Darmstadt, Germany), samples were diluted to 20ng/µl in sterile double distilled water, and quality checks of all DNA samples were carried out by gel electrophoresis on a 96 capillary Fragment Analyser (Advanced Analytical, Ames, IA, USA). Genotyping on the 60k *Brassica* Illumina SNP array was outsourced to TraitGenetics GmbH (Gatersleben, Germany).

2.6 *In silico* SNPs mapping and quality control

All the called SNPs were mapped to the *Brassica napus* cv. Darmor-bzh reference genome (Chalhoub et al. 2014) using the basic local alignment search tool (BLAST) with no mismatches permitted in the flanking oligonucleotides. All SNPs showing multiple BLAST hits or a non-random distribution were removed and a total of 28,286 single-position SNPs remained. Furthermore, all the SNP markers were removed with the following criteria (a) individuals having more than 20 % missing calls and SNP markers having more than 20 % missing calls across the whole panel (b) SNPs having less than 0.05 minor allele frequencies (MAF). Finally a total of 24,442 'unique', single-copy SNPs were considered for the downstream genomic analysis and predictions.

2.7 Determination of population structure

In the absence of clearly defined heterotic pools in *B. napus*, I analysed genetic relatedness between the parental lines using the genome-wide SNP data. Principal coordinate analysis (PCoA) was performed based on Roger's genetic distances (Roger, 1972) using the whole panel of 24,442 filtered, single copy SNPs.

Clusters of genetically related individuals were identified using the *K*-means method, following the algorithm of (Hartigan and Wong, 1979). A diagnosis of the optimal number of clusters in the dataset was also performed using the method described by (Caliński and Harabasz, 1974) and Saitou and Nei (1987). The software *SelectionTools* (Hofheinz and Frisch, 2014; www.uni-giessen.de/population-genetics/downloads) and R (http://www.r-project.org) were used for the PCA and *K*-means clustering.

2.8 Genome-wide association studies (GWAS)

2.8.1 General combining ability (GCA) estimates used in GWAS

F1 hybrid data on three important traits, including seed yield (dt/ha), flowering time or days to onset of flowering (DTF; measured as number of days from sowing until 50% flowering plants per plot) and seed oil content (% volume per seed dry weight) was considered during my GWAS for their respective trait heterosis.

Coefficient of variation was shown as a percentage for each trait. Following Becker's (2011) method, general combining ability (GCA) for each pollinator was estimated from the F1 hybrid BLUE values separately for each trait. In this study, GCA estimates for each pollinator were considered in the following GWAS and haplotype analyses instead of *per se* F1 hybrid BLUE values.

2.8.2 Linkage disequilibrium (LD)

LD between all pairs of markers was determined by calculating the squared of correlation coefficient (r^2) using software package TASSEL (Trait Analysis by aSSociation, Evolution and Linkage) version 2.1 (Bradbury et al. 2007). A total of 24,442 single positions, unique SNPs with MAF ≥ 0.05, were used to calculate chromosome-wide LD. To have a meticulous view of the extent of LD decay in both the subgenomes, we measured mean LD decay against the physical distance in KB and MB, chromosome-wise. Physical map distances for LD decay were divided into 20 different mean intervals ranging from 10 KB to 20 MB for the purpose of better visualisation.

LD decay plots against physical map distance were generated in Microsoft Excel for each chromosome separately in both A- and C- subgenomes. We followed a standard cutoff value of r^2= 0.1 for the comparison of LD decay. LD triangle blocks of the significant SNPs regions were generated using 'genetics' package in R version 1.3.8.1 (Warnes et al. 2012,

http://cran.r-project.org/web/packages/genetics/genetics.pdf). In cases of exceptional LD, r^2 > 0.40 (Hatzig et al. 2015), the whole block was surveyed for candidate genes. I also reported the number of polymorphic markers per chromosome in each A and C subgenomes and calculated SNP density (I divided the total number of SNPs per chromosome by its physical distance (bp)).

2.8.3 Genome-wide association mapping

The following mixed effects model was used to test associations between the SNP loci and the trait.

$$y = X\alpha + P\beta + K\mu + e$$

where y is the vector of phenotype observations, α is the vector of SNP effects, β is the vector of population structure effects, μ is the vector of kinship background effects, e is the vector of residual effects, P is the PCA matrix relating y to β, X and K are incidence matrices relating y to α and μ, respectively (Yu et al. 2006).

To avoid spurious signals (*type I errors*) due to subpopulation structure and to increase statistical power, we followed *P+K* method of the mixed effects model approximation where the population structure (*Q matrix*) was substituted by the principal components (*P matrix*) as suggested in Price et al. 2006. *K* represents kinship matrix. In this mixed effects model, principal components are treated as fixed effects and kinship as random effects (Yu et al. 2006). Mixed effects model takes into account both population structure and cryptic relationships and therefore, renders asmaller number of false positives compared to other models used in GWAS (Larsson et al. 2013). All these analyses were conducted using R '*GenABEL-package*' (Aulchenko et al. 2007). Quantile-quantile (Q-Q) and Manhattan plots for all the association analyses were generated in R package 'common' version 0.1.2 by Turner, 2014. Positive associations were corrected for multiple testing in each trait using the false discovery rate (FDR: $p \leq 0.05$) following the method of Benjamini and Hochberg

(1995). I used '*fdr tool'*, an R package to determine a significant threshold. All SNPs with a $-\log_{10}$ (*p*-value) > 2.5 were considered as significant.

My approach was to compare significant signals within different traits on all the chromosomes and then identify those SNP loci that were either in closely located regions (<500 Kb) or overlapping significant SNPs between the two traits on the same chromosomes. I then, investigated for strong LD blocks between or around the close-by or overlapping SNPs between the two traits and correlate those SNPs with genomic regions contributing to heterosis. I searched for candidate genes in the strong LD regions associated with significant SNPs in each of the traits chromosome-wise. *Brassica napus* candidate gene sequences were taken using *Brassica napus* genome browser *GENOSCOPE* (Chalhoub et al. 2014) and aligned (BLAST) these to the *Arabidopsis* gene database, the *Arabidopsis* information resource (TAIR). The idea was to search for *Brassica napus* orthologue genes in *Arabidopsis.*

2.8.4 Identification of significant SNP haplotype groups

I identified SNP haplotype diversity groups or blocks in the flanking regions of significant SNPs trait-wise, which could significantly contribute to hybrid performance. The haplotype block identification was done in the case of those chromosomes where I had close-by located or overlapping significant SNPs associated with candidate genes within strong LD regions on both the A- and C- subgenomes. I also calculated haplotype frequencies by dividing the number of specific haplotypes found in certain accessions by the total number of lines (475). Only haplotype diversity groups with frequency more than 1 % were considered and the rest were removed. Finally a significant 'Welch' two sample t-tests ($p \leq 0.05$) considering unequal variances were carried out between haplotypes within each chromosome. All the SNP haplotypes (hereafter referred to as Hap 1, Hap 2, Hap 3 etc.) corresponding to their respective phenotype values (GCA in this case) are shown in

boxplots using R. At the end, I reconstructed predicted F1 hybrid genotypes from genotypes of the significant haplotype groups of the respective male (pollinator) and female (tester) parental lines.

2.9 Genomic prediction

2.9.1 Phenotype data used in genomic prediction analysis

Seven important agronomic traits in my study were considered for the genomic predictions including seed yield (dt/ha), oil yield (dt/ha), seed oil content (% volume per seed dry weight), content of total seed glucosinolate (GSL; µmol/g seed), seedling emergence (visual observation ranging from a minimum value of 1 to maximum 9), lodging resistance (visual observation ranging from a minimum value of 1 to maximum 9) and days to onset of flowering (DTF; measured as number of days from sowing until 50% flowering plants per plot).

General combining ability (GCA) estimates were calculated from the F1 hybrid BLUE values for each trait following the method of Becker (2011), and were used as a phenotype matrix in the following genomic prediction analysis.

2.9.2 Scenarios for the genomic prediction of breeding values

Three independent scenarios based on the population structure were applied to estimate marker effects by genomic prediction. In scenario 1 the genomic prediction was performed across the whole population (WP). To investigate genomic prediction accuracy separately in the different genetic backgrounds of cluster 1 (C1) and cluster 2 (C2), respectively, we developed scenario 2 (prediction within C1) and scenario 3 (prediction within C2) (Figure 2.1). However, we did not directly compare prediction accuracies among these three prescribed scenarios due to confounding caused by their different TP and VP sizes, and rather reported them separately. Results from predictions within subpopulation C3 alone

are not reported due to the very small size of the test and validation populations in this case. We, further tested the prediction accuracy across the whole population using a model that included the population substructure as a covariate.

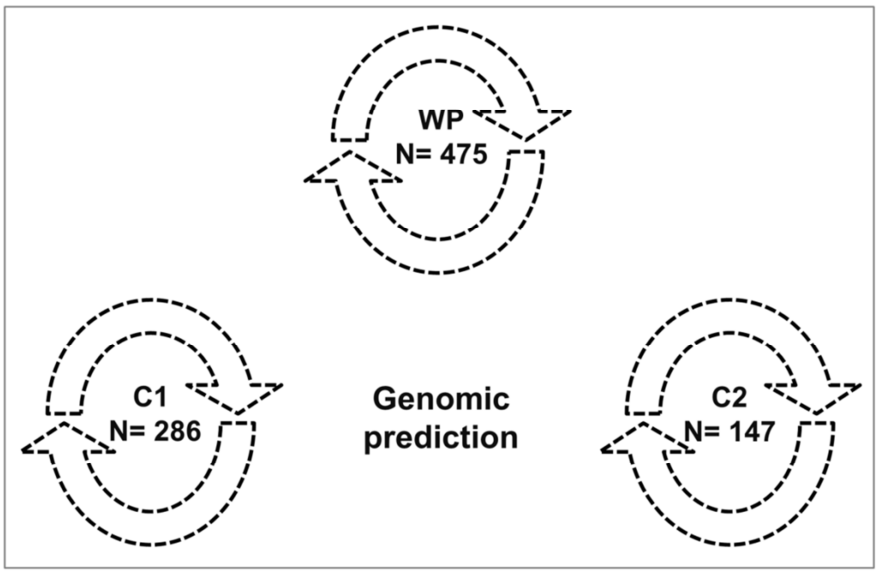


Figure 2.1: Schematic illustration of three independent genomic prediction scenarios. Genomic prediction across the WP (whole population) and genomic predictions within C1 (cluster 1) and C2 (cluster 2) separately are represented by dotted circular arrows.

2.9.3 Genomic prediction using the RR-BLUP mixed model

Genomic prediction accuracies were estimated using the RR-BLUP model described by (Whittaker et al. 2000; Meuwissen et al. 2001), assuming the same distribution of marker effects across the whole-genome. The following statistical model was used:

$$y = \mu + \sum_{i=1}^{N_m} X_i a_j + e,$$

where:

y is a Nx1 vector of phenotype (vector of BLUEs across locations);

μ is the overall mean;

N_m is the number of SNPs;

a_j is the effect of the j^{th} marker;

X_i is an Nx1 vector of genotypes (coded as 0,-1, +1) of the lines for each marker j, and variance of a_j is assumed to be uniformly distributed and is σ^2_G / N_m (Meuwissen et al. 2001).

2.9.4 Imputation of genotype data

Monomorphic SNPs and markers having more than 20% missing data were removed from the dataset. The rr-BLUP package in R (Endelman, 2011) was used to estimate genomic predictions with the remaining missing data replaced using the default method (mean imputation). Genomic prediction accuracy, denoted as r_{GPA} was calculated for each trait as the Pearson's correlation, $r_{(y, \hat{y})}$ between the estimated breeding values, ŷ, and observed BLUEs values, y, using the rr-BLUP package (Endelman, 2011).

2.9.5 Model cross validation

For determination of the optimum composition of training population size, we tested the prediction accuracies for each of the seven traits in the whole population under incremental increase of the training population from 10% up to 90 % of the 475 lines. Based on the results of this test (see below), the training population for all further analyses and scenario testing was set up at 70% of the total lines in the given dataset. Hence, in each run, the dataset was divided into a random 70 percent training population (TP) containing both genotyped and phenotyped data, and 30 percent validation (VP) or prediction population

having only SNP data and SNP effects with no consideration of phenotype values. For each scenario the data for each trait was cross-validated for 500 *iterations* and a mean value was subsequently considered.

Chapter 3: Results

3.1 Statistical analysis of the phenotype

3.1.1 Broad sense heritability and variance components

Broad sense heritabilities, along with first and second degree statistics for all the traits under consideration, are shown in Table 3.1. Heritability values ranged from 32% for seedling emergence to 90% of seed oil content. Seed oil content had the highest genetic variance and DTF had the lowest. Seed yield, however showed the highest coefficient of variation and DTF the lowest. Best linear unbiased estimators (BLUEs) of each trait followed approximately the normal distribution expected for quantitative traits. This was further confirmed by Q-Q plots drawn individually for each trait using R (Appendix I: Supplementary figures. A). The highest genetic variance was observed for seed oil content, while seedling emergence had the lowest genetic variance. As expected, a positive correlation was observed between oil yield and oil content (r = 0.66) followed by seed yield and oil yield (r = 0.57). Highly negative correlation was observed between seed oil content and seed GSL (r = -0.34) followed by correlation between seedling emergence and lodging resistance (r = -0.03) (Appendix I: Supplementary figures. B).

Table 3.1: First and second degree statistics for seed yield (dt/ha), oil yield (dt/ha), seed oil content (%), seed glucosinolate content (GSL; µmol/g), seedling emergence (visual observation scale 1-9; good=9), lodging resistance (visual observation scale 1-9; good=9) and days to onset of flowering (DTF) in field trials with 950 spring canola F1 testcross phenotypes in 4 independent field locations throughout Europe. δ^2_g : genetic variance, δ^2_ε : estimated error variance, H^2: broad sense heritability.

Traits	Mean	Min	Max	σ (SD)	σ^2_g	σ^2_ε	C_V (%)	H^2 (%)
Seed yield (dt/ha)	31.17	23.94	38.38	±1.97	1.56	7.95	6.3	44
Oil yield (dt/ha)	14.54	4.5	24.55	±1.38	0.56	0.95	9.5	70
Seed oil content (%)	48.41	44.08	52.73	±1.93	1.91	0.81	4	90
GSL (µmol/g)	9.22	6.91	11.57	±1.83	1.35	3.12	19.8	63
Emergence (good=9)	6.66	4.53	08.80	±0.48	0.048	0.41	7.2	32
Lodg.resistance (good=9)	7.16	5.05	08.65	±0.56	0.119	0.47	7.8	50
DTF	171.26	160.23	182.28	±1.57	0.808	5.26	0.92	38

3.2 Statistical analysis of the genotype data

3.2.1 Population structure among pollinators

The results of the principal component analysis (PCA) based on Roger's genetic distances (Roger, 1972) between the parental inbred lines using SNP markers are shown in Figure 3.1 (A). The molecular variance explained by the first three principal components comprised 25.12 %, 18.43% and 8.01%, respectively, making a total of 51.56 %.

The PCA indicated the existence of subpopulations within the dataset. The *K*-means clustering revealed a tendency to two main clusters and one relatively smaller cluster. This assumption was supported by the results of the Caliński-Harabasz (Caliński and Harabasz, 1974) clustering, which also suggested three optimum clusters, as shown in Figure 3.1(B) and Neighbour-joining (NJ) tree in Figure 3.1 (C). These clusters are subsequently referred to as cluster 1 (C1; n=286), cluster 2 (C2; n=147) and cluster 3 (C3; n=42), respectively.

3.2.2 Distribution of SNP markers and pattern of linkage disequilibrium decay

The use of large numbers of molecular markers greatly affects the resolution of genome-wide LD estimation (Van-Esbroeck and Bownam, 1998). Figure 3.2 shows the distribution of polymorphic markers per chromosome. The total number of SNPs on A- subgenome was 10,944 while on C- subgenome was 13,498. Chromosome A3 on subgenome A carried the highest number of SNPs (1637) with an SNP density of one SNP per 17.8 KB while chromosome A8 had the least number of SNPs (728) with an SNP density of one marker per 25.9 KB. Similarly, on subgenome C, highest number of SNPs (2523) was anchored on chromosome C4 with a SNP density of one SNP per 19 KB and C5 with the lowest number of SNPs (706) with one SNP per 60 KB (Table 3.2). Population sub-structure along with other several factors strongly influences LD (Flint-Garcia et al. 2003).

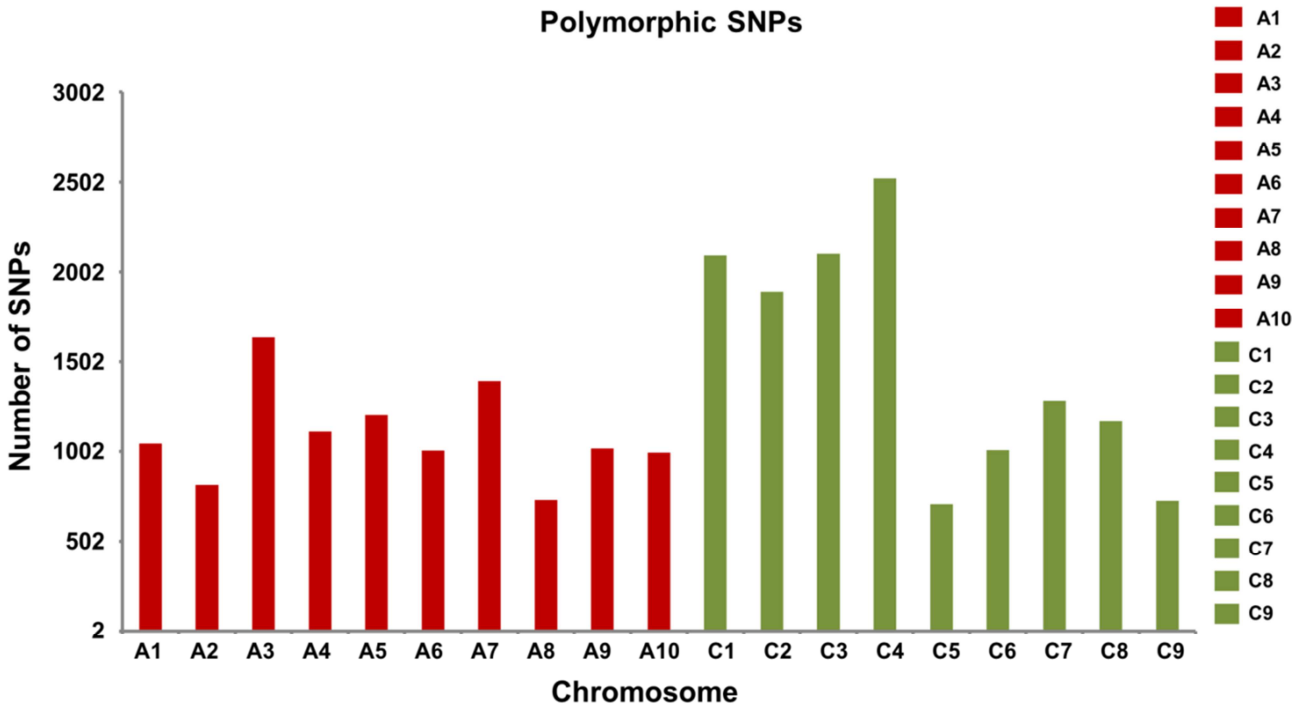


Figure 3.2: Distribution of polymorphic SNPs per chromosome in A- and C- subgenomes of spring-type *Brassica napus*.

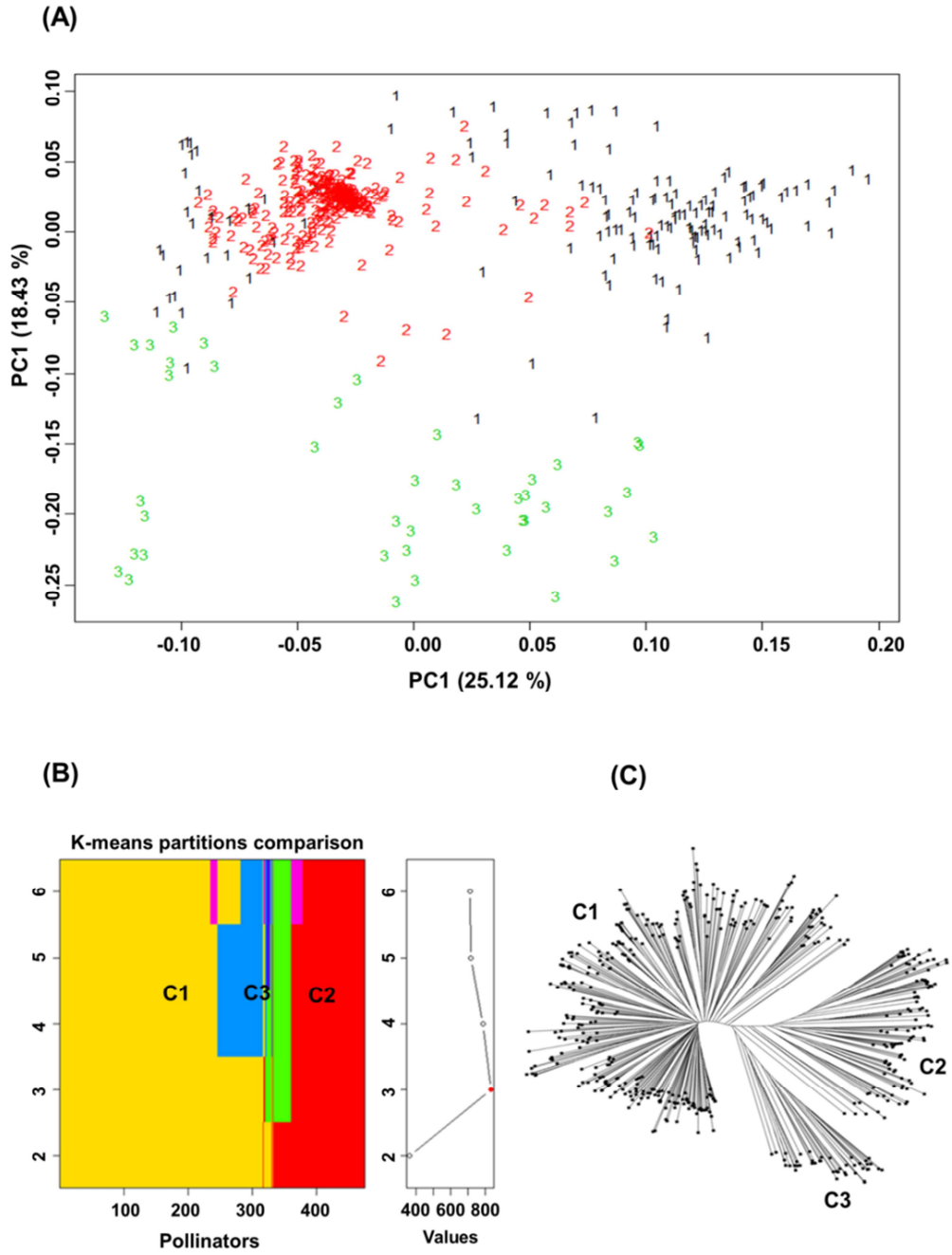


Figure 3.1: (A) Principal component analysis (PCA) among the population of 475 spring-type canola pollinators used for the testcross production. The PCA is based on Roger's distances estimated using a subset 3540 single nucleotide polymorphism (SNP) markers. The proportions of explained variance of principal components 1 and 2 are given in parentheses. (B) *K*-means clustering of the 475 pollinator lines using the method of Caliński-Harabasz (1974) showing cluster 1 (C1), cluster 2 (C2) and cluster 3 (C3) respectively. (C) Neighbour-joining tree of the same dataset depicting three clusters (C1, C2 & C3).

LD was calculated for all the chromosomes individually across A- and C- subgenomes, Figures 3.3: (A) and (B). There was a general trend of low mean LD (r^2) per chromosome in subgenome A compared to subgenome C in spring-type canola and the difference in LD per chromosome in both the subgenomes was remarkable (Table 3.2). To have a meticulous assessment of the patterns of LD decay, I measured the physical distance at which the pair-wise genotypic association in the filtered SNP dataset decays below a threshold of $r^2 = 0.1$. LD decay in majority of chromosomes started at ~400 KB regions in subgenome A while in C subgenome, it was at 3MB region. In A subgenome, slowest LD decay was observed from 4MB-5MB regions (A9) to 8MB-10MB (A8), Figure 3.3: (A). Phenomenally low LD decay was observed on C subgenome in the chromosomes C2, C1 and C4 within 15MB, 15MB and 10MB regions, respectively, Figure 3.3: (B).

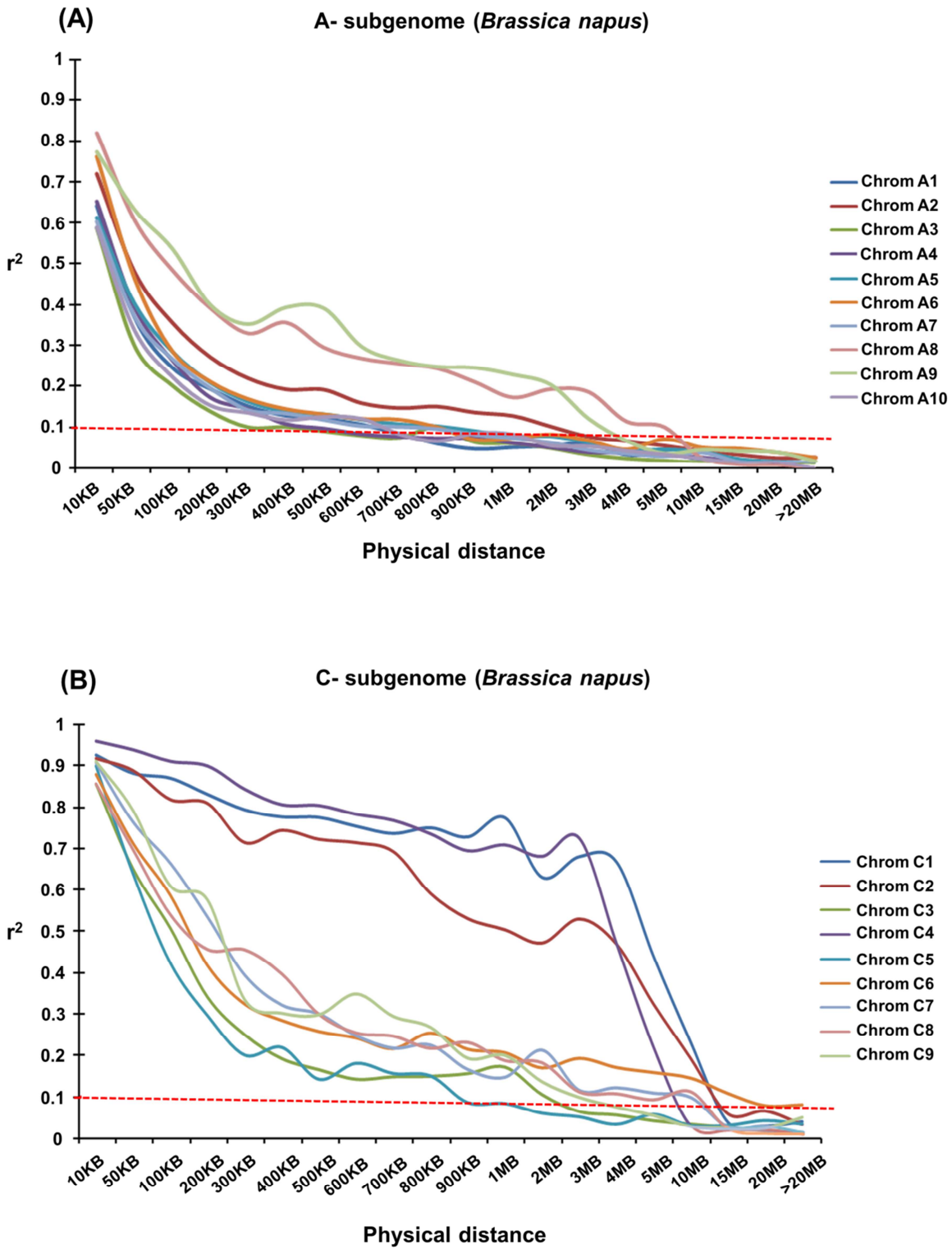


Figure 3.3: (A) Individual chromosome decay in subgenome A of linkage disequilibrium (LD) expressed as r^2 in spring-type *Brassica napus* in a set of 475 male lines against a physical distance. Horizontal 'Dashed red line' along the x-axis shows a standard cutoff r^2 value of 0.1. (B) Extent of LD decay in C- subgenome within each chromosome.

Table 3.2: Summary of the estimated chromosome-wise mean LD decay (r^2=0.1) employing 24,442 unique, single-copy SNPs within A and C subgenomes, number of SNPs per chromosome and SNP density.

A subgenome	Chromosome	LD decay (MB)	No. of SNPs	SNP density (KB/SNP)
	A01	0.70-0.80	1044	22.0
	A02	2.00-2.50	811	30.5
	A03	0.20-0.30	1637	17.8
	A04	0.40-0.50	1110	15.5
	A05	0.80-0.90	1203	19.0
	A06	0.80-0.90	1005	24.0
	A07	0.60-0.70	1395	17.0
	A08	4.00-05.0	728	25.9
	A09	4.00-4.50	1018	32.9
	A10	0.70-0.75	993	32.9
C subgenome				
	C1	15.0-15.5	2094	18.0
	C2	15.0-15.5	1894	24.0
	C3	2.00-2.50	2103	28.7
	C4	10.0-10.5	2523	19.0
	C5	0.90-0.95	706	60.0
	C6	20.0-20.5	1007	36.9
	C7	10.0-10.5	1280	34.9
	C8	4.00-5.00	1167	32.8
	C9	3.00-3.50	724	66.9

3.3 Genome-wide association studies (GWAS)

3.3.1 Association analysis

GWAS was conducted using mixed effects model accounting for population structure with a panel of 24,442 polymorphic SNPs distributed across the whole genome. Manhattan plots from all the three traits (GCA for seed yield; GCA for DTF and GCA for seed oil content) show various significant peaks on different chromosomes (Figure 3.4: (A), (B) and (C)). Quantile-Quantile (Q-Q) plots were generated along with their respective Manhattan plots

for all the three traits under study that show reduced inflation of the *p*-values caused by population sub-structure except the most significant markers that deviate the null hypothesis of no causative markers.

A total of 316 significant SNP loci within the three traits under study (GCA for seed yield: 94, GCA for DTF: 124 and GCA for seed oil content 98) above the genome-wide significant *p*-value threshold of ($-\log_{10}$ 2.5) under (FDR: $p \leq 0.05$) for each trait was observed.

I found two close-by (<500 KB) located significant SNP loci related to GCA for seed yield and GCA for DTF on two different regions on chromosome A3 (Table 3.3) and two close-by located SNPs related to GCA for seed yield and GCA for seed oil content on A4 and C3, respectively (Table 3.4). Similarly, we identified one significant SNP related to GCA for seed yield and GCA for DTF on chromosome A9 (Table 3.5) and 4 overlapping significant SNPs related to GCA for seed yield and GCA for seed oil content anchored on chromosomes A3, A4, A5 and C3 (Table 3.6).

3.3.2 Prediction of candidate genes associated with hybrid performance

I received significant signals on various chromosomes. For brevity, we report one detailed example each from closely located two separate significant SNPs related to GCA for seed yield and GCA for DTF on chromosome A3 and one from an overlapping significant SNP related to GCA for seed yield and GCA for seed oil content on the same chromosome (A3) along with their respective associated *B.napus* orthologues of *Arabidopsis thaliana* candidate genes. I also report haplotype diversity analyses for each of the two examples on chromosome A3.

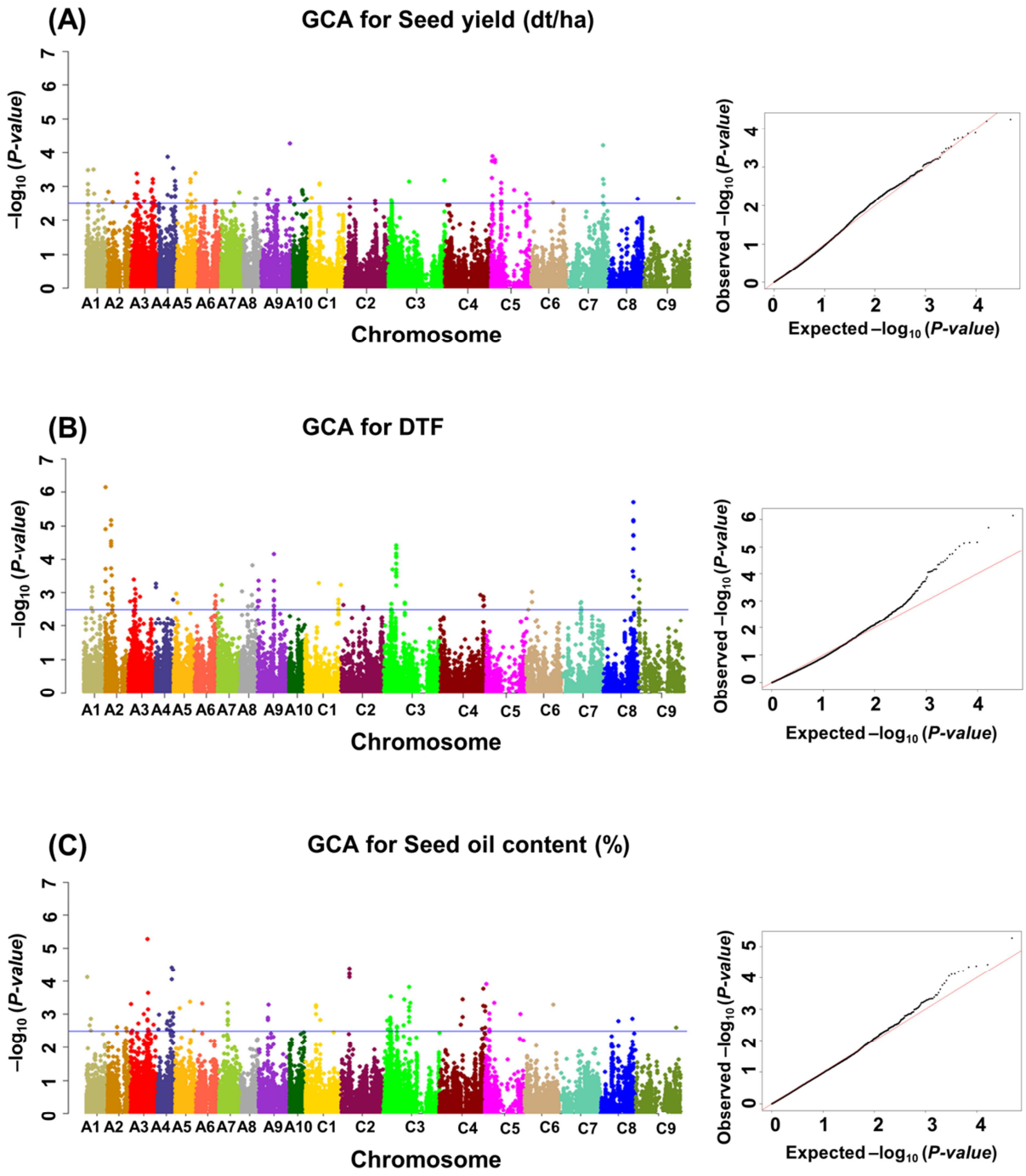


Figure 3.4: Genome-wide association study of hybrid performance in (A) GCA for seed yield (B) GCA for DTF and (C) GCA for seed oil content. In these Manhattan plots, (-$\log_{10}$ *p*-values on y-axis from the mixed model are plotted against the SNP position on the x-axis for each of the 19 spring-type canola chromosome. The horizontal 'light blue' line represents the significant *p*-value threshold under FDR ($p \leq 0.05$) for each trait. In the right corner of each Manhattan plot, the Q-Q plot for each trait indicates the region of inflated *p*-values (deviation from the 1:1 line)..

Table 3.3: Closely located SNPs between the two traits

(GCA for seed yield and GCA for DTF)

Trait	Closest SNP ID	Chromosome	Position	-$\log_{10}$ p value
GCA for seed yield	Bn-A03-p6744344	A3 (region1)	6017807	3.372290122
GCA for DTF	Bn-A03-p6898220	A3 (region1)	6179649	3.38391294
GCA for seed yield	Bn-A03-p7501352	A3 (region2)	6799644	2.633281076
GCA for DTF	Bn-A03-p7672403	A3 (region2)	6972869	3.110051589

Table 3.4: Closely located SNPs between the two traits

(GCA for seed yield and GCA for seed oil content)

Trait	Closely located SNP ID	Chromosome	Position	-$\log_{10}$ p value
GCA for seed yield	Bn-A04-p14807150	A4	15257698	3.5375
GCA for oil content	Bn-A04-p14756001	A4	15205214	4.0493
GCA for seed yield	Bn-scaff_18936_1-p922423	C3	3459404	2.5354
GCA for oil content	Bn-scaff_18936_1-p867397	C3	3398494	2.8126

Table 3.5: Overlapping SNPs between the two traits

(GCA for seed yield and GCA for DTF)

Trait	Overlapping SNP ID	Chromosome	Position	-$\log_{10}$ p value
GCA for seed yield	Bn-A09-p20652846	A9	17583800	2.599219
GCA for DTF		A9	17583800	4.144566

Table 3.6: Overlapping SNPs between the two traits

(GCA for seed yield and GCA for seed oil content)

Trait	Overlapping SNP ID	Chromosome	Position	-$\log_{10}$ p value
GCA for seed yield	Bn-A03-p26833841	A3	25271758	2.577185685
GCA for seed oil content		A3	25271758	2.680452323
GCA for seed yield	Bn-A04-p12283561	A4	13272610	2.70748188
GCA for seed oil content		A4	13272610	3.031950266
GCA for seed yield	Bn-A05-p18161591	A5	16447216	2.532394854
GCA for seed oil content		A5	16447216	3.382027267
GCA for seed yield	Bn-scaff_17521_1-p299252	C3	21921202	3.139823761
GCA for seed oil content		C3	21921202	3.449097092

Table 3.7: Closely located significant SNPs associated with candidate genes

(GCA for seed yield and GCC for DTF)

Chrom.	Trait	Significant SNP ID	Significant SNP position	-log$_{10}$ p value	Distance from the gene (KB)	Gene ID (*B.napus*)	Strat position	End position	*A.thaliana* orthologue gene	Gene name (*A.thaliana*)
A3	GCA for seed yield	Bn-A03-p6744344	6017807	4.24E-04	19	BnaA03g13220D	5998751	6007985	AT5G51230	EMBRYONIC FLOWER 2
					0.975	BnaA03g13230D	6018782	6020213	AT5G51160	Ankyrin repeat family protein
					32.95	BnaA03g13310D	6050759	6052938	AT5G51100	FSD2
					35.3	BnaA03g13320D	6053113	6055294	AT4G00650	FRIGIDA
					47.32	BnaA03g13340D	6065129	6070227	AT5G51060	ROOT HAIR DEFECTIVE 2
A3	GCA for DTF	Bn-A03-p6898220	6179649	4.13E-04	180.89					
					128.89					
					126.53					
					114.52					

Table 3.8: Overlapping significant SNP associated with candidate genes

(GCA for seed yield and GCC for seed oil content)

Chrom.	Trait	Significant SNP ID	Significant SNP position	-log$_{10}$ p value	Distance from the gene (KB)	Gene ID (*B.napus*)	Strat position	End position	*A.thaliana* orthologue gene	Gene name (*A.thaliana*)
A3	GCA for seed yield	Bn-A03-p26833841	25271758	2.65E-03	107.2	BnaA03g49250D	25378962	25382653	AT2G20610	ALF1
	GCA for seed oil content			2.09E-03	105.32	BnaA03g48960D	25166439	25167980	AT4G28030	

Table 3.9: Comparison of haplotypes with their respective means corresponding GCA and trait (BLUE) values on chromosome A3

Haplotypes	GCA for seed yield	GCA for DTF	Seed yield (F1:PollinatorsxM1)	Seed Yield (F1:PollinatorsxM2)	DTF (F1:PollinatorsxM1)	DTF (F1:PollinatorsxM2)
Hap1	-0.637	1.512	31.437	29.629	171.9	173.6
Hap2	-0.654	0.006	31.573	29.460	170.8	171.7
Hap3	0.520	-0.523	32.086	31.295	170.7	170.8
Hap4	-0.750	-0.787	30.441	30.398	170.5	170.5
Hap5	0.766	-1.180	32.407	31.465	169.6	170.5
Hap6	0.513	-0.036	31.592	31.774	171	171.5
Hap7	0.327	-0.170	31.852	31.143	170.7	171.5
Hap8	0.242	-0.600	31.488	31.336	170.4	170.9
Hap9	-0.821	-0.450	30.801	29.897	170.7	171
Hap10	-0.805	-0.535	31.042	29.687	170.4	171
Hap11	-0.293	0.621	31.073	30.681	171.6	172.2

'Dashed lines' show significant haplotype (Hap5) for GCA for seed yield

Table 3.10: Comparison of haplotypes with their respective means corresponding GCA and trait (BLUE) values on chromosome A3

Haplotypes	GCA for seed yield	GCA for seed oil content	Seed yield (F1:PollinatorsxM1)	Seed yield (F1:PollinatorsxM2)	Seed oil content (F1:PollsxM1)	Seed oil content (F1:PollinatorsxM2)
Hap1	-2.186	-1.131	29.657	28.312	47.4803	47.078
Hap2	0.351	0.713	31.952	31.089	49.0987	49.148
Hap3	-1.678	-1.140	30.850	28.134	47.500	47.040
Hap4	0.0197	1.525	31.391	30.988	49.287	50.583
Hap5	0.072	0.927	31.415	31.070	49.310	49.365
Hap6	0.882	0.830	32.399	31.705	49.277	49.203
Hap7	0.396	1.502	31.795	31.338	49.542	50.283
Hap8	-0.006	0.188	31.651	30.678	48.962	48.234
Hap9	0.395	-0.986	32.411	30.718	48.037	46.812
Hap10	-0.201	-1.784	31.878	30.061	47.403	45.850
Hap11	-0.564	-0.703	31.233	29.979	48.463	46.950
Hap12	-0.946	-1.564	30.673	29.775	47.63	46.062

'Dashed lines' show significant haplotype (Hap6) for GCA for seed yield

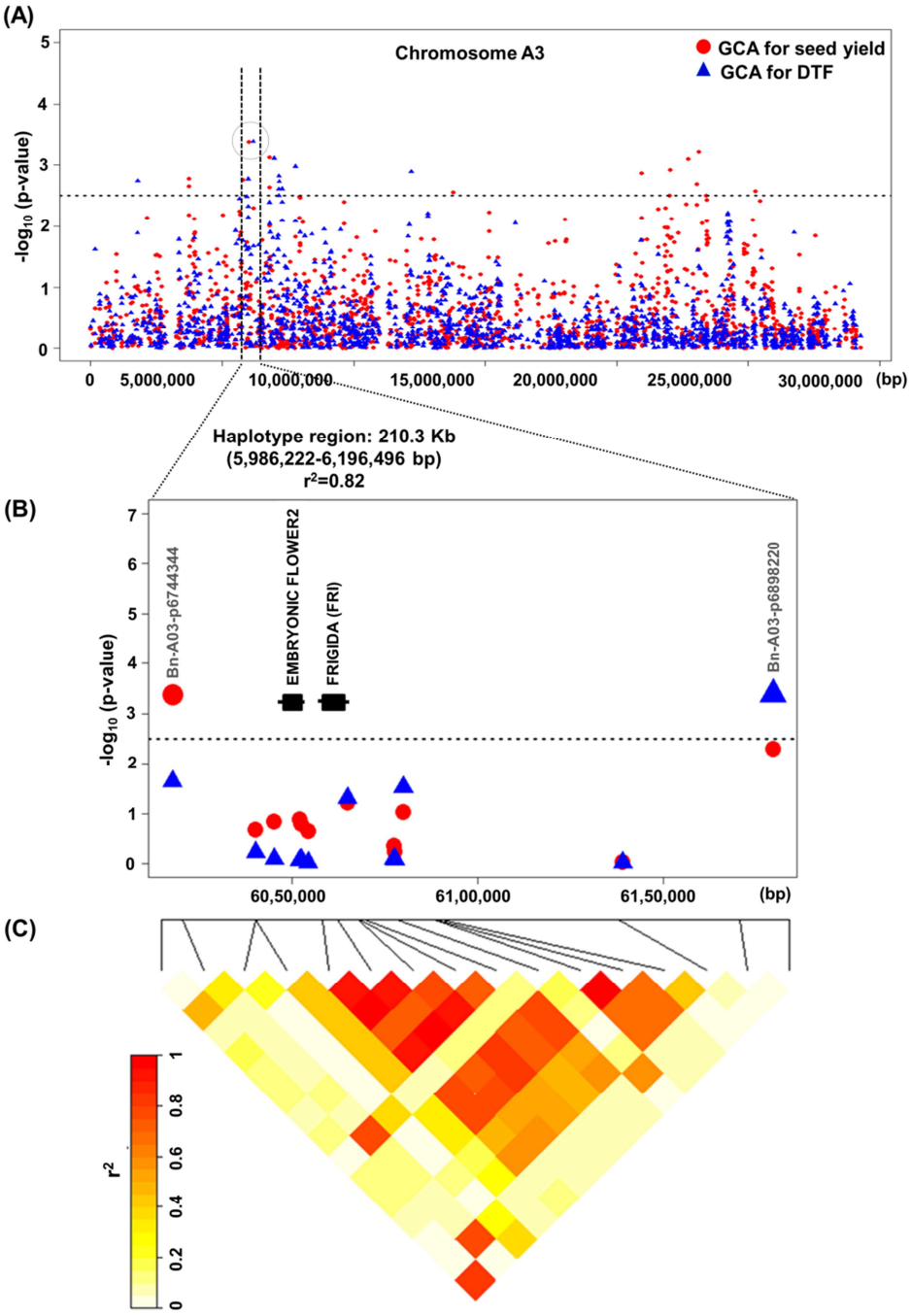


Figure 3.5: Schematic overview of the candidate genes in strong LD region: (A) Peak of transformed ($-\log_{10}$ *p*-values are shown on chromosome A3 in a region of closely located two significant SNPs for the two different traits (i.e. GCA for seed yield and GCA for DTF). (B) This is a zoom in figure on chromosome A3 which illustrates two closely located significant SNPs as 'red dot' (left) and 'blue dot' (right) found in close proximities associated with two candidate genes '*FRI*' and '*EMBRYONIC FLOWER 2*'. (C) Haplotype of 210.3 Kb in strong LD region. Different colours in the triangle block show extent of LD.

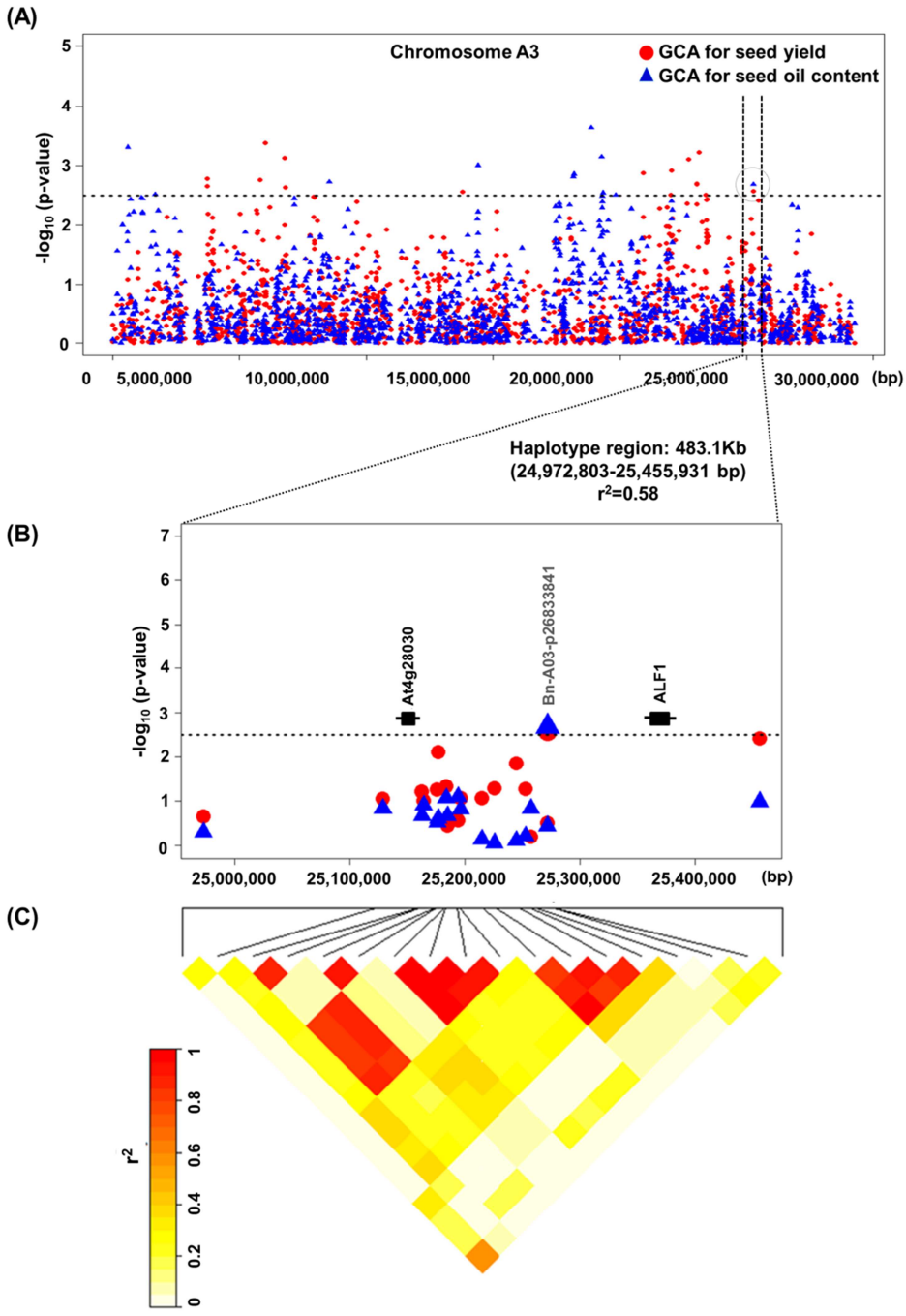


Figure 3.6: Schematic overview of the candidate genes in strong LD region: (A) Peak of transformed (-$\log_{10}$ *p*-values are shown on chromosome A3 in a region of overlapping two significant SNPs for the two different traits (i.e. GCA for seed yield and GCA for seed oil content). (B) This is a zoom in the figure on chromosome A3 which illustrates two overlapping significant SNPs as 'blue dot' (top) and 'red dot' (down) found and associated with two candidate genes '*ALF1*' and '*AT4G28030*'. (C) Haplotype of 483.1 Kb in strong LD region. Different colours in the triangle block show extent of LD.

In the case of GCA for seed yield and GCA for DTF on A3, we found a total of 38 reported genes within a region of 210.3 KB with a strong LD (r^2=0.82) haplotype block region (5,998,751-6,007,985 bp). These genes are involved in different pathways and activities. In this region two closest significant SNP loci (GCA for seed yield: Bn-A03-p6744344) and (GCA for DTF: Bn-A03-p6898220) with strong LD located at 19 and 35.3 KB away from the two important *Arabidopsis* orthologue candidate genes related to flowering time and yield, for example, *EMBRYONIC FLOWER 2* (AT5G51230) and *FRIGIDA* (AT4G00650) respectively (Figure 3.5; Table 3.7). Similarly, in the case of GCA for seed yield and GCA for seed oil content, we found a total of 65 reported genes (related to different pathways and activities) within a region of 483.1 KB with a strong LD (r^2=0.58) haplotype block region (24,972,803-25,455,931). The significant overlapping SNP between (GCA for seed yield and GCA for oil content: Bn-A03-p26833841) located at 105.32 and 107.2 KB away from two important *Arabidopsis* orthologue candidate genes (*AT4G28030*) and *ALF1* (AT2G20610) related to yield and seed oil content, respectively (Figure 3.6; Table 3.8).

3.3.3 Haplotype diversity analysis

In a panel of 475 accessions of *Brassica napus*, a total of 46 SNP haplotype diversity blocks (GCA for seed yield: 11, GCA for DTF: 11) and (GCA for seed yield: 12, GCA for oil content: 12) corresponding to their respective calculated GCA values and estimated trait (BLUE) values. These haplotype blocks were manually traced within the strong LD (r^2> 0.40) flanking regions of significant SNPs for the three trait heterosis under study.

(A)

Haplotype diversity | **Haplotype frequency (%)**

Hap1	C	G	A	T	C	G	G	G	T	C	C	G	T	C	C	G	1.68
Hap2	C	G	A	T	C	G	G	G	T	C	C	G	T	T	C	G	3.37
Hap3	T	A	A	T	T	T	A	T	C	T	T	A	T	C	A	T	4
Hap4	T	A	A	T	T	T	A	T	C	T	T	A	T	C	N	T	1.05
Hap5	T	G	G	C	C	G	G	G	T	C	C	G	T	C	C	T	1.47
Hap6	T	G	G	C	C	G	G	G	T	C	C	G	T	N	C	T	3.58
Hap7	T	G	G	C	C	G	G	G	T	C	C	G	T	T	C	T	19.58
Hap8	T	G	G	C	T	T	A	N	C	C	T	A	C	C	C	T	1.47
Hap9	T	G	G	T	T	T	A	N	C	C	T	A	C	C	C	T	2.53
Hap10	T	G	G	T	T	T	A	T	C	C	C	G	T	T	C	T	2.32
Hap11	T	G	G	T	T	T	A	T	C	C	T	A	C	C	C	T	17.68

(B)

GCA for seed yield and DTF

Hap5	T	G	G	C	C	G	G	G	T	C	C	G	T	C	C	T
M1	T	G	G	T	T	T	A	T	C	C	T	A	C	C	C	T
M2	T	G	G	T	T	T	A	T	C	C	C	G	T	N	C	T

F1: Hap 5 x M1

F1	TT	GG	GG	CT	CT	GT	AG	GT	CT	CC	CT	AG	CT	CC	CC	TT

F1: Hap 5 x M2

F1	TT	GG	GG	CT	CT	GT	AG	GT	CT	CC	CC	GA	TT	C/N	CC	TT

(C)

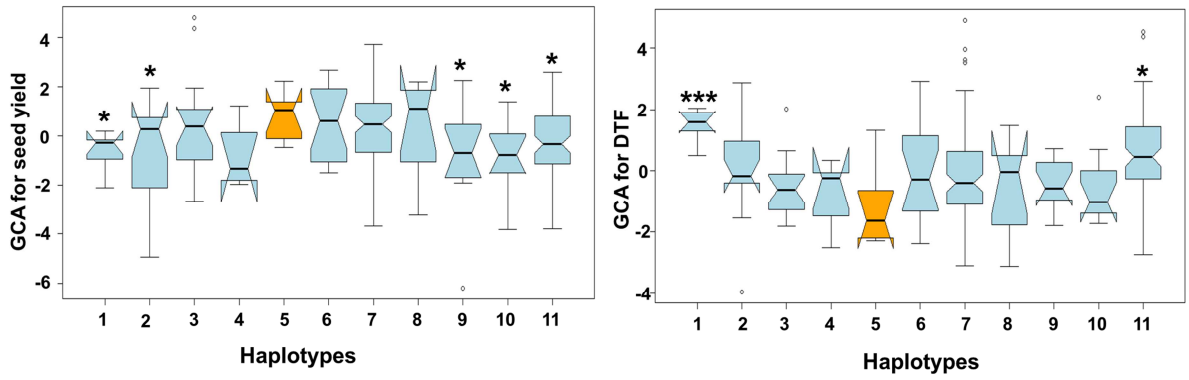


Figure 3.7: SNP haplotypes. (A) Haplotype diversity block contributing to hybrid performance for GCA for seed yield and GCA for DTF on chromosome A3 (B) Reconstruction of the expected F1 genotypes from the Hap5 and its respective two mother line (M1 and M2) regions. (C) SNP boxplots of each haplotype: Boxplots of Hap5 with 'orange' colours correspond to the highest phenotype value of GCA for seed yield (left) and lowest phenotype value of GCA for DTF (right). The asterisks above the different haplotypes represent significant ($p \leq 0.05$) differences between the 'orange' box plots and the rest. ($p \leq 0.05$*, $p \leq 0.01$**, $p \leq 0.001$***).

(A)

Haplotype diversity

																		Haplotype frequency (%)
Hap1	A	C	G	C	G	T	C	G	G	T	C	A	G	C	T	C	C	T	1.26
Hap2	A	C	G	C	G	T	C	G	G	T	C	A	G	C	T	T	C	C	5.47
Hap3	A	C	G	C	G	T	C	G	G	T	C	A	G	C	T	T	C	T	1.47
Hap4	A	C	G	C	G	T	C	G	G	T	C	A	G	C	T	T	T	C	2.53
Hap5	A	T	G	C	T	G	C	G	G	T	T	G	A	T	C	T	C	C	2.11
Hap6	A	T	T	A	T	G	A	A	T	G	T	G	A	T	T	T	T	C	9.47
Hap7	G	T	G	C	G	T	C	G	G	T	C	A	G	C	T	C	C	T	3.79
Hap8	G	T	G	C	T	G	C	G	G	T	T	G	A	T	C	T	C	T	6.74
Hap9	G	T	T	A	G	T	A	A	T	G	T	G	A	T	C	C	C	T	3.16
Hap10	G	T	T	A	G	T	A	A	T	G	T	G	A	T	C	T	T	T	3.16
Hap11	G	T	T	A	T	G	A	A	T	G	C	A	G	C	T	C	C	T	2.74
Hap12	G	T	T	A	T	G	A	A	T	G	T	G	A	T	C	C	C	T	2.95

(B)

GCA for seed yield

Hap6	A	T	T	A	T	G	A	A	T	G	T	G	A	T	T	T	T	C
M1	G/A	T/C	G	C	G	T	C	G	G	T	C	A	G	C	T	C	C	T
M2	A	C	G	C	G	T	C	G	G	T	C	A	G	C	T	T	C	C

F1: Hap 6 x M1

F1	AG/AA	TT/CT	TG	AC	GT	GT	AC	AG	GT	GT	CT	AG	AG	CT	TT	CT	CT	CT

F1: Hap 6 x M2

F1	AA	CT	GT	AC	GT	GT	AC	AG	GT	GT	CT	AG	AG	CT	TT	TT	CT	CC

GCA for seed oil content

Hap4	A	C	G	C	G	T	C	G	G	T	C	A	G	C	T	T	T	C
M1	G/A	T/C	G	C	G	T	C	G	G	T	C	A	G	C	T	C	C	T
M2	A	C	G	C	G	T	C	G	G	T	C	A	G	C	T	T	C	C

F1: Hap 4 x M1

F1	AG/AA	CT/CC	GG	CC	GG	TT	CC	GG	GG	TT	CC	AA	GG	CC	TT	CT	CT	CT

F1: Hap 4 x M2

F1	AA	CC	GG	CC	GG	TT	CC	GG	GG	TT	CC	AA	GG	CC	TT	TT	CT	CC

(C)

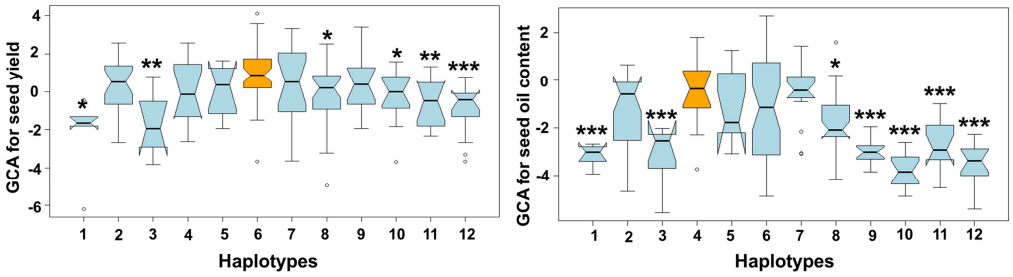


Figure 3.8: SNP haplotypes. (A) Haplotype diversity block contributing to hybrid performance for GCA for seed yield and GCA for seed oil content on chromosome A3 (B) Reconstruction of the expected F1 genotypes from the Hap6 and its respective two mother line (M1 and M2) regions for seed yield and Hap4 and its respective two mother line (M1 and M2) regions for seed oil content. (C) SNP boxplots of each haplotype: Boxplots of Hap6 with 'orange' colours correspond to the highest phenotype value of GCA for seed yield (left) and highest corresponding phenotype value of the GCA for seed oil content (right). The asterisks above the different haplotypes represent significant ($p \leq 0.05$) differences between the 'orange' box plots and the rest. ($p \leq 0.05$*, $p \leq 0.01$**, $p \leq 0.001$***).

Finally, the significant haplotypes from the pollinators were compared with their corresponding haplotype regions from the two testers (mother lines) and predicted F1 genotypes were reconstructed.

In the case of GCA for seed yield and GCA for DTF, Hap 5 had the highest corresponding GCA value for GCA for seed yield while the lowest corresponding GCA value for GCA for DTF with a frequency of 1.47 % (Table 3.9). There was a significant difference ($p \leq 0.05$) between Hap 5 and Hap 1, Hap 2, Hap 9 and Hap 11 for GCA for seed yield and a significant difference ($p \leq 0.05$) between Hap 5 and Hap 1 and Hap 11 for GCA for DTF (Figure 3.7: A, B, C).

While comparing haplotype diversity of GCA for seed yield and GCA for seed oil content, Hap 6 had the highest corresponding GCA value (Table 3.10) and was significantly different ($p \leq 0.05$) from Hap 1, Hap 3, Hap 8, Hap 10, Hap 11 and Hap 12 with a haplotype frequency of 9.47 %. However, in GCA for oil content, I identified Hap 4 having highest corresponding GCA value and was significantly different ($p \leq 0.05$) from Hap 1,Hap 3,Hap 8,Hap 9,Hap 10,Hap 11 and Hap 12 with a haplotype frequency of 2.53 % (Figure 3.8: A, B, C).

3.4 Estimation of genomic prediction accuracy for different traits

3.4.1 Prediction across whole population

Figure 3.9 and Table 3.11 shows the accuracies of genomic prediction for the respective traits, along with their respective standard errors for testcross performance using the whole population (WP) without consideration of population structure. For the seven traits considered, the highest prediction accuracy was recorded for seed oil content ($r_{GPA} = 0.81$) followed by oil yield ($r_{GPA} = 0.75$), seed glucosinolate content ($r_{GPA} = 0.61$), days to flowering ($r_{GPA} = 0.56$), seed yield ($r_{GPA} = 0.45$), lodging resistance ($r_{GPA} = 0.39$) and the

least heritable trait, seedling emergence (r_{GPA} = 0.29). Scatter plots showing the correlations between true observed trait values and genomic predicted values for all the traits are shown in Figure 3.10 (A,B,C,D,E,F,G).

Table 3.11: Average genomic prediction accuracies (r_{GPA}) and standard errors (SE) for seed yield (dt/ha), oil yield (dt/ha), seed oil content (%), seed glucosinolate content (GSL; µmol/g), seedling emergence (visual observation scale 1-9; good=9), lodging resistance (visual observation scale 1-9; good=9) and days to onset of flowering (DTF) derived from 500 iterations of cross-validation across the whole-population.

Traits	**Seed yield (dt/ha)**	**Oil yield (dt/ha)**	**Seed oil content (%)**	**GSL (µmol/g)**	**Seedling emergence (good=9)**	**Lodging resistance (good=9)**	**DTF**
r_{GPA} ±SE	45±0.002[a]	0.75±0.001	0.81±0.001	0.61±0.002	0.29±0.002	0.39±0.003	0.56±0.002

[a]Approximate standard errors (SE) attached

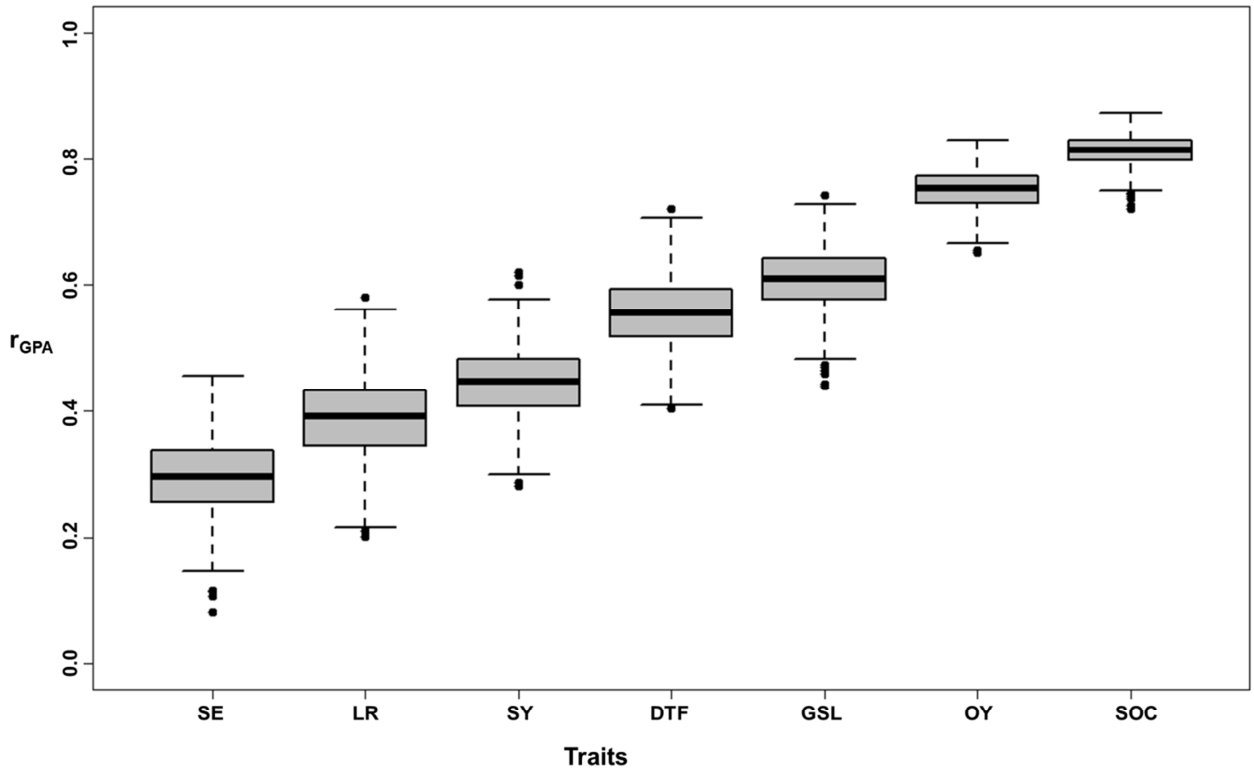


Figure 3.9: Mean genomic prediction accuracies (r_{GPA}) across the whole test population for seedling emergence; SE, lodging resistance; LR, seed yield; SY, days to flowering; DTF, seed glucosinolate content; GSL, oil yield; OY and seed oil content; SOC, respectively.

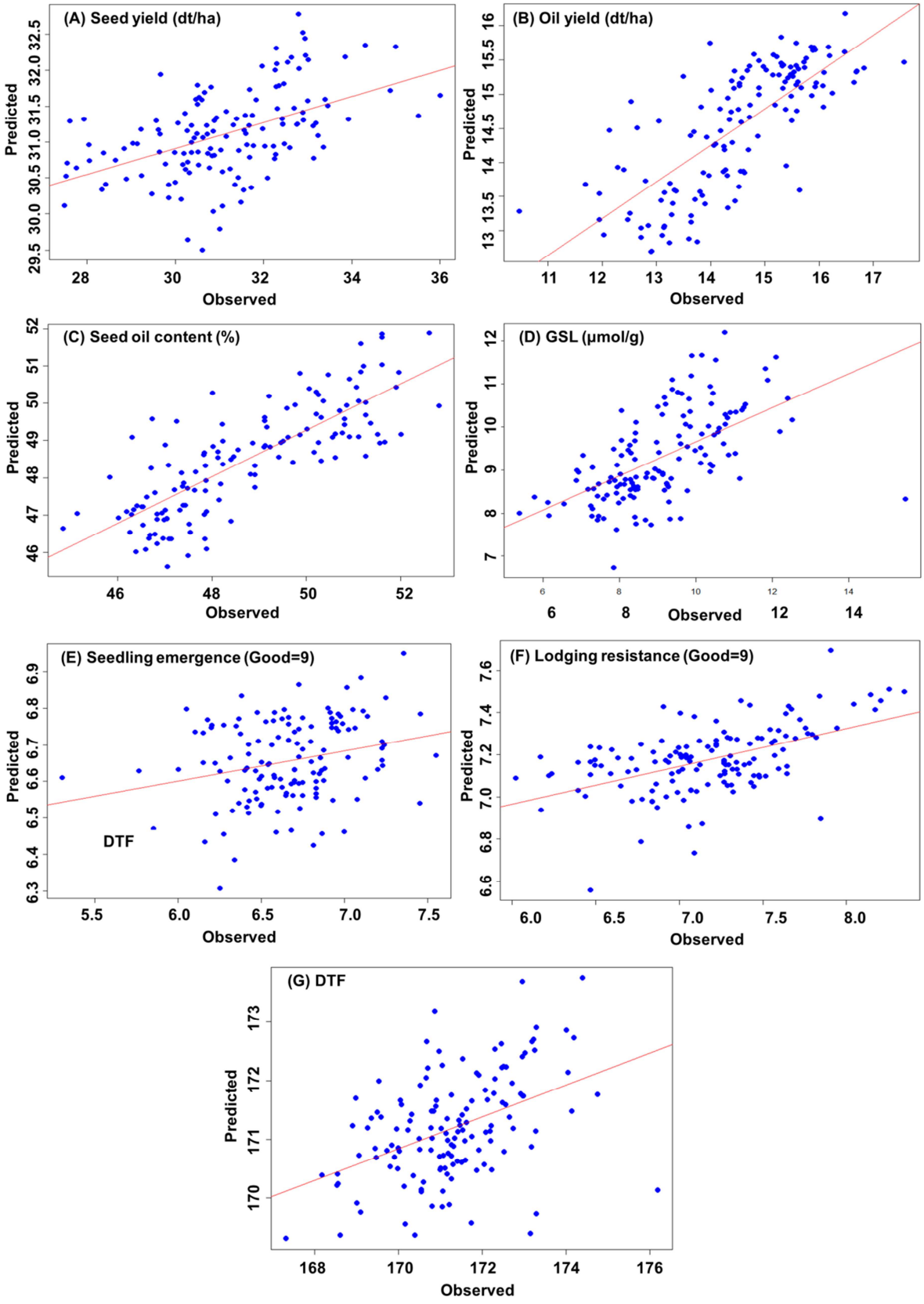


Figure 3.10: Scatter plots showing correlations between true observed mean trait values (observed) and genomic predicted (predicted) for the seven traits evaluated under scenario 1.

3.4.2 Genomic predictions within subpopulations

Figure 3.11 (A, B) and Appendix (Table 1 and Table 2) shows the independent prediction accuracies within subpopulations C1 and C2, respectively. Interestingly, an improved prediction accuracy (r_{GPA} = 0.39) was observed for the low-heritability trait seedling emergence within subpopulation C1, the largest subpopulation but also the narrowest in terms of genetic diversity. Predictions accuracies also improved for two other traits with low to moderate heritability, seed yield (r_{GPA} = 0.47) and DTF (r_{GPA} = 0.59) (Figure.3.11: A), Appendix (Table 1). Similarly within the second-largest subpopulation, C2, the prediction accuracies improved to r_{GPA} = 0.65 for GSL and r_{GPA} = 0.49 for lodging resistance, respectively (Figure 3.11: B), and Appendix (Table 2). For seed oil content and oil yield, I observed no improvement in prediction accuracy within subpopulations compared to the whole population.

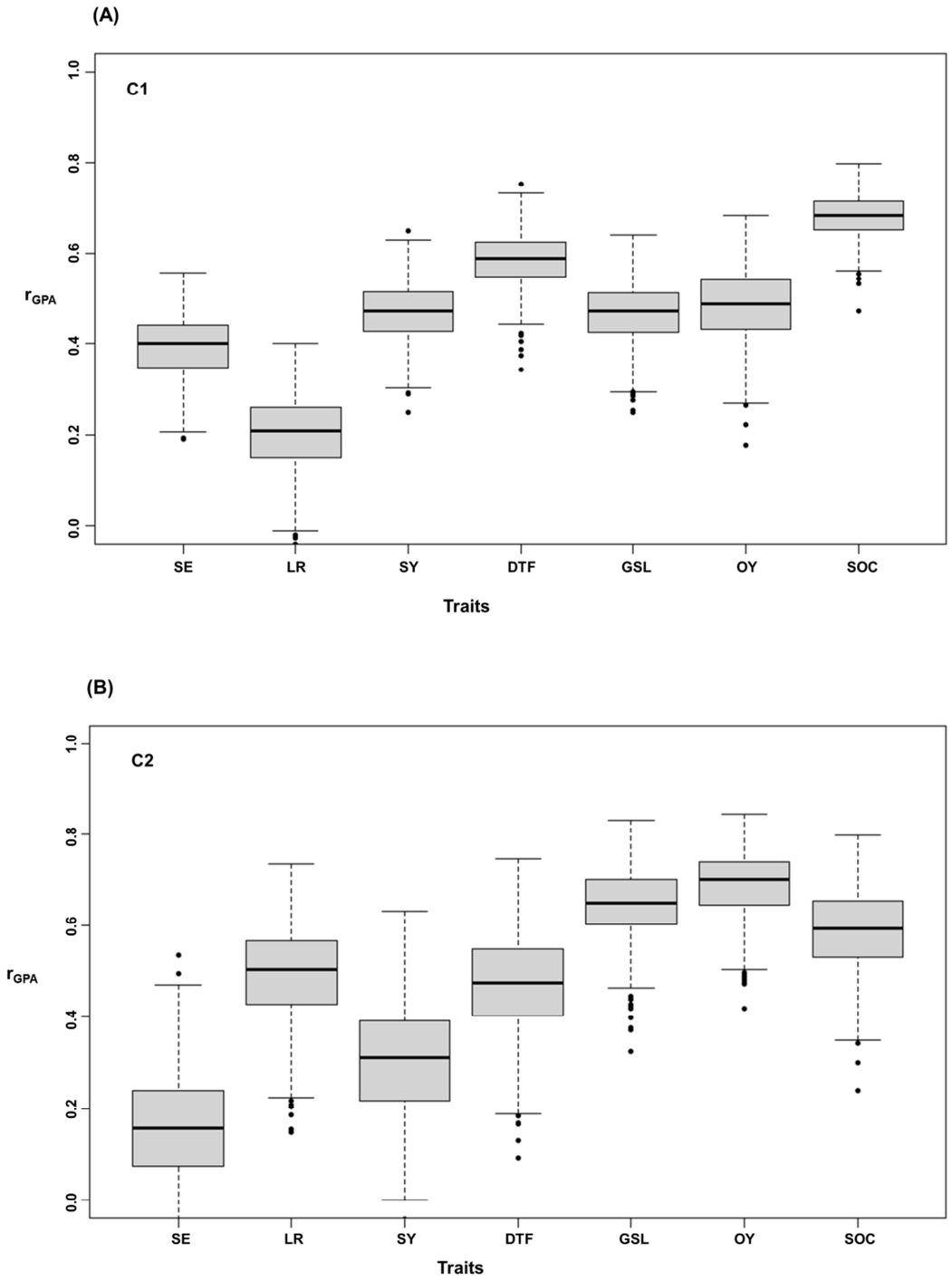


Figure 3.11: (A) Genomic prediction accuracies (r_{GPA}) within cluster 1 (C1) for seedling emergence; SE, lodging resistance; LR, seed yield; SY, days to flowering; DTF, seed glucosinolate content; GSL, oil yield; OY and seed oil content; SOC, respectively. (B) Genomic prediction accuracies (r_{GPA}) within cluster 2 (C2) for seedling emergence; SE, lodging resistance; LR, seed yield; SY, days to flowering; DTF, seed glucosinolate content; GSL, oil yield; OY and seed oil content; SOC, respectively.

3.4.3 Prediction accuracy and training population (TP) size

As expected, increasing the size of the TP resulted in the improvement of the genomic prediction accuracy for all the traits (Figure 3.12). All the traits showed a plateau of prediction accuracy at a TP proportion of 80%, except days to flowering, and only insignificant increases in accuracy were observed as the TP size increased from 70% to 90%. I, therefore optimised my model with an arbitrary TP size of 70% for all subsequent analyses and scenario testing.

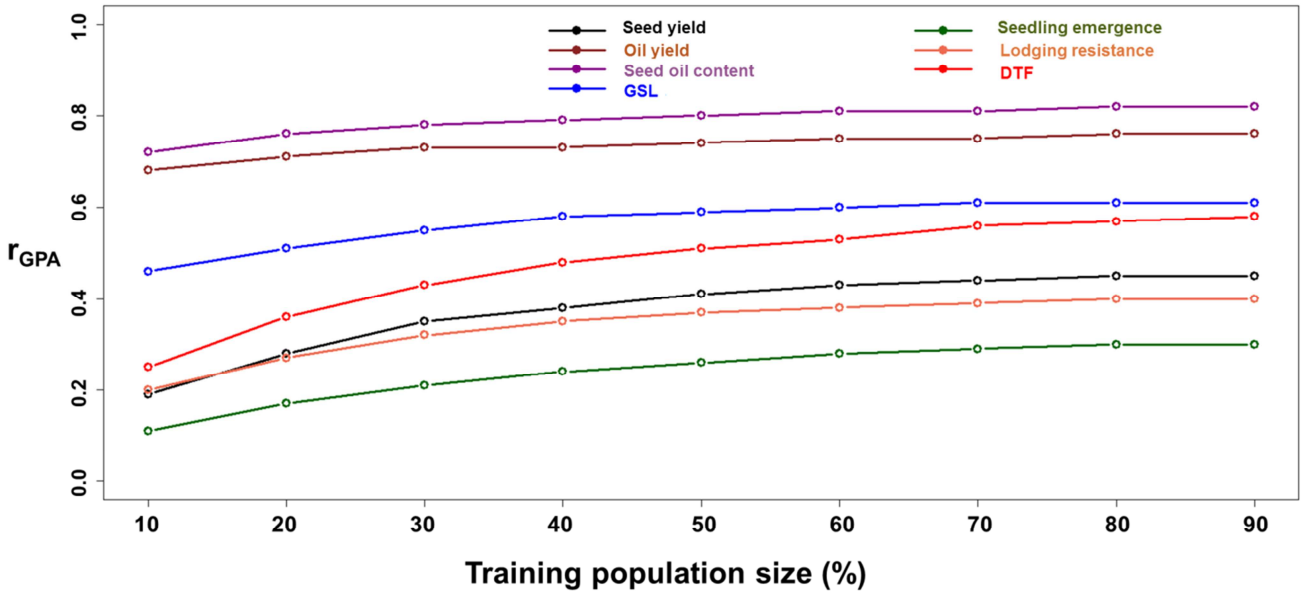


Figure 3.12: Influence of the size of the training population (TP; % of whole population size) on the genomic prediction accuracy (r_{GPA}) for the seven traits seedling emergence, lodging resistance, seed yield, days to flowering (DTF), seed glucosinolate content (GSL), oil yield and seed oil content.

Chapter 4: Discussion

4:1 Genome-wide association study (GWAS) in hybrid *Brassica napus*

Large scale genome-wide association study was carried out to identify interesting genomic regions linked to candidate genes governing hybrid performance across three important agronomic traits under study. A mixed effects model approach was used to avoid confounding effects caused by cryptic genetic background in the samples. SNP haplotype diversity was also investigated in the flanking regions of significant SNPs. I found some interesting candidate regions which may have pleiotropic effects and are involved in the co-regulation of different phenotypes. To the best of my knowledge, this is the first approach where GCA estimates are used in association mapping to associate markers with the traits controlling hybrid performance in rapeseed.

4.1.1 Linkage disequilibrium and population structure

Optimum genetic diversity is a pre-requisite to genetically improve crops through breeding programmes. The overall genetic diversity of spring-type rapeseed in both Canada and Australia is low (Fu and Gugel, 2010, Cowling, 2007). This decline in allelic variations may be due to the persistent breeding practices in favour of certain agronomic traits. Hybrid breeding in spring rapeseed is a better option to enlarge the genetic base and sustainability (Bus et al. 2011, Rahman and Kebede, 2012). The diverse gene pools of *Brassica rapa* and *Brassica oleracea*, the progenitor species of *B.napus*, could be used as a rich source to enlarge the narrow genetic pool of *B. napus* and also to bolster its resistance capabilities via resynthesised *B.napus* (Rygulla et al. 2007).

I performed LD analysis using 24,442 unique polymorphic SNP panel across the A and C subgenomes, respectively. Certain chromosomes on both the subgenomes i.e. A08, A09, C1, C2 and C4 carry long highly conserved LD blocks which suggest an artificial selection for some traits of the corresponding regions on these chromosomes. In my study, the general pattern of LD was more conserved on C- subgenome than A- subgenome. This is

in congruence with the recent study on LD reported in semi-winter type rapeseed by Qian et al. 2014. The overall low mean LD in majority of A- subgenome chromosomes indicates high recombination rates after the interspecific hybridisation of this alloploid several thousand years ago. Keeping in view the genetic diversity of *B.rapa*, breeders have been taking advantage of the fact and try to introgress new genetic diversity into *B.napus* (Bennett et al. 2012). Similar attempts were also made to boost up the gene pool of subgenome C by crossing *B.napus* to *B. Oleracea* but have not been so successful (Leflon et al. 2010). Chalhoub et al. 2014 argues, that the presence of a large number of transposable elements on C- subgenome may cause lower recombination rate and hence a big difference in LD decay between the two subgenomes is observed. The two relatively large subpopulations in my dataset of spring type rapeseed indicate introgression of adapted and exotic materials from its progenitors. Due to its vernalisation requirements, winter-type canola germplasm is not suitable to introgress into spring-type canola but winter-type canola can still be used as a diverse genetic source to improve spring-sown canola especially to increase heterotic potential for seed yield by transferring 'super alleles' at some specific loci (Quijada et al. 2006). Udall et al. 2004 reported that seed yield in spring-type canola can also be improved through introgression of some alleles from resynthesised *B. napus*.

4.1.2 Association analysis and trait heterosis in *Brassica napus*

Association mapping has emerged as a strong method for the dissection of both simple and complex trait architectures. In my study, I tried to decipher different interesting genomic regions bearing both candidate genes possibly involved in regulating specific traits heterosis but also co-regulating two or more different traits. I also show haplotype diversity within all the three traits that could be linked to hybrid performance of spring-type rapeseed. I followed the LD-block based approach while searching for candidate genes instead of

using ‘fixed-window’ approach which is based on a specific genomic region from the significant SNP markers on both the sides. The LD-block based approach has advantages over ‘fixed-window’ in terms of including true candidate genes (Chen et al. 2012; Courtois et al. 2013). I report some of the important candidate genes which have a pleiotropic effect and are involved in both co-regulating heterosis for seed yield and DTF components and some genes related to fatty acid biosynthesis for oil content and yield production. The strong close-by or overlapping signals on various chromosomes, especially on chromosome A3 in each trait and their fine mapping with crucial candidate genes suggest that these specific regions of the genome on A- subgenome might play an important role in hybrid performance of spring-type rapeseed. In this association mapping approach, I identified numerous SNP markers significantly associated with traits of interest that could be used effectively in future breeding strategies. The candidate genes identified in this study appear to have a role in the co-regulations of two or more important and complex traits like flowering time and seed yield heterosis. For example, the identification of *FRIGIDA (FRI)* gene in a strong LD region associated with two closely located significant SNP loci (Figure 3.5; Table 3.7). This gene is located 35.3 KB away from the significant SNP of GCA for seed yield and 114.52 KB away from the significant SNP of GCA for DTF. It is a key upstream regulator gene in *Arabidopsis thaliana* that has an important role in the activation of *FLOWERING LOCUS C (FLC)* gene and thus, inhibits floral transition. Both *FLC* and *FRI* genes play an essential role in vernalisation requirement and life cycle adaptation in different climatic conditions (Stinchcombe et al. 2004; Shindo et al. 2005; Werner et al. 2005). Wang et al. 2011 reported to have identified four *FRI* homologues in *B.napus*, one of which co-localises with a major flowering time QTL on chromosome A3. The flowering time genes *FRIGIDA (FRI)* and *FLOWERING LOCUS C (FLC)* have a major role in flowering time trait heterosis and they interact epistatically (Moore and Lukens,

2011). Furthermore, it has been reported that mutations within these two flowering time genes in *A. thaliana* may cause changes in the other traits, for example, total number of seeds, silique number and floral development etc. (Tienderen et al. 1996; Koornneef et al. 1998; Alonso-Blanco et al. 1999). Therefore, I postulate that *FRI* and *FLC* might have a role in the co-regulation of flowering time as well as seed yield heterosis. Similarly, I report another candidate gene *(EMBRYONIC FLOWER 2)*, which is just 19 KB away from the significant SNP of GCA for seed yield and 180.89 KB away from GCA for DTF in this region and is involved in both vernalisation response and flower development in *A. thaliana* (Figure 3.5; Table 3.7). Biotic and abiotic stresses can badly affect crop production (Atkinson and Urwin, 2012). Abiotic stresses like drought, salinity, heat, cold, and nutrient deficiency could reduce crop average yield by more than 50 % (Wang et al. 2003). Salt stress being one of the major abiotic stresses hampers the seed yield of *Brassica napus* (Long et al. 2015). In response to biotic stresses (bacteria, fungi, viruses, insects, etc.) and abiotic stresses, a cascade of cellular and molecular responses is evoked within the plant that often leads to lower growth and yield production (Hammond-Kosack and Jones, 2000; Herm and Mattson, 1992). Another candidate gene *ROOT HAIR DEFECTIVE 2 (RHD2)* is found on chromosome A3 (Table 3.7) which has a role in root hair elongation and defense response. Koscienly and Gulden, 2012 compared open-pollinated and hybrid *B.napus* genotypes for their relationship between different root parameters including root length and area with seed yield and reported a strong relationship between them. Similarly, another *A.thaliana* orthologue gene *(AT5G51160)* found in this area is involved in response to nitrate, nitrate transport and cellular response to iron ion starvation (Table 3.7). Nitrate despite being an essential nutrient; also plays an important role in breaking seed dormancy by serving as signaling molecules and also regulating lateral root development (Almagro et al. 2008, Alboresi et al. 2005). It is also involved in regulation of lateral root development (Zhang and

Forde, 2002). Therefore, it is also hypothesised that these genes might be indirectly involved in affecting and regulating seed yield performance.

To investigate candidate genes involved in co-regulation of heterosis of both seed yield and seed oil content, I report a very important *A.thaliana* orthologue candidate gene *(ABERRANT LATERAL ROOT FORMATION 1: ALF1)* which is associated with a significant overlapping SNP (Bn-A03-p26833841) in both the traits and is involved in a plethora of essential regulations and pathways, for example, unsaturated fatty acid biosynthetic process, adventitious root development, defense response, sulfur compound biosynthetic process and glucosinolate biosynthetic process (Figure 3.6; Table 3.8). Unsaturated fatty acids which make up about 93 % (with a density of 0.91 g/cm^3) of the total fatty acids in *B.napus* being of great significance for human health (Omidi et al. 2010). The binary system of glucosinolate-myrosinate in canola and other members of the *Brassicaceae* make a unique defense system against herbivores and pathogens (Ahuja et al. 2011). A better defence system could ultimately boost yield production. I therefore believe that this candidate gene might have a pleiotropic role in the co-regulation of both oil biosynthesis and yield performance heterosis.

The effects of drought and water deprivation on *B.napus* are severe which are accompanied by loss of grain and affecting all yield components (Andersen et al. 1996; Norouzi et al. 2008). I report another orthologue gene *(AT4G28030)* of *A.thaliana* that is involved in water transport, response to salt stress and indoleacetic acid biosynthetic process (Table 3.8). This gene is in is in the strong LD region with the significant overlapping SNP and is located at 105.32 KB away. Water stress also badly affects seed oil content and brings changes in the lipid profile (Danesh-Shahraki et al. 2008 and Boucherean et al. 1996). Indoleacetic acid is an important constituent of auxin, an essential class of phytohormones which has also a role in seed dormancy in *A.thaliana* (Liu et al.

2013). Collectively, my finding on this candidate gene suggests that it has both direct and indirect roles in heterosis for both oil biosynthesis and seed yield production.

4.1.3 Haplotypes and hybrid performance in *Brassica napus*

I identified specific haplotype diversity blocks that might contribute to trait heterosis in each of my examples. Each haplotype within the strong LD block on each chromosome may be considered as one 'allele'. These 'heterotic haplotype' architectures could be used to identify common SNP markers that capture the optimum diversity observed in the population. In the classical hybrid crops, for example in maize, where decades of hybrid breeding shaped a strong heterotic pool. In contrast, canola germplasm lack such strong heterotic gene pool. Therefore, to establish a novel heterotic pool within the adapted canola germplasm through introgression from either its two diploid progenitor species or exotic materials could be an important step. To capture haplotype diversity at the F1 hybrid level could be an effective strategy for predictive breeding (Snowdon et al. 2015). The identification and reconstruction of expected specific possible F1 genotypes between the parental haplotypes of the pollinators and two testers in both of my studied examples might have a potential role in heterosis for the three traits.

In conclusion, the genomic regions, candidate genes and the SNP haplotypes diversity groups identified within all the three traits in this study provide interesting information about their possible role in the respective trait heterosis. Therefore, these interesting genomic regions could be tracked and considered during future marker assisted selection or 'precision predictive hybrid breeding' for the improvement of allopolyploid spring-type canola.

4.2 Genomic predictions in hybrid *Brassica napus*

The first investigation of the potential of genomic selection in *B. napus* breeding (Würschum et al. 2014) investigated a relatively narrow set of winter oilseed rape breeding lines derived from 9 elite parental lines that were genotyped with only 253 SNP markers. To my knowledge, this study is the first report of testcross performance prediction in this important oil crop species. The population size, the represented genetic diversity and the number of SNP markers used in my analysis were all considerably larger than the previous study of (Würschum et al. 2014).

I investigated genomic prediction accuracies for seven key agronomic traits, including seed yield, oil content and quality related traits using a diverse population of spring-type canola. The RR-BLUP method used for the prediction modeling has been shown to be effective in accounting for both major and minor effect quantitative trait loci (QTL) in plant breeding (Würschum et al. 2013; Reif et al. 2013; Würschum et al. 2014).

4.2.1 Independent genomic prediction across the whole population

First, I investigated genomic prediction accuracy for each trait within the whole-population. Taking the whole population under consideration, the lowest genomic prediction accuracy was estimated for seedling emergence and highest for seed oil content. The low genomic prediction accuracy for seedling emergence under scenario 1 may be explained by the low heritability and genetic variance for this trait. To increase prediction accuracy in such traits as future strategy is to combine these with other correlated highly heritable traits in a multi-trait genomic prediction model. In the case of seed oil content, the prediction accuracy remained high across the whole population. This is presumably due to the high heritability and the comparatively simple genetic architecture underlying this trait, where a few major QTL control maximum phenotypic variance (Wu et al. 2006 and Delourme et al. 2006).

Riedelsheimer et al. (2012) and Saatchi et al. (2011) reported that population substructure might affect genomic prediction accuracies. In my dataset, implementation of independent prediction within subpopulations increased prediction accuracies in specific subpopulations for low to moderate heritability traits like seed glucosinolate content, lodging resistance, DTF and seedling emergence. This is in line with previous studies that reported higher prediction accuracies when genetically closely individuals were used in the TP and VP (Habier et al. 2007; Hayes et al. 2009). The most straightforward explanation for such improvements might be that these traits are affected by variants at major-effect loci in some subpopulations that are rare or absent in the remainder of the materials. For some traits no improvement in accuracy was observed within subpopulations. This indicates that a large TP, in which the captured diversity strongly represents the diversity in the corresponding VP, may overcome the potential disadvantage caused by the use of genetically distant individuals in the TP and VP. Adding a covariate to the prediction model which identified the clusters in the whole population did not improve the overall prediction accuracy for any trait. This scenario may be rather specific for canola, in which modern, adapted breeding pools have a particularly narrow genetic basis (Hasan et al. 2006; Cowling et al. 2007; Bus et al. 2011). The situation is very different in maize or cattle, for example, where genetic differentiation among sub-populations or races are highly pronounced and population differences in gene and allele content are therefore often decisive (Hayes et al. 2009; Habier et al. 2007; Technow et al. 2012). I conclude that adjustment of prediction models on a case-by case basis in canola can potentially give a small improvement in prediction of specific traits depending on the variance within a given breeding population.

For the high-value traits of seed oil content, oil yield and seed glucosinolate content, for which high heritabilities can be attributed to a better rank correlation among locations, I consistently obtained very high prediction accuracies in predictions across the entire

population regardless of substructure. This may be further due to the modulating maternal influence of the two common male-sterile testers on embryo-related traits like seed size and oil content. As noted by (Heffner et al. 2011; Saatchi et al. 2011), using combined training populations for hybrid prediction from genetically diverse parental lines can increase prediction accuracy compared to predictions based on individuals arising from the same heterotic pool.

4.2.2 Effect of TP sample size on genomic prediction accuracy

It has been shown earlier, using both simulation studies (Habier et al. 2007) as well as real datasets (Heffner et al. 2011; Saatchi et al. 2011; Zhao et al. 2012), that an increase in the training population size has a positive impact on the overall genomic prediction accuracy. In predictions across the entire test population a TP comprising 70 % of the overall population size (333 lines from 475) was sufficient to accurately predict the performance of the remaining lines for testcross performance. With the exception of flowering time, where the prediction accuracy still did not achieve a plateau even with 90% TP, only small or insignificant increases in accuracy were achieved with a TP proportion greater than 70%. The failure to achieve a plateau for flowering time suggests the presence of some accessions with distinctly different genetic control of flowering time. From a breeder's viewpoint a smaller TP size is of course advantageous to reduce phenotyping costs. The most satisfying solution is the one in which adequate selection gains are achieved without surpassing current phenotyping costs.

4.2.3 Genomic selection prospects in hybrid rapeseed

At the dawn of canola hybrid breeding, various authors reported considerable heterosis in F1 hybrids (Grant and Beversdorf, 1985; Lefort-Buson et al. 1987; Brandle and McVetty,

1989). Plant researchers and breeders take advantage of the information on the genetic diversity to exploit heterosis in the available gene bank by using cross combinations to bring a tangible improvement in important agronomic traits. With the availability of recent inexpensive genomic sequence data and *Brassica* 60k SNP chip array, large scale genotyping is no more a limiting factor which dictates and facilitates towards genome-based predictive molecular breeding in *Brassica napus.* Effective integration of multidisciplinary research areas is required to get significant improvement in the yield and yield related trait performance in this important crop species. Prediction of heterosis is vital in hybrid breeding practices. Efficient pre-selection of the optimum combinations between parental lines to produce the most promising hybrids is quite challenging. The use of molecular markers (Li et al. 2006; Badani et al. 2006; Radoev at al. 2008; Mei et al. 2011) to accelerate the differentiation of hybrid pools and investigate the genetic basis of heterosis in canola has further increased hybrid performance. However, levels of yield improvement seen in more classical hybrid crops like maize are still to achieve in canola because so far this struggle for the development of heterotic pools in canola has made only slow progress due to the generally low diversity within the species. The highly complex allopolyploid genome of *B. napus*, with multiple interacting homoeologous copies of almost all the genes (Chalhoub et al. 2014), increases the difficulty in prediction of individual gene actions (Hasan et al. 2006).

The main purpose of genomic selection is the utilization of large and inexpensive DNA marker datasets to bring an improvement to the mean performance of a certain population (Bernardo and Yu, 2007). Seed yield, seed oil content and other polygenic traits are under the influence of complex genetic and biochemical interactions, and hundreds or thousands of small-effect QTL might be involved in their expression.

From a breeder's perspective the implementation of genomic prediction is only worthwhile if equivalent or greater selection gain can be achieved with equal or reduced time and/or cost than using conventional selection methods (generally a multiple-year, multiple-location field evaluations). Depending on the selection intensity, the results presented in my study clearly demonstrate the value of performance predictions based on high-density SNP markers in hybrid canola. Of course, with the integration of additional data on the transcriptome, metabolome and epigenome may improve prediction accuracies further in the coming time. Therefore, I anticipate that the active uses of system biology approaches into canola hybrid breeding practices in the future would be rewarding. Even where little improvement on phenotypic selection gain is achieved through genomic prediction, the method is still of considerable value for traits like seedling emergence, where the very low heritability seedlots generated in multiple maternal environments combined with multi-location field evaluations. In such cases an increase in genetic gain might still be achieved if the early pre-selection approach enables a shortening of the breeding cycle. The results of my study suggest that prediction of testcross performance in canola breeding, where molecular variants are used across the whole genome, taking both large and small QTL effects into account, could be a promising avenue for improving important commercial agronomic traits without consideration of detailed *a priori* knowledge of their underlying genetic architecture by saving both time and resources.

Chapter 5: Summary

Canola/rapeseed (*Brassica napus* L., (AACC, 2n=38) is one of the world's most important oilseed crops and is used as human food, i.e. cooking oil and as animal feed. In Europe, winter-type canola is also used as a sustainable source of bioenergy. Canola was naturally formed ~7500 years ago from spontaneous inter-specific hybridisations between cabbage (*Brassica oleracea*) and turnip rape (*Brassica rapa*). Recently, the reference genome of the *B. napus* 'Darmor-bzh 'cultivar was sequenced and published in Science (Chalhoub et al. 2014) which provides new insights to be explored, to further improve this important oil crop in the coming time.

Commonly used breeding materials of canola have developed a narrow gene pool due to continuous strong conscious selection by breeders for quality traits, i.e. low seed glucosinolate, low erucic acid, etc. Attempts have been made over the years to boost up the genetic diversity of canola through introgression from its progenitor species or other exotic materials. Breeders practice hybrid breeding in canola to exploit heterosis for improved agronomic traits, especially for seed yield gain and yield stability. Canola is considered to have a well-defined pollination control system, for example, cytoplasmic male sterility system (CMS), genic male sterility system (GMS), etc. and can be used for the production of hybrid seed with optimum success.

Due to the recent advances in high-throughput genomic technologies, an avalanche of inexpensive single nucleotide polymorphism (SNP) markers is now available. These genome-wide markers have made molecular predictive breeding possible and lucrative in different crop species, i.e. Maize, rice, etc. I used the 60k *Brassica* SNP Illumina genotyping array in my study. After rigorous quality checks, a panel of single position 24,442 polymorphic SNPs distributed across the whole genome were used in my genomic

analyses. First, I investigated the population structure in my dataset using the whole set of filtered SNP markers. Based on the *K* means clustering method, two main clusters along with one small cluster were diagnosed. I also explored chromosome-wise linkage disequilibrium (LD) decay within both the subgenomes A and C. The general pattern of LD was more conserved on C- subgenome than A- subgenome. This was in congruence with the previous reported studies in canola.

Genome-wide association studies (GWAS) have emerged as a useful approach in genetics and has been used to correlate molecular markers with phenotypic variations in various crop populations. I used the GWAS approach to unravel genomic regions contributing to hybrid performance in canola and have identified candidate genes that have pleiotropic effects for two or more different traits. It has been reported already (Qian et al. 2007) that in canola hybrid breeding, additive gene effects are the main contributors to heterosis. General combining ability (GCA) accounts for additive gene effects. Therefore, GCA values were estimated for each pollinator in a set of 475 male lines and used in my genomic analyses instead of the *per se* F1 phenotype data. I used a mixed effects model approach which effectively accounts for the cryptic population structure. For GWAS, we considered three important agronomic traits, i.e. GCA for seed yield, GCA for DTF (days to flowering) and GCA for seed oil content.

On chromosome A3, I found some *Arabidopsis* orthologue candidate genes with pleiotropic effects associated with significant SNP loci related to GCA for seed yield and GCA for DTF. For example, *FRIGIDA (FRI)* and *EMBRYONIC FLOWER 2* genes which have been shown already in their direct role in flowering time and indirect role in yield related traits. Similarly, I reported a very important *A.thaliana* orthologue candidate gene (*ABERRANT LATERAL ROOT FORMATION 1: ALF1*) significantly associated with an overlapping SNP between GCA for seed yield and GCA for seed oil content. This gene is involved in various

biochemical pathways, for example, unsaturated fatty acid biosynthetic process, adventitious root development, defense response, sulfur compound biosynthetic process and glucosinolate biosynthetic process. I also identified significant SNP haplotype diversity groups or blocks in the flanking regions of the significant SNPs in each trait that might contribute to trait heterosis in each of my examples. At the end, I reconstructed predicted F1 genotypes from the genotypes of the significant haplotypes from male lines (pollinators) and their corresponding haplotypes on the two tester lines (M1 and M2) in each trait. The genomic regions, candidate genes and the predicted F1 hybrid genotypes identified in my study provide worthwhile information that could be used in the future hybrid breeding strategies.

My second project focused on the whole-genome prediction of hybrid performance in canola instead of identifying individual genes. Genome-wide selection (GS) or genomic prediction of unphenotyped germplasms (Meuwissen et al. 2001) is now rapidly making its way into plant breeding. In GS, molecular markers are employed across the whole genome simultaneously and genomic breeding values (GEBVs) are estimated. Pre-selection of the unphenotyped material is made on the basis of these GEBVs. Genomic prediction of test-cross hybrid performance in canola using widely-tested ridge-regression best linear unbiased prediction (RR-BLUP) model was carried out in this study taking seven important agronomic traits under consideration. These were seed yield (dt/ha), oil yield (dt/ha), seed oil content (% volume per seed dry weight), content of total seed glucosinolate (GSL; µmol/g seed), seedling emergence (visual observation ranging from a minimum value of 1 to maximum 9), lodging resistance (visual observation ranging from a minimum value of 1 to maximum 9) and days to onset of flowering (DTF; measured as number of days from sowing until 50% flowering plants per plot).

Based on the observed population stratification in my dataset, I devised three scenarios for the genomic prediction. First, I considered the whole population, including all the pollinators (475) and then across two main clusters independently. In the whole population scenario, the highest prediction accuracy was achieved for seed oil content (r_{GPA} = 0.81) and lowest for the least heritable trait, seedling emergence (r_{GPA} = 0.29). No uniform improvement was seen in genomic prediction accuracies across individual clusters. The results of my study, however, suggest that prediction of testcross performance in hybrid spring-type canola breeding, where molecular variants are used across the whole genome, could be an efficient and cost-effective breeding approach to improve this important allopolyploid species.

Chapter 6: Zusammenfassung

Raps (*Brassica napus* L., (AACC, 2n=38) zählt zu den weltweit wichtigsten Ölsaaten und spielt eine bedeutende Rolle sowohl als Nahrungsmittel - vor allem in Form von Speiseöl – wie auch als Futtermittel in der Tierernährung. Zudem wird in Europa Winterraps verstärkt als nachwachsender Bioenergierohstoff angebaut. Die Kulturart Raps entstand vor etwa 7500 Jahren im Zuge einer spontanen interspezifischen Hybridisierung zwischen Kohl (*Brassica oleracea*) und Rübsen (*Brassica rapa*). Das vor kurzem sequenzierte und in Sciene veröffentlichte *B. napus* Referenzgenom „Darmor-bzh“ (Chalhoub et al. 2014) stellt derzeit eine beispiellose Möglichkeit zur Erforschung und Weiterentwicklung dieser bedeutenden Ölpflanze dar.

Durch kontinuierliche, artifiziell gerichtete Selektion auf Qualitätsmerkmale, darunter Glucosinolatreduktion und Erucasäurefreiheit im Samen, sind heutige Züchtungspools durch eine relativ geringe genetische Diversität charakterisiert. Zahlreiche Versuche wurden in der Vergangenheit unternommen, diese reduzierte Diversität durch Introgression genetischen Materials aus den Raps-Vorfahren sowie aus exotischen Genotypen zu erweitern. Ein enormer Vorteil der Hybridzüchtung im Raps besteht darin, dass unter Ausnutzung des Heterosiseffekts wichtige agronomische Merkmale optimiert werden können, insbesondere Kornertrag und Ertragsstabilität. Für Raps sind umfassend charakterisierte Bestäubungskontrollmechanismen, wie z.B. die cytoplasmatische männliche Sterilität (CMS, INRA-Ogura) und andere kerngenisch bedingte männliche Sterilitätssysteme (MS) beschrieben und werden für die Produktion von F1-Hybridsorten erfolgreich eingesetzt.

Infolge der ständig fortschreitenden Weiterentwicklung von genomischen Hochdurchsatzmethoden ist gegenwärtig eine große Anzahl an Einzelnukleotid-

Polymorphismus Markern (Single Nucleotide Polymorphism; SNP) verfügbar. Der Einsatz dieser genomweiten molekularen Marker ermöglicht die Prädiktion von Qualitäts- und Leistungsparametern in Zuchtprogrammen, was bereits in verschiedenen Kulturen erfolgreich demonstriert werden konnte, beispielsweise in Mais und Reis. Nach einer strikten Qualitätskontrolle ergab sich für unsere genomischen Analysen ein Set aus 24.442 polymorphen, über das gesamte Genom verteilten SNP-Markern. Zunächst wurde unter Verwendung dieses Markersets die Populationsstruktur der Genotypen untersucht. Basierend auf der „k means clustering“ Methode konnten zwei große Hauptcluster und ein kleines Cluster identifiziert werden. Zudem wurde der Verfall des Kopplungsphasenungleichgewichts (Linkage Disequilibrium decay; LD-decay) auf allen Chromosomen berechnet. Im Allgemeinen war das LD im C-Subgenom stärker ausgeprägt als im A-Subgenom, was im Einklang mit anderen Studien in Raps steht.

Die Genomweite Assoziationskartierung (Genome-Wide Association Studies; GWAS) konnte in den letzten Jahren als nützliche Methode zur Identifikation von Assoziationen zwischen molekularen Markern und phänotypischen Merkmalen in verschiedenen landwirtschaftlichen Kulturen etabliert werden. Im Rahmen dieser Arbeit wurde dieses Verfahren zur Lokalisierung genomischer Regionen angewendet, welche Einfluss auf die Hybridleistung in Raps haben. Dabei wurden verschiedene Kandidatengene mit pleiotropen Effekten auf zwei oder mehr Merkmale entdeckt. Qian et al. (2007) schlussfolgerten bereits, dass hauptsächlich additive Geneffekte an der Ausprägung von Heterosis in Rapshybriden beteiligt sind. Dabei werden die additiven Geneffekte durch die Schätzwerte der Allgemeinen Kombinationsfähigkeit (General Combining Ability; GCA) repräsentiert. Dementsprechend wurden die GCA-Werte für alle 475 Vaterlinien errechnet und anstelle der phänotypischen *per se* Leistung der F1-Hybriden in den genomischen Analyseverfahren verwendet. Es wurde ein Gemischtes Lineares Modell verwendet,

welches die Populationsstruktur berücksichtigt. Bei den GWAS-Analysen wurden drei agronomisch wichtige Merkmale betrachtet: GCA für Kornertrag, GCA für den Blühzeitpunkt (Days to Flowering; DTF) und GCA für Ölgehalt im Samen.

Auf Chromosom A3 konnten mehrere *Arabidopsis*-orthologe Kandidatengene mit pleiotropen Effekten gefunden werden, welche mit SNP Loci assoziiert sind, die mit GCA für Kornertrag und GCA für den Blühzeitpunkt in Verbindung stehen. Dazu zählten beispielsweise die Gene *FRIGIDA (FRI)* und *EMBRYONIC FLOWER 2*, deren direkte Rolle für den Blühzeitpunkt sowie indirekte Rolle für ertragsbildende Merkmale bereits demonstriert werden konnte. Weiterhin wurde ein wichtiges, *A. thaliana*-orthologes Kandidatengen gefunden (*ABERRANT LATERAL ROOT FORMATION 1: ALF1*) welches mit einem SNP assoziiert ist, der sowohl mit GCA für Kornertrag wie auch GCA für Ölgehalt überlappt. Dieses Gen ist in verschiedenen biochemischen Prozessen involviert, beispielsweise in der Biosynthese ungesättigter Fettsäuren, der Entwicklung von Adventivwurzeln, in Abwehrreaktionen sowie der Biosynthese von schwefelhaltigen Verbindungen und Glucosinolaten. Außerdem konnten signifikante SNP Haplotyp-Diversitätsblöcke in den flankierenden Regionen der entsprechenden SNPs identifiziert werden, die zur Heterosis in den untersuchten Merkmalen beitragen. Schließlich wurden die vorhergesagten F1 Genotypen basierend auf den signifikanten Haplotypen der männlichen Bestäubungslinien und den entsprechenden Haplotypen der beiden Mutterlinien (M1 und M2) für jedes Merkmal rekonstruiert. Die hier identifizierten Genomregionen, Kandidatengene und prognostizierten Genotypen der F1-Hybriden liefern wertvolle Informationen für zukünftige Hybridzüchtungsstrategien.

Der Fokus des zweiten Projekts lag auf der genombasierten Prädiktion der Hybridleistung in Raps. Genomweite Selektion (GS) bzw. genomische Prädiktion von nicht-phänotypisiertem Zuchtmaterial (Meuwissen et al. 2001) erhält fortschreitend Einzug in die

Pflanzenzüchtung. In der genomischen Selektion werden zahlreiche, über das gesamte Genom verteilte molekulare Marker simultan für die Schätzung genomischer Zuchtwerte (Genomic Breeding Value; GEBV) genutzt. Die Vorselektion nicht-phänotypisierten Zuchtmaterials wird anschließend auf Basis des GEBVs durchgeführt. In dieser Studie wurde die Testkreuzungsleistung mittels ridge-regression best linear ubiased prediction (RR-BLUP) Modell für insgesamt sieben agronomisch wichtige Merkmale vorhergesagt. Diese umfassten den Kornertrag (dt/ha), Ölertrag (dt/ha), Ölgehalt im Samen (Volumen-% in der Trockenmasse) und Glucosinolatgehalt im Samen (µmol/g), den Aufgang (Bonitierung von 1-9), die Standfestigkeit (Bonitierung von 1-9) und die Dauer bis zur Blüte (Anzahl an Tagen bis sich 50% der Pflanzen je Parzelle in der Blüte befanden).

Basierend auf der im Datensatz beobachteten Populationsstruktur wurden drei verschiedene Szenarios für die Durchführung der Genomischen Prädiktion entwickelt. Dabei wurde zunächst die gesamte Population aller 475 Bestäuberlinien betrachtet, anschließend die beiden großen Hauptcluster separat. In der Gesamtpopulation wurde die höchste Prädiktionsgenauigkeit für den Ölgehalt im Samen beobachtet (r_{GPA} = 0.81), die geringste Genauigkeit für den Aufgang (r_{GPA} = 0.29), das Merkmal mit der geringsten Heritabilität. Eine allgemeine Steigerung der Prädiktionsgenauigkeit bei Betrachtung der beiden einzelnen Clustern konnte nicht festgestellt werden. Die Ergebnisse dieser Studie zeigen, dass die Vorhersage der Testkreuzungsleistung in Zuchtprogrammen mittels genomischer Prädiktion eine effiziente und ökonomische Methode zur Optimierung von Sommerraps darstellt.

Chapter 7: References

Ahuja I, Borgen BH, Hansen M, Honne BI and Müller C (2011) Oilseed rape seeds with ablated defence cells of the glucosinolate–myrosinase system. Production and characteristics of double haploid MINELESS plants of *Brassica napus* L., J of Exp Bot. (62): 4975–4993.

Albrecht T, Wimmer V, Auinger HJ, Erbe M, Knaak C, et al. (2011) Genome-based prediction of testcross values in maize. Theor Appl Genet. (123): 339–50.doi: 10.1007/s00122-011-1587-7.

Alboresi A, Gestin C, Leydecker MT, Bedu M and Meyer C, et al. (2005) Nitrate, a signal relieving seed dormancy in *Arabidopsis*. Plant Cell and Environment (28): 500–512.

Almagro A, Lin SH and Tsay YF (2008) Characterization of the *Arabidopsis* nitrate transporter NRT1.6 reveals a role of nitrate in early embryo development. The Plant Cell (20): 3289–3299.

Alonso-Blanco C, Blankestijn-de Vries H, Hanhart CJ and Koornneef M (1999) Natural allelic variation at seed size loci in relation to other life history traits of *Arabidopsis thaliana*. Proc Natl Acad Sci. USA. (96): 4710–4717.

Andersen MN, Heidmann T and Plauborg F (1996) The effects of drought and nitrogen on light interception, growth and yield of winter oilseed rape. Acta Agric Scand, Sect B, Plant Soil Sci. (46): 55–67.

Asoro FG, Newell MA, Beavis WD, Scott MP, Jannink J (2011) Genomic Selection Accuracy using Multifamily Prediction Models in a Wheat Breeding Program. The Plant Genome (4):65–75.

Atkinson NJ and Urwin PE (2012) The interaction of plant biotic and abiotic stresses: from genes to the field. J Exp Bot. (63): 3523–3543.

Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, et al. 2010, Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. Nature (465): 627–31.

Aulchenko YS, Ripke S, Isaacs A and van Duijn (2007) GenABEL: an R library for genome-wide association analysis. Bioinformatics 23 (10):1294-1296.

Avendano S, Woolliams JA, Villanueva B (2005) Prediction of accuracy of estimated Mendelian sampling terms. J Anim Breed Genet. (122): 302–308.

Badani AG, Snowdon RJ, Wittkop B, Lipsa FD, Baetzel R, et al. (2006) Colocalization of a partially dominant gene for yellow seed colour with a major QTL influencing acid detergent fibre (ADF) content in different crosses of oilseed rape (*Brassica napus*). Genome (49): 1499–1509.doi: 10.1139/g06-091.

Bastiaansen JWM, Coster A, Calus MPL, van Arendonk JAM, and Bovenhuis H (2012) Long-term response to genomic selection: effects of estimation method and reference population structure for different genetic architectures. Genet Sel Evol. (44): 1-3.

Basunanda P, Spiller TH, Hasan M, Gehringer A, Schondelmaier J, et al. (2007) Marker-assisted increase of genetic diversity in a double-low seed quality winter oilseed rape genetic background. Plant Breeding 126 (6): 581–587. doi:10.1111/j.1439-0523.2007.01404.x.

Beck DL, Vaal SK and Crossa J (1990) Heterosis and combining ability of CIMMYT's tropical early and intermediate maturity maize (*Zea mays*) germplasm. Maydica (35): 279-285.

Becker H (2011) Pflanzenzüchtung. Verlag Eugen Ulmer Stuttgart., pp.287-289. ISBN.978-3-8001-2940-9.

Bekele WA, Fiedler K, Shiringani A, Schnaubelt D, Windpassinger S, et al. (2014) Unravelling the genetic complexity of sorghum seedling development under low-temperature conditions. Plant Cell & Environment. 37 (3): 707–723.doi: 10.1111/pce.12189.

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

Bennett RA, Seguin-Swartz G and Rahman H (2012) Broadening Genetic Diversity in Canola Using the C-Genome Species *Brassica oleracea* L. Crop Sci. (52): 2030-2039.

Bernardo R and Yu J (2007) Prospects for genome wide selection for quantitative traits in maize. Crop Sci. (47): 1082–1090. doi: 10.2135/cropsci2006.11.0690

Borlaug NE (1983) Contributions of conventional plant breeding to food production. Science 219 (4585): 689-93.

Bouchereau A, Clossais-Besnard N, Bensaoud A, Leportb L and Renard M, et al. (1996) Water stress effects on rapeseed quality. Eur J Agron. (5): 19–30. doi: 10.1016/S1161-0301(96)02005-9.

Brachi B, Morris GP, Borevitz JO (2011) Genome-wide association studies in plants: the missing heritability is in the field. Genome Biology 12: 232. doi: 10.1186/gb-2011-12-10-232.

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

Brandle JE, McVetty PBE (1989) Heterosis and Combining Ability in Hybrids Derived from Oilseed Rape Cultivars and Inbred Lines. Crop Science 29 (5): 1191. doi: 10.2135/cropsci1989.0011183X002900050020x.

Bus A, Körber N, Snowdon RJ, Stich B (2011) Patterns of molecular variation in a species-wide germplasm set of *Brassica napus*. Theor Appl Genet. 123 (8): 1413–1423. doi: 10.1007/s00122-011-1676-7.

Buzza GC (1995) Plant breeding. In: Kimber DS, McGregor DI. (Ed.) Brassica oilseeds: Production and utilization. CABI publishing, Wallingford, CT. pp. 153-175. doi: 10.1002/lipi.19960980908.

Cabrera-Bosquet L, Crossa J, von Zitzewitz J, DolorsSerret M, Araus JL (2012) High-throuput phenotyping and genomic selection: The frontiers of crop breeding converge. Journal of Integrative Plant Biology 54(5): 312-320.

Cai D, Xiao Y, Yang W, Ye W, Wang B, et al. (2014) Association mapping of six yield related traits in rapeseed (*Brassica napus* L.). Theor Appl Genet. (127): 85–96.

Caliński T & Harabasz J (1974) A dendrite method for cluster analysis. Commun. Stat. (3): 1 – 27. doi:10.1080/03610927408827101.

Calus ML, Veerkamp RF (2007) Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. J Anim Breed Genet. (124):362-368.

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

Cavalli-Sforza LL, Feldman MW (2003) The application of molecular genetic approaches to the study of human evolution. Nat Genet. 33(Supplement): 266–275. Historical article.

Chalhoub B, Denoeud F, Liu S, Parkin, Isobel AP, et al. (2014) Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. Science 345 (6199): 950–953. doi: 10.1126/science.1253435.

Chen C, DeClerck G, Tian F, Spooner W, McCouch S, et al. (2012) PICARA, an analytical pipeline providing probabilistic inference about a priori candidates genes underlying genome-wide association QTL in plants. PloS One, 7:e46596.

Cho YC, Jeung JU, Park HJ, Yang CI, Choi YH, et al. (2008) Haplotype diversity and durability of resistance genes to blast in Korean Japonica rice varieties. J Crop Sci Biotech. (11): 205–214.

Coster A, Bastiaansen JWM, Calus MPL, van Arendonk JAM, Bovenhuis H (2010) Sensitivity of methods for estimating breeding values using genetic markers to the number of QTL and distribution of QTL variance. Genet Sel Evol. 42:9

Courtois B, Audebert A, Dardou A, Roques S and Ghneim- Herrera T, et al. (2013) Genome-wide association mapping of root traits in a japonica rice panel. PloS One, 8: e78037.

Cowling WA (2007) Genetic diversity in Australian canola and implications for crop breeding for changing future environments. Field Crops Research 104 (1-3): 103–111. doi.org/10.1016/j.fcr.2006.12.014.

Crossa J, Perez P, Hickey J, Burgueno J, Ornella L, et al. (2013) Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity 1-13.

Crossa J, de los Campos G, Perez P, Gianola D, Burgueno J, et al. (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics (186): 713–724.

Daetwyler HD, Pong-Wong R, Villanueva B, Wooliams JA (2010) The impact of genetic architecture on genome-wide evaluation methods. Genetics (185): 1021–1031.

Danesh-Shahraki A, Nadian H, Bakhshandeh A, Fathi G and Alamisaied K, et al. (2008) Optimization of irrigation and nitrogen regimes for rapeseed production under drought stress. J Agron. (7): 321–326. doi: 10.3923/ja.2008.321.326.

Darwin C (1859) On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. Nature (1st Ed.) (London: John Murray) 5 (121): 502.

Delourme R, Falentin C, Huteau V, Clouet V, Horvais R, et al. (2006) Genetic control of oil content in oilseed rape (*Brassica napus* L.). Theor Appl Genet. 113 (7): 1331–1345. doi: 10.1007/s00122-006-0386-z.

de los Campos G, Gianola D, Rosa GJ, KA Weigel KA, Crossa J, (2010) Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genet Res. (92): 295– 308.

de los Campos G, Naya H, Gianola D, Crossa J, Legarra A, et al. (2009) Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics (182): 375–385

Duvick DN, (1999) Heterosis: Feeding people and protecting natural resources. Genetics and Exploitation of Heterosis in Crops, Coors J.G., Pandey S., editors. , (Madison, WI: American Society of Agronomy and Crop Science Society of America;), pp. 19–29.

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

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

FAO (2011) FAOSTAT statistical database (http://faostat3.fao.org)

FAO (2013) FAOSTAT statistical database (http://faostat3.fao.org)

Falconer DS and Mackay TFC (1996) Introduction to Quantitative Genetics. Ed 4. Longmans Green, Harlow, Essex, UK.

Fehr WR (1987) Principles of cultivar development: volume 1. Theory and Technique. Macmillan Publishing Company, New York.

Flint-Garcia SA (2003) Structure of linkage disequilibrium in plants. Annu Rev Plant Biol. (54): 357-74. doi: 10.1146/annurev.arplant.54.031902.134907

Fisher RA (1918) The correlation between relatives on the supposition of Mendelian inheritance. Trans R Soc. Edinb. (52): 399-433.

Franklin I and Lewontin RC (1970) Is the gene the unit of selection? Genetics (65):7007-734.

Franklin I, Lewontin RC (1970) Is the gene the unit of selection? Genetics 65(4):707–734.

Frisch M, Thiemann A, Fu J, Schrag TA, Scholten S, et al. (2010) Transcriptome-based distance measures for grouping of germplasm and prediction of hybrid performance in maize. Theor Appl Genet. (120):441–450.

Fu YB and Gugel RK (2010) Genetic diversity of Canadian elite summer rape (*B. napus*) cultivars from the pre- to post-canola quality era. Can J Plant Sci. (90): 23-33.

Gaut BS, and Long AD (2003) The lowdown on linkage disequilibrium. Plant Cell (15): 1502–1506.

Girke A, Schierholt A, Becker HC (2012) Extending the rapeseed gene pool with resynthesized *Brassica napus* II: Heterosis. Theor Appl Genet. (124): 1017-1026.doi: 10.1007/s00122-011-1765-7.

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

Goddard ME (2009) Genomic selection: prediction of accuracy and maximisation of long term response. Genetica (136): 245-257.

Grant I, Beversdorf WD (1985) Heterosis and combining ability estimates in spring-planted oilseed rape (*Brassica napus* L.). Can J Genet Cytol. 27 (4): 472–478. doi: 10.1139/g85-069.

Gupta PK, Rustgi S, and Kulwal PL (2005) Linkage disequilibrium and association studies in higher plants: Present status and future prospects. Plant molecular biology (57): 461-485.

Habier D, Fernando RL, Dekkers, J C M (2007) The impact of genetic relationship information on genome-assisted breeding values. Genetics 177 (4): 2389–2397. doi: 10.1534/genetics.107.081190.

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

Hammond-Kosack KE and Jones JDG (2000) Response to plant pathogens. In: Buchannan B, Gruissem W, Jones R, eds. Biochemistry and molecular biology of plants. Rockville, MD: American Society of Plant Physiol. 1102–1157.

Hartigan JA, Wong MA (1979) Algorithm AS 136: A K-Means Clustering Algorithm. Applied Statistics 28 (1): 100.

Hasan M, Seyis F, Badani AG, Pons-Kühnemann J, Friedt W, et al. (2006) Analysis of Genetic Diversity in the *Brassica napus* L. Gene Pool Using SSR Markers. Genet Resour Crop Evol. 53 (4): 793–802. doi: 10.1007/s10722-004-5541-2.

Hartigan JA, Wong MA (1979) Algorithm AS 136: A K-Means Clustering Algorithm. Applied Statistics 28 (1): 100.

Hatzig SV, Frisch M, Breuer F, Nesi N and Ducournau S, et al. (2015) Genome-wide association mapping unravels the genetic control of seed germination and vigor in *Brassica napus*. Front Plant Sci. 6: 221.

Hayes BJ, Lewin HA and Goddard ME (2013) The future of livestock breeding: genomic selection for efficiency, reduced emissions intensity, and adaptation .Trends Genet. (29): 206 – 214.

Hayes BJ, Visscher PM, Goddard ME (2009) Increased accuracy of artificial selection by using the realized relationship matrix. Genetics research 91 (1): 47–60. doi: 10.1017/S0016672308009981.

Haussmann BIG, Hess DE, Omanya GO, Folkertsma RT, Reddy BV, et al. (2004) Genomic regions influencing resistance to the parasitic weed Strigahermonthica in two recombinant inbred populations of sorghum. Theor Appl Genet. (109): 1005–1016.

Hayes BJ, Visscher PM, Goddard ME (2009) Increased accuracy of artificial selection by using the realized relationship matrix. Genetics research 91 (1): 47–60. doi: 10.1017/S0016672308009981.

Hayes BJ, Cogan NOI, Pembelton LW, Goddard ME, Wang J, et al. (2013) Prospects for genomic selection in forage plant species. Plant Breeding (132): 133-143.

Hayward A, Dalton-Morgan J, Mason A, Zander M, Edwards D, et al. (2012) SNP discovery and applications in *Brassica napus* . J Plant Biotech (39):49–61.

Heffner EL, Sorrells ME, Jannink J (2009) Genomic Selection for Crop Improvement. Crop Sci. 49 (1): 1-12. doi:10.2135/cropsci2008.08.0512.

Heffner EL, Jannink J, Sorrells ME (2011) Genomic selection accuracy using multifamily prediction models in a wheat breeding program. The Plant Genome 4: 65–7. doi: 10.3835/plantgenome2010.12.0029.

Henderson CR (1975) Use of relationships among sires to increase accuracy of sire evaluation. J Dairy Sci. (58): 1731–1738.

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

Herms DA and Mattson WJ (1992) The dilemma of plants - to grow or defend.Quarterly Review of Biology,67: 283–335.

Hill WG and Robertson A (1968) Linkage disequilibrium in finite populations .Theor Appl Genet. (33): 226-231.

Hofheinz N and Frisch M (2014) Heteroscedastic ridge regression approaches for genome-wide prediction with a focus on computational efficiency and accurate effect estimation. G3 (Bethesda, Md.); 4 (3): 539–546. doi: 10.1534/g3.113.010025.

Hofheinz N, Borchardt D, Weissleder K, Frisch M (2012) Genome-based prediction of test cross performance in two subsequent breeding cycles. Theor Appl Genet (125): 1639-1645.

Huang X, Zhao Y, Wei X, Canyang L, Wang A, et al. (2012) Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. Nat Genet. (44): 32–9.

Jain HK (2010) The Green Revolution: History, Impact and Future. Studium Press, Houston, TX.

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

Jannink JL (2010) Dynamics of long-term genomic selection. Genet Sel Evol. 42:35.

Jeffreys AJ, Kauppi L and Neumann R (2001) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nat Genet. (29): 217–222.

Jonas E and Koning D de (2013) Does genomic selection have a future in plant breeding. Trends in Biotechnol. 31 (9): 497–504. http://dx.doi.org/10.1016/j.tibtech.2013.06.003.

Koebner RMD and Summers RW (2003) 21st century wheat breeding: plot selection or plate detection? Trends in Biotechnol. (21): 59–63.

Koscielny CB and Gulden RH (2012) Seedling root length in *Brassica napus* L. is indicative of seed yield, Can J Plant Sci. (92): 1229–1237.

Kebede B, Thiagarajah M, Zimmerli C, Rahman H (2010) Improvement of open-pollinated spring rapeseed (*Brassica napus* L.) through introgression of genetic diversity from winter rapeseed. Crop Sci. (50): 1236–1243.

Koornneef M, Alonso-Blanco C, Peeters AJM and Soppe W (1998) Genetic control of flowering time in Arabidopsis. Annu. Rev. Plant Physiol. (49): 345–370.

Kumar S, Marco C A, Bink M, VolzR K,Vincent G, et al. (2012) Towards genomic selection in apple (Malus × domesticaBorkh.) breeding programmes: Prospects, challenges and strategies. Tree Genetics & Genomes (8):1–14.

Lander ES and Schork NJ (1994) Genetic dissection of complex traits. Science (265): 2037–2048.

Lander ES and Botstein D (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics (121): 185–199.

Larsson SJ, Lipka, AE and Buckler ES (2013) Lessons from Dwarf8 on the strengths and weaknesses of structured association mapping. PLoS Genetics, 9, e1003246.

Leflon M, Grandont L, Eber F, Huteau V, Coriton O, et al. (2010) Crossovers get a boost in *Brassica* allotriploid and allotetraploid hybrids. Plant Cell (22): 2253–2264.

Lefort-Buson M, Guillot-Lemoine B, Dattee Y (1987) Heterosis and genetic distance in rapeseed (*Brassica napus* L.): crosses between European and Asiatic selfed lines. Genome 29 (3): 413–418. doi: 10.1139/g87-072.

Lewontin C, and Kojima K (1960) The evolutionary dynamics of complex polymorphisms. Evolution (14): 458-472.

Lewontin RC (1964) The interaction of selection and linkage I. General considerations; heterotic models. Genetics (49): 49–67.

Li CC (1969) Population subdivision with respect to multiple alleles. Ann. Hum. Genet. 33: 23–29

Li Y, Ma C, Fu T, Yang G, Tu J, et al. (2006) Construction of a molecular functional map of rapeseed (*Brassica napus* L.) using differentially expressed genes between hybrid and its parents. Euphytica 152 (1): 25–39. doi: 10.1007/s10681-006-9173-9.

Li F, Chen B, Xu K, Wu J, Song W, et al. (2014) Genome-wide association study dissects the genetic architecture of seed weight and seed quality in rapeseed (*Brassica napus* L.) Genome-wide association study dissects the genetic architecture of seed weight and seed quality in rapeseed (*Brassica napus* L.) DNA res. (4): 355-67.doi:10.1093/dnares/dsu002.

Liu X, Zhang H, Zhao Y, Feng Z and Li Q, et al. (2013) Auxin controls seed dormancy through stimulation of abscisic acid signaling by inducing ARF-mediated ABI3 activation in Arabidopsis. Proc Natl Acad Sci. U.S.A. (110): 15485–15490.

Long W, Zou X, Zhang X (2015) Transcriptome Analysis of Canola (*Brassica napus*) under Salt Stress at the Germination Stage. PLoS ONE 10(2): e0116217. doi:10.1371/journal.pone.0116217.

Lorenz AJ, Chao S, Asoro FG, Heffner EL, Hayashi T, et al. (2011) Genomic selection in plant breeding: knowledge and prospects. Adv Agron. (110): 77–123.

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

Mackay I and Powell W (2007) Methods for linkage disequilibrium mapping in crops. Trends Plant Sci. (12): 57–63.

Makowsky R, Pajewski NM, Klimentidis YC, Vazquez AI, Duarte CW, et al. (2011) Beyond missing heritability: prediction of complex traits. PLoS Genet. 7: e1002051.

MalthusTR (1798) An essay on the Principle of Population. London: J. Johnson.

Mandel JR, Nambeesan S, Bowers JE, Marek LF, Ebert D, et al. (2013) Association mapping and the genomic consequences of selection in sunflower. PLoS Genet. 9(3):e1003378.

Mei J, Fu Y, Qian L, Xu X, Li J, et al. (2011) Effectively widening the gene pool of oilseed rape (*Brassica napus* L.) by using Chinese B. rapa in a 'virtual allopolyploid' approach. Plant Breeding 130 (3): 333–337. doi:10.1080/03610927408827101.

Meuwissen T, Hayes B, Goddard M (2013) Accelerating Improvement of Livestock with Genomic Selection. Annual Review of Animal Biosciences (1): 221–237. doi: 10.1146/annurev-animal-031412-103705.

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

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

Moose SP and Mumm RH (2008) Molecular plant breeding as the foundation for 21st century crop improvement. Plant Physiology (147): 969–977.

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

Nakaya A, and Isobe SN (2012) Will genomic selection be a practical method for plant breeding? Ann Bot. (Lond.) (110): 1303–1316.

Nordborg M and Weigel D (2008) Next-generation genetics in plants. Nature (456): 720–3.

Norouzi M, Toorchi M, Salekdeh GH, Mohammadi SA and Neyshabouri MR, et al. (2008) Effect of water deficit on growth, grain yield and osmotic adjustment in rapeseed. J Food Agric Environ. (6): 312–318.

OGTR (2011). The biology of Brassica napus L. (canola) v2.1. Document prepared by the Office of the Gene Technology Regulator, Canberra, Australia, available online at http://www.ogtr.gov.au/

Omidi H, Tahmasebi Z, Naghdi Badi HA, Torabid H and Miransarid M (2010) Fatty acid composition of canola (*Brassica napus* L.), as affected by agronomical, genotypic and environmental parameters. Comptes Rendus Biologies (333): 248–254.

Pasam RK, Sharma R, Malosetti M, van Eeuwijk FA, Haseneyer G, et al. (2012) Genome-wide association studies for agronomical traits in a worldwide spring barley collection. BMC Plant Biol. 12:16.

Patterson N, Price AL, and Reich D (2006) Population structure and Eigenanalysis. PLoS Genet 2: e190. doi:10.1371/journal.pgen.0020190.

Pérez P, de los Campos G, Crossa J, and Gianola D (2010) Genomic-enabled prediction based on molecular markers and pedigree using the Bayesian Linear Regression Package in R. Plant Gen. (3) :106-116.

Piepho HP, Möhring J, Melchinger AE, and Bŭchse A (2007) BLUP for phenotypic selection in plant breeding and variety testing. Euphytica (161): 209-228.

Piepho HP (2009) Ridge Regression and Extensions for Genomewide Selection in Maize. Crop Science 49 (4): 1165. doi: 10.2135/cropsci2008.10.0595.

Piepho HP, Ogutu JO, Schulz-Streeck T, Estaghvirou B,Gordillo A, et al. (2012) Efficient computation of ridge-regression BLUP in genomic selection in plant breeding. Crop Sci. (52):1093– 1104.

Price AL, Patterson NJ, Plenge RM, Michael EW, Shadick NA, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. (38): 904–9.

Pritchard JK, and Przeworski M (2011) Linkage Disequilibrium in Humans: Models and Data. Am J Hum Genet. (69):1–14.

Pritchard J, Stephens M and 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. Frontiers Plant Sci. 2(51).

Pungpapong V, William Muir M, Xianran Li, Zhang D, and Zhang M (2012) A Fast and Efficient Approach for Genomic Selection with High-Density Markers. G3. 2(10): 1179–1184.

Qian W, Chen X, Fu D, Zou J, Meng J (2005) Intersubgenomic heterosis in seed yield potential observed in a new type of *Brassica napus* introgressed with partial *Brassica rapa* genome. Theor Appl Genet. 110 (7): 1187–1194. doi: 10.1007/s00122-005-1932-9.

Qian W, Sass O, Meng J, Li M, Frauen M, et al (2007) Heterotic patterns in rapeseed (*Brassica napus* L.): I. Crosses between spring and Chinese semi-winter lines. Theor Appl Genet. (115): 27-34.

Qian L, Qian W and Snowdon RJ (2014) Sub-genomic selection patterns as a signature of breeding in the allopolyploid *Brassica napus* genome. BMC Genomics.15:1170. doi: 10.1186/1471-2164-15-1170.

Qian W, Chen X, Fu D, Zou J, Meng J (2005) Intersubgenomic heterosis in seed yield potential observed in a new type of *Brassica napus* introgressed with partial *Brassica rapa* genome. Theor Appl Genet. (110): 1187-94. doi: 10.1007/s00122-005-1932-9.

Quijada PA, Udall JA, Lambert B, Osborn TC (2006) Quantitative trait analysis of seed yield and other complex traits in hybrid spring rapeseed (*Brassica napus* L.): Identification of genomic regions from winter germ plasm. Theor Appl Genet. (113): 549-561.

R 3.1.0 Development Core Team R (2014) A language and environment for statistical computing. R Foundation for Statistical Computing, GWDG Gottingen, Germany, http://www.R-project.org (accessed 15 April, 2014).

Radoev M, Becker HC, Ecke W (2008) Genetic analysis of heterosis for yield and yield components in rapeseed (*Brassica napus* L.) by quantitative trait locus mapping. Genetics 179 (3): 1547–1558. doi: 10.1534/genetics.108.089680.

Rahman H and Kebede B (2012) Improvement of spring canola *Brassica napus* (L.) by use of winter canola. J Oilseed *Brassica* (3):1–17.

Reif JC, Zhao Y, Würschum T, Gowda M, Hahn V, et al. (2013) Genomic prediction of sunflower hybrid performance. Plant Breeding 132 (1): 107–114. doi: 10.1111/pbr.12007.

Reif JC, Hahn V and Melchinger AE (2012) Genetic basis of heterosis and prediction of hybrid performance. Helia 35, Nr. 57, pp. 1-8. doi: 10.2298/hel1257001r.

Renard M, Delourme R and Pierre J (1997) Market introduction of rapeseed hybrid varieties. GCIRC Bulletin (14): 114-119.

Resende M F Jr, Munoz P, Resende MD, Garrick DJ, Fernando RL, et al. (2012) Accuracy of genomic selection methods in a standard data set of loblolly pine (*Pinus taeda* L.). Genetics (190): 1503–1510.

Riedelsheimer C, Brotman Y, Meret M, Melchinger AE and Willmitzer L, et al. (2013) The maize leaf lipidome shows multilevel genetic control and high predictive value for agronomic traits. Sci. Rep. (3): 1−7.

Riedelsheimer C, Czedik-Eysenberg A, Grieder C, Lisec J, Technow F, et al. (2012) Genomic and metabolic prediction of complex heterotic traits in hybrid maize. Nature Genetics (44): 217-220. doi: 10.1038/ng.1033.

Rogers JS (1972) Measures of genetic similarity and genetic distances. Studies in Genetics VII. University of Texas. Publication 7213. Univ of Texas, Austin, 145-153.

Rygulla W, Friedt W, Seyis F, Lühs W, Eynck C, et al. (2007) Combination of resistance to Verticillium longisporum from zero erucic acid *Brassica* oleracea and oilseed *Brassica rapa* genotypes in resynthesized rapeseed (*Brassica napus*) lines. Plant Breeding (126): 596-602.

Saatchi M, McClure MC, McKay SD, Rolf MM, Kim J, et al. (2011) Accuracies of genomic breeding values in American Angus beef cattle using K-means clustering for cross-validation. Genet Sel Evol. 43: 40. doi:10.1186/1297-9686-43-40.

Saitou N and Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution, volume 4, issue 4, pp. 406-425.

Schrag TA, Melchinger AE, Sørensen AP, Frisch M (2006) Prediction of single-cross hybrid performance for grain yield and grain dry matter content in maize using AFLP markers associated with QTL. Theor Appl Genet.113 (6): 1037–1047. doi: 10.1007/s00122-006-0363-6.

Shindo C, Aranzana MJ, Lister C, Baxter C and Nicholls C, et al. (2005) Role of FRIGIDA and FLOWERING LOCUS C in determining variation in flowering time of *Arabidopsis*. Plant Physiol. (138): 1163–1173.

Shull GH (1908) The composition of a field of maize. Am. Breeders Assoc. Rep. (4): 296–301.

Snowdon RJ, Abbadi A, Kox T, Schmutzer T, Leckband G (2015) Heterotic Haplotype Capture: precision breeding for hybrid performance. Trends in Plant Science 20 (7): 410–413. idoi:10.1016/j.tplants.2015.04.013.

Snowdon R, Friedt W (2012) Renewable energy: European biodiesel can be sustainable. Nature 490 (7418): 37. doi:10.1038/490037d.

Snowdon RJ, Iñiguez-Luy FL (2012) Potential to improve oilseed rape and canola breeding in the genomics era. Plant Breeding (131): 351-60.doi: 10.1111/j.1439-0523.2012.01976.x.

Snowdon RJ, Lühs W, Friedt W (2006) Oilseed rape. In Kole C, ed., Genome Mapping and Molecular Breeding, Vol. 2: Oil-seeds. Berlin: Springer Verlag, pp. 55-114.

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

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

Solberg T, Sonesson AK, Woolliams JA, Meuwissen THE (2009) Reducing dimensionality for prediction of genome-wide breeding values. Genet Sel Evol. 41:29.

Soller M (1978) The use of loci associated with quantitative effects in dairy cattle improvement. Anim. Prod. (27): 133–139.

Soller M, and Plotkin-Hazan J (1977) The use of marker alleles for the introgression of linked quantitative alleles. Theor Appl Genet. (51): 133–137.

Sonessen AK, Meuwissen THE (2009) Testing strategies for genomic selection in aquaculture breeding programs. Genet Sel Evol.41:37.

Stinchcombe JR, Weinig C, Ungerer M, Olsen KM and Mays C, et al. (2004) A latitudinal cline in flowering time in *Arabidopsis thaliana* modulated by the flowering time gene FRIGIDA. Proc Natl Acad Sci.USA. (101): 4712–4717.

Stuber CW, Goodman MM and Moll RH (1982) Improvement in yield and ear number resulting from selection at allozyme loci in a maize population. Crop Sci. (22): 737-740.

Stuber CW, Goodman MM, Shaffer HE and Weir BS (1980) Allozyme frequency changes associated with selection for increased grain yield in maize. Genetics (95): 225-336.

Stumpf MPH (2004) Haplotype diversity and SNP frequency dependence in the description of genetic variation. Eur J Hum Genet. 12: 469–77.

Sved JA (1971) Linkage disequilibrium and homozygosity of chromosome segments in finite populations. Theor Popul Biol. (2): 125-141.

Tanksley SD and Hewitt J (1988) Use of molecular markers in breeding for soluble solids content in tomato- a reexamination. Theor Appl Genet. (75): 811-823.

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 Genet125 (6): 1181–1194. doi: 10.1007/s00122-012-1905-

Tian F, Bradbury PJ, Brown PJ, Hung H, Sun Q, et al. (2011) Genomewide association study of leaf architecture in the maize nested association mapping population. Nat. Genet. (43): 159–62.

Tienderen PHV, Hammad I and Zwaal FC (1996) Pleiotropic effects of flowering time genes in the annual crucifer Arabidopsis thaliana (*Brassicaceae*). Am J Bot. (83): 169–174.

Turner SD (2014) qqman: an R package for visualizing GWAS results using Q-Q and Manhattan plots (http://biorXiv.org). doi: 10.1101/005165.

van Raden P M (2008) Efficient methods to compute genomic predictions. J Dairy Sci. (91): 4414–4423.

Varshney RK, Graner A and Sorrells ME (2005) Genic microsatellite markers in plants: features and applications, Trends Biotechnol. 23(1): 48–55. doi: 10.1016/j.tibtech.

Van-Esbroeck G, and Bownam DT (1998) Cotton germplasm diversity and its importance to cultivar development, J Cotton Sci. (2): 121–129.

Wang Y, Mette MF, Miedaner T, Gottwald M, Wilde P, et al. (2014) The accuracy of prediction of genomic selection in elite testcross rye populations surpasses the accuracy of marker-assisted selection and is equally augmented by multiple field evaluation locations and test years. BMC Genomics 15: 556.

Wang N, Qian W, Suppanz I, Lijuan W and Bizeng M, et al. (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 WX, Vinocur B and Altman A (2003) Plant responses to drought, salinity and extreme temperatures: towards genetic engineering for stress tolerance. Planta. (218): 1-14

Warnes G, Gorjanc G, Leisch F and Man M (2012) Package 'genetics'. Rochester, NY.

Werner JD, Borevitz JO, Uhlenhaut NH, Ecker JR and Chory J, et al. (2005) FRIGIDA-independent variation in flowering time of natural *Arabidopsis thaliana* accessions. Genetics (170): 1197–1207.

Windhausen VS, Atlin GN, Hickey JM, Crossa J, Jannink J, et al. (2012) Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. G3 (Bethesda, Md.) 2 (11): 1427–1436. doi: 10.2135/cropsci2012.08.0463.

Whittaker JC, Thompson R, Denham MC (2000) Marker-assisted selection using ridge regression. Genet Res Camb. (75): 249–252. doi:10.1111/j.1469-1809.1999.ahg634_0351_17.x.

Wu J, Shi C, Zhang H (2006) Partitioning genetic effects due to embryo, cytoplasm and maternal parent for oil content in oilseed rape (*Brassica napus* L.). Genet Mol Biol. 29 (3): 533–538. doi:10.1590/S1415-47572006000300023.

Würschum T, Abel S, Zhao Y, Léon J (2014) Potential of genomic selection in rapeseed (*Brassica napus* L.) breeding. Plant Breeding 133 (1): 45–51. doi: 10.1111/pbr.12137.

Würschum T, Langer SM, Longin C, Friedrich H, Korzun V, et al. (2013) Population structure, genetic diversity and linkage disequilibrium in elite winter wheat assessed with SNP and SSR markers. Theor Appl Genet. 126 (6): 1477–1486. doi: 10.1007/s00122-013-2065-1.

Xiao L, Zhao Z, Du DZ, Yao YM, Xu L, et al. (2012) Genetic characterization and fine mapping of a yellow-seeded gene in Dahuang (a *Brassica rapa* landrace). Theor Appl Genet. (124): 903–909. doi: 10.1007/s00122-011-1754-x.

Yu J, Pressoir G, Briggs WH, Bi IV, Yamasaki M, et al. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. (38): 203–208. doi: 10.1038/ng1702.

Zhang E, Yang Z, Wang Y, Hu Y, Song X, et al. (2013) Nucleotide polymorphisms and haplotype diversity of RTCS gene in China elite maize inbred lines. PLoS One 8, e56495. doi: 10.1371/journal.pone.0056495.

Zhao Y, Zeng J, Fernando RL and Reif JC (2013) Genomic prediction of hybrid wheat performance, Crop Sci . (53): 802-810. doi: 10.2135/cropsci2012.08.0463.

Zhao Y, Gowda M, Liu W, Würschum T, Maurer HP, et al. (2012) Accuracy of genomic selection in European maize elite breeding populations. Theor Appl Genet. (124): 769–776.doi: 10.1007/s00122-011-1745-y.

Zhao, K, Tung CW, Eizenga GC, Wright MH, Ali AL, et al. (2011) Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. Nat. Commun. 2: 467.

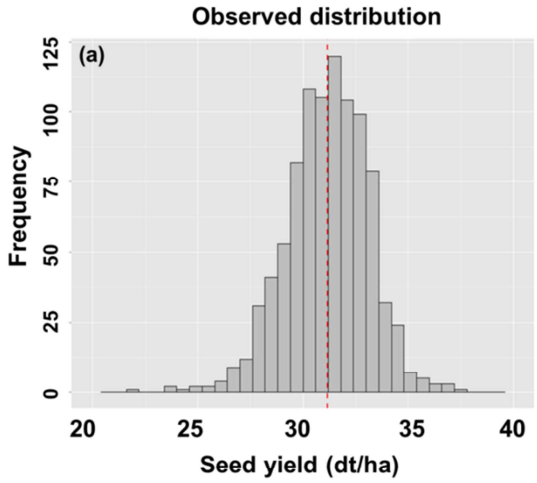
Zhang H and Forde BG (2000) Regulation of Arabidopsis root development by nitrate availability. J of Exp Bot. (51): 51–59.

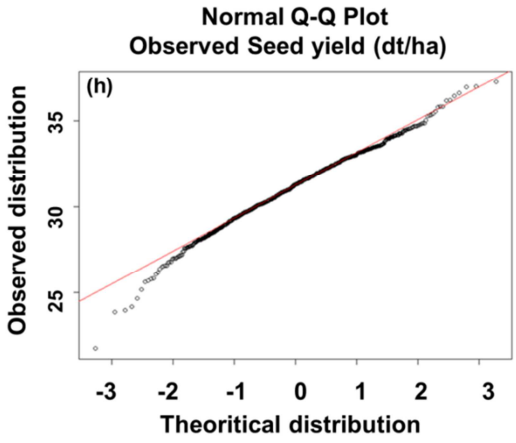
Zhang D, Lin Y, and Zhang M (2009) Penalized orthogonal-components regression for large p small n data. Electronic Journal of Statistics (3): 781–796.

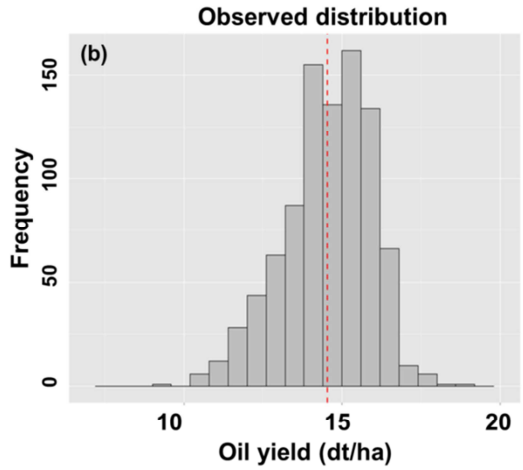
Zou J, Zhu J, Huang S, Tian E, Xiao Y, et al. (2010) Broadening the avenue of intersubgenomic heterosis in oilseed *Brassica*. Theor Appl Genet. 120 (2): 283–290. doi: 10.1007/s00122-009-1201-4.

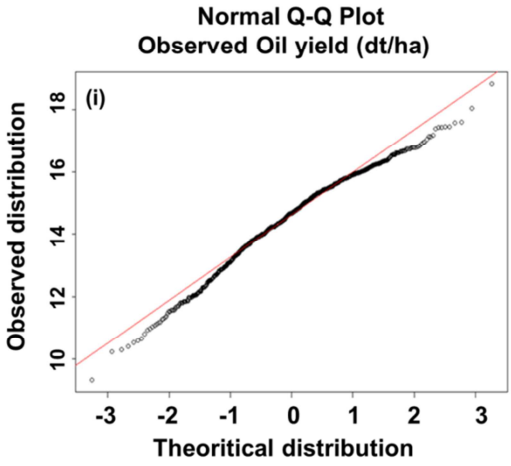
Chapter 8: Appendices

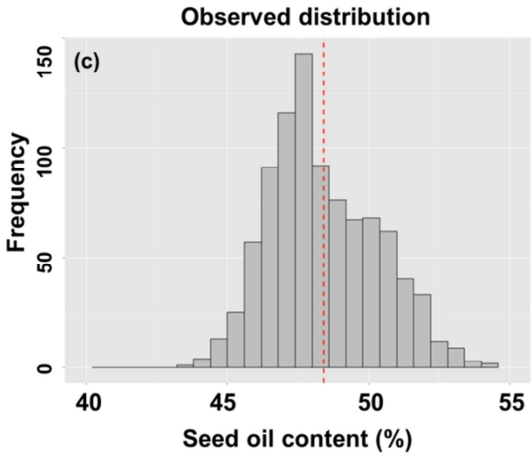
Appendix I. Supplementary figures. A: Histograms and Q-Q plots of best linear unbiased estimators (BLUEs) of each trait

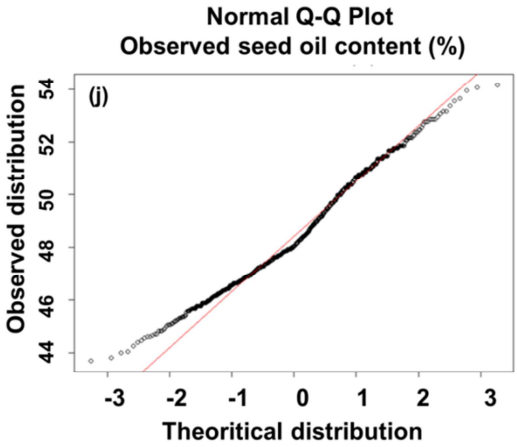


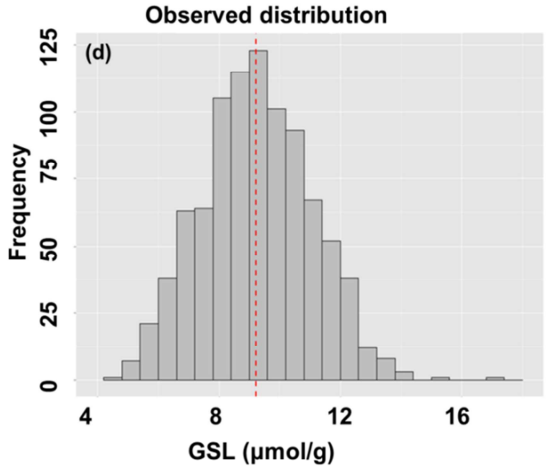


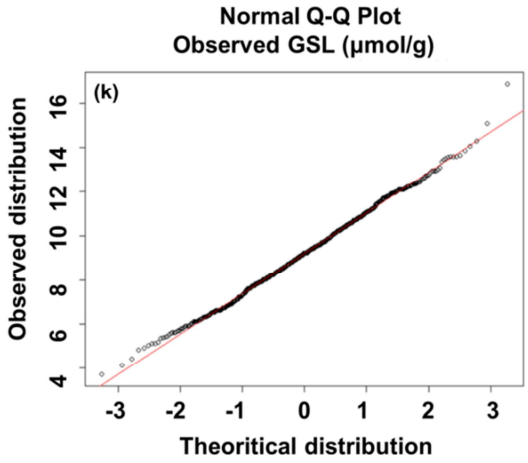


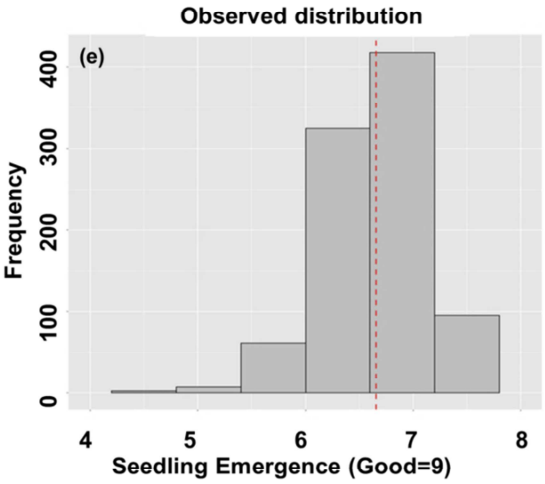


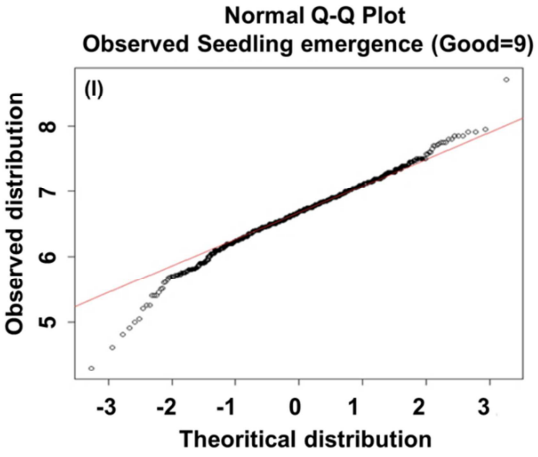


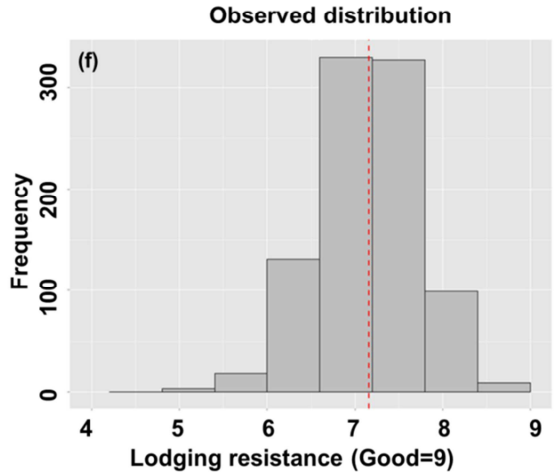


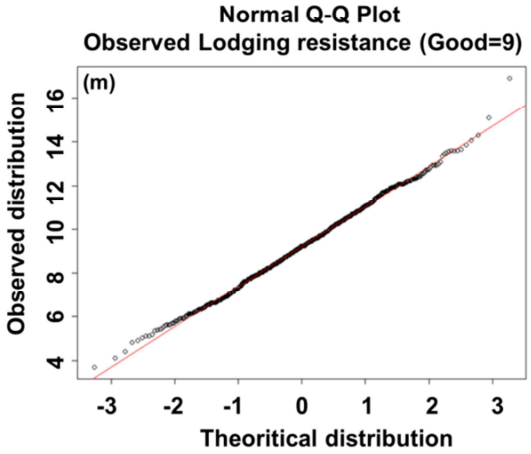

Observed distribution
(d)
Frequency
0
25
50
75
100
125
4
8
12
16
GSL (µmol/g)

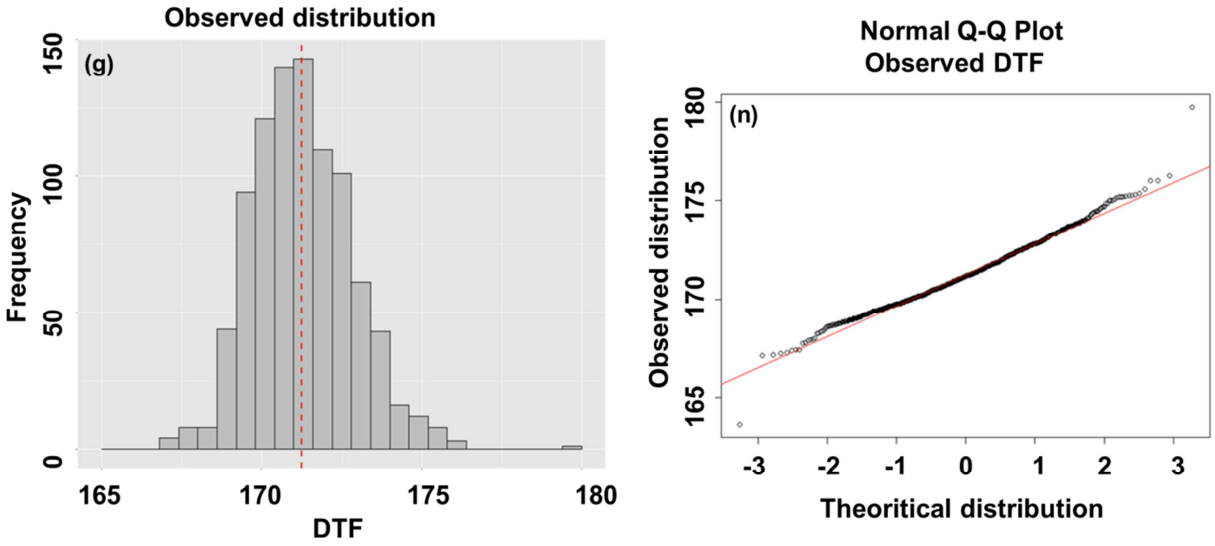

Normal Q-Q Plot
Observed GSL (µmol/g)
(k)
Observed distribution
4
6
8
10
12
14
16
-3
-2
-1
0
1
2
3
Theoritical distribution


Observed distribution
(e)
Frequency
0
100
200
300
400
4
5
6
7
8
Seedling Emergence (Good=9)


Normal Q-Q Plot
Observed Seedling emergence (Good=9)
(l)
Observed distribution
5
6
7
8
-3
-2
-1
0
1
2
3
Theoritical distribution

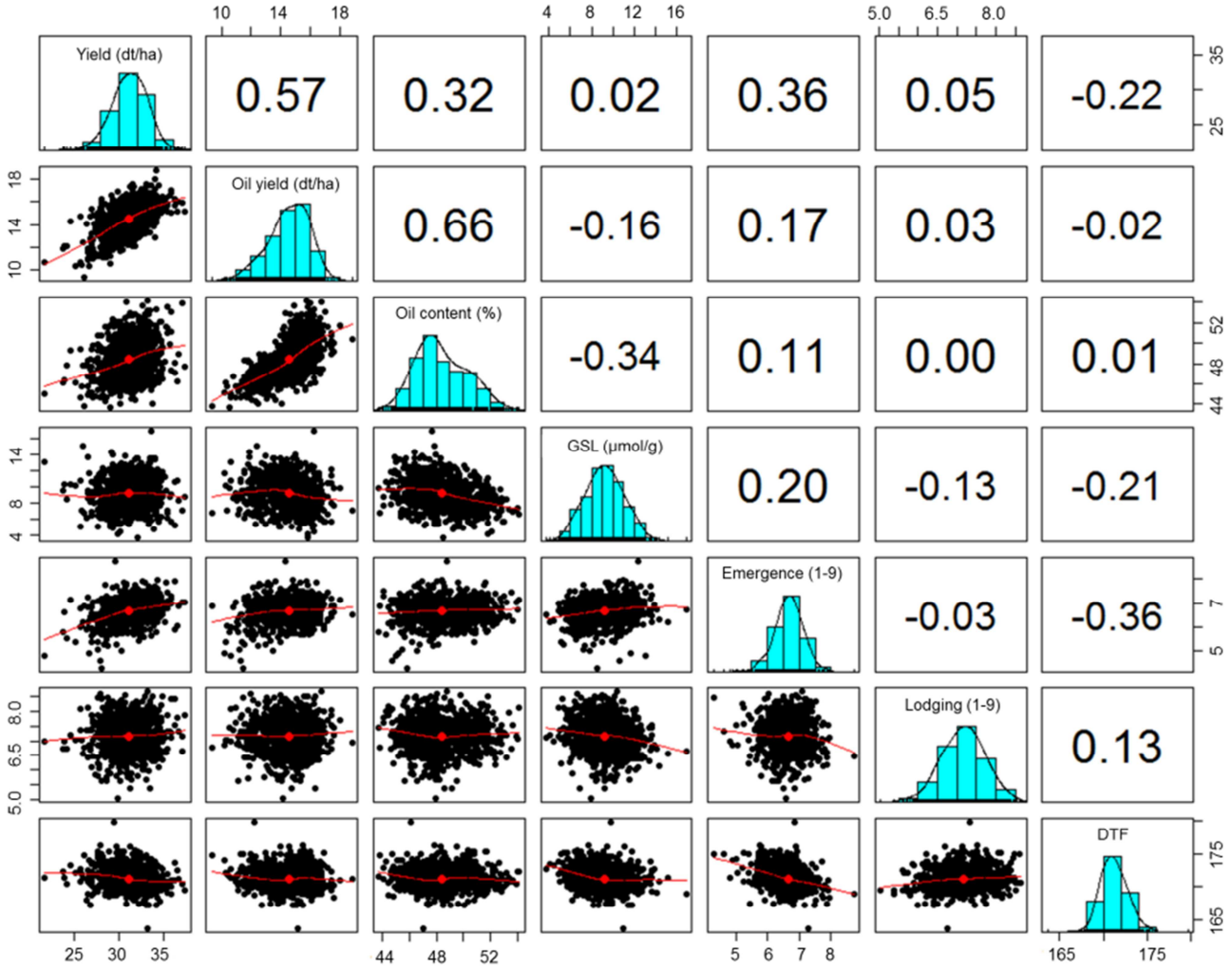

Observed distribution
(f)
Frequency
0
100
200
300
4
5
6
7
8
9
Lodging resistance (Good=9)


Normal Q-Q Plot
Observed Lodging resistance (Good=9)
(m)
Observed distribution
4
6
8
10
12
14
16
-3
-2
-1
0
1
2
3
Theoritical distribution



Supplementary figures (A) : Trait distribution: a-g) Histograms and h-n) Q-Q plots of best linear unbiased estimates (BLUEs) for a,h) seed yield (dt/ha), b,i) oil yield (dt/ha), c,j) seed oil content (%), d,k) seed glucosinolate content (GSL; µmol/g), e,l) emergence (visual observation scale 1-9; good=9), f,m) lodging resistance (visual observation scale 1-9; good=9) and g,n) days to onset of flowering (DTF) in field trials from 8 independent locations.

Appendix I. Supplementary figure. B: The Pearson's correlation coefficients between all the seven traits



Appendix II:

Table 1: Average prediction accuracies (r_{GPA}) and standard errors (SE) for seed yield (dt/ha), oil yield (dt/ha), seed oil content (%), seed glucosinolate content (GSL; µmol/g), seedling emergence (visual observation scale 1-9; good=9), lodging resistance (visual observation scale 1-9; good=9) and days to onset of flowering (DTF) derived from 500 iterations of cross-validation across the cluster 1(C1).

Traits	Seed yield (dt/ha)	Oil yield (dt/ha)	Seed oil content (%)	GSL (µmol/g)	Seedling emergence (good=9)	Lodging resistance (good=9)	DTF
r_{GPA}±SE	0.47±0.003a	0.49±0.004	0.68±0.002	0.47±0.003	0.49±0.003	0.2±0.004	0.59±0.003

[a]Approximate standard errors (SE) attached

Table 2: Average prediction accuracies (r_{GPA}) and standard errors (SE) for seed yield (dt/ha), oil yield (dt/ha), seed oil content (%), seed glucosinolate content (GSL; µmol/g), seedling emergence (visual observation scale 1-9; good=9), lodging resistance (visual observation scale 1-9; good=9) and days to onset of flowering (DTF) derived from 500 iterations of cross-validation across the cluster 2(C2).

Traits	Seed yield (dt/ha)	Oil yield (dt/ha)	Seed oil content (%)	GSL (µmol/g)	Seedling emergence (good=9)	Lodging resistance (good=9)	DTF
r_{GPA}±SE	0.30±0.003a	0.69±0.003	0.59±0.004	0.65±0.004	0.16±0.006	0.49±0.005	0.47±0.002

[a]Approximate standard errors (SE) attached

List of Abbreviations and Symbols

ALF1 : Aberrant Lateral Root Formation

BLAST : Basic local alignment tool

BLR : Bayesian Linear Regression

BLUE : Best linear unbiased estimates

BLUP : Best linear unbiased prediction

bp : base pair

DNA : Deoxyribonucleic acid

DTF : Days to onset of flowering

FAO : Food and Agricultural Organisation

FLC : FLOWERING LOCUS C gene

FRI : FRIGIDA gene

GBLUP : Genomic best linear unbiased prediction

GCA : General combining ability

GS : Genomic selection

GEBV : Genomic estimated breeding values

GSL : Glucosinolate

GWAS : Genome-wide association study

IBD : Identity by descent

IBS : Identity by state

LASSO : Least absolute shrinkage selection operator

LD : Linkage disequilibrium

MAF : Minor allele frequency

MAS : Marker assisted selection

MSL : Male sterility *Lembke*

N_e : Effective population size

NJ : Neighbour joining

PCA : Principal component analysis

QTL : Quantitative trait loci

Q-Q : Quantile-quantile

REML : Restricted maximum likelihood

RHD2 : Root Hair Defective 2

RKHS : reproductive kernel Hilbert spaces

RR-BLUP : Ridge regression best linear unbiased prediction

RRM : Realized relationship matrix

r_{GPA} : Genomic prediction accuracy

r^2 : Coefficient of linkage disequilibrium

SE : standard error

SNP : Single nucleotide polymorphism

TAIR : The *Arabidopsis* information resource

TASSEL : Trait Analysis by aSSociation, Evolution and Linkage

TP : Training population

VP : Validation population

WP : Whole population

σ^2_ε : error variance

σ^2_g : genotypic variance

Declaration (Erklärung)

I declare that the dissertation here submitted is entirely my own work, written without any illegitimate help by any third party and solely with materials as indicated in the dissertation. I have indicated in the text where I have already published sources, either word for word or in substance, and where I have made statements based on oral information given to me. At all time during the investigations carried out by me and described in the dissertation, I have followed the principles of good scientific practice as defined in the “Statutes of the Justus Liebig University Gießen for the Safeguarding of Good Scientific Practice”.

Habib Jan

Gießen,

November 3, 2015

Acknowledgments

I would like to extend my immense sense of gratitude to my worthy supervisor Prof. Dr. Rod Snowdon for his continuous supervision, support, guidance, encouragement and invaluably constructive criticism throughout the course of my PhD research. You have been a tremendous mentor for me. I would like to thank you for giving me this opportunity and for allowing me to grow as a research scientist.

I would like to express my appreciation and thanks to my second supervisor Prof. Dr. Richard Nichols (QMUL) for his sincere and valuable help and guidance during my placement in London. You have been extremely cooperative, kind and patient during my data analysis. Thanks to all my colleagues in *INTERCROSSING* for the great time we spent together across Europe during our various training sessions.

I would like to take this opportunity to thank all the members of the Plant Breeding Department, Giessen for their help and support. I am thankful to all the Postdoc members of our department, i.e. Birgit Samans, Christian Obermeier, Wubishet Bekele, Benjamin Wittkop, Annaliese Mason, Anna Stein and Sven Gottwald for the informal helpful discussions during the course of my research. Thanks to my colleagues Sarah Schießl, Sarah Hatzig, Kidist Kibret, Andreas Stahl, Christian Werner, and Steffen Windpassinger etc. who helped me in one way or the other. I am especially grateful to Christian Werner, who helped me translating my thesis summary into German.

It has been a great time, sharing my lab with my colleagues Lunwen Qian, Stefanie Lück, Kai Voss-Fels, and Roman Gäbelein. Thank you all for your help and support. Thanks to Frau Sabine Schomber, Frau Ulla Riedmeier and Frau Ingeborg Scholz (now retired), our department admin staff, for your help, support and politeness. I would like to thank technical assistants, especially Stavros Tzigos, for their help during my work in the lab and greenhouse.

I offer my thanks to Prof. Dr. Wolfgang Friedt (*Professor Emeritus*) for his continuous encouragement and for being a source of inspiration to me. Thanks to Prof. Dr. Matthias Frisch for his useful discussion during the analysis. I would also like to thank all the committee members who participated in my defense.

I would like to thank my parents, brothers, sisters and other relatives and friends back home for their moral support and prayers for me during my stay abroad. I acknowledge and appreciate the role of my wife for being extremely supportive during my research work.

Sufficient financing is always required for a successful venture. I appreciate Prof. Nichols, his team and all the principal investigators (PIs) for launching and executing *Marie Curie INTERCROSSING* consortium under which I completed my PhD research. I am thankful to our commercial partner *NPZ* Lembke, Germany and German Research Foundation (*DFG*) for the experimental materials and financial support during my lab work.

Habib Jan

Giessen

Der Lebenslauf wurde aus der elektronischen Version der Arbeit entfernt.

The curriculum vitae was removed from the electronic version of the paper.