Phenotypic and genetic diversity in cauliflower genebank accessions

University of Hohenheim

Faculty of Agriculture

Institute of Plant Breeding, Seed Science and Population Genetics (350)

Department of Crop Biodiversity and Breeding Informatics (350b)

Eltohamy Ali Ahmed Yousef

Cuvillier Verlag Göttingen
Internationaler wissenschaftlicher Fachverlag

Phenotypic and genetic diversity in cauliflower genebank accessions

Phenotypic and genetic diversity in cauliflower genebank accessions

Dissertation
submitted in fulfillment of the requirements for the degree of "Doktor der Agrarwissenschaften"
(Dr. sc. Agr.)

Hohenheim University

Faculty of Agriculture

Institute of Plant Breeding, Seed Science and Population Genetics (350)

Department of Crop Biodiversity and Breeding Informatics (350b)

Eltohamy Ali Ahmed Yousef

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der
Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet
über http://dnb.d-nb.de abrufbar.
1. Aufl. - Göttingen : Cuvillier, 2015
Zugl.: Hohenheim, Univ., Diss., 2015

D100

Gedruckt mit Unterstützung des Deutschen Akademischen Austauschdienstes

Nonnenstieg 8, 37075 Göttingen
Telefon: 0551-54724-0
Telefax: 0551-54724-21
www.cuvillier.de

1. Auflage, 2015
Gedruckt auf umweltfreundlichem, säurefreiem Papier aus nachhaltiger Forstwirtschaft.

ISBN 978-3-7369-9016-6
eISBN 978-3-7369-8016-7

University of Hohenheim
Faculty of Agricultural Sciences
Institute of Plant Breeding, Seed Science and Population Genetics (350)
Department of Crop Biodiversity and Breeding Informatics (350 b)
Prof. Dr. Karl Schmid

Phenotypic and genetic diversity in cauliflower genebank accessions

Dissertation
submitted in fulfillment of the requirements for the degree of "Doktor der Agrarwissenschaften"
(Dr. sc. Agr. / Ph. D. in Agricultural Sciences)

to the
Faculty of Agricultural Sciences at the University of Hohenheim

Presented by:

Eltohamy Ali Ahmed Yousef

from Kafrelsheikh, Egypt

Stuttgart-Hohenheim
2015

This thesis was accepted as a doctoral dissertation in fulfillment of the requirements for the degree “Doktor der Agrarwissenschaften, Dr.sc.agr.” by the Faculty of Agricultural Sciences at the University of Hohenheim on 21th April 2015

Date of oral examination: 19th May 2015

Examination committee

Chair of the committee	Prof. Dr. Harald Grethe
Examiner and Reviewer:	Prof. Dr. Karl Schmid
Co-reviewer	Prof. Dr. Heiko Becker
Examiner:	Prof. Dr. Jens Wünsche
Examiner:	Prof. Dr. Simone Graeffe-Honninger

Gedruckt mit Unterstützung des Deutschen Akademischen Austauschdienstes

Table of contents

List of tables

List of figures

List of abbreviations

G x E	Genotype-by-environment interaction
GBS	Genotyping by sequencing
SNP	Single nucleotide polymorphism
GWAS	Genome-wide association study
NGS	Next generation sequencing
AFLP	Amplified fragment-length polymorphism
IPK	Leibniz-Institut fürPflanzengenetik und Kulturpflanzenforschung
USDA	United States Department of Agriculture
LD	Linkage disequilibrium
F_{ST}	Genetic differentiation
BLUP	Best linear unbiased predictors
RRBLUP	Random regression best linear unbiased prediction
EMMAX	Efficient mixed-model association
MLMM	Multi locus mixed model
GS	Genomic selection
MAS	Marker-assisted selection
GLM	Generalized linear model
FDR	False discovery rate
GEBV	Genomic breeding value
QTL	Quantitative trait loci
Bayes B	Bayesian B

1 General introduction

1.1 Cauliflower importance

Cauliflower (*Brassica oleracea* var. *botrytis*) is a cool season crop and a member of the *Brassicaceae* family. It is thought that cauliflower originated over 2,000 years ago in the Mediterranean and Asia Minor region and the oldest written record of it dates back to 6th century BC. European writers mention cauliflower in Turkey and Egypt in the 16th century (http://aggie-horticulture.tamu.edu/archives/parsons/publications/vegetabletravelers/broccoli.html).

Cauliflower can contribute positively to human health because of its high glucosinolates content (Kushad et al. 1999; Schonhof et al. 2004), which can be converted by plant enzymes to other compounds, such as isothiocyanates and indole-3-carbinol, these having potential anticancer properties. Several epidemiological studies suggested that consumption of cruciferous crops, including cauliflower, can significantly reduce the risk of different types of cancer (Kirsh et al. 2007; Tang et al. 2008; Lee et al. 2008). Cauliflower and broccoli are cultivated on at least 1,204,257 hectares worldwide, with an annual production of over 21,266,789 tons (FAO, 2012).

1.2 Ex situ conservation

At the beginning of the last century, the need to conserve genetic diversity in crop species was recognized and also different conservation techniques were developed: *ex situ* and *in situ* (Maxted et al. 1997). Briefly, with *in situ* conservation the genetic material is preserved and maintained in their habitats, whereas *ex situ* conservation preserves and maintains genetic material outside their habitats, such as in zoos, botanical gardens and genebanks (Kasso and Balakrishnan 2013). At the present time, genebanks are the most common and essential means of *ex situ* conservation. Worldwide, genebanks are considered important reservoirs for natural genetic variation that originated from historical genetic events such as responses to environmental stresses as well as from selection through crop domestication. There are 7.4 million *ex situ* plant germplasm samples, consisting of breeding materials, wild plant species, modern cultivars, landraces, hybrids, and old cultivars conserved in world genebanks (FAO, 2010). Nevertheless, only a very small proportion of these genebank materials has been used. Effective exploitation of such *ex situ* genetic materials is therefore important in overcoming the problems associated with the narrow genetic basis of modern cultivars (Abdurakhmonov and Abdukarimov 2008). Moreover, understanding and assessing the genetic diversity existing in the

germplasm of a crop species is essential for genetic resource organizations such as genebanks. It could help in making decisions about what, where and how genetic materials should be conserved, as well as in developing and improving protocols for regeneration of germplasm (Rao and Hodgkin 2002), which in turn could help to avoid the negative effects of *ex situ* conservation on the conserved materials, such as loss of genetic diversity and inbreeding depression (*Hagenblad et al. 2012*; Brütting et al. 2013).

1.3 Genetic diversity in crop plants

Genetic diversity is the key to adaptability of populations to environmental changes. It therefore has a very essential role in the evolving and survival of populations over time. In the last century, the general trend of agriculture was to develop and use improved cultivars, which tended to be highly uniform (Rao and Hodkin 2002; Fernie et al. 2006). Consequently, this led to a reduction of genetic diversity in crop species. Nowadays, this low genetic diversity is one of the big challenges that face plant breeder with regard to sustaining and improving crop productivity (Zamir 2001). Without a broad base of heterogeneous genetic materials, it is impossible for plant breeders to produce new cultivars that meet the different needs of farmers (high productivity, high adaptation to specific growing conditions, resistance to biotic and abiotic stresses) and consumers (specific quality requirements).

1.4 Genetic diversity in cauliflower

Cauliflower cultivars exhibited high similarity and very low genetic diversity (Zhao et al. 2014; Tonguc and Griffith 2004), which hinders modern breeders from producing new cauliflower varieties with high yield and specific qualities. Therefore, an efficient assessment of the genetic diversity existing in cauliflower germplasm could help to broaden the genetic basis of cauliflower. Practically, it could provide valuable information for several applications in cauliflower breeding, like in other crop species, such as utilization of heterosis, selection of parental lines and legal protection of germplasm. Different types of molecular markers were employed to quantify the genetic diversity level in cauliflower (Astarini et al. 2006; Truong et al. 2012; Izzah et al. 2013; Zhao et al. 2014). However, due to limitations in former marker-based genotyping approaches and high similarity among cauliflower genotypes, several studies reported that development of high polymorphic marker systems, such as new sequence based

methods, are needed to facilitate a differentiation of cauliflower genotypes (Tonug and Griffiths 2004; Zhao et al. 2014).

1.5 Genotyping by sequencing

Thanks to advances in next-generation sequencing (NGS) technologies, several promising approaches have emerged that aim to combine the discovery, sequencing and genotyping of thousands of high-quality markers simultaneously (Stapley et al. 2010). One of these promising methods is genotyping by sequencing (GBS; Elshire et al. 2011; Poland et al. 2012a). The key point of GBS is the reduction of genome complexity, which can be achieved by digesting genomic DNA of different samples with a restriction enzyme. The main steps of GBS can be summarized as follows: digestion of genomic DNA with one or two restriction enzyme(s), ligation of barcode adapters, pooling the fragments, PCR-based amplification, and sequencing of the amplified pools (Figure 1).

GBS has been shown to be able to generate tens of thousands to hundreds of thousands of molecular markers at low cost in several crop species, such as maize (Elshire et al. 2011), barley and wheat (Poland et al. 2012a), soybean (Sonah et al. 2013) and oat (Huang et al. 2014). Several studies show that GBS can be used efficiently in several applications and research questions in genetics and breeding studies. For instance, GBS was successfully used for exploring genetic diversity and population structure in different crop species, such as maize, switch grasses and oilseed rape (Romay et al. 2013; Lu et al. 2013; Fu et al. 2014). Also, GBS was applied successfully to study genome-wide association (GWAS) for different traits in different crops (Morris et al. 2013; Sonah et al. 2014, Bastien et al. 2014) and several studies demonstrated the usability of GBS data to perform genomic prediction analysis (Poland et al. 2012b; Crossa et al. 2013; Jarquin et al. 2014). In addition, it was successfully performed to rapidly, efficiently and economically identify the alleles at the soybean maturity gene E3 (Tardivel et al. 2014). Therefore, GBS is a powerful tool that could effectively exploit the large reservoir of *ex situ* genetic materials conserved in genebanks (FAO 2010).

Despite all the mentioned advantages of GBS, one major drawback in its use is the large amount of incomplete single nucleotide polymorphism (SNP) data, with up to 90% of missing observations (Elshire et al. 2011; Poland et al. 2012a), which could make the different genetic

analyses difficult and less trustworthy (Fu et al. 2014). Sequencing with high coverage, by using more lanes on the sequencer or by reducing the multiplexing per lane, was suggested to overcome this problem (Poland and Rife 2012). However, this increases the cost, causing GBS to lose one of its advantages, i.e. cost efficiency. Therefore, imputation of missing values (Poland and Rife 2012; Romay et al. 2013) was introduced as a possible solution to overcome high missing values associated with GBS data. Several softwares have been successfully developed, such as random forest (*Breiman 2001*), fastPHASE (Scheets and Stephens; 2006) and BEAGLE (Browning and Browning 2007) to recover the missing values. But there is a debate among scientists about the choice of imputation method and whether using imputation improves results compared to simply selecting SNPs without or with low rates of missing data (Poland et al. 2012b; Rutkoski et al. 2013; Fu 2014). Therefore, it will be informative to assess the accuracy of various genetic analyses with respect to complete, incomplete and imputed GBS data.

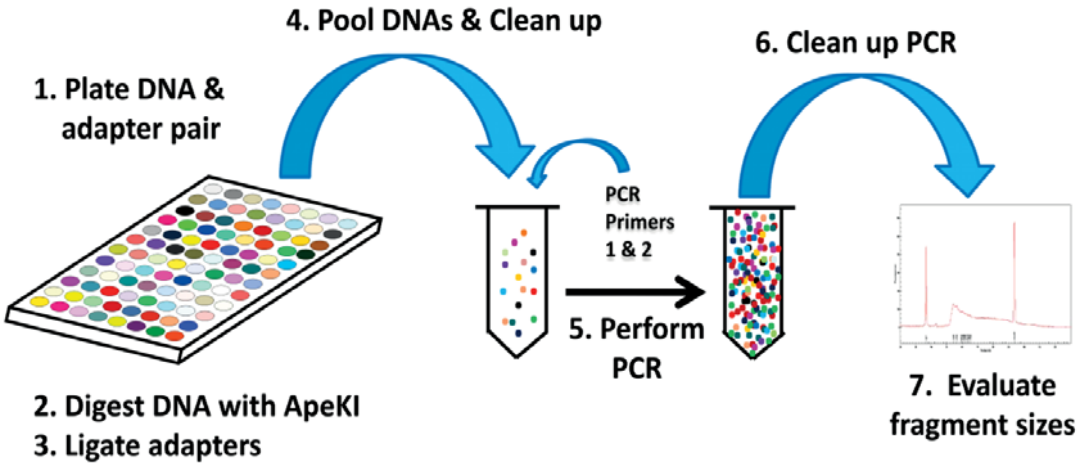

Figure 1.1 Basic scheme of used protocol for performing GBS (see Elshire et al. 2011)

1.6 Organic breeding in cauliflower

Nowadays, organic agriculture is gaining public interest, scientific attention and support by many governments, due to its positive effects on human and environmental health. However, organic agriculture is characterized by low output compared to conventional agriculture. Several empirical comparisons reported that yields under organic cultivation were lower than under conventional cultivation (Seufert et al. 2012). The low productivity of organic agriculture may be a consequence of low chemical inputs such as fertilizers and pesticides into the organic system. However, it might also be due to the fact that organic farmers depend on conventional varieties, which were bred and selected under conventional practices, including high inputs of artificial chemicals. Lammerts van Bueren et al. (2011) noted that more than 95% of the varieties used in organic farming have been bred under conventional practices. In the same regard, Banziger and Cooper (2001) reported that cultivars developed through conventional strategies may be not adapted to the side-effects of organic farming, such as low chemical inputs, or they lack the traits which allow for optimal production under organic farming. Consequently, breeding crop varieties specifically for organic cultivation is very important because these varieties are expected to realize their full high-yielding and high stability potential under organic cultivation.

Several studies reported a difference in performance between organic and conventional farming in wheat cultivars and they observed that direct selection in organic cultivation resulted in higher yields (5-31%) than indirect selection in conventional cultivation (Kirk et al. 2012; Murphy et al. 2007; Reid et al. 2009). This indicates that direct selection in the organic system could result in a significant advantage and could help in producing varieties specifically for organic cultivation (Kirk et al. 2012; Reid et al. 2009). On the other hand, some studies reported that selection should be performed under optimum conditions to avoid reduced heritability due to the greater environmental variance component under organic conditions and they mentioned that selection was similar between organic and conventional farming for some traits (Cerccarelli 1996; Wolf et al. 2008). Therefore, it will be informative to obtain information on whether direct or indirect selection under conventional conditions is preferable in cauliflower breeding programs for organic cultivation.

In contrast to new commercial varieties, which are genetically highly homogeneous (Almekinders and Elings 2001; Fernie et al. 2006), old varieties and landraces may have

agronomic value to organic farmers, as they were developed before synthetic inputs were available, or due to their genetic heterogeneity (Finckh 2008; Dawson et al. 2011). Thus, these varieties may be able to evolve specific adaptations to unpredictable and harsh environmental conditions under organic farming. Most of these genotypes have been preserved and stored in *ex situ* genebanks for some time now. Consequently, they could not interact with environmental changes and more recent high-synthetic-input agricultural practices and may have preserved useful genetic variation for low-input cultivation. One aim of this study is therefore to investigate the variability in genetically diverse cauliflower accessions from two *ex situ* genebanks: the United States Department of Agriculture (USDA) in the USA and the Leibniz-Institut für Pflanzengenetik und Kulturpflanzenforschung (IPK) in Germany, to identify genotypes that are better adapted to organic farming or that may be suitable as starting material for a breeding program focusing on organic cauliflower agriculture.

1.7 Genome wide association study (GWAS)

The major goal of genetic mapping approaches is to identify inherited genetic markers that are located close to genes controlling the complex quantitative traits (Abdurakhmonov and Abdukarimov 2008). The widely used methods to dissect quantitative traits are linkage mapping and association mapping (Flint-Garcia et al. 2005). Both linkage and association mapping methods use the linkage disequilibrium (LD; which is defined as the non-random association between alleles at different loci) between genes controlling a trait and closely linked markers. Linkage mapping is performed using segregating bi-parental populations and consequently captures only a small proportion of genetic variability, while association mapping utilizes the higher number of historical genetic events, such as selection, recombination, mutation and migration, that have occurred throughout the evolutionary history of mapping population (Flint-Garcia et al. 2003; Nordborg and Weigel 2008). Therefore, association mapping provides the opportunity to identify quantitative trait loci (QTL) with high mapping resolution as well as less research effort than the linkage mapping approach. With the dramatic changes in sequencing technologies and the decreasing costs of sequencing, association mapping has become a powerful and promising approach for understanding the genetic basis of different complex traits across a wide range of crop species, such as maize, wheat and rapeseed (Li et al. 2013; Edae et al. 2014; Li et al. 2014; Cai et al. 2014).

In *Brassica oleracea,* several studies have been performed to localize QTL for quantitative traits such as inflorescence (curd)-related traits, plant morphological traits and quality traits (Lan and Paterson 2000; 2001; Gu et al. 2008; Walley et al. 2012; Brown et al. 2014). All of them were linkage mapping studies using F2 or F3 generations. So far, no whole GWAS has been performed in *Brassica oleracea*, although it has been shown that it can dissect the genetic architecture of different complex traits in *Brassica napus,* such as disease resistance, seed oil content and quality, seed weight and quality, seed glucosinolate content and morphological traits (Jestin et al. 2011; Zou et al. 2010; Rezaeizad et al. 2011; Li et al. 2014; Hasan et al. 2008; Cai et al. 2014). Therefore, GWAS has a good potential to detect the genetic basis of complex traits successfully in *Brassica oleraceae* in general, and specifically in cauliflower.

Despite the well documented advantages of the association mapping approach, one of its major limitations is the existence of subpopulations in the mapping population, which could lead to false marker-trait associations if the trait under consideration is correlated with the subpopulation structure (Wright and Gaut 2005). Several statistical models, which differed in their way of accounting for population structure in the mapping population were introduced to overcome this limitation. Another aim of this study is therefore to test the potential of different statistical models that account differently for population structure and kinship in a very diverse collection of cauliflower genebank accessions, in order to identify the optimum model for association mapping analysis in cauliflower breeding programs.

1.8 Genomic selection

In contrast to association mapping, which aims to identify specific loci that affect the trait under consideration, genomic selection (GS) assumes that every locus in the genome contributes to the trait (Meuwissen et al. 2001). GS is a form of marker-assisted selection (MAS; Goddard and Hayes 2007). However, MAS is efficient only for traits controlled by low number of loci with large effects (Moreau et al. 1998). As a result, MAS might still miss a large proportion of genetic variation caused by several loci with small effects. This additional variation can be captured through the use of genome-wide SNPs implemented in GS (Bao et al. 2014). In brief, genomic selection consists of two steps: firstly, a training population with genotypic and phenotypic information is used to estimate marker effects. This step is called the training phase. Secondly, calculation of the genomic estimated breeding values (GEBVs) using only genotypic data of

breeding materials. This step is called the prediction phase. Thus GEBVs can be used directly in breeding programs to select individuals without the need for phenotypic data (Figure 2). This allows breeders to select the best genotypes based on predictions rather than observations, which increases genetic gains by shortening the time needed for the breeding cycle and reduces the costs of the field trials (Schaeffer 2006; König et al. 2009).

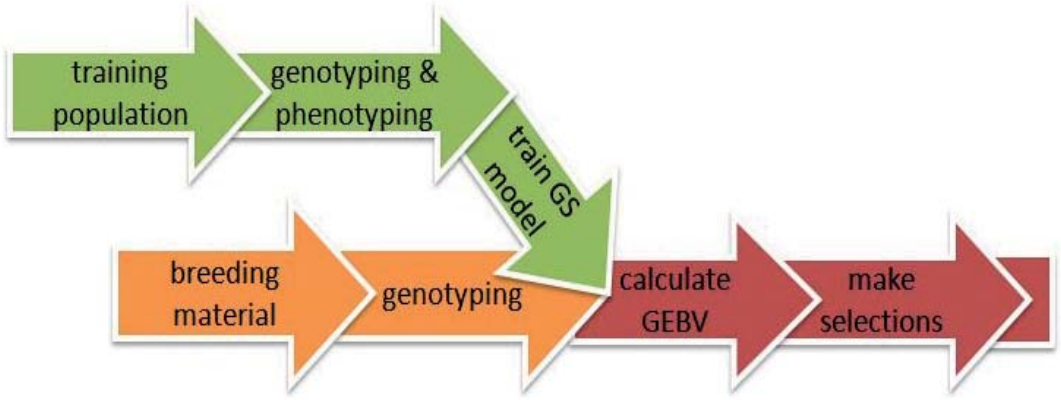

Figure 1.2 The basic scheme for GS, starting from the training population and selection candidates continuing through to genomic estimated breeding value (GEBV)–based selection (modified from Heffner et al. 2009).

Given the high-density marker panels, GS has become a superior approach using molecular markers for selection of complex traits in breeding programs (Lorenz et al. 2011). Currently, there are a large number of GS studies of different crop species, such as maize (Riedelsheimer et al. 2012a,b), barley (Heslot et al. 2012), wheat (Poland et al. 2012b) and soybean (Jarquin et al 2014). However, knowledge of GS in *Brassicaceae* is still limited, particularly in *Brassica oleracea*. Therefore, it would be desirable to evaluate the accuracy of GS for some important yield traits in cauliflower.

Different statistical models were successfully employed to perform genomic selection in plant and animal breeding. However, among these models the random regression best linear unbiased prediction (RRBLUP) and BayesianB (BayesB) are considered to be the major approaches for

performing GS because they show good performance (Meuwissen et al. 2001). These two models vary considerably in their assumptions about marker effects. For instance, the theory underlying RRBLUP is that all marker effects are normally distributed and all markers have the same variance. Bayes B assumes the variance of markers to equal zero with probability π, and the complement with probability (1–π) follows an inverse X^2 distribution, with v degree of freedom and scale parameter S (reviewed by Resende et al. 2012). Daetwyler et al. (2010) reported that performance of each method relies on the genetic architecture controlling the studied trait. Bayesian approaches are recommended for traits that are affected by a few QTL with large effects, whereas RRBLUP approaches are preferred for traits that are affected by several QTL with small effects (Hayes et al. 2009; VanRaden et al. 2009; Daetwyler et al. 2010). Therefore, it will be informative to test the potential of different statistical models to identify the optimum one for performing the genomic prediction analyses in cauliflower breeding programs.

1.9 Objectives:

The overall goal of this research thesis was to study the phenotypic and genotypic diversity as well as to perform association mapping and genomic prediction in a large number of cauliflower genebank accessions. In particular, the objectives were:

1. To quantify the extent of genotype × environment interaction (GxE) that influences some yield and maturity traits under organic and conventional conditions in cauliflower;
2. To obtain information whether direct or indirect selection under conventional conditions for organic cultivation is preferable;
3. To examine the population structure and the genetic diversity in genebank accessions of cauliflower using GBS;
4. To examine the efficiency of GBS in detection of genetic diversity and population structure in a large number of cauliflower genebank accessions;
5. To identify chromosomal regions affecting curd-related traits using genome-wide association mapping;
6. To quantify the ability of genomic prediction using GBS data with curd-related traits in cauliflower;
7. To study the effect of GBS data imputation on genetic diversity, association mapping and genomic prediction results;

8. To identify the optimum model for performing the association mapping and genomic prediction analyses in cauliflower breeding programs.

2 Evaluation of cauliflower genebank accessions under organic and conventional cultivation in Southern Germany

Eltohamy A. Yousef[+1,2], Christian Lampei[+1] and Karl J. Schmid[*1]

[1]Department of Crop Biodiversity and Breeding Informatics (350b), University of Hohenheim, Fruwirthstraße 21, D - 70599 Stuttgart, Germany

[2]Department of Horticulture, Faculty of Agriculture, University of Suez Canal, Ismailia (41522), Egypt.

* Corresponding author: karl.schmid@uni-hohenheim.de

+ These authors contributed equally to this work

This paper is published in Euphytica 201:389-400

2.1 Abstract

In recent years, public attention increased towards products from organic farming due to their presumed higher quality and health benefits. Frequently, organic farming is characterized by lower yields than conventional farming. One reason may be the use of varieties that were bred for conventional cultivation and are not adapted to organic farming. This raises the question if high yielding varieties differ in their performance under different cultivation methods allowing the selection of varieties with superior performance in organic cultivation. To answer this question and to identify suitable genotypes we evaluated a collection of 178 cauliflower genebank accessions under organic and conventional farming conditions. Two traits (curd width and time to budding) were evaluated for mean and stability. We observed a significant genotype × cultivation method interaction because genotypes differed in their performance between cultivation methods. Of the two traits investigated, curd width showed a lower heritability (H^2_{org}= 0.26, H^2_{conv}= 0.37) and low genotypic correlation between organic and conventional systems, compared to days to budding that show high heritability (H^2_{org}= 0.86, H^2_{conv}= 0.87) and a high correlation between the two farming systems. Our results demonstrate that the selection for curd width should be preferably conducted under organic conditions, whereas selection for number of days can be carried out under organic or conventional conditions. The evaluation of genotypes at both environments identified genotypes that may be used as parental lines for breeding under organic conditions.

Keywords Cauliflower, Organic farming, Genotype-by-environment interaction, Heritability, Stability

2.2 Introduction

Cauliflower (*Brassica oleracea* ssp. *botrytis)* is a major horticultural crop with an annual production of over 18,164,958 tons (FAO 2010). One reason for the popularity of cauliflower is the high glucosinolate content, which presumably has anti-cancerogenic properties (Lee et al. 2008; Krish et al. 2007; Tang et al. 2008). Therefore, cauliflower enjoys a high reputation among proponents of healthy nutrition and sustainable food production. A central component of the latter is the cultivation system, in particular conventional *versus* organic agriculture. Despite the reduced negative impact of organic farming on ecosystem services (e.g. soil conservation, quality of water resource, species diversity) and the growing demand of consumers for organic products, as well as the support by many of European governments (Willer and Kilcher 2010), organic farming shares only a small portion (5.3 million hectares) of the total utilized agricultural land (3%) in Europe, Germany had the biggest area of organic crop production (850,000 ha), covering 16% of the total organic area of Europe (FiBL and SOEL 2010). The reason is probably a lower productivity of organic cultivation (Trewavas 2004), which may lead to a reduced income for organic compared to conventional farmers and is one of the major reasons for deregistration form organic farming (Koesling et al. 2012). Consequently, scientists argue about the perspective of organic farming for providing food for the growing human society (Tilman et al. 2002; de Ponti et al. 2012).

The extent of yield reduction in organic cultivation was investigated in numerous studies. Different studies found that yields under organic conditions were typically lower than under conventional cultivation, but the extent of yield reduction differed among crops (de Ponti et al. 2012; Seufurt et al. 2012). For cauliflower, florets weight, yield and curd diameter showed a significant decrease in organic farming (about 25%, 20% and 15%) compared to conventional farming (Lo Scalzo et al. 2008 and Maggio et al. 2013). The yield reduction under organic cultivation may be a direct consequence of reduced inputs in fertilizer and pesticides, but is also influenced by the fact that the varieties used had been bred and selected under conventional farming. Lammerts van Bueren et al. (2011) noted that more than 95% of the varieties used in organic food production have been bred under conventional conditions. Such varieties might be maladapted to side effects of organic farming like an increased herbivore pressure or reduced availability of inorganic nitrogen.

Several researchers began to search for more diverse varieties such as old varieties that were grown before the widespread use of chemical inputs to conduct mass selection within these varieties or mix varieties to let natural selection create populations that are adapted to particular environments or agricultural practices (Dawson et al. 2013, Serpolay et al. 2011). It is frequently assumed that such varieties evolve specific adaptations to organic conditions because their genetic heterogeneity may buffer crop responses to unpredictable and harsh conditions (Finckh 2008). Currently, diverse varieties such as landraces, old varieties and mixtures are of greater relevance in organic agriculture than new commercial varieties because most new commercial varieties are genetically highly homogeneous (Almekinders and Jongerden 2002). Moreover, old varieties and landraces may have agronomic and quality traits of interest to organic farmers as they were developed before synthetic inputs were available (Dawson et al. 2011). Most of these genotypes are now stored in *ex-situ* genebanks, and these genotypes could not co-evolve with changing environmental conditions and agricultural practices and may have preserved useful genetic variation for low-input cultivation.

Banziger and Cooper (2001) pointed out that cultivars developed by current plant breeding usually are not adapted to low-input conditions. Consequently, cultivars that were explicitly bred and adapted to organic farming should better realize their full potential as high-yielding alternatives to conventional farming. Direct selection in organic cultivation resulted in higher yields (5-31%) than indirect selection in conventional cultivation suggesting that direct selection of genotypes for organic farming may be advantageous over indirect selection in conventional conditions (Kirk et al. 2012; Murphy et al. 2007; Reid et al. 2009). However, some studies concluded that selection for organic farming may take place under conventional conditions (Kokare et al. 2014; Wolfe et. 2008).

In addition to yield, yield performance is an important trait in organic farming because the low-impact philosophy of organic cultivation implies that the local environment has a stronger effect on yield than in conventional cultivation (Backes and Ostergard 2008). Yield stability can be defined as a reaction to environmental change which depends on unpredictable variance components (Kang 2002). Genotype × environment interaction (G×E) analysis allows to characterize the response of genotypes to changing environments and to determine the best genotypes for specific target environments or those that have a good performance for a wide

range of environments (Mohammadi and Amri 2009). Significant G×E interactions for several agronomic traits were found in cauliflower (Crisp 1977; Crisp and Kesvan 1978; Kesavan et al. 1976; Tharuk 2006), and Renaud et al. (2010) reported some evidence for GxE interactions in organic and conventional conditions for broccoli. However, little is known about the specific requirements of *Brassica* crops for an adaptation to organic cultivation (Myers et al. 2012).

The goal of the present study was to use information on G×E interactions from genetically diverse cauliflower accessions to allow the selection of suitable genotypes for further improvement to organic cultivation. We did not restrict the analysis to a narrow collection of varieties commonly used in a geographic region such as Southern Germany, where the field trials were carried out, because material from other regions may be better adapted to organic cultivation. Instead, we used a wide selection of genotypes from the USDA (USA) and the IPK (Germany) genebanks. The objectives of this study were: (1) to quantify the extent of GxE interactions that influences cauliflower yield (curd width) and number of days under organic and conventional conditions, (2) to study the stability of genotypes for curd width and number of days, (3) to identify genotypes that are better adapted to organic farming, and (4) to obtain information whether direct or indirect selection under conventional conditions for organic cultivation is preferable.

2.3 Materials and Methods

2.3.1 Plant materials and experimental stations

A total of 200 cauliflower accessions were obtained from the United States Department of Agriculture (USDA) genebank and the German genebank in Gatersleben (IPK) (Supplementary Table S2.1). The accessions were randomly selected and represent genotypes from over 28 countries. They had different biological status (cultivars, landraces, hybrid, breeding materials and unverified material). Accessions were evaluated by two cultivation methods (organic and conventional) for three growing seasons (June 2011, April 2012 and August 2012), in accordance with the optimal time for cauliflower production in Germany. Field trials were carried out at two experimental stations in Stuttgart, Germany: Heidfeldhof (conventional farm; Latitude: 48.71509 Longitude: 9.18992) and Kleinhohenheim (organic farm; Latitude 48.73797 Longitude: 9.20078), which has been run as organic farm since 1994 and is Demeter and Bioland

certified. The two farms lay within 3 km distance from each other and share the same clay soil (Soil analysis of each field in each farm is provided in Supplementary Table S2.2). Within each farm a different field site was chosen for each cultivation season which were up to 700 m apart from each other.

2.3.2 Agricultural practices

Seeds were cultivated in organic medium (KKS Bio-potgrond; Recipe-No: 025) and normal medium (Euflor Aussaaterde) for one month in the greenhouse to produce the transplants for organic and conventional cultivation, respectively. At the organic farm (Kleinhohenheim), sheep manure (240 dt/ha) was added before soil plowing. Also, 500 kg/ha of Bioilsa 11 (N, P_2O_5, K_2O: 11%, 1.2%, 0.5%; BIOFA Bio-Farming-System) were added once per cultivation during soil preparation before the transplantation of the seedlings. At the conventional farm (Heidfeldhof), no organic manure was added. Instead, 200 kg/ha of calcium-ammonium-nitrate (total N 27%, NO_3 50%, NH_3 50%) was added to the field two times (20 and 40 days after transplantation), adding up to a total of 110 kg/ha of total nitrogen. On both farms, irrigation was mainly rain-fed but occasionally additional irrigation was necessary (automated sprinkler) to prevent plants from suffering drought stress. In these cases, the experiments at both farms were irrigated with the same amount of water. At the organic farm, plants were covered for 40 days with a fine mesh net directly after planting to protect them from flea beetle *(Phyllotreta striolata)*. In the season April 2012, plants were first covered with a frost protection fleece (RANTAI 48, Rudolf Schachtrupp KG Gmbh & Co) for 20 days, which was then replaced with a net for another 20 days. At the conventional farm, no protective net was used, but similar to Kleinhohenheim, plants were also covered in the April 2012 season by the frost protection fleece for 20 days.

For plant protection, plants at the organic farm were treated once with Spruzit insecticide (Neudorff Natural Gardening) in the June 2011 season against flea beetle because it was very abundant in this season. On the conventional farm, plants were treated two times each season with two different insecticides. During the first days after transplanting, plants were sprayed with Biscaya (Bayer CropScience) and 15 days later with Plenum 50 W (Syngenta).

2.3.3 Phenotypic measurements

The field experiment was designed as a randomized complete block design (RCBD) with two replications in each location (6 plants per plot for each genotype). Two traits were measured: curd width because of its direct effects on yield (Jindal and Thakur 2003) and maturity, because early maturity reduces the chance of insect herbivory, fungal infections or other negative effect on yield. In addition, genotypes with a shorter time to maturity increase the income per land unit by increasing the potential of use per year. At maturity, five random plants per plot were evaluated for curd width (cm), each curd was cut into two equal parts and curd width was measured with a ruler according to Lan and Paterson (2000). For the trait number of days from planting to budding (the appearance of the first floral bud), the first five budded plants per plot were evaluated. Both traits were measured every three days.

2.3.4 Data analysis

In our data analysis we used two types of linear mixed-effects models. In the first model we fitted all variables except the block in environment which was regarded random as fixed effects to evaluate their contribution to the explained variation and their significance. In the contrary, the second model contained only random effects to estimate variance components. In the genotype-by-environment interaction literature the factors genotype and environment are often considered fixed effects (Gauch, 2006). However, modeling them as random effects is also justified as they can be considered a random sample of a larger population of genotypes or possible target environments (Atlin et al. 2000).

Linear mixed-effects models were fitted with restricted maximum likelihood (REML) using the *lme* function in the R package *nlme* (version 3.1-113, Pinheiro et al. 2013). The model was adjusted for variance heterogeneity by weighting the variance estimate for each location and growing season (using the *varIdent* and *varComb* functions) to account for heteroscedasticity with the following model equation:

$$y_{ijk} = \mu + S_i + M_j + G_k + B_l + SM_{ij} + SG_{ik} + MG_{jk} + SMG_{ijk} + \varepsilon_{ijkl} \qquad \varepsilon_{ijkl} \sim N\left(0, \sigma^2_{jk}\right)$$

where μ is the overall mean, S_i is the effect of the growing season I, M_j is the effect of the cultivation method j, G_k is the effect of the genotype k, B_l the effect of block l and SM_{ij}, SG_{ik} and

MG_{jk} are the two way interactions and SMG_{ijk} is the three way interaction. The only factor regarded random was block in cultivation method and growing season. The residuals ε_{ijkl} were modeled with a different variance for each cultivation method and growing season. Significance of factors was evaluated with a Wald-F-Test on the full model.

To further evaluate the effect of growing seasons the *glht* function from the R package *multcomp* (version 1.3-3, Hothorn et al. 2014) was used to calculate multiple comparison as *post hoc* comparison after a mixed model with the fixed factor environment (consisting of cultivation method and growing season) and the random factor genotype. Comparisons were corrected for false discovery rate with the Benjamini and Hochberg (1995) method.

To estimate the variance components associated with each of the two traits in each cultivation system, a mixed effects model was fitted with restricted maximum likelihood (REML) for each of the two cultivation methods with the *lmer* function from the R package *lme4* (version 1.0-5., Bates et al. 2013) with the model equation:

$$y_{ijk} = \mu + S_i + B(S)_{j(i)} + G_k + SG_{ik} + \varepsilon_{kj(i)}$$

with μ being the overall mean, B_i the effect of block i, S_j the effect of the j^{th} growing season and G_k the effect of the genotype k and SG_{jk} as the genotype x growing season interaction. In this model, the fixed effects part contained only the overall intercept whereas all other effects were considered random. Following Atlin et al. (2000), no constraints were applied to the terms of the model.

Variance components of this model were used to calculate broad sense heritability for each trait separately in each cultivation method according to Atlin et al. (2000):

$$H = \frac{\sigma_G^2}{\sigma_G^2 + \frac{\sigma_{GS}^2}{s} + \frac{\sigma_E^2}{sr}}$$

where σ^2_G is the variance explained by genotype, σ^2_{GS} is the variance due to genotype × growing season interaction, σ^2_E is the error variance. Further, s is the number of growing seasons and r the number of replicates in each growing season. Standard errors for the heritability estimates were

obtained from 1000 parametric bootstrap samples by applying the *simulate* and *refit* functions of the *lme4* package.

From the same model, best linear unbiased predictors (BLUPs) were extracted as estimated by the random effect genotype and used to estimate the genetic correlation (r_G) among the cultivation methods for both traits by the Pearson correlation coefficient.

To assess the impact of differences in heritability between cultivation systems against a low genetic correlation between cultivation systems, the response to direct (r_d) and indirect (r_{id}) selection and their ratio (relative efficiency of indirect selection) was calculated as according to Messmer et al. (2012):

$$\frac{r_{id(org)}}{r_{d(org)}} = \frac{H_{con} \times r_{G(con/org)}}{H_{org}}$$

where H is the heritability and r_G is the genetic correlation between two cultivation systems.

Genotype stability was calculated as the variance of growing season genotype means within cultivation method which is equivalent to type 4 stability as described in Lin and Binns (1991). This variability of genotype means over time provides the advantage over other stability parameters that it is easily interpreted and possesses higher heritability (Lin and Binns 1991). All statistical analysis reported here were conducted with R programming language (R Development Core Team 2013)

2.3.5 Selection according mean and stability

The selection of genotypes under the two cultivation methods was conducted simultaneously for yield and stability with equal weighting by standardizing BLUPs and stability measures, adding and sorting the values to obtain a ranking variable which allowed to select for largest (curd) or smallest (days to budding) values.

2.4 Results

One hundred seventy eight accessions were used in the analysis because twenty two accessions were excluded from the initial set of accessions as they entered the flowering stage without producing a curd in at least one cultivation season. The genotype mean of curd width ranged

from 11.5 - 15.3 cm with a global average of 13.7 cm (Table S2.3 and S2.4). For days to budding, the genotype mean ranged from 45 to 111 days with a global average of 76.2 days (Table S2.5 and S2.6). The global means of both traits were 13.69 cm and 75.15 days under conventional cultivation and 13.64 cm and 77.15 days under organic cultivation. Among the three cultivation seasons, the August 2012 season showed a 2 cm reduced curd width on average and time to budding was on average 15 days longer (Figure 2.1).

2.4.1 Evaluation of fixed effects

The linear mixed-effects models for curd width and days to budding showed that in both traits growing season was the factor that contributed most to the explained variation (Table 2.1). After growing season also genotype was a strong factor, however, genotype was more important in days to budding ($F_{(177/8203)}$ = 241.2) than in curd width ($F_{(177/8203)}$=7.85). Further, in both traits the three way interaction and the two way interactions genotype × cultivation method and genotype × growing season were significant. However, the main effect cultivation method was the least contributing main effect in days to budding and was even not significant in curd width. The interaction cultivation method × growing season was not significant in curd width, but a strongly contributing and significant factor in days to budding (Table 2.1).

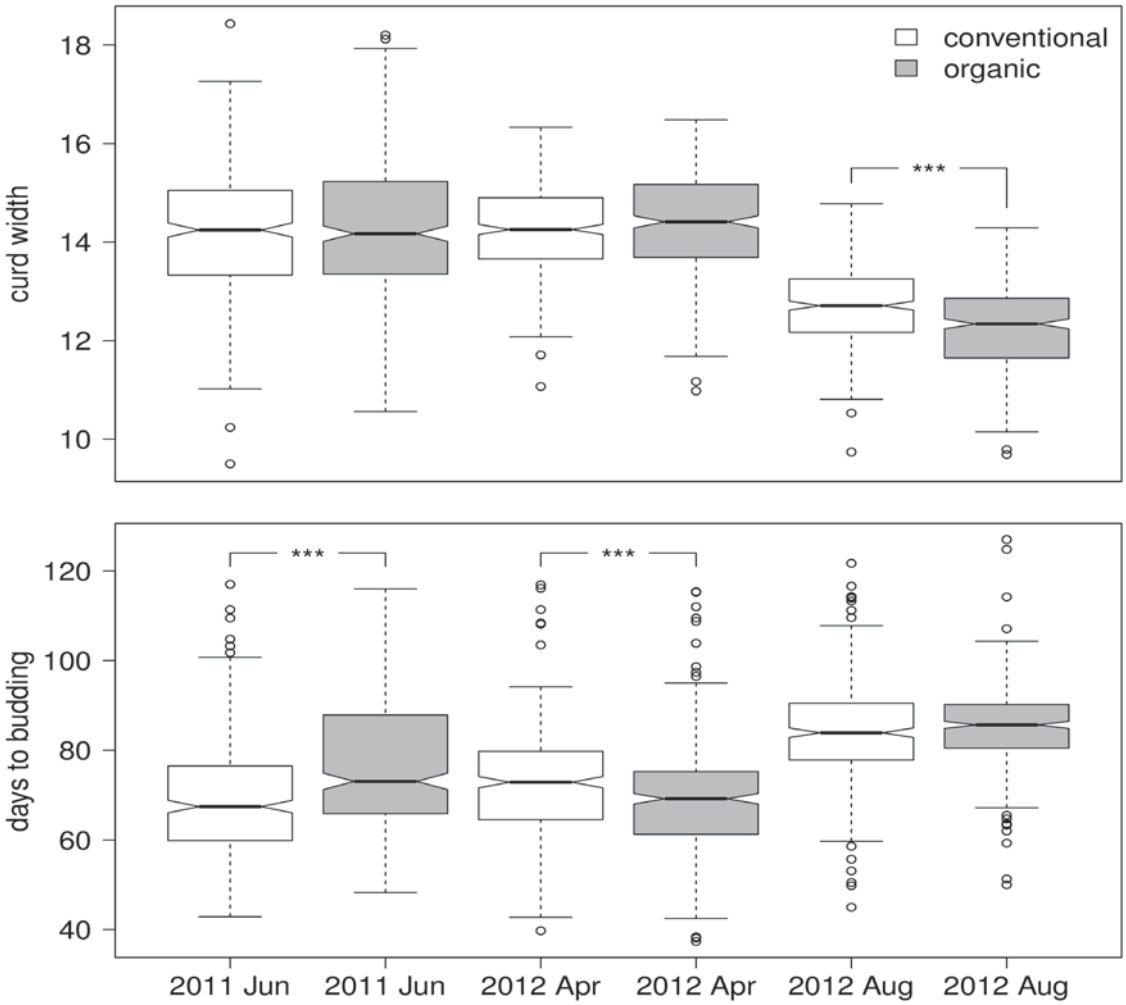

Figure 2.1 Boxplot graph for curd width (cm) and number days to budding under conventional and organic cultivation method over three cultivation times displaying the method × time interaction. Significant differences between cultivation methods as observed from multiple comparison tests are indicated with stars between the whiskers (***<0.001).

In such a large selection of genotypes it would be theoretically possible that the genotype × cultivation method and genotype × growing season would be driven by few outliers that were adapted to exotic environments. The distribution of the two traits in the two cultivation methods and three growing seasons (Supplementary Figure S2.1) indicates that this did not happen in the present experiment.

Table 2.1 Wald test-F test of fixed effects in a linear mixed-effect model fitted with REML for curd width and days to budding

Source of variation	**Curd width**		**Days to budding**	
	F-Value(DF/DF)	**P-Value**	**F-Value(DF/DF)**	**P-Value**
Cultivation method (M)	0.26(1/6)	0.63	33.1(1/6)	<0.01
Growing season (S)	149.56(1/6)	<0.001	838.4(1/6)	<0.001
Genotype (G)	7.85(177/9606)	<0.001	41.2(177/8203)	<0.001
MxS	4.21(2/6)	0.072	110.2(2/6)	<0.001
MxG	3.68(177/9606)	<0.001	9.8(177/8203)	<0.001
SxG	4.80(354/9606)	<0.001	24.9(354/8203)	<0.001

2.4.2 Genetic correlation, heritability and selection efficiency

The genetic correlation between the two cultivation methods (organic and conventional) was low for curd width 0.28 (95% CI: 0.14 and 0.41) and high for number of days to budding 0.94 (95% CI: 0.92 and 0.95). Also, heritability estimates differed strongly between the two traits. Days to budding showed a high heritability (organic: 0.86 ± 0.02, conventional: 0.87 ± 0.02) in both environments (Table 2.2). In contrast, curd width showed a low heritability and which tended to be lower in the organic cultivation, however this was not significant (organic: 0.26 ± 0.09, conventional: 0.37 ± 0.08). The difference among the cultivation methods was mainly due to a reduced variance component for genotype (Table 2.2). We further calculated the predicted relative response to direct (r_d) and indirect (r_{id}) selection for both traits. For days to budding, the difference between direct and indirect selection was marginal with $r_d = 9.64$ and $r_{id} = 9.15$. For curd width, the response to selection was very low in general but was twice as high for direct selection ($r_d = 0.099$) as for indirect selection ($r_{id} = 0.039$). The efficiency of indirect selection was $r_{id} / r_d = 0.95$ for days to budding but only $r_{id} / r_d = 0.39$ for curd width.

Table 2.2 Estimates of variance components and heritability for curd width and number of days to budding grown under conventional and organic management.

Variance components	Curd width		Days to budding	
	Conventional	**Organic**	**Conventional**	**Organic**
Genotype (G)	0.20	0.14	122.9	124.3
Growing season (S)	0.72	1.39	54.5	65.6
GxS	0.69	0.87	48.7	54.1
Residuals	0.62	0.64	8.7	9.2
Heritability (SE)	0.37 (0.08)	0.26 (0.09)	0.87 (0.02)	0.86 (0.02)

2.4.3 Selection of genotypes and stability analysis

The selection of the ten best genotypes for each trait and each cultivation method showed overlap between the cultivation methods for days to budding (five genotypes) and for curd width (four genotypes). Table 2.3 shows mean and stability for the top ten genotypes for curd width over three conventional and three organic environments. The corresponding values for all genotypes are displayed in Figure 2.2a, b and the Supplementary Table S2.3. Based on mean and stability, BRA 1346 (116) and PI 267722 (59) appear to be the most stable and high yielding genotypes for conventional cultivation, whereas PI 462225 (100) and BRA 2097 (199) are most suitable for organic cultivation. BRA 1406 (174) grew well in both cultivation methods.

Mean and stability of the top ten genotypes for number of days to budding over the three conventional and three organic growing seasons are provided in Table 2.4. An overview over the whole population and the selected accessions is provided in Figure 2.2c, d and the supplementary Table S2.4. PI 244663 (47) was the earliest flowering and most stable genotype in both cultivation methods. PI 277273 (69) and BRA 1750 (191) were the next in ranking under the conventional method, whereas PI 321001 (86) and PI 250127 (54) were stable and performing well under the organic cultivation method.

Table 2.3 Genotypic mean and stability for the top ten genotypes evaluated across three organic and conventional environments for curd width.

Organic					**Conventional**				
Genotype	**Number**	**Yield**	**Stability**	**Rank**	**Genotype**	**Number**	**Yield**	**Stability**	**Rank**
PI 462225	100	14.92	0.614	2.64	BRA 1346	116	15.02	0.203	2.89
BRA 1406	**174**	14.94	0.802	2.58	PI 267722	59	15.14	0.757	2.66
BRA 2097	199	15.00	1.491	2.34	BRA 1406	**174**	14.86	0.279	2.62
PI 269312	**62**	15.06	1.782	2.29	BRA 1481	186	14.68	0.448	2.26
BRA 290	200	15.05	1.843	2.26	BRA 1337	**107**	14.40	0.011	2.18
BRA 1386	**155**	14.45	0.549	2.02	BRA 1376	146	14.50	0.219	2.17
BRA 1480	185	14.54	0.977	1.95	BRA 1413	181	14.58	0.561	2.03
BRA 1370	140	14.18	0.018	1.90	BRA 1386	**155**	14.55	0.580	1.98
BRA 1337	**107**	14.15	0.066	1.83	PI 269312	**62**	14.25	0.014	1.97
BRA 1359	129	14.50	1.158	1.81	BRA 1334	104	15.01	1.492	1.97

Bold font indicates genotypes which were selected under both cultivation methods

The selection of genotypes with large curd and early budding and high stability yielded no genotype among the top ten with these attributes under both cultivation methods. Genotype 186 showed early flowering and large curd together with high stability in the conventional cultivation but in the organic cultivation it ranks only on place 23 in curd width and place 49 in days to budding. The variety that is first selected in both traits and both cultivation methods is genotype 69 (PI 277273) which ranks high in number of days under both conditions (Table 2.4) and on rank 16 in curd under organic cultivation and rank 21 under conventional cultivation (Supplementary Tables S2.3-S2.6). The variety is called “Early Patna” and was commonly used in India.

2.5 Discussion

Key characteristics of organic agriculture are a reduced control of pathogens, herbivores and weeds, as well as a lack of inorganic input. As a consequence, crops grow under harsher and more variable conditions than in conventional agriculture. Hence, varieties adapted to organic cultivation should not only show a good performance but also a high yield stability. In the present experiment, average curd width and time to budding did not differ strongly between

cultivation methods (Figure 2.1) suggesting that the genebank accessions tested were not very sensitive to the cultivation method, which is in contrast to the results of Maggio et al. (2013), who reported a significant decrease in curd diameter (about 15%) under organic cultivation compared to conventional cultivation. Yield reduction under organic cultivation was reported for other crops (de Ponti et al. 2012; Seufurt et al. 2012). However, Seufurt et al. (2012) further observed that the yield differences among farming systems varied strongly between crops. In our experiment, only in one cultivation time (Aug 2012) curd width was significantly reduced in the organic system. This season showed a reduction in curd width of about 2 cm and 15 days later budding compared to the other seasons and was the only one to show a difference among the farming systems. This finding suggests that organic cultivation may lead to yield reduction especially when the cultivation environment is of suboptimal quality. As the region of Filder plains, in which experiments were conducted, is well-known for cabbage production due to the fertile clay soil, the general suitability of the planting area for cauliflower cultivation may be also a reason for the low overall decrease of curd width under organic conditions.

Table 2.4 Genotypic mean and stability for the top ten genotypes evaluated across three organic and conventional environments for number of days to budding.

Organic					**Conventional**				
Genotype	**Number**	**Yield**	**Stability**	**Rank**	**Genotype**	**Number**	**Yield**	**Stability**	**Rank**
PI 244663	**47**	47.02	59.32	-3.21	PI 244663	**47**	44.14	26.45	-3.28
PI 321001	86	51.92	88.50	-2.49	PI 277273	**69**	56.78	4.37	-2.38
PI 250127	**54**	59.87	27.48	-2.47	BRA 1750	**191**	53.86	38.28	-2.37
BRA 1750	**191**	52.53	93.42	-2.38	BRA 1385	154	59.70	11.55	-2.08
PI 269315	64	64.60	3.52	-2.34	PI 244831	49	59.70	19.04	-2.02
PI 277273	**69**	58.44	58.00	-2.27	BRA 2057	197	58.32	38.21	-1.99
BRA 1388	157	66.73	21.14	-1.97	PI 343478	90	53.12	107.64	-1.90
PI 291993	**79**	57.70	96.00	-1.92	BRA 1481	186	58.34	55.12	-1.86
PI 267724	60	69.32	16.08	-1.81	PI 250127	**54**	60.05	37.08	-1.85
PI 320998	84	63.14	70.17	-1.75	PI 291993	**79**	57.22	69.23	-1.84

Bold font indicates genotypes which were selected under both cultivation methods

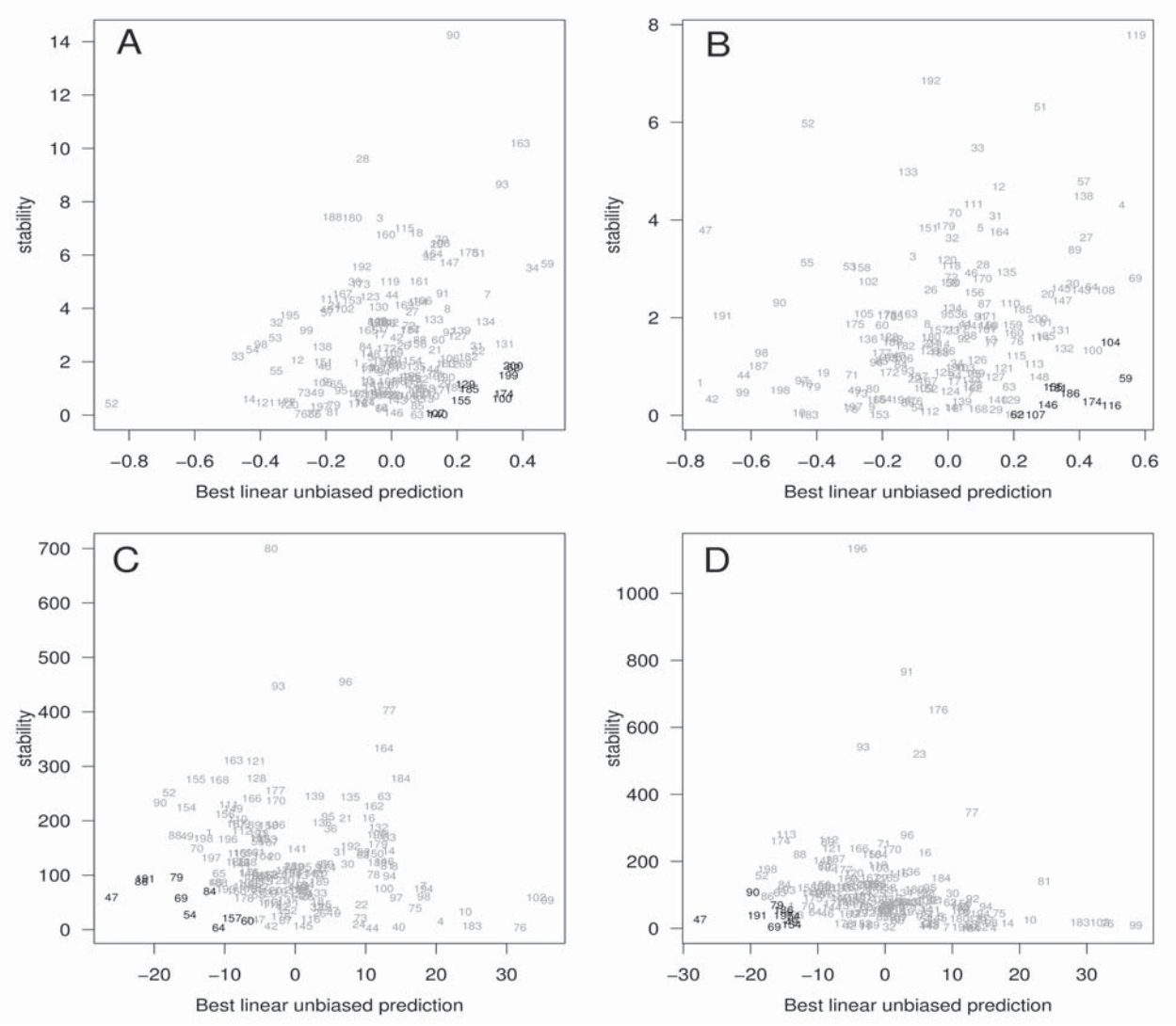

Figure 2.2 Selection scatterplots showing yield stability (T4) against the best unbiased predictors (BLUP) of genotypes yield. The selection plots a and b show stability and yield 0f 178 cauliflower accessions for curd width under organic (a) and conventional (b) cultivation. Accordingly, the selection plots c and d show stability and yield of 178 cauliflower accessions for days to budding. Selected accessions under each condition are shown in black.

For both traits, the variable explaining most of the variation is growing season. This is in agreement with the finding of Rendaud et al. (2010) who reported that performance of broccoli cultivars was influenced by cultivation time. Also, Elwan and Abd El-Hamed (2011) found a

strong effect of growing season on yield in broccoli. They further reported a strong interaction of growing season with genotype which was also found in our study. One reason for the different performance of cauliflower genotypes at different growing seasons may be temperature sensitivity during their development (Wiebe 1975

In both traits the effect of cultivation method was rather small. This was especially obvious in curd width. Part of the weak main effect of cultivation method in our study may be explained by the fact that the two farms (conventional and organic) were in close proximity to each other (less than 3 km) and share very similar soil and microclimatic conditions. On the other hand, this also suggests that in our experiment the effect of the cultivation method is rather unbiased, because the seasonal climate was the same for both sites. Significant genotype × cultivation method interaction indicated that some genotypes performed differently in both cultivation methods. Consequently, a test of whether direct selection should be preferred over indirect selection for organic cultivation is justified.

Most cauliflower varieties were bred under conventional breeding systems and consequently, using high performing varieties for an organic cultivation system may be regarded a case of indirect selection (Messmer et al. 2009). Indirect selection, however, is most efficient for traits with a high heritability that show a strong correlation between cultivation systems (Baenziger et al. 2011, Goldstein et al. 2012, Messmer et al. 2009). In our study, the two traits greatly differed in both heritability and genetic correlation between the two cultivation systems. Heritability was low in curd width but high in days to budding (Table 2.2), which is in agreement with previous studies that reported low heritability for curd width (Singh et al. 2010) and a high heritability for number of days to budding (Lan and Paterson, 2000). Similarly, the genetic correlation between cultivation systems was low in curd width but very high in days to budding, which also corresponds to earlier studies (Messmer et al. 2009; Przystalski et al. 2008). Furthermore, for curd width heritability tended to be lower in the organic environment, but this reduction was not significant. Due to an increased environmental variation in organic conditions, a reduction in heritability is not unexpected (Ceccarelli 1996), however in our experiment the effect of heritability on selection efficiency is of minor importance. Further, the estimate of the efficiency of indirect selection indicated that indirect selection may be only half as efficient as direct selection for curd width, which is mainly due to the low genetic correlation in this trait among

the cultivation methods. It must be noted though that heritability was generally low for curd width in this population, which may be due to a long selection history for large curds in cauliflower. Hence, it may be difficult to select for a larger curd in organic cultivation. Instead, it may be interesting to select for yield stability, especially considering that the variance component of growing season was twice as high in the organic system compared to the conventional. For days to budding, on the other hand, indirect selection from conventional farming may be nearly as efficient as direct selection. For this trait, a separate breeding program is likely more promising.

One aim of this multi-environment trial was to identify varieties that may be suitable as starting material for a breeding program focusing on organic agriculture. Among the top ten genotypes for curd width, four genotypes (Table 2.3) were selected in both cultivation systems. In days to budding, five genotypes were under the ten best in both cultivation systems (Table 2.4). In contrast to this result, in a selection based only on genotype trait means 7 out of 10 genotypes were chosen in both cultivation systems for days to budding and only 3 out of 10 were similar between cultivation systems for curd width. This finding demonstrates the effect of low heritability and low genotypic correlation on indirect selection. Furthermore, the difference between the two selection approaches suggests that also the heritability of stability is important when joint selection is performed. Furthermore, it also shows that despite our earlier result that indirect selection may be more efficient than the direct, the differences in ranking in the two cultivation systems make it rather unlikely to select the most well performing genotypes via indirect selection. A solution may be the strategy proposed by Löschenberger et al. (2008) to use conventional low input trials which exhibit high genetic correlations for indirect selection in early generations, followed by direct selection in later generations.

Interestingly, many of the best performing genotypes for both traits (under organic or conventional farming) came from locations distant to Germany suggesting the potential value of exotic germplasm to identify breeding material that may be better adapted to a specific cultivation system.

Since in the two farms in our study different fields were used for different seasons and the farms were situated in close proximity with similar in soil and weather conditions, we assume that the

cultivation system is the strongest environmental difference between the farms. However, this also means that our observations may be limited to the environment of fertile clay soil in Southern Germany. Further investigations including more locations are needed to validate the observed small effect of the cultivation method on the performance of cauliflower in general and the suitability of selected lines for organic cultivation in different soil and climate conditions.

2.6 Conclusion

The results of this study indicate that number of days to budding was mainly affected by genotypes and genotype accounted for the largest part of the explained variance while curd width was mainly affected by the growing season and its interaction with genotype. The cultivation method (organic and conventional) had no main effect on both traits which may be explained by the fact that the two farms were in close proximity to each other and share very similar soil types and microclimatic conditions. A significant genotype x cultivation method interactions indicated that some genotypes performed better under conventional conditions and some accessions under organic conditions. Based on the high genotypic correlations between two cultivation methods and high heritability for number of days to budding, the selection for this trait could be done in organic or conventional farming systems (direct or indirect selection). Conversely, for curd width, the selection should be done in an organic farming system (direct selection) because it has low heritability and a low genotypic correlation between the organic and the conventional farming system.

Acknowledgments

We express our thanks to Nayyef Al-Jaar and Elfadil Mukhtar Adam for their assistance with the recording and analysis of data. Also, we thank the Klasmann company for their offering of the organic media. This work was funded by a DAAD GERLS Fellowship to E.Y. and by an endowment of the Stifterverband der deutschen Wissenschaft to K. J. S.

2.7 References

Almekinders CJM, Elings A (2001) Collaboration of farmers and breeders: participatory crop improvement in perspective. Euphytica 122:425–438

Atlin, G., Baker, R., McRae, K., & Lu, X. (2000). Selection response in subdivided target regions. *Crop Science*, 40: 7–13.

Backes G, Ostergard H (2008) Molecular markers to exploit genotype–environment interactions of relevance in organic growing systems. Euphytica 169: 523-531

Baenziger PS, Ibrahim S, Little RS, Santra DK, Regassa T, Wang MY (2011) Structuring an efficient organic wheat breeding program. Sustainability 3: 1190–1205

Banziger M, Cooper M (2001) Breeding for low input conditions and consequences for participatory plant breeding: Examples from tropical maize and wheat. Euphytica 122:503–519.

Bates D, Maechler M, Bolker B, Walker S (2013) lme4: Linear mixed-effects models using Eigen and S4. R package version 1.0-5.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society. Series B 57:289-300

Ceccarelli S (1996) Adaptation to low high input cultivation. Euphytica 92:203–214

Crisp P (1977) Genotype x environment interactions in early winter cauliflowers in south-west Britian. J Hortic Sci 52: 357-366

Crisp P, Kesavan V (1978) Genotypic and environmental effects on the weight of the curds of autumn-maturing cauliflowers. J Agric Sci 90:11-17

Dawson JC, Murphy KM, Huggins DR, Jones S (2011) Evaluation of winter wheat breeding lines for traits related to nitrogen use under organic management. Org Agric 1:65–80

Dawson JC, Serpolay E, Giuliano S, Schermann N, Galic N, Berthellot J-F, Chesneau V (2013) Phenotypic diversity and evolution of farmer varieties of bread wheat on organic farms in EuropeGenetic Resources and Crop Evolution 60:145-163

de Ponti T, Rijk B, van Ittersum MK (2012) The crop yield gap between organic and conventional agriculture. Agric Syst 108: 1–9

Elwan MWM, Abd El-Hamed KE (2011) Influence of nitrogen form, growing season and sulfur fertilization on yield and the content of nitrate and vitamin C of broccoli. Sci Hortic 127: 181-187

FAO (2010) Food and Agriculture Organization of the United Nation. The Statistics Division, http://www.fao.org.

FiBL, SOEL (2010) European Organic Farming Statistics. http://www.organic-europe.net/europe_eu/statistics.

Finckh MR (2008) Integration of breeding and technology intodiversification strategies for disease control in modern agriculture. Eur J Plant Pathol 121:399 – 409

Gauch HG (2006) Statistical analysis of yield trails by AMMI and GGE. Crop Sci 46:1488 - 1500

Goldstein WA, Schmidt W, Burger H, Messmer M, Pollak LM, Smith ME, Goodman MM, Kutka FJ, Pratt RC (2012) Maize: breeding and field testing for organic farmers. In: Myers JR, Lammerts van Bueren ET (eds.) Organic Crop Breeding. Wiley-Blackwell, West Sussex, pp 175-188

Hothorn T, Bretz F, Westfall P (2008) Simultaneous inference in general parametric models. Biometrical Journal 50:346-363.

Jindal SK, Thakur JC (2003) Interrelationship of curd weight and other characters in November cauliflower. J of Res 40: 358-362

Kang MS (2002) Genotype-environment interaction: Progress and prospects. In: M.S. Kang (Ed.), Quantitative genetics, genomics and plant breeding, CAB International: Wallingford: England 221-243

Kesavan V, Crisp P, Gray AR, Dowker BD (1976) Genotypic and environmental effects on the maturity time of autumn cauliflowers. Theor Appl Genet 47:133-140

Kirk AP, Fox SL, Entz MH (2012) Comparison of organic and conventional selection environments for spring wheat. Plant Breed 131:687-694

Kirsh VA, Peters U, Mayne ST (2007) Prospective study of fruit and vegetable intake and risk of prostate cancer. J Natl Cancer Inst 99:1200-1209

Koesling M, Løes A, Flaten O, Kristensen NH, Hansen MW (2012) Farmers' reasons for deregistering from organic farming. Org Agric 2:103–116

Kokare A, Legzdine L, Beinarovica I, Mallie paard C, Niks RE, Lammerts van Bueren ET (2014) Performance of spring barely (*Hordeum vulgare*) varieties under organic and conventional conditions. Euphytica 1-15. doi:10.1007/s10681-014-1066-8

Lammerts van Bueren ET, Jones SS, Tamm L, Murphy KM, Myers JR, Leifert C, Messmer MM (2011) The need to breed crop varieties suitable for organic farming, using wheat, tomato, and broccoli as examples: a review. NJAS-Wageningen J Life Sci 58:193-205

Lan TH, Paterson AH (2000) Comparative mapping of quantitative trait loci sculpting the curd of *Brassica oleracea*. Genetics 155: 1927–1954

Lee SA, Fowke JH, Lu W, Ye C, Zheng Y, Gu K, Gu YT, Gao XO, Shu X, Zheng W (2008) Cruciferous vegetables, the GSTP1 Ile^{105}Val genetic polymorphism, and breast cancer risk. Am J clinnutr 87:753–760

Lin C, Binns M (1991) Genetic properties of four types of stability parameters. TAG 82:505-509

Lo Scalzo R, Iannoccari T, Genna A, Di Cesare LF, Viscardi D, Ferrari V, Campanelli G (2008) Organic vs. conventional field trials: the effect on cauliflower quality. In Proceedings of

the 16th IFOAM organic world congress, cultivate the future based on science, 2nd conference of the international society of organic agriculture research ISOFAR, Modena, Italy, ID code 11758

Löschenberger F, Fleck A, Grausgruber G, Hetzendorfer H, Hof G, Lafferty J, Marn M, Neumayer A, Pfaffinger G, Birschitzky J (2008) Breeding for organic agriculture the example of winter wheat in Austria. Euphytica 163:469–480

Maggio A, De Pascale S, Paradiso R, Babieri G (2013) Quality and nutritional value of vegetables from organic and conventional farming. Scientia Horticulture 164: 532-539

Messmer MM, Burger H, Schmidt W, Geiger HH (2009) Importance of appropriate selection environments for breeding maize adapted to organic farming systems. Tagung der Vereinigung der Pflanzenzücher und Saatgutkaufleute Österreichs. Pp. 49-51

Messmer M, Hildermann I, Thorup-Kristensen K, Rengel Z (2012) Nutrient management in organic farming and consequences for direct and indirect selection strategies.. In: Lammerts van Bueren ET, Myers JR (eds.) Organic Crop Breeding. Wiley-Blackwell, New York. pp. 15-32

Mohammadi R, Amri A (2009) Analysis of Genotype × Environment Interactions for Grain Yield in Durum Wheat. Crop Sci 49:1177– 1186

Murphy KM, Campbell KG, Lyon SR, Jones SS (2007) Evidence of varietal adaptation to organic farming systems. Field Crop Res 102:172-177

Myers JR, McKenzei L, Voorrips RE (2012) Brassica: breeding cole crops for organic Agriculture. In: Lammerts van Bueren ET, Myers JR (eds.) Organic Crop Breeding. Wiley-Blackwell, New York. pp. 251-262

Pinheiro J, Bates D, DebRoy S, Sarkar D, the R Core team (2013) *nlme*: Linear and Nonlinear Mixed Effects Models. R package version 3.1-113.

Przystalski M, Osman A, Thiemt EM et al (2008) Comparing the performance of cereal varieties in organic and non organic cropping systems in different European countries. Euphytica 163:417–433

R Development Core Team (2013) R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing. Vienna (Austria): R Foundation for Statistical Computing, Vienna

Reid T, Yang R-C, Salmon DF, Spaner D (2009) Should spring wheat breeding for organically managed systems be conducted on organically managed land? Euphytica 169:239-252

Renaud ENC, Lammerts van Bueren ET, Jiggins C, Maliepaard J, Paulo JA, Juvik JR, Myers JR (2010) Breeding for specific bioregions. a genotype by environment study of horticultural and nutritional traits integrating breeder and farmer priorities for organic broccoli cultivar improvement. In. Goldringer I. Dawson J, Rey F, Vettoretti A (eds) Breeding for resilience. A strategy for organic and low-input farming systems? EUCARPIA 2nd Conference of the Organic and Low-Input Agriculture Section. Paris, pp. 127-145

Serpolay E, Dawson JC, Chable V, Lammerts Van Bueren E, Osman A, Pino S, Silveri D, Goldringer I (2011) Diversity of different farmer and modern wheat varieties cultivated in contrasting organic farming conditions in Western Europe and implications for European seed and variety legislation. Organ Agric1:127–145

Seufert V, Ramankutty N, Foley JA (2012) Comparing the yields of organic and conventional agriculture. Nature 485:229-234

Singh G, Singh DK, Bhardwaj SB (2010) Variability studies on November maturity group of cauliflower (*Brassica olearaceaevar botrytis L.*). Pantrnagrar J Res 8:202-205

Tang L, Gary RZ, Guru K, Kirsten BM, Zhang Y, Ambrosone CB, McCann SE (2008) Consumption of raw cruciferous vegetables is inversely associated with bladder cancer risk. Cancer Epidemiol Biomark prev 17: 938–944

Tharuk BS (2006) Adaptability for yield in some mid-late and late group cauliflower (*Brassica olearaceaevar botrytis*) genotypes under the mid-hill conditions of Himachal Pradesh. Indian J AgrSci 76:37-40

Tilman D, Cassman KG, Matson PA, Naylor R, Polasky S (2002) Agricultural sustainability and intensive production practices. Nature 418:671–677

Trewavas A (2004) A critical assessment of organic farming and food assertions with particular respect to the UK and the potential environmental benefits of no-till agriculture. Crop Prot 23: 757–781

Wiebe HJ (1975) The morphological development of cauliflower and broccoli cultivars depending on temperature. Sci Hortic 3: 95-101

Willer H, Kilcher L (Eds.) (2010) The World of Organic Agriculture. Statistics and Emerging Trends 2010. IFOAM, Bonn and FiBL, Frick

Wolfe MS, Baresel JP, Desclaux D, Goldringer I, Hoad S, Kovacs G, Lo schenberger F, Miedaner T, Østergard H, Lammerts van Bueren ET (2008) Developments in breeding cereals for organic agriculture. Euphytica 163:323–346

2.8 Supplementary Materials

Table S2.2 Soil analysis of two farms

Kleinhohenheim (organic)

Over all	T1	T2	T3	Mean	Unit
pH-Wert (CaCl2-Suspension):	6.0	6.9	6.4	6.43	--
P2O5 (CAL-Extract VDLUFA):	12	17	17	15.28	mg/100g
K2O (CAL-Extract VDLUFA):	44	22	42	35.89	mg/100g
Mg (CAL-Extract VDLUFA):	18	16	15	16.22	mg/100g
Humus (Elementaranalyse):	4.2	2.6	2.6	3.12	%

Heidfieldhof (conventional)

Over all	T1	T2	T3	Mean	Unit
pH-Wert (CaCl2-Suspension):	6.7	5.9	6.8	6.47	--
P2O5 (CAL-Extract VDLUFA):	27	28	26	27.00	mg/100g
K2O (CAL-Extract VDLUFA):	26	49	27	34.00	mg/100g
Mg (CAL-Extract VDLUFA):	13	14	13	13.33	mg/100g
Humus (Elementaranalyse):	1.8	4.3	1.9	2.67	%

T1 and T2 and T3 represent growing season June 2011, April 2012 and August 2012, respectively.

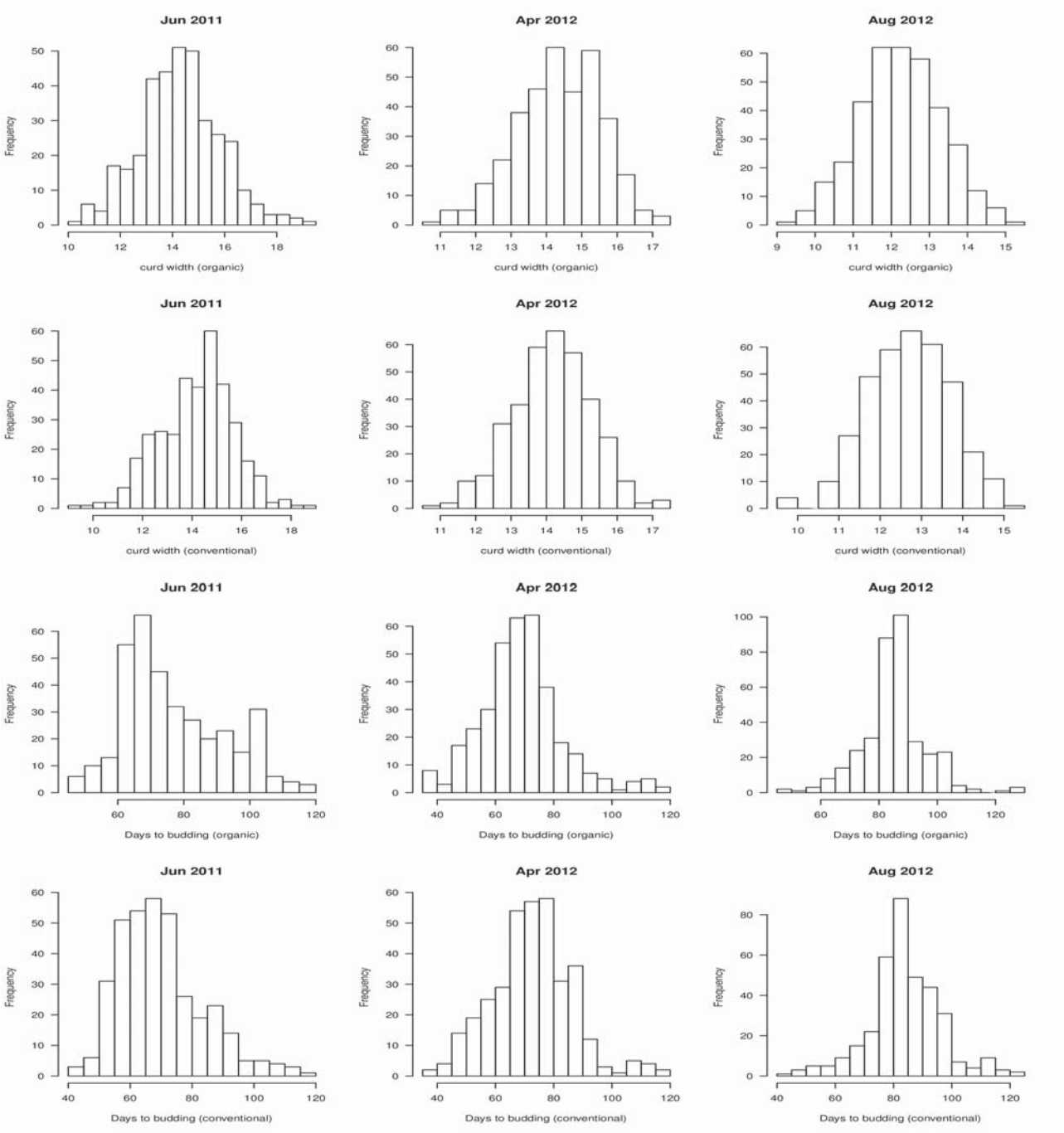

Figure S2.1 The distribution of each trait at two cultivation methods (organic and conventional) and three planting times (June 2011, April 2012 and August 2012)

3 Evidence for strong population structure caused by germplasm regeneration in *ex situ* genebank collections of cauliflower (*Brassica oleracea* var. *botrytis*)

Eltohamy A. A. Yousef,[1,2] Thomas Müller[1], Andreas Börner[3] and Karl J. Schmid[1]

[1] Department of Crop Biodiversity and Breeding Informatics, Faculty of Agriculture, University of Hohenheim, Stuttgart, Germany

[2] Department of Horticulture, Faculty of Agriculture, University of Suez Canal, Ismailia, Egypt

[3] Department of Genebank, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), OT Gatersleben, Germany

*Corresponding author

Email: karl.schmid@uni-hohenheim.de

This article is submitted to PloS ONE jornal.

3.1 Abstract

Cauliflower (*Brassica oleracea* var. *botrytis*) is an important vegetable crop and characterized by a low genetic diversity. We used genotyping-by-sequencing (GBS) to characterize the genetic diversity of 191 cauliflower accessions from the USDA and IPK genebanks. They originated from 20 different countries and represent about 50% of all accessions in both genebanks. The analysis of genetic diversity revealed that accessions did not cluster by country of origin but formed two major groups that represented the two genebanks. This finding was robust with respect to the clustering method that included principal component analysis, model-based clustering with STRUCTURE and neighbor-joining trees. Genetic diversity was higher in the USDA accessions, because of a higher proportion of genetically diverse landraces. The results indicate that seed regeneration procedures by genebanks have a strong effect on the structure of genetic diversity. Possible reasons are genetic drift, inbreeding as well as strong selection for adaptation to the regeneration conditions. Fst-based outlier tests of genetic differentiation identified only a small proportion (<1%) of SNPs that are highly differentiated between the two genebanks in response to selection, which suggests genetic drift as main cause of differentiation between genebanks. In summary, GBS is a useful method for characterizing genetic diversity in cauliflower and our results suggest that it may be useful to incorporate genotyping into seed regeneration to reduce the loss of genetic diversity by genetic drift.

3.2 Introduction

The extent and type of genetic variation present in the germplasm of a crop is an important component of efficient breeding programs, because it provides useful information for the broadening of breeding pools, the utilization of heterosis and the selection of parental lines. Also, this information helps breeders to narrow the search for new alleles at loci of interest and assists in the identification of markers linked to desirable traits for introgression into new varieties [1]. An assessment of genetic diversity is also essential for the organization, conservation and use of genetic resources to develop strategies for optimal germplasm collection, evaluation and conservation and to develop improved protocols for the regeneration of seeds [2].

Various research projects were initiated to exploit the genetic diversity of *ex situ* conserved genetic resources with the goal to increase the low genetic diversity of modern cultivars (e.g., [3]). *Ex situ* conserved populations are an important resource for the introgression of new and exotic genetic variation into breeding pools, but they experience a loss of genetic diversity, inbreeding depression (especially outcrossing crops), accumulation of deleterious alleles and adaptation to habitats used for regeneration [4-6]. These processes often result from the small population sizes of individual genebank accessions and may negatively affect the success of *ex situ* conservation after several cycles of *ex situ* regeneration [5,7-8].

Cauliflower (*Brassica oleracea var. botrytis)* is an important vegetable crop worldwide and considered to be an important component of a healthy diet because of a high content of glucosinolates that have anticancer properties [9-10]. Cauliflower and broccoli are currently cultivated worldwide on about 1.2 Mio hectares, with an annual production of over 21 Mio tons [11]. In previous studies, different marker types were used to analyze genetic diversity in cauliflower that included amplified polymorphic DNA (RAPD) [12], simple sequence repeat (SSR) [13-14] and inter-simple sequence repeat (ISSR) [4]. All of these marker systems have certain limitations [15]. Several studies reported a low genetic diversity for cauliflower [16-17, 14], and it was suggested that highly polymorphic marker systems are required to enable the differentiation of cauliflower genotypes.

In recent years, single nucleotide polymorphisms (SNP) emerged as a highly useful marker type for characterizing plant genetic resources (PGR) because they provide a higher quality and

quantity of number information than marker system like AFLPs and SSRs [18-19]. Among SNP genotyping methods, genotyping-by-sequencing (GBS) was developed as a cost effective alternative because it can be highly multiplexed. Tens of thousands of polymorphisms are identified in a single reaction without the need of a reference sequence [20-22]. In the context of PGR, GBS was used to characterize the genetic variation of maize, sorghum and switchgrass with respect to their known ancestral history and geographical origin [23-25]. In Brassicaceae, GBS was used to analyse genetic diversity in yellow mustard [26].

The objectives of this project were: (i) to assess the genetic diversity of genebank accessions of the USDA and IPK *ex situ* collections by genotyping randomly selected accessions with GBS; (ii) to investigate the ability of GBS to detect the genetic diversity and population structure in a large number of cauliflower accessions; and (iii) to study the effect of missing values and imputation on the results of genetic diversity and population structure. To achieve these objectives, we tested whether genetic diversity is affected by the geographic origin of cauliflower accessions or other factors and compared levels of genetic and phenotypic variation. We found different average levels of diversity among the accessions from the two genebanks and a strong genetic differentiation between the two genebanks that suggested that germplasm regeneration procedures in *ex situ* genebank collections of cauliflower have a strong effect on their genetic composition.

3.3 Material and Methods

3.3.1 Plant materials

A total of 191 cauliflower accessions were randomly selected and ordered from the genebanks of the United States Department of Agriculture (USDA), USA and IPK Gatersleben, Germany. The selection included traditional cultivars (85), breeding materials (4), hybrids (1), unverified genotypes (90), collector materials (4), commercial vegetable seeds (1) and landraces (6). The accessions represent 26 countries of origin and eleven accessions of unknown geographic origin. A plant individual of the wild type of *Brassica oleracea* was obtained from Heidelberg Botanic Garden and Herbarium (HEID), Germany. All accessions in this study, their origin and their biological status are listed in Table S3.1.

3.3.2 Field experiment and phenotypic measurements

All accessions were evaluated for six morphological traits at six environments (two cultivation methods: organic and conventional, and three growing seasons: June 2011 and April/August 2012). The field experiment was established in a randomized complete block design (RCBD) with two replicates in each environment. For more details about the phenotyping see Yousef et al. (2014) [27]. Five random plants per plot were evaluated every three days for the following traits: curd width (cm), cluster width (cm), number of branches, length of the apical meristem (cm), length of the nearest branch to apical meristem (cm) and number of days from planting to appearance of the floral buds. Morphological traits were measured according to Lan and Paterson [28]. The mean over locations and seasons was calculated.

3.3.3 DNA extraction

Genomic DNA was extracted from leaf tissue three weeks after sowing in the green house from a single individual of each genotype according to a standard CTAB protocol [29]. The quality and quantity of extracted DNA were checked with Nanodrop 2000c (Thermo scientific), Qubit 2.0 Fluorometer (Life Technologies) and 3% agarose gel. The final concentration of each DNA sample was adjusted to 100 ng/ml for DNA digestion.

3.3.4 Genotyping by sequencing (GBS)

GBS was carried out according to the protocol of Elshire et al. [20] with minor modifications as described below. The DNA was digested with the *Ape*K1 restriction enzyme. A total of 96 barcodes were used. Sixty-four barcodes were generated with the web tool provided by http://www.deenabio.com/services/gbs-adapters. Thirty-two barcodes were taken from Elshire et al. [20]. All barcodes have an even distribution in length (4-8 nucleotides) and nucleotide composition of nucleotides at each position (Table S3.2). The 192 genotypes were divided into two libraries, each consisting of 96 genotypes.

Before sequencing the GBS libraries, the distribution of fragment sizes were determined with an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA) to verify that adapter dimers are absent and DNA fragments range between 170–350 bp [20]. The two libraries were sequenced on two lanes of an Illumina HiSeq1000 at the Kompetenzzentrum Fluoreszente Bioanalytik (KFB), Regensburg, Germany to produce 100 bp long paired-end reads.

3.3.5 Sequence data analysis

Sequence reads were filtered for sequencing artifacts and low quality reads with custom Python scripts, bwa [30] and FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads mapping to the PhiX genome were identified and removed with bwa. All reads with ambiguous 'N' nucleotides and reads with low quality values were discarded. Remaining sequence reads were demultiplexed into separate files according to their barcodes. After removal of the barcode sequence and end-trimming, the reads had a length of 88 bp.

3.3.6 Read mapping and SNP calling

The pre-processed reads were aligned to the available genome of *Brassica oleracea sp. capitata*, a cabbage line 02-12 [31] with bwa. SNP calling was performed with SAMtools [32], bcfutils, vcfutils and custom python scripts. The .vcf file was parsed to filter out SNP positions with a coverage of at least 30, whereby at least ten reads had to confirm the variant nucleotide. Positions not fulfilling these criteria were marked and considered as missing data. A distance matrix was calculated using the SNP data as input (Note S1).

3.3.7 Analysis of population structure and genetic diversity

Population structure was investigated by various methods for comparison. One method was principal component analysis (PCA), as implemented in the R package *adegenet* [33]. Also, a PCA of six morphological traits was performed with the *prcomp* function in the *stats* R package [34]. The correlation between genetic and phenotypic distances based on the PCA was determined with a Mantel test [35] in the *ade4* R package [36]. A *t*-test was used to check if there are significant differences between the two seed sources (genebanks) for six traits. Principal coordinate analysis (PCoA) based on pairwise Fst values between genotypes calculated with R *adegenet* R package [33], was used for further analysis of population structure with the *ape* R package [37]. Discriminant analysis of principal components (DAPC), a multivariate method that combines PCA and discriminant analysis (DA) to identify clusters of genetically related individuals [38], was performed with the *dapc* function of the R package *adegenet* [33]. Briefly, DAPC depends on data transformation using PCA as a prior step to DA, which ensures that variables submitted to DA are perfectly uncorrelated, and that their number is less than that of analyzed individuals [38]. In addition, the genetic relationship among accessions was assessed

with a neighbor-joining tree (NJ tree) based on a pairwise distance matrix with *ape* R package [37].

Population structure was further inferred with STRUCTURE 2.3.4 [39]. Numbers of subpopulations were allowed to range from *K*=1-10 with ten runs for each *K* using the admixture model and assuming correlated allele frequencies [40], and a burn-in length of 50,000 followed by 50,000 iterations. The best number of subpopulations given the data was determined with the method of Evanno et al. [41]. Population structure was also inferred with ADMIXTURE [42]. The number of subpopulations analyzed ranged from *K*=1-10 and cross validation was used to estimate the optimum number of clusters *K* [43]. An Analysis of Molecular Variance (AMOVA) was carried out with ARLEQUIN v3.5.3.1 [44] and the extent of genetic differentiation (Fst) between two the genebanks was estimated with the *pairwise.fst* function in the R package *adegenet* [33]. Also, F_{ST} was calculated overall accessions of each genebank accessions separately [33].

For each group of accessions from USDA and IPK genebanks, we calculated the observed and expected heterozygosity, inbreeding coefficient (*F*) with *adegenet*. Also we computed the percentage of polymorphic loci (%P) and nucleotide diversity with the *R* package *pegas* [45].

3.3.8 Data imputation

To compare the effects of missing values and of genotype imputation on the population structure inference and genetic diversity estimates, we carried out the analyses with three data sets: 1) data without missing values, where all markers with missing values were excluded; 2) data with missing values, in which all markers with missing values were retained; and 3) imputed data, in which the missing values were imputed with fastPHASE [46].

3.3.9 Detection of outlier SNPs

Outlier tests for highly differentiated SNPs between the two populations (USDA and IPK) were based on a coalescent-based simulation method [47] implemented in LOSITAN [48]. This method identifies putative targets genes that differentiated in response to selection based on the distributions of expected heterozygosity and Fst values under an island model. LOSITAN was implemented in two steps: in the first step, an initial run with 50 000 simulations was performed

with all loci and using the mean Fst. After excluding a candidate subset of selected loci determined in the initial run, the distribution of neutral Fst values were calculated. We also used ARLEQUIN 3.5 [44] to detect outlier loci by accounting for the hierarchical genetic structure of the accessions that were obtained with the population inference methods. The outlier detection was based on 10,000 simulations under a hierarchical island model. For both methods, loci outside the 99 and 1 % confidence areas were identified as candidate genes affected by directional and balancing selection, respectively. A further test of differentiation by selection was conducted with Bayenv2 [49]. The analysis with Bayenv2 was also performed in two steps. First, the variance–covariance matrix for the groups of accessions was calculated. The variance–covariance matrix was used from the final run of the MCMC after 100,000 iterations. Second, we used variance–covariance matrix to control for the evolutionary history for the groups in the calculation of X^TX for each SNP by using 100,000 iterations of the MCMC.

Data availability

Raw sequence data have been submitted to Short Read Archive under accession number WILL BE PROVIDED IN PUBLISHED PAPER. Aggregated data (e.g., SNP calls) and analysis scripts are available on DataDryad under accession WILL BE PROVIDED IN PUBLISHED PAPER

3.4 Results

3.4.1 Patterns of genetic diversity

The total number of raw 100 bp sequence reads was 580,520,246. After filtering out 5.2% reads that mapped to the PhiX genome or were low quality reads, the number of reads ranged from only 80 to 7,136,388 per accession with an average of 2,369,313 reads. A total of 133,423,496 reads were mapped to the *B. oleracea* reference genome. The percentage of mapped reads per genotype against *B. oleracea* ranged from 14.49% to 34.68% with an overall average of 29% (Figure 3.1 and Table S3.3). Eighteen accessions had less than 300,000 reads and were excluded from further analysis. Based on the mapping to the *Brassica oleracea* reference genome, 120,693 SNPs were detected in the remaining 174 samples (120,693 SNPs with missing data, 1,444 SNPs without missing data in any of the accessions). The mean percentage of missing data across all genotypes was 42% and values ranged from 19% to 77% per genotype. The number of SNPs and percentage of missing data per genotype are shown in Table S3.4.

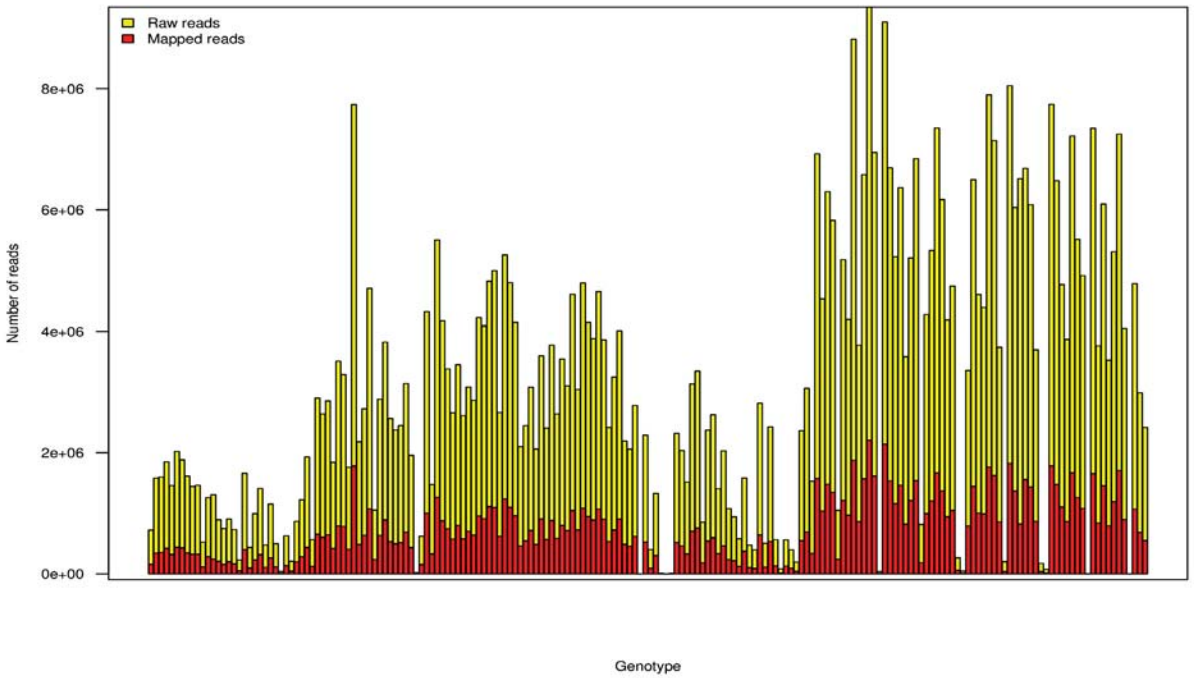

Figure 3.1 Distribution of the number of raw and mapped reads across 192 barcoded cauliflower genotypes.

3.4.2 Population structure of sample

The genetic structure of the whole collection (*n* =174) was analyzed using PCA, PCoA, DAPC, NJ tree, STRUCTURE and ADMIXTURE. Here, we present the results for the GBS data set consisting of 120,693 SNPs with missing data, but the same results were obtained with the other two data sets without missing data or imputed data, which are provided as supplementary information.

The PCA showed a clear differentiation between the two genebanks, despite some overlap (Figure 3.2). The first two axes explained 15.05 % of the overall variance and separated USDA genotypes from IPK genotypes. The two data sets without missing values and imputed values differentiate between the two genebanks (Figure S3.1). The morphological traits did not show the same strong clustering into USDA and IPK accessions in the PCA (Figure 3.3) as the genotyping data, which may reflect the inherent phenotypic diversity and a high of genotype × environment interactions that was observed in the different accessions (Yousef et al., submitted). However, a *t*-test that compared trait values found a significant differentiation between the two genebanks in three of the six traits (curd width, cluster width and number of days to budding: $p< 0.001$; Figure 3.4). Mantel tests showed a positive correlation between phenotypic and genetic ($r = 0.291$, $p < 0.001$).

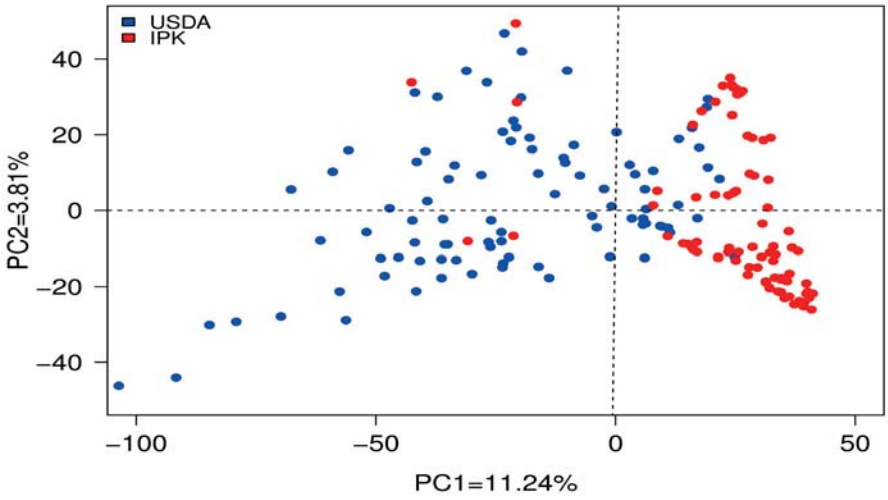

Figure 3.2 Principal component analysis of 174 accessions of cauliflower based on data with missing values (120,693 SNPs).

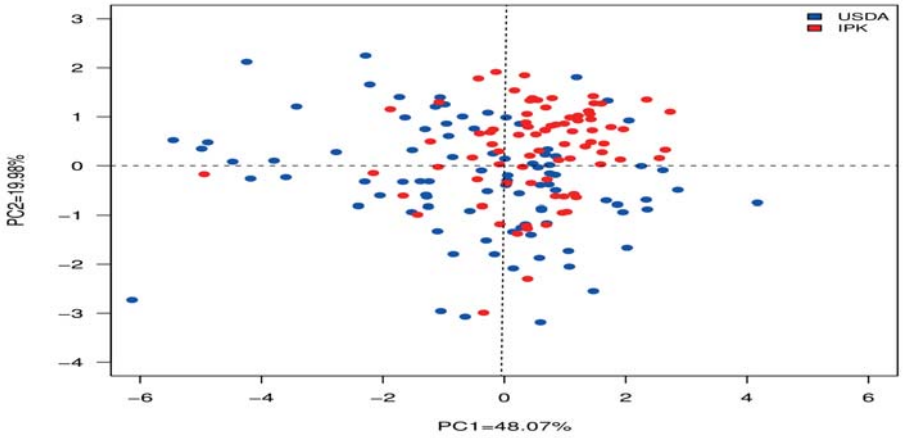

Figure 3.3 Principal component analysis of the 174 cauliflower accessions based on six morphological traits.

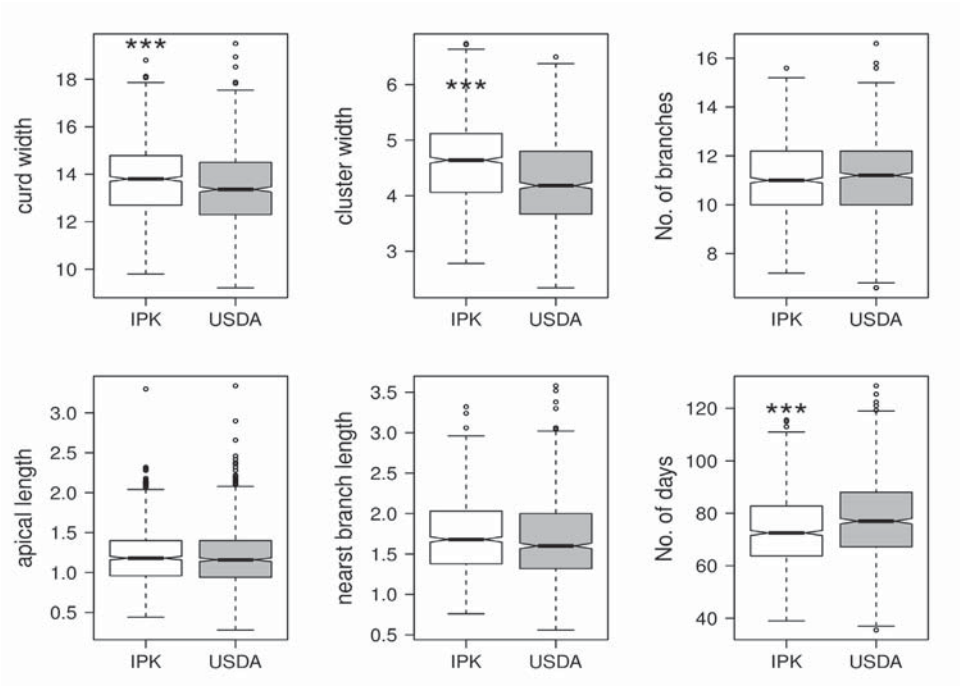

Figure 3.4 Box plots of six curd-related traits in accessions grouped by seed source (USDA and IPK). Significant differences between USDA and IPK as observed from t test are indicated with stars above the whiskers (***<0.001). Individual phenotype values are averages of 6 environments (2 locations and three growing seasons)

Further tests of genetic differentiation confirmed the strong differentiation between the two genebanks. A PCoA based on pairwise Fst values clearly separated the USDA from the IPK accessions on the first principal component axis, which explained 24% of the overall variance, whereas the second axis explained only 8% of the variance (Figure 3.5). In the DAPC analysis, fifty components and three discriminant functions were retained. According to the Bayesian information criterion (BIC; Figure S3.2), a grouping into five clusters is most consistent with the data, which explains 62.5% of the variance (Figure 3.6). Although a plot of the DAPC analysis of the genotype data with missing values does not show a strong differentiation between the USDA and IPK accessions (Figure 3.6), the two data sets without missing values and imputed values differentiate between the two gene banks (Figure S3.3 and S3.4). We also observed sub-groups within the USDA and IPK, but they did not cluster accessions by the country of origin.

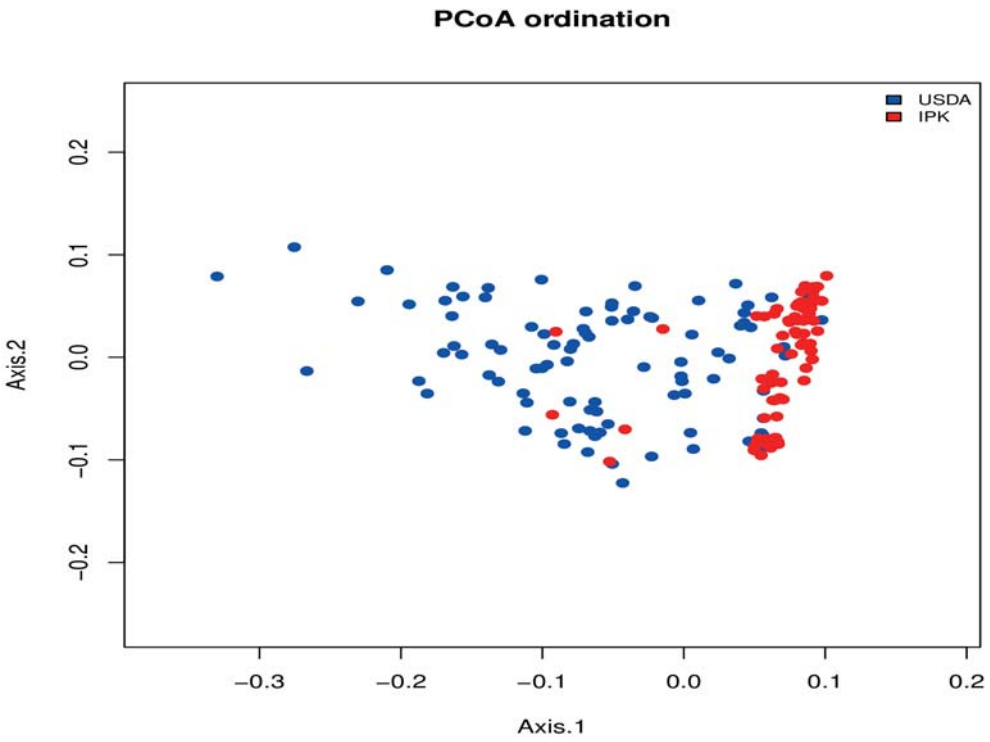

Figure 3.5 Scatter plot from PCoA based on the pairwise Fst between the genotypes using data without missing values.

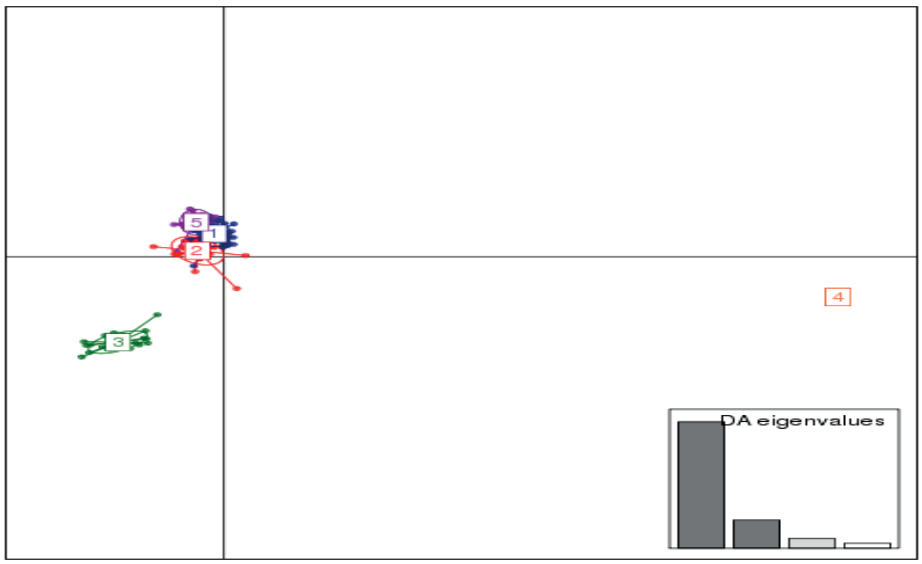

Figure 3.6 Scatter plot of DAPC analysis showing the first two principal components of the analysis using data with missing values. Cluster No. 3 consists of 9 accessions that are from USDA. While cluster No. 3 consists of 30 accessions (27 from USDA and 3 from IPK. The other clusters are mix between the IPK and USAD.

A neighbor joining tree based on a pairwise distance matrix separated the 174 accessions into two distinct groups (Figure 3.7A) representing the two genebanks. In both groups, accessions are not differentiated into well-supported subgroups that may reflect the distribution of the six phenotypic trait values or the country of origin (Figure 3.7B). Also, the other two data sets, without missing values and imputed values, differentiate between the two gene banks (Figure S3.5 and S3.6).

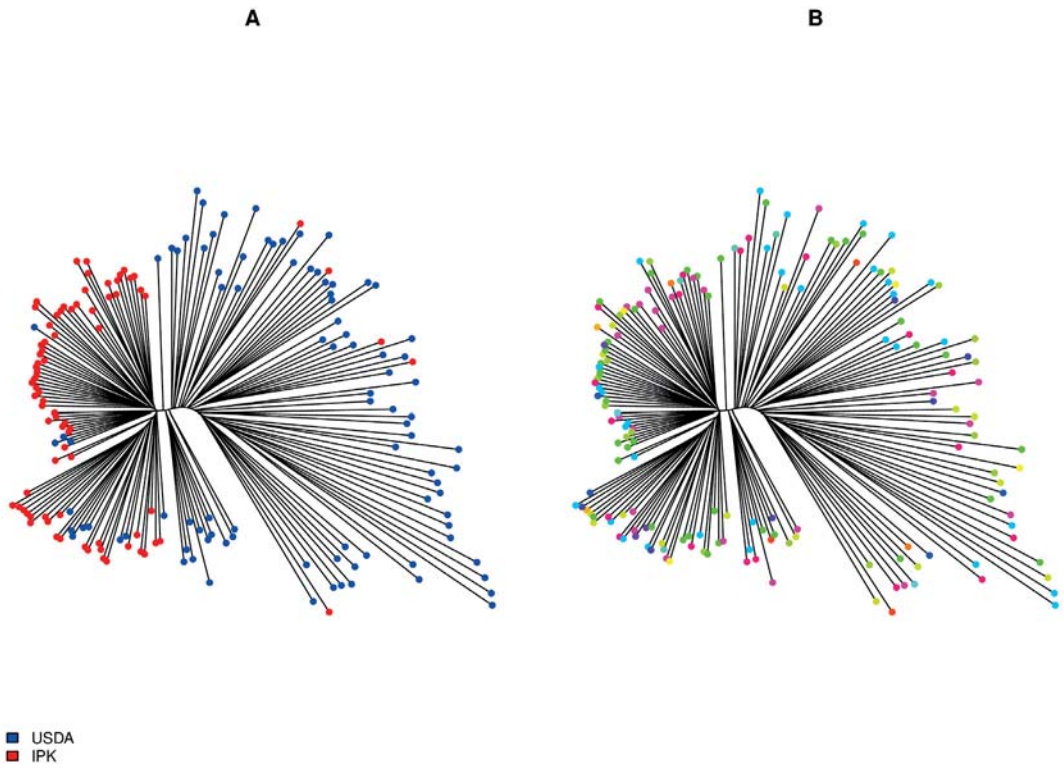

Figure 3.7 Neighbor joining tree for 174 cauliflower accessions based on the pairwise distance matrix using data with missing values. In Figure 3.7A each genebank was represented by single color. In Figure 3.7B each country was represented by single color.

Finally, we used the model-based STRUCTURE and ADMIXTURE approaches for population structure inference. With STRUCTURE, the most likely number of clusters was K=2 according to the ΔK criterion [41] (Figure S3.7), and the two groups mainly differentiated between the two genebanks (Figure 3.8). Based on cross-validation, ADMIXTURE identified different five clusters (Figure S3.8), and similar to STRUCTURE, the number of clusters among IPK

accessions remained essentially the same whereas they increased among USDA accessions for K ranging from *K* = 4 to *K* = 10 (Figure S3.9 – S3.11)

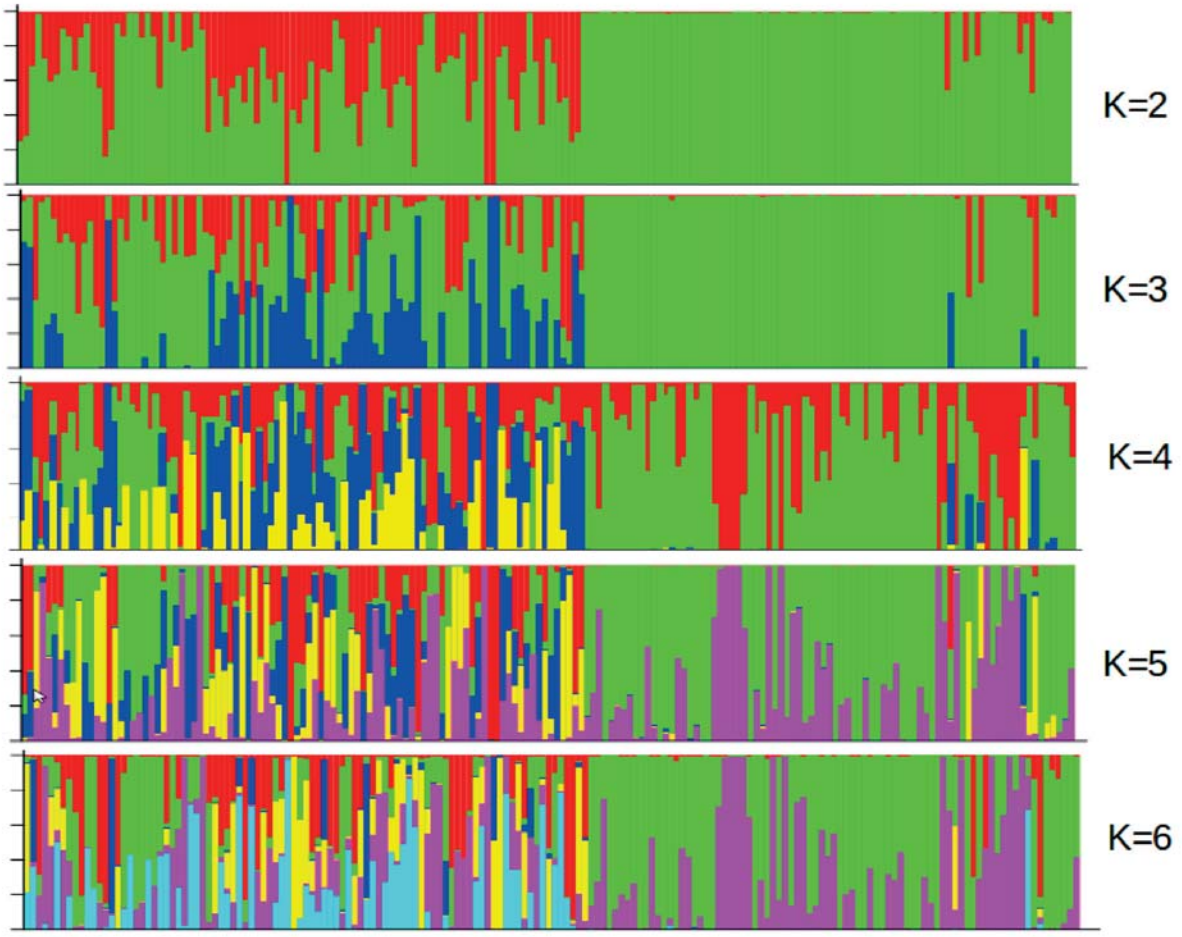

Figure 3.8 Population structure generated by STRUCTURE 2.3 among the 174 cauliflower genotypes (K=2, 3, 4, 5, 6) using data with missing values. Each horizontal bar represents one genotype, which is partitioned into up to K colored segments. The codes for the most probable population structure K=2 are as follows: USDA accessions (number 1-93) and IPK accessions (number 94-174).

Despite the strong differentiation between the two genebanks, an analysis of molecular variance (AMOVA) showed that a larger proportion of genetic variation (87%) segregates within rather than between genebanks (13%; $p < 0.001$; Table S3.5-S3.7). The overall genetic differentiation between the two genebanks was Fst = 0.029. The mean pairwise Fst of accessions within each

genebank was 0.301 for the USDA and 0.160 for the IPK accessions (calculated for the data set without missing values), showing that the USDA accessions are more differentiated from each other than the IPK accessions.

3.4.3 Genetic diversity

For a further comparison between the accessions from the two genebanks, we calculated several genetic diversity parameters (Table 3.1). Values for expected heterozygosity, observed heterozygosity, percentage of polymorphic loci and nucleotide diversity were all larger in the USDA than in the IPK accessions. In both genebanks, accessions tend to have a high inbreeding coefficient (>0.5).

Table 3.1 Sample size and measures of diversity within two genebanks based three different data sets.

	Without missing values (1,444 SNPs)		**With missing values (120,693 SNPs)**		**Imputed data (120,693 SNPs)**	
	USDA	**IPK**	**USDA**	**IPK**	**USDA**	**IPK**
N	93	81	93	81	93	81
He	0.157	0.088	0.214	0.149	0.201	0.107
Ho	0.167	0.127	0.419	0.339	0.094	0.077
F	0.501	0.500	0.501	0.501	0.495	0.507
%P	0.945	0.464	0.833	0.516	0.848	0.603
Π	0.158	0.089	0.157	0.095	0.202	0.110

N= number of genotypes, He = Expected heterozygosity, Ho = Observed heterozygosity, F = Inbreeding coefficient, %P = Percent of polymorphic loci, π = nucleotide diversity.

3.4.4 Detection of highly differentiated outlier SNPs

To identify polymorphisms whose frequencies are strongly different between the two genebank collections, we performed outlier tests only with the GBS data without missing values (1,444 SNPs). We identified one 182 (12.6%) SNPs as outliers with LOSITAN based on Fst values among the two populations in 10 iterations (FDR < 0.1). All outliers appeared in the upper part

of the Fst distribution figure, which is consistent with a signature of differentiation by directional selection (Figure S3.12A). ARLEQUIN identified only 79 (5.5%) loci as outliers (Figure S3.12 B). Since Bayenv2 does not calculate *p*-values for the X^TX statistic, we ordered SNPs by the rank order of X^TX values, as suggested by Günter and Coop [49], to detect strongly differentiated SNPs. A total of 72 SNPs were in among the top 5% (Table S3.8). All detected outliers were distributed over all nine chromosomes (Figure S3.9).

3.5 Discussion

3.5.1 Assessment of genetic diversity by GBS

Previously, several marker types were used to assess the genetic diversity of cauliflower, these marker genotypes have several disadvantages that can be overcome by sequence-based methods like GBS [20]. Although we generated a large number of raw reads, a substantial proportion of reads did not align to the *B. oleracea* reference genome, similar to what has been observed in maize [23]. The low proportion may reflect that the reference genome was produced with the subspecies *B. oleracea spp. capitata* (cabbage) instead *B. oleracea var. botrytis* (cauliflower), or that the reference genome is still incomplete. It may be also influenced by the limited sensitivity of the Burrows-Wheeler Alignment (BWA) software or a high proportion of presence/absence variation (PAV) as suggested by Romay et al. [23].

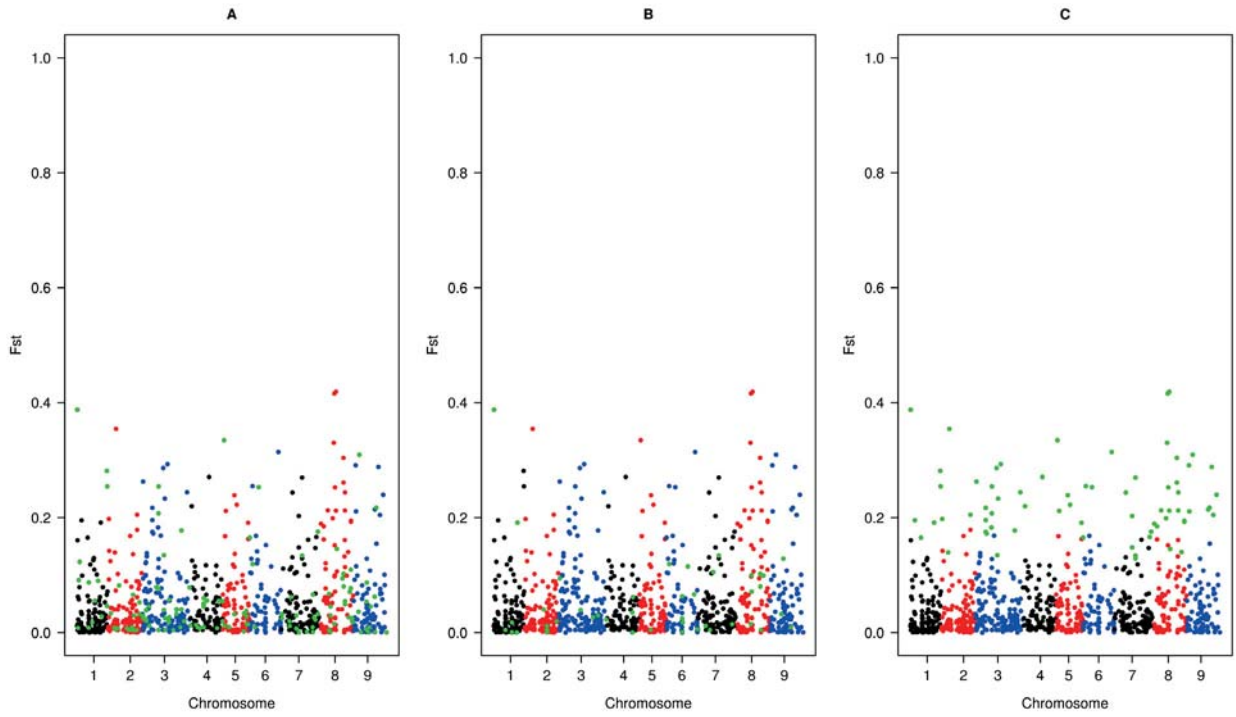

Figure 3.9 Distribution of Fst values in the *B. oleraceae* genome and Fst outliers which resulted form LOSITAN (A), ARLEQUIN (B) and Bayenv2 (C) which are represented in green color.

Although GBS was used for a wide range of species and is an effective method for generating tens of thousands SNP markers [20, 22], the high proportion of missing data, which in our study varied between 19% and 77%, is a major disadvantage [20]. Two strategies were suggested to overcome the high proportion of missing data that include the imputation of missing values and the sequencing to higher read depths [21].

3.5.2 Patterns and causes of genetic structure in cauliflower accessions

Despite the high proportion of missing data, our study demonstrates the utility of GBS for analyzing genetic diversity and population structure in *B. oleracea* genebank accessions. All analysis methods suggested that the worldwide samples of cauliflower cluster into two distinct groups that do not reflect their geographic origin but the seed source (i.e., genebank). A geographic population structure was found in previous survey of cauliflower cultivars [12]

despite a smaller sample size than our study. Furthermore, the genotyping of switchgrass, maize and sorghum genebank accessions with GBS revealed patterns of genetic diversity that reflected the known ancestral history, morphological types and geographic distribution of three crops [23-25]. Since each set of accessions we analysed from both genebanks represent independent worldwide samples of cauliflower genotypes (Table S3.1), a strong clustering by genebank suggest that other effects influence the genetic composition of the accessions used in out study.

The lower genetic diversity of the IPK accessions material and the population structure of the complete sample has several possible explanations. First, although accessions from both genebanks represent a global sample, the initial sample of varieties may have led to sets of accessions with different levels of genotypic and phenotypic diversity. This explanation is consistent with a higher proportion of exotic material (landraces) in the USDA than in the IPK set of accessions. Five accessions in the IPK set differed markedly from the other IPK accessions in all analyses and they clustered together with the set of USDA accessions. Since the five genotypes consist of four landraces and one hybrid (Table S3.1), this connection between the two genebanks support the notion that the USDA genbank contains more genetically distinct accessions than the IPK genbank. Unfortunately, limited passport data did not allow us to utilize information about the breeding history and relationships of varieties as co-variate in the population structure analysis. In addition, the absence of a strong geographic structure may result from a combination of low genetic diversity in cauliflower [14, 16], an exchange of seeds over large distance in historical time, and a high level of gene flow due to outcrossing with other varieties [50]. These processes are difficult if not impossible to reconstruct and confound the analysis of genetic diversity.

A second explanation for the low genetic diversity and structure may be multiple effects of seed regeneration procedures because genetic changes resulting from seed regeneration can be substantial [51-54]. In several species, the *ex situ* conserved genetic resources had a lower diversity than *in situ* conserved populations [55-57] or historical material [50]. A reduction in the diversity of *ex situ* genebank material is mainly caused by a small number of individuals per accession that are usually conserved. Such accessions are exposed to genetic bottlenecks, inbreeding depression, the accumulation of mildly deleterious mutations and a loss of genetic diversity by random drift [9, 56, 58-59]. The effects of small population size are indicated by a

high inbreeding coefficient of >0.5 for all accessions estimated from the GBS data. However, there are also significant differences between the regeneration procedures at the USDA and IPK genebanks that may contribute to the observed patterns of diversity. The multiplication in the USDA genebank is carried out in 12 x 24 ft cages (corresponding to 26.8 m^2) with mesh covers and at least 100 plants per accession propagated in each regeneration cycle. At IPK, cauliflower accessions are cultivated in small glass houses containing other species as well. The total area is about 6 m^2 and 20-25 plants are regenerated in each cycle. At IPK, seeds are regenerated after 20 years and at USDA after 15 years. Taken together, the effect of a small population size appears to be significantly stronger at the IPK than USDA genebank, which is consistent with the observed levels of diversity.

A third process affecting genetic diversity in genebank material that is closely linked to the seed propagation is natural and/or artificial selection during seed propagation. Since the accessions were collected from different regions of the world, one expects strong selection on genes controlling traits photoperiod sensitivity, flowering time, and other traits responding to the specific propagation environment. Taken together, the effects of selection on fitness and genetic diversity can be quite strong. Since pollination is managed with commercial pollinators like bumblebees, the reproductive success is not closely monitored. To test the potential impact of selection, we used three outlier tests to identify highly differentiated SNPs. Out of 1,444 of tested SNPs without missing data, 12.5% (LOSITAN), 5.5% (ARLEQUIN) and 5% (Bayenv2) were classified as highly differentiated between the two sets of accessions. The three methods differ in their approach to control for population structure and kinship to reduce the proportion of false positives. Shimada et al. [60] suggested to consider only SNPs that were identified by more than one method as true outliers, and such an approach was further confirmed in a simulation study of non-equilibrium populations [61]. Hence, in a comparison of outliers identified in our data (Table S3.9), we identified 14 SNPs as outliers by LOSITAN and Bayenv2, eight SNPs by LOSITAN and ARLEQUIN, three SNPs by Bayenv2 and ARLEQUIN, and 1 SNP by all three methods. This is a small proportion (<1%) of the total number of SNPs tested, which suggests that if natural or unintentional artificial selection during seed regeneration has an effect on genetic differentiation, it likely affects only few genes or short genomic regions. However, the effect of selection requires further study and needs to consider that selection tests developed equilibrium populations may not apply to sets of genebank accessions.

3.5.3 Characterizing genebank accessions with GBS

A main advantage of GBS is its applicability to any species because it does not require setup costs like SNP arrays and has low cost per individual reaction. On the other hand, GBS has a high proportion of missing data that may reduce the power for correct estimation of population parameters. Data imputation was suggested as a solution because it can be accurate and then increase the quality of genomic selection or association mapping [62,63]. In our study, however, a comparison between three data GBS sets consisting of SNPs without missing values, SNPs with missing values, and imputed SNPs revealed only a minor effect of data imputation on the ability to infer the population structure, although diversity estimates differed significantly between imputed and non-imputed data (Table 3.1). This result confirms a study of Fu [64] who reported that the estimation of heterozygosity and inbreeding coefficients was less accurate with a high proportion of missing data and that estimation biases were much smaller for data sets with missing values than for imputed data sets. Therefore, data imputation on GBS data obtained from highly diverse populations cannot be recommended.

3.6 Conclusion

Our study outlined the usefulness of GBS to characterize the genetic diversity of genebank accessions of a minor crop like cauliflower. The strong differentiation between genebanks points to the importance of seed regeneration procedures in maintaining genetic variation. The low cost of GBS and its technical simplicity suggest that it can be used as a tool to real-time monitoring of genetic diversity during seed regeneration in order to select individuals of an accession that maximize the genetic diversity to mitigate some disadvantages of small population sizes of *ex situ* conserved plant genetic resources.

Acknowledgments

We express our thanks to Dr. Fabian Freund, Dr. Christian Lampei, Dr. Dounia Saleh, Dr.Torsten Günther, Patrick Thorwarth and Linda Homann for their assistance with the analysis of data. Also, we thank the USDA and IPK genebanks for seed stocks. We thank Prof. Dr. Marcus Koch, Heidelberg Botanic Garden and Herbarium HEID, for offering the seeds of wild type of *Brassica oleraceae*.

3.7 References

1. Lu X, Liu L, Gong Y, Zhao L, Song X, et al. (2009) Cultivar identification and genetic diversity analysis of broccoli and its related species with RAPD and ISSR markers. Scientia Horticulturae 122:645–648.

2. Rao VR, Hodgkin T (2002) Genetic diversity and conservation of plant genetic resources. Plant Cell, Tissue and Organ Culture 68:1-19.

3. de Jesus ON, Silva, SO, Amorim EP, Ferreira CF, de Campos JM, et al. (2013) Genetic diversity and population structure of Musa accessions in ex situ conservation. BMC Plant Biology 13: 41.

4. Parzies HK, Spoor W, Ennos RA (2000) Genetic diversity of barley landrace accessions (*Hordeum vulgare ssp. vulgare*) conserved for different lengths of time in ex situ gene banks. Heredity 84:476–486.

5. Hakansson J (2004) Genetic Aspects of Ex Situ Conservation: Introductory Paper, Department of Biology, Linköping University. Available: https://www.ifm.liu.se/biology/zoology/avian/phd-literature-essays/Jennie-Hintroduktionsuppsats.pdf. Accessed:16 November 2014.

6. Kasso M, Balakrishnan M (2013) Ex Situ Conservation of Biodiversity with Particular Emphasis to Ethiopia. ISRN Biodiversity, vol. 2013, Article ID 985037, 11 pages. Doi:10.1155/2013/985037.

7. Kjaer ED, Graudal L, Nathan I (2001) Ex situ conservation of commercial tropical trees: Strategies, options and constraints; A paper presented at ITTO International Conference on ex situ and in situ Conservation of Commercial Tropical Trees, Yogyakarta, Indonesia; 2001.

8. Lauterbach D, Burkart M, Gemeinholzer B (2012) Rapid genetic differentiation between ex situ and their in situ source populations: an example of the endangered *Silene otites* (*Caryophyllaceae*). Botanical Journal of the Linnean Society 168: 64–75.

9. Lee SA, Fowke JH, Lu W, Ye C, Zheng Y, et al. (2008) Cruciferous vegetables, the GSTP1 Ile105Val genetic polymorphism, and breast cancer risk. The American Journal of Clinical Nutrition 87:753-760.

10. Tang L, Gary RZ, Guru K, Kirsten BM, Zhang Y, et al. (2008) Consumption of raw cruciferous vegetables is inversely associated with bladder cancer risk. Cancer Epidemiology, Biomarkers and Prevention 17: 938-44.

11. FAO (2012) Food and Agriculture Organization of the United Nation. The Statistics Division, http://www.fao.org.

12. Astarini IA, Plummer JA, Lancaster RA, Yan G (2006) Genetic diversity of Indonesian cauliflower cultivars and their relationships with hybrid cultivars grown in Australia. Scientia Horticulturae 108: 143-150.

13. Izzah NK, Lee J, Perumal S, Park JY, Ahn K, et al. (2013) Microsatellite-based analysis of genetic diversity in 9 commercial *Brassica oleracea* L. cultivars belonging to six varietal groups. Genetic Resources and Crop Evolution 60:1967-1986.

14. Zhao Z, Gu H, Sheng X, Yu H, Wang J, et al. (2014) genetic diversity and relationships among loose-curd cauliflower and related varieties as revealed by microsatellite markers. Scientia Horticulturae 166: 105-110.

15. Truong HT, Ramos AM, Yalcin M, de Ruiter M, van der Poel HJA, et al. (2012) Sequence-based genotyping for marker discovery and co-dominant scoring in germplasm and populations. PLoS One 7:e37565.

16. Tonguc M, Griffiths PD (2004) Genetic relationships of *Brassica* vegetables determined using database derived sequence repeats. Euphytica 137: 193-201.

17. Louarn S, Torp AM, Holme IB, Andersen SB, Jensen BD (2007) Database derived microsatellite markers (SSRs) for cultivar differentiation in *Brassica oleracea*. Genetic Resources and Crop Evolution 54: 1717-1725.

18. Varshney RK, Chabane K, Hendre PS, Aggarwal RK, Graner A (2007) Comparative assessment of EST-SSR, EST-SNP and AFLP markers for evaluation of genetic diversity and conservation of genetic resources using wild, cultivated and elite barleys. Plant Science 173: 638-649.

19. Jones ES, Sullivan H, Bhattramakki D, Smith JSC (2007) A comparison of simple sequence repeat and single nucleotide polymorphism marker technologies for the genotypic analysis of maize (Zea mays L.). Theoretical and Applied Genetics 115:361-37.

20. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, et al. (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6: e19379.

21. Poland J, Rife TW (2012) Genotyping-by-sequencing for plant breeding and genetics. Plant Genome 5: 92-102.

22. Poland J, Brown PJ, Sorrells ME, Jannink J (2012) Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS ONE 7:e32253.

23. Romay MC, Millard M, Glaubitz JC, Peiffer JA, Swarts KL, et al. (2013) Comprehensive genotyping of the USA national maize inbred seed bank. Genome Biology 14: R55.

24. Morris GP, Ramu P, Deshpande SP, Hash CT, Shah T, et al. (2013) Population genomic and genome-wide association studies of agroclimatic traits in sorghum. PNAS 110: 453-458.

25. Lu F, Lipka AE, Elshire RJ, Glaubitz JC, Cherney J, et al. (2013) Switchgrass genomic diversity, ploidy and evolution: novel insights from a network-based SNP discovery protocol. PLoS Genetics 9: e1003215.

26. Fu YB, Cheng B, Peterson GW (2014) Genetic diversity analysis of yellow mustard (*Sinapis alba* L.) germplasm based on genotyping by sequencing. Genetic Resources and Crop Evolution 61: 579-594.

27. Yousef AAE, Lampei C, Schmid KJ (2015) Evaluation of cauliflower genebank accessions under organic and conventional cultivation in Southern Germany. Euphytica 201:389-400.

28. Lan TH, Paterson AH (2000) Comparative mapping of quantitative trait loci sculpting. sculpting the curd of *Brassica oleracea*. Genetics 155: 1927-1954.

29. Saghai-Maroof MA, Soliman KM, Jorgensen RA, Allard RW (1984) Ribosomal DNA spacer-length polymorphisms in barley: Mendelian inheritance, chromosomal location, and population dynamics. Proceedings of the National Academy of Sciences of the United States of America 81: 8014-8018.

30. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics 25:1754-60.

31. Liu S, Liu Y, Yang X, Tong C, Edwards D, et al. (2014) The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes. Nature communications 5:3930. 3930.

32. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics 25: 2078-2079.

33. Jombart T, Ahmed I (2011) adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. Bioinformatics 27:3070-3071.

34. R Core Team (2014) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

35. Mantel N (1967) The detection of disease clustering and a generalized regression approach. Cancer Research 27: 209-220.

36. Dray S, Dufour AB (2007) The ade4 package: implementing the duality diagram for ecologists. Journal of Statistical Software 22:1-20.

37. Paradis E, Claude J, Strimmer K (2004) APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20: 289-290.

38. Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: A new method for the analysis of genetically structured populations. MBC Genetics 11:94.

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

40. Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164:1567-1587.

41. Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software Structure: a simulation study. Molecular Ecology 14: 2611–2620.

42. Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome Research 19:1655-1664.

43. Alexander DH, Lange K (2011) Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC Bioinformatics 12: 246.

44. Excoffier L, Lischer HL (2010) Arlequin suite ver3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Molecular Ecology Resources 10:564-567.

45. Paradis E (2010) pegas: an R package for population genetics with an integrated-modular approach. Bioinformatics 26: 419-420.

46. Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. American journal of human genetics 78:629–644.

47. Beaumont MA, RA Nichols (1996) Evaluating loci for use in the genetic analysis of population structure. Proceeding the Royal Society of London 263:1619–1626.

48. Antao T, Lopes A, Lopes RJ, Beja-Pereira A, Luikart G (2008) LOSITAN: a workbench to detect molecular adaptation based on an Fst-outlier method. BMC Bioinformatics 9:323

49. Günther T, Coop G (2013) Robust identification of local adaptation from allele frequencies. Genetics 195:205-220.

50. Hagenblad J, Zie J, Leino MW (2012) Exploring the population genetics of genebank and historical landrace varieties. Genetic Resources and Crop Evolutions 59:1185-1199.

51. Dulloo ME, Hunter D, Borelli T (2010) Ex situ and in situ conservation of agricultural biodiversity: major advances and research needs. Notulae Botanicae Horti Agrobotanici Cluj-Napoca 38:123–135.

52. Börner A, Chebotar S, Korzun V (2000) Molecular characterization of the genetic integrity of wheat (*Triticum aestivum* L.) germplasm after long-term maintenance. Theoretical and Applied Genetics 100:494-497.

53. Chebotar S, Röder MS, Korzum V, Saal B, Weber WE, et al. (2003) Molecular studies on genetic integrity of open-pollinating species rye (*Secale cereale* L.) after long-term gene bank maintenance. Theoretical and Applied Genetics 107:1469-1476.

54. van Hintum TJL, Van De wiel CCM, Visser DL, Van Treuren R, Vosman B (2007) The distribution of genetic diversity in a *Brassica oleracea* gene bank collection related to the effects on diversity of regeneration, as measured with AFLPs. Theoretical and Applied Genetics 114: 777-786.

55. Gómez OJ, Blair MW, Frankow-Lindberg BE, Gullberg U (2005) Comparative study of common bean (*Phaseolus vulgaris* L.) landraces conserved ex-situ in genebanks and in-situ by farmers. Genetic Resources and Crop Evolution 52: 371-380.

56. Rucinska A, Puchalski J (2011) Comparative molecular studies on the genetic diversity of an ex situ garden collection and its source population of the critically endangered Polish endemic plant Cochlearia polonica E. Fröhlich. Biodiversity and Conservation 20: 401–413.

57. Brütting C, Hensen I, Wesche K (2013) Ex situ cultivation affects genetic structure and diversity in arable plants. Plant Biology15: 505-513.

58. Crossa J (1995) Sample size and effective population size in seed regeneration of monoecious plants, pp.140-143. In: Engels, J. M. M., R. R. Rao Eds. Regeneration of seed crops and their wild relatives. Proceedings of a consultation meeting, 4-7 December 1995, ICRISAT, Hyderabad, India. IPGRI, Rome, Italy.

59. Frankham R, Ballou JD, Briscoe DA (2002) Introduction to conservation genetics. Cambridge University Press, Cambridge, UK.

60. Shimada Y, Shikano T, Merila J (2011) A high incidence of selection on physiologically important genes in the three-spined stickleback, *Gasterosteus aculeatus*. Molecular Biology and Evolution 28:181-193.

61. Lotterhos KE, and Michael CW (2014) Evaluation of demographic history and neutral parameterization on the performance of FST outlier tests. Molecular Ecology 23:2178-2192.

62. Rutkoski JE, Poland J, Jannink J-L, Sorrells ME (2013) Imputation of unordered markers and the impact on genomic selection accuracy. Genes Genomes Genetics 3:427-439.

63. Marchini J, B Howie (2010) Genotype imputation for genome-wide association studies. Nature Reviews Genetics 11: 499-51.

64. Fu Y-B (2014) Genetic diversity analysis of highly incomplete SNP genotype data with imputations: an empirical assessment. Genes Genomics Genetics 4:891-899.

3.8 Supplementary Materials

Table S3.1. Numbers, accessions Id, accessions names, gene bank source and origin country of used accessions characterized in this study.

Table S3.2. GBS Barcode IDs and usage of barcodes.

Table S3.3. Number of raw read, mapped reads and percentage of mapped reads per genotype.

Table S3.4. Number of SNPs and percentage of missing data per genotype.

Table S3.5 Analysis of molecular variance (AMOVA) of given different groupings for data without missing values.

Level of variation	Sum of squares	Variance consonant	% variance explained	P
Among groups (genebanks)	1031.689	5.43273	5.64	<0.001
Within groups	31448.01	90.89022	94.36	<0.001
Total	32479.704	96.32294		

Table S3.6 Analysis of molecular variance (AMOVA) of given different groupings for data with missing values.

Level of variation	Sum of squares	Variance consonant	% variance explained	P
Among groups (genebanks)	105533.225	586.283	12.77	<0.001
Within groups	1385793.155	4005.18253	87.23	<0.001
Total	1491326.379	4591.46553		

Table S3.7 Analysis of molecular variance (AMOVA) of given different groupings for imputed data.

Level of variation	Sum of squares	Variance consonant	% variance explained	P
Among groups (genebanks)	131193.633	733.6279	15.02	<0.001
Within groups	1435733.927	4149.52002	84.98	<0.001
Total	1566927.56	4883.14792		

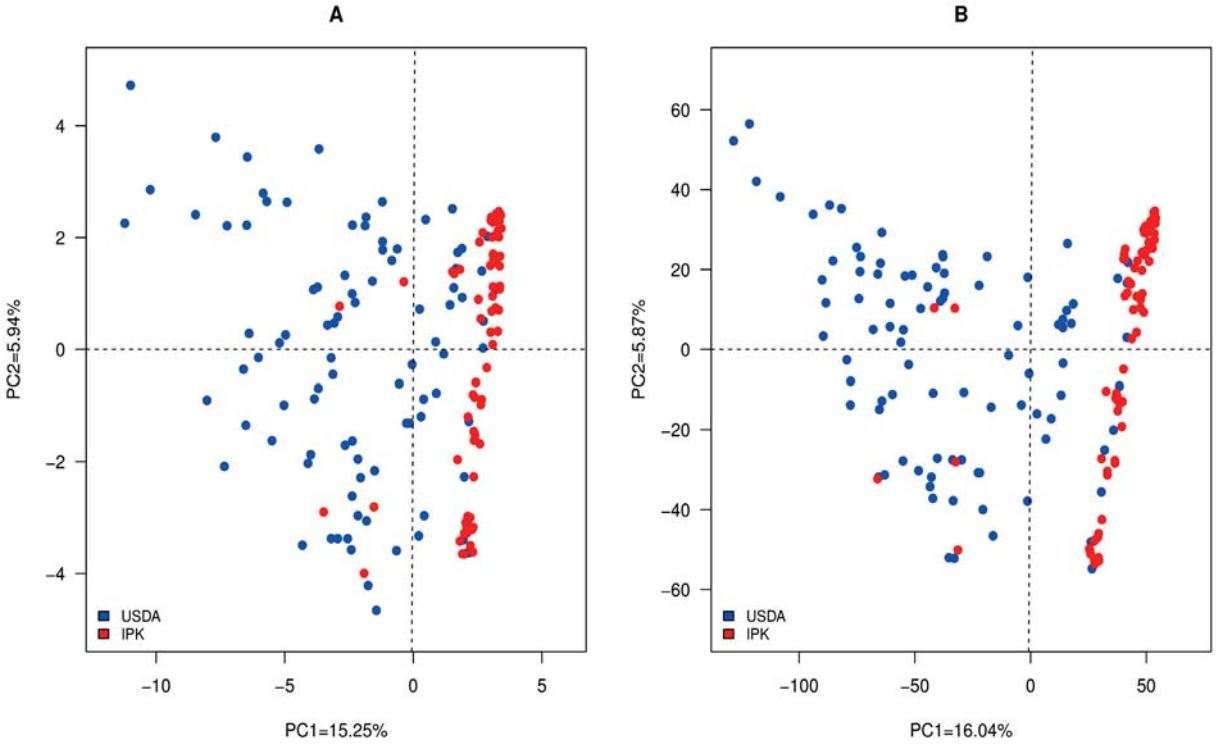

Figure S3.1 Principal component analysis of 174 accessions of cauliflower based on data without missing values (A) and imputed data (B).

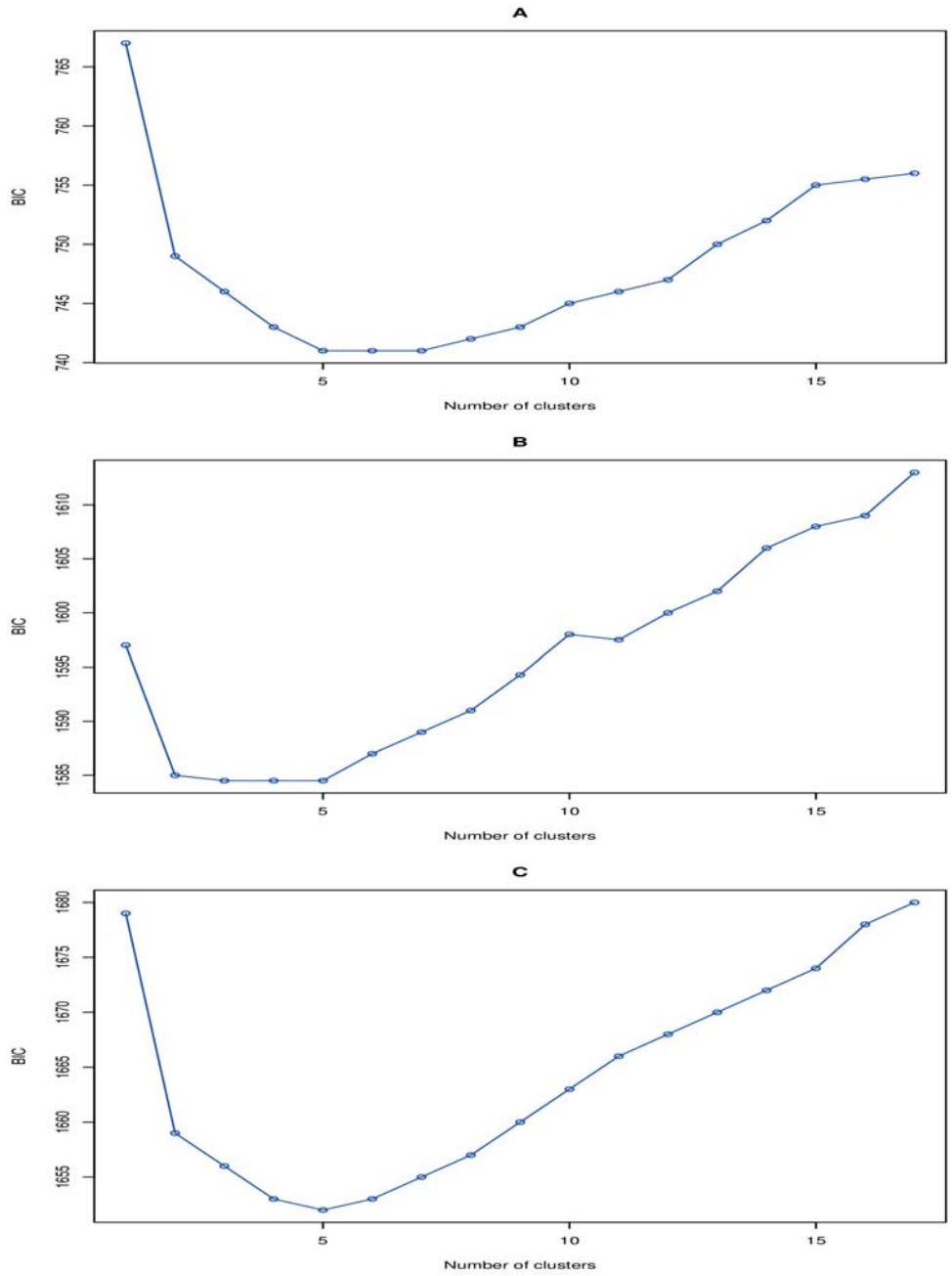

Figure S3.2 Plot of BIC estimates using DAPC to infer the number of clusters using data without missing values (A), data with missing values (B) and imputed data (C).

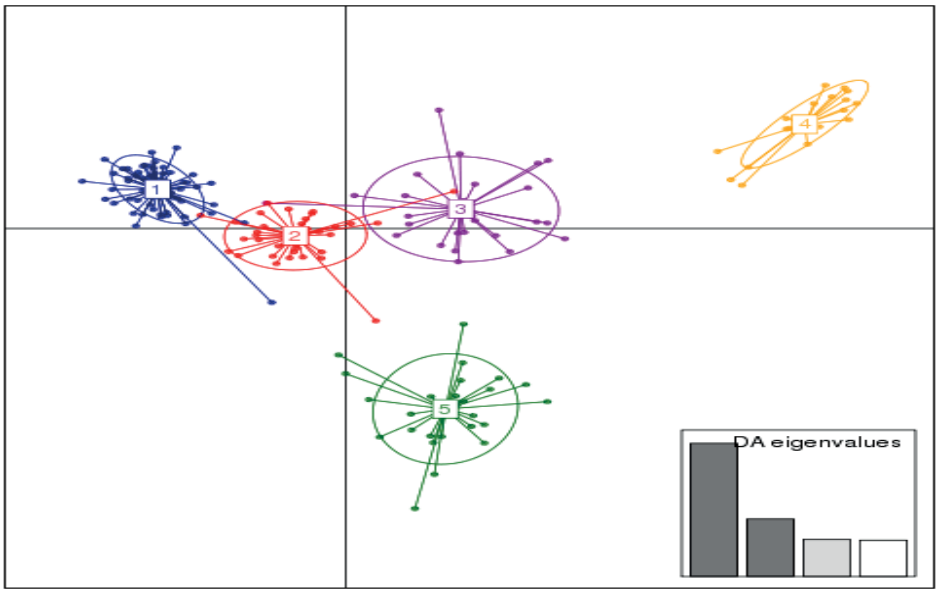

Figure S3.3 Scatter plot of DAPC analysis showing the first two principal components using data with missing values.

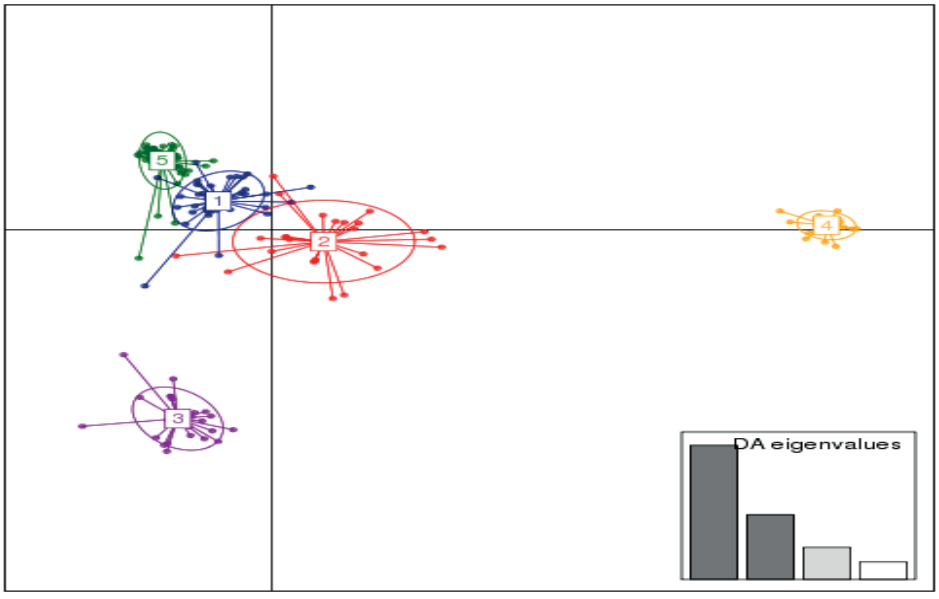

Figure S3.4 Scatter plot of DAPC analysis showing the first two principal components using imputed data.

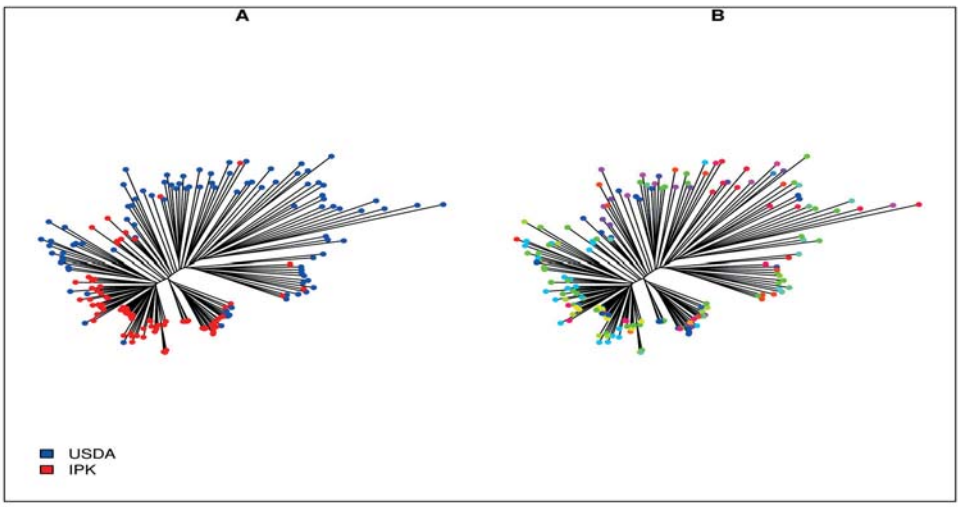

Figure S3.5 Neighbor joining tree for 174 cauliflower accessions based on the pairwise distance matrix of data without missing values. In Figure S3.5A genebanks are represented by different colors. In Figure S3.5B each of origin country is represented by a different color.

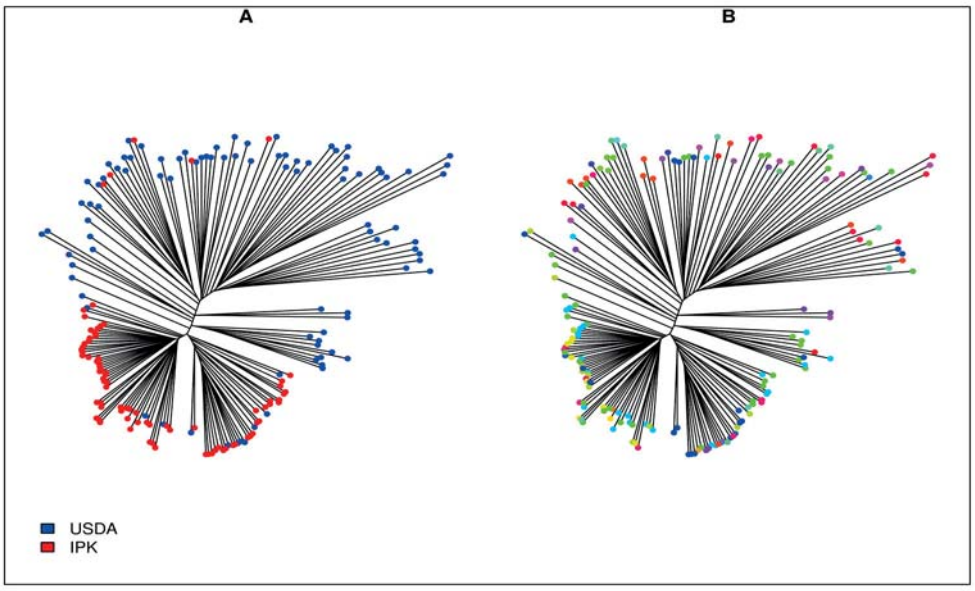

Figure S3.6 Neighbor joining tree for 174 cauliflower accessions based on the pairwise distance matrix of imputed data. In Figure S3.6A genebanks are represented by different colors. In Figure S3.6B each of origin country is represented by a different color.

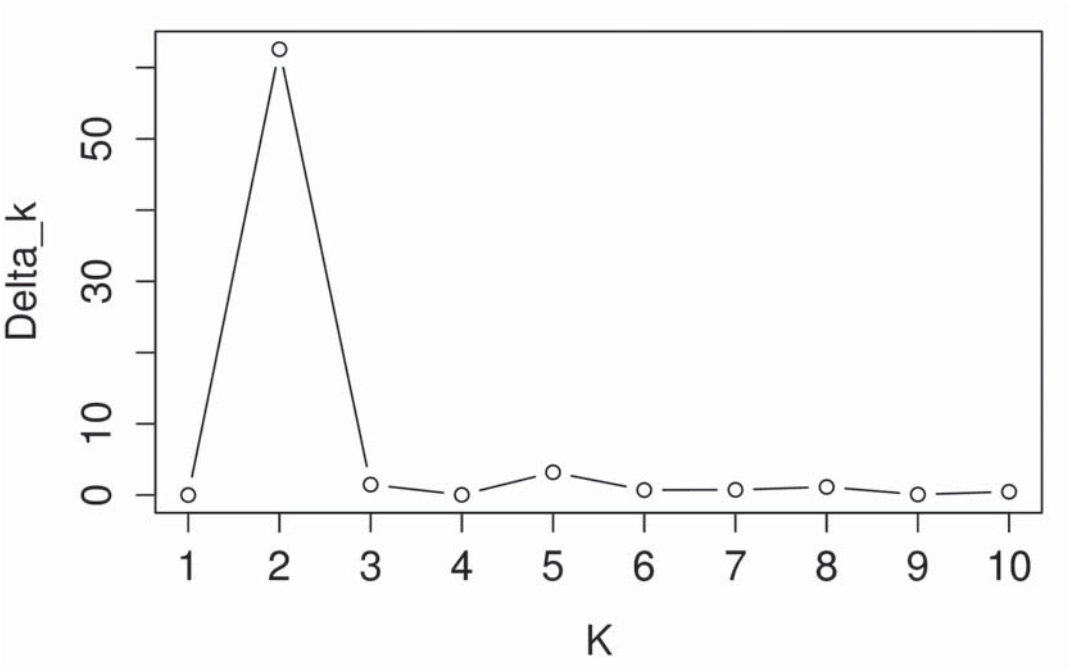

Figure S3.7 Graphical plot of Delta K values from ten runs of STRUCTURE using data without missing values according to Evanno et al. (2005).

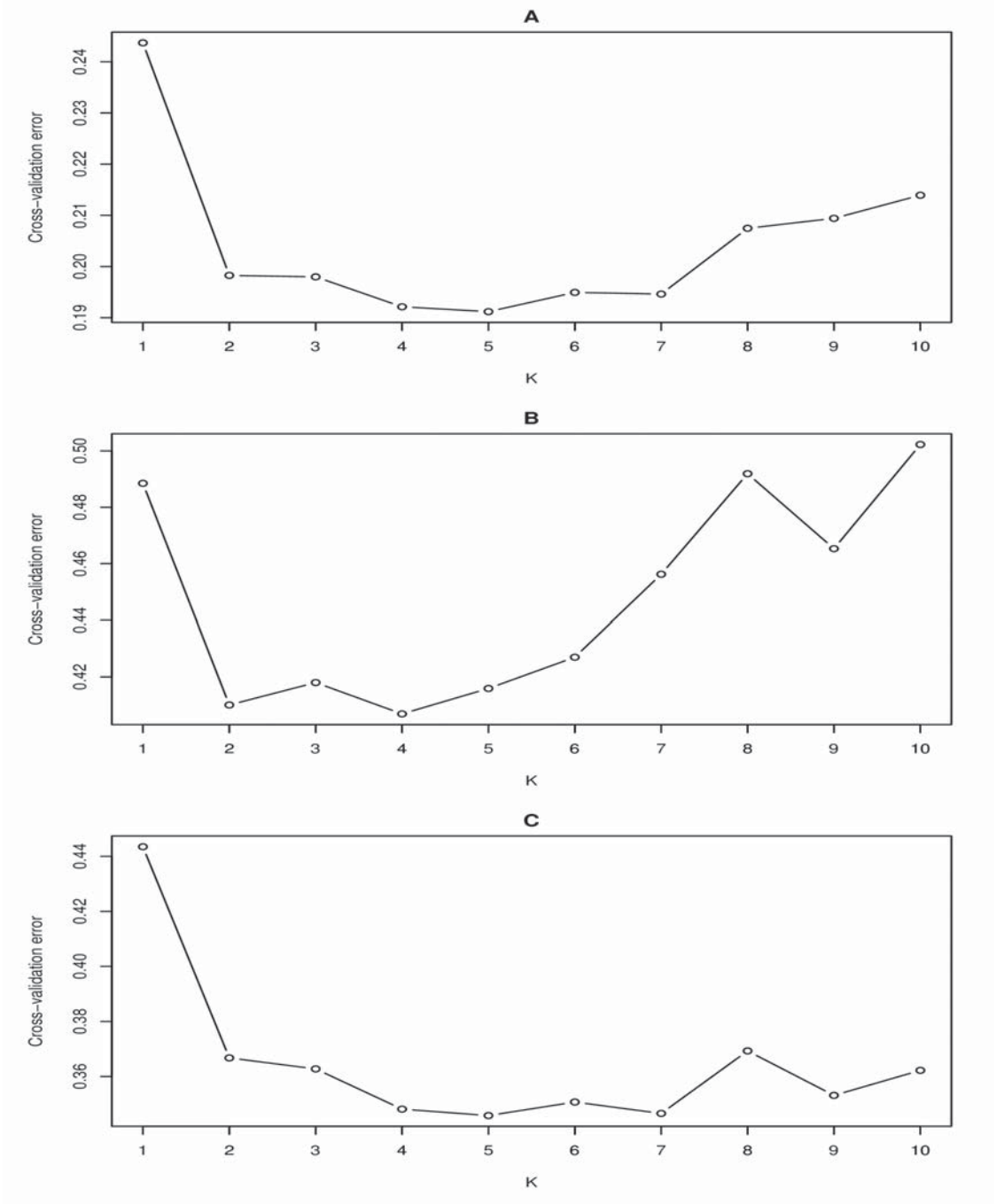

Figure S3.8 Cross validation plot for inference of the best K using data without missing values (A), data with missing values (B) and imputed data (C).

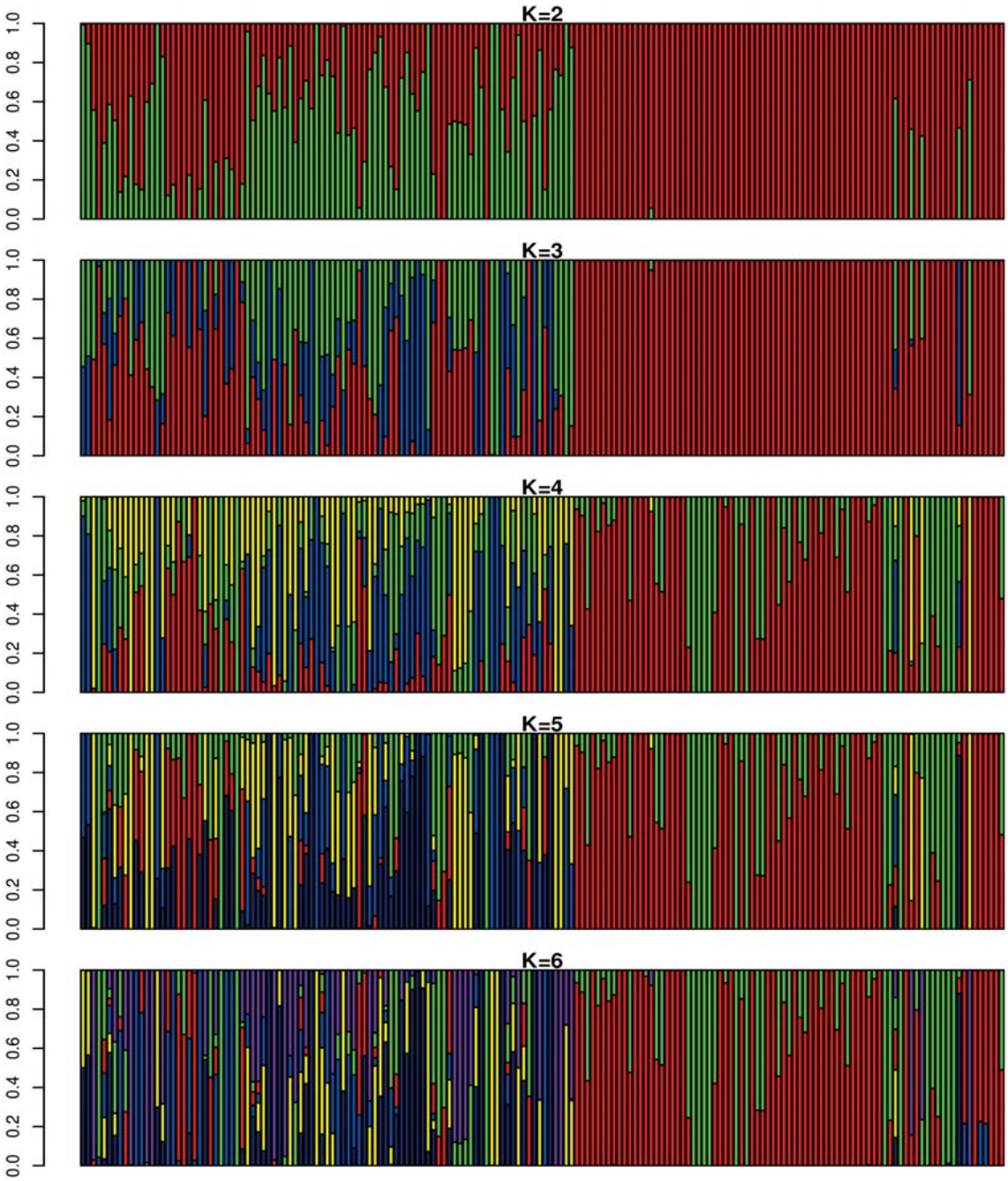

Figure S3.9 Population structure for 174 cauliflower accessions (K=2, 3, 4, 5, 6) generated by ADMIXTURE software using data with missing values. Each horizontal bar represents one genotype.

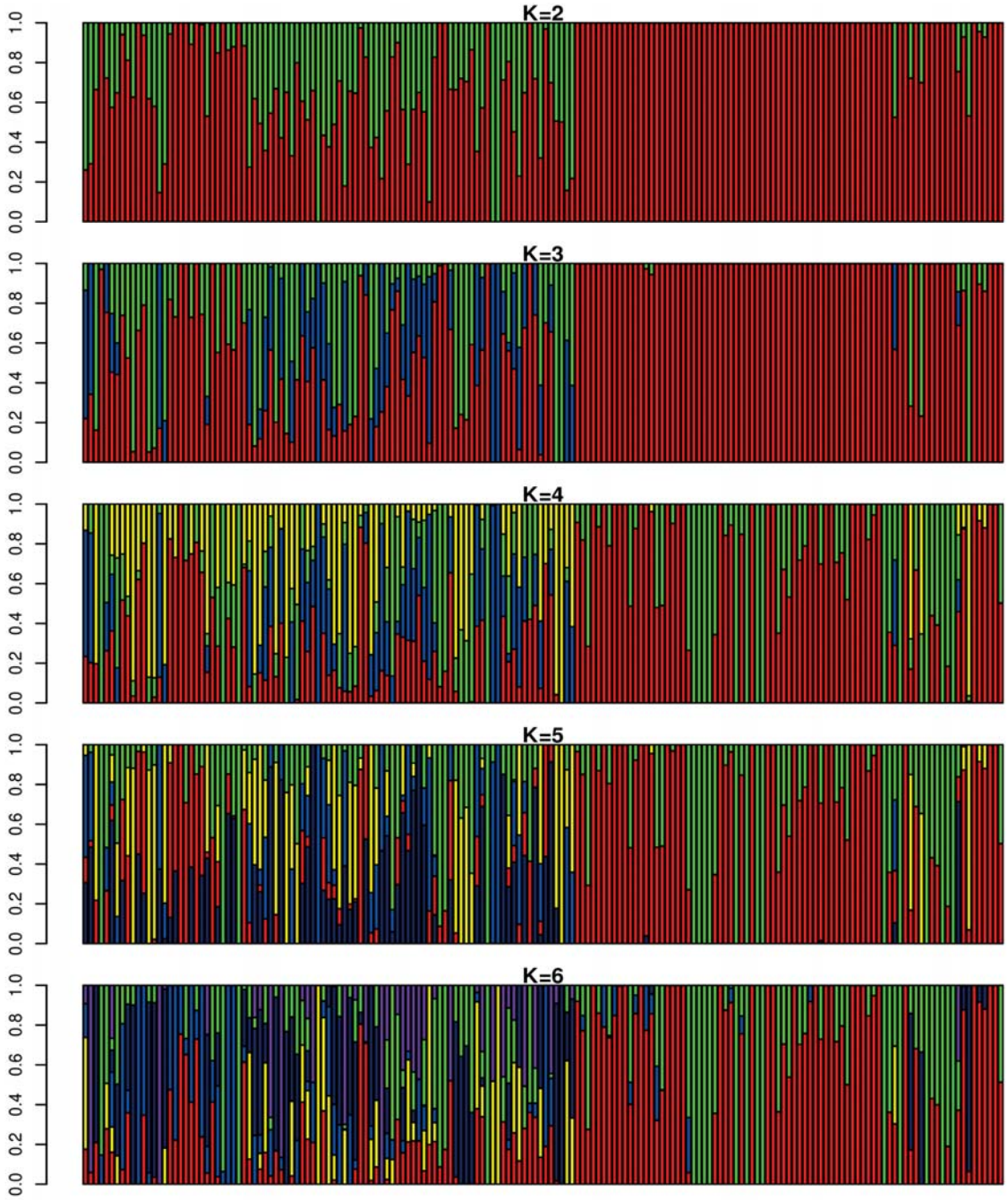

Figure S3.10 Population structure for 174 cauliflower accessions (K=2, 3, 4, 5, 6) generated by ADMIXTURE software using data without missing values. Each horizontal bar represents one genotype.

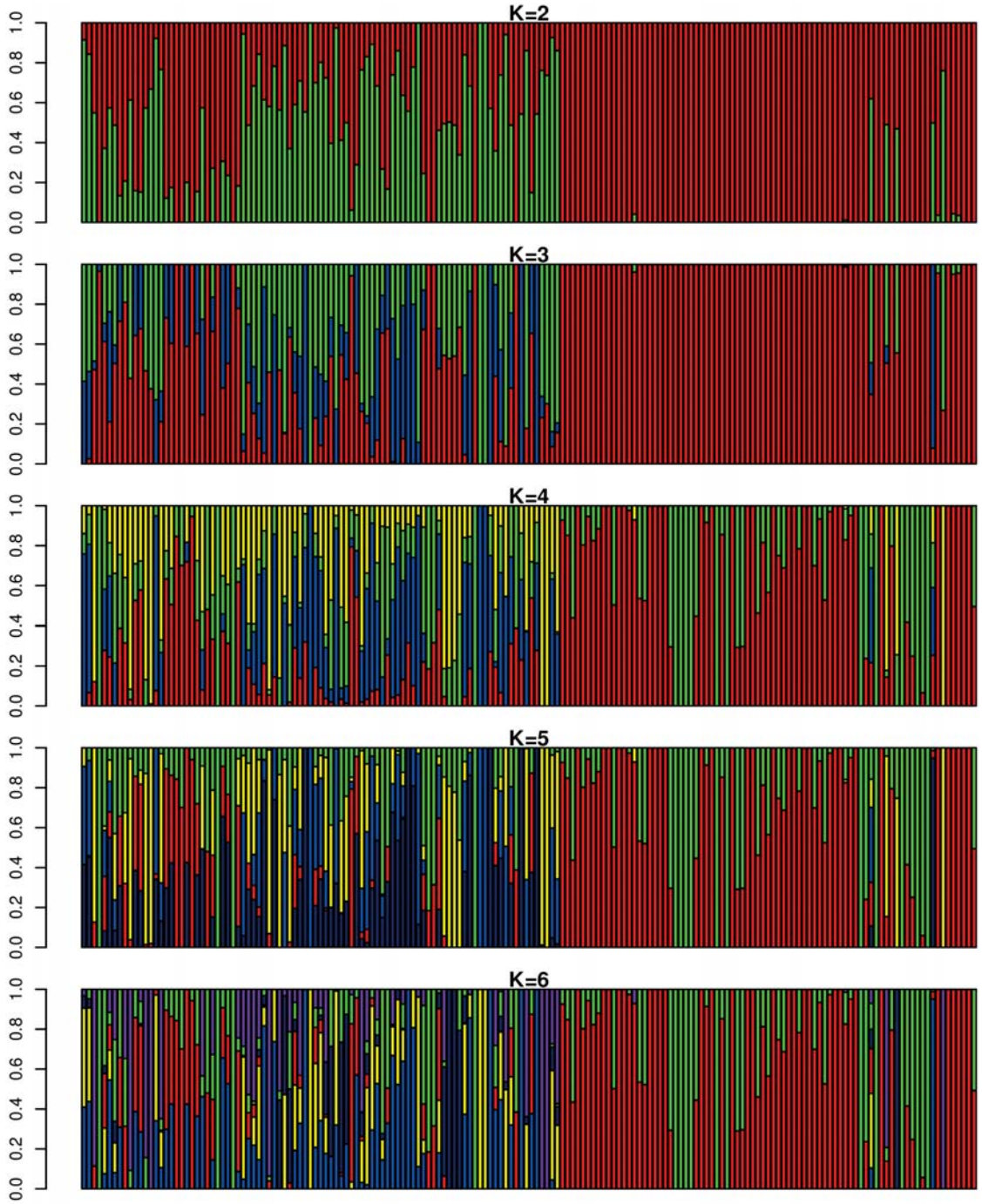

Figure S3.11 Population structure for 174 cauliflower accessions (K=2, 3, 4, 5, 6) generated by ADMIXTURE software using imputed data. Each horizontal bar represents one genotype.

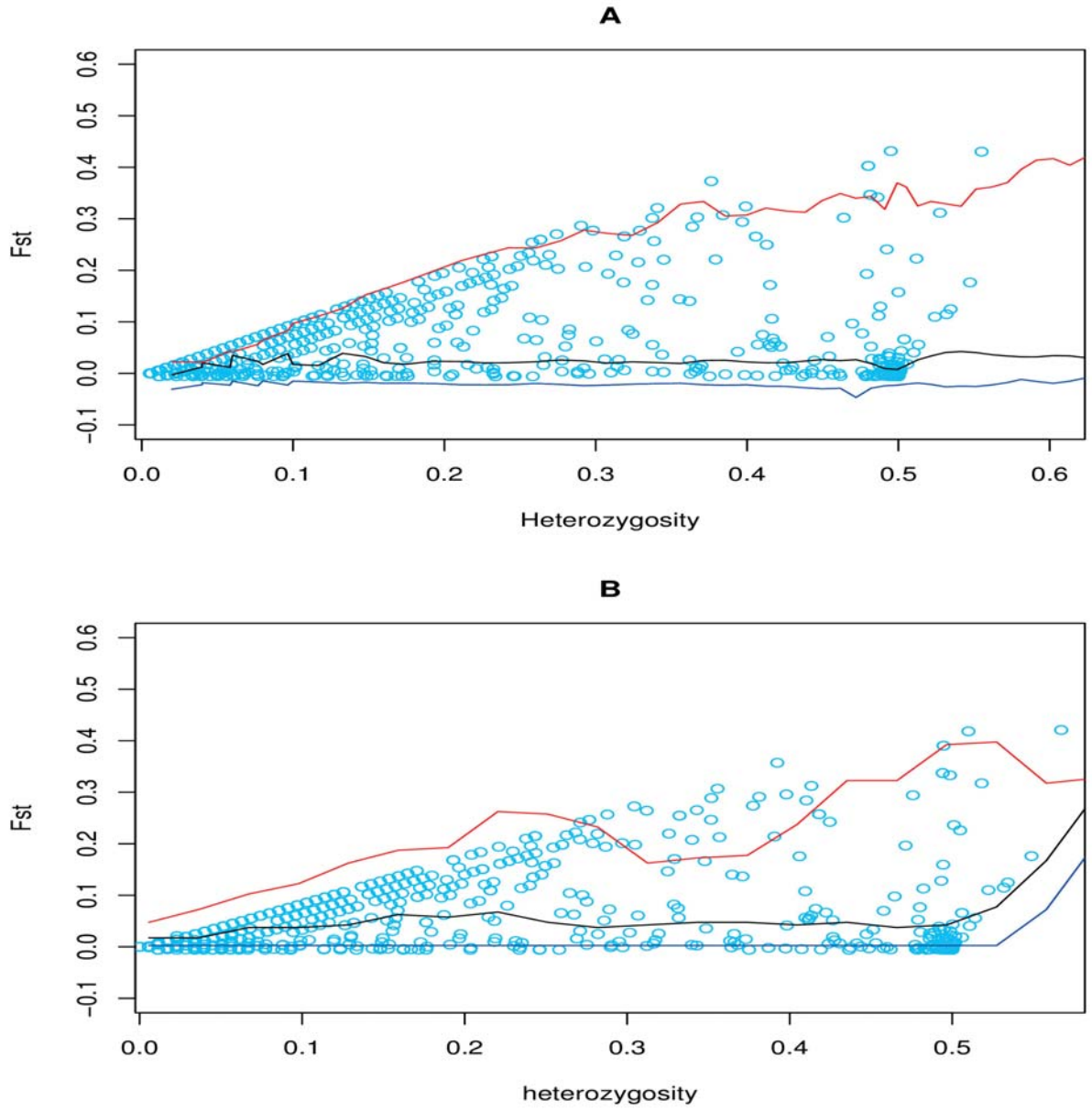

Figure S3.12 SNP neutrality test in two structure sub-populations with LOSITAN (A) and ARLEQUIN (B). Distribution of empirical Fst values is shown as function of expected hetrozigosity (He). The red and blue line indicates 99 and 1% confidence limits, respectively. The black line represents the median.

4 Genome-wide association and genomic prediction of curd-related traits in genebank accessions of cauliflower (*Brassica oleracea var botrytis* L.)

Eltohamy A.A. Yousef [1,2*], P. Thorwarth[1*] and K. J. Schmid[1]

[1] Department of Crop Biodiversity and Breeding Informatics (350b), University of Hohenheim, Fruwirthstraße 21, D - 70599 Stuttgart, Germany

[2] Department of Horticulture, Faculty of Agriculture, University of Suez Canal, Ismailia (41522), Egypt

Corresponding author: karl.schmid@uni-hohenheim.de

* Equal contribution of both authors.

This paper is submitted to Theoretical and Applied Genetics Journal

4.1 Abstract

Key message: Genome-wide association and genomic prediction studies with a sample of 174 genetically diverse accessions of cauliflower for six curd-related traits and >120,000 SNP markers demonstrated that these methods are useful for the characterization of genetically diverse cauliflower genotypes for breeding purposes.

Abstract: Genome-wide association studies (GWAS) and genomic prediction are established methods to detect the genetic basis of phenotypic variation and to predict genomic breeding values of complex traits in crop plants. The goal of our study was to explore the potential of GWAS and genomic prediction for improving yield-related traits in cauliflower (*Brassica oleracea var. botrytis)*. We genotyped a collection of 174 randomly selected cauliflower genebank accessions with genotyping-by-sequencing (GBS) and measured six curd-related traits at two locations and three growing seasons. The genome-wide association analysis based on 120,693 SNPs and five different statistical models with imputed and non-imputed data was performed. The potential for genomic prediction was assessed with a random regression best linear unbiased prediction model (RRBLUP) and BayesB with non-imputed data and imputed data. We identified 24 significant associations for curd-related traits and obtained prediction abilities in the range from 0.10 to 0.66 for different traits. The prediction abilities of RRBLUP and BayesB were not significantly different. Data imputation had no significant effect on the prediction ability or accuracy and there was no significant difference between imputation methods. The results suggest that for traits with a sufficiently high heritability, GWAS and genomic prediction in combination with GBS and phenotyping can be applied to genetically diverse genebank materials to identify useful quantitative trait loci (QTL) and genotypes for utilization as genetic resources in cauliflower breeding.

Key words: Cauliflower, Genome-wide association, Genomic prediction, Genotyping by sequencing, RRBLUP, BayesB, Data imputation

4.2 Introduction

Yield improvement is one of the main goals in the breeding of cauliflower, *Brassica oleracea* var. *botrytis* (Singh et al. 2013). It can be achieved by improving traits with direct or indirect effects on yield, which is commonly measured as curd weight. Yield-related traits are curd width and number of days to maturity. They are easily measured and are positively correlated with curd weight (Sheemar et al. 2012; and Jindal and Thakur 2003).

Uncovering the genetic basis of yield and related traits by genetic mapping will contribute to the breeding of cauliflower varieties with higher yields. The main goal of genetic mapping is to identify markers in close linkage to genes controlling complex quantitative traits with methods like linkage or association mapping. In cauliflower, linkage mapping was employed to characterize QTL for curd traits. For example, Lan and Paterson (2000) identified 86 QTLs controlling eight curd-related traits in three segregating populations. Linkage mapping is widely used, but it is time-consuming to develop the mapping populations and its resolution is poor for detecting QTLs because of the low number of recombination events that occur in bi-parental mapping populations (Stich and Melchinger 2010). Association mapping overcomes some of these limitations because it allows higher mapping resolution, includes more diverse populations and requires less time for QTL detection (Stich and Melchinger 2010; Flint-Garcia et al. 2003). Genome-wide association studies (GWAS) were successfully carried out in the major cereal crops, like maize (Li et al. 2013), rice (Norton et al. 2014), barley (Cai et al. 2013) and wheat (Edae et al. 2014). In *Brassicaceae*, GWAS were mainly conducted in rapeseed (*Brassica napus)* to dissect the genetic basis of disease resistance (Jestin et al. 2011), seed oil content and quality (Zou et al. 2010; Rezaeizad et al. 2011), seed weight and quality (Li et al. 2014), seed glucosinolate content (Hasan et al. 2008) and several morphological and phenological traits (Cai et al. 2014). To our knowledge, GWAS have not yet been conducted in *Brassica oleracea*.

In addition to genetic mapping, genomic selection (GS) as described by Meuwissen et al. (2001) has become an important method in plant breeding (Schmid and Thorwarth 2014). GS is based on the prediction of breeding values based on marker information alone (Meuwissen et al. 2001). A training population with genotypic and phenotypic information is used to estimate marker effects from a statistical model, which is then applied to calculate the breeding values of the potential selection candidates in the breeding population (Heffner et al. 2009). Cross-validation

allows the best model to be selected for a certain population, trait, and genetic architecture (Crossa et al. 2010). Genomic selection provides several benefits in both animal and plant breeding (Hayes et al. 2009, Heffner et al. 2009, Jannink et al. 2010), including more rapid breeding cycles and fewer field trials leading to increased genetic gain per time unit at a lower cost (Schaeffer 2006; König et al. 2009, Heffener et al. 2010). Genomic prediction was successfully applied in wheat (Poland et al. 2012b), maize (Crossa et al. 2013) and soybean (Jarquin et al. 2014). In *Brassicaceae*, one study investigated the potential of genomic selection in rapeseed (Würschum et al. 2014), and concluded that it could be a powerful method for rapeseed breeding.

Advances in sequencing technology allow the detection of single nucleotide polymorphisms (SNPs) from large and diverse germplasm collections (Crossa et al. 2013). Genotyping-by-sequencing (GBS) generates tens of thousands of molecular markers at low cost (Elshire et al. 2011; Poland et al. 2012a; Sonah et al. 2013) and is an effective tool for many applications in plant genetics and breeding (Poland et al. 2012b; Fu et al. 2014; Tardivel et al. 2014). GBS was successfully applied to study genomic selection in wheat, maize and soybean (Poland et al. 2012b; Crossa et al. 2013; Jarquin et al. 2014), and in GWAS studies for morphological traits and flavonoid pigmentation in maize and sorghum, respectively (Romay et al. 2013; Morris et al. 2013). One disadvantage of GBS is a high proportion of missing data, although imputation may overcome this problem (Poland and Rife 2012). However, the effects of imputation on the quality of association mapping and genomic prediction are still disputed (Marchini et al. 2007; Poland et al. 2012b; Rutkoski et al. 2013; Jarquin et al. 2014).

Over the last century, thousands of genotypes, such as old and new varieties, landraces and breeding stocks have been collected and are now preserved in *ex situ* genebanks. This material has great potential for association mapping and genomic prediction because of the high genetic variation developed through adaptation to different regions of the world. It will therefore be useful to use cauliflower *ex situ* genetic resources for association mapping and genomic prediction, as these methods can be directly applied to natural populations as well as collections of old cultivars, or breeding material for mapping interesting QTLs or to select genotypes for future breeding .

In this context, the main objectives of this study have been: (1) to identify SNP markers that are associated with phenotypic variation in curd-related traits using GWAS, (2) to quantify the predictive ability of genomic prediction models for curd-related traits in diverse cauliflower genotypes obtained from *ex situ* genebanks and (3) to evaluate the effect of data imputation on association mapping results and genomic prediction accuracy.

4.3 Material and Methods

4.3.1 Plant materials and phenotyping

A total of 192 cauliflower accessions representing a wide range of morphological diversity and geographical origins were obtained from the genebanks of the United States Department of Agriculture (USDA), USA and from the Leibniz-Institut für Pflanzengenetik und Kulturpflanzenforschung (IPK) Gatersleben, Germany. Detailed information about accessions is given in Table S4.1. All accessions were phenotyped for curd-related traits in replicated field trials at two locations (experimental stations: Heidfeldhof and Kleinhohenheim in Stuttgart, Germany), for three successive growing seasons (June 2011, April 2012 and August 2012). The field experiment was conducted in randomized complete block design (RCBD) with two replications as described by Yousef et al. (2015). Five ripened curds were harvested from each plot and used to measure six traits that reflect various aspects of curd development and morphology according to Lan and Paterson (2000):

1. Curd width (cm): the width of the curd.
2. Cluster width (cm): the width of the largest floral cluster.
3. Number of branches: number of branches within the curd that originated from the main stem.
4. Apical shoot length (cm): stem length from the apical meristem to where the closest first-rank branch originated from the main stem.
5. Nearest branch length (cm): length of the branch that is nearest to the apical meristem.
6. Days to budding: number of days from planting to appearance of the first floral bud.

4.3.2 Genotyping

Genomic DNA was isolated from leaf tissue sampled three weeks after sowing from a single plant of each genotype with a CTAB protocol (Sagahi-Maroof et al. 1984). Genotyping-by-sequencing (GBS), including DNA digestion with the *ApekI* restriction enzyme, adapter ligation and PCR amplification was performed according to Elshire et al. (2011). A total of 96 barcodes were used (Supplementary Table S4.2), of which 64 were designed with a web-based tool (www.deenabio.com/services/gbs-adapters) and 32 barcodes were taken from Elshire et al. (2011). The 192 genotypes were divided into two libraries, each consisting of 96 genotypes tagged with a different barcode. Sequencing of 100 bp reads was performed on two lanes of Illumina HiSeq 1000 at the Kompetenzzentrum Fluoreszente Bioanalytik (KFB), Regnesburg, Germany.

After sequencing, reads were filtered for sequencing artifacts and low quality reads with Python scripts and the Burrows-Wheeler Aligner software (BWA; Li and Durbin 2009) and FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Eighteen genotypes with <300,000 reads were excluded from further analysis, which resulted in a total sample of 174 accessions for further analyses. The pre-processed reads were aligned to the genome of *Brassica oleracea sp. capitata* (Liu et al. 2014), using BWA (Li and Durbin 2009). SNP calling was performed with SAMtools (Li et al. 2009), bcfutils, vcfutils and custom Python scripts. The .vcf file was parsed to retain only SNP positions with a coverage of >=30, and >=10 reads confirming the variant nucleotide. In the end, 120,693 SNPs were detected with 19.02% to 76.73% of missing values per genotype (Yousef et al., in reversion).

4.3.3 Analysis of phenotypic variation

The effects of the genotype and environment (environment was treated as the combination of location and season) and their interactions on phenotypic variation were evaluated by using analysis of variance (ANOVA) as implemented in the *aov* function in R (R Development Core Team, 2014). Details about the phenotypic analysis can be found in Yousef et al. (2015). A mixed-effect model was fitted using restricted maximum likelihood (REML) with the *lmer* function from the R package *lme4* (version 1.0-5., Bates et al. 2013) with the model equation:

$$y_{ij} = \mu + G_j + E_i + GE_{ji} + e_{ij} \quad (1)$$

where μ being the overall mean, G_j the effect of the j^{th} genotype and E_i the effect of the i^{th} environment and GE_{ji} the environment x genotype interaction. In this model, all effects were considered random, whereas only the overall intercept was treated as a fixed effect. Variance components of this model were used to calculate broad sense heritability for each trait according to Nyquist (1991) as:

$$H^2 = \frac{\sigma_g^2}{\sigma_g^2 + \left(\frac{\sigma_{ge}^2}{e}\right) + \left(\frac{\sigma_e^2}{re}\right)} \quad (2)$$

where H^2 represents the heritability, σ_g^2 the genetic variance, σ_{ge}^2 the genotype-by-environment variance, σ_e^2 the error variance, *r* the number of replications, and *e* the the number of environments, respectively. Best linear unbiased predictors (BLUPs) were extracted from model (1) and estimated by the random effect G_j. BLUPs were used to estimate the genetic correlation (r_G) among all traits. The genetic and phenotypic correlation coefficients are based on the Pearson correlation coefficient. A *t*-test was used to check for significant differences between the IPK and USDA accessions for all phenotypic traits.

4.3.4 Population structure

The genetic structure of the collection was analyzed by multivariate, phylogenetic and model-based methods (Yousef et al., in reversion) as follows: (1) Principal component analysis (PCA) as implemented in the R package *adegenet* (Jombart and Ahmed, 2011), (2) Principal coordinate analysis (PCoA) based on pairwise Fst values between genotypes as implemented in the *ape* R package (Paradis et al. 2004), (3) Discriminant analysis of principal components (DAPC) with the *dapc* function of the R package *adegenet* (Jombart and Ahmed, 2011), (4) Neighbor-joining tree based on a pairwise distance matrix as performed in *ape* R package (Paradis et al. 2004), (5) STRUCTURE 2.3.4 software (Pritchard et al. 2000). The data set was tested for a numbers of

subpopulations ranging from *K*=1-10 with ten runs for each *K* with a burn-in length of 50,000 followed by 50,000 iterations, (6) ADMIXTURE analysis (Alexander et al. 2009) and cross-validation was used to estimate the optimum number of clusters *K* (Alexander and Lange, 2011).

4.3.5 Association analysis

GWAS was performed with five different statistical methods that included: A generalized linear model (GLM) without correction for population structure (*Q)* and kinship (*K) (*naïve GLM), with correction for only *Q*, which is estimated by EIGENSTRAT (Price et al. 2006) and a model with correction for Q and K. Additionally, a linear mixed model that used a kinship matrix *K* only was calculated with EMMA (Kang et al. 2010). Further we tested a Multi Locus Mixed Model (MLMM; Segura et al. 2012). GLM, EIGENSTRAT and *Q+K* analyses were carried out using the R package *GenABEL* (GenABEL project developers 2013). The EMMA analysis was performed using the Efficient Mixed-Model Association eXpedited (EMMAX) software (Kang et al. 2010) and MLMM was analyzed using R scripts available at https://github.com/Gregor-Mendel-Institute/mlmm.

4.3.6 Control of false positive rates

Confounding effects due to population structure were visually assessed with quantile-quantile (*qq*) plots. In *qq*-plots the estimated $-\log_{10}(p)$ values are displayed against the expected values following a χ^2 distribution with one degree of freedom. The model whose observed $-\log_{10}(p)$ values are closest to the expected $-\log_{10}(p)$ values control more efficiently for the occurrence of type I and type II errors, which are expected to be more frequent in a structured population. The inflation factor λ was calculated as the ratio of the median of the empirically distribution of the data to the expected median, where values close to 1 indicate no inflation.

The significance threshold was set for the non-imputed and imputed data for each trait separately using the false discovery rate (FDR; Benjamini and Hochberg 1995), where FDR values are computed from the *p*-values. The FDR was set to 0.2 for all data sets.

4.3.7 Genomic prediction

A random regression best linear unbiased prediction (RRBLUP; Meuwissen et al. 2001) model, implemented in the *synbreed* R package (Wimmer et al. 2012) was used to estimate the marker effects and to calculate the prediction ability. The model is of the form:

$$y = X\beta + Wm + e \quad (3)$$

as described by Wimmer et al. (2012), with X as the design matrix, the fixed effects β, W the $n*p$ matrix of markers, m a vector of the marker effects with $m \sim N\left(\mathbf{0}, I\sigma_m^2\right)$ and the term for the residuals e.

Additionally, a BayesB (Meuwissen et al. 2001) model as implemented in the *BGLR* R package (de los Campos and Rodriguez 2014) was used to estimate the marker effects and to calculate the prediction ability. BayesB was chosen over other Bayesian models such as BayesA and BayesC, because it assumes an *a priori* distribution of marker effects following a mixture distribution with point mass at zero and a scaled-*t* slab similarly to BayesA, and it can utilize both shrinkage and variable selection methods, similar to BayesC (de los Campos and Rodriguez 2014). The hyperparameters were chosen according to the default values in BGLR.

The prediction ability, defined as the correlation between observed phenotypic and predicted genotypic values ($cor\left(y,\hat{g}\right)$), was assessed via five-fold cross-validation with ten replications for each trait, respectively. The prediction accuracy was expressed as $cor\left(y,\hat{g}\right)/h$ where h is the square root of heritability (Lande and Thompson 1990; Dekkers 2007). A heritability of one would result in the most accurate prediction, which then equals the prediction ability. We used a *t*-test to check if there were significant differences in prediction ability and accuracy between imputed and non-imputed data, as well as between the imputation methods and between the different models.

To compare the genotypic effects of SNPs calculated with RRBLUP and BayesB as well as the association result, the SNPs with the highest effect for each model were compared to the significant markers detected in the GWAS.

4.3.8 Marker imputation

We used two methods, fastPHASE (Scheets and Stephens; 2006) and BEAGLE (Browning and Browning 2007) to impute missing genotypes. The two imputation methods use a Hidden Markov Model (HMM) to cluster haplotypes, but differ in the underlying model. FastPHASE uses an Expectation-Maximization (EM) algorithm to estimate parameters for cluster configurations, whereas BEAGLE uses empirical frequencies as parameters. In addition, fastPHASE fixes the number of haplotype clusters in the model, but BEAGLE allows the cluster number to be changed at each locus for a better fit to the localized linkage disequilibrium (Pei et al. 2008). We decided to use fastPHASE as the imputation method, because we expected it to perform better than BEAGLE with our data which is characterized by a low LD, small sample size and high marker density. However, we included BEAGLE for comparison for some analyses, but used only fastPHASE imputed data for the GWAS to keep the number of analyses manageable.

GWAS was carried out with 120,693 SNP markers for all traits using the mean values of three growing seasons and two locations. First, SNP markers with minor allele frequency (MAF) <0.05 and missing values were excluded from further analysis, which resulted in a total of 675 markers (non-imputed data set). Second, SNP alleles were imputed with fastPHASE (Scheets and Stephens; 2006) and markers with a MAF <0.05 were excluded, which resulted in a total of 64,372 SNPs (imputed data set with fastPHASE). Genomic prediction was conducted with a third data set, in which SNP alleles were imputed with BEAGLE 4 (Browning and Browning 2007) and markers with a MAF <0.05 were excluded, which resulted in a total of 62,566 SNPs (imputed data set with BEAGLE). The distribution of non-imputed and imputed SNPs is shown in Figure S4.1, Figure S4.2 and Figure S4.3.

4.3.9 Linkage disequilibrium

Linkage disequilibrium (LD) between markers was calculated with the *LDcorSV* R package (Mangin et al. 2012). Besides the well-known definition of LD as correlation between alleles at two loci (r^2), Mangin et al. (2012) introduced a corrected measure for the LD in structured populations (r^2_S) by adding a Bernoulli random variable for population assignment, which accounts for the effect that r^2 between different populations should be zero for unlinked loci. A

further method corrects for kinship within populations (r^2_V) with a variance-covariance matrix and a combination of the correction for population structure and kinship denoted as r^2_{VS} to correct for confounding effects.

Data availability

The phenotypic data and the aggregated genotypic data are available under DataDryad DOI: (WILL BE ASSIGNED AFTER ACCEPTANCE OF THE MANUSCRIPT)

4.4 Results

4.4.1 Phenotypic analysis of the six yield-related traits

The 174 lines were evaluated for six curd-related traits at two locations over three consecutive seasons. A large phenotypic variation was observed for all traits (Yousef et al. 2015). Number of days to budding varied 2.6 fold, ranging from 45 to 118 days with an average of 75.23 ± 11.68 cm. Curd width showed a 1.4-fold change, ranging from 11.11 to 15.29 cm with an average of 13.52 ± 0.73 cm (Table 4.1). The distribution with the means of the six curd-related traits over two locations and three growing seasons is shown in Figure S4.4.

Analysis of variance showed that all traits were strongly affected by genotype (G), environment (E) and genotype by environment interaction (G × E; $P < 0.001$; Table 4.1). Broad-sense heritability (H^2) differed strongly between traits (Table 4.1). Two traits, cluster width and number of days to budding, had moderate and high heritability (56% and 94%), indicating that these traits are stably inherited. Large positive phenotypic ($r_p = 0.69$) and genotypic ($r_g = 0.59$) correlations were observed between curd width and cluster width, respectively (Table 4.2), as well as between apical length and nearest branch length ($r_p = 0.79$ and $r_g = 0.71$). The number of days to budding showed negative phenotypic ($r_p = -0.22$) and genotypic correlations ($r_g = -0.23$) with number of branches.

Table 4.1 Descriptive statistics and broad sense heritability of six curd-related traits.

Trait	Mean ±SD	Range	G	E	GxE	H^2 (%)
Curd width	13.52±0.73	11.11-15.29	***	***	***	0.437
Cluster width	4.38±0.39	2.86-5.28	***	***	***	0.564
No. of branches	11.17±0.59	9.71-13.40	***	***	***	0.264
Apical length	1.19±0.14	0.74-1.65	***	***	***	0.111
Nearest branch	1.68±0.17	1.19-2.23	***	***	***	0.05
No. of days	75.23±11.68	45.57-118.43	***	***	***	0.943

*** Significant at $P < 0.001$ for the effect of genotype (G), environment (E) and genotype by environment interaction (GxE) on phenotypic variance estimated by two-way ANOVA.
H^2: broad-sense heritability.

Table 4.2 Phenotypic and genotypic correlation among six curd related-traits.

	Curd Width	Cluster width	No. of branches	Apical length	Nearest branch	No. of days
Curd width		0.59	0.004	0.27	0.24	0.04
Cluster width	0.69***		0.065	0.25	0.32	-0.21
No. of branches	-0.27***	-0.16*		-0.27	-0.21	-0.23
Apical length	0.64***	0.42***	-0.41***		0.71	-0.03
Nearest branch	0.50***	0.50***	-0.37***	0.79***		-0.07
No. of days	0.09	-0.13	-0.22**	0.005	0.11	

The values below and above the diagonal represent phenotypic correlation coefficient based the means of the 2 locations and three seasons and genetic correlation coefficient based on the trait values in 2 locations and 3 seasons.

** Significant at $P < 0.01$, *** Significant at $P < 0.001$

4.4.2 Model comparison to control for false associations

A previous analysis of genotyping data of accessions used indicated the presence of a strong population structure of two main groups on the genetic level that reflect the two genebanks from which the seeds were obtained (Yousef et al., in reversion). Also, there was significant differentiation between the two genebanks in three of the six morphological traits (curd width, cluster width and number of days to budding: $p< 0.001$; Figure S4.5). Therefore, GWAS were performed for the six yield-related traits using five models that differed with respect to

controlling for population structure and kinship. In all traits, observed *p*-values for the GLM markedly deviated from the expected *p*-values, followed by the EIGENSTRAT and *Q*+*K* models with the two data sets (Figure 4.1 and 4.2, respectively). The lambda values for all models are in Supplementary Table S4.3. The observed distribution of *p*-values for EMMAX and MLMM were closer to the expected distribution of *p*-values with both data sets than with the other models (Figure 4.1 and 4.2), indicating that these models efficiently controlled for false positive associations. For this reason we selected both models for the GWAS.

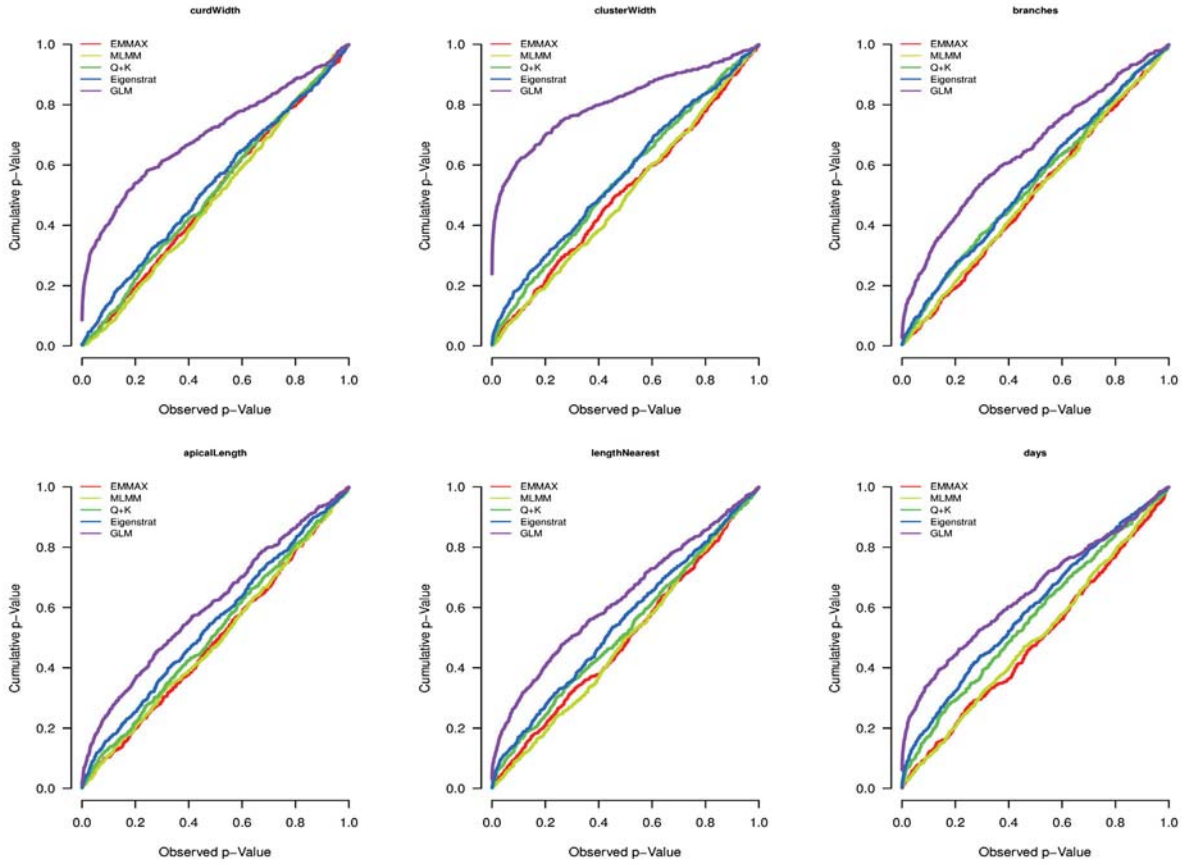

Figure 4.1 Quantile–quantile plots of estimated P value vs. cumulative P value from association analysis of six yield-related traits for non-imputed data. The purple, blue, green, yellow and red lines represent observed P values using the GLM, EIGENSTRAT, Q+K, MLMM and EMMAX models, respectively.

4.4.3 GWAS of six yield-related traits

Overall, 24 markers were significantly associated with the six curd-related traits (Table 4.3). With EMMAX, 6 of the 675 non-imputed SNPs were associated with the four traits: curd width, cluster width, number of branches and number of days to budding (Figure 4.3; Table 4.3). Only 3 SNPs of the imputed SNPs were associated with the number of days to budding trait (Figure 4.4; Table 4.3). With MLMM, 8 SNPs associated with number of branches, nearest branch and number of days to budding (Figure S4.6; Table 4.3). Seven SNPs of the imputed data were associated with the two traits: apical length and number of days (Figure S4.7; Table 4.3).

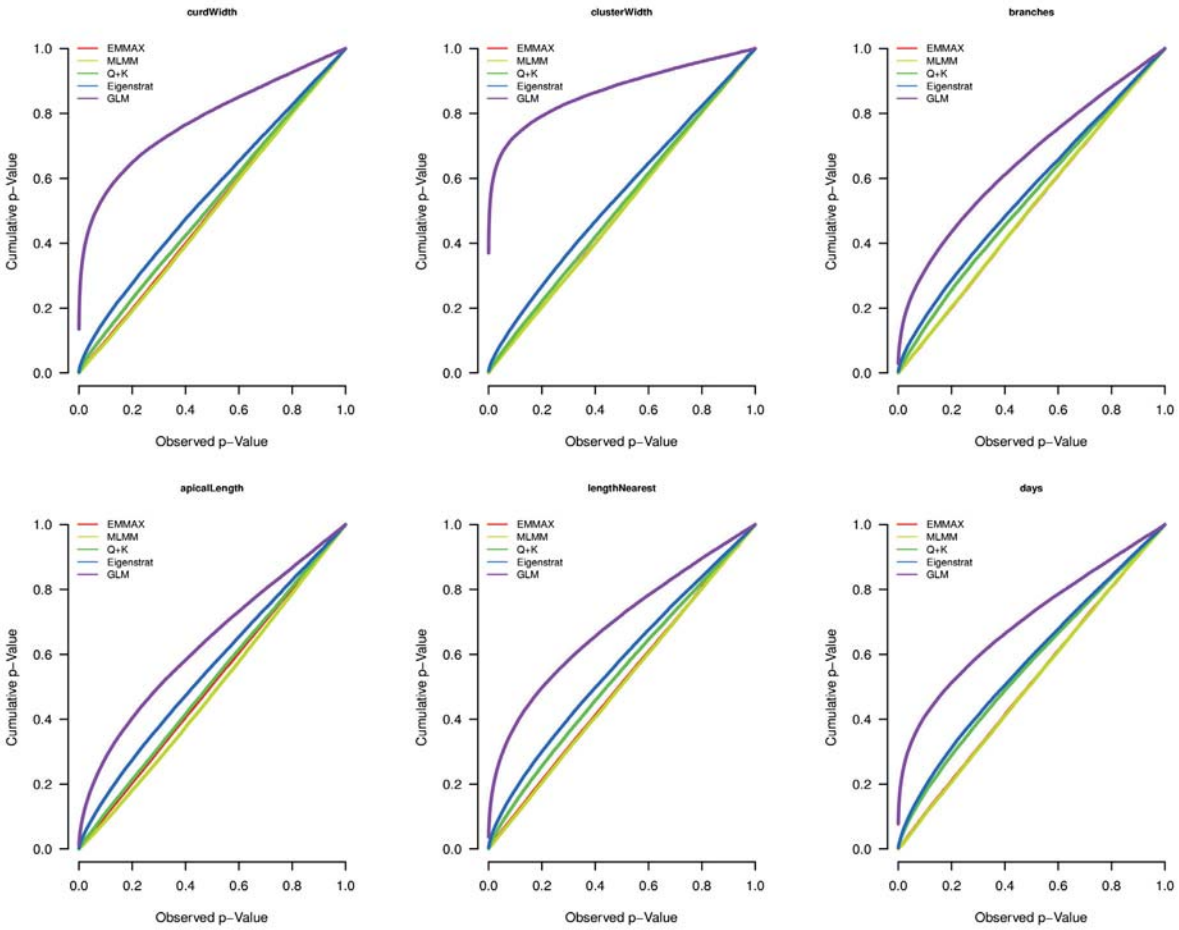

Figure 4.2 Quantile–quantile plots of estimated P value vs. cumulative P value from association analysis of six yield-related traits for imputed data. The purple, blue, green, yellow and red lines represent observed P values using the GLM, EIGENSTRAT, Q+K, MLMM and EMMAX models, respectively.

Table 4.3 SNP markers significantly associated with six curd-related traits.

	EMMAX						MLMM					
	Non-imputed data	RRBLUP Rank	BayesB Rank	Imputed data	RRBLUP Rank	BayesB Rank	Non-imputed data	RRBLUP Rank	BayesB Rank	Imputed data	RRBLUP Rank	BayesB Rank
Curd width	C02:5063181	6	4									
	C02:3528844	1	1									
Cluster width	C02:5063181	8	4									
No. of branches	C06:2323306	2	2				C06:2323306	2	2			
	C07:41524584	1	1				C07:41524584	1	1			
Apical length										C03:20986662	2397	1712
										C06:4914166	4902	1785
										C06:34494415	2111	1194
										C02:26787029	30559	29378
Nearest branch							C09:25012587	2	2			
No. of days	C01:37688065	1	1	C07:936770	1	1	C02:2708182	7	4	C07:936770	1	1
				C06:2949314	1861	572	C02:2708163	6	5	C06:2949314	1861	572
				C07:936738	2	5	C02:2708156	5	6	C07:936738	2	5
							C07:41524584	2	3			
							C01:37688065	1	1			

Rank indicates to the SNP effect by RRBLUP and BayesB.
Italic are common markers detected with the two methods (EMMAX and MLMM).
Bold italic are common markers detected in two traits.

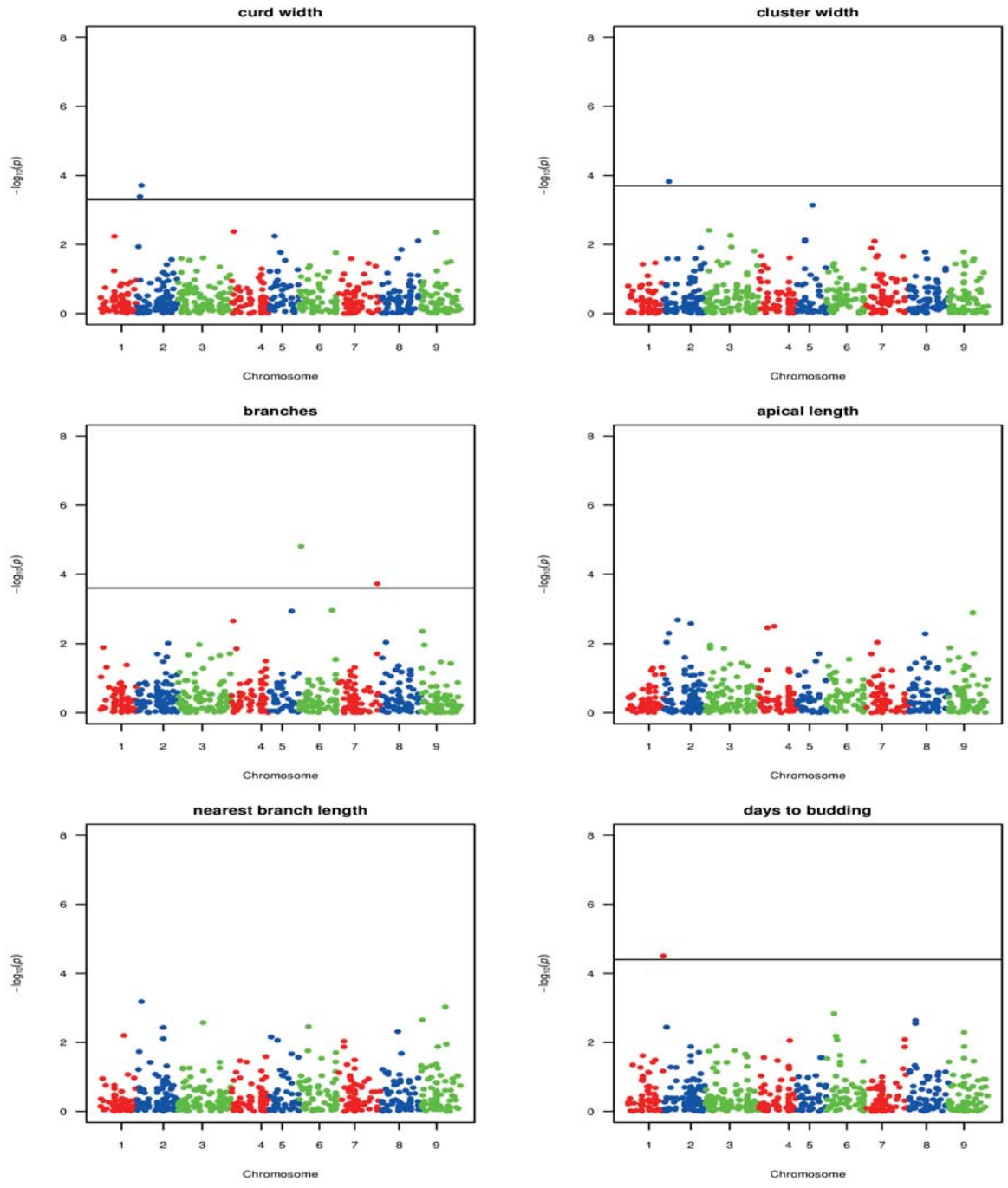

Figure 4.3 Manhattan plots of association analysis using 675 SNPs and the EMMAX method for six curd-related traits for imputed data. Each dot represents a SNP. The horizontal line represents significance threshold with FDR.

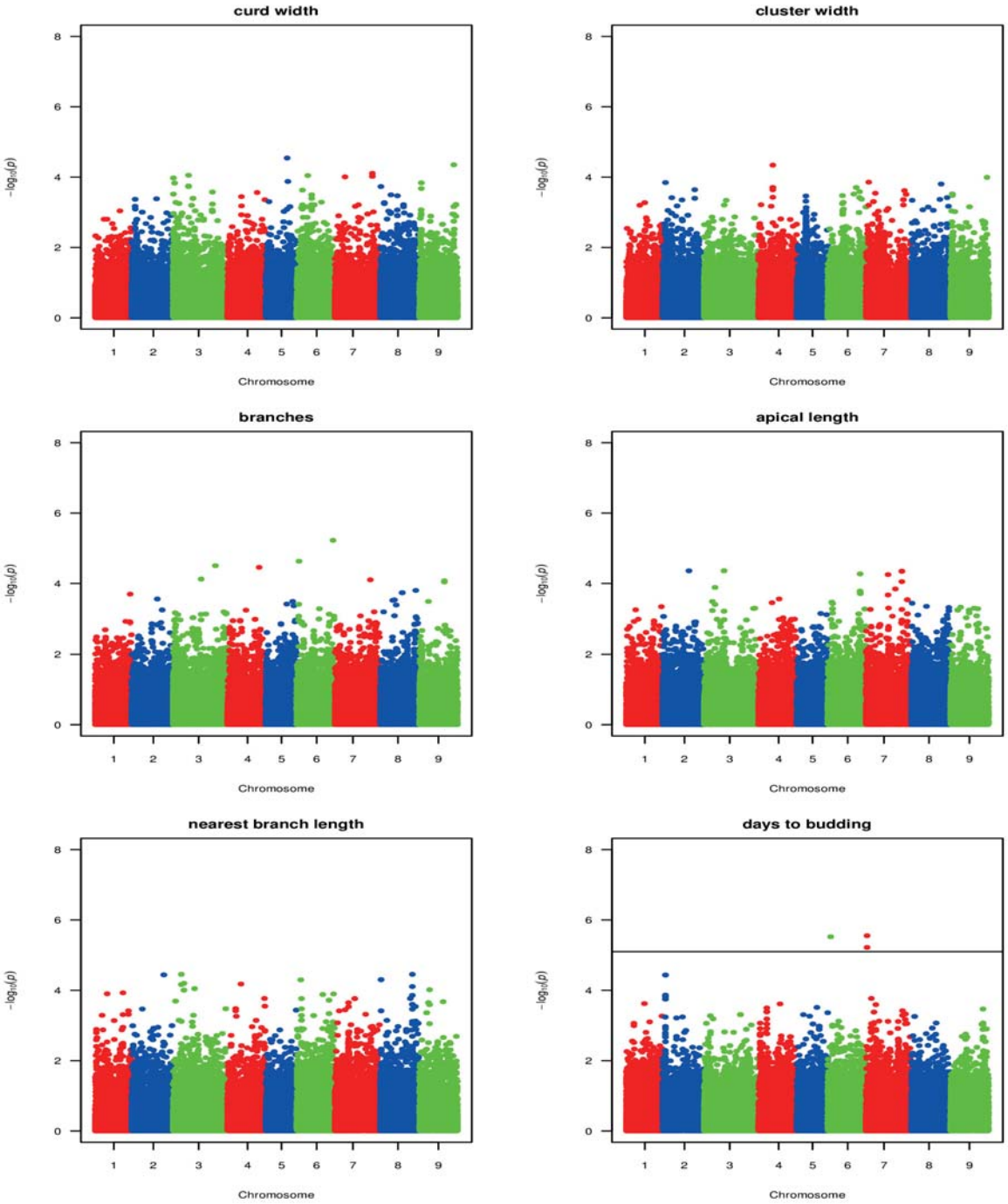

Figure 4.4 Manhattan plots of association analysis using 64,372 SNPs and the EMMAX method for six curd-related traits for imputed data. Each dot represents a SNP. The horizontal line indicates the significance threshold with FDR.

Out of the 24 SNPs significantly associated with the six traits, six SNP were detected by both EMMAX and MLMM (Table 4.3). Moreover, one SNP (C02:5063181) was associated with both curd width and cluster width. Another SNP (C07:41524584) was associated with number of days and number of branches (Table 4.3). Taken together, the results indicate that the imputation did not increase the number of significant associations and that the SNPs identified as being associated differed with the two methods used.

4.4.4 Genomic prediction with RRBLUP and BayesB

Prediction ability was assessed using five cross-validations with ten replications. The prediction ability ranged from 0.128 to 0.652 with RRBLUP (Table 4.4). For all traits, prediction abilities were slightly larger with imputed than non-imputed markers. The prediction ability of the BEAGLE was similar to the fastPHASE imputation for all traits (Table 4.4).

Table 4.4 Prediction ability and accuracy for six curd-related traits with different data sets using RRBLUP.

Trait	Non-imputed data		Imputed data (BEAGLE)		Imputed data (fastPHASE)		Trait Mean	
	Ability	accuracy	ability	accuracy	ability	accuracy	ability	accuracy
Curd width	0.378	0.572	0.448	0.678	0.448	0.678	0.425	0.642
Cluster width	0.620	0.826	0.647	0.862	0.652	0.868	0.646	0.852
No. of branches	0.335	0.652	0.384	0.747	0.382	0.743	0.367	0.714
Apical length	0.128	0.384	0.134	0.402	0.140	0.420	0.134	0.402
Nearest branch	0.221	0.988	0.273	1.221	0.282	1.261	0.259	1.157
No. of days	0.631	0.650	0.632	0.651	0.638	0.657	0.634	0.653
Mean	**0.386**	**0.679**	**0.420**	**0.760**	**0.424**	**0.771**	**0.409**	**0.737**

With BayesB, the genomic prediction ability ranged from 0.090 to 0.660 (Table 4.5). Imputation resulted in slightly higher prediction abilities for all traits except number of days to budding, and the prediction ability of fastPHASE was slightly higher than with BEAGLE imputed data for all traits except number of days to budding and apical length (Table 4.5).

An overview of SNP effects calculated by RRBLUP and BayesB is provided in Figure S4.8-S4.13. For comparison we ranked marker effects calculated with RRBLUP and BayesB in descending order and compared them to the 24 most significant associations detected by EMMAX and MLMM. Of these, 19 SNPs also produced the largest marker effects and were among the top eight SNPs with the highest *p*-values in the GWAS (Table 4.3).

Table 4.5 Prediction ability and accuracy for six curd-related traits with different data sets using BayesB.

Trait	Non-imputed data		Imputed data (BEAGLE)		Imputed data (fastPHASE)		Trait Mean	
	Ability	accuracy	ability	accuracy	ability	accuracy	ability	accuracy
Curd width	0.354	0.536	0.396	0.599	0.444	0.672	0.398	0.602
Cluster width	0.604	0.804	0.641	0.854	0.660	0.879	0.635	0846
No. of branches	0.378	0.736	0.353	0.687	0.414	0.806	0.382	0.743
Apical length	0.090	0.270	0.120	0.360	0.104	0.312	0.105	0.314
Nearest branch	0.230	1.029	0.278	1.243	0.285	1.275	0.264	1.182
No. of days	0.660	0.680	0.658	0.678	0.614	0.632	0.644	0.663
Mean	**0.386**	**0.676**	**0.408**	**0.737**	**0.420**	**0.762**	**0.405**	**0.725**

4.4.5 Linkage disequilibrium analysis

In the analysis of LD, means of r^2 and r^2_{VS} were 0.029 and 0.010 for non-imputed data, respectively. Due to very long running times, we did not calculate LD for the imputed data. The r^2 and r^2_{VS} for each chromosome are presented in Table 4.6. The r^2 values were much larger than r^2_{VS} values. Some pairs of markers with significant associations in the GWAS also showed high LD levels. For example, SNPs C02:2708182, C02:2708163, C02:2708156 are in complete linkage with both measures of LD. Overall, LD levels between markers were low and decreased rapidly with increasing distance which reflects the out-crossing nature of cauliflower (Figure 4.5 and 4.6).

Table 4.6 The LD for each chromosome with two data sets.

Chromosome	r^2	r^2_{VS}
Chr 1	0.0268	0.0100
Chr 2	0.0193	0.0082
Chr 3	0.0293	0.0097
Chr 4	0.0209	0.0067
Chr 5	0.0324	0.0139
Chr 6	0.0275	0.0103
Chr 7	0.0248	0.0091
Chr 8	0.0459	0.0130
Chr 9	0.0309	0.0097
Mean	**0.0287**	**0.0101**

4.5 Discussion

We conducted association mapping and genomic prediction in 174 genotypes of cauliflower reflecting a wide genetic diversity in order to dissect the genetic basis and assess the predictive potential of six curd-related traits.

4.5.1 Identification of significant marker-trait associations with GWAS

GWAS is a powerful method of establishing the contribution of known and novel genes to plant phenotypic diversity (Li et al 2013; Cai et al. 2013; Norton et al. 2014). We compared five GWAS methods using two data sets (non-imputed and imputed SNP data) and detected at total of 24 SNPs that were associated with six curd-related traits. To our knowledge this is the first GWAS study in *Brassica oleracea*, and for this reason we could not compare our results with other studies in this species. A comparison to *B. napus* was also not possible, because the traits we analyzed are specific for cauliflower. The genetic base of curd-related traits has been thoroughly dissected in *B. oleracea,* and several QTLs with major effects have been identified using the strategy of linkage mapping (Lan and Paterson 2000). Unfortunately, the results are also not directly comparable to our results as due to the different resolutions of the linkage -map used in their study and the physical map used in this study.

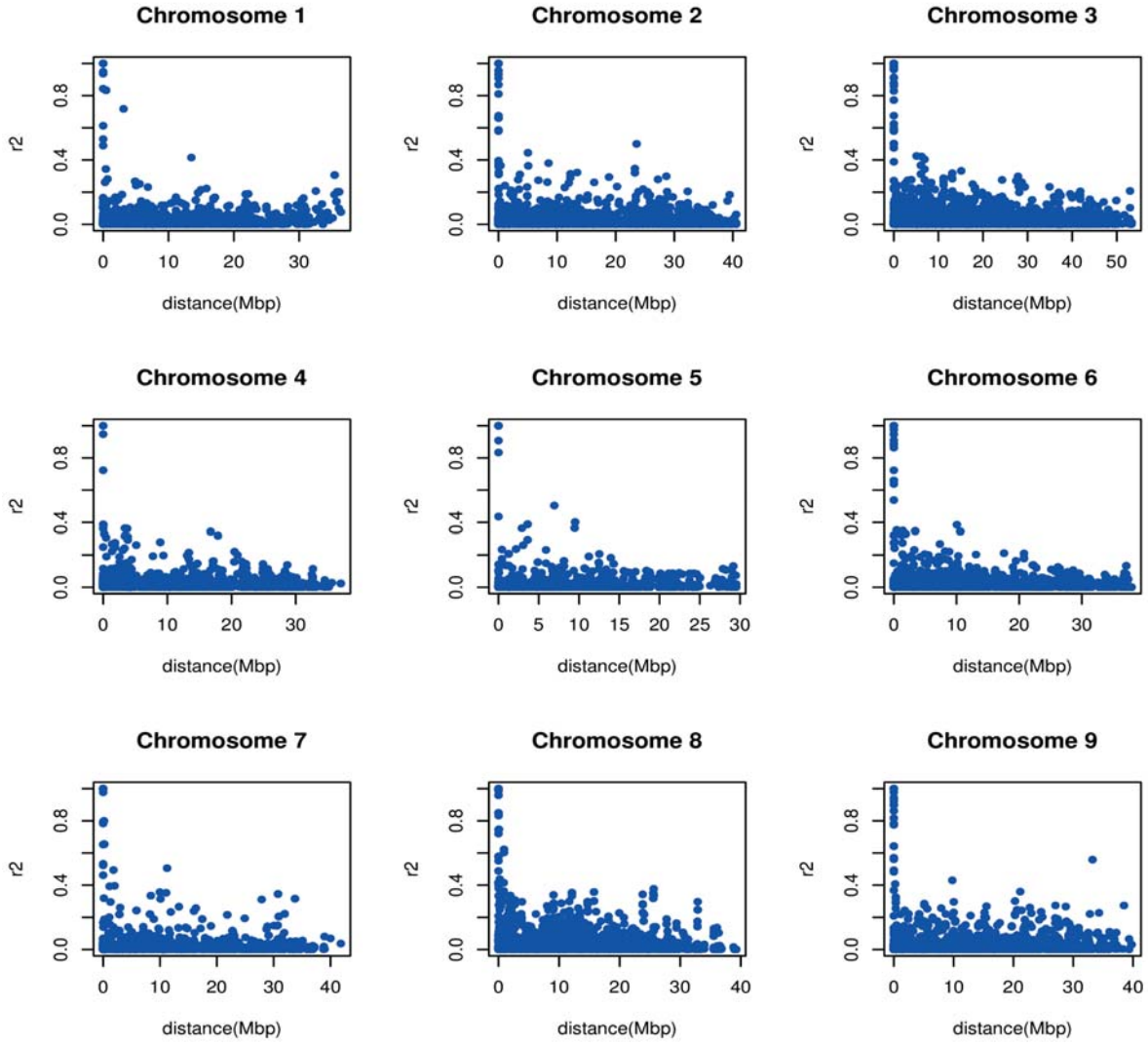

Figure 4.5 Distribution of pairwise LD values (r^2) of SNPs identified in 174 cauliflower accessions.

In association mapping, the presence of population structure leads to spurious marker-trait associations. Since the population genetic analysis indicated the presence of a population structure (Yousef et al., in revision), we compared the performance of the five GWAS methods because they differentially account for population structure and kinship. The genome-wide correction for the confounding effects of kinship and structure was most efficient with EMMAX

and MLMM (Figure 4.1 and Figure 4.2) and we used these two methods for further analysis. Only six of the 24 significant marker-trait associations are shared between EMMAX and MLMM (Table 4.3), which indicates our GWAS strongly depended on the method used, but also suggests that several strong QTLs for the measured traits segregate among the *B. oleracea* genebank accessions, because we identified them with both methods. The number of marker-trait associations differed between EMMAX and MLMM, whereas EMMAX detected 9 significant SNP markers and MLMM 15 significant SNP markers in both the non-imputed and imputed data. This difference between the two methods is caused by MLMM over-correcting for kinship compared to EMMAX, as indicated by the lambda values (Table S4.3), which is defined as the observed median value of the chi-squared statistic divided by the expected median value of the chi-squared statistic for the null markers values (Hinrichs et al. 2009); the lambda values with MLMM model were < 1 for most traits while they were almost one with EMMAX.

4.5.2 Evaluation of genomic prediction ability

We used two genomic prediction models (RRBLUP and BayesB) because of their good performance and stable prediction ability (Meuwissen et al. 2001; Heslot et al. 2012). The number of days to budding was the best trait for prediction as the prediction accuracy was close to the prediction ability (Table 4.4 and 4.5), which was expected, due to the high heritability of this trait (0.94; Table 4.1). The traits apical length and length of nearest branch had the lowest prediction accuracy (Table 4.4 and 4.5), because of the low heritability (0.11 and 0.05; Table 4.1) and low prediction ability (Table 4.4 and 4.5, respectively) that reflect the genetic architecture of these traits. Additionally, the given marker density and the low level of LD in this sample reduced the prediction ability. Moreover, the two traits are not easy to measure and measurement errors may result in biased phenotypic measurements which affect heritability estimates. Prediction accuracies for the traits curd width, cluster width and number of branches were not as accurate as for number of days to budding (Table 4.4 and 4.5) but they show a good prediction ability. Therefore, high quality phenotypic measurements will improve GS for these traits. Prediction accuracies for number of days to budding (0.653 and 0.663; Table 4.4 and 4.5) and curd width (0.642 and 0.602; Table 4.4 and 4.5) are comparable to flowering time (0.70) and grain yield (0.50) in *B. napus* (Würschum et al. 2014). The prediction ability values found in this

study are sufficiently good to suggest the implementation of GS in cauliflower breeding programs.

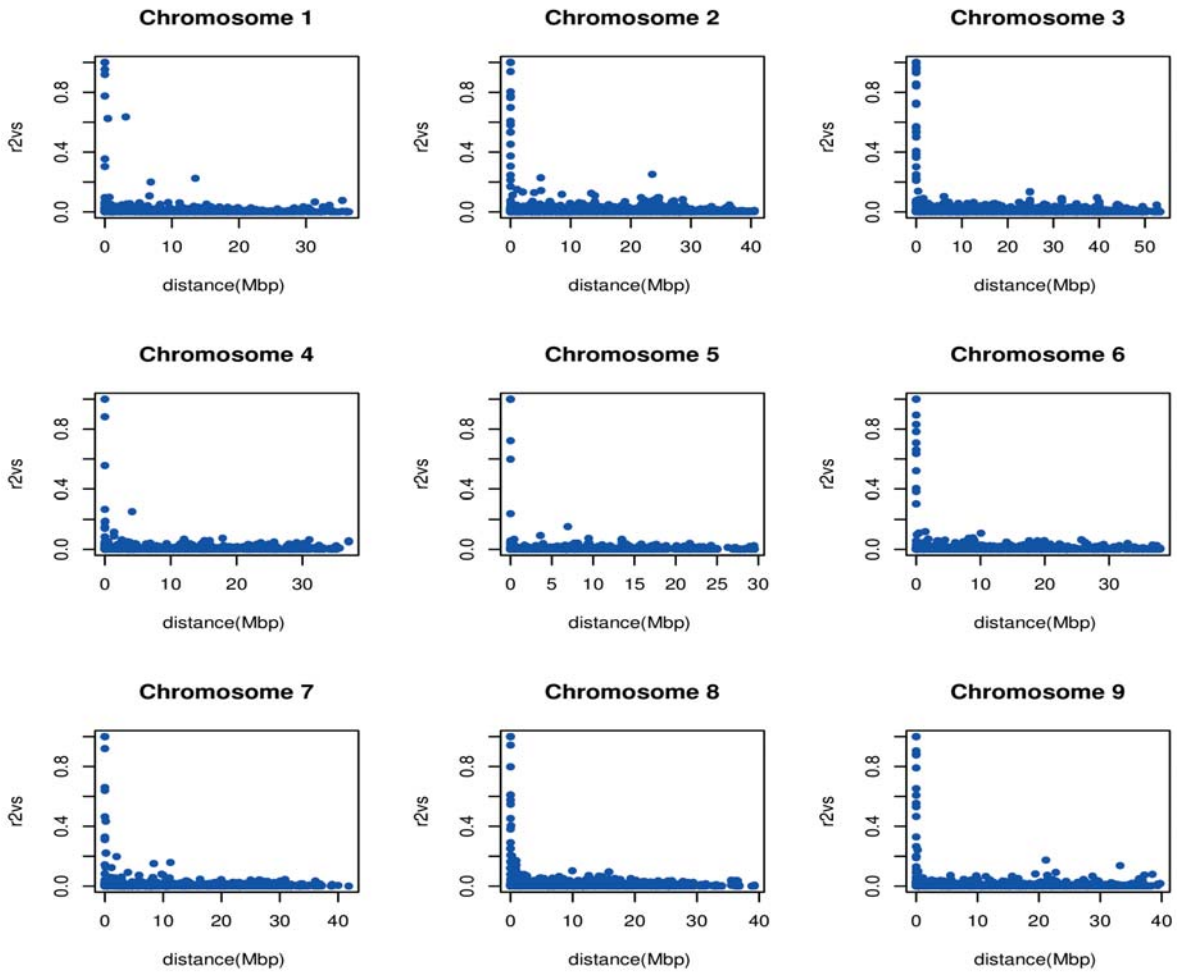

Figure 4.6 Distribution of pairwise LD values corrected for population structure and kinship (r^2_{VS}) of SNPs identified in 174 cauliflower accessions.

There was no significant difference between the RRBLUP and BayesB regarding their mean prediction ability and accuracy over all traits and different data sets, (Table 4.4 and 4.5). In agreement with our results, Bao et al (2014) reported that Bayesian models did not outperform RRBLUP. However, some traits showed a difference between the two models (Table 4.4 and 4.5). These differences in prediction ability are due to model specifications. A key assumption of

RRBLUP is that all markers have an effect that explains the same proportion of the genotypic variance and that these effects are normally distributed (Meuwissen et al. 2001), while in the BayesB model, marker effects follow a mixture distribution with point mass at zero and a scaled-*t* slab with a marker-specific variance parameter s_β. A parameter $pi \sim Beta\left(p_{0,} pi_0\right)$ (de los Campos and Rodriguez 2014) is introduced as the prior proportion of non-zero effects, offering the potential to perform marker selection. Since RRBLUP is computationally less expensive and much simpler to implement than the BayesB model, RRBLUP can be recommended for genomic selection in cauliflower breeding.

A general point to be considered while performing genomic prediction analysis is the size of the training set, because larger training sets improve prediction ability and allow a robust estimation of marker effects. Furthermore, the population structure may influence prediction ability. If a population structure exists in the training and validation sets, a correction for population structure reduces prediction ability (Guo et al. 2014). Based on this conclusion we assume that the predictive potential of our data may be inflated by the presence of a strong population structure among the accessions, which consists of two major pools that reflect the respective origin of the accessions from the USDA or IPK genebanks. However, because the population size in our study was small, a subdivision into smaller groups would not have improved prediction ability within subpopulations. Our analysis provides a first assessment of genomic prediction in diverse cauliflower genebank material, but we conclude that much larger sample sizes would be required for traits with a low heritability and that the presence of a population structure likely has a strong effect on prediction ability, which needs to be considered.

A high proportion of SNPs (19 out of 24) with significant associations to genotypes in the GWAS were also SNPs with the highest marker effects in the genomic prediction with RRBLUP and BayesB (Table 4.3). This result reflects the similarity of the statistical models used for GWAS and genomic prediction and suggests that these SNPs are linked to robust QTLs. A comparison of markers identified by a GWAS and estimated marker effects in genomic prediction can therefore be used as a quick validation of the robustness of detected QTLs.

4.5.3 Imputation effect on GWAS and genomic prediction results

Our set of *B. oleracea* genotypes exhibited a low LD level (0.0287), as expected for an outcrossing species (Flint-Garcia et al. 2003; Wright et al. 2008). The majority of markers were not in high LD with each other, consistent with previous estimates of 0.037 (Deluorme et al. 2013) and 0.0176 (Li et al. 2014) in *Brassica napus*. The low LD requires high marker densities to perform GWAS effectively in *B. oleracea,* which justified the use of imputed data sets to enhance the power of the GWAS analysis (Marchini et al. 2007; Howie et al. 2009; Marchini and Howie 2010). However, our results show that the imputation of missing genotypes does not improve the identification of significant associations (Table 4.3). The number of significant associations was lower (10) with imputed data than with non-imputed data (14) as presented in Table 4.3.

The difference in the number of significant associations between the non-imputed and imputed data sets may be based on the differences in the significant threshold, which is strongly related to the number of markers used in the study. In the case of a GWAS with a quantitative trait we expect to find several QTLs with a small effect. The Bonferroni correction, an often used threshold in GWAS, can lead to a high rate of false negatives (type II errors; Perneger 1998). Therefore, we used the FDR as a threshold, which is conceptually different from a Bonferroni correction. The FDR is the expected proportion of significant associations that are false positives. It is defined as $\frac{i}{n} * Q$, where i is the rank of the ascending *p*-values, *n* is the number of markers (tests) and Q is the false discovery rate (Benjamini and Hochberg 1995). We have chosen a FDR of $Q = 20\%$, which means we expect that 20% of the significant associations we observed are false positives. A smaller or higher Q value will lead to a more relaxed or more stringent threshold, respectively. For example when the FDR value was set to 5% and 30%, we detected 10 and 35 significant associations respectively with both EMMAX and MLMM. The threshold calculated with the FDR is specific for a data set. Therefore, different thresholds are used for the different traits, or for imputed vs. non-imputed data. Overall, the identification and implementation of optimal significance thresholds is debated in the literature (Sham 2014).

The difference in the number of significant associations between the non-imputed and imputed data sets may be also due to the imputation method. A drawback of fastPHASE is the limited

usability for GBS data as a parameter vector of allele frequencies has to be estimated for each SNP (the columns of θ in the original publication*)*. Thus, fastPHASE performs best at SNPs with many genotyped individuals but not for sites with no or little genotypic information available (Scheets and Stephens 2006).

The imputation of missing markers only slightly improved prediction ability and accuracy for most of traits (Table 4.4 and 4.5). There was no significant difference in prediction ability or accuracy between the non-imputed data set and imputed data sets. In the same context, Rutkoski et al. (2013) reported that imputation could offer a slight advantage in genomic prediction over non-imputed data, but a prediction based on imputed SNPs in maize was not better than that with non-imputed SNPs (Crossa et al. 2013). Even in self-pollinated crops such as wheat and soybean, for which extensive LD is expected, imputation improved prediction accuracy only minimally (Poland et al. 2012b; Jarquin et al. 2014). In addition, prediction ability and accuracy were similar for both BEAGLE and fastPHASE imputation methods (Table 4.4 and 4.5), although they differ in their underlying model as mentioned above. BEAGLE performs well with medium to large sample sizes (> 1000 individuals) but not as well as fastPHASE with small samples (around 100 individuals) and marker density is low with small marker numbers (~100) in a region (Browning and Browning 2011). BEAGLE's clustering approach flexibly changes cluster number to better accommodate local LD patterns, but these parameters may be biased in the case of low LD.

In summary, imputation did not provide an advantage for GWAS and genomic prediction with our material. Therefore data imputation by fastPHASE and BEAGLE is not recommended for performing GWAS and genomic prediction in cauliflower.

4.6 Conclusion

Cauliflower is characterized by a low genetic diversity (Zhao et al. 2014), which limits the potential for improving varieties that meet the expectations of growers and consumers. With this study we wanted to test modern molecular breeding methods for the first time in such cauliflower material. Overall, our results indicate that association mapping and genomic selection are potentially useful for cauliflower breeding by enabling breeders to understand the genetic basis of complex traits and to select promising genotypes based on genomically

estimated breeding values rather than phenotypic analysis. This study also provides a perspective with respect to the utilization of *ex situ* conserved genebank accessions. Although the set of materials we used was randomly selected and was geographically very broad, we achieved reasonable genomic prediction abilities. For this reason, genotyping whole collections of genebank accessions and the phenotyping of a sufficiently large subset of these may allow the prediction of relevant phenotypic traits in the whole collection and the subsequent selection of accessions for further use as genetic resources.

Author contributions

EY, PT and KS designed the study. EY coordinated the field experiments and conducted the genotyping. PT wrote the scripts for the GWAS, the genomic prediction models and the imputation, EY and PT analyzed the data, EY, PT and KS wrote the manuscript.

Acknowledgments

We express our thanks to Dr. Julie Jacquemin and Getu Bekele for helpful comments and suggestions on earlier versions of the manuscript. Also, we thank Thomas Müller for his support in genotypic data analysis. We are grateful to the USDA and IPK genebanks for sending us the seeds. This work was funded by a DAAD GERLS Fellowship to E.Y. and an endowment of the Stifterverband der deutschen Wissenschaft to K. J. S.

Conflict of Interest: The authors declare that they have no conflict of interest.

4.7 References

Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19:1655-1664

Alexander DH, Lange K (2011) Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC Bioinformatics 12: 246

Bao Y, Vuong T, Meinhardt C, Tiffin P, Denny R et al (2014) Potential of Association Mapping and Genomic Selection to Explore PI 88788 Derived Soybean Cyst Nematode Resistance. Plant Genome 7:3

Bates D, Maechler M, Bolker B, Walker S (2014) lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-6

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

Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing data inference for whole genome association studies by use of localized haplotype clustering. Am J Hum Genet 81:1084-97

Browning SR, Browning BL (2011) Haplotype phasing: existing methods and new developments. Nat Rev Genet 12:703-714

Cai D, Xiao Y, Yang W, Ye W, Wang B, Younas M, Wu J, Liu K (2014) Association mapping of six yield-related traits in rapeseed (*Brassica napus* L.). Theor Appl Genet 127:85-96

Cai S, Wu D, Jabeen Z, Huang Y, Huang Y, Zhang G (2013) Genome-wide association analysis of aluminum tolerance in cultivated and Tibetan wild barley. PloS One 8:e69776

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

Crossa J, Beyene Y, Kassa S, Perez P, Hickey JM et al (2013) Genomic prediction in maize breeding populations with genotyping-by-sequencing. G3-Genes Genomes Genet 3:1903-1926

de los Campos Gdl, Rodriguez PP (2014) BGLR: Bayesian Generalized Linear Regression. R package version 1.0.3. http://CRAN.R-project.org/package=BGLR

Dekkers JCM (2007) Prediction of response to marker-assisted and genomic selection using selection index theory. J Anim Breed Genet 124:331-341

Delourme R, C Falentin, BF Fomeju, M Boillot, G Lassalle et al (2013) High-density SNP-based genetic map development and linkage disequilibrium assessment in Brassica napus L. BMC Genomics 14:120

Edae EA, Byrne PF, Haley SD, Lopes MS, Reynolds MP (2014) Genome-wide association mapping of yield and yield components of spring wheat under contrasting moister regimes. Theor Appl Genet 127:791-807

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

Flint-Garcia SA, Thornsberry JM, Buckler ES (2003) Structure of linkage disequilibrium in plants. Ann Rev Plant Biol 54:357-374

Fu YB, Cheng B, Peterson GW (2014) Genetic diversity analysis of yellow mustard (*Sinapis alba* L.) germplasm based on genotyping by sequencing. Genet Resour Crop Evol 61: 579-594

GenABEL project developers (2013) GenABEL: genome-wide SNP association analysis. R package version 1.8-0

Guo Z, Tucker DM, Basten CJ, Gandhi H, Ersoz E et al (2014) The impact of population structure on genomic prediction in stratified populations. Theor Appl Genet 127:749–762

Hasan M, Friedt W, Freitag NM, Link K, Pons-Kühnemann J, Snowdon RJ (2008) Association of gene-linked SSR markers to seed glucosinolate content in oilseed rape (*Brassica napus ssp. napus*). Theor Appl Genet 116:1035-1049

Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009) Genomic selection in dairy cattle: Progress and challenges. J Dairy Sci 92:433-443

Hefner EL, Lorenz AJ, Jannink J-L, Sorrells ME (2010) Plant breeding with genomic selection: Gain per unit time and cost. Crop Sci 50:1681-1690

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

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

Hinrichs AL, Larkin EK, Suarez BK (2009) Population stratification and patterns of linkage disequilibrium. Genet Epidemiol 33:S88–S92

Howie BN, Donnelly P, Marchini J (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet 5:e1000529

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

Jarquin D, Kocak K, Posadas L, Hyma K, Jedlicka J et al. (2014) Genotyping by sequencing for genomic prediction in a soybean breeding population. BMC Genomics 15:740

Jestin C, Lodé M, Vallée P, Domin C, Falentin C et al. (2011) Association mapping of quantitative resistance for Leptosphaeria maculans in oilseed rape (*Brassica napus L.*). Mol Breed 27:271-287

Jindal SK, Thakur JC (2003) Interrelationship of curd weight and other characters in November cauliflower. J Res 40:358-362

Jombart T, Ahmed I (2011) adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. Bioinformatics 27:3070-3071

Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY et al. (2010) Variance component model to account for sample structure in genome-wide association studies. Nat Genet 42:348-354

König S, Simianer H, Willam A (2009) Economic evaluation of genomic breeding programs. J Dairy Sci 92:382-391

Lan TH, Paterson AH (2000) Comparative mapping of quantitative trait loci sculpting the curd of *Brassica oleracea*. Genetics 155:1927-1954

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

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.*). DNA Res 21:355-367

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics 25:1754-1760

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J et al (2009) The sequence alignment/map (SAM) format and SAMtools. Bioinformatics 25: 2078-2079

Li H, Peng Z, Yang X, Wang W, Fu J et al (2013) Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. Nat Genet 45:43-50

Liu S, Liu Y, Yang X, Tong C, Edwards D et al (2014) The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes. Nature Communications DOI: 10.1038/ncomms4930

Mangin B, Siberchicot A, Nicolas S, Doligez A, This P, Cierco-Ayrolles C (2012) Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. Heredity 108:285-291

Marchini J, Howie B (2010) Genotype imputation for genome-wide association studies. Nat Rev Genet 11:499-511

Marchini J, Howie B, Myers S, McVean G, Donnelly PA (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 39:906-913

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

Moriss GP, Rhodes DH, Brenton Z, Ramu P, Thayil VM et al (2013) Dissecting genome-wide association signals for loss-of-function phenotypes in sorghum flavonoid pigmentation traits. G3-Genes Genomes Genet 3:1903-1926

Narum SR (2006) Beyond Bonferroni: Less conservative analyses for conservation genetics. Conservation Genetics 7:783-787

Norton GJ, Douglas A, Lahner B, Yakubova E, Guerinot ML et al. (2014) Genome Wide Association Mapping of Grain Arsenic, Copper, Molybdenum and Zinc in Rice (*Oryza sativa* L.) Grown at Four International Field Sites. PLoS One 9:89685

Nyquist WE (1991) Estimation of heritability and prediction of selection response in plant populations. Cit Rev Plant Sci 10:235-322

Paradis E, Claude J, Strimmer K (2004) APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20: 289-290

Pei Y-F, Li J, Zhang L, Papasian CJ, Deng H-W (2008) Analyses and comparison of accuracy of different genotype imputation methods. PLoS One 3: e3551

Perneger TV (1998) What's wrong with Bonferroni adjustments. BMJ 316: 1236-1238

Poland J, Brown PJ, Sorrells ME, Jannink J (2012a) Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS One 7:e32253

Poland J, Endelman J, Dawson J, Rutkoski J, Wu S et al (2012b) Genomic selection in wheat breeding using genotyping-by-sequencing. Plant Genome 5: 103-113

Poland J, Rife TW (2012) Genotyping-by-sequencing for plant breeding and genetics. Plant Genome 5: 92-102

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

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

R Core Team (2014) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Rezaeizad A, Wittkop B, Snowdon R, Hasan M, Mohammadi V et al (2011) Identification of QTLs for phenolic compounds in oilseed rape (*Brassica napus L.*) by association mapping using SSR markers. Euphytica 177:335-342

Rice WR (1989) Analyzing tables of statistical tests. Evolution 43:223-225

Romay MC, Millard M, Glaubitz JC, Peiffer JA, Swarts KL et al (2013) Comprehensive genotyping of the USA national maize inbred seed bank. Genome Biol 14: R55

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

Saghai-Maroof MA, Soliman KM, Jorgensen RA, Allard RW (1984) Ribosomal DNA spacer-length polymorphisms in barley: Mendelian inheritance, chromosomal location, and population dynamics. Proc Natl Acad Sci USA 81:8014-8018

Schaeffer LR (2006) Strategy for applying genome-wide selection in dairy cattle Strategy for applying genome-wide selection in dairy cattle. J Animal Breed Genet 123:218-223

Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 78:629-644

Schmid KJ, Thorwarth P (2014) Genomic Selection. In: N. Stein and J. Kumlehn (ed.) Biotechnological approaches to Barley Improvement. Springer Science and Business Media, New York, pp. 367-378

Segura V, Vilhjálmsson BJ, Platt A, Korte A, Seren Ü et al. (2012) An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat Genet 44: 825-830

Sham PC, Purcell SM (2014) Statistical power and significance testing in large-scale genetic studies. Nat Rev Genet 1: 335-346

Sheemar G, Singh D, Malik A, Kumar A (2012) Correlation and Path analysis studies of economic traits in cauliflower (*Brassica oleracea var botrytis* L.). J Agric Tech 8:1791-1799

Singh PK, Pandey V, Singh M, Sharma SR (2013) Genetic improvement of cauliflower. Vegetable Sci 40:121-136

Sonah H, Bastien M, Iquira E, Tardivel A, Légaré G et al (2013) An Improved Genotyping by Sequencing (GBS) Approach Offering Increased Versatility and Efficiency of SNP Discovery and Genotyping. PLoS One 8:e54603

Stich B, Melchinger AE (2010) An introduction to association mapping in plants. CAB Rev 5:039

Tardivel A, Sonah H, Belzile F, O'Donoughue LS (2014) Rapid Identification of Alleles at the Soybean Maturity Gene E3 using genotyping by Sequencing and a Haplotype-based Approach. Plant Genome 7:1-9

Wimmer V, Albrecht T, Auinger HJ and Schoen CC (2012) synbreed: a framework for the analysis of genomic prediction data using R. Bioinformatics 28:2086-2087

Wright SI, Ness RW, Foxe JP, Barrett SCH (2008) Genomic consequences of outcrossing and selfing in plants. Int J Plant Sci 169:105-118

Würschum T, Abel S, Zhao Y (2014) Potential of genomic selection in rapeseed (Brassica napus L.) breeding. Plant Breed 133:45-51

Yousef EAA, Lampei C, Schmid K (2015) Evaluation of cauliflower genebank accessions under organic and conventional cultivation in Southern Germany. Euphytica 201:389-400

Yousef EAA, Muller T, Borner A, Schmid K (2015) Evidence of strong population structure induced by germplasm regeneration in *ex situ* genebank collections of cauliflower (*Brassica oleracea var. botrytis*)

Zhao Z, Gu H, Sheng X, Yu H, Wang J, Zhao J, Cao J (2014) genetic diversity and relationships among loose-curd cauliflower and related varieties as revealed by microsatellite markers. Sci Hortic 166: 105-110

Zou J, Jiang C, Cao Z, Li R, Long Y, Chen S, Meng J (2010) Association mapping of seed oil content in *Brassica napus* and comparison with quantitative trait loci identified from linkage mapping. Genome 53:908-916

4.8 Supplementary Materials

Table S4.3 Lambda values for used models with the two data sets (none- mputed data and imputed data).

	None-imputed data					Imputed_data				
	GLM	Q+K	Eigens	EMMAX	MLMM	GLM	Q+K	Eigens	EMMAX	MLMM
Curd width	3.74	0.97	1.25	1	0.87	4.44	1.16	1.40	1.00	0.98
Cluster width	3.74	1.25	1.64	1.08	0.98	8.29	1.19	1.45	1.02	1.02
No. of branches	2.41	1.30	1.45	1.09	1.13	2.47	1.21	1.46	0.995	1.00
Apical length	1.92	1.19	1.35	1.03	1.05	2.03	1.06	1.35	1.02	0.93
Nearest branch length	2.48	1.28	1.44	1.11	0.99	2.79	1.24	1.50	1.04	1.02
No. of days	3.54	1.54	1.72	1.18	1.18	3.61	1.40	150	1.01	1.00

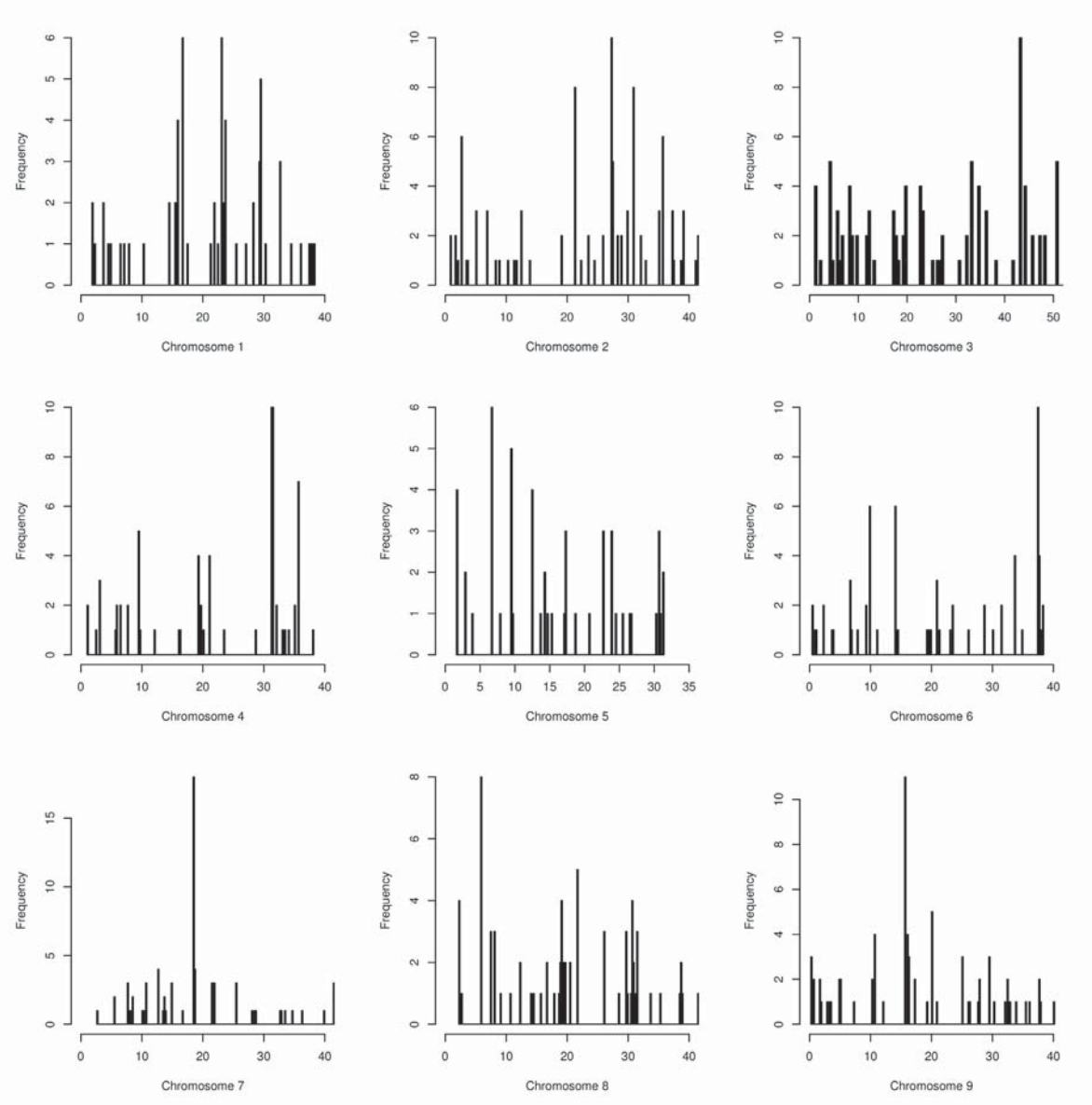

Figure S4.1 The distribution of non-imputed SNPs over nine chromosomes.

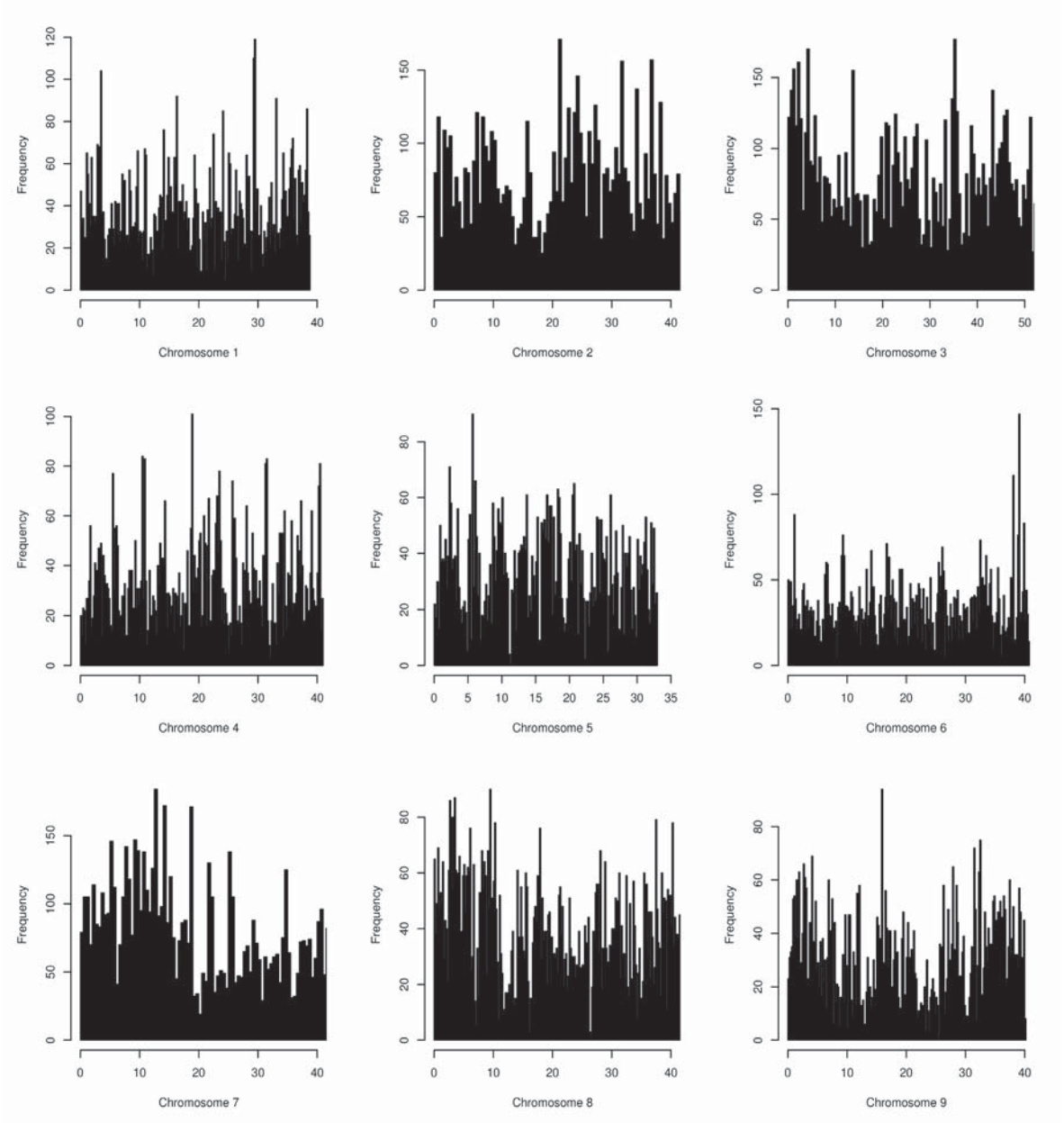

Figure S4.2 The distribution of imputed SNPs using BEAGLE over nine chromosomes.

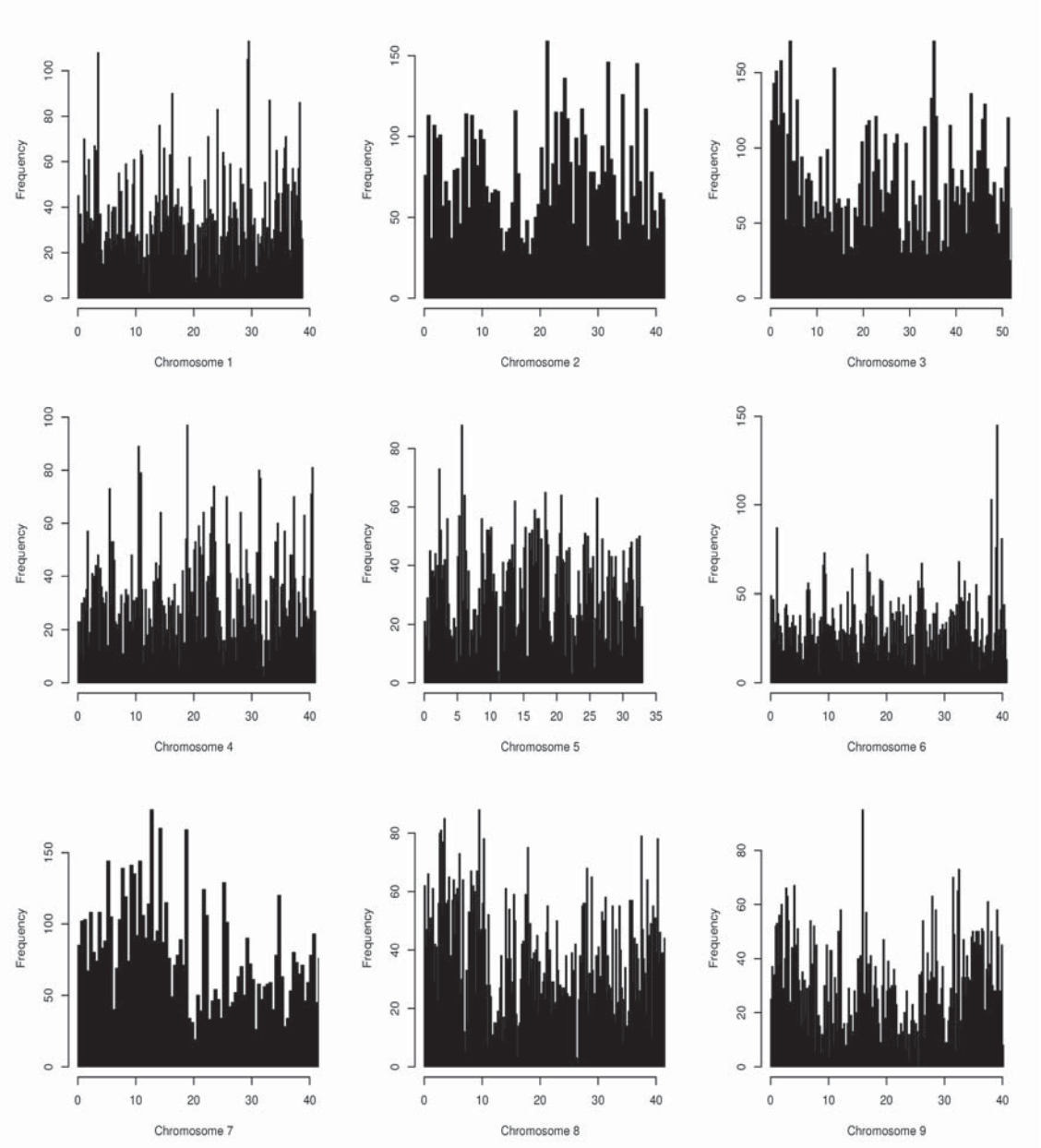

Figure S4.3 The distribution of imputed SNPs using fastPHASE over nine chromosomes.

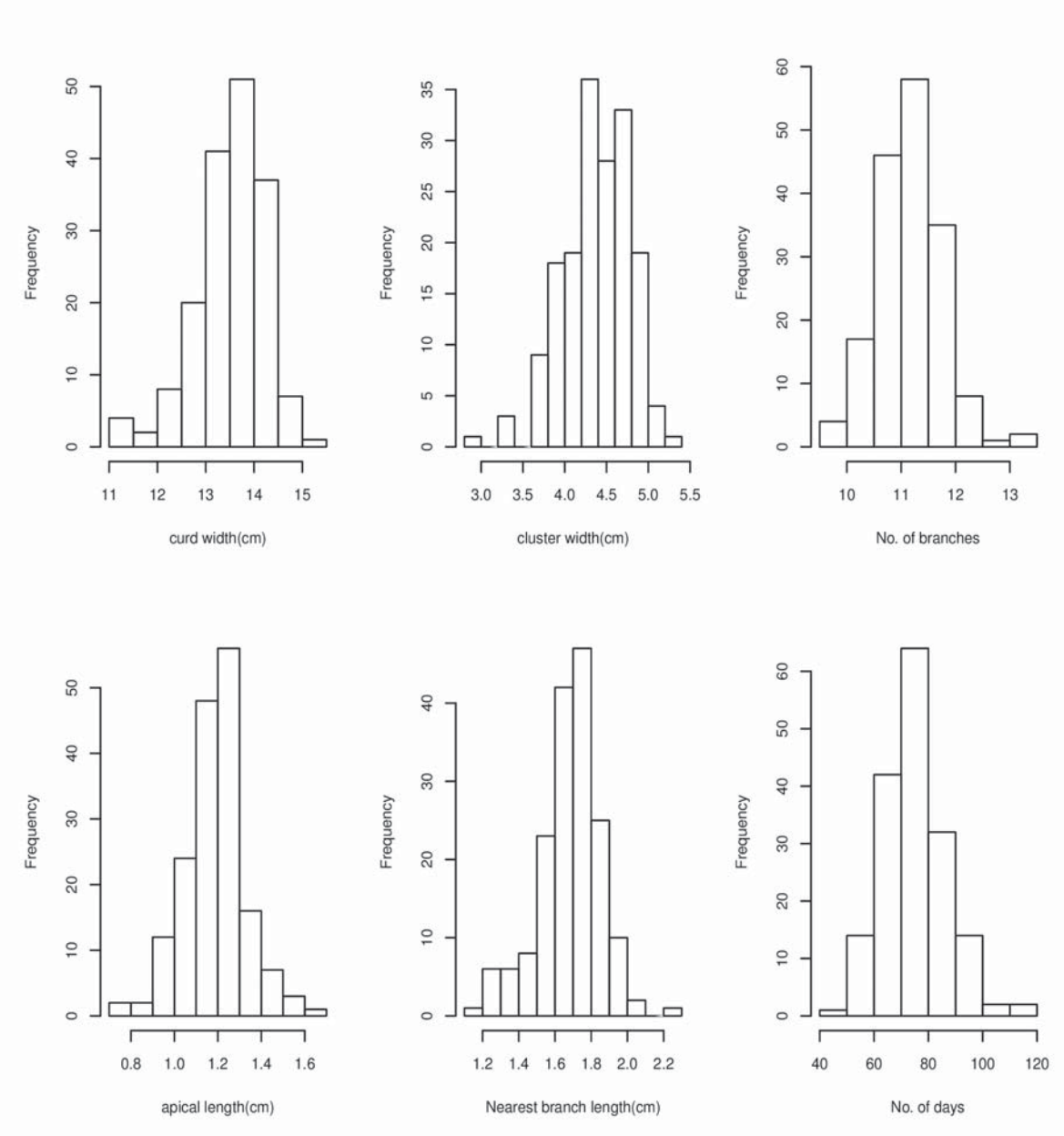

Figure S4.4 The distribution of curd related traits over two locations and three growing seasons.

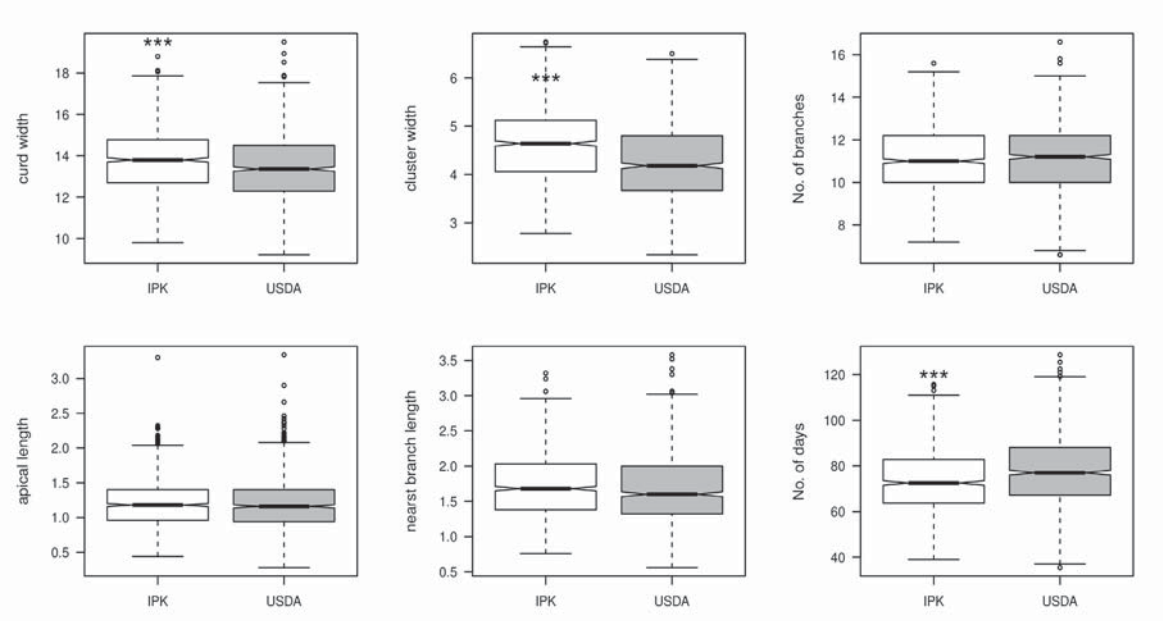

Figure S4.5 Boxplot graph for six curd-related traits between accessions of two genebanks over two locations and three growing seasons. Significant differences between the two genebanks are indicated with stars on whiskers (***<0.001).

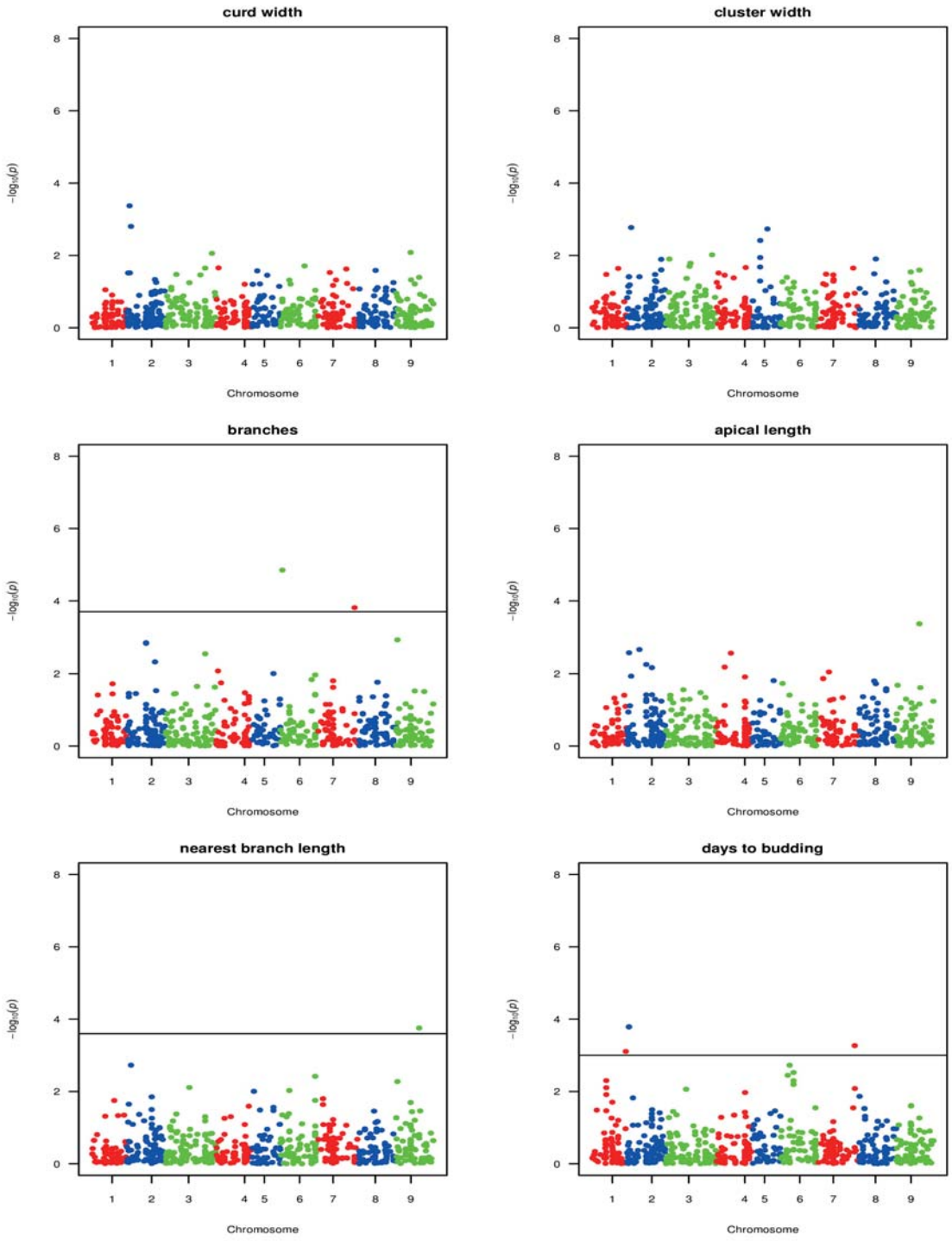

Figure S4.6 Manhattan plots of association analysis using 675 SNPs and the MLMM method for six curd-related traits for imputed data. Each dot represents a SNP. The horizontal line represents significance threshold with FDR.

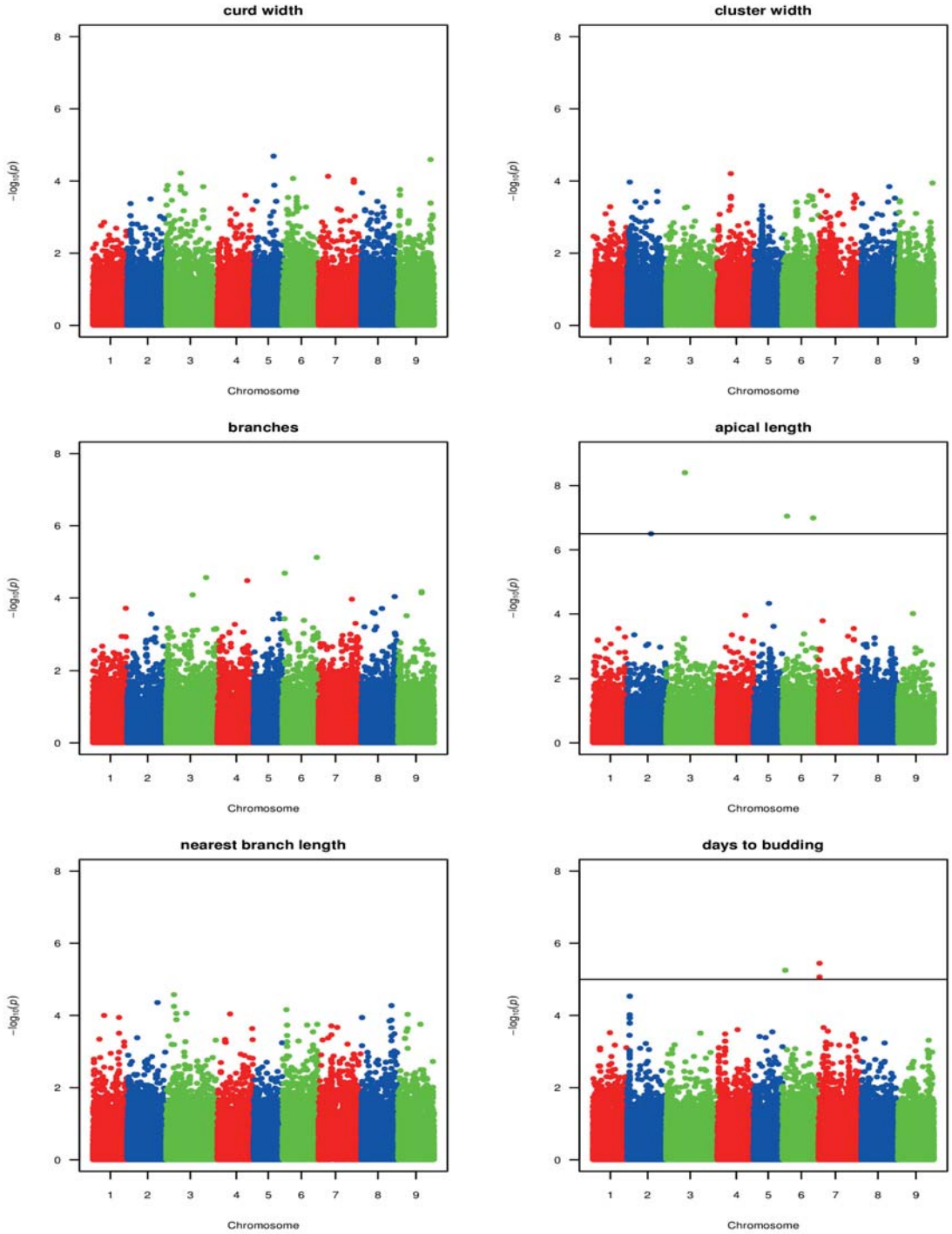

Figure S4.7 Manhattan plots of association analysis using 64,372 SNPs and the MLMM method for six curd-related traits for imputed data. Each dot represents a SNP. The horizontal line indicates the significance threshold with FDR.

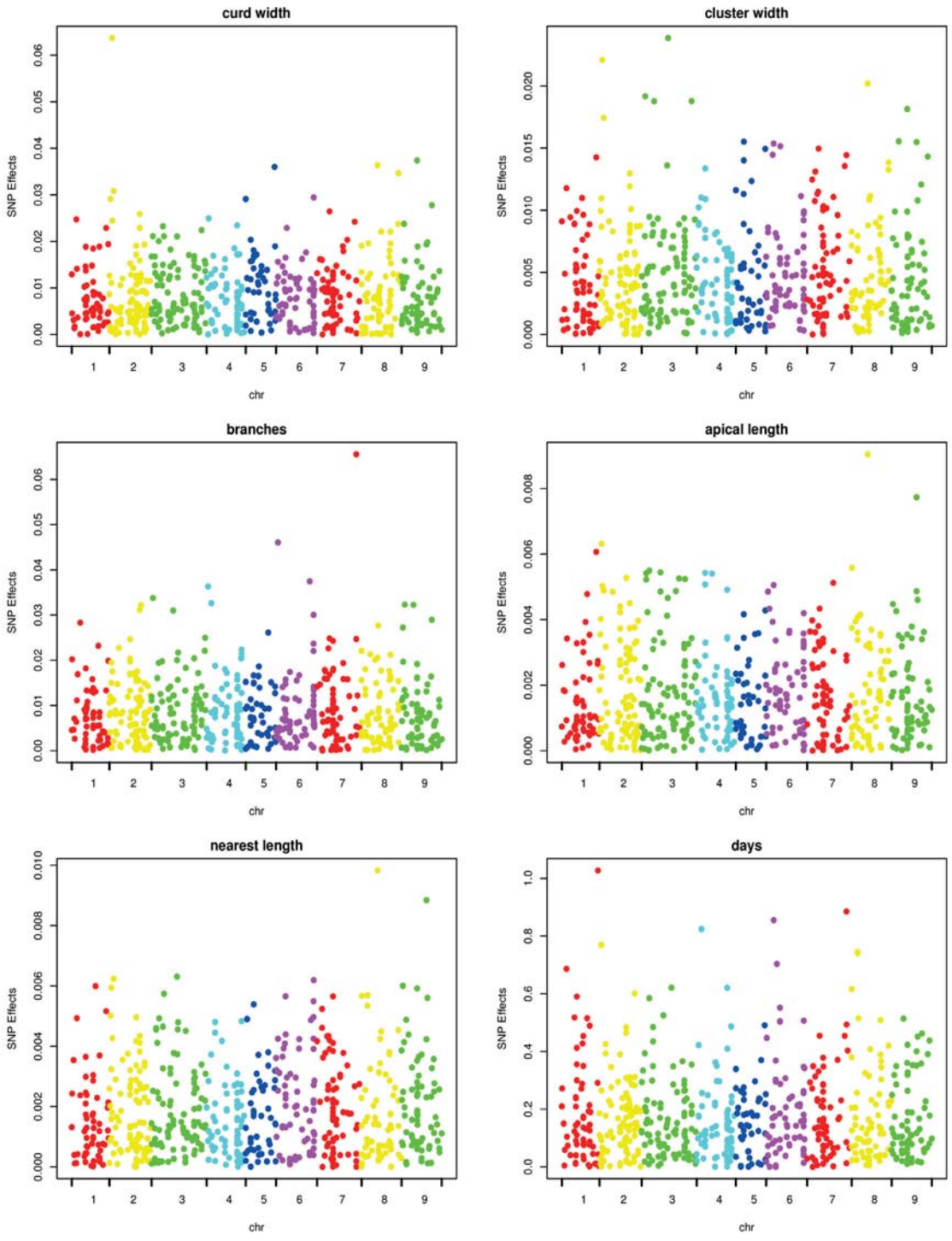

Figure S4.8 Manhattan plots for marker effects using RRBLUP with non-imputed data for six curd-related traits.

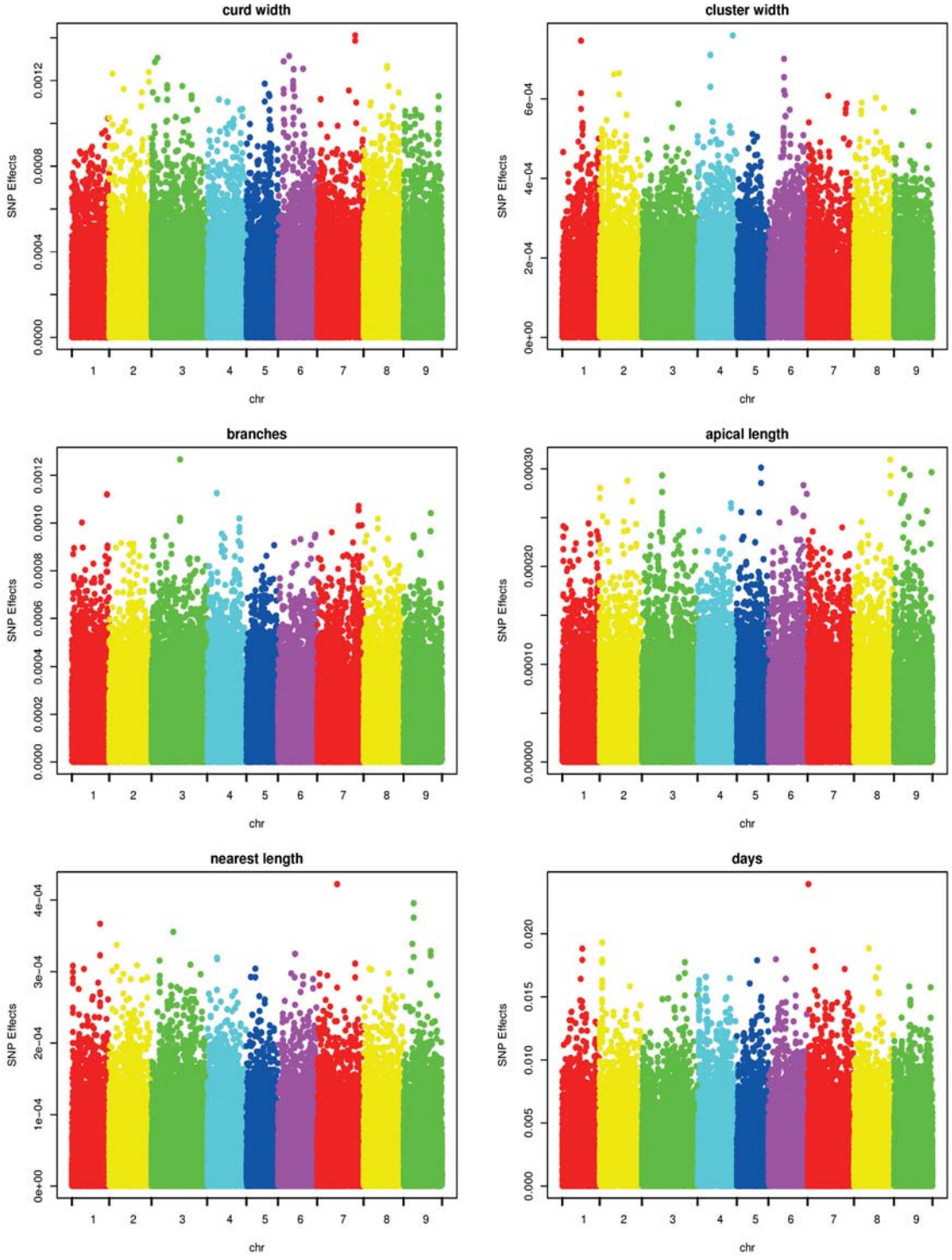

Figure S4.9 Manhattan plots for marker effects using RRBLUP with imputed data by BEAGLE for six curd-related traits.

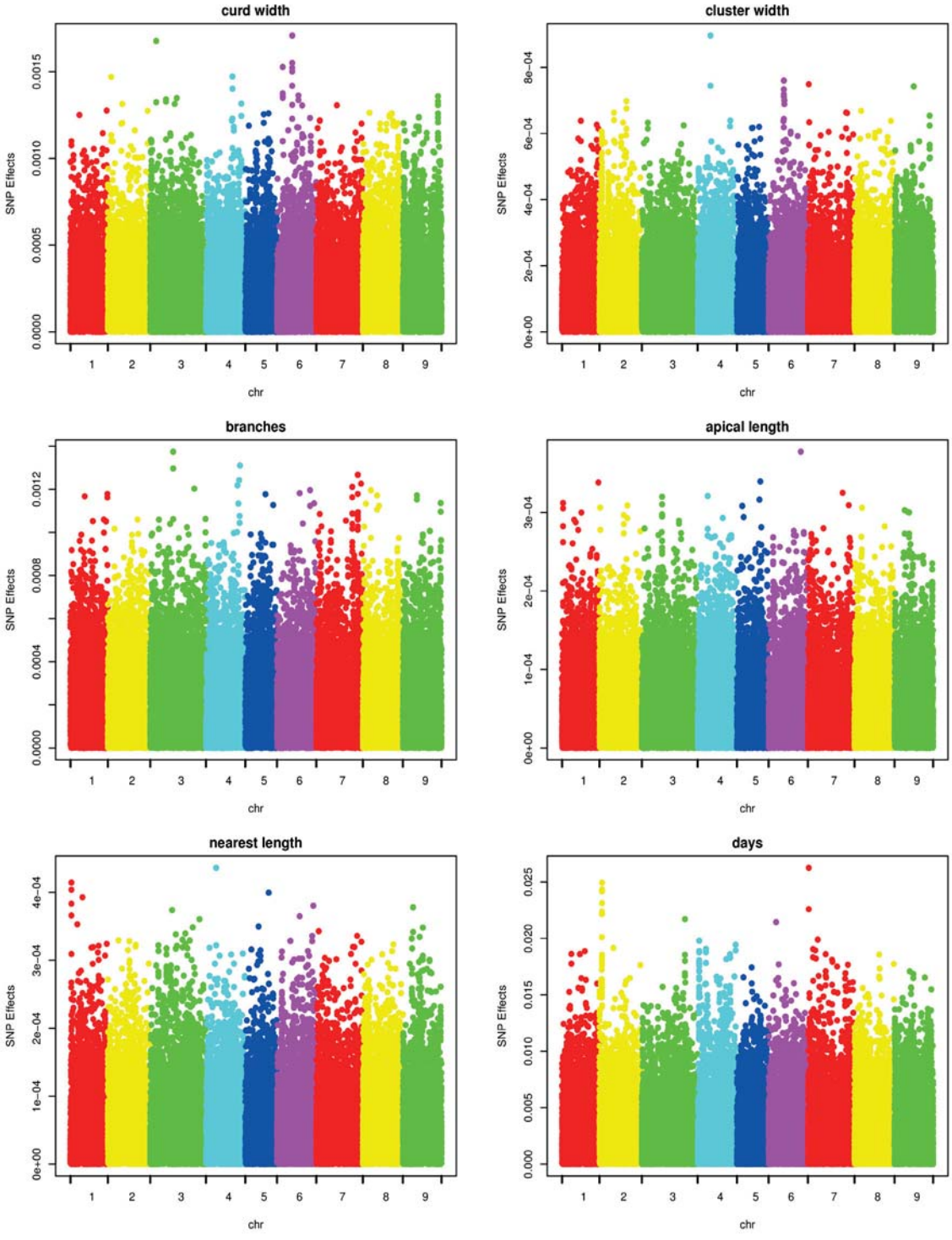

Figure S4.10 Manhattan plots for marker effects using RRBLUP with imputed data by fastPHASE for six curd-related traits.

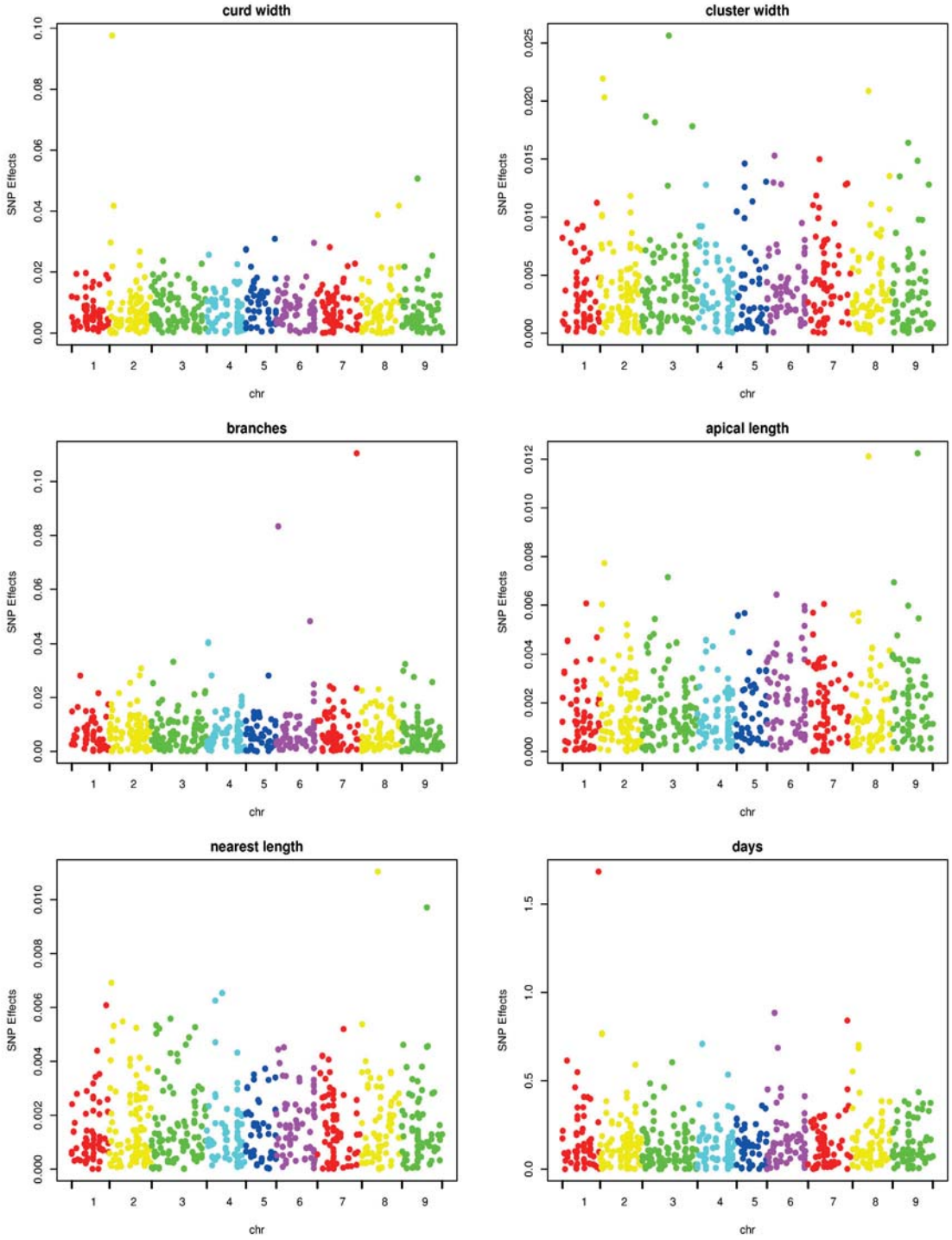

Figure S4.11 Manhattan plots for marker effects using BayesB with non-imputed data for six curd-related traits.

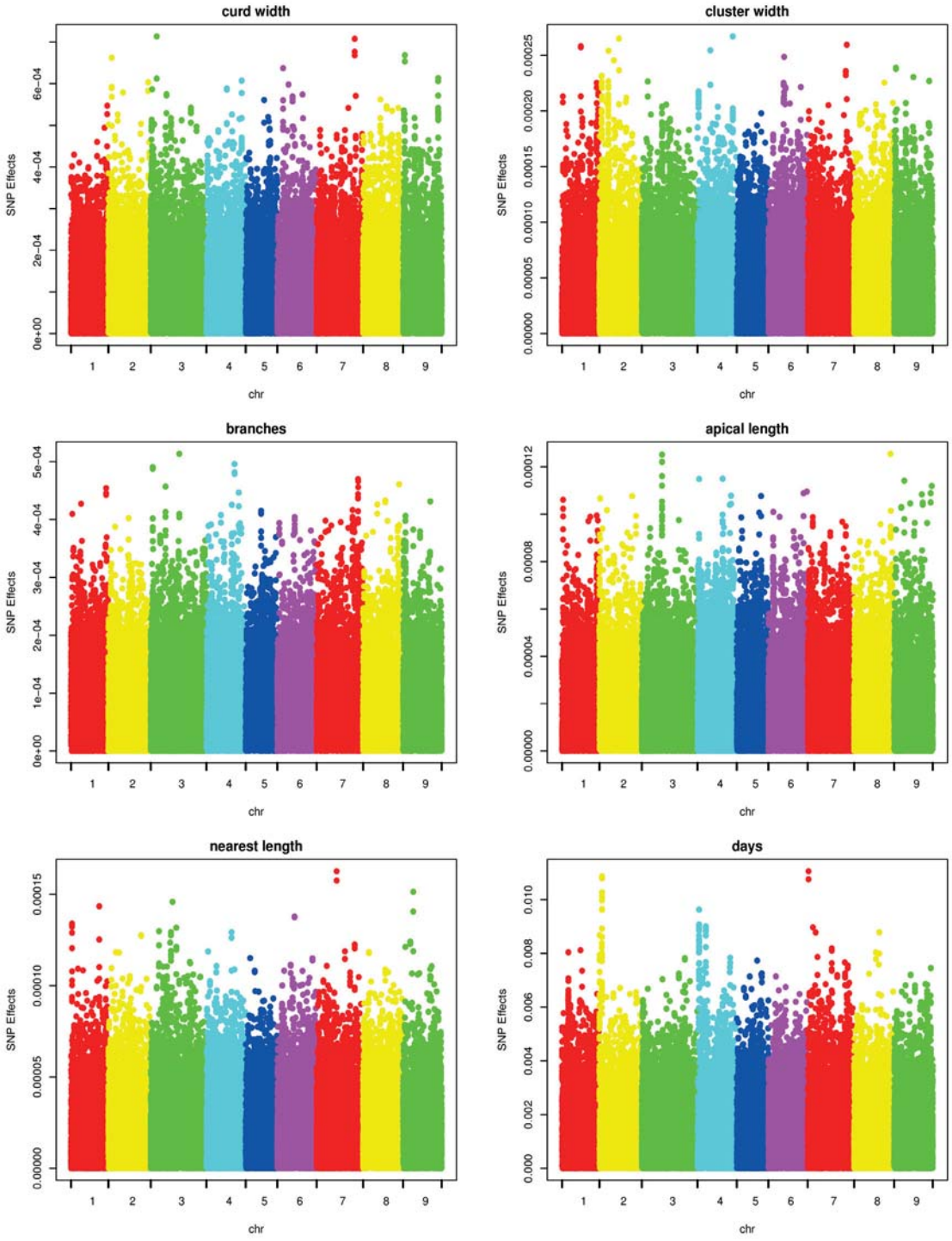

Figure S4.12 Manhattan plots for marker effects using BayesB with imputed data by BEAGLE for six curd-related traits.

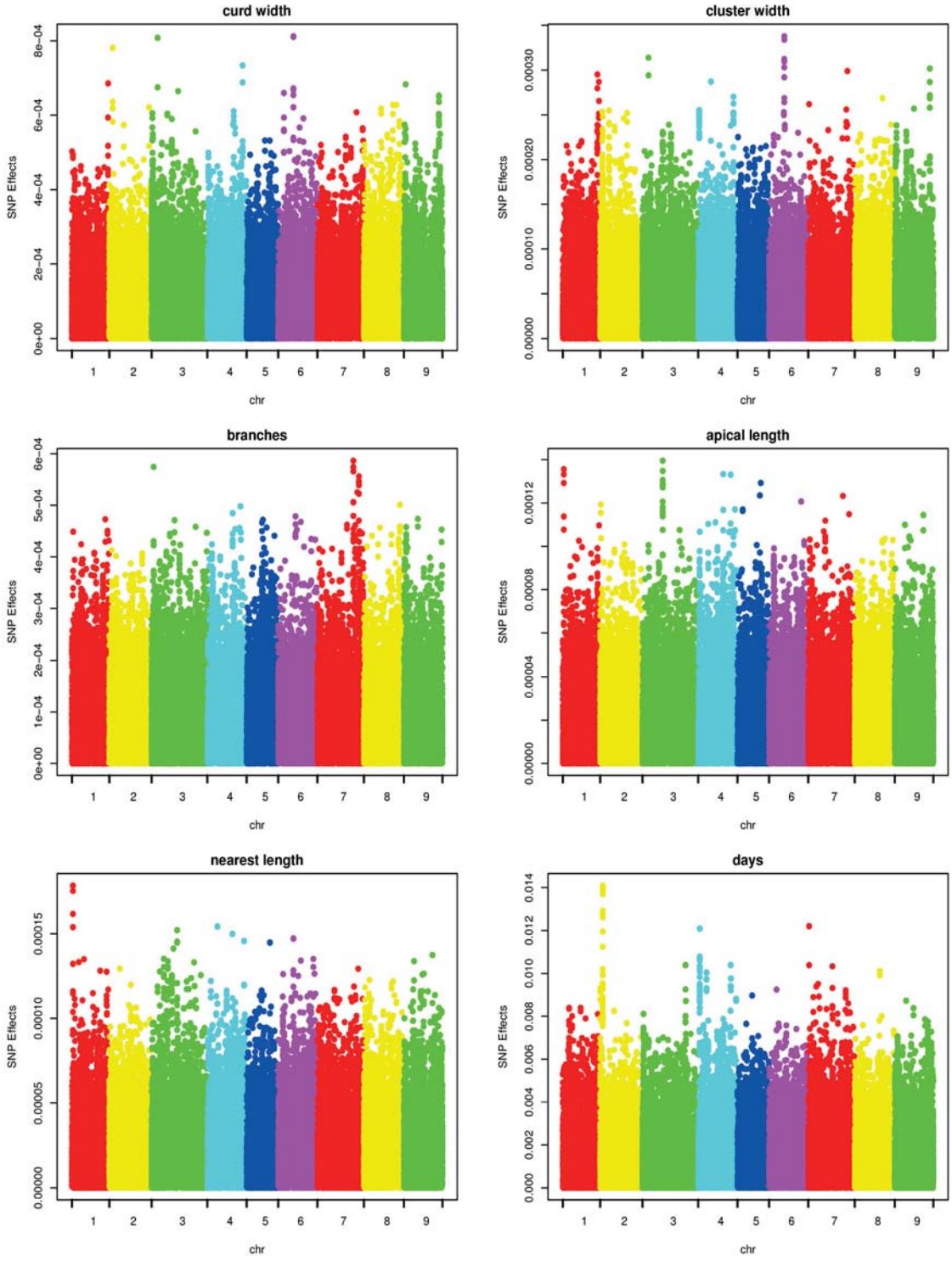

Figure S4.13 Manhattan plots for marker effects using BayesB with imputed data by fastPHASE for six curd-related traits.

5 General discussion

Genetic resources ensure the genetic materials for enhancing crop productivity through plant breeding. Without a broad base of diverse genetic materials, it is impossible for plant breeders to produce new cultivars to feed the world's rapidly rising population. Therefore, broadening the genetic base for breeders is of fundamental importance for crop improvement so as to overcome the current problems associated with the narrow genetic base of modern cultivars (Rao and Hodkin 2002; Fernie 2006). Broadening of the genetic base could be achieved by assessing the phenotyoic and genotypic diversity of a wide range and number of *ex situ* conserved genetic materials, which are considered an important source of genetic variations (Abdurakhmonov and Abdukarimov 2008).

5.1 Organic breeding in cauliflower

Due to an acute lack of varieties adapted to organic farming, more than 95% of the varieties used in organic food production nowadays have actually been bred under conventional conditions (Lammerts van Bueren et al. 2011). However, there is some uncertainty about the performance of these varieties under the regimen of organic farming, such as the absence or reduced use of chemical inputs. Therefore, organic growers place greater confidence in breeding new crop varieties suitable for organic farming that are developed for optimal production in organic systems (Myres et al. 2012). The genetic variability existing in old varieties and landraces, which are now preserved and stored in *ex situ* genebanks, are considered an important reservoir for organic breeding because of their genetic heterogeneity and they were developed before synthetic inputs were available. Consequently, they may have the potential to evolve specific adaptations to organic farming conditions or have agronomic and quality traits of interest to organic farmers (Dawson et al. 2011).

In this thesis, one hundred seventy eight cauliflower genebank accessions were evaluated under organic and conventional cultivation for three successive growing seasons to estimate the effect of the cultivation method on some curd-related traits that have a direct effect on yield and maturity in cauliflower, and to identify genotypes that may be used as parental lines for future breeding in organic farming. Surprisingly, in contrast to Maggio et al. (2013), who reported significant reduction in curd width under organic conditions, this study did not find a strong effect of cultivation methods on the investigated traits of curd width and number of days to

budding. This result may be explained by the fact that both farms observed (one using organic and one using conventional cultivation methods) were located in a very fertile cabbage-growing region of southern Germany and also that the two farms are close to each other and share very similar soil types and microclimatic conditions. However, only the Aug 2012 cultivation season showed a significant difference between the two systems, and also showed a reduction in curd width of about 2 cm as well as 15 days delayed budding compared to the other seasons. This finding suggests that the effect of organic cultivation may be stronger on less fertile soil or otherwise worse environmental conditions. Moreover, our results show that both traits were more strongly affected by growing seasons, which indicates that the cauliflower breeder should take the growing season into account.

Although the effect of cultivation method was rather small, this study observed a significant genotype x cultivation-method interaction. Consistent with this result, Mason et al. (2007) also found some evidence for the existence of a genotype x cultivation-method interaction. This indicates that some genotypes perform better under organic conditions and some accessions perform better under conventional conditions. Based on this finding, it should be possible to select genotypes with good performance under organic conditions. However, one question that has been discussed in the scientific community is whether cultivars for organic cultivation should be selected under organic or conventional conditions. Cerccarelli (1996) and Wolf et al. (2008) mentioned that selection should be performed under optimum conditions to avoid reduced heritability, moreover they found a similarity between selection under organic and under conventional farming for some traits. Murphy et al. (2007), Reid et al. (2009) and Kirk et al. (2012) on the other hand mentioned that direct selection of genotypes for organic farming is better than indirect selection in conventional conditions and they reported that the direct selection resulted in higher yields (5-31%) compared to indirect selection in conventional cultivation.

Baenziger et al. (2011) and Goldstein et al. (2012) provided an intermediate and reasonable solution to this question. They reported that indirect selection is most efficient for traits with a high heritability and show a strong correlation between cultivation systems in wheat and maize. The present study found that the genetic correlation between the two cultivation methods (organic and conventional) was low for curd width but it was high for number of days to

budding. Also, this study found that the number of days to budding had a high heritability, while curd width showed a low heritability in both environments. Based on these results we calculated the predicted response to direct and indirect selection for both traits. For days to budding, response to direct selection and indirect selection was high (9.64 and 9.15, respectively), and the difference between direct and indirect selection was marginal. For curd width, the response to selection was very low in general (0.099 and 0.039, respectively) but it was two times higher for direct selection than for indirect selection. Moreover, we calculated the efficiency of indirect selection for both traits. It was quite high (0.95) for number of days to budding and low (0.39) for curd width. These findings demonstrate that the selection for curd width should be conducted under organic conditions, whereas selection for number of days to budding can be carried out under organic or conventional conditions.

To identify varieties that may be suitable as starting material for a breeding program focusing on organic agriculture in cauliflower, we calculated the genotype trait "means" (best linear unbiased predictors; BLUPs) and the genotype "stability". Based on these values, four and five genotypes were found to have high adaptability with both cultivation methods for curd width and for number of days to budding, respectively. However, when the selection was based only on genotype trait "means" (BLUPs), it resulted in three genotypes for curd width and seven genotypes for number of days to budding in both cultivation systems. This finding further demonstrates the effect of low heritability and low genotypic correlation on indirect selection. Also, it suggests that heritability of stability is important when joint selection is performed. Furthermore, it shows that despite our earlier results, that indirect selection may be more efficient than direct, the differences in ranking in the two cultivation systems make it rather unlikely that the best performing genotypes will be selected via indirect selection.

5.2 Inference of genetic diversity and population structure

Knowledge about germplasm diversity and population structure is needed before studying several topics such association mapping, heterosis and the selection of parental lines. Also, it is highly important not only in the context of breeding but also in that of germplasm conservation. For instance, it is required for decision-making in plant genetic resource organizations such as gene banks, to decide what, how and where genotypes to preserve, as well as to make core collections (Rao and Hodgkin 2002).

Several molecular marker systems are available and are widely used to perform analyses of genetics and population structure. But one very important concern is whether these markers can unravel genetic relatedness, especially in crop species that have very narrow genetic diversity, such as cauliflower (Tonguc and Griffith 2004; Louarn et al. 2007; Zhao et al. 2014). Therefore, we investigated the population structure and genetic diversity of cauliflower accessions collected worldwide (174 accessions from two genebanks) using a promising GBS approach, in order to overcome the limitations of previously used marker-based genotyping techniques which were time consuming, expensive and not suitable to producing large number of markers.

The 174 accessions used were collected as samples from twenty six countries and stored and maintained in two genebanks. Surprisingly, we discovered the accessions clustered in two distinct groups belonging to the USDA and IPK genebanks and not to the geographic origin. These two accession groups showed strong divergence. Also, our results show a close similarity between all analyses we used. Moreover, the same result was obtained by using different data sets (data without missing values, data with missing values, and with imputed data). One possible explanation for this unexpected result is that strong natural or artificial selection has acted upon the samples in each genebank, something that might indeed occur during the propagation of the accessions. We therefore investigated signals of selection using three different software programs (LOSITAN, ARLEQUIN and Bayenv2). Interestingly, these three outlier tests detected different numbers of loci as outliers. The outliers represent regions of candidate genes exhibiting selection signatures. These findings indicate that selection may indeed act on the genebank material. If this is true, genebank material may change considerably over years of storage, which can have unwanted effects. However, further investigations are needed to clarify this point.

The results of this study show that the collection of accessions we used is characterized by low genetic diversity. However, the low genetic diversity in this collection could be mirroring a generally low diversity of cauliflower, as cauliflower cultivars generally exhibit high similarity (Zhao et al. 2014; Tonguc and Griffith 2004). On the other hand, the low genetic diversity can be a consequence of the maintenance procedures of the genebanks. There is considerable empirical data showing loss of genetic variation within the preserved genebank collections as a consequence of repeated rounds of rejuvenation (Parzies et al. 2000; Gómez et al. 2005;

Rucin'ska and Puchalski 2011; Hagenblad et al.2012 and Brütting et al.2013). This loss in genetic variation in *ex situ* genebanks might be due to the low number of parents contributing to the rejuvenation process per accession due to the high number of accessions that need to be rejuvenated. This in turn could expose the genebank collections to problems such as genetic bottleneck, inbreeding depression or deleterious mutation accumulation, which lead to losses of genetic diversity (Crossa 1995; Frankham et al. 2002; Rucin'ska and Puchalski 2011; Kasso and Balakrishnan 2013).

One of the interesting points of this study is that all population structure analyses and genetic diversity parameters with three data sets indicate that the USDA's accessions were more diverse than those of the IPK. Moreover, the mean pairwise F_{ST} within the USDA collection was higher (F_{ST} = 0.301) than within the IPK collection (F_{ST} = 0.160), which confirms that the USDA's accessions were indeed more diverse than the IPK's. This result might be explained by two factors. First, the difference between the two genebanks with regard to genetic diversity parameters might be due to their regeneration procedures. It was reported that genetic changes resulting from regeneration can be quite strong (Börner et al. 2000, Chebotar et al. 2003; van Hintum et al. 2007). To our knowledge, the regenerations of the USDA are performed in cages (12 x 24 ft) with mesh covers. At least 100 plants per accession for each regeneration cycle are propagated. The IPK, on the other hand, uses small glass houses containing several species together and the area for the cauliflower multiplication is around 6 m^2. They propagate only 20-25 plants per accession in each regeneration. In this regard, Parzies et al. (2000) stated that a too low size of rejuvenated populations could lead to loss in the genetic variation and accumulate the deleterious mutation that may be difficult to monitor. They also reported that 200 parents are needed in a self-fertilizing species to avoid such deleterious effects. Second, the difference between the two genebank collections may indicate that the USDA harbors more exotic material (landraces) than the IPK. Unfortunately, detailed information about the biological status of the USDA accessions is not available, since most accessions were classified as unverified material.

5.3 Genome-wide association study in cauliflower

Association mapping was introduced as an alternative to linkage mapping, to overcome the known limitations of the latter approach (Flint-Garcia et al. 2003). Association mapping was successfully used, for example, to identify QTL in *Brassica olearacea (*Lan and Paterson 2000;

2001; Gu et al. 2008; Walley et al. 2012; Brown et al. 2014). It utilizes historical LD occurring in diverse germplasm or in natural populations, thus there is more time and a higher probability for recombination events to happen by chance than in linkage mapping, which also utilizes LD, but in segregating biparental populations (Flint-Garcia et al. 2003; Ersoz et al. 2007).

Several studies showed that association mapping is a promising and effective technique to unravel the genetic architecture of complex traits in several crop species (Li et al. 2014; Cai et al. 2014, Norton et al. 2014; Edae et al. 2014). For instance, it was successfully used to study the genetic architecture of several traits in *Brassica napus (*Rezaeizad et al. 2011; Cai et al. 2014; Li et al. 2014). Several types of important cruciferous vegetable crops belonging to the *Brassica oleracea* show high diversity with respect to roots, stems, leaves, inflorescences and apical or lateral buds. Due to this remarkable morphological diversity *B. oleracea* is an ideal species for association mapping studies. Nevertheless, no association mapping studies have yet been conducted in *B. oleracea*.

As discussed above, investigation of the population structure and genetic diversity is very important for different aspects of breeding and genetic conservation. Ignoring the population structure and genetic diversity in association mapping could lead to pseudo-marker trait associations, because they fabricate LD between unlinked loci (Wright and Gaut 2005). We therefore examined five different statistical models used to perform association analysis, which account differently for population structure and kinship. In the comparison of the different models, the *p*-values observed with a GLM (generalized linear model) strongly deviated from the expected *p*-values for all traits in both data sets (non-imputed data and imputed data), which is expected because a GLM model does not account for population structure or relatedness. GLM was followed by the EIGENSTRAT and Q+K models. However, for some traits EIGENSTRAT and Q+K did not deviate as strongly as the GLM from the expected *p*-values, which indicates that these approaches efficiently controlled for confounding effects due to population structure. This can be explained by the fact that both of them are correcting for population structure (Yu et al. 2006). The observed distribution of *p-values* of EMMAX (Efficient Mixed-Model Association) and MLMM (Multi Locus Mixed Model), which are correcting for kinship, were closer to each other as well as to the expected distribution of *p-values,* with both data sets, than for the other models. This result indicates that the control of both type I and II error rates, which

are expected to occur more frequently in a structured population, was improved compared to the other methods, which in turn indicates that these models effectively controlled for false positive associations and avoided false negative associations. Thus the EMMAX and MLMM models are recommended as the first choice for association analyses in *B. oleracea*.

In this study, twenty four SNP markers were found to be significantly (based on false discovery rate, FDR) associated with six curd-related traits. Out of these 24 significant associations, six associations were shared by EMMAX and MLMM models. Moreover, common association between the traits of curd width and cluster width, and between the traits of number of branches and number of days to budding were detected. These results indicate several strong QTL segregating in this *Brassica oleracea* material, which proved to be robust over different methods. All of these previous results suggest that association mapping can be used successfully to dissect polygenic traits in *Brassica oleracea*. Also, they indicate that the associations identified for the curd-related traits might be successfully used in marker MAS programs for cauliflower breeding.

5.4 Genomic prediction in cauliflower

As we mentioned above, after identifying marker-trait associations, breeders could use the identified markers in their breeding programs in a genetic approach, which is called MAS. However, MAS has some limitations. It is less efficient for complex traits controlled by several loci with small effects (Moreau et al. 1998). Because of the inadequacy of MAS for improving complex traits, GS was introduced by Meuwissen et al. (2001) as an alternative approach. The theory underlying GS is that the effects of all markers throughout the entire genome are assessed by using genomic prediction models. The estimates thus obtained are then used to predict the breeding values of the individuals of the next generations (Jannink et al. 2010). The genomic estimated breeding values can be used to select the individual without the need of phenotypic data. To date, only one study has investigated the potential of genomic selection in *Brassicaceae,* i.e. in rapeseed (Würschum et al. 2014). In this study, the authors concluded that GS can be a powerful tool for breeding in rapeseed.

In the current study, we successfully performed genomic prediction analysis with two promising statistical models: RRBLUP and BayesB (Meuwissen et al. 2001) for six different traits in cauliflower. The obtained prediction abilities were promising. We found prediction ability in a range from 0.128 to 0.652 with an average of 0.409 using RRBLUP, and prediction ability in the

range of 0.09 to 0.660 with an average 0.405 using BayesB. Interestingly, both models show a promising prediction ability for all traits, except apical length and length of the nearest branch. The lower predictive power for the these two traits might be due to their low heritability, the lack of enough markers to cover the whole genome evenly or the strong genetic structure of the plant material used, as the collection analyzed was classified into two separate groups, as discussed above. The high prediction ability for number of days to budding is comparable to results for number of days to flowering trait in rapeseed (Würschum et al. 2014). Our results demonstrate that GS possesses the potential to effectively accelerate cauliflower breeding by increasing genetic gain per unit time and reducing the need for extensive phenotyping in cauliflower breeding programs.

In this study, there was no significant difference in predictive ability or accuracy between the RRBLUP and Bayes B models. Similarly, Bao et al. (2014) found that the sophisticated Bayesian models did not outperform RRBLUP in soybean. This may indicate that the RRBLUP model captures most of the genetic variation existing in our germplasm. Also, several studies suggested slight differences among different genomic prediction models in oat and maize (Asoro et al. 2011; Lorenzana and Bernardo 2009). Altogether, the RRBLUP had a mean of prediction ability and accuracy equivalent to BayesB and is moreover computationally less demanding than the Bayes B model. RRBLUP has been shown to provide high, accurate, efficient and stable prediction accuracy for different complex traits in several crop species (Heslot et al. 2012; Ould Estaghvirou et al. 2013; Würschum et al. 2013; Würschum et al. 2014). All of these advantages suggest that RRBLUP would be a robust and efficient tool for performing genomic prediction in applied cauliflower breeding.

5.5 Imputation of missing GBS values

In the current study, GBS has produced 120693 SNPs, which confirms its flexibility and ability to produce high density information in *Brassicaceae* at low cost. However, GBS data is known to have a large proportion of missing data. The overall percentage of missing data in our study ranged from 19% to 77% with an average of 42%. This finding is in accordance with the results of Elshire et al. (2011) and Poland et al. (2012a). Imputation of missing values was suggested as a potential solution to overcome this problem and several methods have been developed and implemented in different programs to perform the imputation.

To study the effect of imputation on the results of the population structure analysis, the association mapping and the genomic prediction methods, a comparison between non-imputed and imputed data was performed, and missing values were imputed by fastPHASE (Scheets and Stephens 2006) and BEAGLE (Browning and Browning 2007) in this study. We can conclude from the results obtained with this comparison, that there is no big difference between non-imputed data and imputed data with regard to the population structure and genetic diversity. Both imputed and non-imputed data sets revealed that the accessions used were clustered into two distinct groups as we discussed above. Also, diversity estimates with different data sets, i.e. imputed and non-imputed data sets, indicate that the USDA's collection was more diverse than that of the IPK. In this regard, Fu (2014) reported that estimation of heterozygosity and inbreeding coefficient (from original, missing, and imputed) were less accurate with larger numbers of missing data, but that the estimation biases were much smaller for missing data without imputation compared to imputed data. Therefore, we conclude that GBS has the ability to assess the genetic diversity and population structure in *Brassicacaea* and imputed data is not encouraged.

Interestingly, the accuracy of association mapping was not improved by using imputed data compared to non-imputed data. The number of significant markers was lower (10) in the case of imputed data than number of significant markers in the case of non-imputed data (14). This result is in contrast to the findings of Marchini et al. (2007); Howie et al. (2009); Marchini and Howie (2010), who reported that marker imputation could increase the power of association mapping studies. Also, we failed to find a significant difference in the prediction ability and accuracy between the non-imputed and imputed data sets or between the imputation methods. Consistent with our results, Crossa et al. (2013) report that data imputation did not significantly improve prediction ability in maize compared to non-imputed SNPs. Also, Poland et al. (2012b) and Jarquin et al. (2014) showed that data imputation slightly improved prediction ability and accuracy, but did not have a significant effect on the genomic prediction ability and accuracy nor were there any significant differences between the imputation methods.

All previous findings suggest that GBS with non-imputed data could be used effectively to study population structure, association mapping and genomic prediction in cauliflower without loss of accuracy. However, our analyses were restricted to the imputation methods described here − or

essentially to only one type of imputation using LD information, which was found to be low in cauliflower. Thus, different results might be explored with other imputation methods.

5.6 General conclusion

In order to understand the phenotypic and genotypic diversity of cauliflower as well as to accelerate marker-assisted breeding, two hundred cauliflower genebank accessions were phenotyped at two locations (organic farm and conventional farm) and 3 growing seasons and were characterized with 120,693 SNPs using GBS. The following conclusions can be drawn, which are relevant to cauliflower breeding targeting for yield improvement:

1. No significant difference was found between the cultivation methods (organic and conventional) for curd-related traits (curd width and number of days to budding).
2. Selection for organic farming is feasible and preferable due to genotype x cultivation-method interactions.
3. Selection for curd width should be conducted under organic conditions, whereas selection for number of days to budding can be carried out under organic or conventional conditions.
4. The accessions used had low genetic diversity.
5. Surprisingly, the accessions used were clustered into two distinct groups. These two groups were not geographically determined, but based on the genebanks.
6. USDA's accessions were more diverse than those of the IPK.
7. Twenty four markers were found to be significantly associated with one or more of the six curd-related traits.
8. Prediction abilities in the range of 0.128 to 0.652 and 0.09 to 0.66 using RRBLUP and Bayes B models was found.
9. No significant difference was found between RRBLUP and BayesB. RRBLUP is recommended for genomic selection in cauliflower breeding.
10. GBS is a suitable and powerful tool to uncover the genetic diversity and population structure of cauliflower, as well as to perform association mapping and genomic prediction in cauliflower specifically and in *Brassica oleracea* in general.
11. Imputation of missing values did not change the detected population structure or increase the accuracy of association mapping and genomic prediction.

12. Association mapping and genomic selection in combination with the GBS approach could be used to improve genetic gain in cauliflower breeding programs.

6 General summary

Low genetic diversity in cauliflower is one of the big challenges that hinders plant breeders in developing cultivars that meet the different needs of farmers (high yield, high adaptation and resistance to biotic and abiotic stresses) and consumers (high quality). The most efficient way to improve the performance of cauliflower varieties is to access large, diverse pools of genetic material of this crop. Therefore, the overall aims of this thesis have been to assess the phenotypic and genotypic diversity of a broad collection of genebank cauliflower accessions, as well as to perform association mapping and genomic prediction on this crop. To achieve this goal, a large collection of cauliflower accessions (200) was ordered from two genebanks (USDA in the USA and IPK in Germany). They were phenotyped at two locations and over three growing seasons and genotyped with GBS (genotyping by sequencing).

As consumers nowadays are becoming more interested in organic products and more growers are switching their production to organic farming, it has become inevitable that we try to answer two questions: 1) Can cultivars developed for conventional farming perform well in organic farming? and 2) Should breeders establish new cultivars that are adapted explicitly to organic farming? To answer these questions, we evaluated 200 accessions of cauliflower in two locations − one organic farm and one conventional farm − in Southern Germany, a region well-known for cruciferous vegetables production. The evaluation was conducted over three consecutive growing seasons (June 2011, April 2012 and August 2012).

The results of this study indicate that the cultivation method (organic vs conventional) had no major effect on curd-related traits (curd width and number of days to budding). But this finding might be explained by the fact that the two farms were both in a very fertile cabbage region in southern Germany and had very similar soil types and microclimatic conditions. The comparison between the three planting dates shows that the August 2012 season was suboptimal and was the only growing season showing a significant difference between the two cultivation methods for curd-width traits. However, this study does report significant genotype x cultivation method interactions, thus demonstrating a difference in performance of accessions under the two cultivation methods. This in turn means that genotypes can be selected for good performance under organic conditions. Moreover, we report a low genotypic correlation between the two cultivation methods and low heritability for curd width. On the other hand, this investigation

found a high genotypic correlation between the two cultivation methods and high heritability for the number of days to budding. In addition, the efficiency of indirect selection was quite high (0.95) for number of days to budding, but was low (0.39) for curd width. These findings suggest that selection for curd width should be conducted under organic farming conditions, while selection for number of days to budding can be performed under either organic or conventional farming systems. This thesis also identifies some genotypes that could be used as starting materials for breeding under organic farming conditions.

Using earlier marker-based genotyping systems, several studies have reported that cauliflower is characterized by a notably low genetic diversity and high similarity. These studies concluded that the development of highly polymorphic marker systems is needed to reveal the differentiation in cauliflower genotypes. We therefore examined the genetic diversity and population structure in a wide range of cauliflower genebank materials, employing a promising approach of next-generation sequencing technology, i.e. genotyping by sequencing (GBS).

Our results show that the collection analyzed (174 cauliflower accessions) was characterized by low genetic diversity and was surprisingly clustered into two distinct groups that were not geographically structured but rather characterized by seed source (genebanks). Also, our results indicated that the USDA's accessions were much more diverse than that of the IPK. According to the genetic analyses of this study, we can conclude that both genebanks, especially that of the IPK, should rethink their regeneration methods in order to keep the present level of diversity in cauliflower accessions available to breeders. However, the low genetic diversity of the genebank material could also be intrinsic to cauliflower itself, as it is one of the crops with a very low genetic diversity. Further, such low diversity could be due to the selection strategy of genebanks. To investigate this further, we tested signatures of selection using three different software programss (LOSITAN, ARLEQUIN and Bayenv2). Interestingly, these three tests detected several loci as outliers. These outliers represent regions of candidate genes exhibiting signatures of selection, which reveals that selection may indeed have an impact on the diversity of the genebank material. This finding shows that the conservation procedures of genebanks may have a strong influence on the conserved materials.

Genetic markers which were obtained by GBS in the current study allowed us to study the diversity as well as the correlation between phenotypic and genotypic data. Association mapping

and genomic prediction were used successfully to detect the genetic basis and predict the genomic breeding values of complex traits in several crop species. Neither of these two methods has so far been investigated for yield or yield-related traits in cauliflower. This study therefore investigated the feasibility of association mapping and genomic prediction in cauliflower gene bank accessions. The results indicate that GBS can be used effectively to generate enough markers for association mapping and genomic prediction studies in cauliflower germplasm. Also, by applying GWAS to six different traits, we found that 24 SNPs were associated significantly with the different curd traits. In addition, our genomic prediction results report low, medium and high prediction abilities in cauliflower. These findings imply that association mapping and genomic selection might be beneficial genomic approaches for complex traits in cauliflower breeding.

The current study shows that GBS is a powerful and suitable tool for revealing population structure and genetic diversity as well as for performing association mapping and genomic prediction in cauliflower. GBS yielded 120,693 SNPs, which confirms the flexibility of GBS and its ability to produce high density information at low cost in *Brassicaceae*. Furthermore, the current study found no big difference between the non-imputed data and imputed data with regard to population structure. Also, neither association mapping nor genomic prediction ability was improved by using the imputed data compared to non-imputed data. There was no difference in prediction ability found between the imputation methods. One explanation for this finding is that the imputation methods used here rely on linkage disequilibrium (LD), which was found to be low in this study. All of these findings indicate that non-imputed GBS data can be applied to study population structure, association mapping and genomic prediction without loss of accuracy of these analyses.

Finally, the findings laid out in this thesis show that GBS could be used effectively for several tasks in the commercial improvement of cauliflower in particular and in B. oleraceae in general, without the need for imputation of missing values. These tasks include genetic diversity analysis, marker trait association and genomic prediction. Also, the current thesis affirms that both association mapping and genomic prediction have the potential to accelerate the genetic gain in cauliflower breeding programs. In addition, several promising genetic materials (for high yield and stability) were found, which could be used as starting materials in cauliflower breeding programs, and especially for organic farming.

7 Zusammenfassung

In Blumenkohl ist geringe genetische Diversität eine der großen Schwierigkeiten mit der Pflanzenzüchter zu kämpfen haben um neue Sorten zu züchten, welche die Ansprüche von Landwirten (hoher Ertrag, gute Anpassung und Resistenz gegen biotischen und abiotischen Stress) und Verbrauchern (hohe Qualität) erfüllen. Für die Verbesserung von Blumenkohlsorten ist es daher wichtig diverse genetische Pools zugänglich zu machen.

Aus diesem Grund war die Bestimmung der phänotypischen und genetischen Diversität ein Hauptbestandteil dieser Dissertation. Des Weiteren wurde Assoziationskartierung und genomische Selektion von Genbankmaterial von Blumenkohl durchgeführt. Hierfür wurden 200 Blumenkohlakzessionen aus 2 Genbanken (USDA und IPK) bestellt und an zwei Standorten mit drei Pflanzterminen phänotypisiert. Darüber hinaus wurden alle Akzessionen mit Genotyping-by-Sequencing (GBS) genotypisiert.

Da der ökologische Landbau über die letzten Jahre stark an Beliebtheit gewonnen hat und immer mehr Landwirte auf diese Anbaumethode umsteigen, ist es wichtig folgende zwei Fragen zu beantworten: Sind konventionelle Sorten auch im ökologischen Landbau leistungsfähig, oder sollten speziell auf diese Anbauphilosophie angepasste Sorten entwickelt werden. Um diese Fragen zu beantworten haben wir 200 Blumenkohlakzessionen an zwei Standorten, einem konventionellen und einem biologisch-dynamischen Betrieb, in Süddeutschland evaluiert. Die Auswertung wurde über drei aufeinanderfolgende Pflanztermine (Juni 2011, April 2012, August 2012) durchgeführt.

Die Ergebnisse dieser Studie weisen darauf hin, dass die Anbaumethode (biologisch-dynamisch oder konventionell) keine starken Effekte auf den gemittelten Ertrag (Durchmesser des Blumenkohlkopfes) und den mittleren Blühzeitpunkt (Anzahl der Tage bis Knospung) hat. Dieses Ergebnis kann dadurch zustande gekommen sein, dass die beiden Standorte sich in einer bevorzugten Kohlanbauregion befinden, der Anbau also unter optimalen Bedingungen mit sehr nährstoffreicher Erde stattfand. Ein Vergleich der drei Pflanztermine zeigt, dass im August 2012 der Ertrag unter beiden Methoden suboptimal war und zugleich ist dies der einzige Anbau in dem die biologisch-dynamische Anbaumethode einen signifikant reduzierten Mittelwert für Ertrag aufzeigt. Trotz der geringen Unterschiede zwischen den Anbaumethoden, fand sich eine

signifikante "Genotypen x Anbaumethode-Interaktion“, welche auf eine unterschiedliche Leistungsfähigkeit der Akzessionen unter den beiden Anbaumethoden hinweist. Das bedeutet, dass die Ergebnisse eine Selektion von Akzessionen erlaubt, die unter biologisch-dynamischer Anbaumethode überdurchschnittlich hohen Ertrag bringen.

Darüber hinaus beschreiben unsere Ergebnisse für Ertrag eine geringe genetische Korrelation zwischen den zwei Anbaumethoden sowie eine niedrige Heritabilität. Für Blühzeitpunkt hingegen zeigte sich eine hohe genetische Korrelation zwischen den Anbaumethoden und eine hohe Heritabilität. Die daraus berechnete Effizienz für indirekte Selektion war gering für Ertrag (0.39) aber hoch für Blühzeitpunkt (0.95). Diese Ergebnisse zeigen, dass die Selektion auf Ertrag unter einer biologisch-dynamischen Anbaumethode durchgeführt werden sollte, während die Selektion auf Blühzeitpunkt unter beiden Anbaubedingungen stattfinden kann. Außerdem haben wir einige Genotypen identifiziert, welche als Ausgangsmaterial für biologisch-dynamische Zuchtprogramme genutzt werden könnten.

Mehrere Studien haben bereits berichtet, dass Blumenkohl durch eine geringe genetische Diversität gekennzeichnet ist und dass stark polymorphe Markersysteme benötigt werden, um die Differenzierung von Blumenkohlgenotypen festzustellen. Daher haben wir die genetische Diversität und Populationsstruktur von Blumenkohlgenbankmaterial mit Hilfe eines vielversprechenden Next Generation Sequencing Ansatzes, Genotying-by-Sequencing, (GBS) untersucht.

Unsere Ergebnisse beschreiben, dass die Kollektion von 174 Blumenkohlakzessionen eine geringe genetische Diversität hatte und sich erstaunlicherweise in zwei Untergruppen aufteilte. Diese Gruppierung stimmte jedoch nicht mit der geografischen Herkunft überein, sondern repräsentierte die zwei Genbanken von denen das Saatgut erhalten wurde. Darüber hinaus zeigten unsere Ergebnisse, dass die Akzessionen der USDA Genbank wesentlich diverser waren als Akzessionen der deutschen Genbank. Die geringe genetische Diversität kann verschiedene Ursachen haben. Teilweise zeigt sie sicher die bereits bekannte geringe Diversität dieser Art. Zusätzlich aber zeigten verschiedene Tests auf Selektion (LOSITAN, ARLEQUIN, Bayenv2), dass bestimmte Regionen des Genomes in den beiden Genbanken unter unterschiedlicher Selektion stehen. Dies zeigt, dass die Art und Weise, wie Akzessionen in Genbanken erhalten werden einen starken Einfluss auf das Material haben kann.

Genomische Marker, wie die durch GBS in dieser Studie erhalten, erlauben über Diversitätsstudien hinaus Untersuchungen zum Zusammenhang zwischen genetischen und phänotpyischen Unterschieden. Die Anwendung von Assoziationskartierung und genomischer Selektion für Ertrag und Ertragskomponenten von Blumenkohl wurde bisher nicht untersucht. Daher haben wir die Durchführbarkeit von Assoziationskartierung und genomischer Selektion in Blumenkohl mit Hilfe von Genbankmaterial untersucht. Die Ergebnisse weisen darauf hin, dass GBS effektiv dafür genutzt werden kann, genügend Marker für Assoziationskartierung und genomische Selektion in Blumenkohl zu generieren. Außerdem konnten wir durch die Anwendung von Assoziationskartierung auf sechs verschiedene Merkmale 24 SNPs identifizieren die signifikant mit Blütensprossmerkmalen assoziiert waren. Zusätzlich zeigten unsere Ergebnisse der genomischen Selektionsstudie, geringe, mittlere und hohe Vorhersagbarkeit für Blumenkohl. Diese Ergebnisse deuten an, dass Assoziationskartierung und genomische Selektion vorteilhafte genomische Methoden für komplexe Merkmale in der Blumenkohlzucht sein können.

Die Analyse der GBS Daten zeigt, dass genomische Ansätze, die das gesamte Genom in reduzierter Form repräsentieren sich sowohl dazu eignen die genetische Diversität von Blumenkohlvarianten zu studieren, als auch dazu Assoziationsstudien und Genomic-Prediction durchzuführen. Die GBS-Methode identifizierte 120,693 Punktmutationen, was die Fähigkeit von GBS zeigt, viel Information bei relativ geringen Kosten zur Verfügung zu stellen. Des weiteren zeigten unsere Versuche zur Imputation von fehlenden Markerdaten, dass im Falle von GBS die Imputation nicht zu einer Verbesserung der Ergebnisse führt. Dies galt sowohl für die Assoziationsstudie, als auch für die Genomic-Prediction Studie, die mit imputierten Daten keine verbesserte Vorhersage machen konnte. Eine Erklärung dafür, dass Imputation nicht zu einer Verbesserung führte könnte am geringen Kopplungsungleichgewicht in dieser Studie liegen, da die Imputation auf dieses basiert. Nichtsdestotrotz, zeigen unsere Ergebnisse, dass nicht imputierte GBS Daten sich eignen um die Populationsstruktur zu studieren, Assoziationskartierungen durchzuführen oder mittels Genomic-Prediction den Zuchterfolg vorherzusagen.

Zusammenfassend, zeigt diese Studie, dass GBS eine gute Technik ist um viele Aufgaben in der kommerziellen Züchtung in Blumenkohl im Speziellen und vermutlich auch im Genus *Brassica oleracea* zu unterstützen. Es bedarf dabei keiner Verbesserung der Daten durch Imputation. Die

zu lösenden Aufgaben umfassen die Analyse genetischer Diversität, die Evaluierung von Phänotyp-Genotyp-Assoziationen, sowie die Vorhersage von Züchtungserfolgen mittels Genomic-Prediction. Die Ergebnisse meiner Dissertation zeigen auch, dass sowohl Assoziationsstudien wie auch Genomic-Prediction das Potential haben die Blumenkohlzüchtung voranzubringen. Des Weiteren, konnten in dieser Studie Blumenkohl-Varianten identifiziert werden, die als gutes Startmaterial für ein Blumenkohl-Züchtungsprogramm speziell für biologisch-dynamischen Ackerbau verwendet werden können.

8 References

Abdurakhmonov IY, Abdukarimov A (2008) Application of association mapping to understanding the genetic diversity of plant germplasm resources. Int J Plant Genomics 2008:574927

Aggie-Horticulture (2014) Retrieved December 25, 2014, http://aggie-horticulture.tamu.edu/archives/parsons/publications/vegetabletravelers/broccoli.html

Almekinders CJM, Elings A (2001) Collaboration of farmers and breeders: participatory crop improvement in perspective. Euphytica 122:425-438

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

Astarini IA, Plummer JA, Lancaster RA, Yan G (2006) Genetic diversity of Indonesian cauliflower cultivars and their relationships with hybrid cultivars grown in Australia. Sci Hortic 108: 143-150

Bao Y, Vuong T, Meinhardt C, Tiffin P, Denny R et al. (2014) Potential of Association Mapping and Genomic Selection to Explore PI 88788 Derived Soybean Cyst Nematode Resistance. Plant Genome 7: doi:10.3835/plantgenome2013.11.0039

Banziger M, Cooper M (2001) Breeding for low input conditions and consequences for participatory plant breeding examples from tropical maize and wheat. Euphytica 122:503-519

Baenziger PS, Ibrahim S, Little RS, Santra DK, Regassa T, Wang MY (2011) Structuring an efficient organic wheat breeding program. Sustainability 3:1190-1205

Bastien M, Sonah H, Belzile F (2014) Genome wide association mapping of *Sclerotinia sclerotiorum* resistance in soybean with a genotyping by sequencing approach. Plant Genome 7:1-13

Börner A, Chebotar S, Korzun V (2000) Molecular characterization of the genetic integrity of wheat (*Triticum aestivum* L.) germplasm after long-term maintenance. Theor Appl Genet 100:494-497

Breiman L (2001) Random forests. Mach. Learn. 45: 5-32

Brown AF, Yousef GG, Chebrolu KK, Byrd RW, Everhart KW, Thomas A, Reid RW, Parkin IAP, Sharpe A, Oliver R, Guzman I, Jackson EW (2014) High-density single nucleotide polymorphism (SNP) array mapping in *Brassica oleracea*: identification of QTL associated with carotenoid variation in broccoli florets. Theor Appl Genet 127:2051-2064

Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing data inference for whole genome association studies by use of localized haplotype clustering. Am J Hum Genet 81:1084-97

Brütting C, Hensen I, Wesche K (2013) Ex situ cultivation affects genetic structure and diversity in arable plants. Plant Biol 15: 505-513

Cai D, Xiao Y, Yang W, Ye W, Wang B, Younas M, Wu J, Liu K (2014) Association mapping of six yield-related traits in rapeseed (*Brassica napus* L.). Theor Appl Genet 127:85-96

Ceccarelli S (1996) Adaptation to low high input cultivation. Euphytica 92:203-214

Chebotar S, Röder MS, Korzum V, Saal B, Weber WE, Borner A (2003) Molecular studies on genetic integrity of open-pollinating species rye (*Secale cereale* L.) after long-term gene bank maintenance. Theor Appl Genet 107:1469-1476

Crossa J (1995) Sample size and effective population size in seed regeneration of monoecious plants, pp.140-143. In: Engels, J. M. M., R. R. Rao Eds. Regeneration of seed crops and their wild relatives. Proceedings of a consultation meeting, 4-7 December 1995, ICRISAT, Hyderabad, India. IPGRI, Rome, Italy

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

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

Dawson JC, Murphy KM, Huggins DR, Jones S (2011) Evaluation of winter wheat breeding lines for traits related to nitrogen use under organic management. Org Agric 1:65-80

Edae EA, Byrne PF, Haley SD, Lopes MS, Reynolds MP (2014) Genome-wide association mapping of yield and yield components of spring wheat under contrasting moisture regimes. Theor Appl Genet 127:791-807

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

Ersoz ES, Yu J, Buckler ES (2007) Applications of linkage disequilibrium and association mapping. R.K. Varshney, R. Tuberosa (Eds.), Genomics-Assisted Crop Improvement. Vol. 1: Genomics Approaches and Platforms, Springer, New York. 97-120

FAO (2012) Food and Agriculture Organization of the United Nation. The Statistics Division, http://www.fao.org

FAO (2010) The Second Report on the State of the World's Plant Genetic Resources for Food and Agriculture. FAO, Rome, Italy

Fernie AR, Tadmor Y, Zamir D (2006) Natural genetic variation for improving crop quality. Curr Opin Plant Biol 9: 196-202

Finckh MR (2008) Integration of breeding and technology into diversification strategies for disease control in modern agriculture. Eur J Plant Pathol 121:399 -40

Flint-Garcia SA, Thornsberry JM, Buckler ES (2003) Structure of linkage disequilibrium in plants. Annu Rev Plant Biol 54:357-374

Flint-Garcia SA, Thuillet AC, Yu J, Pressoir G, Romero SM, et al (2005) Maize association population: a high-resolution platform for quantitative trait locus dissection. Plant J 44: 1054-1064

Frankham R, Ballou JD, Briscoe DA (2002) Introduction to conservation genetics. Cambridge University Press, Cambridge, UK

Fu Y-B (2014) Genetic diversity analysis of highly incomplete SNP genotype data with imputations: an empirical assessment. G3 4:891-899

Fu YB, Cheng B, Peterson GW (2014) Genetic diversity analysis of yellow mustard (*Sinapis alba* L.) germplasm based on genotyping by sequencing. Genet Resour Crop Evol 61: 579-594

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

Goldstein WA, Schmidt W, Burger H, Messmer M, Pollak LM, Smith ME, Goodman MM, Kutka FJ, Pratt RC (2012) Maize: Breeding and field testing for organic farmers. In: Lammerts van Bueren ET, and Myers JR(eds.) Organic Crop Breeding. Wiley-Blackwell, West Susse, pp. 175-189

Gömez OJ, Blair M, Frankow-Lindberg BE, Gullberg U (2005) Comparative study of common bean (*Phaseolus vulgaris* L.) landraces conserved *ex situ* in gene banks and *in situ* by farmers. Genet Resour Crop Evol 52:371-380

Gu Y, Zhao QC, Sun DL, Song WQ (2008) A genetic linkage map based on AFLP and NBS markers in cauliflower (*Brassica oleracea var. botrytis*). Bot Stud 49:93-99

Hagenblad J, Zie J, Leino MW (2012) Exploring the population genetics of genebank and historical landrace varieties. Genet Resour Crop Evol 59:1185-1199

Hasan M, Friedt W, Freitag NM, Link K, Pons-Kühnemann J, Snowdon RJ (2008) Association of gene-linked SSR markers to seed glucosinolate content in oilseed rape (*Brassica napus* ssp. *napus*). Theor Appl Genet 116:1035-1049

Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009) Invited review:Genomic selection in dairy cattle. Progress and challenges. J Dairy Sci 92:433-443

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

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

Howie BN, Donnelly P, Marchini J (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet 5:e1000529

Huang Y-F, Poland JA, Wight CP, Jackson EW, Tinker NA (2014) Using Genotyping-By-Sequencing (GBS) for Genomic Discovery in Cultivated Oat. PLoS One 9(7): e102448

Izzah NK, Lee J, Perumal S, Park JY, Ahn K, Fu D, Kim GB, Nam YW, Yang TJ (2013) Microsatellite-based analysis of genetic diversity in 9 commercial Brassica oleracea L. cultivars belonging to six varietal groups. Genet Resour Crop Evol 60:1967-1986

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

Jarquín D, Kocak K, Posadas L, Hyma K, Jedlicka J, Graef G, Lorenz A (2014) Genotyping by sequencing for genomic prediction in a soybean breeding population. BMC Genomics 15:740

Jestin C, Lodé M, Vallée P, Domin C, Falentin C, Horvais R, Coedel S, Manzanares-Dauleux M, Delourme R (2011) Association mapping of quantitative resistance for Leptosphaeria maculans in oilseed rape (Brassica napus L.). Mol Breed 27:271-287

Kasso M, Balakrishnan M (2013) Ex Situ Conservation of Biodiversity with Particular Emphasis to Ethiopia. ISRN Biodiversity. Doi:10.1155/2013/985037

Kirk AP, Fox SL, Entz MH (2012) Comparison of organic and conventional selection environments for spring wheat. Plant Breed 131:687-694

Kirsh VA, Peters U, Mayne ST (2007) Prospective study of fruit and vegetable intake and risk of prostate cancer. J Natl Cancer Inst 99:1200-1209

König S, Simianer H, Willam A (2009) Economic evaluation of genomic breeding programs. J Dairy Sci 92:382-391

Kushad MM, Brown AF, Kurlich AC, Juvik JA, Klein BP, Wallig MA, Jeffery EH (1999) Variation of glucosinolates in vegetable crops of *Brassica oleracea.* J Agric Food Chem 47:1548-1571

Lammerts van Bueren ET, Jones SS, Tamm L, Murphy KM, Myers JR, Leifert C, Messmer MM (2011) The need to breed crop varieties suitable for organic farming, using wheat, tomato, and broccoli as examples: A review. NJAS-Wageningen J Life Sci 58:193-205

Lan TH, Paterson AH (2000) Comparative mapping of quantitative trait loci sculpting the curd of *Brassica oleracea*. Genetics 155: 1927-1954

Lan TH, Paterson AH (2001) Comparative mapping of QTLs determining the plant size of Brassica oleracea. Theor Appl Genet 103:383-397

Lee SA, Fowke JH, Lu W, Ye C, Zheng Y, Gu K, Gu YT, Gao XO, Shu X, Zheng W (2008) Cruciferous vegetables, the GSTP1 IIe105Val genetic polymorphism, and breast cancer risk. Am J clinnutr 87:753-760

Li H, Peng Z, Yang X, Wang W, Fu J, Wang J, Han Y, Chai Y, Guo T,Yang N, Liu J, Warburton ML, Cheng Y, Hao X, Zhang P, Zhao J, Liu Y, Wang G, Li J, Yan J (2013) Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. Nat Genet 45:43–50

Li F, Chen B, Xu K, Wu J, Song W, Bancroft I, Harper AL, Trick M, Liu S, Gao G, Wang N, Yan G, Qiao J, Li J, Li H, Xiao X, Zhang T, Wu X (2014) Genome-wide association study dissects the genetic architecture of seed weight and seed quality in rapeseed (*Brassica napus* L.). DNA Res. doi:10.1093/dnares/dsu002

Lorenz AJ, Chao S, Asoro FG, Heffner EL, Hayashi T, Iwata H, Smith KP, Sorrels ME, Jannink J-L (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 120:151-161

Louarn S, Torp AM, Holme IB, Andersen SB, Jensen BD (2007) Database derived microsatellite markers (SSRs) for cultivar differentiation in *Brassica oleracea*. Genet Resour Crop Evol 54: 1717-1725

Lu F, Lipka AE, Elshire RJ, Glaubitz JC, Cherney J, et al. (2013) Switchgrass genomic diversity, ploidy and evolution: novel insights from a network-based SNP discovery protocol. PLoS Genet 9: e1003215

Maggio A, S. De Pascale, R Paradiso, G Babieri (2013) Quality and nutritional value of vegetables from organic and conventional farming. Sci Hort 164: 532-539

Marchini J, Howie B (2010) Genotype imputation for genome-wide association studies. Nature Rev Genet 11:499-511

Marchini J, Howie B, Myers S, McVean G, Donnelly PA (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 39:906–913

Mason HE, Navabi A, Frick BL, et al (2007) The weed competitive ability of Canada Western Red Spring Wheat cultivars grown under organic management. Crop Sci 47:1167-1176

Maxted N, Ford-Lloyd BV, Hawkes JG (1997) Complementary conservation strategies. In: Maxted, N., Ford-Lloyd, B.V., Hawkes, J.G. eds. Plant Genetic Conservation: The In-situ Approach. Chapman and Hall, London, pp. 15-40

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

Moreau L, Charcosset A, Hospital F, Gallais A (1998) Marker-assisted selection efficiency in populations of finite size. Genetics 148: 1353-1365

Morris GP, Ramu P, Deshpande SP, Hash CT, Shah T, et al. (2013) Population genomic and genome-wide association studies of agroclimatic traits in sorghum. PNAS 110: 453-458

Murphy KM, Campbell KG, Lyon SR, Jones SS (2007) Evidence of varietal adaptation to organic farming systems. Field Crop Res 102:172-177

Myres JR, McKenzie, Voorrips RE (2012) Brassica: Breeding Cole crops for organic agriculture. In: Lammerts van Bueren ET, and Myers JR(eds.) Organic Crop Breeding. Wiley-Blackwell, West Sussex, pp 251-262

Nordborg M, Weigel D (2008) Next-generation genetics in plants. Nature 456: 720-723

Norton GJ, Douglas A, Lahner B, Yakubova E, Guerinot ML, et al. (2014) Genome Wide Association Mapping of Grain Arsenic, Copper, Molybdenum and Zinc in Rice (*Oryza sativa* L.) Grown at Four International Field Sites. Plos One 9:89685

Ould Estaghvirou SB, Ogutu JO, Schulz-Streeck T, Knaak C, Ouzunova M, et al. (2013) Evaluation of approaches for estimating the accuracy of genomic prediction in plant breeding. BMC Genomics 14:860

Parzies HK, Spoor W, Ennos RA (2000) Genetic diversity of barley landrace accessions (*Hordeum vulgare ssp. Vulgare*) conserved for different lengths of time in ex situ gene banks. Heredity 84:476-486

Poland J, Rife TW (2012) Genotyping-by-sequencing for plant breeding and genetics. Plant Genome 5: 92-102

Poland J, Brown PJ, Sorrells ME, Jannink J (2012a) Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS One 7:e32253

Poland J, Endelman J, Dawson J, Rutkoski J, Wu S, et al. (2012b) Genomic selection in wheat breeding using genotyping-by-sequencing. Plant Genome 5: 103-113

Rao VR, Hodgkin T (2002) Genetic diversity and conservation of plant genetic resources. Plant Cell Tissue Organ Cult 68:1-19

Reid T, Yang R-C, Salmon DF, Spaner D (2009) Should spring wheat breeding for organically managed systems be conducted on organically managed land? Euphytica 169:239-252

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

Rezaeizad A, Wittkop B, Snowdon R, Hasan M, Mohammadi V, et al. (2011) Identification of QTLs for phenolic compounds in oilseed rape (*Brassica napus* L.) by association mapping using SSR markers. Euphytica 177:335-342

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

Riedelsheimer C, Technow F, Melchinger AE (2012b) Comparison of whole-genome prediction models for traits with contrasting genetic architecture in a diversity panel of maize inbred lines. BMC Genomics 13:452

Romay MC, Millard M, Glaubitz JC, Peiffer JA, Swarts KL et al. (2013) Comprehensive genotyping of the USA national maize inbred seed bank. Genome Biol 14: R55

Rucinska A, Puchalski J (2011) Comparative molecular studies on the genetic diversity of an *ex situ* garden collection and its source population of the critically endangered Polish endemic plant *Cochlearia polonica* E. Fröhlich. Biodivers Conserv 20: 401-413

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

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

Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 78:629-644

Schonhof I, Krumbein A, Brückner B (2004) Genotypic effects on glucosinolates and sensory properties of broccoli and cauliflower. Nahrung/Food 48:25-33

Seufert V, Ramankutty N, Foley JA (2012) Comparing the yields of organic and conventional agriculture. Nature 485:229-234

Sonah H, Bastien M, Iquira E, Tardivel A, Légaré G, et al (2013) An Improved Genotyping by Sequencing (GBS) Approach Offering Increased Versatility and Efficiency of SNP Discovery and Genotyping. PLoS One 8: e54603.

Sonah H, O'Donoughue L, Cober E, Rajcan I, Belzile F (2014) Identification of loci governing eight agronomic traits using a GBS-GWAS approach and validation by QTL mapping in soya bean. Plant Biotechnol J doi:10.111/pbi.12249

Stapley J, Reger J, Feulner PGD, Smadja C, Galindo J, Ekblom R, Bennison C, Ball AD, Beckerman AB, Slate J (2010) Adaptation genomics: the next generation. Trends Ecol. Evol. 25:705-712

Tang L, Gary RZ, Guru K, Kirsten BM, Zhang Y, Ambrosone CB, McCann SE (2008) Consumption of raw cruciferous vegetables is inversely associated with bladder cancer risk. Cancer Epidemiol Biomarkers Prev 17: 938-44

Tardivel A, Sonah H, Belzile F, O'Donoughue LS (2014) Rapid identification of alleles at the soybean maturity gene E3 using genotyping by sequencing and a haplotype-based approach. Plant Genome 7:1-9

Truong HT, Ramos AM, Yalcin M, de Ruiter M, van der Poel HJA, Huvenaars KHJ (2012) Sequence-based genotyping for marker discovery and co-dominant scoring in germplasm and populations. PLoS One 7:e37565

Tonguc M, Griffiths PD (2004) Genetic relationships of Brassica vegetables determined using database derived sequence repeats. Euphytica 137:193-201

van Hintum TJL, Van De wiel CCM, Visser DL, Van Treuren R, Vosman B (2007) The distribution of genetic diversity in a *Brassica oleracea* genebank collection related to the effects on diversity of regeneration, as measured with AFLPs. Theor Appl Genet 114:777-786

VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel DR et al. (2009) Invited review: Reliability of genomic prediction for north American holstein bulls. H Dairy Sci 92:16-24

Walley PG, Carder J, Skipper E, Mathas E, Lynn J, Pink D, Buchanan-Wollasto V (2012) A new broccoli × broccoli immortal mapping population and framework genetic map: tools for breeders and complex trait analysis. Theor Appl Genet 124:467-484

Wolfe MS, Baresel JP, Desclaux D, Goldringer I, Hoad S, Kovacs G, Lo schenberger F, Miedaner T, Østergard H, Lammerts van Bueren ET (2008) Developments in breeding cereals for organic agriculture. Euphytica 163:323-346

Wright SI, Gaut BS (2005) Molecular population genetics and the search for adaptive evolution in plants. Mol Biol Evol 22:506-519

Würschum T, Abel S, Zhao Y (2014) Potential of genomic selection in rapeseed (*Brassica napus* L.) breeding. Plant Breed 133:45-51

Würschum T, Reif JC, Kraft T, Janssen G, Zhao Y (2013) Genomic selection in sugar beet breeding populations. BMC Genetics 14: 85-93

Yu J, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 2:203-208

Zamir D (2001) Improving plant breeding with exotic genetic libraries. Nat Rev Genet 2:983-989

Zhao Z, Gu H, Sheng X, Yu H, Wang J, Zhao J, Cao J (2014) Genetic diversity and relationships among loose-curd cauliflower and related varieties as revealed by microsatellite markers. Sci Hort 166: 105-110

Zou J, Jiang C, Cao Z, Li R, Long Y, Chen S, Meng J (2010) Association mapping of seed oil content in *Brassica napus* and comparison with quantitative trait loci identified from linkage mapping. Genome 53:908-916

9 Acknowledgements

First and foremost, I would like to express my sincere gratitude to my academic supervisor Prof. Dr. Karl Schmid for giving me the opportunity to work in his scientific group, his advice, suggestions and continuous support during this thesis work.

Sincere thanks to my colleagues and co-authors, Dr. Christain Lampei, Thomas Müller and Patrick Thorwarth for their assistance in the data analysis and proofreading of my work.

Many thanks to all of my colleagues at group of "Crop Biodiversity and Breeding Informatics (350b)" whom I shared the good times and enjoyed four years working and also for all kind of help, interest and support I received from them.

I am grateful to Mrs. Elisabeth Kokai-Kota for for her stimulating support and work in the laboratory. Also, many thanks to Mrs. Viola Abraham for her help in the field trials.

Especial thanks to Egyptian ministry of higher education and German Academic Exchange Service (DAAD) for financial support.

Last but the least, I am more than thankful to my father, my mother, my wife, my daughter "Habiba" and my son "Hozifa" for their endless love and support.

10 Curriculum vitae

Personal information

Name	Eltohamy Ali Ahmed Yousef
Date of Birth	17 August 1982
Place of Birth	Kafrelsheikh, Egypt

Education

School education	1988–1992	Elementary school (Alaiash elementary school, Egypt)
	1994–1996	Secondary school (Alzahraa secondary school, Egypt)
	1997–1999	High school (Baltim high school, Egypt)
University education	10/1999- 7/2003	B.Sc. in Agricultural Sciences, University of Suez Canal, Egypt
	10/2004 - 12/2007	M.Sc. in Horticultural Sciences, University of Suez Canal, Egypt
	4/2011 - 05/2015	Ph.D. Student, Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim
Work experience	4/2004 – 12/2007	Teaching assistant at Department of Horticulture, Faculty of Agriculture, Suez Canal University, Ismailia, 41522, Egypt
	1/2008 - 10/2010	Assistant lecturer at Department of Horticulture, Faculty of Agriculture, Suez Canal University, Ismailia, 41522, Egypt Ismailia, 41522, Egypt

11 Erklärung

Hiermit erkläre ich an Eides statt, dass die vorliegende Arbeit von mir selbst verfasst und lediglich unter Zuhilfenahme der angegebenen Quellen und Hilfsmittel angefertigt wurde. Wörtlich oder inhaltlich übernommene Stellen wurden als solche gekennzeichnet.

Die vorliegende Arbeit wurde in gleicher oder ähnlicher Form noch keiner anderen Institution oder Prüfungsbehörde vorgelegt.

Insbesondere erkläre ich, dass ich nicht früher oder gleichzeitig einen Antrag auf Eröffunung eines Promotionsverfahrens unter Vorlage der hier eingereichten Dissertation gestellt habe.

Stuttgart-Hohenheim, Januray 2015

Eltohamy Ali Ahmed Yousef