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.

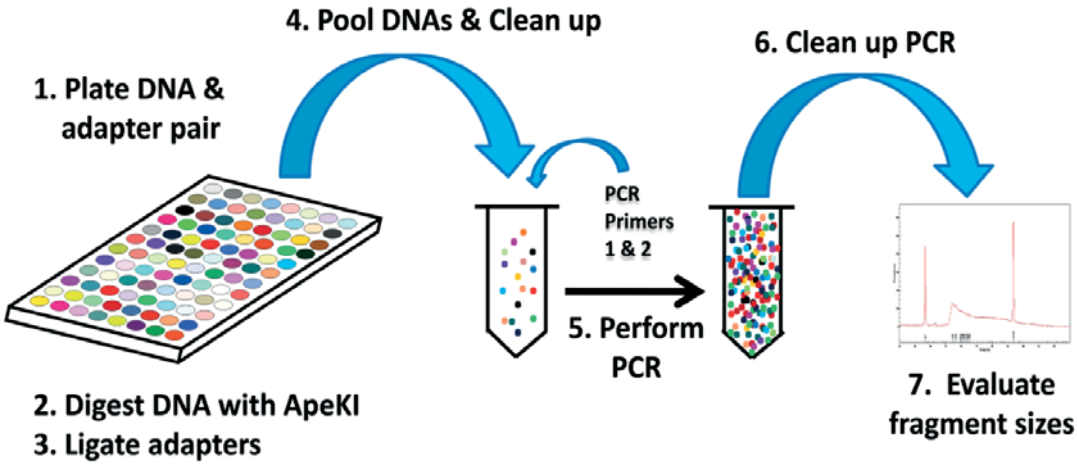

Figure 1.1 Basic scheme of used protocol for performing GBS (see Elshire et al. 2011)

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).

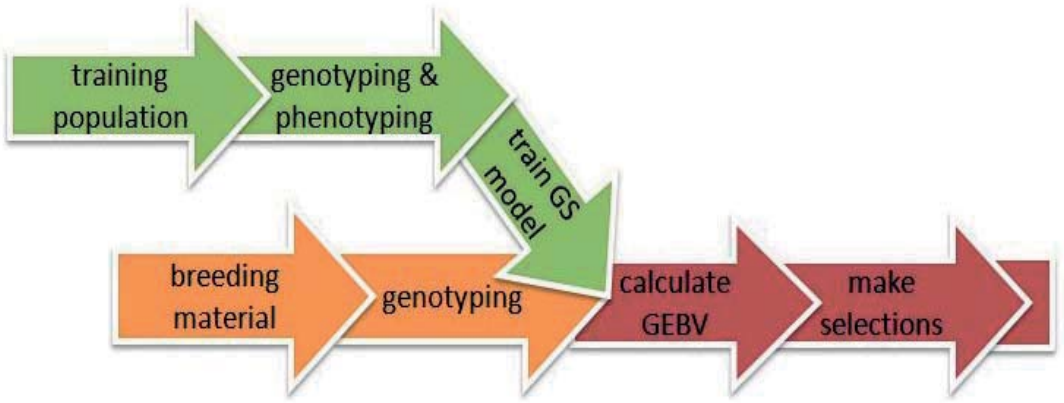

Figure 1.2 The basic scheme for GS, starting from the training population and selection candidates continuing through to genomic estimated breeding value (GEBV)–based selection (modified from Heffner et al. 2009).

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.