RESEARCH ARTICLE

Comparative analysis of genetic diversity and differentiation of cauliflower (*Brassica oleracea* var. *botrytis*) accessions from two *ex situ* genebanks

Eltohamy A. A. Yousef[1,2☯], Thomas Müller[1☯¤], Andreas Börner[3], 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** Research (IPK), Seeland OT Gatersleben, Germany

☯ These authors contributed equally to this work.
¤ Current address: Universitätsspital Zurich, Zurich, Switzerland
* karl.schmid@uni-hohenheim.de

Abstract

Cauliflower (*Brassica oleracea* var. *botrytis*) is an important vegetable crop for human nutrition. We characterized 192 cauliflower accessions from the USDA and IPK genebanks with genotyping by sequencing (GBS). They originated from 26 different countries and represent about 44% of all cauliflower accessions in both genebanks. The analysis of genetic diversity revealed that accessions formed two major groups that represented the two genebanks and were not related to the country of origin. This differentiation was robust with respect to the analysis methods that included principal component analysis, ADMIXTURE and neighbor-joining trees. Genetic diversity was higher in the USDA collection and significant phenotypic differences between the two genebanks were found in three out of six traits investigated. GBS data have a high proportion of missing data, but we observed that the exclusion of single nucleotide polymorphisms (SNPs) with missing data or the imputation of missing SNP alleles produced very similar results. The results indicate that the composition and type of accessions have a strong effect on the structure of genetic diversity of *ex situ* collections, although regeneration procedures and local adaptation to regeneration conditions may also contribute to a divergence. F_{st}-based outlier tests of genetic differentiation identified only a small proportion (<1%) of SNPs that are highly differentiated between the two genebanks, which indicates that selection during seed regeneration is not a major cause of differentiation between genebanks. Seed regeneration procedures of both genebanks do not result in different levels of genetic drift and loss of genetic variation. We therefore conclude that the composition and type of accessions mainly influence the level of genetic diversity and explain the strong genetic differentiation between the two *ex situ* collections. In summary, GBS is a useful method for characterizing genetic diversity in cauliflower genebank material and our results suggest that it may be useful to incorporate routine genotyping into accession management and seed regeneration to monitor the diversity present in *ex situ* collections and to reduce the loss of genetic diversity during seed regeneration.

OPEN ACCESS

Citation: Yousef EAA, Müller T, Börner A, Schmid KJ (2018) Comparative analysis of genetic diversity and differentiation of cauliflower (*Brassica oleracea* var. *botrytis*) accessions from two *ex situ* genebanks. PLoS ONE 13(2): e0192062. https://doi.org/10.1371/journal.pone.0192062

Editor: Manoj Prasad, National Institute of Plant Genome Research, INDIA

Received: October 9, 2017

Accepted: January 16, 2018

Published: February 8, 2018

Data Availability Statement: All 3 files are available from the Zenodo database (DOI 10.5281/zenodo.887955).

Funding: This work was supported by F. W. Schnell Endowed Professorship of the Stifterverband (www.stifterverband.org) and GERLS Fellowship of the German Academic Exchange Service (DAAD).

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 seed regeneration [2].

Ex situ conserved plant genetic resources (PGR) are plant genotypes that are stored in central storage facilities. PGR are utilized to improve modern cultivars by the introgression of new and exotic genetic variation into breeding pools (e.g., [3]). However, PGR often experience a loss of genetic diversity, stronger inbreeding depression (especially in outcrossing crops) and accumulation of deleterious alleles because of small population sizes of individual genebank accessions. These processes may negatively affect the success of *ex situ* conservation after several regeneration cycles [4, 5]. In addition, strong selection caused by adaptation to the seed regeneration environment may further reduce genetic variation.

Cauliflower (*Brassica oleracea* var. *botrytis*) is an important vegetable crop worldwide and a valuable component of a healthy diet because of a high content of glucosinolates with anticancer properties [6, 7]. Cauliflower and broccoli are currently cultivated worldwide on about 1.2 Mio hectares, with an annual production of over 21 Mio. tons [8]. Genetic diversity of cauliflower was analyzed with a diversity of marker systems like amplified polymorphic DNA (RAPD; [9]) or simple sequence repeats (SSRs; [10, 11]). These initial genotyping studies indicated that genetic diversity for cauliflower was limited [11–13]. More recently, whole genome resequencing revealed the genetic structure of cauliflower germplasm and identified genomic regions with low and high genetic diversity, respectively [14]. Whole genome sequencing approaches are still too expensive for large numbers of genebank accessions, or not necessary because smaller polymorphism numbers are sufficient for the analysis of diversity and genetic relationships. Reduced representation sequencing methods such as genotyping-by-sequencing (GBS) are a cost-effective alternative because these methods allow a high degree of multiplexing and tens of thousands of polymorphisms are identifed in a single reaction without the need of a reference sequence [15–17]. 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 [18–20]. In Brassicaceae, GBS was used to analyse genetic diversity in yellow mustard [21].

The density of polymorphisms identified by reduced representation sequencing methods like GBS are sufficient to conduct genome-wide analyses of diversity and to construct core collections for further phenotyping or breeding. Our objective was to use GBS for assessing the genetic diversity and genetic relationship of randomly selected accessions of the USDA and IPK *ex situ* genebanks, which harbor large collections of cauliflower accessions. We also investigated whether imputation of missing genotypes improves diversity estimates and whether genetic and phenotypic diversity among cauliflower accessions are correlated. Cauliflower is a predominately outcrossing species whose pollination depends on insects. A self-incompatibility (SI) system prevents self-fertilization, but high variation in the extent of SI was reported in different landraces and varieties [22]. This variation has been used to select highly inbred and highly homozygous lines for breeding. Both SI and cytoplasmatic sterility (CMS) mechanisms are employed in hybrid breeding of modern varieties, which is now the predominant breeding method. For these reasons, genebank accessions of cauliflower may be highly variable in their genetic diversity. In the present study, we focused on characterising species-wide diversity and included only a single individual of each accession to maximise diversity of number of accessions given the resources available. We observed a high level of population structure among accessions, and in particular a strong genetic differentiation between accessions from the two genebanks.

Materials and methods

Plant material

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. They represent 47% (100 of 212) of the USDA and 40% (91 of 227) of the IPK cauliflower accessions, respectively. We selected accessions from these two large genebanks because they harbor large collections of cauliflower that are expected to reflect the worldwide diversity of this vegetable crop. According to the passport information, the sample consists of traditional cultivars, breeding material, hybrids, unverified genotypes, collector material, commercial vegetable seeds and landraces (Table 1). Accessions originate from 26 countries and 11 accessions are of unknown geographic origin. In addition, a single plant of the wild type of *Brassica oleracea* was obtained from Heidelberg Botanic Garden and Herbarium (HEID), Germany. All accessions of this study including accession type and country of origin are listed in Table A in S2 File.

Field experiment and phenotypic measurements

All accessions were evaluated for six morphological traits at six environments consisting of two cultivation methods (organic and conventional) and three growing seasons (June 2011 and April/August 2012) in a randomized complete block design (RCBD) with two replicates in each environment [23]. 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 Patterson (2000) [24]. Phenotypic values were calculated as mean over locations and seasons.

DNA extraction and genotyping by sequencing

Genomic DNA was extracted from leaf tissue three weeks after sowing in the green house from a single individual of each accession according to a standard CTAB protocol [25]. The

Table 1. Counts of different types of genebank accessions genotyped.

Accession type	Genebank		Combined sets
	USDA	IPK	
Traditional cultivars	3	79	82
Breeding material	0	4	4
Hybrids	0	1	1
Unverified genotypes	92	1	93
Collector materials	4	0	4
Commercial vegetable seeds	1	0	1
Landraces	0	6	6
Wild relative	0	0	1
Total	100	91	192

https://doi.org/10.1371/journal.pone.0192062.t001 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. GBS was carried out according to the protocol of Elshire et al. [15] with minor modifications. The DNA was digested with the ApeK1 restriction enzyme. A total of 96 barcodes were used, of which 64 barcodes were obtained with the web tool at http://www.deenabio.com/services/gbs-adapters and 32 barcodes were taken from [15]. All barcodes have an even distribution in length (4-8 nucleotides) and nucleotide composition of nucleotides at each position (Table B in S2 File). 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 [15]. 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.

Sequence data analysis

Sequence reads were filtered for sequencing artifacts and low quality reads with custom Python scripts, bwa [26] and FastQC [27]. Reads mapping to the PhiX genome, which is used for calibration in Illumina sequencing, 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, reads had a length of 88 bp. The pre-processed reads were then aligned to the genome of *Brassica oleracea* sp. *capitata* cabbage line 02-12 [28] with bwa. SNP calling was performed with SAMtools [29], bcfutils, vcfutils and custom Python scripts. The VCF file was parsed to retain 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 (Supplementary Note 1).

Analysis of genetic structure and genetic diversity

The genetic structure of the sample was investigated by various methods for comparison. They included principal component analysis (PCA), as implemented in the R package adegenet [30]. Also, a PCA of six morphological traits was performed with the prcomp function in the stats R package [31] and a multivariate analysis of variation (MANOVA) with the manova function of R. For post-hoc tests of phenotypic differentiation, a single-factor ANOVA was calculated and a Bonferroni correction applied by applying a critical threshold of $p = 0.05/6 = 0.0083$ to test results. The correlation between genetic and phenotypic distances based on the PCA was determined with a Mantel test [32] in the ade4 R package [33]. Principal coordinate analysis (PCoA) based on pairwise F_{st} values between genotypes calculated with R adegenet R package [30], was used for further analysis of population structure with the ape R package [34]. 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 [34]. Population structure was inferred with ADMIXTURE [35]. The number of subpopulations analyzed ranged from $K = 1–10$ and cross-validation was used to estimate the value of K which best fits the data [36]. An Analysis of Molecular Variance (AMOVA) was carried out with Arlequin v3.5.3.1 [37]. The extent of genetic differentiation (F_{st}) between the two genebanks was estimated with the pairwise.fst function in the R package adegenet [30]. For each group of accessions from USDA and IPK genebanks, we calculated the observed and expected heterozygosities, H_o and H_{exp}, as well as the inbreeding coefficient (F) with adegenet. We also calculated percent polymorphic loci, %P, and nucleotide diversity, π, with the R package pegas [38]. 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 [39].

Detection of outlier SNPs

Outlier tests for highly differentiated SNPs between the sets from the two genebanks (USDA and IPK) were based on a coalescent-based simulation method [40] implemented in LOSITAN [41]. This method identifies putative targets genes that differentiated in response to selection based on the distributions of expected heterozygosity and F_{st} values under an island model. LOSITAN was run in two steps: in the first step, an initial run with 50,000 simulations was performed with all SNPs and using the mean F_{st}. After excluding a candidate subset of selected SNPs determined in the initial run, the distribution of neutral F_{st} values were calculated. We also used Arlequin 3.5 [37] to detect outlier SNPs by accounting for a hierarchical genetic structure based on 10,000 simulations under a hierarchical island model. For both methods, SNPs outside the 99 and 1% confidence areas were identified as candidate genes potentially affected by directional and balancing selection, respectively. A further test of differentiation by selection was conducted with BayPass v2.1 [42] which is a fast implementation of the Bayenv2 algorithm [43]. We first estimated the population covariance matrix Ω, which is a variance-covariance matrix of allele frequencies, for the two groups of accessions from each genebank was calculated from the final run of the MCMC after 100,000 iterations. We used 360 (25%) randomly selected SNP markers from the SNP dataset without missing values. To evaluate the robustness of this matrix, we repeated the calculation with five randomly drawn SNP sets and calculated the average correlation coefficient of estimated variance and covariance parameters from all 10 pairwise comparisons of matrices. The resulting average correlation coefficient was 0.99, which indicates a high convergence of Ω matrices based on different random SNP sets. The Ω matrix was then used to control for the genome-wide genetic relationship among groups in the calculation of the X^TX statistics for each SNP by using 100,000 iterations of the MCMC. The X^TX statistic is equivalent to the F_{st} value as a measure of population differentiation.

Results

SNP identification by GBS analysis

To analyse genetic diversity in the sample we included one individual plant per genebank accession. The sequencing data of the 192 accessions consisted of 455 Mio. reads with a length of 100 bp. Quality filtering removed 5.2% of reads because they mapped to the PhiX genome or were of low quality, and read number per accession ranged from only 80 to 7.1 Mio with an average of 2.4 Mio reads. Eighteen accessions with less than 200,000 reads were excluded from further analysis because the proportion of missing data in these accessions was too high. A total of 133 Mio reads mapped to the *B. oleracea* reference genome (Fig 1). The percentage of mapped reads per genotype against *B. oleracea* ranged from 14% to 35% with an overall average of 29% (Table C in S2 File). 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 and 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

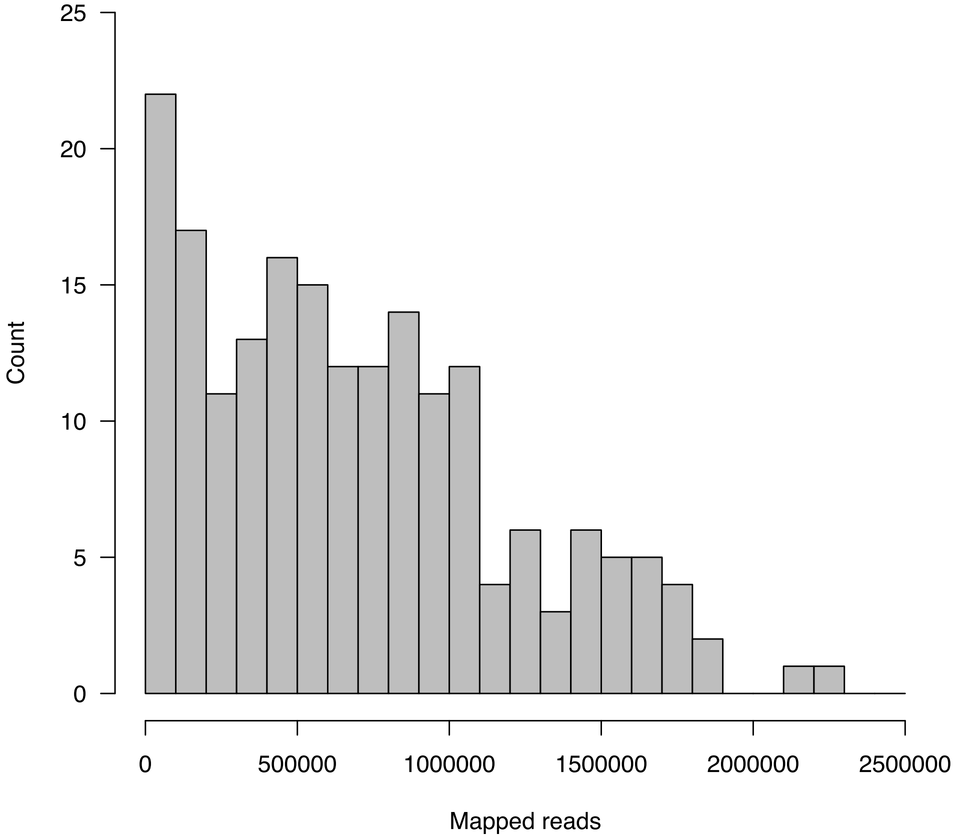

Fig 1. Histogram of read counts mapped to the reference genome.

https://doi.org/10.1371/journal.pone.0192062.g001

genotype. The number of SNPs and percentage of missing data per genotype are shown in Table D in S2 File.

Analysis of genetic population structure

The genetic structure of the whole collection (*n* = 174) was analyzed with PCA, PCoA and ADMIXTURE. Here, we present the results for the set of 1,444 SNPs without missing data, but the same results were obtained with the other two data sets with missing or imputed data (in both cases, *n* = 120,693), which are provided as Supplementary Information.

The marker-based PCA showed a clear differentiation between the two genebanks for the SNP data (Fig 2A). The first two axes explained 21% of the overall variance and separated accessions from the USDA and IPK genebanks. The SNP data sets with missing and imputed values showed the same genetic structuring (Fig A in S1 File).

Further analyses confirmed the genetic differentiation of accessions and reveal that the largest proportion of genetic variance is explained by the difference between the two genebanks. A PCoA based on pairwise F_{st} values 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 (Fig B in S1 File). A neighbor joining (NJ) tree based on a pairwise distance matrix separated the 174 accessions into two distinct groups (Fig 3A) representing the two genebanks. In both groups, accessions are not differentiated into well-supported subgroups that reflects the country of origin (Fig 3B). The NJ trees based on the SNP

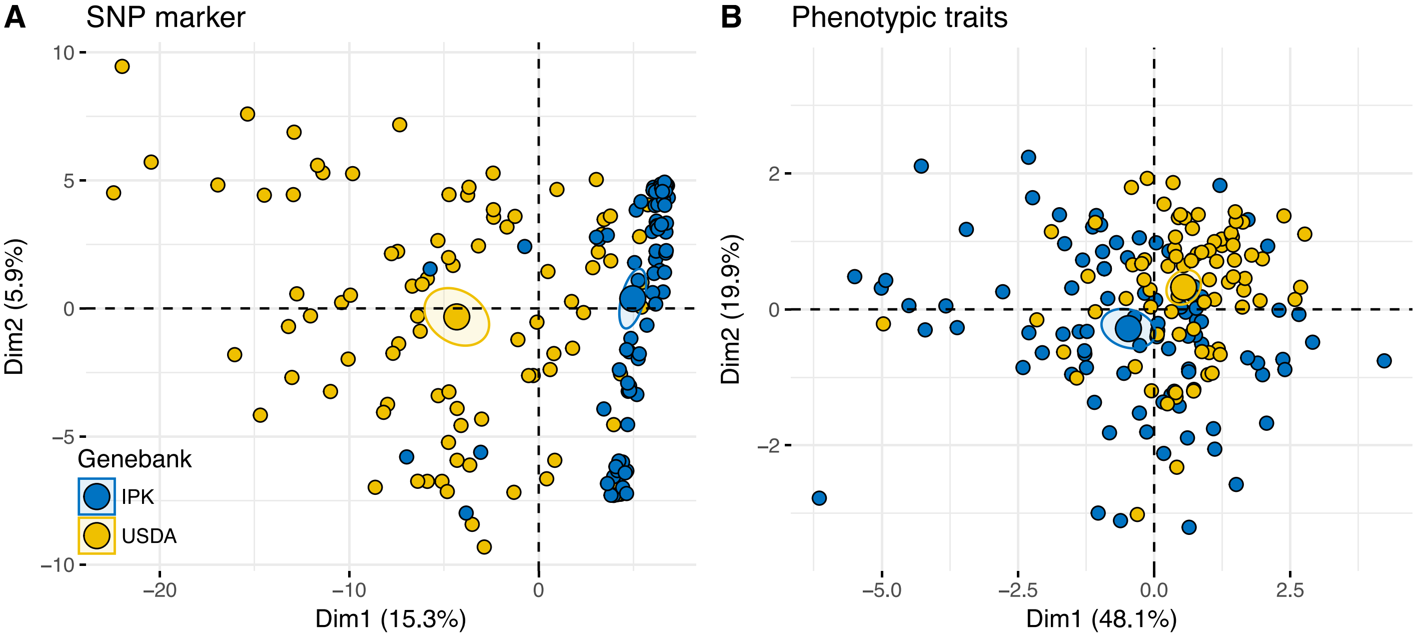

Fig 2. Principal component analysis (PCA) of SNP markers and phenotypic traits. (A) PCA of 1,444 GBS-derived SNPs without missing data in any of the accessions. (B) PCA of six phenotypic traits measured in two locations over three growing periods. In both PCA analyses, the two larger dots indicate the centroids and their confidence regions for the genebanks.

https://doi.org/10.1371/journal.pone.0192062.g002

data with missing or imputed data (120,693 SNPs in both data sets) confirm these results (Fig C in S1 File).

In addition to the differentiation between genebanks the previous analyses also indicate the presence of additional clusters. We therefore used ADMIXTURE to infer population structure and to estimate the number of genetic clusters that is most consistent with the data. For $K = 2$, two groups mainly differentiated between the two genebanks (Fig 4, Figs D and E in S1 File). Based on cross-validation, ADMIXTURE identified five genetically different clusters as most consistent with the GBS data without missing values With $K = 5$, the IPK accessions cluster

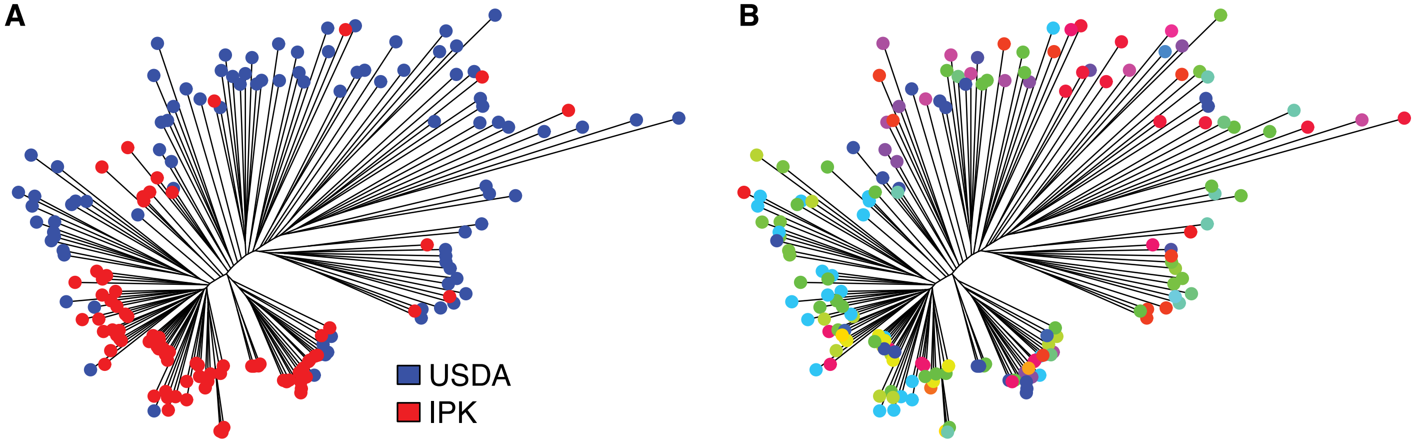

Fig 3. Neighbor joining tree of 174 cauliflower accessions. The NJ tree is based on the pairwise distance matrix using data without missing values (1,444 SNPs). (A) The two genebanks are represented as different colors. (B) Each country of origin of the original seeds as stated in the passport information is represented with a different color.

https://doi.org/10.1371/journal.pone.0192062.g003

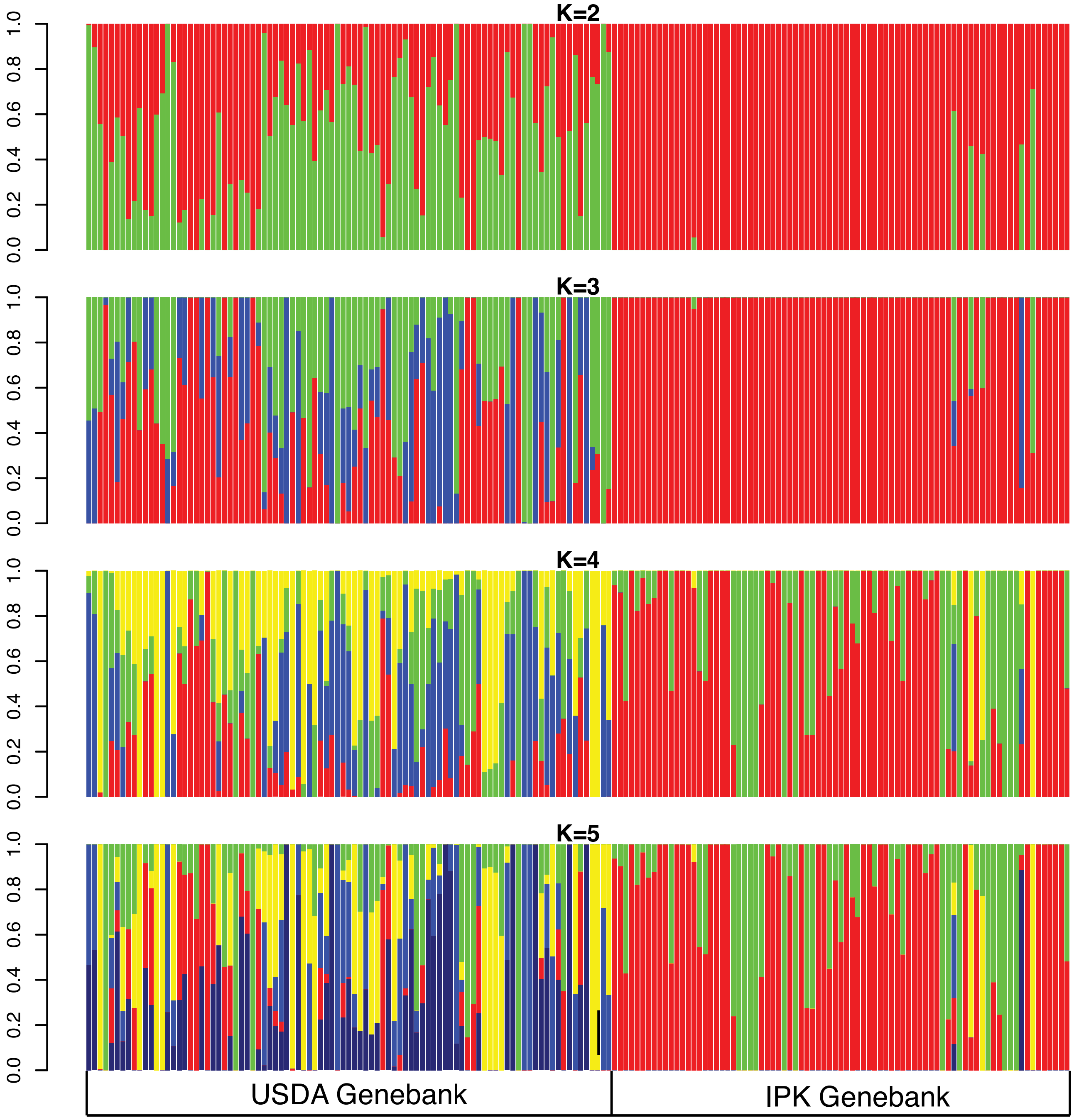

Fig 4. Genetic structure of 174 cauliflower genotypes inferred with ADMIXTURE. The number of predefined clusters ranged from $K = 2 - 6$ and was inferred using SNPs without missing values ($n = 1,444$).

https://doi.org/10.1371/journal.pone.0192062.g004

into two distinct groups with some degree of admixture, whereas the USDA accessions do not form distinct clusters and show a high level of admixture (Fig 4).

Despite the clear genetic differentiation between the two genebanks, a very large proportion of genetic variation (90%) segregates within rather than between genebanks (5%; AMOVA, $p < 0.001$; Table 2, Figs E and F in S1 File). This is consistent with a low overall genetic differentiation between the two genebanks ($F_{st} = 0.029$). The mean pairwise F_{st} of accessions within

Table 2. Analysis of molecular variance (AMOVA) of different groups based on SNP data without missing values.

Level of variation	Sum of Squares	Variance component	% variance explained	*P*-value
Among groups (genebanks)	1,032	5.43	6	<0.001
Within groups	31,448	91	94	<0.001
Total	32,489	96		

https://doi.org/10.1371/journal.pone.0192062.t002

each genebank was 0.301 for the USDA and 0.160 for the IPK accessions (estimated from 1,444 SNPs without missing values), showing that the USDA accessions are more differentiated from each other than the IPK accessions.

Levels of genetic diversity

For a further comparison between accessions from the two genebanks, we calculated various genetic diversity parameters (Table 3). Values for expected heterozygosity, observed heterozygosity, percentage of polymorphic SNPs and nucleotide diversity were all larger in the USDA than in the IPK accessions. Accessions showed a high inbreeding coefficient ($F > 0.5$), which did not differ between both genebanks.

Identification of highly differentiated outlier SNPs

The strong genetic clustering of accessions from the two genebanks despite a low overall F_{ST} value suggests that a small number of SNPs are responsible for the differentiation. To identify polymorphisms whose allele frequencies differ between the two genebank collections, we performed outlier tests using the GBS data without missing values (1,444 SNPs). LOSITAN identified 182 (12.6%) SNPs as outliers (Table G in S2 File) based on F_{st} values between the two populations (FDR < 0.1). Arlequin identified only 79 (5.5%) SNPs as outliers (Table H in S2 File). Since no *p*-values can be calculated for the X^TX statistic, we ordered SNPs by the rank order of X^TX values, as suggested by [43] to detect strongly differentiated SNPs. A total of 73 SNPs were among the top 5% (Table I in S2 File). The Venn diagram in Fig 5 shows little overlap of highly differentiated SNPs detected between pairs of methods, and only one SNP was identifed as an outlier by all three methods.

Phenotypic differentiation between genebanks

Since the genetic analysis indicated a strong differentiation of accessions from the two genebanks, we also investigated their phenotypic differentiation. A PCA of phenotypic traits indicated a weaker differentiation between USDA and IPK accessions (Fig 2B) than the genotyping data, although the centroids of each cluster are clearly differentiated. The IPK

Table 3. Measures of diversity within two genebanks based three different data sets. Sample size of USDA accessions, *n* = 93 **and of IPK accession,** *n* = 81.

Statistic	SNPs without missing values (1,444 SNPs)		SNPs with missing values (120,693 SNPs)		Imputed SNPs (120,693 SNPs)	
	USDA	IPK	USDA	IPK	USDA	IPK
H_e	0.157	0.088	0.214	0.149	0.201	0.107
H_o	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
π	0.158	0.089	0.157	0.095	0.202	0.110

H_e = Expected heterozygosity, H_o = Observed heterozygosity, F = Inbreeding coefficient, π = nucleotide diversity.

https://doi.org/10.1371/journal.pone.0192062.t003

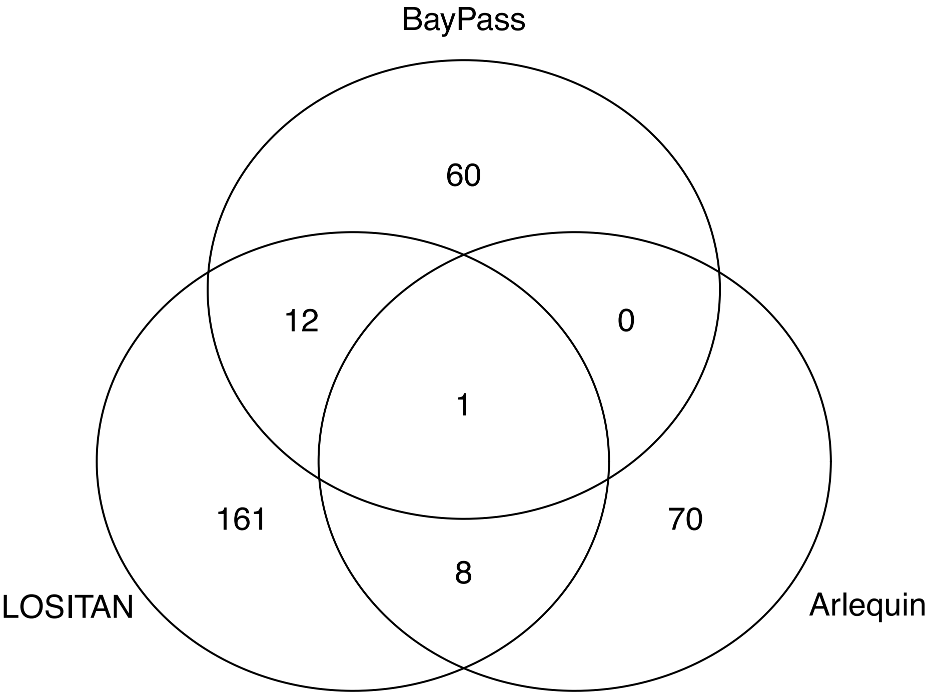

Fig 5. Overlap between outlier SNPs. Venn diagram of all significantly differentiated SNPs detected with Arlequin ($p < 0.05$), LOSITAN ($p < 0.05$) and BayPass (top 5% $X^T X$ values).

https://doi.org/10.1371/journal.pone.0192062.g005

accessions clustered more strongly than the USDA accessions which indicates a lower phenotypic variation and is also consistent with the genotyping data. The lower phenotypic than genetic differentiation between the genebanks reflects genotype x environment (GxE) interactions of phenotypic traits as inferred in a previous study [23]. The first principal component explained almost half of phenotypic variation in the total sample and the second about 20% (Fig 6A). The variable correlation plot indicates that the curd related traits are strongly related to the first principal component, whereas days to bolting (flowering time) shows a very high correlation with the second principal component (Fig 6B).

To test which traits are significantly differentiated between genebanks (Fig 7), we used MANOVA for comparison because some traits are strongly correlated (Fig F in S1 File). The strongest correlation is between length of nearest branch to apical meristem and length of apical meristen ($r = 0.798$, $p = 10^{-12}$). The phenotypic variance was different between accessions from both genebank, and the MANOVA with genebank as grouping factor strongly supported the phenotypic differentiation (Pillai test statistic: 0.3158, $p < 10^{-11}$). A post-hoc analysis of single-factor ANOVA with Bonferroni correction revealed that only the traits curd width ($F = 15.4$, $p = 0.0001254$) and cluster width ($F = 67.8$, $p < 10^{-14}$) differed between genebanks.

Finally, a Mantel test showed a positive correlation of phenotypic and genetic similiarities between pairs of accessions ($r = 0.291$, $p < 0.001$).

Discussion

Assessment of genetic diversity by GBS

Previously, the genetic diversity of cauliflower was assessed with different types of markers and based on small data sets. This limitation can be overcome by sequencing-based methods like GBS [15]. Although we generated a large number of raw reads, a substantial proportion

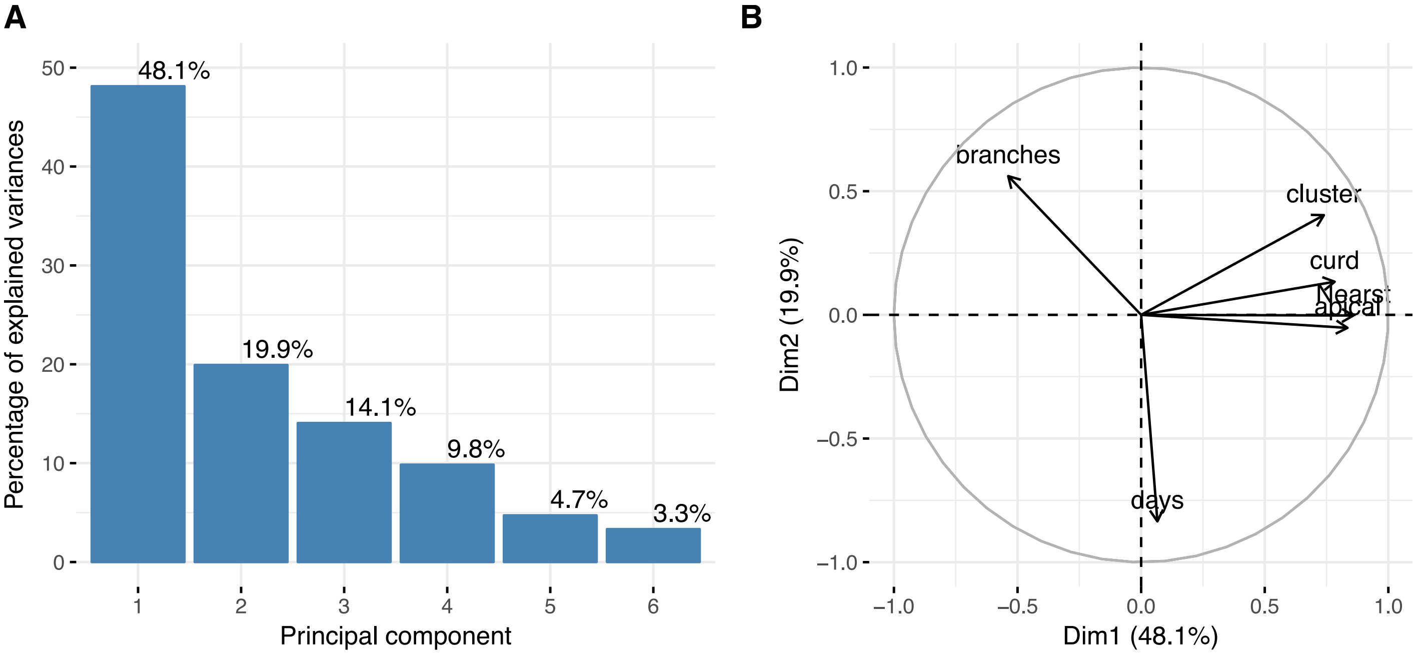

Fig 6. Parameters of PCA with phenotypic variation. (A) Percent variance explained by each of the six principal components. (B) Variable correlation plot on the first two principal components. The quality of traits on the PC map is expressed as distance between traits and the origin. Traits that are distant from the origin are well represented on the factor map.

https://doi.org/10.1371/journal.pone.0192062.g006

of reads (71% on average) did not align to the *B. oleracea* reference genome, similar to what has been observed in maize [18]. Such a high proportion of unmapped reads may result from the evolutionary divergence between the subspecies *B. oleracea* spp. *capitata* (cabbage) from which the reference genome was produced and cauliflower, *B. oleracea* var. *botrytis*. This interpretation is confirmed by the pan-genome analysis of *B. oleracea* cultivars [14, 44]. Another explanation is that the reference genome is still incomplete. The proportion of matching reads may be also influenced by a limited sensitivity of the Burrows-Wheeler Alignment (BWA) algorithm or a high proportion of presence/absence variation (PAV) [18].

GBS was used for a wide range of species and is an effective method for generating tens of thousands SNP markers [15, 17], but the high proportion of missing data, which in our study varied between 19% and 77% per accessions, is a major disadvantage [15]. Possible solutions to reduce the proportion of missing data include the sequencing to higher read depths [16] or the imputation of missing values, which we pursued in this study. A comparison of diversity estimates obtained with the three SNP sets indicates that estimates of observed heterozygosity, H_o, are influenced by missing data, whereas estimates of the inbreeding coefficient and nucleotide diversity, π are very similar (Table 3). Therefore, genome-wide parameters like diversity or genetic structure can be estimated with or without missing data. Due to the rapidly decreasing sequencing costs and protocols for low-cost preparation of whole genome sequencing libraries [45], low coverage genome sequencing is rapidly becoming the method of choice for characterizing the genomic diversity of species with moderate genome sizes such as *B. oleracea*. The correlation of phenotypic and genotypic distance as indicated by the Mantel test indicates that genotyping can be used to assemble phenotypically diverse core collections for further evaluation.

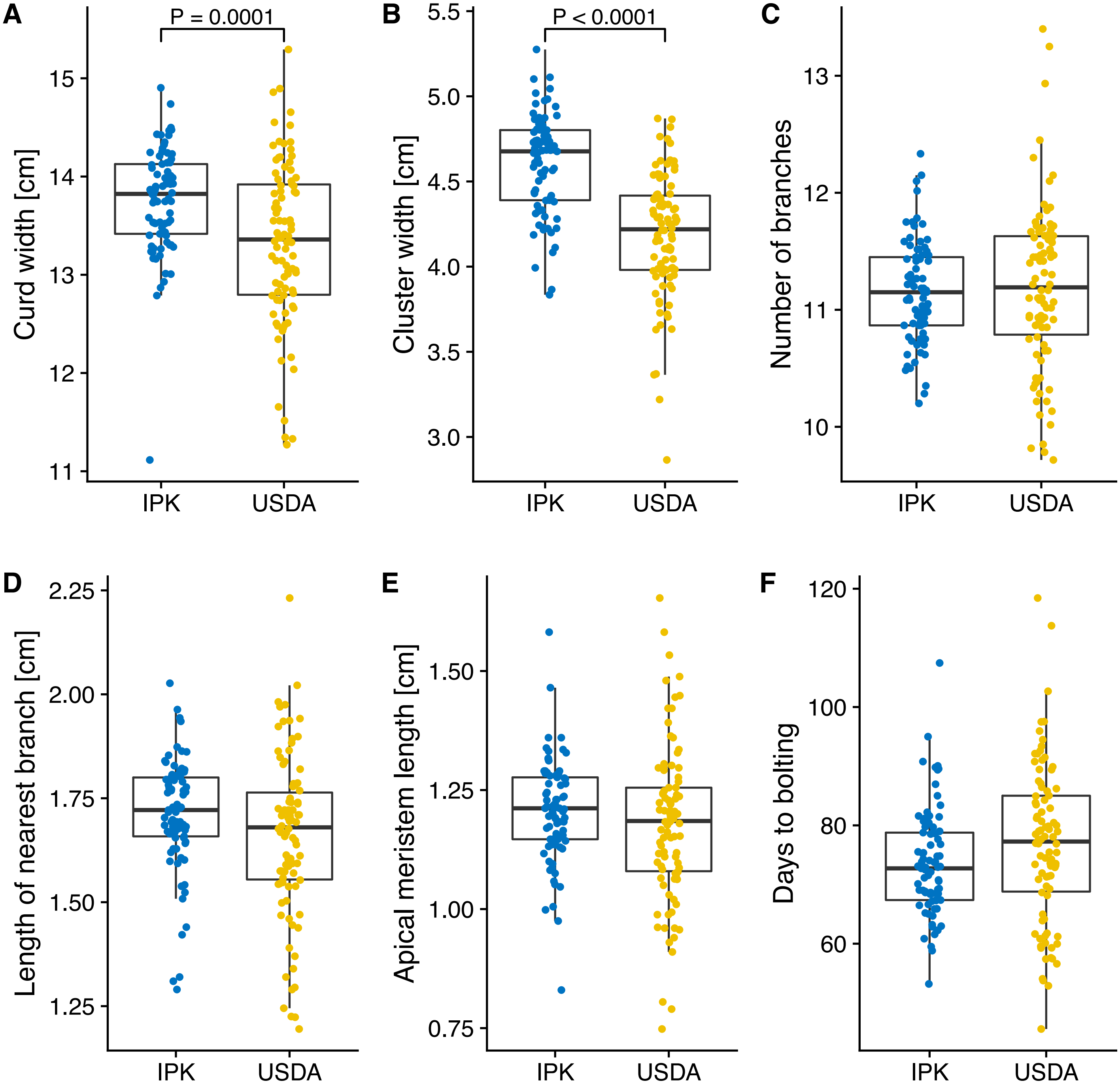

Fig 7. Box plots of six curd-related traits in accessions grouped by seed source (USDA and IPK genebanks). Significant differences between the two genebanks as observed from a single-factor ANOVA with Bonferroni correction of *p*-values. The *p*-values for the two traits with significant differences are shown. Individual phenotype values are averages of 6 environments (2 locations and 3 growing seasons each). The upper and lower hinges correspond to the first and third quartiles (the 25th and 75th percentiles), and the whiskers extend to the largest value no further than 1.5 * IQR (inter-quartile range), or distance between the first and third quartiles. Phenotypic data are from [23].

https://doi.org/10.1371/journal.pone.0192062.g007

Patterns and causes of genetic structure in cauliflower accessions

Despite the high proportion of missing data, GBS allowed to analyze genetic diversity and population structure in *B. oleracea* genebank accessions. Our sample of cauliflower accessions form groups that do not reflect their geographic origin but the seed source (i.e., *ex situ* genebank). A geographic population structure was found in previous survey of cauliflower cultivars [9] despite a smaller sample size than our study. Furthermore, patterns of genetic diversity of switchgrass, maize and sorghum genebank accessions obtained with GBS was consistent with the ancestral history, morphological types and geographic distribution of these crops [18–20]. The clustering and the different levels of genetic diversity between genebanks suggest that other factors than geographic origin determine the genetic relationship of accessions.

First, the accession types (landraces vs. cultivars) and collection dates may result from different collection strategies between genebanks that may be responsible for the observed clustering of accessions. For example, the USDA sample includes a much higher proportion of unverified material than the IPK sample which mostly consist of modern varieties. Passport data suggest that many accessions in the 'unverified' group of USDA and the 'cultivar' group of IPK reflect common cultivars (or possibly landraces) that were commercially available during the time of collection. 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. They consist of four landraces and one hybrid (Table A in S2 File), supporting the notion that the USDA genbank included more sources of germplasm that resulted in a larger collection of genetically diverse landraces (which were labeled as 'unverified') than the IPK genbank. The significant difference in average collection date between both genebanks may further contribute to their differentiation. According to the passport data, the USDA accessions are on average more than 15 years older than the IPK accessions. The USDA accessions were collected between 1948 to 1981 with an average of 1959, whereas the IPK accessions were collected between 1957 to 2002 with an average of 1974. For this reason, the observed population differentiation may also be caused by various process such as breeding progress, seed management strategies, inbreeding or genetic drift of *B. oleracea* germplasm. Unfortunately, limited passport information did not allow us to account for the breeding history and relationships among varieties as co-variate in the population structure analysis and to disentangle the effects collection date and geographic origin on diversity estimates. In addition, the absence of a strong geographic structure may result from a combination of low genetic diversity in cauliflower [11, 12], an exchange of seeds over large distance in historical time, and a high level of gene flow due to outcrossing with other varieties [46]. On the other hand, the lower genetic diversity of the IPK relative to the USDA accessions may be caused by a higher proportion of modern cultivars, which have a narrower genetic basis resulting from modern breeding methods. These different processes are difficult if not impossible to reconstruct and confound the analysis of genetic diversity.

A second explanation for differentiation between genebanks are seed regeneration procedures which may affect genetic diversity [47–50]. In several species, *ex situ* conserved genetic resources had a lower diversity than *in situ* conserved populations [51–53] or historical material [46]. 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 [6, 52, 54, 55]. A high overall inbreeding coefficient of >0.5 for all accessions estimated from the GBS data suggest an impact of small population size on genetic diversity. The inbreeding coefficient depends on the breeding history of the material before inclusion into the genbank and the seed propagation protocols. The seed regeneration procedures at the USDA and IPK genebanks likely did not contribute much to the observed differentiation. Seed regeneration at the USDA genebank is carried out in 12 × 24 ft cages (corresponding to 26.8 m^2) with mesh covers to prevent cross-pollination by insects and a population size of at least 100 plants per accession. At IPK, cauliflower accessions are cultivated in small glass houses containing other species as well. The total area is about 6 m^2 and population size is 20-25 plants. At IPK, seeds are regenerated after 20 years and at USDA after 15 years, which on average corresponds to 1.5 and 4 regeneration cycles for the material included in this study. Using the formula for calculating the expected decay in heterozygosity, $H_t = H_0 \times (1 - 1/N_e)^t$ [56], the expected relative decay in heterozygosity of USDA genebank accessions is $\Delta H = (1 - 1/100)^4 = 0.96$ under the assumption that on average each accession has undergone 4 regeneration cycles, and of IPK accessions $\Delta H = (1 - 1/20)^{1.5} = 0.93$. Although a more rapid decay in heterozygosity is expected in the IPK collection, the difference is rather small.

Natural and/or artificial selection during seed propagation of genebank accessions may also impact genetic diversity and differentiation in *ex situ* genebanks. The cauliflower accessions were collected in different regions of the world, and selection of genes controlling photoperiod sensitivity, flowering time, and other traits caused by local adaptation to the propagation environment may have occured. The effect of selection on fitness and genetic diversity can be quite strong. A significant correlation between genetic and phenotypic distance shown by the Mantel test and a high heritability of traits like flowering time [23] indicate a strong phenotype-genotype relationship that facilitates a rapid evolutionary response to selection. During seed regeneration, pollination is managed with commercial pollinators like bumblebees, but the reproductive success is not closely monitored and the effect of selection on the relationship structure is unknown. To test the potential impact of selection, we used three outlier tests to identify highly differentiated SNPs. Out of 1,444 tested SNPs without missing data, 12.5% (LOSITAN), 5.5% (Arlequin) and 5% (BayPass) 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.* [57] 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 [58]. Hence, in a comparison of outliers identified in our data (Fig 5), we identified only 0.8% (12 out of 1,444 SNPs) of SNPs as outliers by LOSITAN and BayPass, 0.6% (0) by LOSITAN and Arlequin, 0% (0) of SNPs by BayPass and Arlequin, and only a single SNP by all three methods. Overall, this is a small proportion (<1%) of the total number of SNPs tested. In conclusion, if natural or unintentional artificial selection during seed regeneration contribute to genetic differentiation, it may either weak selection, affect only few genomic regions or occur in regions that were not tagged by the SNPs of this study. The identification of strongly differentiated SNPs rests on the assumption that both collections derive from the same ancestral population, which likely is not true for our sample. The comparison of allele frequencies of accessions over seed regeneration cycles with higher marker densities is a more powerful approach to detect the effect of selection on *ex situ* genebank material, and facilitates the close monitoring of allele frequency changes and a better management of *ex situ* germplasm collections.

Characterizing genebank accessions with GBS

A major advantage of GBS and related methods is their applicability to any species. These methods do not entail setup costs like SNP arrays and do not cost much per individual genotype, but provide sufficient power for genome-wide analyses of population structure and genetic relationships. 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 [59, 60]. In our study, however, a comparison between three GBS-derived data sets consisting of SNPs without missing values, SNPs with missing values, and imputed SNPs revealed only a minor effect of missing data and data imputation on the ability to infer the population structure, although diversity estimates differed significantly between imputed and non-imputed data (Table 2). This result confirms a previous study [61] in which the estimation of heterozygosity and inbreeding coefficients was less accurate with a high proportion of missing data and estimation biases were much smaller for data sets with missing values than for imputed data sets. Furthermore, the density of GBS-derived markers are frequently too low to detect footprints of selection [62] caused by the different history of genebank collections or ongoing selection during seed regeneration. The correlation of genetic and phenotypic differentiation of collections (Fig 7) indicates that GBS is a highly suitable approach for defining core collections and sets of genetically differentiated genebank accessions that are further used for whole genome sequencing, phenotypic characterization or the establishment of (pre-)breeding populations.

Conclusions

Our study outlined the usefulness of GBS to characterize the genetic diversity of genebank accessions of a minor crop like cauliflower. A key result was the strong differentiation of genetic diversity between the two genebanks which most likely reflects the different collection histories of the two genebanks. Due to a lack of detail in the passport information, factors influencing genetic diversity like sampling strategy, regeneration procedures and selection during regeneration could not be well reconstructed, although the type of accessions included (landraces vs. cultivars) likely has a strong influence. The low cost of GBS and low-coverage genome sequencing suggest that a lack of passport information can be substituted by high-resolution genotyping and suitable analysis methods to characterize the diversity of germplasm from different sources. This facilitates the exchange of material between genebanks and the construction of core collections that harbor a high proportion of species-wide genetic and phenotypic diversity for a more efficient utilization of plant genetic resources [63]. GBS-derived polymorphisms may facilitate an exchange of germplasm between genebanks, but this requires an infrastructure for genomic data management similar to the information system already in place for passport data. Our work also demonstrated monitoring of genetic diversity during seed regeneration allows to manage diversity within accessions to mitigate some disadvantages of small population sizes of *ex situ* conserved plant genetic resources, in particular for outbreeding crops such as cauliflower.

Supporting information

S1 File. Supporting file with figures and accompanying text.
(PDF)

S2 File. Supporting spreadsheet file with tables.
(XLSX)

Acknowledgments

Keygene N. V. owns patents and patent applications protecting its Sequence Based Genotyping Technologies. We express our thanks to Fabian Freund, Christian Lampei, Dounia Saleh, Torsten Günther, Patrick Thorwarth and Linda Homann for discussions on data analysis methods. Also, we thank the USDA and IPK genebanks for prividing seed stocks. We thank Marcus Koch, Heidelberg Botanic Garden and Herbarium HEID, for seeds of the wild ancestor of *Brassica oleraceae*. This work was supported by a GERLS Fellowship of the DAAD to E. Y. and the F. W. Schnell Endowed Professorship of the Stifterverband für Deutsche Wissenschaft to K. J. S.

Author Contributions

Conceptualization: Eltohamy A. A. Yousef, Andreas Börner, Karl J. Schmid.

Data curation: Eltohamy A. A. Yousef, Thomas Müller, Andreas Börner.

Formal analysis: Eltohamy A. A. Yousef, Thomas Müller.

Investigation: Thomas Müller.

Methodology: Eltohamy A. A. Yousef, Thomas Müller, Andreas Börner.

Project administration: Karl J. Schmid.

Software: Thomas Müller.

Supervision: Karl J. Schmid.

Writing – original draft: Eltohamy A. A. Yousef, Thomas Müller, Karl J. Schmid.

Writing – review & editing: Eltohamy A. A. Yousef, Karl J. Schmid.

References

1. Lu X, Liu L, Gong Y, Zhao L, Song X, Zhu X. Cultivar identification and genetic diversity analysis of broccoli and its related species with RAPD and ISSR markers. Sci Hort. 2009; 122(4):645–648. https://doi.org/10.1016/j.scienta.2009.06.017
2. Rao VR, Hodgkin T. Genetic diversity and conservation and utilization of plant genetic resources. Plant Cell Tissue Organ Cult. 2002; 68(1):1–19. https://doi.org/10.1023/A:1013359015812
3. de Jesus ON, e Silva SdO, Amorim EP, Ferreira CF, de Campos JMS, de Gaspari Silva G, et al. Genetic diversity and population structure of Musa accessions in *ex situ* conservation. BMC Plant Biol. 2013; 13(1):41. https://doi.org/10.1186/1471-2229-13-41 PMID: 23497122
4. Kjaer ED, Graudal L, Nathan I. Ex-situ Conservation of Commercial Tropical Trees: Strategies, Options and Constraints. Danida Forest Seed Centre; 2001.
5. Lauterbach D, Burkart M, Gemeinholzer B. Rapid genetic differentiation between *ex situ* and their *in situ* source populations: an example of the endangered *Silene otites* (Caryophyllaceae). Bot J Linn Soc. 2012; 168(1):64–75. https://doi.org/10.1111/j.1095-8339.2011.01185.x
6. Lee SA, Fowke JH, Lu W, Ye C, Zheng Y, Cai Q, et al. Cruciferous vegetables, the GSTP1 Ile105Val genetic polymorphism, and breast cancer risk. Am J Clin Nutr. 2008; 87(3):753–760. PMID: 18326615
7. Tang L, Zirpoli GR, Guru K, Moysich KB, Zhang Y, Ambrosone CB, et al. Consumption of raw cruciferous vegetables is inversely associated with bladder cancer risk. Cancer Epidemiol Biomarkers Prevention. 2008; 17(4):938–944. https://doi.org/10.1158/1055-9965.EPI-07-2502
8. Food and Agriculture Organization of the United Nations. FAOSTAT Database; 2017. Available from: http://faostat.fao.org.
9. Astarini IA, Plummer JA, Lancaster RA, Yan G. Genetic diversity of Indonesian cauliflower cultivars and their relationships with hybrid cultivars grown in Australia. Scientia Hort. 2006; 108(2):143–150. https://doi.org/10.1016/j.scienta.2006.01.033
10. Izzah NK, Lee J, Perumal S, Park JY, Ahn K, Fu D, et al. Microsatellite-based analysis of genetic diversity in 91 commercial *Brassica oleracea* L. cultivars belonging to six varietal groups. Genetic Res Crop Evol. 2013; 60(7):1967–1986. https://doi.org/10.1007/s10722-013-9966-3
11. Zhao Z, Gu H, Sheng X, Yu H, Wang J, Zhao J, et al. Genetic diversity and relationships among loose-curd cauliflower and related varieties as revealed by microsatellite markers. Scientia Horticulturae. 2014; 166:105–110. https://doi.org/10.1016/j.scienta.2013.12.024
12. Tonguç M, Griffiths PD. Genetic relationships of Brassica vegetables determined using database derived simple sequence repeats. Euphytica. 2004; 137(2):193–201. https://doi.org/10.1023/B:EUPH.0000041577.84388.43

13. Louarn S, Torp AM, Holme I, Andersen SB, Jensen BD. Database derived microsatellite markers (SSRs) for cultivar differentiation in *Brassica oleracea*. Gen Res Crop Evol. 2007; 54(8):1717–1725. https://doi.org/10.1007/s10722-006-9181-6

14. Golicz AA, Bayer PE, Barker GC, Edger PP, Kim H, Martinez PA, et al. The pangenome of an agronomically important crop plant *Brassica oleracea*. Nat Commun. 2016; 7:13390. https://doi.org/10.1038/ncomms13390 PMID: 27834372

15. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLOS ONE. 2011; 6(5):e19379. https://doi.org/10.1371/journal.pone.0019379 PMID: 21573248

16. Poland JA, Rife TW. Genotyping-by-sequencing for plant breeding and genetics. Plant Genome. 2012; 5(3):92–102.

17. Poland JA, Brown PJ, Sorrells ME, Jannink JL. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLOS One. 2012; 7(2): e32253. https://doi.org/10.1371/journal.pone.0032253 PMID: 22389690

18. Romay MC, Millard MJ, Glaubitz JC, Peiffer JA, Swarts KL, Casstevens TM, et al. Comprehensive genotyping of the USA national maize inbred seed bank. Genome Biol. 2013; 14(6):R55. https://doi.org/10.1186/gb-2013-14-6-r55 PMID: 23759205

19. Morris GP, Ramu P, Deshpande SP, Hash CT, Shah T, Upadhyaya HD, et al. Population genomic and genome-wide association studies of agroclimatic traits in sorghum. Proc Natl Acad Sci USA. 2013; 110 (2):453–458. https://doi.org/10.1073/pnas.1215985110 PMID: 23267105

20. Lu F, Lipka AE, Glaubitz J, Elshire R, Cherney JH, Casler MD, et al. Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. PLOS Genet. 2013; 9(1):e1003215. https://doi.org/10.1371/journal.pgen.1003215 PMID: 23349638

21. Fu YB, Cheng B, Peterson GW. Genetic diversity analysis of yellow mustard (*Sinapis alba* L.) germplasm based on genotyping by sequencing. Genetic Res Crop Evol. 2014; 61(3):579–594. https://doi.org/10.1007/s10722-013-0058-1

22. Watts LE. Investigations into the breeding system of cauliflower *Brassica oleracea* var. *botrytis* (L.). Euphytica. 1963; 12(3):323–340.

23. Yousef EA, Lampei C, Schmid KJ. Evaluation of cauliflower genebank accessions under organic and conventional cultivation in Southern Germany. Euphytica. 2015; 201(3):389–400. https://doi.org/10.1007/s10681-014-1225-y

24. Lan TH, Paterson AH. Comparative mapping of quantitative trait loci sculpting the curd of *Brassica oleracea*. Genetics. 2000; 155(4):1927–1954. PMID: 10924486

25. Saghai-Maroof MA, Soliman KM, Jorgensen RA, Allard R. Ribosomal DNA spacer-length polymorphisms in barley: Mendelian inheritance, chromosomal location, and population dynamics. Proc Natl Acad Sci USA. 1984; 81(24):8014–8018. PMID: 6096873

26. Li H, Durbin R. Fast and accurate short read alignment with Burrows—Wheeler transform. Bioinformatics. 2009; 25(14):1754–1760. https://doi.org/10.1093/bioinformatics/btp324 PMID: 19451168

27. Andrews S. FastQC: a quality control tool for high throughput sequence data. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

28. Liu S, Liu Y, Yang X, Tong C, Edwards D, Parkin IA, et al. The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes. Nat Commun. 2014; 5.

29. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009; 25(16):2078–2079. https://doi.org/10.1093/bioinformatics/btp352 PMID: 19505943

30. Jombart T, Ahmed I. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. Bioinformatics. 2011; 27(21):3070–3071. https://doi.org/10.1093/bioinformatics/btr521 PMID: 21926124

31. Team RC. R: A language and environment for statistical computing. Vienna, Austria. URL http://www.R-project.org. 2015.

32. Mantel N. The detection of disease clustering and a generalized regression approach. Cancer Res. 1967; 27(2 Part 1):209–220. PMID: 6018555

33. Dray S, Dufour AB, et al. The ade4 package: implementing the duality diagram for ecologists. J Stat Software. 2007; 22(4):1–20. https://doi.org/10.18637/jss.v022.i04

34. Paradis E, Claude J, Strimmer K. APE: analyses of phylogenetics and evolution in R language. Bioinformatics. 2004; 20(2):289–290. https://doi.org/10.1093/bioinformatics/btg412 PMID: 14734327

35. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009; 19(9):1655–1664. https://doi.org/10.1101/gr.094052.109 PMID: 19648217

36. Alexander DH, Lange K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC Bioinf. 2011; 12(1):246. https://doi.org/10.1186/1471-2105-12-246

37. Excoffier L, Lischer HE. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour. 2010; 10(3):564–567. https://doi.org/10.1111/j.1755-0998.2010.02847.x PMID: 21565059

38. Paradis E. pegas: an R package for population genetics with an integrated—modular approach. Bioinformatics. 2010; 26(3):419–420. https://doi.org/10.1093/bioinformatics/btp696 PMID: 20080509

39. Scheet P, Stephens M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet. 2006; 78(4):629–644. https://doi.org/10.1086/502802 PMID: 16532393

40. Beaumont MA, Nichols RA. Evaluating loci for use in the genetic analysis of population structure. Proc Roy Soc London B: Biol Sci. 1996; 263(1377):1619–1626. https://doi.org/10.1098/rspb.1996.0237

41. Antao T, Lopes A, Lopes RJ, Beja-Pereira A, Luikart G. LOSITAN: a workbench to detect molecular adaptation based on a Fst-outlier method. BMC Bioinf. 2008; 9(1):323. https://doi.org/10.1186/1471-2105-9-323

42. Gautier M. Genome-wide scan for adaptive divergence and association with population-specific covariates. Genetics. 2015; 201(4):1555–1579. https://doi.org/10.1534/genetics.115.181453 PMID: 26482796

43. Günther T, Coop G. Robust identification of local adaptation from allele frequencies. Genetics. 2013; 195(1):205–220. https://doi.org/10.1534/genetics.113.152462 PMID: 23821598

44. Cheng F, Sun R, Hou X, Zheng H, Zhang F, Zhang Y, et al. Subgenome parallel selection is associated with morphotype diversification and convergent crop domestication in *Brassica rapa* and *Brassica oleracea*. Nat Genet. 2016; 48(10):1218–1224. https://doi.org/10.1038/ng.3634 PMID: 27526322

45. Therkildsen NO, Palumbi SR. Practical low-coverage genomewide sequencing of hundreds of individually barcoded samples for population and evolutionary genomics in nonmodel species. Mol Ecol Resour. 2016; 12(2):194–208.

46. Hagenblad J, Zie J, Leino MW. Exploring the population genetics of genebank and historical landrace varieties. Genet Res Crop Evol. 2012; 59(6):1185–1199. https://doi.org/10.1007/s10722-011-9754-x

47. Dulloo ME, Hunter D, Borelli T. *Ex situ* and *in situ* conservation of agricultural biodiversity: major advances and research needs. Notulae Botanicae Horti Agrobotanici Cluj-Napoca. 2010; 38(2):123–135.

48. Börner A, Chebotar S, Korzun V. Molecular characterization of the genetic integrity of wheat (*Triticum aestivum* L.) germplasm after long-term maintenance. Theor Appl Genet. 2000; 100(3-4):494–497. https://doi.org/10.1007/s001220050064

49. Chebotar S, Röder M, Korzun V, Saal B, Weber W, Börner A. Molecular studies on genetic integrity of open-pollinating species rye (*Secale cereale* L.) after long-term genebank maintenance. Theor Appl Genet. 2003; 107(8):1469–1476. https://doi.org/10.1007/s00122-003-1366-1 PMID: 12898026

50. van Hintum TJ, van De Wiel C, Visser D, Van Treuren R, Vosman B. The distribution of genetic diversity in a *Brassica oleracea* gene bank collection related to the effects on diversity of regeneration, as measured with AFLPs. Theor Appl Genet. 2007; 114(5):777–786. https://doi.org/10.1007/s00122-006-0456-2 PMID: 17273846

51. Gómez OJ, Blair MW, Frankow-Lindberg B, Gullberg U. Comparative Study of Common Bean (*Phaseolus vulgaris* L.) Landraces Conserved *ex situ* in Genebanks and *in situ* by Farmers. Genetic Res Crop Evol. 2005; 52(4):371–380. https://doi.org/10.1007/s10722-005-2249-x

52. Rucińska A, Puchalski J. 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. 2011; 20(2):401–413. https://doi.org/10.1007/s10531-010-9965-z

53. Brütting C, Hensen I, Wesche K. *Ex situ* cultivation affects genetic structure and diversity in arable plants. Plant Biol. 2013; 15(3):505–513. https://doi.org/10.1111/j.1438-8677.2012.00655.x PMID: 22882447

54. Crossa J. Sample size and effective population size in seed regeneration of monoecious species. In: Regeneration of Seed Crops and Their Wild Relatives: Proceedings of a Consultation Meeting, 4-7 December 1995, ICRISAT, Hyderabad, India. Bioversity International; 1998. p. 140.

55. Frankham R, Briscoe DA, Ballou JD. Introduction to conservation genetics. Cambridge University Press; 2002.

56. Hartl DL, Clark AG. Principles of Population Genetics. 4th ed. Sinauer Associates; 2007.

57. Shimada Y, Shikano T, Merilä J. A high incidence of selection on physiologically important genes in the three-spined stickleback, *Gasterosteus aculeatus*. Mol Biol Evol. 2011; 28(1):181–193. https://doi.org/10.1093/molbev/msq181 PMID: 20660084

58. Lotterhos KE, Whitlock MC. Evaluation of demographic history and neutral parameterization on the performance of FST outlier tests. Mol Ecol. 2014; 23(9):2178–2192. https://doi.org/10.1111/mec.12725 PMID: 24655127

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

60. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat Rev Genet. 2010; 11(7):499–511. https://doi.org/10.1038/nrg2796 PMID: 20517342

61. Fu YB. Genetic diversity analysis of highly incomplete SNP genotype data with imputations: An empirical assessment. Genes, Genomes, Genetics. 2014; 4(5):891–900.

62. Lowry DB, Hoban S, Kelley JL, Lotterhos KE, Reed LK, Antolin MF, et al. Breaking RAD: An evaluation of the utility of restriction site associated DNA sequencing for genome scans of adaptation. Molecular Ecology Resources. 2016; https://doi.org/10.1111/1755-0998.12635

63. Wang C, Hu S, Gardner C, Lüberstedt T. Emerging Avenues for Utilization of Exotic Germplasm. Trends in Plant Science. 2017; 22(7):624–637. https://doi.org/10.1016/j.tplants.2017.04.002 PMID: 28476651