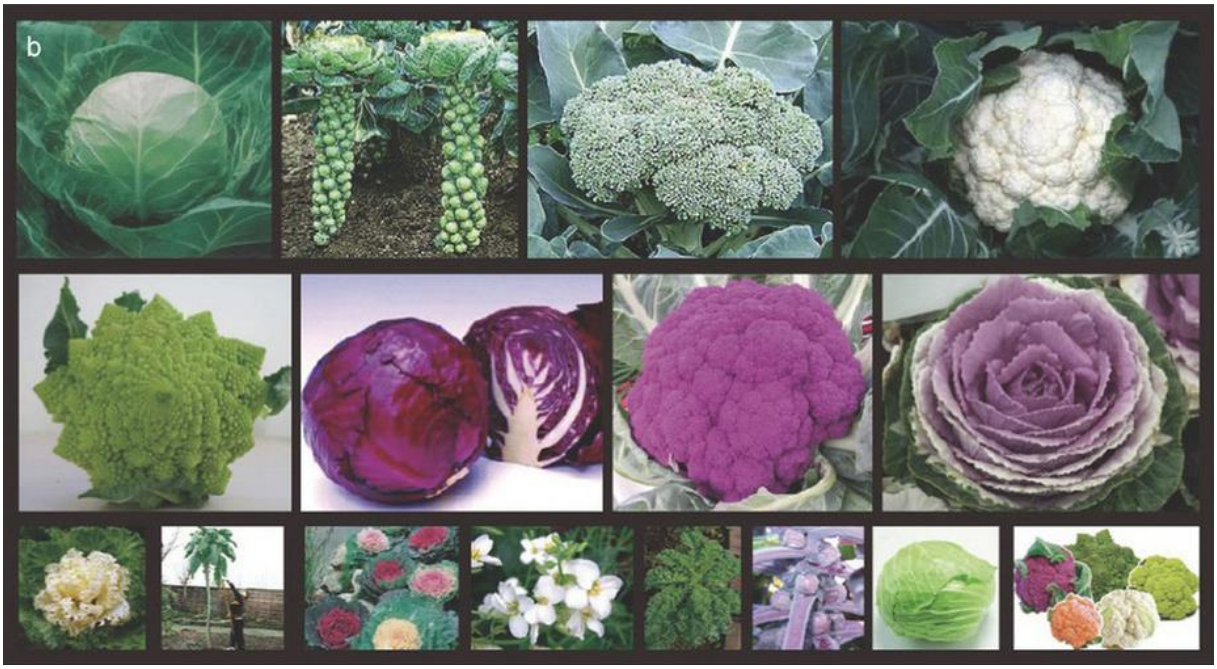
Uncover selective sweeps for variation in diverse morphotypes of *Brassica oleracea*



Cheng *et al.*, 2014

Yonina Hendrikse
910827325070
Guusje Bonnema,
Theo Borm
& Christian Bachem
Wageningen university and research centre
Plant Breeding, growth and development
13-5-2015

Abstract

123 accessions from different morphotypes of *B. oleracea* (cabbage, cauliflower, broccoli, kohlrabi and kale) were collected from multiple companies and the gene bank in Wageningen, CGN. The genomes have been sequenced, mapped against the reference genome published by Liu *et al.* (2014) and variants were called using GATK. The obtained SNPs were used for calculating SweeD (Pavlidis *et al.*, 2013), Linkage Disequilibrium (LD), Fixation index (Fst), Tajima's D and variable sites (π). When several of the parameters showed the same results, the region was marked as putative selective sweep. LD was not taken into account as some regions showed unexpected high LD-decay values, possibly because of incorrect mapping to the reference genome due to the triplicated genome. Genes were found according to their synteny with *A. thaliana* and the associated GO-terms. GO-enrichment was calculated with the fisher exact test and 49 GO-terms showed significance. According to number of occurrence and groups of enriched GO-terms, some were further analyzed. For the traits involved in enlarged inflorescence and inward curling of the leaf (heading trait) several orthologous genes of *A. thaliana* were found. The results also showed more background noise for *B. oleracea* compared to *B. rapa* data from Cheng *et al.* (pers. comm., 2014), suggesting *B. oleracea* fixation occurred later than *B. rapa*.

Table of Contents

1. Introduction

1.1 Domestication

Twelve to ten thousand years ago humans started a transition from hunting-gathering of food to agriculture, where humans started to intentionally propagate specific animal and plant species. This took place on several areas around the world, giving rise to various crops according to human preferences and needs (Gepts, 2014; Lenser and Theissen, 2013). These domesticated plants gained a variety of morphological and physiological changes through a process of plant breeding (selecting plants that best met desired phenotypes) (Gepts, 2014; Lenser and Theissen, 2013). The morphological and physiological changes are based on processes and genes connected to reproduction, development and adaptation related traits (Gepts, 2014; Ross-Ibarra *et al.*, 2007). These genes are important sources of information to understand the often extreme phenotypes of crops; it would therefore be a great advantage for breeders to gain knowledge on these genes and identify them (Gepts, 2014).

1.2 Selective sweep

A consequence of domestication is when a beneficial trait increases in frequency within a population (Schaffner & Sabeti, 2008). Under natural selection, the trait must increase the organisms probability of survival, reproduction and it must be heritable to the offspring (Schaffner & Sabeti, 2008), while in domestication the trait increases due to selection by the farmer/breeder. Nearby located alleles are unintentionally also selected, due to linkage, and "hitchhike" together with the selected allele. Through this process the allele and surrounding alleles in the following generations become very similar to each other and are swept to fixation (McVean, 2006), the particular region shows less variation and is therefore called a "selective sweep" (Nimmakayala, 2014; Qi *et al.*, 2013; Nielsen *et al.*, 2005) (fig 1).

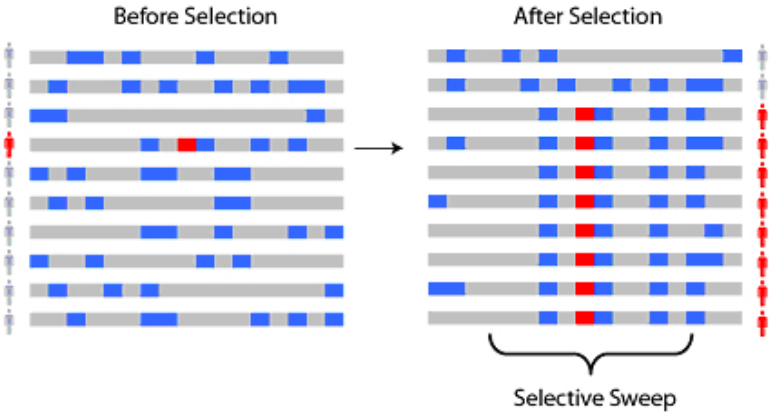


Figure 1. Selective sweep, positive selection of genes and the "hitchhiking" genes results in less variation on the locus (Schaffner & Sabeti, 2008).

1.3 Genetic background

The family *Brassicaceae* (*Cruciferae*) is widely distributed and includes the for ages worldwide cultivated genus *Brassica* as well as the extensively studied model plant *Arabidopsis thaliana* (Lagercrantz, 1998; de Jong *et al.*, 2007). The genus *Brassica* contains a great diversity of economically important crops, used as vegetable, oil, fodder and condiments (de Jong *et al.*, 2007; Cheng *et al.* pers. comm., 2014). From six species the basic genome structure and their interrelationships are known, three are diploid (*B. oleracea*, n=9; *B. nigra*, n=8; *B. rapa,* n=7) and the other three allotetraploid (*B. napus,* n=19; *B. juncea*, n=18; B. *carinata, n=17*) (Cheng *et al.* pers. comm., 2014). The diploid species are considered the basic genomes A, B and C, while the allotetraploid species contain two of the basic genomes, resulting in the U's triangle model (Nagahara, 1935; Cheng *et al.,* pers. comm., 2014) (fig. 2). Research shows that a whole genome

triplication (WGT) must have occurred approximately 9-15 million years ago, as diploid *Brassica* genome encompass three rearranged variants of the ancestral genome and seem to descent from a hexaploid ancestor (Lysak *et al.*, 2005; Cheng *et al.* pers. comm., 2014). *A. thaliana*, descended from the same ancestor, containing only one copy of the genome (Lagercrantz, 1998).

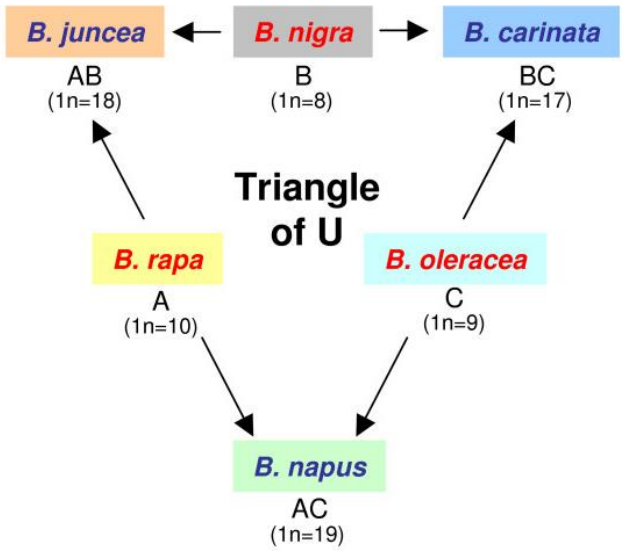


Figure 2. U-triangle of six *Brassica* species, which shows that three species consist of a pair of the basic genomes A, B and C. (Ostengaard and King, 2008)

The WGT of *Brassica* occurred according to a two step theory proposed by Cheng *et al.* (2014) (fig. 3). The theory suggests first two of the ancestors, translocated Proto-Calepine karyotype (tPCK, n=7), More Fractionated (MF)1 and MF2 merged together. Reshuffling and fractionation of the genes took place, resulting in a new diploid. Second, the third tPCK Least Fractionated (LF) genome merged with the MF1/MF2, initiating a second round of reshuffling and fractionation of genes. The biased genome fractionation cause loss of dominant gene expression, so genes in LF show a higher expression level compared to their paralogues in the MFs (Cheng *et al.*, 2014). The functional mutation level on LF is also lower, illustrated by the fact that LF regions display less reshuffling and fractionation of genes (Cheng *et al.*, 2014). Connected to this research a hypothesis rises; the genomic rearrangement, gene losses and gene evolution initiated by WGT created a situation in which new types of *Brassica* plants could be selected (Cheng *et al.*, 2014; Liu *et al.*, 2014). Because the three genome copies have many genes still present in triplicate or duplicate, it's possible for only one copy to change to another function, putatively initiating a different morphotype through a gain of function. But a loss of function is also a possibility, as natural selection occurs since 9-15 million years ago, while human interference started 12-10 thousand years ago (Gepts, 2014). Triplicated or duplicated genes could have been lost by natural selection before human interference, therefore only one gene remains. When a mutation occurs in such a gene, a loss of function could occur and also cause certain morphotypes. As the MF genomic regions are more fractionated, those genes could have been lost by natural selection, leaving only one copy in the LF region. Also for a gain of function mutation, statistically it could have occurred in MF, as these region are more fractionated.

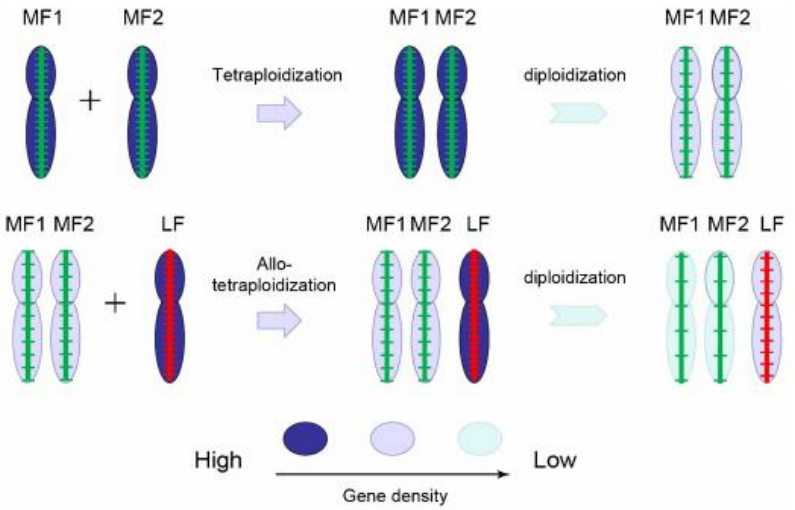


Figure 3. Polyploidization takes place in two steps during the WGT of *Brassica.* (Cheng *et al.*, 2014)

1.4 Morphological variation

B. oleracea comprises some distinct morphotypes, including leafy heads (cabbage), enlarged stems (kohlrabi, kales and marrow kales), enlarged inflorescences (Cauliflower and broccoli) and enlarged axillary buds (Brussels sprouts) (Bonnema *et al.,* 2011). Also, each *Brassica* species has evolved into different morphotypes, analogues which can be found in multiple of these species (Bonnema *et al.*, 2011). Leafy head is observed in *B. oleracea*, but this is also observed in *B. rapa* and *B. juncea*. Similarly, enlarged tubers are found in turnips (*B. rapa*), kohlrabi (*B. oleracea*) and rutabaga (*B. napus*). This raises the question whether the genomic changes underlying this morphotype as found in *B. oleracea* can also be present in other *Brassica* species. It is also observed by Zhao (2005) that different morphotypes in *B. rapa* from the same geographical region are often genetically more similar than similar morphotypes from other regions, like the turnips. This suggests an independent origin in both regions or a long and separate domestication and breeding history in regions (Zhao, 2005; Cheng *et al.,* pers. comm., 2014).

1.5 Identify selective sweeps

1.5.1 Five parameters

To identify the genes involved with domestication and therefore the selective sweep regions, different parameters will be measured and comparisons will be made. These comparisons are between different morphotypes and between groups of morphotypes with similar traits and the remaining group. According to the paper by Nielsen (2005), the effect of selective sweep on genetic variation is based on three parameters (fig. 4); variable sites (pi/π), Linkage Disequilibrium (LD) and the frequency of the spectrum, measured by Tajima's D. In the region of the selective sweep the variable sites are reduced, as the hitchhiking genes and fixation of the region, results in less variation (fig. 1). If a haplotype is overrepresented in the population, due to selection and hitchhiking of genes less recombination occurred, and therefore LD is increased (Bamshad and Wooding, 2003). Tajima's D is a neutrality test, it measures the rare allele excess compared to the neutral model (Jensen *et al.*, 2005). The model is rejected when the excess is bigger than the neutral model, for a selective sweep this will give a negative value (Nielsen, 2005). There are also some papers using Fixation index (Fst) between groups to determine selective sweeps (Hufford *et al.*, 2012; Nimmakayala *et al.*, 2014). Fst, measured by population differentiation, is the proportion of genetic variation in a subpopulation compared to the total genetic variation. Due to directional selection of certain genes in selective sweeps and therefore isolation of alleles from one morphotype compared to the other, differentiation between subpopulations is high. And yet another method for detecting selective sweeps is based on allelic frequencies and the site frequency spectrum (SFS), implemented in the

tool SweeD (Pavlidis *et al.*, 2013). SFS calculates the distribution of allele frequency at segregating sites, which is shifted from its neutral expectation to a high frequency in a selective sweep (Pavlidis *et al.*, 2010). SweeD computes the composite likelihood ratio (CLR) for the region being a selective sweep over the neutral model (Pavlidis *et al.*, 2013).

In this thesis report, I report the selection of selective sweeps in *B. oleracea* by previous described five parameters. To get an impression of the genes in the selective sweep, GO-enrichment was performed. The selected genes are further analyzed on function within *B. oleracea* based on the synteny with *A. thaliana*. The function of the genes should give more insight in the extreme phenotypes of the crops and shed some light on crop domestication. This will be helpful for further crop breeding and varietal breeding and it increases fundamental insight in growth and development.

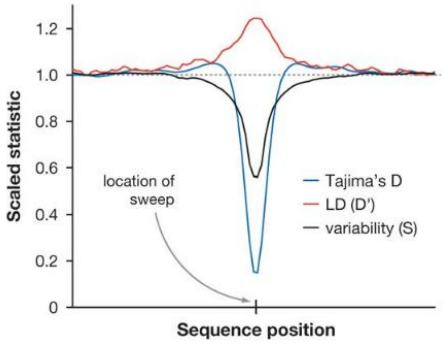


Figure 4. The effect of selective sweep on Tajima's D, LD and pi. It's based on averaging 100 simulations of strong selective sweep. Once the advantageous allele has reached frequency 1 in the population, the statistics are calculated according to a sliding window along the sequence and they are scaled so the expected value under neutrality equals one. (Nielsen, 2005)

2. Material and methods

To find the potential selective sweeps in the genomes of selected *B. oleracea* morphotypes, various bioinformatics tools were used and combined with annotation data to eventually find the genes putatively underlying the traits of interest. This is done by writing a pipeline performing all the necessary tasks. The first tasks in the pipeline have been performed by Theo Borm (pers. comm., 2014), which are quality control, mapping and variant calling.

2.1 The materials

123 genotypes of *Brassica oleracea* have been sequenced. The accessions, coming from different sources, consist of the morphotypes cauliflower (42), broccoli (28), cabbage (31), kohlrabi (18), kale (2) and wild type (2) (tab. 1). Most of the genomes are parents from commercial F1 hybrid cultivars, also called inbred lines (produced through repeated selfing or DH step), which are expected to be homozygous. The remaining are hybrids (3 from Syngenta, 1 from Monsanto and 1 from Bejo, all cauliflower) and landraces (5 times 2 accessions from CGN) for which the degree of heterozygosity is unknown.

Table 1. Genomes from different morphotypes, obtained from multiple companies.

	Broccoli	Cauliflower	Cabbage	Kohlrabi	Kale	Wild type	Total
Rijkzwaan	21	14	11	8			54
BeJo	5	21+1*	18	8	2		55
CGN	2	2	2	2		2	10
Syngenta		3*					3
Monsanto		1*					1
Total	28	42	31	18	2	2	123

* the cauliflower hybrid lines

The reference genome is the white cabbage genome according to Liu *et al.* (2014). With the CTAB extraction method, DNA from the leaves of the *B. oleracea sp. capitata* homozygous line 02-12 was obtained. With Illumina Genome Analyser whole-genome shotgun and GS FLX Titanium sequencing technology a draft genome was constructed. The paired-end reads were assembled using SOAPdenovo to produce contigs and scaffolds. Using genetic mapping many scaffolds and contigs were assigned to chromosomal positions, producing nine pseudo-chromosome scaffolds in the final assembly.

2.2 The pipeline

2.2.1 Resequencing

For each commercial and inbred accession, a batch of ~ 100 seeds were grown into small plantlets, from which hypocotyls and leaf tissue was collected in bulk and used in a DNA extraction. For the landrace and wild accessions, which are heterogeneous, only one plant was grown for DNA extraction. Genome sequences were generated by IlluminaHiSeq 2000 platform with paired-end reads of a length of 100bp. After an initial assessment of data quality using a K-mer based approach (pers. comm., Theo Borm), it was decided to not filter prior to mapping them back to the reference genome. As the data was from a high quality, the untrimmed data was expected to affect variant calls to a *lesser effect* than the loss of overall coverage caused by nucleotide quality trimming would have. In addition, filtering for clonallity (exact duplicate read pairs putatively caused by excessive PCR-ing) was not done before mapping as the tool for this generally requires exact matches in the first bases of a read-pair, thereby becoming sensitive to sequencing errors and trimming, and retain only the first of a set of clonal reads rather than the best quality read. Instead, after mapping reads, duplicate reads were mapped as such, which means that full information is retained for variant calling.

2.2.3 Mapping

The sequence reads were mapped to the reference genome of *B. oleracea*. There are two published reference genomes, white cabbage by Liu *et al.* (2014) and rapid cycling kale by Parkin *et al.* (2014). Because our data contains many cabbage genotypes and because Liu *et al.* (2014) made their genome sequence available, their genome was used as reference. The mapping target consists of nine anchored pseudo-chromosome sequences complemented by the genetically unanchored contigs and scaffolds by Theo Borm (per. comm., 2014). The genomes of the different morphotypes were mapped to the reference genome according to Burrows-Wheeler alignment tool (BWA)(Li *et al.*, 2009). Duplicated reads that were mapped to exactly the same position on the reference genome were marked as such during mapping and have been accounted for during variation calling.

2.2.4 Variation calling

The output of mapping were Binary Alignment/Map (BAM) format files containing all the sequence reads and their assigned positions on the reference genome as well as essential quality metrics.

These BAM files were used to call SNPs and InDels using the Genome Analysis Toolkit (GATK) (McKenna *et al.*, 2010). The variant calling was done on all accessions simultaneously and results were stored in a single Variant Call Format (VCF) file. Prior to working with the file, for all accessions the heterozygotic SNP calls were changed to unknown calls.

2.2.5 Phylogenetic tree

To group and visualize the association of the genotypes/accessions to one another a phylogenetic tree is constructed and visualized in Seaview (Gouy *et al.*, 2010). The age and order of divergence of the genotypes was not taken into account, as this was already done by Cheng *et al.* (pers. comm., 2014). The VCF file was too big for Seaview, so the size was reduced by 15%, only the SNP calls having more that 50% known SNP calls for the 123 genotypes remained. This file was transformed to a tab file by the vcf-to-tab function within vcftools (version 0.1.12b) and sequences were derived per genotype by aligning all the SNP calls. From the constructed fasta file a phylogenetic tree was obtained.

2.2.6 Parameters

Selective sweep regions were selected according to combinations of five different parameters. The parameters were SweeD, variable sites (pi/π), Linkage Disequilibrium (LD) estimated by LD-decay, Tajima's D and pairwise fixation index (Fst).

2.2.6.1 SweeD

For SweeD (Pavlidis *et al.*, 2013) estimation, the data file was decreased to 18% and only contained SNP calls having more than 95% known calls for the 123 genotypes. For the tool SweeD a grid size of 50 000 was used, which is the number of divisions of the data for which the calculations were conducted. With a grid size of 50Kb, the created window sizes are between 650 and 1200 base pairs. In excel graphs were made from the CLR values.

2.2.6.2 Variable sites

The variable sites (pi/ π) was calculated by vcftools (version 0.1.12b) with a 200Kb window size and 5kb stepping size. In excel graphs were made from the output.

2.2.6.3 Linkage Disequilibrium

The to 18% reduced data file was also used for LD-decay estimation. The r^2 for each pair of SNPs within a distance of 100Kb was analysed with PLINK v1.9 (Purcell and Chang). For every 100Kb the average of r^2 against increasing distance for each pair of SNP was calculated. For every averaged 100Kb the half- r^2 LD-decay was obtained by fitting the formula y=exp(-ax) to every 100Kb. These values were plotted in excel to determine the low LD decay values, indicating a big region of LD.

2.2.6.4 Tajima's D

Tajima's D was estimated by vcftools (version 0.1.12b) in a window size of 50Kb. Graphs were made in excel.

2.2.6.5 Fixation index

The Fixation index of the genotypes was estimated by vcftools (version 0.1.12b), with a window and stepping size of 200 and 5 Kb respectively. Fst calculations from this function were according to Weir and Cockerham's paper (Weir and Cockerham, 1984). It was calculated for every possible comparison of the morphotypes cabbage, cauliflower, broccoli and kohlrabi and between the enlarged inflorescence trait in cauliflower and broccoli against the rest (cabbage, kohlrabi and kale). Graphs were made in excel from the mean Fst of three different comparisons between the morphotypes. In total four different groups of three comparisons and the comparison between enlarged inflorescence and the rest were obtained. Plotting three morphotypes at a time made a triangular comparison of the morphotypes based on low or high differentiation to one another possible. This pointed out the

similar morphotypes with a low differentiation and the one diverging from them with a high differentiation value. Figure 6 shows an example, cauliflower versus cabbage and broccoli versus cabbage show high differentiation and broccoli versus cauliflower show low differentiation around the base position of 26.475.001. In a triangular comparison cauliflower and broccoli are closely linked, while cabbage is the diverging one. This region would be considered a putative selective sweep for cabbage.

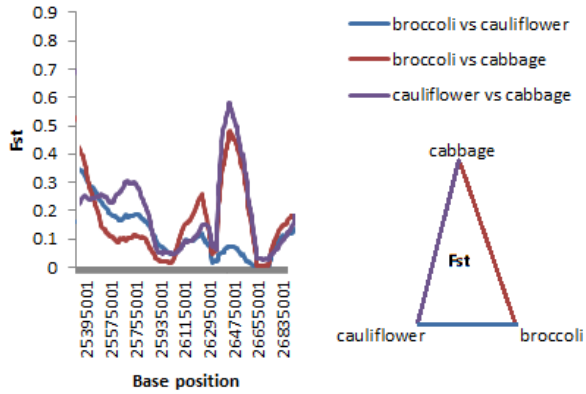


Figure 6. Example of triangular comparison between cabbage, cauliflower and broccoli for Fst from chromosome 6.

2.2.7 Selective sweep region

The selective sweep regions were selected from the graphs by eye based on the differences the parameters show between morphotypes and between the morphotype enlarged inflorescence (broccoli and cauliflower) and the rest. SweeD signals with a likelihood score above 100 were selected as potential selective sweep regions. The belonging value of LD-decay was read from the LD-decay graphs. For pi the selected regions were deemed interesting when it diverged as a low value compared to the other morphotypes or overall pattern in the graph for that chromosome. Tajima's D was deemed interesting when the value was below zero and like pi diverged from other morphotypes or the overall pattern in the graph for that particular chromosome. For Fst the regions where two of the three comparisons showed a high differentiation, one showed a low differentiation and when also the distance between high and low was prominent compared to the remaining graph of the chromosome it was selected as potential selective sweep.

The SweeD selected regions were taken as reference towards the other parameters. When one or more of the other parameters also showed interesting results related to a possible selective sweep, the region was kept for further analysis. For every chromosome and single morphotype at least one and also a certain maximum of regions were selected. There were also some regions selected based on the Fst, pi and Tajima's D, excluding SweeD. The belonging LD value is according to morphotype, but for some selective sweeps the morphotype it applied to was unclear. The results could contradict each other and therefore the LD values were no longer taken into account. For example figure 6 shows a potential selective sweep for cabbage, a contradicting result would be if pi showed a high number of variable sites for all morphotypes and a low pi for kohlrabi. For pi the region would be considered a sweep for kohlrabi, therefore the results contradict each other. When according to Tajima's D the region is a putative sweep for cabbage, the region is according to the majority selected as a potential sweep for cabbage.

2.2.8 Gene onthology

A list of the positions of the *B. oleracea* genes with synteny to *A. thaliana* and the GO-terms of *A. thaliana* were given by dr. Cheng Feng from the Institute of Vegetables and Flowers in Beijing, China (2015). From this list the corresponding genes to the putative selective sweeps and the GO-terms based on the synteny were selected. The number of genes annotated to every GO-term was counted for the putative selective sweeps (gene set) and for the whole genome. This was compared to the

number of genes in the gene set and the total genome. Based on these numbers a fisher exact test was performed to determine the enriched GO-terms. The genes containing the enriched GO-terms were selected for further literature research to explain the gene involvement to crop domestication.

3. Results and discussion

3.1 Data quality and filtering

In total 24 million variable sites were found, consisting of 3 million InDels and 21 million SNPs. The SNP calls showed a high quality, as the phred score increased from a starting value of 30, corresponding to an accuracy of 99.9% the SNP per genotype call is correct. Although the quality was high, many heterozygous calls were observed. The first 8.000 SNP calls on chromosome one for all genotypes contained 11.5% heterozygous SNP calls. Prior to variation calling, heterozygosity was tested with K-mer (31 mer) data, which didn't show much heterozygosity (Borm, pers. comm., 2014). This was also expected, as the accession were chosen based on homozygosity. The percentage of heterozygous SNP calls was therefore surprising and could not be used for calculating the parameters. Heterozygous calls would influence parameter calculations for detecting selective sweeps, as heterozygous calls show two different allele SNP calls for one position, unclear which of the two alleles is the domesticated allele. Unknown SNP alleles were ignored in calculating the parameters, so the heterozygous calls were changed to unknown calls. This caused the loss of some genotypes per morhpotype, but as the quality of the data was high the chance the homozygous calls were correct was high as well and therefore a high quality f SNP calls remained. The high number of heterozygous calls could have been obtained due to the triplicated genome. Mapping a sequence to the cabbage reference genome has probably two or three possibilities, but it will align to only the first possibility it comes across. Theo Borm (per. comm., 2014) revealed that aligning a small sequence of 100bp to the *B. oleracea* genome has a 15% chance of also finding it somewhere else. When there is also a SNP expected within the sequence the chance for finding it somewhere else increases, and more with every added SNP. The sequenced reads had a length of 100bp, but were paired reads with a known distance. This would decrease the 15% chance, but it should also be assumed SNPs could be present, again increasing the chance for finding it somewhere else. The percentage could be decreased by extending the read length, but due to the extent of duplicated regions this will only be of small effect. Also K-mers could be used as another mapping method, as for all the morphotype accessions K-mers are available (Theo Borm, pers. comm., 2014). K-mer count table (K=31) were obtained per accession. These tables were combined, subjected to a set of K-mer count-thresholds, producing a single table showing patterns of presence and absence of each individual K-mer throughout the population. These patterns were then essentially used as markers, and correlations to presence and absence of a specific morphotype (combination) was scored. Highly correlated K-mers were then used to select their underlying read-pairs, and based on these read-pairs'(BWA) map position, a graphical representation of chromosomes was drawn, essentially highlighting those areas with evidence of morphotype specific sequences. As this method essentially performs association analysis prior to read mapping, problems caused by false heterozygous calls putatively caused by duplicated regions in the genome are circumvented. In addition, K-mer selected reads can be assembled de-novo prior to mapping, possibly allowing a distinction between paralogues to be made. The obtained sequences and knowledge of paralogues could be compared to the used sequences from BWA. This will show if the mapping was correct and give some insight in how BWA handles a triplicated genome.

3.2 Window sizes

Window sizes were a point of concern to the calculations of the parameters to identify selection signals. Too small window sizes would in presence of background noise lead to false positive signals and when too large they may lead to a failure to detect small signals that are being averaged-out. Cheng *et al.* (pers. comm., 2014) found a LD of 10Kb for the *B. oleracea* genome and used a window size of 100Kb. A LD of 10Kb suggests not to use a window size smaller than 10Kb, as the LD is expected to be increased. With this knowledge and the successful use of a 100Kb window, different window sizes around 100Kb were tested per parameter to select the most optimal window size for detecting selective sweeps.

3.3 Phylogenetic tree

To show the inter relationships of the *B. oleracea* accessions, a phylogenetic tree was constructed (fig. 6). From the total dataset all SNP calls having less than 50% of the genotype calls were discarded, this subset was viewed in Seaview.

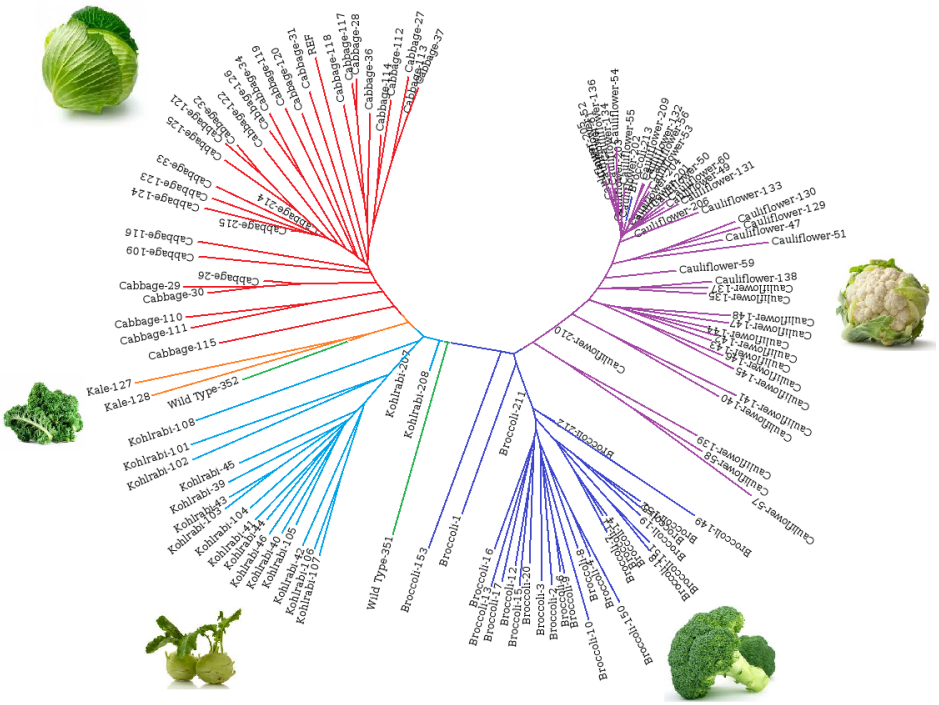


Figure 6. Phylogenetic tree of cabbage (red), kale (yellow), kohlrabi (light blue), wild type (green), broccoli (dark blue) and cauliflower (purple). The data had a cut off for 50% of missing genotypes SNP calls. Values between 0-100 and 213 are from Rijk Zwaan, between 100-200, 203 and 204 are from Bejo and WBol201-202, WBol205-212 and WBol214-215 are from CGN, Monsanto and Syngenta.

Figure 6 shows the grouping of the different morphotypes. The reference genome is grouped with the cabbages ('REF' in red), one wild type is close to kale and the other wild type is more a standalone with the closest link to kohlrabi. Many cauliflowers form a tight cluster consisting of all the different companies, while within the cluster the variation is low. The remaining five bigger cauliflowers clusters are from Bejo, with two clusters also containing Rijk Zwaan accessions. Broccoli also shows two main clusters, where the biggest cluster consists mainly of Rijk Zwaan accessions and the other cluster is a mix of Rijk Zwaan and Bejo. The main cluster of kohlrabi and cabbage is a mix of both companies. The branch length is longest for most cabbage, diverging of the branches starts close to the origin, so the variation is big within this group. Branch length for kohlrabi, broccoli and cauliflower is similar and the majority split from a single lineage. Only the biggest cluster for cauliflower has the shortest branches, this group shows the least variation towards one another. The accessions from CGN, Monsanto and Syngenta (WBol201-202, WBol205-212 and WBol214-215) show the shortest branches, these accessions are similar to the origin and show the least variation towards

the others. Cheng *et al.* (pers. comm., 2014) performed a related study about selective sweeps related to the heading trait in *B. rapa* and *B. oleracea*. Their phylogenetic tree showed a difference between the two species, *B. rapa* showed more variation and a longer branch length. Also the data for their used parameters showed more background noise for *B. oleracea* than *B. rapa*. This could suggest the fixation for *B. oleracea* occurred more recent than *B. rapa,* explaining the difference in background noise and also the short branch length and low variation within clusters.

3.4 LD calculation

The LD results (app. B) is based on the rate of loss (decay) of correlation r^2 between pairs of loci as the distance between the loci increases. The r^2 values between pairs of loci within an interval of 100Kb were computed, binned per 1Kb distance interval and averaged and plotted according to the distance between the loci. A negative exponential curve was fitted to the points, which was characterized by a single parameter directly related to the decay; lambda. The steepness of decay corresponds to higher values of lambda. Also for any given lambda, the distance at which r^2 is halved can be computes as follows: LD50%=-ln(0.5)/lambda in Kb. The results of interest are the low LD-decay values, as for selective sweeps the LD is increased. The LD results shows for every chromosome and morphotype some high LD-decay values and some also span over a broad distance up to 500Kb. A high LD-decay of lambda 7 is an LD50% of 99 bases, this then over a broad distance means many recombinations must have occurred across the genes in the corresponding region and will therefore have no linkage to neighbouring genes. Such a low LD50% and many recombinations in a region up to 500kb was highly unexpected, as for a selective sweep an increased LD (low LD decay) was expected. Even so, some of the putative selective sweeps show a high LD-decay (app. A). As there were only a few regions with a high LD-decay, it was suspected the merging of contigs was the cause. The genome sequence consists of aligned contigs, with as an alignment a small predicted sequence. These small sequences do not belong to either contigs and therefore can cause high LD-decay values. This hypothesis was verified, but the aligned regions did not correspond to the high LD-decay regions. Another reason could be the incorrect alignment due to triplication. As stated before, mapping of the sequences to the reference genome has a chance to be found at a duplicated or triplicated region as well. When reads map to the wrong positions, two regions can be aligned that actually should be aligned far apart from one another and therefore are not connected. This causes the high LD-decay values in some regions. As contigs consist of scaffold and the scaffolds consist of the reads, the scaffold borders could also be checked with the high LD-decay regions. Possibly the scaffolds were wrongly aligned, based on too little available information. As mentioned before, also the mapping method could be validated at how it handles paralogs. As when the mapping method consistently swap paralogs, the genotype calls should also be consistent and therefore it could be less of an influence on r^2 calculations and not be cause for the high LD-decay values. However, if the mapping method doesn't show consistency and shows to be the possible cause for the high LD-decay values, the reference genome should also be questioned for its correctness. This could be tested with information obtained from the K-mer data, by selecting the paralogs aligning to multiple regions and removing them. Calculation can be done for the new dataset and compared to previous results. Because the LD results were unreliable, they were not implemented in further analysis.

3.5 Selective sweep

Selective sweep regions were mainly chosen based on the SweeD (App. C), as this was the only parameter with an included statistical test and therefore the most reliable. When one of the other parameters Fst (App. D), variable sites (pi/ π) (App. E) or Tajima's D (App. F) conformed the region by showing an extreme value as well, it was defined as a putative selective sweep. However the results could contradict each other. Not always did the signal for a specific morphotype from SweeD also show differences for the other parameters and sometimes another morphotype did show a difference for the same region with the other parameters. The region would still be classified to the morphotype according to SweeD. The selective sweep region could also be only based on Fst, pi or Tajima's D when results show extreme values and/or differences, without a SweeD signal. Here the

parameters could contradict each other on the morphotype as well, so classification of the selective sweep to a morphotype was according to the majority. Also some selective sweeps belonged to the rest group (mainly cabbage and kohlrabi) and sometimes it was not possible to determine to which of the two morphotypes the sweep belonged to, therefore also an 'undecided' group was created.
Putative selective sweeps were chosen by eye based on differences and the values of the results in the graphs. The data showed a lot of background noise and was not normally distributed (fig. 7,8,9), this troubled making use of a normalization method. From the frequency histograms in figure 7, 8 and 9 per test, morphotype and chromosome an average and standard deviation for normalization could be read from one of the peaks in the multimodal histogram. For the Fst on linkage group C09 of kohlrabi, for example, we see a Fst distribution with three peaks (fig. 7). A high value was expected, so the average and standard deviation of the third peak of the multimodal histogram could be taken. Based on these values a cut-off at for instance 50% could be taken, as already a certain cut-off is made by taking an average from the third peak. For the variable sites of cauliflower on linkage group C04 (fig. 8) and Tajima's D of cabbage for C01 (fig. 9) the lowest values were of interest. However, both show not a clear number of peaks making it impossible to take an average from the graphs. Therefore no normalization of the data was conducted.

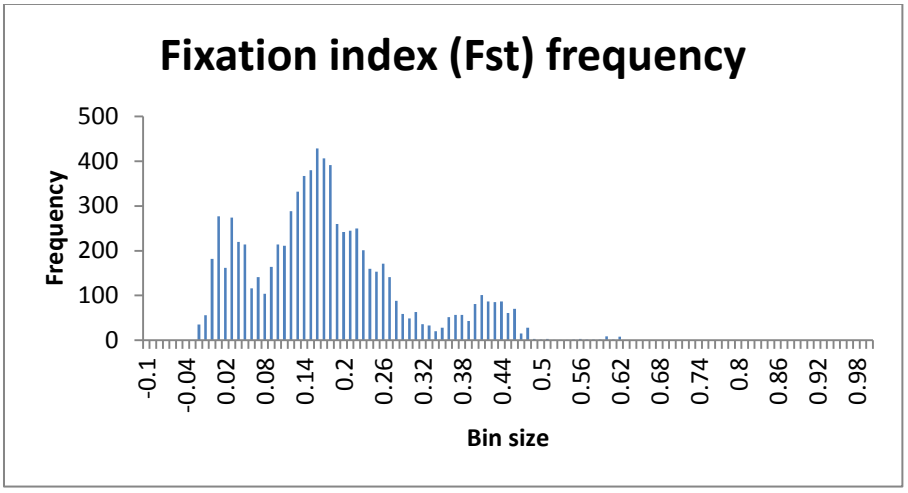


Figure 7. Histogram of fixation index values for cabbage versus kohlrabi for linkage group C09.

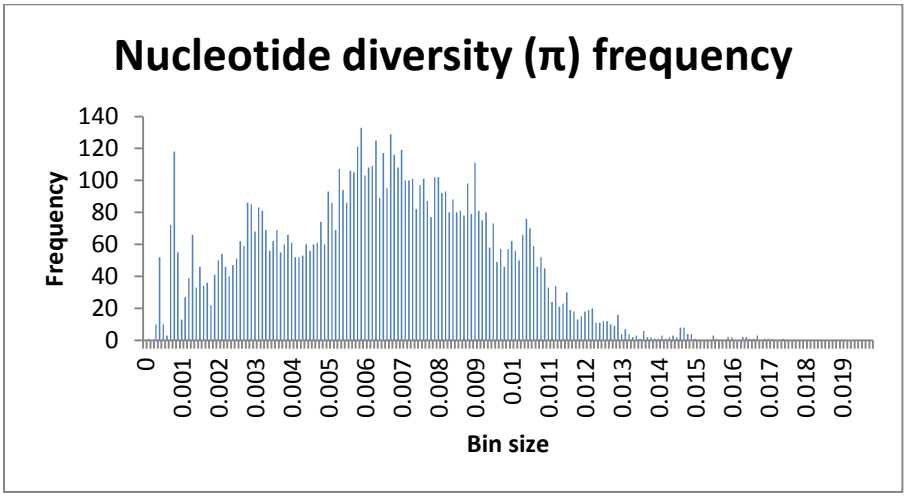


Figure 8. Histogram of variable sites values of cauliflower for linkage group C04.

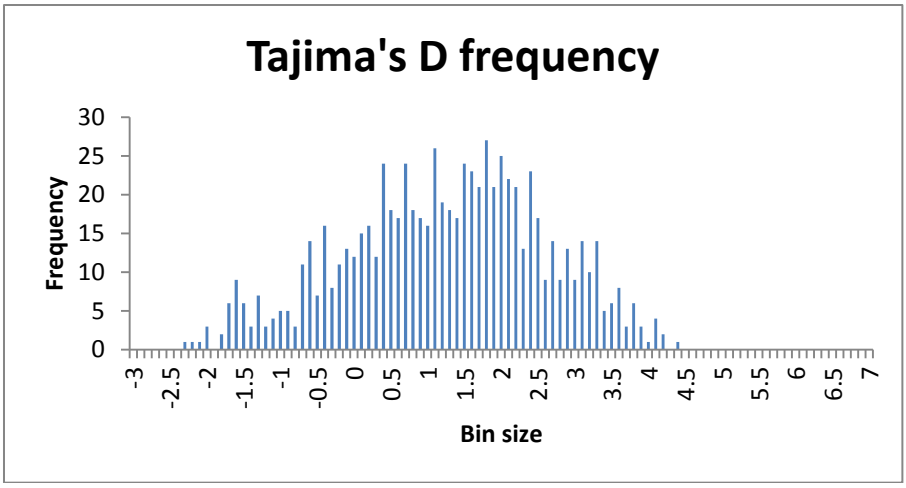


Figure 9. Histogram of Tajima's D values of cabbage for linkage group C01.

The chosen regions are given in the tables 8-13 in appendix A. Because the regions were only chosen by eye of one individual and not all putative regions were selected due to the amount and available time, there're still many regions left to analyse. The selected regions should also be validated, as in this case, by eye selection corresponds to a subjective selection, due to lack of threshold values.
In total 102 putative selective sweep regions were found. Based on the synteny with *A. thaliana* from Cheng Feng (pers. comm., 2015) for 75 regions genes were found. For 27 regions no synteny with *A. thaliana* was found. As only the cabbage reference genome was compared, the 27 uncharacterized regions could contain yet to be discovered genes from the other morphotypes. They could also have gained different genes through evolution which lost their synteny to *A. thaliana* or are not associated with *A. thaliana*. The regions are still a potential selective sweep, but for now not further analysed due to lack of genes with synteny to *A. thaliana*. The Bolbase database (Yu *et al.*, 2013) is a database for *B. oleracea*, where annotation of the reference genome has been done. This annotation is not only based on *A. thaliana* and therefore would give more possibilities. It was not used as there was no GO-term list of the total genome available and therefore only the annotation to *A. thaliana* could be used for GO-enrichment analysis. If an annotation for the whole genome would be available also the 27 unknown regions could be tested for enriched GO-terms.

3.6 Enriched GO-terms

The domesticated genes are the genes of interest, but the selective sweep regions also contain the hitchhiking genes. It is expected for hitchhiking genes to appear randomly and domesticated genes to show an enrichment of functionally selected traits. So to divide the domesticated genes from the hitchhiking genes, GO enrichment analysis was performed on all selected selective sweeps. In total 31.779 *B. oleracea* genes are in synteny with *A. thaliana*, 31.470 are with known GO-terms (Cheng Feng, pers. comm. 2015). There are 554 genes found in the selective sweep gene set (app. A), of which 549 with known GO-terms. Genes can consist of multiple GO-terms, so those 549 genes comprise of 2.294 GO-terms, of which 591 are unique. For those 591 GO-terms the number of occurrence of the GO-terms in the whole genome was 111.382 in total. Based on this data the fisher exact test gives 49 unique significant GO-terms for which the genes in the selected regions are significantly enriched. Tables 2-7 shows the GO-term distribution over the genes per morphotype. In total 89 of the 554 genes contain significant GO-terms, see appendix G. Appendix G also shows for 56 out of 89 genes the known genome fractionation LF, MF1 or MF2 (Cheng *et al.* 2014; Liu *et al.*, 2014). The majority of the domesticated genes belong to the LF genome, this could mean only one copy of the gene was still present and must have undergone a loss of function.
It's notable that the number of "significant regions" found based on either the first parameter SweeD or the three parameters Fst/pi/Tajima's D are almost the same, while the number of selected regions with known GO-terms is higher for genes in regions detected by SweeD than for regions detected by

the other parameters. Eighteen genes with enriched GO terms are found significant from the total of 53 regions based on SweeD and 19 genes with enriched GO terms are significant from the total of 24 regions based on Fst/pi/Tajima's D. SweeD was preferred over the other parameters and therefore more regions were selected, but it was also expected those regions contained more domesticated genes. It was also observed that SweeD showed not as much comparable regions to other parameters as the other parameters towards each other. Fst, pi and Tajima's D showed multiple regions where for all three parameters some extreme values or difference were observed. When comparing SweeD to the other parameters there were only few comparable regions with extreme values or difference, most were moderate differences for only one or two other parameters. From the results the regions based on only Fst, pi and Tajima's D contained more domesticated genes, these parameters therefore showed to be superior parameters to detect selective sweeps over SweeD. As SweeD did not show many similar regions as the other methods, the SweeD regions were mainly based on a single parameter, while the other regions were selected based on multiple parameters. So this could also mean the use of multiple parameters is more accurate than only a single parameter. However, SweeD could also be more sensitive to background noise compared to the other parameters. Background noise could increase the chances for detecting false positives by calculating the parameters. A bigger population size for all morphotypes should test this. As with GO enrichment analysis, background noise is at random and the selective sweeps should show the same sweeps for the morphotypes. With an higher population size the test number increases and also the reliability of the result. A bigger population size should be tested and statistical tests should set a threshold, so also the sweeps close to the threshold are chosen and not only the extreme ones.

Table 2. GO enrichment of the GO-terms in the potential selective sweeps of broccoli, from the fisher exact test (p-value <0.05) and the GO-term names.

GO-term	Present in nr. of genes in geneset	Nr of GO-term in *B. oleracea* based on synteny with *A. thaliana*	Fisher exact test p-value <0.05	GO-term name
GO:0010227	1	33	0.02381206	Floral organ abscission
GO:0010374	1	17	0.004661369	Stomatal complex development
GO:0005507	2	323	0.01476715	Copper ion binding
GO:0009236	1	1	0.03405796	Cobalamin biosynthetic process
GO:0004674	7	881	0.0457431	Protein serine/threonine kinase activity
GO:0004722	1	220	0.01887069	Protein serine/threonine phosphatase activity
GO:0008287	1	58	0.0222001	Protein serine/threonine phosphatase complex
GO:0010050	1	10	0.01739165	Vegetative phase change
GO:0009840	1	9	0.01465925	ChloroplasticendopeptidaseClp complex
GO:0006461	1	13	0.02674085	Protein complex assembly

Table 3.GO enrichment of the GO-terms in the potential selective sweeps of cauliflower, from the fisher exact test (p-value <0.05) and the GO-term names.

GO-term	Present in nr. of genes in geneset	Nr of GO-term in *B. oleracea* based on synteny with *A. thaliana*	Fisher exact test p-value <0.05	GO-term name
GO:0004722	2	220	0.01887069	Protein serine/threonine phosphatase activity
GO:0008287	1	58	0.0222001	Protein serine/threonine phosphatase complex
GO:0010628	1	1	0.03405796	Positive regulation of gene expression
GO:0005669	2	14	0.03021659	Transcription factor TFIID complex
GO:0016706	1	66	0.03278898	Oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, 2-oxoglutarate as one donor, and incorporation of one atom each of oxygen into both donors
GO:0004674	2	881	0.0457431	Protein serine/threonine kinase activity

Table 4. GO enrichment of the GO-terms in the potential selective sweeps of enlarged inflorescence, from the fisher exact test (p-value <0.05) and the GO-term names.

GO-term	Present in nr. of genes in geneset	Nr of GO-term in *B. oleracea* based on synteny with *A. thaliana*	Fisher exact test p-value <0.05	GO-term name
GO:0004379	1	1	0.03405796	Glycylpeptide N-tetradecanoyltransferase activity
GO:0004674	5	881	0.0457431	Protein serine/threonine kinase activity
GO:0004722	2	220	0.01887069	Protein serine/threonine phosphatase activity
GO:0008287	1	58	0.0222001	Protein serine/threonine phosphatase complex
GO:0016117	1	22	0.008821197	Carotenoid biosynthetic process
GO:0016123	1	14	0.002889796	Xanthophyll biosynthetic process
GO:0010020	2	16	0.03766475	Chloroplast fission
GO:0010064	1	1	0.03405796	Embryonic shoot morphogenesis
GO:0019107	1	1	0.03405796	Myristoyltransferase activity
GO:0006461	1	13	0.02674085	Protein complex assembly
GO:0008131	1	25	0.001478845	Primary amine oxidase activity
GO:0005507	1	323	0.01476715	Copper ion binding
GO:0016040	1	1	0.03405796	Glutamate synthase (NADH) activity
GO:0010374	1	17	0.004661369	Stomatal complex development
GO:0001522	1	19	0.04997927	Pseudouridine synthesis
GO:0009954	2	10	0.01739165	Proximal/distal pattern formation
GO:0010022	2	10	0.01739165	Meristem determinacy
GO:0010227	2	33	0.02381206	Floral organ abscission
GO:0010254	2	7	0.009816998	Nectary development
GO:0048439	2	10	0.01739165	Flower morphogenesis

Table 5. GO enrichment of the GO-terms in the potential selective sweeps of cabbage, from the fisher exact test (p-value <0.05) and the GO-term names.

GO-term	Present in nr. of genes in geneset	Nr of GO-term in *B. oleracea* based on synteny with *A. thaliana*	Fisher exact test p-value <0.05	GO-term name
GO:0016706	1	66	0.03278898	Oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, 2-oxoglutarate as one donor, and incorporation of one atom each of oxygen into both donors
GO:0004674	4	881	0.0457431	Protein serine/threonine kinase activity
GO:0004722	1	220	0.01887069	Protein serine/threonine phosphatase activity
GO:0008287	1	58	0.0222001	Protein serine/threonine phosphatase complex
GO:0000338	2	9	0.01465925	Protein deneddylation
GO:0008180	2	19	0.006137474	COP9 signalosome
GO:0010387	2	10	0.001279769	COP9 signalosome assembly
GO:0010388	2	14	0.03021659	Cullindeneddylation
GO:0048564	1	3	0.002855097	Photosystem I assembly
GO:0009840	1	9	0.01465925	ChloroplasticendopeptidaseClp complex
GO:0016043	2	18	0.04573067	Cellular component organization
GO:0010291	2	8	0.01213169	Carotene beta-ring hydroxylase activity
GO:0016117	2	22	0.008821197	Carotenoid biosynthetic process
GO:0016123	2	14	0.002889796	Xanthophyll biosynthetic process
GO:0005507	4	323	0.01476715	Copper ion binding
GO:0008131	3	25	0.001478845	Primary amine oxidase activity
GO:0048038	3	9	0.000997248	Quinone binding
GO:0016667	1	1	0.03405796	Oxidoreductase activity, acting on a sulfur group of donors
GO:0009378	1	1	0.03405796	Four-way junction helicase activity
GO:0043138	1	1	0.03405796	3'-5' DNA helicase activity

Table 6. GO enrichment of the GO-terms in the potential selective sweeps of kohlrabi, from the fisher exact test (p-value <0.05) and the GO-term names.

GO-term	Present in nr. of genes in geneset	Nr of GO-term in *B. oleracea* based on synteny with *A. thaliana*	Fisher exact test p-value <0.05	GO-term name
GO:0004565	3	33	0.02381206	Beta-galactosidase activity
GO:0005507	4	323	0.01476715	Copper ion binding
GO:0004722	2	220	0.01887069	Protein serine/threonine phosphatase activity
GO:0000373	2	6	0.00772341	Group II intron splicing
GO:0004674	1	881	0.0457431	Protein serine/threonine kinase activity
GO:0010374	1	17	0.004661369	Stomatal complex development
GO:0001522	1	19	0.04997927	Pseudouridine synthesis
GO:0019513	2	18	0.04573067	Lactose catabolic process, using glucoside 3-dehydrogenase
GO:0019515	2	19	0.04997927	Lactose catabolic process via UDP-galactose
GO:0045292	1	1	0.03405796	mRNA cis splicing, via spliceosome
GO:0048564	1	3	0.002855097	Photosystem I assembly
GO:0010050	1	10	0.01739165	Vegetative phase change
GO:0016706	1	66	0.03278898	Oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, 2-oxoglutarate as one donor, and incorporation of one atom each of oxygen into both donors

Table 7. GO enrichment of the GO-terms in the potential selective sweeps of undecided for cabbage or kohlrabi, from the fisher exact test (p-value <0.05) and the GO-term names.

GO-term	Present in nr. of genes in geneset	Nr of GO-term in *B. oleracea* based on synteny with *A. thaliana*	Fisher exact test p-value <0.05	GO-term name
GO:0005507	1	323	0.01476715	Copper ion binding
GO:0004411	1	1	0.03405796	Homogentisate 1,2-dioxygenase activity
GO:0006570	1	1	0.03405796	Tyrosine metabolic process
GO:0006572	1	1	0.03405796	Tyrosine catabolic process
GO:0030755	2	5	0.005859395	Quercetin 3-O-methyltransferase activity
GO:0033799	2	5	0.005859395	Myricetin 3'-O-methyltransferase activity
GO:0047763	2	5	0.005859395	Caffeate O-methyltransferase activity
GO:0051555	2	10	0.01739165	Flavonol biosynthetic process
GO:0004674	4	881	0.0457431	Protein serine/threonine kinase activity
GO:0016706	1	66	0.03278898	Oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, 2-oxoglutarate as one donor, and incorporation of one atom each of oxygen into both donors
GO:0004722	1	220	0.01887069	Protein serine/threonine phosphatase activity
GO:0008180	1	19	0.006137474	COP9 signalosome
GO:0010387	1	10	0.001279769	COP9 signalosome assembly

In the following paragraphs a few selected GO terms are discussed based on number of occurrence and if part of a cluster of GO-terms or functions according to tables 2-7. Genes containing the selected enriched GO-terms were further analysed by obtaining the orthologous *A. thaliana* gene. Some orthologous genes showed to be part of a cluster based on function or pathway and were therefore also further analysed.

3.6.1 Kinase activity

The most common GO-terms in this gene set are GO:0004674 (Protein serine/threonine kinase activity), GO:0004722 (Protein serine/threonine phosphatase activity) and GO:0005507 (Copper ion binding). The group of protein serine/threonine kinases and phosphatase activity play an important role in signal transduction pathways (Nemoto *et al.*, 2011). As some orthologous genes of *A. thaliana* seem to be involved in pathways regulated by plant hormone Abscisic acid (ABA). ABA is involved in

many plant growth and development stages as cell differentiation during embryogenesis and this regulates many responses to abiotic stresses as drought, cold and salinity. In *A. thaliana* there are nine clade A protein phosphatase 2Cs (PP2Cs), of which six are known to be negative regulators of ABA signaling (Bhaskara *et al.*, 2012). PP2C10 is ortholog to Bol031987 for kohlrabi or cabbage ('undecided') and PP2C family protein is ortholog to Bol011630 for enlarged inflorescence, Bol014404 for broccoli and Bol015697 for kohlrabi. Another one of these six PP2Cs is ABA Insensitive1 (ABI1), the ortholog of Bol037987 for enlarged inflorescence. *Abi1* mutant affects ABA responses to stomatal closure, maintenance of seed dormancy and inhibition of plant growth (Meyer *et al*., 1994). And according to Leung *et al.* (1994) ABI1 also regulates stomatal aperture in leaves and mitotic activity in root meristems. An ABA-insensitive mutant *abi1-1* reduces Pro accumulation (Bhaskara *et al.*, 2012). Pro is essential for growth and redox buffering when a low soil water potential and salt stress applies (Bhaskara *et al.*, 2012). The function of the remaining three PP2Cs are still uncertain, though mutants showed to have increased Pro and osmoregulatory solute accumulation at a low soil water potential, double or triple mutants showed in contrast to the other six PP2Cs ABA-insensitive seed germination (Bhaskara *et al.*, 2012). One of the three functionally uncertain PP2Cs is HAI3, the ortholog of Bol017589 in cauliflower.

ABA is also suggested to regulate mitogen-activated protein kinase (MAPK) activity, as it induces MAPK activation in barley protoplasts (Knetsch *et al.*, 1996). AP2C1, a PP2C-type phosphatase and ortholog of Bol017649 in cabbage, inactivates the stress-responsive proteins MPK4 and MPK6 (Schweighofer *et al.,* 2007). This suggests PP2Cs regulate MAPKs (Schweighofer *et al.,* 2007), indirectly through ABA or directly. MAPK is involved in cell growth and differentiation (Knetsch *et al.*, 1996) and in the wound signalling pathway, together with the other plant hormones jasmonates and ethylene (Schweighofer *et al.,* 2007). Cell growth and differentiation especially are interesting aspects that can play an important role in the development of the different crops. An enriched gene in cabbage and the ortholog of Bol017650 is a MAPKKK14 protein, involved in the MAPK activity. Another MAPK, ATMAPKK5, ortholog to Bol026625 for broccoli, is involved in the process of programmed separation of entire organs called abscission, which promotes discarding non-functional or infected organs (Cho *et al.*, 2008). MAPKK5 regulates the floral organ abscission in particular (Cho *et al.*, 2008).

AT5G57670 is the ortholog of Bol012196 in cauliflower and is shown to be one of the two protein kinases down regulated by the plant growth repressor DELLA (Chao *et al.*, 2006). DELLA acts according to gibberellin (GA), signalling seed germination, stem elongation and floral development pathways (Chao *et al.*, 2006).

The gene "Barely any meristem 2"(BAM2) is the ortholog of Bol016985 in broccoli and has a serine/threonine kinase activity. Interestingly BAM proteins seem to be involved in shoot and flower meristem stem cell specification, this is initially regulated by CLAVATA1 (CLV1) receptor kinase (DeYoung and Clark, 2008). DeYoung and Clark (2008) showed BAM proteins function similar to CLV1 on a biochemical level and function in parallel with CLV1 in the centre of the meristem, but on their own in the surrounding meristem. BAM seems to be involved in the development of the floral organs and leaf vascular patterning (DeYoung and Clark, 2008).

3.6.2 COP9 signalosome protein

Some GO-terms together belong to the with yet unknown function COP9 signalosome (CSN) protein complex from the ubiquitin-proteasome pathway; GO:0000338 (Protein deneddylation, which is the removal of a protein from the CSN complex), GO:0008180 (COP9 signalosome), GO:0010387 (COP9 signalosome assembly) and GO:0010388 (cullin deneddylation). In *A. thaliana* the COP9 signalosome complex removes the "related to ubiquitin" protein (RUB, also known as NEDD8)('neddylation') from the cullin subunit of SCF (SKP1, Cullin, F-box)-type E3 ubiquitin ligases (Schwechheimer and Deng, 2001). The RUB activating enzyme is encoded by auxin resistant1 (AXR1) and E1 C-terminal-related1 (ECR1) (Nakasone *et al.*, 2012). AXR1 is also believed to be involved with leaf polarity establishment (Li *et al.*, 2007). RUB1 can be removed by COP9 ('deneddylation') and causes repression of the transcription of auxin/indole-3-acetic acid (AUX/IAA) genes (Schwechheimer and Deng, 2001). Auxin

regulates a broad range of cellular and developmental responses, including cell division, expansion and differentiation and distribution of growth between primary and lateral root and shoot meristems (Reed, 2001). Some CSNs are involved in the processes of neddylation and deneddylation, CSN5, CSN6 and CSN9 are orthologs of Bol012321 (CSN5) in a cabbage selective sweep, Bol009898 (CSN6) in an enlarged inflorescence selective sweep and Bol027056 (CSN9) in a cabbage selective sweep respectively. CSN6 is connected to the core subunit of SCF, Cullin, which is modified by RUB (Schwechheimer and Deng, 2001). CSN5 is believed to play a role in the deneddylation process (Schwechheimer and Deng, 2001). Deneddylation causes IAA degradation, Auxin response factors (ARFs) respond to this degradation and changes the patterns of downstream gene expression (Nakasone *et al*., 2012). It is also shown the ARF genes are involved with adaxial-abaxial polarity establishment and this plays a role in the inward curvature of folding leafs (heading trait) (Wu *et al.*, 2008). In a related study by Cheng *et al.* (pers. comm., 2014) in *B. rapa* and *B. oleracea*, AXR1 and ARF genes were also found and considered to be involved in the heading trait of cabbage.

3.6.3 Carotene pathway

There are some GO-terms involved in the carotenoid pathway; GO:0016117 (Carotenoid biosynthetic process) which contains GO:0016123 (Xanthophyll biosynthetic process). They are also part of the pigment biosynthetic process, to which GO:0051555 (Flavonol biosynthetic process) belongs as well. Carotenoids are essential components for the photosynthetic antenna, they provide flower- and fruit colour and in high accumulation attract pollinators (Tian and DellaPenna, 2001). Bol037991 from enlarged inflorescence is homolog of Lutein deficient (LUT) 2, Bol027064 and Bol027080 from cabbage are homologs of LUT5. A mutant line of different Lutein deficient genes, including LUT2 and LUT5, showed photosensitivity due to a failure of photo protection mechanisms (Fiore *et al.*, 2012).

3.6.4 Inflorescence

Some GO-terms involve the inflorescence; GO:0010022 (Meristem determinacy), GO:0010227 (Floral organ abscission), GO:0048439 (Flower morphogenesis), GO:0010254 (Nectary development) and GO:0010064 (Embryonic shoot morphogenesis). Meristem determinacy includes floral meristem determinacy. Floral organ abscission is part of floral organ development. Nectary development and flower morphogenesis are part of flower development and flower morphogenesis also contains establishment of floral organ orientation, floral whorl morphogenesis, flower formation and flower structural organization. And all, except for meristem determinacy, are part of the shoot system development. Bol007570 and Bol007571 in the selective sweep for enlarged inflorescence contain some of these enriched GO-terms and are homolog of Blade on petiole (BOP) 1. BOP1 is involved in the control of determinacy and architecture of floral shoots, activating Apetala1 (AP1) and repressing Agamous-like24 (AGL24) (Xu *et al.*, 2010). Apetala1 is a key regulator of floral meristem identity, while AGL24 is a MADS-box flowering-time gene and when over-expressed lead to an inflorescence meristem identity in *A. thaliana* (Xu *et al.*, 2010). BOP1 can therefore be the cause or one of the causes for the establishment of enlarged inflorescence in cauliflower and broccoli.
Surprisingly, AP1 and CAULIFLOWER (CAL) ortholog regions were not identified as a selective sweep region. Like BOP1, CAL positively regulates AP1 and so they both are involved in flower meristem specification and floral organ specification (Bowman *et al.*, 1993). The floral meristem of a double mutant acts as an inflorescence meristem (Yanofsky, 1995). Cheng *et al.* (pers. comm., 2014) has found the genes in selective sweeps, though for detecting CAL the window size was changed to 20Kb. CAL seemed to be a small signal and therefore traceable with a smaller window size. But this would also give more noise and false positives if the selective sweeps would be traced with a smaller window size and without pre-knowledge on the base position. Though BOP1, an AP1 activating gene was found as domesticated gene, AP1 itself was not found. The cause for this is still unclear, possible reasons could be the window size or the selection by eye was not sensitive enough and missed the AP1 signal.

3.6.5 Oxidoreductase activity

GO:0010291 (Carotene beta-ring hydroxylase activity), GO:0016040 (Glutamate synthase (NADH) activity),GO:0016667 (Oxidoreductase activity, acting on a sulphur group of donors), GO:0016706 (Oxidoreductase activity, acting on paired donor), GO:0008131 (Primary amine oxidase activity) and GO:0004411 (Homogentisate 1,2-dioxygenase activity) are all part of oxidoreductase activity. Oxidoreductase activity, acting on paired donor, contains many processes and a few of them have gibberellins oxidase and dioxygenase activity. Also the degradation of plant hormone cytokinin is one catalysed by oxidase/dehydrogenase (CKX) enzymes, for which CKX3 and CKX5 have shown to regulate the activity of the reproductive meristems of *A. thaliana* (Bartrina *et al.*, 2011). CKX3 is ortholog to Bol037999, a gene in the enlarged inflorescence selective sweep. Bartrina *et al*. (2011) showed an increase of cytokinin in inflorescence would give more flower primordia, as there is formation of a larger inflorescence meristem. They also showed a double mutant ckx3-ckx5 would form larger inflorescence and floral meristems.

3.6.6 Transferase activity

Other GO-terms are part of transferase activity; GO:0004379 (Glycylpeptide N-tetradecanoyltransferase activity), GO:0019107 (Myristoyltransferase activity), GO:0030755 (Quercetin 3-O-methyltransferase activity), GO:0033799 (Myricetin 3'-O-methyltransferase activity) and GO:0047763 (Caffeate O-methyltransferase activity). The last three are part of the methylation process. Caffeic acid *O*-methyltransferase (OMT) 1 is the ortholog of Bol038831 and Bol038832 for either kohlrabi or cabbage ('undecided') and is part of the lignin pathway. Lignin is an important component for cell walls, giving plants strength and the capability of long-distance water transport (Zhang *et al.*, 1997). Lin *et al.* (2014) suggested lignin may be important for the tuber formation in *B. rapa*, which also explains the increased number of xylem vessels in the tuber. The involvement of peroxidases in tuber formation was also shown for potato (vanEck *et al.*).

3.6.7 Lactose metabolic process

GO:0019513 and GO:0019515 are both part of the lactose metabolic process and found in β-galactosidase (BGAL)2. BGAL1 and BGAL2 are orthologs of Bol005945 and Bol005408, Bol005407 respectively and found in the potential selective sweep regions in kohlrabi. In *A. thaliana* BGAL1 seems to be involved with expansion of cotyledons, rosette leaves and cauline leaves, while BGAL2 is active during hypocotyl elongation in light and dark conditions (Albornos *et al.*, 2012). The relevance as domesticated gene is unclear, but could be of interest as two BGALs were found and only in kohlrabi.

4. Conclusion

So, maybe due to incorrect mapping to the reference genome, many heterozygous calls were found. It was shown that a small sequence of 100bp has a 15% chance of aligning to another region and this percentage rises when a SNP is present in the sequence (Borm, pers. comm., 2014). This could also explain the broad and high values of LD decay, as some sequences were closely mapped while in reality their position is much further apart and therefore the region was seen as an high LD-decay region. Though this hypothesis needs further analysis, which can be obtained by identifying paralogs as in the K-mer mapping method and removing them. With the 'problematic' regions removed the analysis can be repeated and compared to previous results. When this shows a difference, it could mean the mapping method cannot handle triplicated genomes in a correct way, which assumes the reference genome is incorrect as well. It also seemed the data of *B. oleracea* showed more background noise than *B. rapa*. The phylogenetic tree (fig. 6) shows short branch length with little variation within groups, while in a related study by Cheng *et al.* (pers. comm., 2014) it was shown the branch length for *B. rapa* was longer and therefore more variable. Their parameter results also showed less background noise for *B. rapa*. This suggests *B. oleracea* fixated more recently than *B. rapa.* The background noise in *B. oleracea* makes it harder for the parameters to detect the true

selective sweeps. This was also observed between the parameters, as SweeD did not show many similar results towards other parameters as the other parameters towards each other, suggesting SweeD is more sensitive to background noise. A bigger population size could overcome this problem, as the tests are still based on a selected phenomenon, which will filter out the random background. Also using a threshold would be more reliable and would give more results, as by eye only the extreme values and differences were selected. The same methods can be repeated for the recently started 1000 genome project of *B. oleracea* (Bonnema, pers. comm., 2015), where the population size will be much bigger and with the advantage of the now obtained knowledge.

A list of all the aligned GO-terms to the *B. oleracea* genes was not yet available from the Bolbase database. The gene selection in the regions and GO-enrichment was therefore restricted to the synteny list with *A. thaliana* from Cheng Feng (pers. comm., 2015). For some putative selective sweeps no genes with synteny to *A. thaliana* were found and therefore were not further analysed. It is also possible the reference genome does not contain any genes for the regions, as they are genes particular for another morphotype than cabbage. Therefore the regions could still contain important domesticated genes and should be further analysed once a GO-term list for Bolbase and reference genomes for the other morphotypes are available. This also applies to the available data of fractionated genes in *B. oleracea,* only for some the fractionation is known. For the known domesticated genes the majority belonged to the LF. This suggests a loss of function in a single gene, where all the paralogs have been lost through natural selection. A loss of function causing extreme morphotypes was already observed for domesticated genes as CAL and AP1, where a loss of function causes the enlarged inflorescence trait.

And as expected, many domesticated genes showed involvement with plant hormones ABA and IAA, as they play important roles in plant growth and development. The majority of the genes also have serine/threonine kinase or phosphatase activity, involved in signal transduction pathways. And there are two major traits represented by several domesticated genes, enlarged inflorescence for broccoli and cauliflower and heading for cabbage. The orthologs of BOP1, BAM2, CKX3, a MAPK involved in involved in floral organ abscission and an AT5G57670 with serine/threonine kinase activity associated with DELLA seem to be involved with enlarged inflorescence. The orthologous genes of several CSNs showed involvement in the heading trait of cabbage. Also the orthologous gene OMT1 was found in a selective sweep in kohlrabi, part of the lignin pathway which is suggested to be involved in tuber formation (Lin *et al.*, 2014). The remaining orthologous genes and their function show some important involvement in plant growth and development, but need more analysis for their exact role in crop domestication. As enlarged inflorescence is extensively studied in *A. thaliana,* while about tuber formation not much is known, it is possible some of the genes containing enriched GO-terms are involved in yet to be discovered pathways involved in specific morhpotypic traits. The found genes should also still be validated for their exact function and involvement in the suggested traits.

To conclude, the majority of the domesticated genes seem to be involved with the plant hormones. Two traits are represented by a group of domesticated genes; enlarged inflorescence and heading. It is suggested the methods of mapping, selection of the selective sweeps and the population size should be revised for better results. This can be done in the recently started 1000 genome project. But overall more fundamental insight in growth and development for *B. oleracea* is obtained, which will be helpful for future breeding.

References

Ahn, Y. O., et al. (2007). "Functional genomic analysis of *Arabidopsis thaliana* glycoside hydrolase family 35." Phytochemistry**68**(11): 1510-1520.

Albornos, L., et al. (2012). "Promoter activities of genes encoding β-galactosidases from Arabidopsis a1 subfamily." Plant Physiology and Biochemistry**60**: 223-232.

Axelsson, E., et al. (2013). "The genomic signature of dog domestication reveals adaptation to a starch-rich diet." Nature**495**:360-364.

Bamshad, M. and S. P. Wooding (2003). "Signatures of natural selection in the human genome." Nature Reviews Genetics**4**(2): 99-111.

Bartrina, I., et al. (2011). "Cytokinin regulates the activity of reproductive meristems, flower organ size, ovule formation, and thus seed yield in Arabidopsis thaliana." The Plant Cell Online 23(1): 69-80.

Bhaskara, G. B., et al. (2012). "Unique drought resistance functions of the highly ABA-induced clade A protein phosphatase 2Cs." Plant physiology**160**(1): 379-395.

Bonnema, G., et al. (2011). "Diversity analysis and molecular taxonomy of Brassica vegetable crops." Genetics, genomics and breeding of crop plants. Enfield, USA: Science Publishers: 81-124.

Bonnema, G. (pers. comm., 2015) Wageningen university and research centre, plant breeding

Borm, T., (pers. comm.., 2014). Wageningen university and research centre, plant breeding

Bowman, J. L., et al. (1993). "Control of flower development in Arabidopsis thaliana by APETALA 1 and interacting genes." DEVELOPMENT-CAMBRIDGE- **119**: 721-721.

Cao, D., et al. (2006). "Gibberellin mobilizes distinct DELLA-dependent transcriptomes to regulate seed germination and floral development in Arabidopsis." Plant physiology 142(2): 509-525.

Cheng, F., et al. (2012). "Biased gene fractionation and dominant gene expression among the subgenomes of Brassica rapa." PloS one **7**(5): e36442.

Cheng, F., et al. (2014). "Genome triplication drove the diversification of *Brassica* plants." Horticulture Research**24**

Cheng, F., et al. (pers. comm., 2014). "The genomic signatures of independent evolution of leaf heading morphotypes in *Brassica rapa* and *Brassica oleracea.*" 1:22.

Cho, S. K., et al. (2008). "Regulation of floral organ abscission in Arabidopsis thaliana." Proceedings of the National Academy of Sciences **105**(40): 15629-15634.

Dharmasiri, S., et al. (2003). "The RUB/Nedd8 conjugation pathway is required for early development in Arabidopsis." The EMBO journal**22**(8): 1762-1770.

vanEck, H.J., et al. "Fine mapping of the *Ro*-locus involved in tuber shape on potato chromosome 10" Wageningen university and research centre, plant breeding

Fiore, A., et al. (2012) "A quadruple mutant of Arabidopsis reveals a β-carotene hydroxylation activity for LUT1/CYP97C1 and a regulatory role of xanthophylls on determination of the PSI/PSII ratio." BMC Plant biology**12**(50)

Gepts, P. (2014). "The contribution of genetic and genomic approaches to plant domestication studies." Curr. Op. in Plant Biology**18**:51-59.

Gouy, M., et al. (2010) "SeaView version 4 : a multiplatform graphical user interface for sequence alignment and phylogenetic tree building." Molecular Biology and Evolution **27**(2):221-224.

Hua, Z. and R. D. Vierstra (2011). "The cullin-RING ubiquitin-proteinligases." Annual review of plant biology **62**: 299-334.

Hufford, M. B., et al. (2012). "Comparative population genomics of maize domestication and improvement." Nature genetics**44**(7): 808-811.

Jensen, J. D., et al. (2005). "Distinguishing between selective sweeps and demography using DNA polymorphism data." Genetics **170**(3): 1401-1410.

deJong, H., et al. (2007). "Association mapping of leaf traits, flowering time, and phytate content in Brassica rapa." Genome**50**(10): 963-973.

Knetsch, M. L., et al. (1996). "Abscisic acid induces mitogen-activated protein kinase activation in barley aleurone protoplasts." The Plant Cell Online**8**(6): 1061-1067.

Lagercrantz, U. (1998). "Comparative mapping between Arabidopsis thaliana and Brassica nigra indicates that Brassica genomes have evolved through extensive genome replication accompanied by chromosome fusions and frequent rearrangements." Genetics**150**(3): 1217-1228.

Lenser, T. and G. Theißen (2013). "Molecular mechanisms involved in convergent crop domestication." Trends in plant science**18**(12): 704-714.

Leung, J., et al. (1994). "Arabidopsis ABA response gene ABI1: features of a calcium-modulated protein phosphatase." Science**264**(5164): 1448-1452.

Li, H. and R. Durbin (2009). "Fast and accurate short read alignment with Burrows–Wheeler transform." Bioinformatics**25**(14): 1754-1760.

Li, L. C., et al. (2007). "Hormonal regulation of leaf morphogenesis in Arabidopsis." Journal of integrative plant biology **49**(1): 75-80.

Lin, K., et al. (2014). "Beyond genomic variation-comparison and functional annotation of three Brassica rapa genomes: a turnip, a rapid cycling and a Chinese cabbage." Bmc Genomics **15**(1): 250.

Liu, S., et al. (2014). "The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes." Nature comm.1:11.

Lysak, M. A., et al. (2005). "Chromosome triplication found across the tribe Brassiceae." Genome research**15**(4): 516-525.

McKenna, A., et al. (2010). "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data." Genome research**20**(9): 1297-1303.

McVean, G. (2007). "The structure of linkage disequilibrium around a selective sweep." Genetics**175**(3): 1395-1406.

Meyer, K., et al. (1994). "A protein phosphatase 2C involved in ABA signal transduction in Arabidopsis thaliana." Science**264**(5164): 1452-1455.

Nagaharu, U. (1935). "Genome analysis in Brassica with special reference to the experimental formation of B. napus and peculiar mode of fertilization." Jap J Bot **7**: 389-452.

Nakasone, A., et al. (2012). "SMALL ACIDIC PROTEIN1 acts with RUB modification components, the COP9 signalosome, and AXR1 to regulate growth and development of Arabidopsis." Plant physiology**160**(1): 93-105.

Nemoto, K., et al. (2011). "Autophosphorylation profiling of Arabidopsis protein kinases using the cell-free system." Phytochemistry **72**(10): 1136-1144.

Nielsen, R. (2005). "Molecular signatures of natural selection." Annu. Rev. Genet.**39**: 197-218.

Nielsen, R., et al. (2005). "Genomic scans for selective sweeps using SNP data." Genome research **15**:1566-1575.

Nimmakayala, P., et al. (2014). "Single nucleotide polymorphisms generated by genotyping by sequencing to characterize genome-wide diversity, linkage disequilibrium, and selective sweeps in cultivated watermelon." BMC genomics**15**(1): 767.

Østergaard, L. and G. J. King (2008). "Standardized gene nomenclature for the Brassica genus." Plant Methods**4**(10): 1-4.

Parkin, I. A., et al. (2014). "Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid Brassica oleracea." Genome biology**15**(6): R77.

Pavlidis, P., et al. (2010). "Searching for footprints of positive selection in whole-genome SNP data from nonequilibrium populations." Genetics **185**(3): 907-922.

Pavlidis, P., et al. (2013) "SweeD: Likelihood-based detection of selective sweeps in thousands of genomes." MolBiolEvol.

Qi, J., et al. (2013). "A genomic variation map provides insights into the genetic basis of cucumber domestication and diversity." Nature genetics.

Reed, J. W. (2001). "Roles and activities of Aux/IAA proteins in Arabidopsis." Trends in plant science**6**(9): 420-425.

Retief, J. D. (1999). "Phylogenetic analysis using PHYLIP."Bioinformatics methods and protocols, Springer: 243-258.

Ross-Ibarra, J., et al. (2007). "Plant domestication, a unique opportunity to identify the genetic basis of adaptation." Proceedings of the National Academy of Sciences**104**(suppl 1): 8641-8648.

Schaffner, S. and P. Sabeti (2008). "Evolutionary adaptation in the human lineage." Nature Education**1**(1)

Schwechheimer, C. and X.-W. Deng (2001). "COP9 signalosome revisited: a novel mediator of protein degradation." Trends in cell biology**11**(10): 420-426.

Schweighofer, A., et al. (2007). "The PP2C-type phosphatase AP2C1, which negatively regulates MPK4 and MPK6, modulates innate immunity, jasmonic acid, and ethylene levels in Arabidopsis." The Plant Cell Online**19**(7): 2213-2224.

Purcell, S. and Chang C. PLINK 1.9 https://www.cog-genomics.org/plink2Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience, 4.

Tian, L. and D. DellaPenna (2001) "Characteriztion of second carotenoid β-hydrocylase gene from Arabidopsis and its relationship to the LUT1 locus."Plant molecular biology**47**(3):379-388

Town, C. D., et al. (2006). "Comparative genomics of Brassica oleracea and Arabidopsis thaliana reveal gene loss, fragmentation, and dispersal after polyploidy." The Plant Cell Online**18**(6): 1348-1359.

Wang, X., et al. (2011). "The genome of the mesopolyploid crop species Brassica rapa." Nature genetics**43**(10): 1035-1039.

Wei, N., et al. (2008). "The COP9 signalosome: more than a protease." Trends in biochemical sciences **33**(12): 592-600.

Wu, G., et al. (2008). "KANADI1 regulates adaxial–abaxial polarity in Arabidopsis by directly repressing the transcription of ASYMMETRIC LEAVES2." Proceedings of the National Academy of Sciences**105**(42): 16392-16397.

Xu, M., et al. (2010). "Arabidopsis BLADE-ON-PETIOLE1 and 2 promote floral meristem fate and determinacy in a previously undefined pathway targeting APETALA1 and AGAMOUS-LIKE24." The Plant Journal**63**(6): 974-989.

Yanofsky, M. F. (1995). "Floral meristems to floral organs: genes controlling early events in Arabidopsis flower development." Annual review of plant biology **46**(1): 167-188.

deYoung, B. J. and S. E. Clark (2008). "BAM receptors regulate stem cell specification and organ development through complex interactions with CLAVATA signaling." Genetics **180**(2): 895-904.

Yu, J., et al. (2013). "Bolbase: a comprehensive genomics database for *Brassica oleracea*." BMC Genomics**14**:664.

Zhang, H., et al. (1997). "An Arabidopsis gene encoding a putative 14-3-3-interacting protein, caffeic acid/5-hydroxyferulic acid O-methyltransferase." Biochimica et Biophysica Acta**1353**(3):199-202

Zhao, J., et al. (2005). "Genetic relationships within Brassica rapa as inferred from AFLP fingerprints." Theoretical and Applied Genetics**110**(7): 1301-1314.

Appendix A

Table 8. The in broccoli classified potential selective sweep positions and the number of found genes based on synteny with *A. thaliana*, found with SweeD or Fst/pi/Tajima's D as first parameter and overlap with one or several parameters. Also the number of genes with significant GO-terms per position and the LD-decay values for all the potential selective sweeps found with SweeD as first parameter.

1st parameter	Linkage group	Position	Nr. genes	Nr. genes with sign. GO-terms	LD-decay
SweeD	C01	15,738,216..15,787,831	-	-	0.05070391-0.03135919
SweeD	C02	13,608,198..13,896,695	10	-	5.835327-0.010581
SweeD	C02	23,417,076..23,439,944	-	-	3.183654
Fst/pi/TD	C03	24,035,001..24,255,001	16	3	
Fst/pi/TD	C04	14,520,001..14,735,001	13	2	
SweeD	C05	19,126,140..19,145,180	-	-	0.058037-0.018887
		19,162,248..19,182,600	2	-	0.030201
		19,192,448..19,221,332	1	-	0.030201
		19,294,204..19,342,128	1	-	0.018658
		19,372,980..19,407,776	2	-	0.013249
SweeD	C06	2,132,219..2,135,119	1	-	0.012427
SweeD	C07	36,580,440..36,700,920	19	2	0.794552-0.393859
		36,704,176..36,830,352	20	4	0.393859-0.049838
		36,838,496..36,850,704	2	-	0.049838-0.038431
		36,854,776..36,936,996	7	-	0.038431
SweeD	C07	37,107,132..37,332,624	26	-	0.22747-4.746025
		37,334,252..37,525,552	25	4	4.746025-0.007101
SweeD	C08	14,636,648..14,937,201	1	-	0.114058-0.386118
SweeD	C09	6,600,719..6,602,324	-	-	0.015955
		6,611,150..6,613,558	-	-	0.015955
SweeD	C09	26,109,856..26,205,344	1	-	0.001469-0.012349

Table 9. The in cauliflower classified potential selective sweep positions and the number of found genes based on synteny with *A. thaliana*, found with SweeD or Fst/pi/Tajima's D as first parameter and overlap with one or several parameters. Also the number of genes with significant GO-terms per position and the LD-decay values for all the potential selective sweeps found with SweeD as first parameter.

1st parameter	Linkage group	Position	Nr. genes	Nr. genes with sign. GO-terms	LD-decay
SweeD	C03	17,328,032..17,332,654	1	-	0.963159
SweeD	C04	16,541,894..16,686,664	4	2	0.075082-1.296956
SweeD	C06	3,403,694..3,404,660	-	-	0.015776
SweeD	C06	18,505,001..18,775,001	20	2	0.001033-0.001897
SweeD	C06	20,234,542..20,313,828	3	1	1.680974-0.011977
SweeD	C07	11,750,001..11,890,001	5	1	0.003096-0.004808
SweeD	C08	32,766,212..32,781,156	3	-	0.030307
		32,817,688..32,820,176	-	-	0.030307
		32,853,388..32,863,352	2	-	0.074119
SweeD	C09	10,573,760..10,601,845	-	-	1.933291
		10,627,523..10,667,644	1	-	1.933291-4.998036
		10,670,052..10,695,730	1	-	4.998036
		10,706,161..10,795,231	-	-	4.998036-1.227434
Fst/pi/TD	C09	25,340,001..25,455,001	7	2	

Table 10. The in enlarged inflorescence (cauliflower and broccoli) classified potential selective sweep positions and the number of found genes based on synteny with *A. thaliana*, found with SweeD or Fst/pi/Tajima's D as first parameter and overlap with one or several parameters. Also the number of genes with significant GO-terms per position and the LD-decay values for all the potential selective sweeps found with SweeD as first parameter.

1st parameter	Linkage group	Position	Nr. genes	Nr. genes with sign. GO-terms	LD-decay
SweeD	C01	18,619,780..18,624,430	-	-	0.01770803
Pi/TD	C01	19,850,000..20,400,000	6	1	

SweeD	C02	13,436,600..13,486,778	5	1	3.590240-3.058549
Fst/pi/TD	C03	6,890,001..7,240,001	27	5	
Fst/pi/TD	C05	4,310,001..4,700,001	16	7	
Fst/pi/TD	C06	26,385,001..26,535,001	9	3	
Fst/pi/TD	C07	26,870,001..27,065,001	5	1	
Fst/pi/TD	C08	3,020,001..4,495,001	21	1	
SweeD/pi	C09	24,095,001..24,625,001	13	2	0.023285-0.018804

Table 11. The in cabbage classified potential selective sweep positions and the number of found genes based on synteny with *A. thaliana*, found with SweeD or Fst/pi/Tajima's D as first parameter and overlap with one or several parameters. Also the number of genes with significant GO-terms per position and the LD-decay values for all the potential selective sweeps found with SweeD as first parameter.

1st parameter	Linkage group	Position	Nr. genes	Nr. genes with sign. GO-terms	LD-decay
SweeD	C01	5,206,617..5,217,470	1	-	5.821979
SweeD	C02	43,607,580..43,658,620	2	1	0.036025-0.01815
SweeD	C03	3,708,693..3,711,004	1	-	0.054333
SweeD	C04	12,795,879..12,796,697	-	-	0.038234
Fst/pi/TD	C04	13,440,001..13,545,001	3	1	
Fst/pi/TD	C04	17,170,001..17,455,001	8	2	
SweeD	C05	11,133,467..11,184,673	1	-	1.515877-1.944451
		11,310,718..11,339,603	-	-	1.897034
		11,392,779..11,498,473	6	-	0.203434-1.691728
		11,528,015..11,556,901	1	-	1.691728-3.392311
SweeD	C05	19,989,476..19,992,102	-	-	0.123894
SweeD	C06	28,446,512..28,452,314	-	-	0.029914
SweeD	C07	19,370,996..19,384,834	-	-	0.610609
Fst/pi/TD	C07	23,790,001..24,335,001	33	5	
SweeD	C08	21,339,312..21,344,292	-	-	0.017935
SweeD	C08	23,215,692..23,217,352	1	1	0.185895
Fst	C08	11,080,001..11,970,001	29	10	
SweeD	C09	24,249,650..24,250,452	-	-	0.008943
SweeD	C09	26,112,838..26,161,786	1	-	3.65757-0.023132

Table 12. The in kohlrabi classified potential selective sweep positions and the number of found genes based on synteny with *A. thaliana*, found with SweeD or Fst/pi/Tajima's D as first parameter and overlap with one or several parameters. Also the number of genes with significant GO-terms per position and the LD-decay values for all the potential selective sweeps found with SweeD as first parameter.

1st parameter	Linkage group	Position	Nr. genes	Nr. genes with sign. GO-terms	LD-decay
SweeD	C01	34,969,560..35,087,384	6	1	0.013819-0.085321
Fst/pi/TD	C02	35,900,001..36,125,001	7	1	
SweeD	C03	28,365,556..28,380,580	1	1	0.009529
SweeD	C04	9,390,930..9,422,828	1	-	1.224823
Fst/pi/TD	C04	13,565,001..13,705,001	3	1	
SweeD	C06	6,411,618..6,499,606	2	-	0.012122-0.012605
		6,516,043..6,546,017	4	-	0.012605-0.004956
		6,565,356..6,661,079	8	-	0.004956-0.003425
SweeD	C06	14,809,175..15,511,147	24	4	0.005273-0.016916
		15,540,154..15,568,194	3	2	0.016916-0.01476
		15,573,996..15,622,341	2	-	0.01476
		15,632,010..15,664,885	2	-	0.01476-0.040171
SweeD	C06	39,591,876..39,610,248	-	-	0.001744
SweeD	C07	13,945,242..13,990,830	-	-	0.450222
		14,025,835..14,083,634	2	-	0.450222-0.012756
SweeD	C07	14,369,371..14,453,220	-	-	1.858985-0.055302
SweeD	C07	14,620,104..15,045,046	15	3	0.522332-0.010593
SweeD	C07	19,589,166..19,590,794	-	-	0.035542
SweeD	C07	20,902,254..20,918,536	-	-	0.301256

SweeD	C07	26,962,980..26,967,050	-	-	0.016915
SweeD	C08	4,845,990..5,161,438	2	-	0.007135-0.018344
SweeD	C08	6,746,148..6,878,138	2	1	0.036087-0.704571
		6,906,362..7,127,176	6	1	0.704571-1.17797
		7,130,496..7,162,041	1	-	1.17797-0.321336
		7,210,188..7,667,587	10	-	0.321336-0.002832
SweeD	C09	13,839,204..13,876,923	1	1	0.014647-0.016878
		13,892,974..13,917,853	-	-	0.016878
SweeD	C09	18,005,214..18,026,884	-	-	0.005589
SweeD	C09	22,307,650..22,323,710	-	-	0.020329

Table 13. The in undecided (kohlrabi or cabbage) classified potential selective sweep positions and the number of found genes based on synteny with *A. thaliana*, found with SweeD or Fst/pi/Tajima's D as first parameter and overlap with one or several parameters. Also the number of genes with significant GO-terms per position and the LD-decay values for all the potential selective sweeps found with SweeD as first parameter.

1st parameter	Linkage group	Position	Nr. genes	Nr. genes with sign. GO-terms	LD-decay
Fst	C01	24,945,001..25,090,001	4	-	
Pi/TD	C01	23,550,000..24,300,000	1	-	
Fst/pi/TD	C02	20,300,001..20,615,001	12	-	
Fst/pi/TD	C05	7,135,001..8,825,001	22	5	
Fst	C05	12,235,001..12,305,001	6	-	
Fst/pi/TD	C06	34,120,001..34,320,001	10	1	
Fst/pi/TD	C08	14,680,001..14,835,001	-	-	
Fst/pi/TD	C09	21,965,001..22,285,001	10	3	
Fst/pi/TD	C09	28,275,001..28,395,001	4	-	

Appendix B

See supplementary information for LD-decay 1 and 2

Appendix C

See supplementary information for SweeD 1 and 2

Appendix D

See supplementary information for Fixation index 1 and 2

Appendix E

See supplementary information for Variable sites 1 and 2

Appendix F

See supplementary information for Tajima's D 1 and 2

Appendix G

Table 14. The genes with enriched GO-terms for broccoli and the known genome fractionation according to Liu *et al.* (2014).

Chr.	B. oleracea gene	Genome fractionation
3	Bol026625	
	Bol026610	

	Bol026606	MF2
4	Bol014401	MF1
	Bol014404	MF1
7	Bol017122	
	Bol017120	LF
7	Bol017089	
	Bol017085	LF
	Bol017082	LF
	Bol017080	LF
7	Bol017009	LF
	Bol017004	LF
	Bol017003	LF
	Bol016985	

Table 15. The genes with enriched GO-terms for cauliflower and the known genome fractionation according to Liu *et al.* (2014).

Chr.	B. oleracea gene	Genome fractionation
4	Bol017589	
	Bol017590	
6	Bol013366	
	Bol013347	MF2
6	Bol039034	LF
7	Bol041437	LF
9	Bol012201	LF
	Bol012196	LF

Table 16. The genes with enriched GO-terms for enlarged inflorescence (cauliflower and broccoli) and the known genome fractionation according to Liu *et al.* (2014).

Chr.	B. oleracea gene	
1	Bol036509	
2	Bol028576	MF1
3	Bol027976	
	Bol027988	MF2
	Bol027998	MF2
	Bol028004	MF2
	Bol028008	MF2
5	Bol037987	LF
	Bol037990	LF
	Bol037991	LF
	Bol037992	LF
	Bol037993	LF
	Bol037994	
	Bol037999	LF
6	Bol007570	MF2
	Bol007571	
	Bol007573	MF2
7	Bol006260	LF
8	Bol011630	
9	Bol009898	LF
	Bol009914	LF

Table 17. The genes with enriched GO-terms for cabbage and the known genome fractionation according to Liu *et al.* (2014).

Chr.	B. oleracea gene	Genome fractionation
2	Bol016445	MF1
4	Bol038612	LF
4	Bol017649	
	Bol017650	

7	Bol012321	
	Bol012326	MF1
	Bol012335	MF1
	Bol005606	LF
	Bol005608	LF
8	Bol007420	
8	Bol027056	
	Bol027062	
	Bol027063	MF2
	Bol027064	MF2
	Bol027071	
	Bol027072	MF2
	Bol027074	MF2
	Bol027079	
	Bol027080	MF2
	Bol027090	MF2

Table 18. The genes with enriched GO-terms for kohlrabi and the known genome fractionation according to Liu *et al.* (2014).

Chr.	B. oleracea gene	Genome fractionation
1	Bol005945	MF1
2	Bol020091	
3	Bol042656	
4	Bol038619	LF
6	Bol015697	LF
	Bol015680	LF
	Bol015674	LF
	Bol015668	LF
6	Bol005408	
	Bol005407	
7	Bol031884	LF
	Bol031887	LF
	Bol031888	
8	Bol033426	
8	Bol033424	
9	Bol007733	

Table 19. The genes with enriched GO-terms for 'undecided' (cabbage and kohlrabi) and the known genome fractionation according to Liu *et al.* (2014).

Chr.	B. oleracea gene	
5	Bol031959	MF1
	Bol031960	
	Bol031975	
	Bol031982	MF1
	Bol031987	MF1
6	Bol040075	MF2
9	Bol038813	LF
	Bol038831	
	Bol038832	