Creating a methodology to map genetic variation using pangenomics to explore temperature sensitivity in Cauliflower.

Jeroen Persoon

GitHub URL:
`https://github.com/jeroenpersoon/Cauliflower_restart`.

Abstract

Farmers experience difficulties in cultivating cauliflower (*Brassica oleracea* var. *botrytis*) due to uneven ripening of the curd. To tackle this problem, a better understanding of the relation between temperature sensitivity and flowering in cauliflower is needed. Several studies have shown that a pangenomic approach can be beneficial over a reference genome-based approach. Using a reference genome does not include the genetic variation that is present in varieties or species. Therefore, in this study, a methodology is created to map the genetic variation within a pangenome to unravel temperature sensitivity in Cauliflower.

To do this, PanTools was used to build a pangenome on multiple evolutionary levels after the found genome assemblies were assessed for their quality. Before analyzing the pangenome, a genome-wide variation study was conducted. To map the variation in the pangenome, a Python script was made that searches all potential Genes of interest (GOI) in the pangenome based on BLAST and homology grouping. This created a variation table where, among other information, gene ID, genome ID, and closest BLAST hit in *Arabidopsis thaliana* are shown. This gives an overview of the CNV of the genes in all the assemblies used in the pangenome. In addition to genome assemblies, resequencing data were used to map the genetic variation between those accessions. The accessions were mapped to the cauliflower genomes to see how the reads were mapped. This was done with a genome browser on a web portal that was built for this research. A use case, the variation between the genomes, and the resequencing data for the genes FLC and ELF3. These genes play a role in the flowering of cauliflower, and their expression is influenced by temperature.

With the methodology demonstrated in this research, genetic variation like CNV or SNPs within a species can be mapped. However, results should always be evaluated and reconsidered since many aspects can influence these results, and calling biological variation is not straightforward. The soft-clipping of the reads of the resequencing accessions mapped to the ELF3 gene is an example of why a pangenomic approach will give better results than using a reference genome.

Keywords: Cauliflower, Pangenomics, Genetic diversity, temperature-sensitivity, harvest control

Introduction

Cauliflower (*Brassica oleracea* var.*botrytis*) is an important crop for the diet of many people worldwide. The countries with the highest cauliflower/broccoli production are China, India, the United States, and Mexico. This list is followed by several European countries, which suggests that cauliflower is globally important (FAO 2024). In addition, cauliflower is important for a healthy human diet as it contains large amounts of bioactive compounds, glucosinolates, vitamins, and phenolic compounds (Picchi, Fibiani, and Lo Scalzo 2020). The combination of the nutritional value of cauliflower and the way it is widely integrated in diets makes it important to safeguard its future cultivation.

The cauliflower cultivation is highly affected by temperature and its fluctuations. Temperature is one of the biggest factors that determines the moment of harvest maturity for cauliflower. Formation of the curd is initiated at the moment of switching from the vegetative stage to the generative stage. Since the edible part of the cauliflower is the curd, consisting of arrested inflorescence meristem, curd maturation is an important process. Previous experiments have shown that harvest maturity is typically delayed when cauliflower is exposed to a period of high temperatures before the onset of the curd. However, temperature had an opposite effect on harvest maturity after the curd induction had taken place. In that case, lower temperatures would delay the moment when the cauliflower curd was ready to harvest. This suggests that the temperature sensitivity regula-

tion in cauliflower is rather complex and hard to easily map out. This challenges the planning of the harvest of cauliflower plants. In addition, not all cauliflower cultivars react the same to temperature fluctuations during cultivation. The difference in temperature sensitivity between different cultivars could be because of differences on the genomic level. That is why, in this research, a methodology will be created to analyze the variation within cauliflowers and other members of the *Brassica* family.

To investigate the variation of loci related to temperature sensitivity in cauliflower, genotypic information can be compared. Until this point, much research on plants was done based on a single reference genome. Nevertheless, a single reference genome is unable to capture the complete genetic diversity of a variety or species. That is why, for this research, a pangenomic approach is chosen. Unlike a reference genome, a pangenome includes the complete variation within individual genomes. With a pangenomic study, it is possible to analyze the variation within a species, for example, cauliflower (Tettelin et al. 2005).

To get insight into the temperature sensitivity of cauliflower, an understanding of the flowering mechanism is needed. Flowering in cauliflower is a complex mechanism that includes many genes and the way they are regulated. For example, the FLC genes are suggested to have a big impact on flowering. FLC genes are known to inhibit flowering in the model plant species, *Arabidopsis thaliana* (AT), by binding to the genes FT and SOC1 (P. Li et al. 2014). FLC is repressed by VIN3, VRN1, and VRN2, which are expressed after the plant undergoes vernalization (Kim and Sung 2013)(Gendall et al. 2001). In cauliflower, there are multiple copies of FLC, of which, for example, BoFLC2 was investigated before (Ridge et al. 2015). Ridge et al. (2015) found that plants which got an early stop codon due to a deletion in the BoFLC2 gene flowered days earlier than plants from which the BoFLC2 gene was intact. Research in AT also shows that VIN3 influences the FLC genes. VIN3 is one of the vernalization genes whose expression is determined by temperature. Due to the function of VIN3 in AT, this gene might be interesting to study in cauliflower (Kim and Sung 2013).

Genetic variation like copy number variation (CNV) in, for example, flowering genes can cause different flowering phenotypes within a species (Schiessl et al. 2017). Scheissl et al. (2017) linked CNV in flowering genes to differences in flowering in *Brassica napus* morphotypes, which also share a subgenome with *Brassica oleracea*. When comparing *Arabidopsis* to *Brassica* species, they have an overlapping evolutionary history. *Brassica* arose from a triplication of a shared ancestral species between *Arabidopsis* and *Brassica* species. The triplication of the common ancestor explains why *Brassica* species have multiple copies of some orthologs compared to *Arabidopsis* species (Akter et al. 2021). This information is crucial when comparing the flowering phenotype and genetic variation of *Brassica* species.

With the pangenomic approach in this research, variations like presence-absence variation (PAV) and CNV can tell something about the genetic diversity within *Brassica* species. In addition, the variation in genes of interest (GOI), like BoFLC2 and VIN3, can be mapped based on homology grouping. Next to the publicly available genome assemblies for constructing a pangenome at different evolutionary levels, resequencing accessions were used. This resequencing data consists of cauliflower cultivars with different tolerance to temperature. In this way, the variation of GOIs can, in the end, be linked to a phenotype. The results of this study are shared with the consortium in the form of a web portal. On this web portal, an overview of the data used is visible. In addition, a JBrowse instance and a blast service are added.

This report is about finding a methodology to analyze the genetic variation across multiple cauliflower genotypes. This starts by making an overview of all the data that is publicly available to construct the pangenome. Before doing this, the quality of the data needs to be assured by doing quality control and filtering low-quality data. Then the question arises: how to use this data to construct a pangenome, or what is the genome-wide variation between? After the pangenome is constructed, the genetic variation is analyzed by mapping the PAV and CNV of the GOIs. This step includes homology grouping, which answers questions like: how are GOIs distributed over homology groups, or: which genes are grouped with certain GOIs? By setting up a web portal, resequencing accessions could be analyzed. Read mapping and variant calling were done for the resequencing data to map the genetic variation in GOIs, and this could be shown via the web portal. As a conclusion, the strengths and weaknesses of this research, and recommendations for further analysis are given.

Materials and methods

This section describes what activities were done to find a methodology to map the variation between 157 cauliflower genotypes. First, the way of collecting and testing the quality of the data is explained. Secondly, the construction of the pangenome and the web portal to share the results with the consortium. Finally, the way of analyzing the genetic variation genome-wide and

for GOIs specifically is explained.

Genome assemblies

The genome assemblies, including their annotations, were searched in publicly available databases, like: NCBI, The Genome Warehouse, Figshare, etc. This resulted in an overview of all publicly available genome assemblies, including annotations for *Brassica species*. The overview is given as a supplementary Excel sheet Brassica_availability.

Quality control and selection

To determine which data needs to be part of the panproteomes and pangenomes, on which the genetic variation is analyzed, the quality of the data was ensured. First, the PanUtils pipeline, available on https://github.com/PanUtils/pantools-qc-pipeline, was run on the genome assemblies, producing multiple statistics, such as genome size, the number of genes, and the average gene length. The PanUtils pipeline was also used to set the minimum genomic FASTA sequence to 10.000 nucleotides, only keep the longest isoform of a gene when alternative splicing was taken into account, and set the minimum length of a peptide to 49.

After running the PanUtils pipeline, a first selection was made on which a BUSCO (Manni et al. 2021) assessment was run. Assemblies or annotations were filtered out in this first selection, for example, because of a low average gene length or because other assemblies were more relevant to represent a variety. To run the BUSCO assessment, a first panproteome was built with all the genome assemblies remaining, with the purpose of only doing the BUSCO assessment. The panproteome was constructed using PanTools v.4.3.3 (Sheikhizadeh et al. 2016). PanTools has a built-in function to run the BUSCO assessment. Running BUSCO checks where highly conserved orthologs are present in the genome assembly; if not, this is an indication of a low-quality assembly. The set of highly conserved orthologs that was used was named: brassicales_odb10. After running the BUSCO assessment, a second selection was made, filtering out all the genome assemblies that had a missing BUSCO percentage of 5.5.

Finally, genome assemblies were selected, in a third selection round, based on their relevance to the research objective. Including every assembly would make the pangenomes too big in terms of storage and the run times of analyses. Relevance for the species- and genus-level pangenomes means an even distribution of every variety/species available, and as many different crops as possible. For the pangenome containing only cauliflower, the genome names are: Korso (T22), T21, T25, and 'Cauliflower'. In the Supplementary file, Brassica_availability, there is a list of all the assemblies that were used per pangenome.

Pangenome/panproteome construction

The panproteomes and pangenomes were constructed using PanTools v.4.3.3 (Sheikhizadeh et al. 2016). Panproteomes were built using the build_panproteome function, which takes a few seconds and enables doing the first analyses more quickly. Not all PanTools' functions are available for panproteomes, so in a later stage, the build_pangenome function was used to construct pangenomes. Add_annotation is a function that can only be done on pangenomes and enables adding the annotation files to the pangenome. Before analyzing the GOIs, they need to be distributed over homology groups. One of the challenges for homology grouping is to find the best relaxation. This determines how similar sequences need to be to end up in the same group. PanTools can run optimal_grouping and will determine the best relaxation for a pangenome based on the number of groups containing only single copy orthologs. For the pangenome that contains only cauliflower assemblies, the relaxation level that resulted in the most single-copy orthologous groups, while the recall and precision were both above 0.97, was chosen. In the end, the relaxation for this pangenome was set to D2 while running the grouping function of PanTools. The grouping function produces a large text file containing all the homology groups and the gene IDs in them, which was later used to analyze the distribution of the GOIs over different homology groups. Supplementary files pantools_commands.sh and pantools_grouping.sh give an overview of the PanTools commands that were used and can be used on one's own data as well.

For the species pangenome, the optimal grouping for relaxation 2 resulted in 15791 single-copy orthologous groups with a recall of 0.979 and a precision of 0.977, so D2 was used as relaxation. Another challenge is to align this relaxation over the pangenome for each evolutionary level. This needs to avoid the fact that, for example, different versions of a GOI are grouped in the *Brassica oleracea* pangenome and not in the cauliflower pangenome. This is done by the busco-validation Python script. This script checks for 100 random BUSCOs in which group they are. Then make a large table listing the homology group numbers for every BUSCO in every genome in the panproteome/pangenome. The last column contains True/False to quickly check where all the BUSCOs are in the same group. If this is not the case, GOIs might also not be grouped together when they should like the BUSCOs should also group together.

WGS data

Part of this research is the incorporation of resequencing

data. In total, 157 resequencing accessions were grown in the field, of which 4 were used for analysis as a proof of principle for analyzing all resequencing accessions. The resequencing accessions consist of a combination of real biological hybrids and in-silico hybrids. The four accessions that were used as proof of principle were named: liria, caniego, TKI-0143, and TKI-0155. In the end, all 157 resequencing accessions will be used to link the desired phenotype to specific genetic variation. A quality control was also done on the resequencing data using MultiQC v1.31 (Ewels et al. 2016). The report is given as a Supplementary file. The MultiQC report was analyzed, and based on the quality of the reads, trimming the reads was considered.

Web portal

At this point, a web portal was set up for conducting analysis and, in the end, sharing the results of this research with the consortium. An in-house web portal template from Wageningen University & Research (WUR) was used and modified. The web portal includes a JBrowse instance to browse over the cauliflower genome and the resequencing accessions. In addition, a BLAST service is included to blast against the used genomes to analyze our GOIs. The web portal is available at localhost:8084 after running: ssh -L 8084:thornton.bioinformatics.nl:8084 myers in the command prompt (for WUR colleagues with access to the servers only)

Read mapping and variant calling

Within this research, the Korso genome was used as a reference genome to compare read mapping to a pangenomic approach and to look at GOIs in the genome browser. First, the reads of all 4 resequencing accessions were mapped to the Korso genome using BWA v.0.7.19(H. Li and Durbin 2010). The resulting SAM files were converted, sorted, and indexed using SAMtools v.1.22.1 (Danecek et al. 2021). Variant calling was done using FreeBayes v.1.3.10 (Garrison and Marth 2012). FreeBayes has an option to use filters for the input base and mapping quality, named –standard_filters, that was used. To analyze the reads and the variant calling, the BAM and VCF files were uploaded to the JBrowse instance in the web portal.

Genome-wide variation

To get a picture of the sequence variation between the species, Mash v.2.3 (Ondov et al. 2016) was run. Mash reduces sequences into smaller sketches from which the distance between sequences can be measured through the fraction of shared k-mers. For all the assemblies passed after the third selection in the QC, one sketch was made. Next, the distance between this sketch with itself was calculated to get the distance between all the genome assemblies. The resulting TSV file was used as input to make a heatmap using RStudio v.2023.06.1. The statistics produced by the PanUtils pipeline were used to make plots in Excel. The total sequence size and total number of genes were plotted in a bar chart to see if there is any variation in these numbers at the species and genus level. To compare the proportion of genes shared among cauliflowers and other *Brassica oleracea* species, PanTools' gene_classification function was run. This function calculates which genes are shared among all assemblies, part of the assemblies, or are unique to each assembly. The output file classified_groups.csv was used in an RStudio script to make an upset plot. The gene_classification function and the RStudio script for making the upset plot were run on the panproteomes at the variety level (containing only cauliflowers) and on the species level (containing only *Brassica oleracea* species).

Variation in Genes of interest

Based on homology grouping, we could see how GOIs grouped together and with which other genes they share a group. However, this is easier said than done, since the grouping is influenced when genes have different copy numbers, different versions, other genes that have similar functions, or similar genes that are not even characterized precisely. This means that GOIs are not always nicely placed in one group with only their orthologs. In addition, the genes found in the homology groups might not be well annotated to know what it is. In most annotations used, the gene name and any functional annotations are absent. Because of this, there is a need to analyze the found homology groups and additional genes to get a full overview of all the candidate GOI. That is why the genes found are linked to the closest gene in *Arabidopsis thaliana* to get an idea about what the found genes are. To get a total overview of any gene and see its PAV, CNV, and which other genes share the same homology group, a Python script was made using Python 3.10.18. This script takes a protein sequence as input and does a BLASTP against the pangenome specified. The BLASTP is run using PanTools' BLAST function. The minimum sequence identity percentage and the alignment length percentage need to be set by the user when running the script. In this research, the minimum sequence identity was set at 25% and the alignment length was set to 65%. The first BLAST hit from this first BLAST is used to do another BLASTP against the specified pangenome. This results in two lists containing potential GOIs, from which the longest list will be used in the rest of the script. A TSV file is made from those potential GOIs, further called variation table, containing the gen ID, homology group number,

genome number, BLAST location, closest gene in AT, the ID% and name of this AT hit, Best hit in the Korso (T22) genome, ID% of the best Korso hit, YES/NO if the BLAST hit was found using the initial query or using the best BLASTP hit from the pangenome, and from which species the potential GOI is. To get the closest gene in AT, a BLASTP was done outside of PanTools using BLAST v.2.17.0+ (Camacho et al. 2009). For the genes that are in the same homology group as any gene that passed the BLASTP, but did not pass the BLASTP itself, a separate variation table is made with the same information. Extending the homology grouping with a BLASTP is chosen to validate that all GOI in the pangenome are found. Without doing a BLAST, it is unable to know beforehand how GOI are grouped, and the risk of missing groups of interest is too big. In this research, the sequences for the GOI to use as input for this Python script are retrieved from NCBI by downloading the protein sequence of the GOI in AT.

To validate and compare CNV and the location of the found *Brassica* genes in the variation tables with the literature, papers were used, or the BRAD database (Chen et al. 2022) was used. The BRAD database contains data from several *Brassica* species. In addition, the BRAD database has a BLAST tool to BLAST GOI on different genomes. In this case, a TBLASTN was done using the protein sequence of the GOI retrieved from NCBI to BLAST against the only two *Brassica oleracea* genomes on the BRAD database, HDEM and JZS V2.0. A TBLASTN was done by uploading the protein sequence to BRAD, and the HDEM and JZS V2.0 CDS were used as the nucleotide database.

Results

In this section, we describe and discuss a pangenomic approach to study variation in cauliflower and its relatives. First, we explain the development of the pangenomic infrastructure required to obtain a holistic view of genetic variation. Then we demonstrate how the infrastructure can be applied to analyze genome-wide variation and to zoom in on genes of interest (GOIs). Finally, we discuss the advantages and disadvantages of a pangenomic approach.

Pangenomic infrastructure

To analyze the genetic variation across 157 cauliflower genotypes using pangenomics, a pangenomic infrastructure had to be developed. This includes everything from data collection to tools that are used to share the results with the consortium. This pangenomic infrastructure forms the base from which genetic variation could be analyzed.

The research started with mapping the availability of *Brassica* genome assemblies, which are publicly available. Table 1 below shows the number of genome assemblies that were found, passed the QC pipeline and BUSCO test, and were selected to construct the pangenome. In total, 78 *Brassica* assemblies, including annotation, were found. Almost all the found genome assemblies were used for the QC pipeline. On the genus level, many genome assemblies not selected for the QC were *Brassica rapa* assemblies. Since this amount would not be relevant for the research, the majority of these assemblies were excluded before running the QC pipeline. After the QC pipeline and BUSCO test, 50 accessions remained of sufficient quality. To avoid storage issues and longer run times for analyses, a maximum of 15 accessions per pangenome was used. Some accessions were used in multiple pangenomes; in the end, 27 different genome assemblies were used in this research.

The selected genome assemblies were first used to construct panproteomes. Constructing panproteomes is a fast way of getting familiar with PanTools and running quick analysis since it only takes a few seconds. Compared to the construction of pangenomes, which took 5 hours for constructing the cauliflower pangenome and took 21 hours for constructing the pangenome at the species level. To the pangenomes, the annotations were added, which only took 5 minutes for the cauliflower pangenome. Constructing a pangenome gave us a collection of all the genome assemblies used. The genomes are collected in a so-called DeBruijn graph, where all the genetic variation between the genomes is incorporated. To this graph, all kinds of nodes can be added, which point to, for example, a gene, a specific phenotype, or a function annotation. Having the pangenome itself is not going to give any answers, so that is why additional scripts were made to map the genetic variation. For example, the script to analyze the homology groups of a GOI to get insight into the CNV and what other similar genes are present in the genomes. In addition to the genome assemblies, four accessions of resequencing data were used. These accessions were used to create use cases regarding genes that play a role in the temperature sensitivity in the flowering mechanism in cauliflower. To analyze the reads for specific regions, the genome browser in the web portal was built. In this way, the variation in these regions between the re-sequencing accessions could be mapped.

Table 1: Number of genome assemblies including annotation, which were collected in the first place, selected after the PanUtils pipeline and the BUSCO assessment, and were selected in the end to construct a pangenome.

	Subspecies *B. oleracea var. botrytis*	Species *B. oleracea*	Genus *Brassica*	Total
Genome assemblies with annotation	5	32	41	78
Passed QC and BUSCO	4	30	16	50
Selected	4	15	15	27

Genome-wide variation

When looking at the evolution of *Brassica*, we expect the biggest genome-wide variation at the genus level. One of the reasons is the allotetraploidization of three *Brassica* species illustrated in Figure S1. Such events did not happen within *Brassica oleracea*, which is why we do not expect to see similar variation at the species level.

With the calculations of the genome size, the number of genes, and the results from the BUSCO test, we could already get more information about the genomic variation comparing the different species across the *Brassica* genus. Figure 1 shows the proportion of duplicated BUSCOs in every sample. The genomes of *B.napus*, *B. carinata*, and *B.juncea* have more duplicated BUSCOs than other *Brassica* species. This is due to the fact that these species contain two subgenomes as a result of allotetraploidization (Yim et al. 2022). This allotetraploidization also doubles many BUSCOs. The allotetraploidization is also the reason for the increased genome size, which can be found in Figure 2. In this Figure, all the *Brassica oleracea* assemblies show more or less the same genome size and the same number of genes. Only at the genus level, these numbers are increased, and more diversification between the assemblies is visible.

However, this does not mean that there is not much diversification between *Brassica oleracea* species or cauliflower genomes. Having roughly the same number of genes does not mean that all these genes have the same sequence or that these genes are in all the genomes. For example, a pangenome made out of nine different *Brassica oleracea* morphotypes contained over 4.8 million SNPs, and 18.7% of the genes were not shared among all the genomes (Golicz et al. 2016). This is an indication that there is still a big diversification within *Brassica oleracea* genotypes. Because of this, we would expect to see variation between our genomes when looking at sequence similarity. Figure 3 is the output of Mash and shows the results on k-mer based similarities for our selected genomes. In this Figure, red means a greater mash distance and dark blue means a lower mash distance, so more sequence similarity. *B. nigra*, *B. juncea*, and *B. rapa* assemblies show the greatest mash distance compared to *B. carinata*, *B. napus*, and *B. oleracea* assemblies indicated by the yellow and red coloring. *B. oleracea* assemblies show more sequence similarity among each other than to *B. carinata* and *B. napus* assemblies indicated by the darker blue coloring, comparing the 17 *B. oleracea* assemblies. In the middle, there is an even darker blue square, which indicates the sequence similarity between cauliflowers. Since there is still a slight difference in dark blue between the cauliflowers, not even the cauliflowers show perfect sequence similarity.

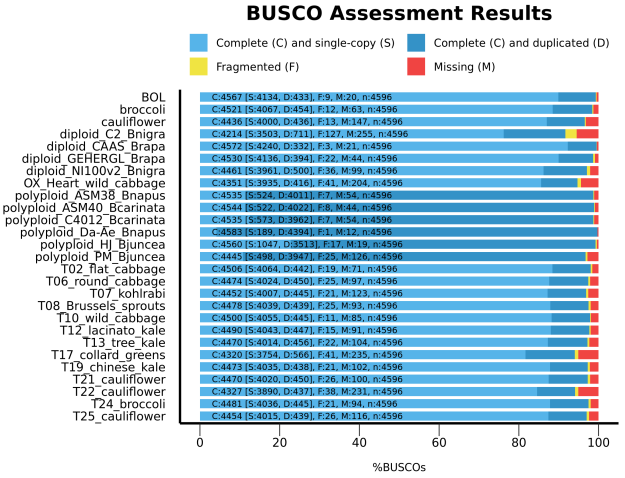


Figure 1: BUSCO results for all the selected genome assemblies. Showing an increased fraction of duplicated BUSCOs for the tetraploid *Brassica* species.

Golicz et al. also found a fraction of 18.7% of the pangenome, which consists of variable genes. In addition to the sequence similarity, we are also interested in the difference in gene repertoire between cauliflowers and when comparing it on the *Brassica oleracea* level. We expect that not all genes will be shared among every cauliflower genotype and that this fraction will be even bigger when comparing it with other *Brassica oleracea* varieties. The core, accessory, and unique parts

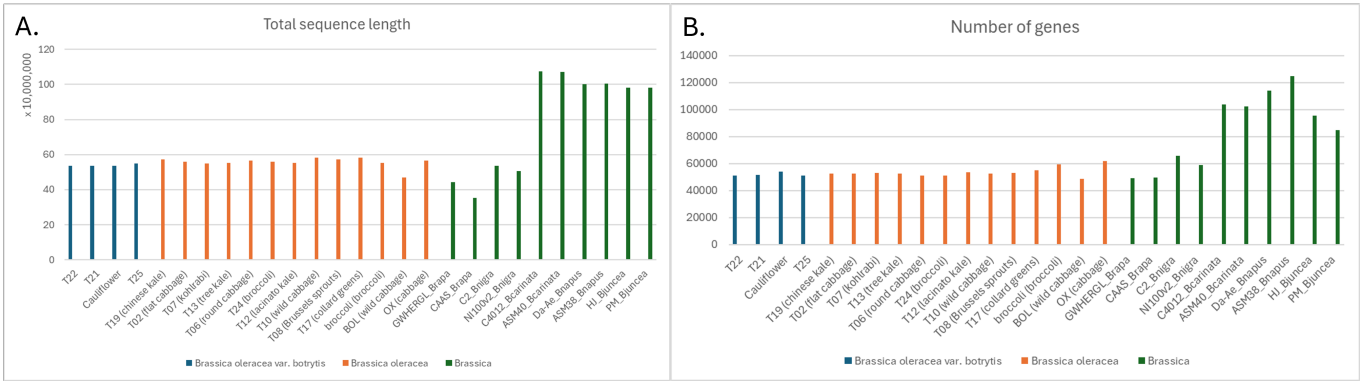


Figure 2: Genome size (A) and number of genes (B) for the genome assemblies used in any pangenome colored at the variety, species, and genus level.

of the genomes are given in Supplementary Figure S2. In Plot A, 28788 genes are shared among all cauliflower genomes. In this plot, Korso has over 9000 more unique genes than the other cauliflower genomes, which is an indication of many poorly annotated genes. When using an updated version of the Korso genome used by (X. Li et al. 2024) (named T22 in their paper), the number of unique genes for Korso decreases to a similar level as the other cauliflower genomes. Doing this increases the number of core genes to 30883. Comparing the core genomes on *Brassica oleracea* level from plot C, we can see 19285 genes are shared among all genomes. This plot also includes OX_Heart and 'broccoli' that have a high number of unique genes. When removing those accessions, the number of core genes is 21352. This means that roughly 9500 genes are part of the core genome for cauliflower but are not shared among all other *Brassica oleracea* varieties. From these plots, we can learn that using poor annotation could lead to many unique genes. We know these unique genes are poorly annotated because using another annotation would decrease the number of unique genes. In addition, the core genome shrank by 9531 genes when comparing the core genome of cauliflowers to the core genome at the *Brassica oleracea* level. This is as expected since from the Mash plot, as in the literature, variation within cauliflowers and *Brassica oleracea* varieties has been shown.

In addition to the genome-wide variation of the genome assemblies, the resequencing data can also show genome-wide variation. In Supplementary figure S3, a position in the genome browser is highlighted showing one SNP in this position. Liria shows a homozygous SNP, Caniego and TKI-0143 show a heterozygous SNP, while TKI-0155 does not show the SNP at all for this position. In table 2, the number of heterozygous SNPs for each accession is shown. Each accession has hundreds of thousands of heterozygous SNPs across the genome. This indicates that within cauliflower, many different loci exist. This increases the expectation that SNPs will also occur in our GOI, which could affect the temperature sensitivity of a cauliflower cultivar.

Variation in Genes of Interest

To make a start, analyzing the variation that could have an effect on temperature sensitivity, we want to map the variation of GOI within the pangenome. For example, we want to know the CNV of GOIs or if orthologs of different accessions are grouped together. In addition, we want to know what the difference is between GOI across different cauliflowers, because variation in these GOI may lead to different temperature sensitivity. To do this, the homology groups containing the GOI needed to be analyzed. This was done by constructing a so-called variation table, which gives information about genes in certain homology groups and genes that come up when BLASTing a GOI. To validate the CNV and the location of the genes found in the variation table, the BLAST tool of the BRAD database (Chen et al. 2022), combined with other literature, was used.

In this section, two GOIs are chosen to be highlighted and used to demonstrate the infrastructure to map the genetic variation within cauliflower. The two GOIs are FLC and ELF3. In short, FLC is an important gene in the vernalization pathway. FLC binds to flowering genes like FT and SOC1, which are then repressed and prevent the plant from flowering (Helliwell et al. 2006). ELF3 is a transcriptional regulator of elongation growth. ELF3 is highly temperature-dependent and tends to repress growth overnight when the temperature drops (Box et al. 2015). ELF3 is part of the circadian clock, a network of interlocking transcriptional feedback loops that regulates hypocotyl growth in this case (Nusinow et al. 2011). For these genes, we show the CNV variation in

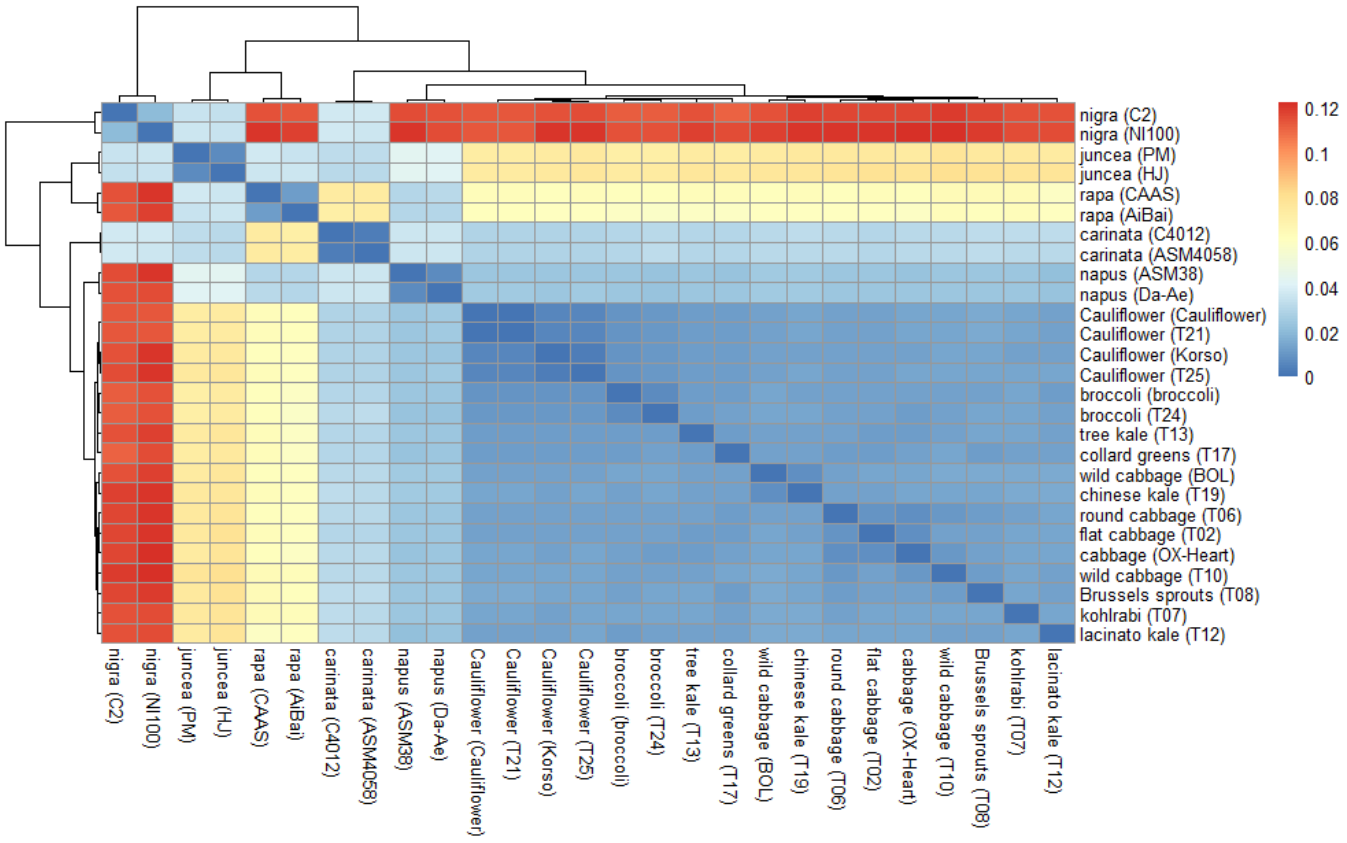


Figure 3: Mash distance between 27 Brassica genomes, including 4 cauliflower genomes. Red indicates a greater mash distance, while dark blue indicates more sequence similarity.

Table 2: Number of heterozygote SNPs called across the four resequencing accessions. Each accession has hundreds of thousands of SNPs, which increases the expectation that SNPs can be found in GOIs.

	Liria	Caniego	TKI-0143	TKI-0155
Number of heterozygous SNPs	915590	784334	1148040	1378007

the used cauliflower genome assemblies, and for ELF3, we show how the reads from the resequencing data are mapped and which variants were called in this region. The expectation is that there will be no difference in CNV variation for these genes. We would expect that the main variation within those genes consists of SNPs, so we would expect to see some variants in the genome browser.

CNV for FLC and ELF3

To analyze the CNV within cauliflower, a variation table was made for the four cauliflower assemblies using the cauliflower pangenome. The complete variation table, including all size initial genes of interest, is available as a Supplementary file. For *Brassica oleracea*, multiple copies of FLC are known. BoFLC1, BoFLC3, and BoFLC5 were previously detected (Schranz et al. 2002) and later also BoFLC4 (Lin et al. 2005) and BoFLC2 (Okazaki et al. 2007) were found in *Brassica oleracea* species.

In the variation table, I found 5 FLCs for the 'Cauliflower' and T25 accession, 6 FLCs for the T21 accession, and 4 FLCs for the Korso accession. Two FLCs of the 'Cauliflower' and one of T25 did not pass the initial BLAST, but are in homology group 35090639, as all other FLCs. This homology group contains 437 genes in total. When looking at the location of those FLCs, three FLCs were on chromosome 9 for all the accessions except 'Cauliflower', which had two. Korso had one FLC on chromosome 3, while T21 and T25 had each two FLCs on chromosome 3, and 'Cauliflower' had

three FLCs on chromosome 3. The sixth FLC in the T21 accessions was located on chromosome 7. The most important aspect to learn from the variation table is that the suggested CNV needs to be interpreted with caution. When looking at the FLCs in T21, the FLC lying on chromosome 7 is not shown for other cauliflower genomes. In addition, 'cauliflower' and T21 each have an FLC lying on chromosome 3, which has a % identity of below 70 compared to the FLC in AT. Because many other FLCs in the variation table have an 80% identity with the AT FLC, having a lower identity % might indicate that this is not a real FLC but more a FLC-like gene. Because the genomes used for these genes are not so well annotated, barcoding what kind of gene is found precisely is more difficult.

For ELF3, the variation table suggests that there is one homolog per cauliflower genome on chromosome 8, all four in homology group 35141310. In the BLAST results in Supplementary Figure S4, there is also a homolog on chromosome 8 with the lowest e-value. The other entries in the variation table have AT3G21320 as the closest *Arabidopsis thaliana* hit and are grouped in homology group 35131985. According to the TAIR database and Uniprot, this gene is suggested to function as an early flowering protein. In the variation table, these homologs are located on chromosomes 1, 3, and 5. The BLAST output also finds homologs on these chromosomes and additionally one on chromosome 4, which is not represented in the variation table. The variation table at the species level for ELF3 was made and is available as a Supplementary file. In there, we can see which genes also have AT3G21320 as the closest *Arabidopsis thaliana* hit, and we see some CNV. Lacinato kale, round cabbage, and the two cauliflowers have one copy of ELF3, tree kale has three copies, whereas all other species have two copies. The third copy in three kale should be interpreted with caution since the % identity is lower than all others and seems to be a neighbouring gene to the other copy found on chromosome 4. For all the accessions, the homolog on chromosome 8 was found. The second homolog found for the other accessions was found on chromosome 4, which was also found in the BLAST results. This suggests that there is a CNV variation between cauliflower, which has one ELF3 copy, and most other *Brassica oleracea* varieties, which have two copies. Now that cauliflower appears to have a single copy, and other *Brassica oleracea* accessions can have two copies, we are interested in whether ELF3 is also a single copy in the resequencing accessions.

ELF3 in resequencing accessions

Before using the reads of the resequencing accessions, the quality of the reads was also analyzed. Based on the MultiQC report, the number of reads for both TKI accessions was more than 80 million. Liria had a total of $\tilde{1}$05 million reads, and Caniego had a total of $\tilde{1}$30 million reads. The only category that did not pass for all accessions was the per base sequence content. Both R1 and R2 FASTQ files for the TKI accessions either failed or showed a warning. For all accessions, the sequence content of the first 10 bp was not nicely distributed. The MultiQC did not suggest any contamination with adapter sequences. For the sake of time, the reads were not trimmed in this research.

To see the CNV of ELF3 in the resequencing accessions, the resequencing accessions were mapped to T22, resulting in Supplementary Figure S5. In this figure, we can see that there is no CNV variation of the whole ELF3 gene in any resequencing accession. Therefore, we should see an increase in coverage over the full gene. Now there are only five regions of 50 to 100 nucleotides and one region of 300 nucleotides that have a higher coverage. Supplementary Figure S6 shows how the reads map and the variants called of the Liria accession. We can see that most of the variants called are in the regions with higher coverage, and that there are many reads that show soft-clipping in these regions. Now we are questioning whether these reads really should be mapped here. To test, the Liria reads in the ELF3 region were isolated and mapped to the other genome assemblies using SAMtools and bedtools v.2.31.1 (Quinlan and Hall 2010). Supplementary Figure S7 shows this mapping to all the genome assemblies. Here we can see that also for T21 and 'Cauliflower' the reads show soft-clipping. Only for T25, there are no soft-clipped reads, and the increased regions also seem to be absent. Using SAMtools idxstats, we could see where all the reads map to the T25 genome, and saw that 337 reads were mapped on chromosome 4. Looking in the genome browser on chromosome 4, we could see that these reads map to the gene BroT25.C04g43900. When making a variation table using the cauliflower pangenome, and using this gene as input, the closest hit with *Arabidopsis thaliana* is AT1G65630. This is a degradation of periplasmic proteins 3 (DEG3) gene, which in the first place does not seem to be specifically related to the ELF3 gene. Why the reads were mapped to this gene in T25 specifically is not further investigated, but this is an example that working with a pangenomic approach would prevent this from happening. Besides the soft-clipped reads that are not present in the ELF3 region when mapping to T25, most of the variants are also absent, shown in Supplementary Figure S8. This indicates that most of the variants called in this region are caused by the soft-clipped reads and could give a false picture of the variants in this gene. The variants that are left when mapping the reads to T25 should also be interpreted carefully. Despite the filter settings used, the suggested SNP from G to T is caused by three reads. The 1bp dele-

tion and insertion, which are on the same position, are caused by 2 and 4 reads, respectively. These variants are not showing the homozygosity or heterozygosity, which is a strong indication of a specific locus.

Read mapping results

Looking at the percentage mapped reads and the average coverage of the reads for the resequencing accessions could give us an indication of how well these accessions map using T22 as a reference genome. In Table 3, these numbers are given where Liria has the lowest percentage of mapped reads with 93.08%. Caniego has 96.18% while both TKI accessions have almost 100% mapped. The average coverage of the reads also does not seem to influence the mapping % since these numbers are not extremely low. The lower mapping percentage of Liria and Caniego could be because these accessions are more distinct from T22 than the TKI accessions are. To test this, Mash could be run to calculate the distance between the resequencing accessions and the genome assemblies. In that case, Liria could show more distance to the T22 than to other genome assemblies. This would suggest that mapping the Liria reads to a pangenome containing these genome assemblies will increase the % of mapped reads.

Conclusions and recommendations

In this research, we demonstrated a method to map the genetic variation within cauliflower by the use of a pangenome. We showed that making a variation table is a way to inspect CNV, but that the results should also be interpreted with caution. The difference in CNV between cauliflower and other *Brassica oleracea* varieties for ELF3 is a result that could help in explaining differences in phenotypes. The web portal helps the consortium with answering questions and providing data. In the research, a use case of ELF3 was shown by showing the variant calling and CNV for ELF3 in the resequencing data. These steps can be used to explore own data or other GOI using this data. In addition, this use case shows how soft-clipped reads can be handled. As a last point, this research includes an overview of the available data and their quality, and a genome-wide analysis which gives a picture of the genome-wide variation existing in the *Brassica* genus.

Some things could be improved in this research. The validation of the grouping between pangenomes of different evolutionary levels can be improved. Validating that certain BUSCOs are in the same homology group across these pangenomes will not guarantee that certain GOIs are also grouped together in both pangenomes. Because BUSCOs are highly conserved, it is more likely that BUSCOs group together than certain GOIs. GOIs might have more similar genes, versions, or CNVs, which can influence the grouping. Ideally, instead of BUSCOs, another single-copy gene would be used that is not highly conserved, so potential uneven grouping between pangenomes can be detected earlier. The variation table can be improved by finding a better way to find potential GOIs in the pangenome than by doing a BLAST. Instead of taking the best BLAST hit, the list of potential GOIs should be more focused on the function of a gene than on sequence similarity. By selecting genes that, for example, share specific protein domains. Before using the variation table results, a thorough evaluation with literature should be done to validate the CNV it suggests. Ideally, a well-annotated cauliflower would be included so genes do not have to be compared over a larger evolutionary distance. The reads for the resequencing accessions should be checked on the per base sequence content for future work. In other resequencing accessions, adapter sequences might be present. At least it is recommended to trim off the first 10bp of the reads when other resequencing accessions show similar results. This report did not mention any reasoning why the soft-clipped reads in ELF3 did map properly to the DEG3 gene in T25. To deeper investigate this, the sequences of the ELF3 and the DEG3 genes of those genome assemblies could be further compared to see which sequence variation caused this soft-clipping. For variant calling, it is recommended to put a filter that calls variants when it is supported by a minimum fraction of the reads. This can be done using FreeBayes. The variants called in ELF3 in T25 do not seem to be useful variants for the breeding process.

For this research, we could see that based on the fact that reads showing soft-clipping for T22, T21, and 'Cauliflower' did not do this for T25. When using a pangenome, these reads would be mapped properly. The lower percentage of mapped reads for the Liria accession to the T22 genome could also be increased when those reads are mapped to a pangenome. This is helpful to map the variation between resequencing samples, which can explain temperature sensitivity in cauliflower. In addition, a pangenome helps when searching for genes of interest across different species based on homology grouping.

The next step is to gain an easier insight into the genetic variations in GOIs for the resequencing accessions than using the genome browser in the web portal. This should be done in a pangenome browser so the accessions can be mapped to the pangenome and variants can be called more accurately. Ideally, this pangenome would consist of more cauliflower assemblies to take more variation into account.

Table 3: Number and percentage of reads mapped for each resequencing accession when mapping it to one single reference genome or to the pangenome.

	total number of reads primarily mapped to T22	% mapped	average coverage
Liria	195,585,207	93.08	53.6943
Caniego	255,213,204	96.18	70.1135
TKI-0143	160,976,814	99.65	44.1353
TKI-0155	166,349,145	99.58	44.934

References

Akter, Ayasha et al. (Feb. 9, 2021). "Genome Triplication Leads to Transcriptional Divergence of FLOWERING LOCUS C Genes During Vernalization in the Genus Brassica". In: *Frontiers in Plant Science* 11. Publisher: Frontiers Media SA. DOI: 10.3389/fpls.2020.619417.

Box, Mathew S. et al. (Jan. 2015). "ELF3 Controls Thermoresponsive Growth in Arabidopsis". In: *Current Biology* 25.2, pp. 194–199. DOI: 10.1016/j.cub.2014.10.076.

Camacho, Christiam et al. (Dec. 2009). "BLAST+: architecture and applications". In: *BMC Bioinformatics* 10.1, p. 421. DOI: 10.1186/1471-2105-10-421.

Chen, Haixu et al. (Jan. 7, 2022). "BRAD V3.0: an upgraded Brassicaceae database". In: *Nucleic Acids Research* 50 (D1), pp. D1432–D1441. DOI: 10.1093/nar/gkab1057.

Danecek, Petr et al. (Jan. 29, 2021). "Twelve years of SAMtools and BCFtools". In: *GigaScience* 10.2, giab008. DOI: 10.1093/gigascience/giab008.

Ewels, Philip et al. (Oct. 1, 2016). "MultiQC: summarize analysis results for multiple tools and samples in a single report". In: *Bioinformatics* 32.19, pp. 3047–3048. DOI: 10.1093/bioinformatics/btw354.

FAO (2024). "Agricultural production statistics 2010–2023". In: Publisher: FAO ;

Garrison, Erik and Gabor Marth (2012). *Haplotype-based variant detection from short-read sequencing*. Version Number: 2. DOI: 10.48550/ARXIV.1207.3907.

Gendall, Anthony R. et al. (Nov. 2001). "The VERNALIZATION 2 Gene Mediates the Epigenetic Regulation of Vernalization in Arabidopsis". In: *Cell* 107.4, pp. 525–535. DOI: 10.1016/S0092-8674(01)00573-6.

Golicz, Agnieszka A. et al. (Nov. 11, 2016). "The pangenome of an agronomically important crop plant Brassica oleracea". In: *Nature Communications* 7.1, p. 13390. DOI: 10.1038/ncomms13390.

Helliwell, Chris A. et al. (Apr. 2006). "The Arabidopsis FLC protein interacts directly *in vivo* with *SOC1* and *FT* chromatin and is part of a high-molecular-weight protein complex". In: *The Plant Journal* 46.2, pp. 183–192. DOI: 10.1111/j.1365-313X.2006.02686.x.

Kim, Dong-Hwan and Sibum Sung (Mar. 26, 2013). "Coordination of the Vernalization Response through a *VIN3* and *FLC* Gene Family Regulatory Network in *Arabidopsis*". In: *The Plant Cell* 25.2. Publisher: Oxford University Press (OUP), pp. 454–469. DOI: 10.1105/tpc.112.104760.

Li, Heng and Richard Durbin (Mar. 1, 2010). "Fast and accurate long-read alignment with Burrows–Wheeler transform". In: *Bioinformatics* 26.5, pp. 589–595. DOI: 10.1093/bioinformatics/btp698.

Li, Peijin et al. (Aug. 1, 2014). "Multiple *FLC* haplotypes defined by independent *cis* -regulatory variation underpin life history diversity in *Arabidopsis thaliana*". In: *Genes & Development* 28.15, pp. 1635–1640. DOI: 10.1101/gad.245993.114.

Li, Xing et al. (Mar. 2024). "Large-scale gene expression alterations introduced by structural variation drive morphotype diversification in Brassica oleracea". In: *Nature Genetics* 56.3, pp. 517–529. DOI: 10.1038/s41588-024-01655-4.

Lin, Shu-I et al. (Mar. 1, 2005). "Differential Regulation of *FLOWERING LOCUS C* Expression by Vernalization in Cabbage and Arabidopsis". In: *Plant Physiology* 137.3, pp. 1037–1048. DOI: 10.1104/pp.104.058974.

Manni, Mosè et al. (Sept. 27, 2021). "BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes". In: *Molecular Biology and Evolution* 38.10. Ed. by Joanna Kelley, pp. 4647–4654. DOI: 10.1093/molbev/msab199.

Nusinow, Dmitri A. et al. (July 2011). "The ELF4–ELF3–LUX complex links the circadian clock to diurnal control of hypocotyl growth". In: *Nature* 475.7356, pp. 398–402. DOI: 10.1038/nature10182.

Okazaki, K. et al. (Feb. 2, 2007). "Mapping and characterization of FLC homologs and QTL analysis of flowering time in Brassica oleracea". In: *Theoretical*

and Applied Genetics 114.4, pp. 595–608. DOI: 10.1007/s00122-006-0460-6.

Ondov, Brian D. et al. (Dec. 2016). "Mash: fast genome and metagenome distance estimation using MinHash". In: *Genome Biology* 17.1, p. 132. DOI: 10.1186/s13059-016-0997-x.

Picchi, Valentina, Marta Fibiani, and Roberto Lo Scalzo (2020). "Cauliflower". In: *Nutritional Composition and Antioxidant Properties of Fruits and Vegetables*. Elsevier, pp. 19–32. DOI: 10.1016/B978-0-12-812780-3.00002-7.

Quinlan, Aaron R. and Ira M. Hall (Mar. 15, 2010). "BEDTools: a flexible suite of utilities for comparing genomic features". In: *Bioinformatics* 26.6, pp. 841–842. DOI: 10.1093/bioinformatics/btq033.

Ridge, Stephen et al. (Jan. 2015). "The role of BoFLC2 in cauliflower (Brassica oleracea var. botrytis L.) reproductive development". In: *Journal of Experimental Botany* 66.1, pp. 125–135. DOI: 10.1093/jxb/eru408.

Schiessl, Sarah et al. (Feb. 6, 2017). "Post-polyploidisation morphotype diversification associates with gene copy number variation". In: *Scientific Reports* 7.1, p. 41845. DOI: 10.1038/srep41845.

Schranz, M Eric et al. (Nov. 1, 2002). "Characterization and Effects of the Replicated Flowering Time Gene *FLC* in *Brassica rapa*". In: *Genetics* 162.3, pp. 1457–1468. DOI: 10.1093/genetics/162.3.1457.

Sheikhizadeh, Siavash et al. (Sept. 1, 2016). "PanTools: representation, storage and exploration of pangenomic data". In: *Bioinformatics* 32.17, pp. i487–i493. DOI: 10.1093/bioinformatics/btw455.

Tettelin, Hervé et al. (Sept. 27, 2005). "Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae* : Implications for the microbial "pan-genome"". In: *Proceedings of the National Academy of Sciences* 102.39, pp. 13950–13955. DOI: 10.1073/pnas.0506758102.

Yim, Won Cheol et al. (Jan. 4, 2022). *The last missing piece of the Triangle of U : the evolution of the tetraploid Brassica carinata genome*. DOI: 10.1101/2022.01.03.474831.

Supplementary figures

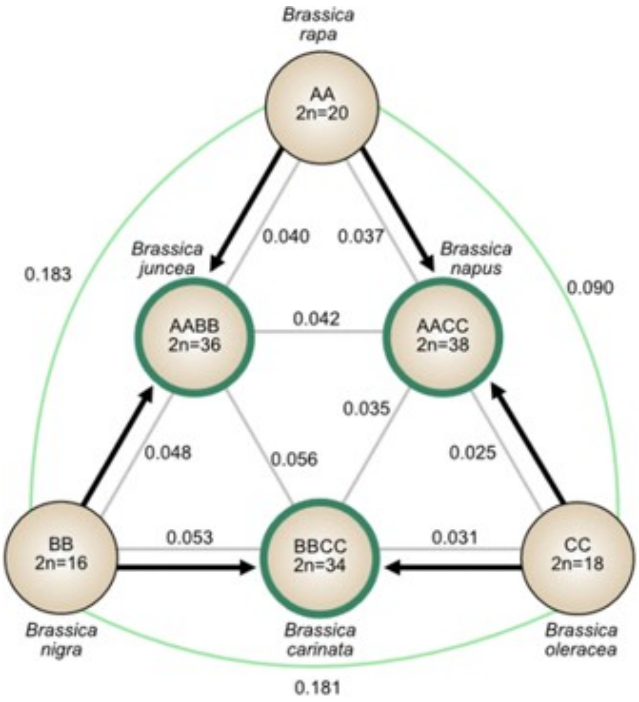


Figure S1: Overview of the evolution of *Brassica* including the hybridization between the diploid species.

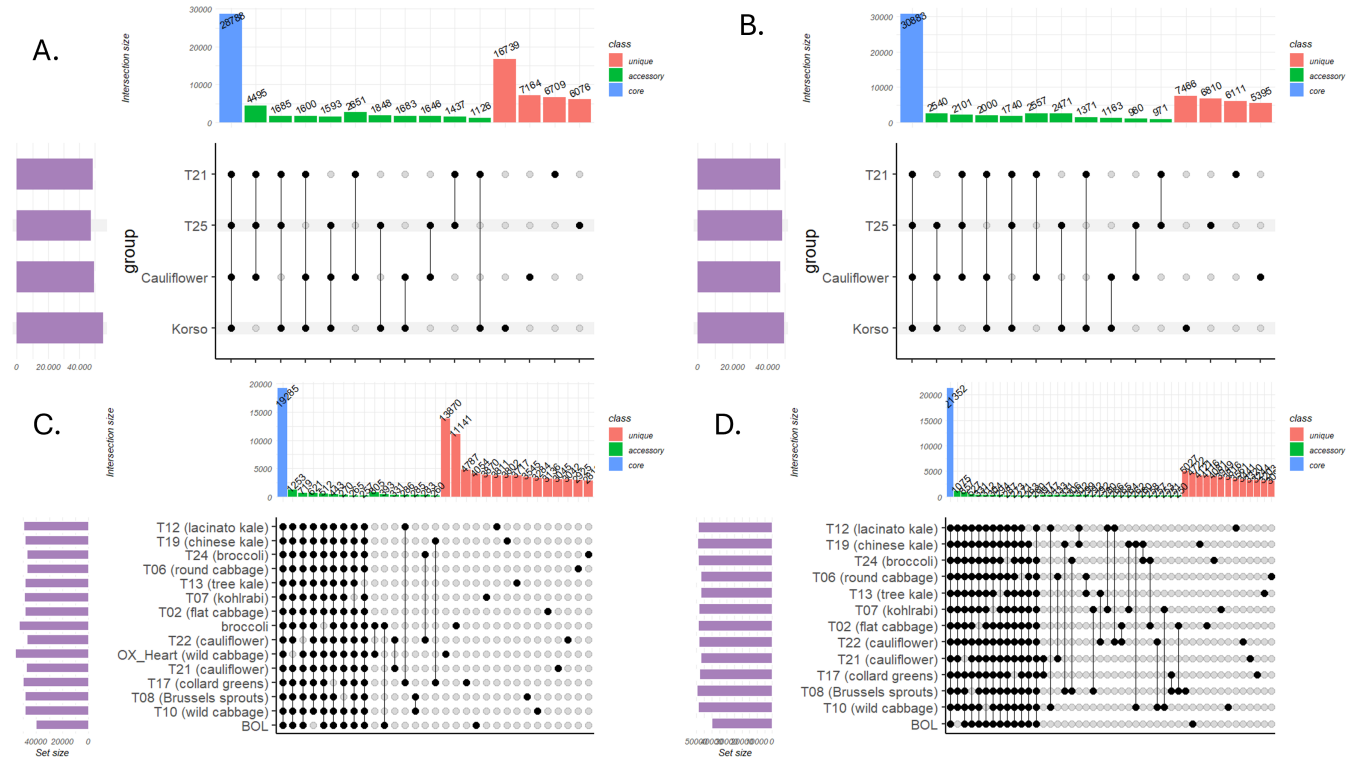


Figure S2: Difference in unique genes regarding the Korso assembly when using a different annotation. Upset plots from the gene classification output. Plot A shows the gene classification output, including the initial Korso

annotation. Plot B shows the results when using the T22 annotation (X. Li et al. 2024). The poor annotation of Korso led to an increase of more than 9000 unique genes. In total, 30883 genes are shared among all cauliflower genomes. In plot C, the results of the gene classification of the genome assemblies used in the *Brassica oleracea* pangenome are given. Here, the OX_Heart and the 'broccoli' accessions show a great number of unique genes. When removing those accessions, the fraction of core genes increases with more than 2000 genes to 21352 genes.

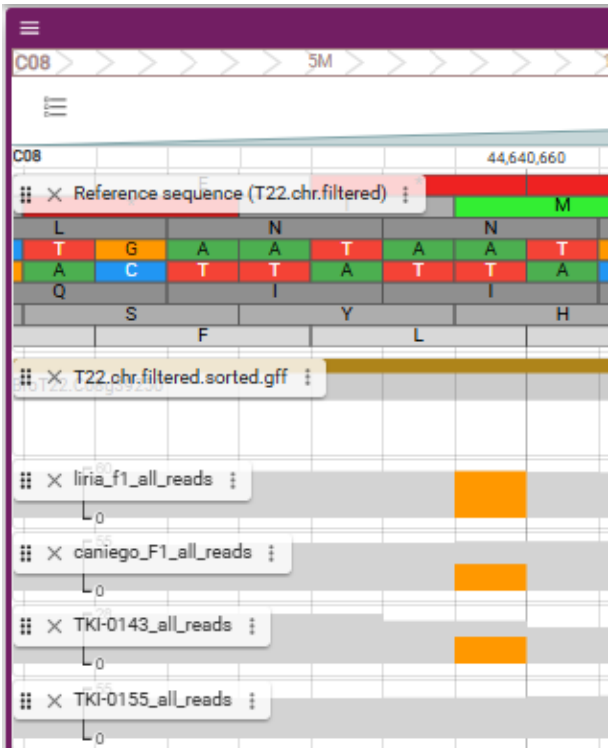
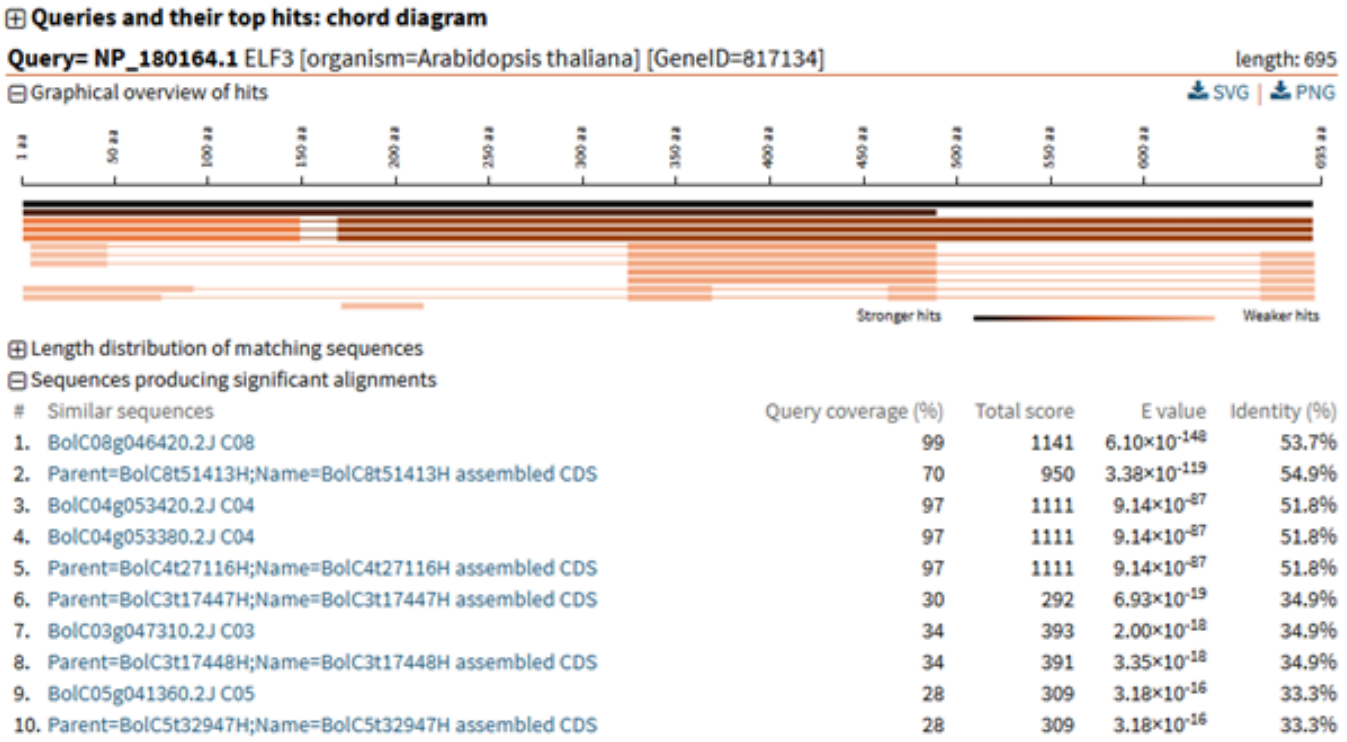


Figure S3: A SNP which is homozygous in liria, heterozygous in caniego and TKI-0143, and absent in TKI-0155, shown in the genome browser of the web portal. The reads are not left out to make the screenshot more visible.



#	Similar sequences	Query coverage (%)	Total score	E value	Identity (%)
1.	BolC08g046420.2J C08	99	1141	6.10×10^{-148}	53.7%
2.	Parent=BolC8t51413H;Name=BolC8t51413H assembled CDS	70	950	3.38×10^{-119}	54.9%
3.	BolC04g053420.2J C04	97	1111	9.14×10^{-87}	51.8%
4.	BolC04g053380.2J C04	97	1111	9.14×10^{-87}	51.8%
5.	Parent=BolC4t27116H;Name=BolC4t27116H assembled CDS	97	1111	9.14×10^{-87}	51.8%
6.	Parent=BolC3t17447H;Name=BolC3t17447H assembled CDS	30	292	6.93×10^{-19}	34.9%
7.	BolC03g047310.2J C03	34	393	2.00×10^{-18}	34.9%
8.	Parent=BolC3t17448H;Name=BolC3t17448H assembled CDS	34	391	3.35×10^{-18}	34.9%
9.	BolC05g041360.2J C05	28	309	3.18×10^{-16}	33.3%
10.	Parent=BolC5t32947H;Name=BolC5t32947H assembled CDS	28	309	3.18×10^{-16}	33.3%

Figure S4: BLAST results using the BRAD database as input the ELF3 *Arabidopsis thaliana* sequence was used to BLAST against the *Brassica oleracea* CDS sequence of HDEM and JVS2.0.

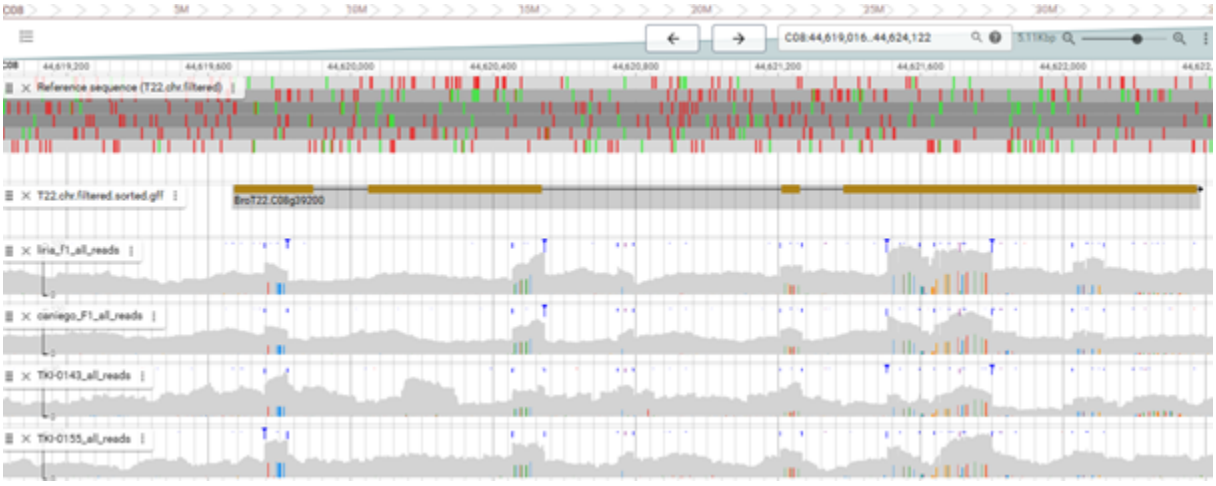


Figure S5: The coverage of the ELF3 gene for all the resequencing accessions mapped to the T22 genome. The ELF3 gene does not seem to show CNV for any accessions but does show regions of higher coverage.

Figure S6: Liria reads mapped to the ELF3 gene. Showing the soft-clipped reads in the regions with coverage. In addition, the VCF file is loaded in the browser. The variant-calling was done using Freebayes with standard filters, which they provide. Nevertheless, many SNPs are called for this region, mainly caused by the clipped reads.

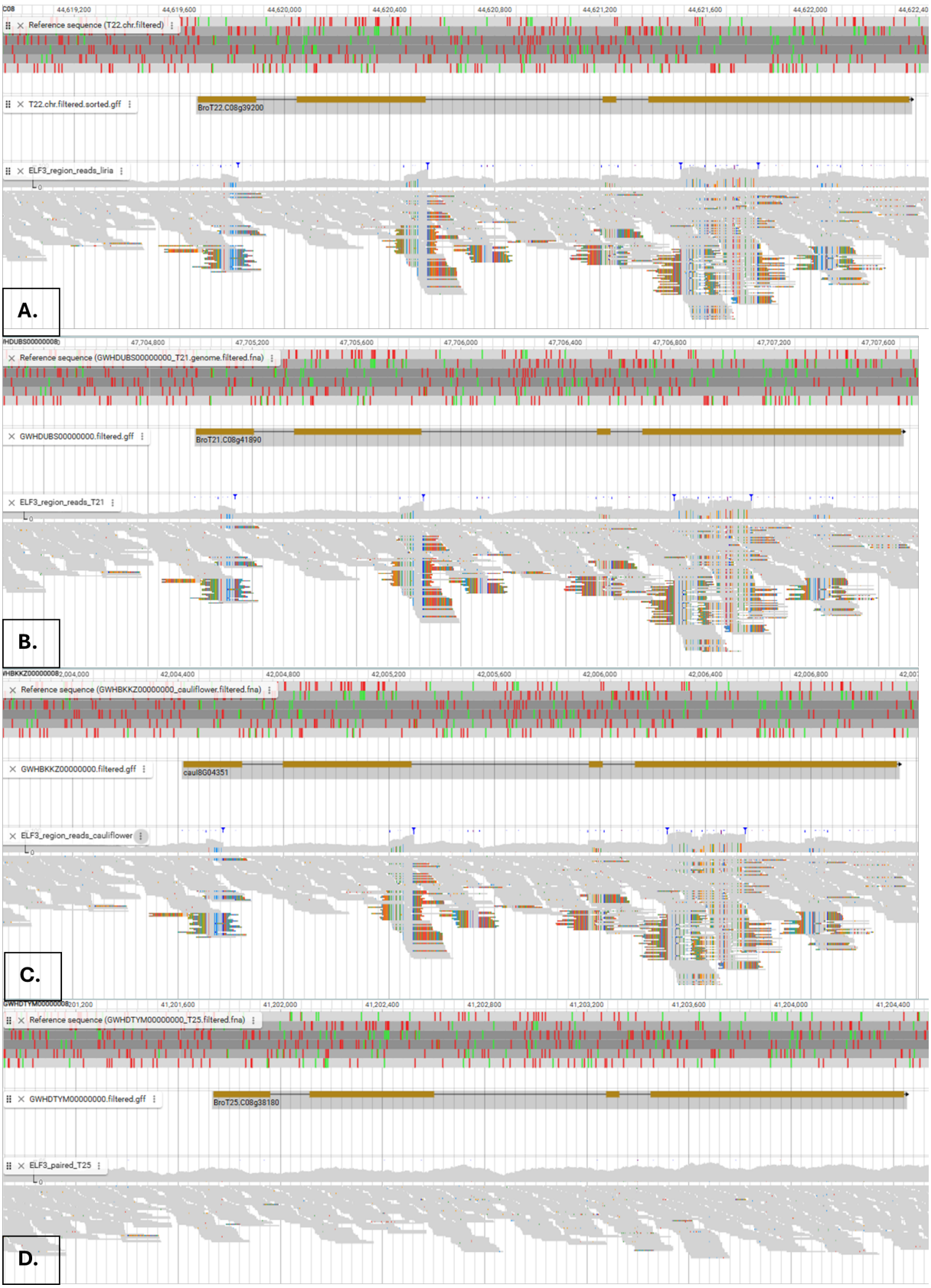


Figure S7: The reads from the ELF3 region gained from the Liria BAM file mapped to all four cauliflower genome

assemblies. In panel A, the reads are mapped to T22, in panel B to T21, in panel C to 'Cauliflower', and in panel D to T25. In the browser, we can see that T22, T21, and 'Cauliflower' show soft clipping for the ELF3 gene, but T25 does not. In addition, the increased coverage region seems to be absent when mapping the reads to T25 compared to the other resequencing accessions.

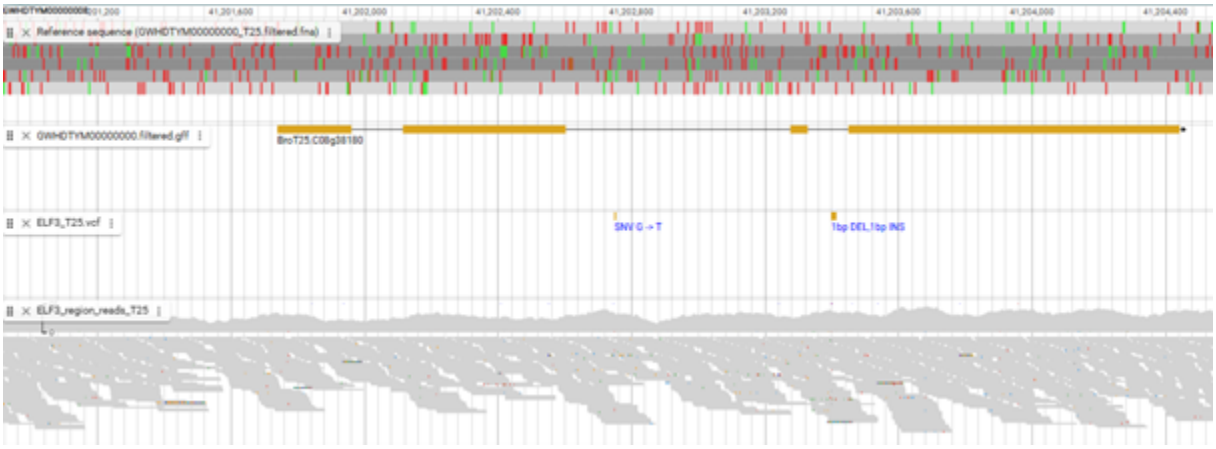


Figure S8: Reads in the ELF3 region gained from the Liria BAM file, mapped to T25. This is the only genome assembly used that did not show soft-clipped reads in ELF3. In addition, the majority of the variants are not called anymore.