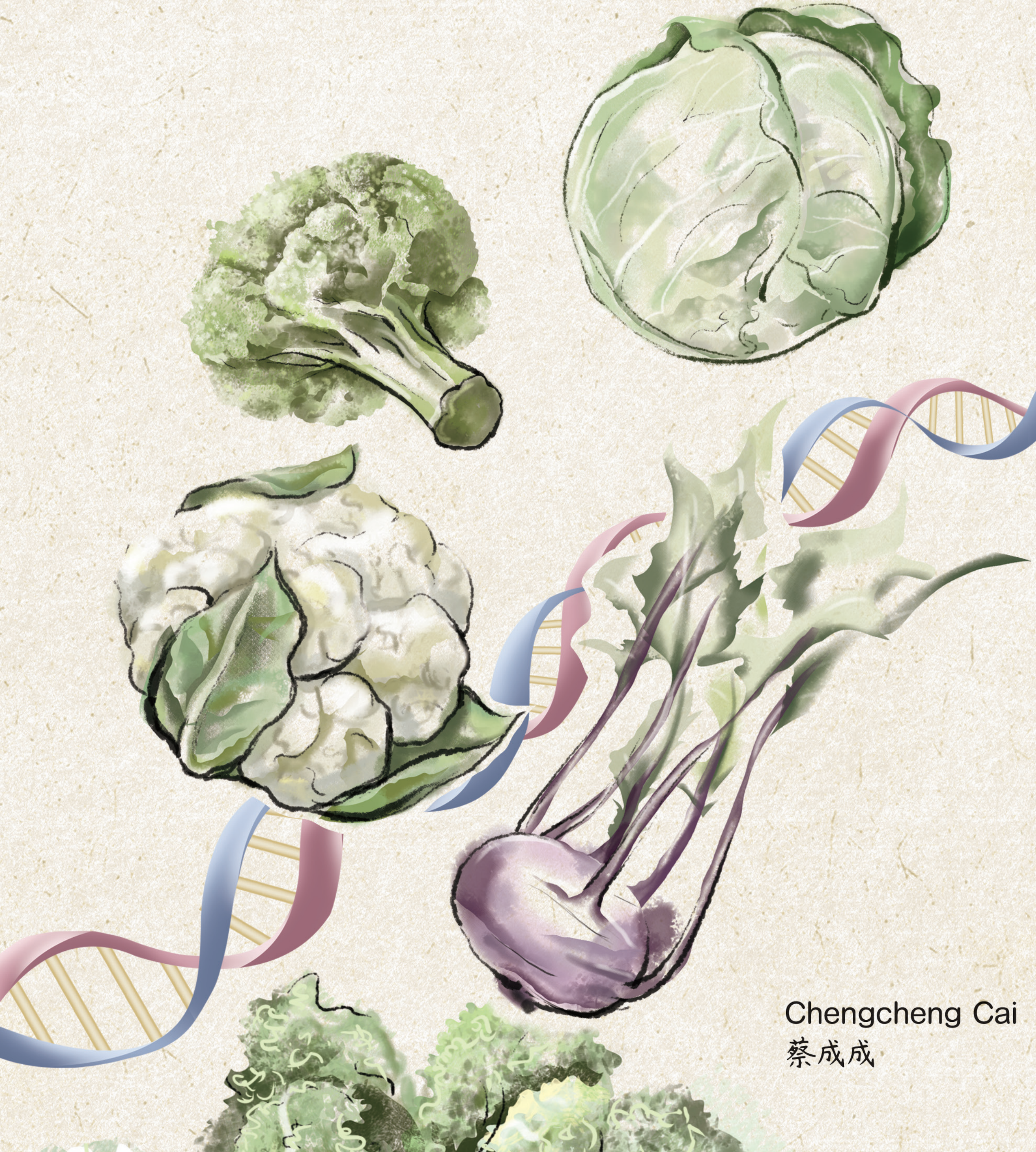
# Unlocking the secrets of *Brassica oleracea* crops: a genomic journey



Chengcheng Cai
蔡成成

**Propositions**

1. The generation of high-quality reference genomes is limited by sequencing technology rather than computational algorithms.
(this thesis)
2. The constrained number and uneven distribution of meiotic crossovers in crops limit the breeding efficiency.
(this thesis)
3. Imbalanced appreciation exists between the scientific impact of research on model plant species and the agronomic impact of research on crops.
4. Unified standards need to be established for scientific journals to facilitate submission and review processes.
5. Office noise is the biggest factor that influences working efficiency.
6. Parents should limit their children's time on using digital devices.

Propositions belonging to the thesis, entitled

Unlocking the secrets of *Brassica oleracea* crops: a genomic journey

Chengcheng Cai
Wageningen, 31 May 2023

# Unlocking the secrets of *Brassica oleracea* crops: a genomic journey

Chengcheng Cai

**Thesis committee**

**Promotor**
Dr A.B. Bonnema
Associate Professor, Plant Breeding
Wageningen University & Research

**Co-promoter**
Dr H.J. Finkers
Senior Scientist, Plant Breeding
Wageningen University & Research

**Other members**
Prof. Dr M.E. Schranz, Wageningen University & Research
Dr C.M. Kreike, Inholland University of Applied Sciences, Amsterdam
Dr A.D.J. van Dijk, Wageningen University & Research
Dr M. Rousseau-Gueutin, INRAE, Institut Agro, Université de Rennes, France

This research was conducted under the auspices of the Graduate School Experimental Plant Sciences (EPS).

# Unlocking the secrets of *Brassica oleracea* crops: a genomic journey

## Chengcheng Cai

**Thesis**
submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus,
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Wednesday 31 May 2023
at 1:30 p.m. in the Omnia Auditorium.

Chengcheng Cai
Unlocking the secrets of *Brassica oleracea* crops: a genomic journey,
199 pages.

PhD thesis, Wageningen University, Wageningen, the Netherlands (2023)
With references, with summary in English

ISBN 978-94-6447-632-3
DOI https://doi.org/10.18174/589859

# Table of contents

# Chapter 1

## General Introduction

## *Brassica* genus

The genus *Brassica* currently consists of 38 species and numerous varieties, many of which are important crops or weeds (Cheng *et al.*, 2013). Six major economically important interrelated species from the *Brassica* genus make up the so-called "Triangle of U" (Nagaharu, 1935b). These species include three diploids *Brassica rapa* (AA, 2n=20), *Brassica nigra* (BB, 2n=16), *Brassica oleracea* (CC, 2n=18) and three allotetraploids formed by pairwise hybridization between the three diploids, *Brassica juncea* (AABB, 2n=36), *Brassica napus* (AACC, 2n=38) and *Brassica carinata* (BBCC, 2n=34). Each of these six species encompasses many widely cultivated crops, making them an ideal model for polyploidy and evolutionary studies.

### Whole genome triplication and subgenomes

It was estimated that the *Brassica* ancestor diverged from the *Arabidopsis thaliana* lineage ~14.5 million years ago (MYA) (Yang *et al.*, 1999, Town *et al.*, 2006, Beilstein *et al.*, 2010, Wang *et al.*, 2011b, Cheng *et al.*, 2017). After this divergence, *Brassica* genomes have experienced the common Brassiceae-specific whole genome triplication (WGT) event ~11 MYA (Lysak *et al.*, 2005, Wang *et al.*, 2011b, Liu *et al.*, 2014a, Parkin *et al.*, 2014a), which is confirmed by extensive comparative analyses between genome sequences of *Brassica* species and *A. thaliana* (Wang *et al.*, 2011b, Chalhoub *et al.*, 2014, Cheng *et al.*, 2014, Liu *et al.*, 2014a, Parkin *et al.*, 2014a). Rediploidization processes following the WGT, such as extensive gene fractionation, genomic reshuffling and chromosome reduction, shaped the genome structure of the extant diploid *Brassica* species (Cheng *et al.*, 2014). The extant *Brassica* diploid genomes are expected to consist of three subgenomes due to the WGT. Indeed, these three subgenomes, the least fractionated (LF), the medium fractionated (MF1) and the most fractionated (MF2) subgenome, have been identified and they were demonstrated to have evolved from a common translocation Proto-Calepineae Karyotype (tPCK) ancestral diploid genome (Cheng *et al.*, 2013, Cheng *et al.*, 2014). A "two-step theory" explaining the WGT process in *Brassica* was proposed, which suggests a tetraploidization event between the tPCK genomes of MF1 and MF2, followed by fractionation and a subsequent hybridization event between the diploidized genome and a third tPCK genome of LF (Cheng *et al.*, 2012a, Cheng *et al.*, 2014, Cai *et al.*, 2021).

### Divergence time between major *Brassica* species

Genome sequencing of *Brassica* species enabled the estimation of divergence time between the diploids and formation time of the allotetraploids. Synonymous substitution (Ks) analysis indicated an emergence time of ~6.5 MYA for *B. nigra* (Cheng *et al.*, 2017). *B. rapa* and *B. oleracea* were estimated to diverge ~4.6 MYA. The formation time of *B. napus* were ~7500 and 38,000-51,000 years ago, based on a Ks estimation and a Bayesian Markov chain Monte Carlo (MCMC) simulation, respectively (Chalhoub *et al.*, 2014, Yang *et al.*, 2016, Lu *et al.*, 2019). *B. juncea* was deduced to have formed 39,000-55,000 and 72,000-80,000 years ago in two independent studies (Yang *et al.*, 2016, Song *et al.*, 2021b). Very recently, a high-quality genome of *B. carinata* became available and it was estimated that *B. carinata* formed about 45,000-49,000 years ago (Song *et al.*, 2021b). Song and colleagues (Song *et al.*, 2021b) concluded that the formation of *B. carinata* is slightly earlier than *B. napus* but later than *B. juncea*.

### Glucosinolates in *Brassica*

*Brassica* crops vary extensively in their metabolite content and composition, which is especially studied for the *Brassica*-specific class of glucosinolates (GSLs). GSLs are a group of secondary plant metabolites comprising sulfur and nitrogen, which are almost exclusively found in the order of Brassicales (Halkier and Gershenzon, 2006, Harun *et al.*, 2020). To date, more than 130 GSL structures have been scientifically documented (Petersen *et al.*, 2018, Harun *et al.*, 2020). Based on the precursor amino acid, GSLs can be classified into three categories: aliphatic, aromatic and indole GSLs (Mithen *et al.*, 2000, Halkier and Gershenzon, 2006, Sønderby *et al.*, 2010, Zhang *et al.*, 2015b). GSLs and their breakdown products attract interest for extensive studies because their activated forms play important roles in defense against pathogens and pests and have health benefits for humans in consumed *Brassica*'s (Zhang *et al.*, 2015b, Harun *et al.*, 2020). GSLs protect the plant from insect and pathogen damage through the hydrolysis products that are deterrent or toxic to attackers (Petersen *et al.*, 2018). For human health, some GSL hydrolysis products, such as isothiocyanates (ITCs), have been shown to be involved in lowering the risk of myocardial infarction and several kinds of cancer (Petersen *et al.*, 2018). However, not all GSL hydrolysis products are beneficial for humans and farm animals that eat feed containing *Brassica* seed meal. One example is progoitrin, which exists in several *Brassica* species, displaying anti-thyroid activity and promoting goitre disease (Voorrips *et al.*, 2000, Bonnema *et al.*, 2019). Also, GSLs provide a series of tastes like bitterness and pungency. Breeding *Brassica* variety with tailored GSL content and composition requires more fundamental knowledge of genetic loci that control GSL biosynthesis and storage in these crops.

### *Brassica oleracea*

*B. oleracea* consists of many subspecies that are usually referred to as morphotypes, which exhibit enormous diversity in their appearances, including var. *capitata* (cabbage), var. *botrytis* (cauliflower), var. *italica* (broccoli), var. *gongylodes* (kohlrabi), var. *gemmifera* (Brussels sprouts), var. *viridis* (Collard green), var. *alboglabra* (Chinese kale), var. *costata* (Tronchuda kale), *var. acephala* (bore and curly kale, marrow stem kale, etc), var. *acephela* (ornamental kale), etc (Fig. 1) (Kole and Henry, 2010, Bonnema *et al.*, 2011, Dias, 2012, Guo *et al.*, 2019). Despite this enormous diversity, *B. oleracea* truly remains one species and morphotypes can be interbred. Although the resulting progenies of inter-morphotype crosses, for example aiming to introgress resistances across crops, are usually viable, problems like reduced recombination and sterility are often met.

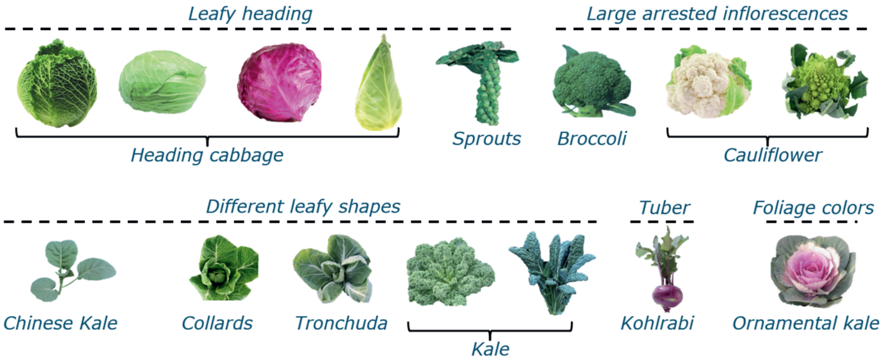


**Fig. 1** Phenotype diversity of *Brassica oleracea* morphotypes.

## Morphological characters of *B. oleracea* crops

Domesticated *B. oleracea* morphotypes are important vegetable and fodder crops, with several crops cultivated almost worldwide (*i.e.* cabbage, cauliflower, broccoli). Some other crops are cultivated only in specific countries depending on local preferences, such as Tronchuda's in Portugal and Collards in the USA. Wild *B. oleracea* plants form a strong stem with large waxy leaves, and only flower after several winters (Crozier, 1891, Kole and Henry, 2010, Bonnema *et al.*, 2011). Unlike the wild plants, domesticated *B. oleracea* crops are usually grown as annual or biennial plants and form their own distinguishable morphological characteristics, as a result of human selection during domestication and breeding. As such, cabbages are characterized by a round or pointed head that is made of leaves surrounding the terminal bud (Maggioni, 2015). Brussels sprouts are axillary buds, resembling tiny cabbage heads, that form at the base of each leaf, alongside a long, spiral stem.

Broccoli and cauliflower are characterized by the typical curd formed by large arrested inflorescences. The edible portion of broccoli is the florets, which are actually the large branching green undeveloped flower buds. The edible head of cauliflower is actually a mass of abortive flowers. The edible part of kohlrabi is the erect stem of kohlrabi, which is swollen at the bottom of the plant. Many other *B. oleracea* crops are mainly characterized by their different shapes of fleshy leaves, such as collard green, Chinese kale, tronchuda kale, bore and curly kale, marrow stem kale, ornamental kale, etc.

### Domestication of *B. oleracea*

Wild *B. oleracea*, its domesticated crops and their closely-related wild relatives (wild 'C9 species', *i.e. Brassica incana*, *Brassica cretica*, *Brassica villosa*, *Brassica macrocarpa* and *Brassica rupestris*) all belong to the so-called 'C-genome group' characterized by nine pairs of chromosome (Maggioni, 2015). Wild *B. oleracea* can be found along the Atlantic and Mediterranean coasts. Wild 'C9 species' mainly occur in the Mediterranean region, particularly in Sicily and Greece (Snogerup *et al.*, 1990, Tribulato *et al.*, 2017). However, the domestication history (origin, timing and route) of cultivated *B. oleracea* is still unclear. Both Northwestern Europe and Mediterranean/Middle East have been hypothesized as potential regions of origin (Maggioni *et al.*, 2018). Based on evidence from ancient literatures, it is more likely that cultivated *B. oleracea* originated from the Mediterranean/Middle East (Kole and Henry, 2010, Maggioni *et al.*, 2018). Besides, Mabry and colleagues recently investigated the evolutionary history of *B. oleracea* using mRNA-seq data of 224 accessions representing 14 different *B. oleracea* crop types and nine potential wild progenitor species (Mabry *et al.*, 2021). They identified *B. cretica* as the closest living relative of cultivated *B. oleracea*, supporting an origin of cultivation in the Eastern Mediterranean region (Mabry *et al.*, 2021). With regard to domestication route, one hypothesis is that kales were the earliest cultivated form of *B. oleracea* by the Celts along the Atlantic and Mediterranean coasts (Gómez-Campo and Prakash, 1999). Subsequently, they were brought to the East Mediterranean region (first and second millenia BC) to become domesticated with an explosive diversification of the cultivated forms. Another hypothesis is that several kales were already distinguished and these diverse kales were introduced to the Middle East through tin trade routes (Bronze Age, around 3300-1000 BC) from Spain and the British islands to the Phoenicians (present-day Libanon) (Penhallurick, 2008, Berger *et al.*, 2019). In ancient writings, the earliest descriptions of kales are from the Greek scholar

Theophrastrus (370-285 BC), suggesting that domestication of *B. oleracea* can date back to as early as around 400 BC.

## Genome sequencing

### Revolutions of sequencing technology

The past few decades have seen several rounds of technological revolution in DNA sequencing. The first revolution came with the advent of Sanger sequencing, which allows deciphering of complete genome sequences for the first time (Sanger *et al.*, 1977, Van Dijk *et al.*, 2018). This technology was considered the gold standard for DNA sequencing for the next 25 years since 1977 (Sanger *et al.*, 1977, Grada and Weinbrecht, 2013). Its read lengths typically range between 600 and 800 bp and the base accuracy is high with only 0.01% error rate (Mardis, 2017). However, the major drawback is that this technology is too labour-intensive, time-consuming and expensive for large-scale sequencing projects (Van Dijk *et al.*, 2018). Sanger sequencing was used in the 13-year-long (1990-2003) and very expensive ($3 billion costs) Human Genome Project (HGP) (1990-2003) (Olsen *et al.*, 2001, Venter *et al.*, 2001, Grada and Weinbrecht, 2013). Starting from 2005, the second revolution came with the appearance of next-generation (NGS) sequencing technologies, also known as short-read sequencing technologies. The advent of NGS was driven by the increasing demand for cheaper and faster sequencing methods since the completion of HGP (Grada and Weinbrecht, 2013, Van Dijk *et al.*, 2018). NGS technologies can provide extremely high throughput data and dramatically decrease the sequencing costs. Additionally, the base accuracy is also very high with error rates of 0.1-0.25% (Demirci, 2021). As a consequence, they almost completely superseded Sanger sequencing (Shendure *et al.*, 2017). However, the major limitation of NGS is the short read length, only ranging from 2 x 100 bp to 2 x 300 bp for paired-end reads depending on current Illumina sequencing platforms. After several years of competition among various commercial companies, the NGS market is currently dominantly occupied by Illumina (Van Dijk *et al.*, 2018). Shortly after the occurrence of NGS, the third DNA sequencing revolution came, marked with the advent of long-read sequencing technologies, also known as third-generation sequencing (TGS), which are characterized by single-molecule sequencing and sequencing in real time (Schadt *et al.*, 2010, Van Dijk *et al.*, 2018). Two representatives of TGS are 'Pacific Biosciences' (PacBio) single-molecule real-time (SMRT) sequencing and 'Oxford Nanopore Technologies' (ONT) nanopore sequencing, which were commercially released in 2011 and 2014, respectively. These two technologies currently dominate the long-read sequencing market (Amarasinghe *et al.*, 2020). The most striking improvement of TGS in comparison with NGS is that they can generate continuous sequences ranging from 10 kilobases to several megabases in length directly from native DNA (Logsdon *et al.*, 2020). When the two long-read sequencing technologies were first introduced, the major drawback was the high error rate, with ~13% for PacBio 'single-pass' sequences and ~15% for ONT raw sequences (Quail *et al.*, 2012, Jain *et al.*, 2017). Currently, due to developments in sequencing chemistry and improvements in DNA preparation, each of PacBio and ONT technology can produce different types of long reads differing in both length and accuracy. The most recent type of PacBio data is the High-fidelity (HiFi) read generated using circular consensus sequencing (CCS) mode, which is both long in length (greater than 10Kb) and high in accuracy (less than 0.1% error rate) (https://www.pacb.com/technology/hifi-sequencing/, accessed November 24, 2022). For ONT, one type of data is the ultra-long read, which typically has read N50s > 50Kb and can reach more than 4Mb (Logsdon *et al.*, 2020). The latest update of ONT duplex data can produce reads with only 0.1% error rate (https://nanoporetech.com/accuracy, accessed November 24, 2022).

**Genome assembly application of DNA sequencing**

One of the key applications of DNA sequencing is *de novo* genome assembly, which computationally produces a representation of the original intact chromosome sequences by putting back together random sampled DNA fragments without any prior knowledge about composition, layout or length of the source DNA (Jung *et al.*, 2019). This process usually requires high depth of sequencing data to be able to cover the whole genome. Sanger sequencing is not a good option for *de novo* genome assembly because of its low output volume and high cost. In comparison, the NGS is more suitable for *de novo* genome assembly. Indeed, the massively parallel NGS technologies together with various short-read assembly algorithms lead to dramatically increased numbers of *de novo* assemblies. The most popular short-read assembly algorithm is the de Bruijn Graph construction, which splits all reads into k-mers that are connected by overlapping prefix and suffix (k-1)-mers (Miller *et al.*, 2010, Nagarajan and Pop, 2013, Sohn and Nam, 2018). After the graph is constructed, an optimal path is identified (Compeau *et al.*, 2011). Subsequently, the optimal path is transformed into contig sequences. However, short-read assembly often results in highly fragmented contigs and incomplete genome sequences due to the weakness of NGS reads to resolve repeats and duplications. In comparison, TGS technologies enable the generation of highly contiguous genome assemblies with unprecedented accuracy, which can span across complex repeat regions (Jung *et al.*, 2019). Overlap-Layout-Consensus (OLC) is the best suited approach for long-read *de novo* assembly, which includes three consecutive steps: 1) overlap graph construction and potentially overlapping reads searching, 2) reads merging and graph simplification, and 3) DNA sequence deriving and error correction (Jung *et al.*, 2019). Although long-read assemblies by far surpass short-read assemblies in contiguity and completeness, solely long reads are still not sufficient to resolve complex genomes at the chromosome level. Usually, additional long-range information is required to create super-scaffolds or chromosome-scale genome assemblies. One of such technology currently commonly used is Bionano Genomics optical mapping, which uses fluorescently labelled enzymes to produce fingerprints of DNA fragments of multiple hundreds kb by imaging the locations of the restriction sites under light microscopes (Jiao *et al.*, 2017, Jiao and Schneeberger, 2017). The labelled molecules can be assembled into genome-wide maps that serve as the skeleton to order and orient contig sequences. Bionano Genomics' new Direct Label and Stain (DLS) technology uses Direct Labeling Enzyme 1 (DLE-1) to attach a single fluorophore to specific sequence motifs and does not make sequence-specific nicks via nicking endonucleases. Thus, this method does not damage DNA molecules at specific sites (Deschamps *et al.*, 2018). The DLS-labeled molecules are much longer than those labelled via the endonuclease approach, with the longest ones becoming larger than 2 Mbp. Another technology is Hi-C, which generates genome-wide libraries from originally close-by loci in the nucleus using a proximity ligation approach (Lieberman-Aiden *et al.*, 2009, Logsdon *et al.*, 2020). The long-range information between pairs of loci provided by Hi-C sequencing data can reach tens of megabases apart on the same chromosome (Ghurye *et al.*, 2019, Logsdon *et al.*, 2020). Currently, the combination of long-read sequencing and long-range scaffolding technologies becomes a common standard to generate high quality genome assemblies (Rousseau-Gueutin *et al.*, 2020).

### Genome assemblies in plants and *B. oleracea*

In 2000, the first plant genome sequence of *A. thaliana* was released (Arabidopsis Genome Initiative, 2000). Initially, it was difficult to assemble plant genomes using short reads, especially those plant species with large and repeat-rich genomes and high levels of ploidy. However, due to advances in sequencing technologies and computational algorithms, the field of plant genome sequencing has grown rapidly in the past 20 years, with genomes of more than 800 species being assembled to date (Marks *et al.*, 2021). Among these plant genomes, sugar pine (*Pinus lambertiana*) has the largest genome size, with an assembly size of 27.6 Gb (Stevens *et al.*, 2016). The *Brassica* species have relatively medium-sized genomes. For example, the genome size of *B. oleracea* was estimated to be around 600 Mb based on flow cytometry analysis (Guo *et al.*, 2020). The first *B. oleracea* draft genome was released in 2014, a cabbage line 02-12 with an assembly size of 515.4 Mb (Liu *et al.*, 2014a). Only about one month later, another draft genome was released, the doubled haploid *B. oleracea* annual kale-like type TO1000DH, with an assembly size of 488.6 Mb (Parkin *et al.*, 2014a). Both the two genomes were generated using NGS technologies and genetic maps. In 2016, a *B. oleracea* pan-genome study based on a reference-guided approach using short-read sequencing technology demonstrates high levels of variation between different morphotypes, with nearly 20% of genes affected by presence/absence variation (Golicz *et al.*, 2016b). Recently, several other *B. oleracea* reference genomes with higher quality have been released, which were primarily created by combining high-coverage long-read sequencing data with long-range scaffolding information, including three cabbage, two cauliflower and one broccoli lines (Belser *et al.*, 2018, Sun *et al.*, 2019a, Cai *et al.*, 2020, Lv *et al.*, 2020, Guo *et al.*, 2021).

## Meiotic recombination

Meiosis is a specialized cell division process that is required for sexually reproducing organisms to generate gametes for fertilization (Mercier *et al.*, 2015, Wang and Copenhaver, 2018). This process consists of a single round of DNA replication followed by two rounds of nuclear division (Meiosis I and Miosis II). Both the two divisions include four phases: prophase, metaphase, telophase and anaphase (Wang and Copenhaver, 2018). There is no interphase between Meiosis I and II. In Meiosis I, pairs of homologous chromosomes segregate. In Meiosis II, sister chromatids are separated into four haploid cells. Meiotic recombination occurs during the first meiotic division, which ensures genome integrity and stability and generates genetic diversity (Pelé *et al.*, 2018). Recombination is initiated in prophase I stage of meiosis by the formation of a large number of DNA Double Strand Breaks (DSBs) on bivalents, which are induced by SPO11 protein (Osman *et al.*, 2011, Demirci, 2021). A minority of DSB repairs on the nonsister homologous chromatids form crossovers (COs), whereas the majority of these DSB repairs lead to noncrossovers (NCOs) (Mercier *et al.*, 2015, Li *et al.*, 2019a). COs involve reciprocal exchanges of large DNA fragments between the homologous chromatids, with two of the four chromatids being modified. NCOs copy a small section of genetic material from the intact donating chromosome to the broken chromosome, with only one of the four chromatids being modified, often resulting in small changes (Li *et al.*, 2019a).

### Pathways of CO and NCO formation

Several known pathways are involved in the formation of COs and NCOs (Fig. 2) (Mercier *et al.*, 2015, Wang and Copenhaver, 2018, Demirci, 2021). The processing of DSB ends produces 3' single-strand DNA ends, which search either one of the two homologous chromatids (inter-homologous invasion) or the intact sister chromatid (inter-sister invasion) as templates for repair. These invasions promote the formation of a D-loop. In the first ZMM pathway, after a successful inter-homologous invasion, DNA synthesis using the second 3' end, followed by ligation, produces the double Holliday junction (dHJ) intermediate. These dHJs can be resolved into the interfering Class I COs, with the occurrence of a CO preventing the formation of close-by COs. In a second MUS81 pathway, after successful inter-homologous invasion, unknown recombination intermediates generate either non-interfering Class II COs or they are resolved into NCOs via Synthesis-dependent strand annealing (SDSA) pathway. This third pathway also produces NCO from the ejected inter-homologous invading strands after D-loop formation. In SDSA pathway, the chromatids are repaired via DNA synthesis relying on  the sister chromatid as a template. In many organisms, these two types of COs coexist, with Class I contributing 85-90% of all COs. Class I COs depend on ZMM complex (*Saccharomyces cerevisiae* Zip1-4, Msh4-5, and Mer3; HEI10 is the Arabidopsis homolog of Zip2) besides MLH1 and MLH3 proteins (Mezard *et al.*, 2007, Osman *et al.*, 2011, Mercier *et al.*, 2015, Pelé *et al.*, 2017, Durand *et al.*, 2022). Class II COs rely on MUS81 and EME1/MMS4 proteins. NCOs are sometimes accompanied by gene conversion, which describes the nonreciprocal transfer of small genetic material between loci, typically allelic loci, in meiosis (Wang and Copenhaver, 2018).

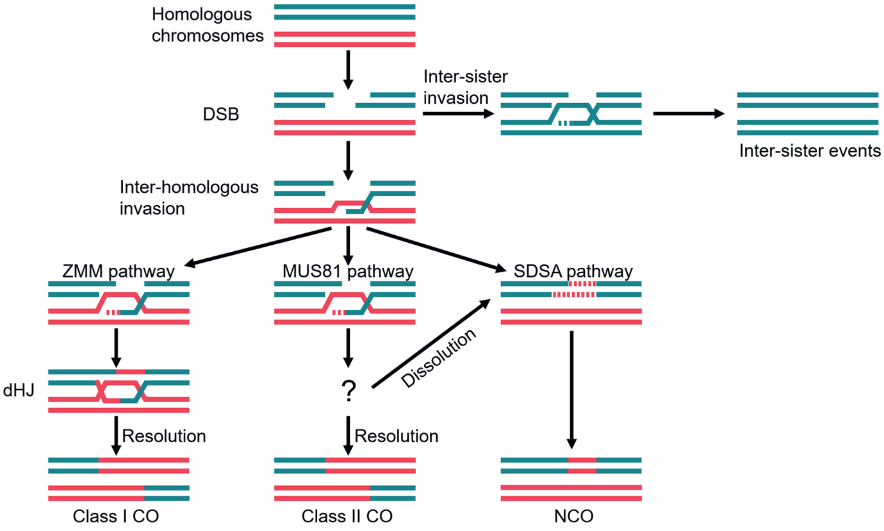


**Fig. 2 A schematic model for meiotic recombination**. Meiotic recombination starts with the formation of double-strand breaks (DSBs). DSB ends are resected to yield 3' single-strand tails, which invade a non-sister chromatid or a sister chromatid to form a D-loop. Inter-homologous intermediates can be processed via different pathways to form either crossovers (COs) or noncrossovers (NCOs). In the ZMM pathway, DNA synthesis (dashed lines) and ligation result in the formation of double Holliday junctions (dHJ), which are primarily resolved as Class I COs that depend on ZMM proteins and are interference sensitive. In the MUS81 pathway, unknown recombination intermediates are resolved to form Class II COs that are non-interference. Alternatively, unknown intermediates can be processed by the synthesis-dependent strand annealing pathway (SDSA), which produces NCOs. This figure was adapted from Mercier *et al.,* Wang *et al.,* and Demirci *et al.* (Mercier *et al.*, 2015, Wang and Copenhaver, 2018, Demirci, 2021).

## Approaches to detect meiotic COs

Meiotic COs establish a physical connection between two chromatids of homologous chromosomes, known as chiasmata, which is required for balanced segregation of chromosomes at meiosis. The absence of at least one CO per chromosome pair results in reduced fertility and aneuploidy (Hunter, 2015, Fernandes *et al.*, 2018). In addition, COs drive genetic exchange between homologous chromosomes, which generates novel allelic combinations and genetic diversity among offspring (Fernandes *et al.*, 2018, Wang and Copenhaver, 2018). The first step towards understanding the underlying mechanism of CO formation is to accurately detect COs. To date, several approaches have been developed to measure the location and frequency of COs in plants, including immunocytological analyses, fluorescence-tagged lines (FTLs) systems and sequencing-based methods (reviewed in Kim and Choi, 2022) (Kim and Choi, 2022). COs can be cytologically detected by counting the number of chiasmata from pollen mother cells (PMCs) based on the bivalent shape using fluorescence *in situ* hybridization (FISH) staining of specific regions or whole chromosomes (Moran *et al.*, 2001, Sanchez-Moran *et al.*, 2002, López *et al.*, 2012). Also, immunostaining of single or combined proteins specifically involved in CO formation can be used to count the number of COs according to immunostained foci (Ziolkowski *et al.*, 2017, Lloyd *et al.*, 2018, Modliszewski *et al.*, 2018, Capilla-Pérez *et al.*, 2021, France *et al.*, 2021, Nageswaran *et al.*, 2021). In a higher throughput manner, segregation assays with transfer DNAs that express fluorescent proteins in seeds or pollen have been extensively developed to detect CO frequency, with the availability of genome-wide sets of seed and pollen fluorescence-tagged lines (FTLs) (Melamed-Bessudo *et al.*, 2005, Francis *et al.*, 2007, Berchowitz and Copenhaver, 2008, Wu *et al.*, 2015). Furthermore, COs can be detected by resequencing individuals of a segregating population, such as F2, recombinant inbred lines (RILs), backcross (BC), etc, followed by extensive bioinformatics analyses to reconstruct the mosaic genetic material of their parents. Recently, CO detection from a pool of pollen DNA from F1 hybrids was developed using Illumina sequencing of 10X Genomics linked-reads (Dréau *et al.*, 2019, Sun *et al.*, 2019b, Rommel Fuentes *et al.*, 2020).

## Tight regulations of CO number and distribution

Meiotic CO number and distribution are both tightly regulated. Besides the obligate occurrence of one CO per chromosome pair, rarely more than three COs are found in most eukaryotes regardless of the chromosome size (Mercier *et al.*, 2015). Both CO interference and anti-CO factors could participate in limiting the number of COs. Indeed, three anti-CO pathways have been identified in *A. thaliana* using forward screen approaches, which rely on the activity of 1) the FANCM helicase and its cofactors (Crismani *et al.*, 2012, Girard *et al.*, 2014), 2) the BLM/Sgs1 helicase homologs RECQ4A and RECQ4B and the associated proteins TOP3α and RMI1 (Séguéla-Arnaud *et al.*, 2015, Séguéla-Arnaud *et al.*, 2017) and 3) the FIGL1 AAA-ATPase (Girard *et al.*, 2015), respectively. Moreover, COs are not uniformly distributed along the chromosomes in almost all studied species (Mézard *et al.*, 2015, Kianian *et al.*, 2018). On a smaller scale, most COs preferentially cluster in small genomic regions of a few kilobases, known as recombination hotspots (Petes, 2001, Mézard *et al.*, 2015). At a larger scale, COs are concentrated in distal regions and always suppressed in centromeric and pericentromeric regions (Marand *et al.*, 2017, Dreissig *et al.*, 2019, Raz *et al.*, 2021). In some plants, like bread wheat, maize and barley, CO distribution gradually decreases from telomeres to centromeres (Liu *et al.*, 2009, Saintenac *et al.*, 2011, Higgins *et al.*, 2012). These observations could be connected with different genomic and epigenomic features. It has been found that the occurrence of COs positively correlates with gene density that is generally high in distal regions and negatively with TE density, which is highest at and next to centromeres (Wu *et al.*, 2003, Erayman *et al.*, 2004, Anderson *et al.*, 2006, Dooner and He, 2008). Small-scale sequence divergence, such as SNPs and InDels, and large structural variations could also locally affect CO formation (Rowan *et al.*, 2019, Boideau *et al.*, 2022, Lian *et al.*, 2022b). With regard to epigenetic factors, it has been reported that CO occurrence correlates with low levels of DNA methylation, low nucleosome density and enrichment in specific histone marks (*i.e.* K3K9me2) (Choulet *et al.*, 2014, Swagatika and Tomar, 2016, Li *et al.*, 2019b, Boideau *et al.*, 2022).

## CO interference and heterochiasmy

Two interesting observations of meiotic recombination are CO interference and heterochiasmy. As already mentioned, Class I COs are subject to interference, with the presence of one CO preventing the formation of additional CO nearby along the same chromosome pair. The so-called phenomenon of heterochiasmy refers to marked differences between male and female in CO rate and distribution (Lenormand and Dutheil, 2005, Dluzewska *et al.*, 2018). Both CO interference and heterochiasmy were discovered more than one century ago (Muller, 1916a, Zickler and Kleckner, 2016). However, the underlying mechanism of these two observations still remains elusive (Dluzewska *et al.*, 2018, Capilla-Pérez *et al.*, 2021). COs are formed in the context of synaptonemal complexes (SC), structures that zip homologous chromosomes together during meiosis. Recent studies suggest that SC imposes CO interference and heterochiasmy in Arabidopsis (Capilla-Pérez *et al.*, 2021).

## Objectives and outline of the thesis

In this thesis, we aim to explore genomic and genetic features of the highly diverse *B. oleracea* species. We generated a core collection representing the *B. oleracea* germplasm for population genetic studies and selected five genotypes representing important crops for in-depth genomics studies. The ultimate goal is to provide valuable genomic resources and insights for improvement of economically important *Brassica* crops. We produce massive DNA and RNA sequencing data and metabolite data for diverse *B. oleracea* crops, with four concrete objectives:

- To investigate genetic diversity and domestication history of *B. oleracea* morphotypes;
- To generate high-quality reference genomes of five diverse *B. oleracea* morphotypes;
- To study meiotic recombination in crosses between the five parental *B. oleracea* morphotypes;
- To study genetic regulation underlying GSL profile variation in *B. oleracea*.

In **Chapter 2**, we generated a collection of 912 globally distributed accessions representing ten morphotypes of *B. oleracea*, wild *B. oleracea* accessions and nine related wild relatives (C9 *Brassica* species). These samples were genotyped using Sequence-Based Genotyping method, which resulted in high-quality SNP markers for population genetics study. We investigated genetic diversity, genealogical relationship, population structure and domestication history of these *B. oleracea* morphotypes.

In **Chapter 3**, we *de novo* assembled high-quality reference genomes for five different *B. oleracea* morphotypes, including broccoli, cauliflower, kale, kohlrabi and white cabbage. The genome assemblies were created using the state-of-the-art long-read sequencing (ONT) and long-range scaffolding (Bionano DLS optical maps) technologies. A catalogue of genome-wide SNPs and structural variations was generated based on comparative genomics analysis. We then performed a pan-genome analysis using our five plus four previously reported *B. oleracea* genomes. Furthermore, we compared evolutionary patterns with regard to intact LTR-RTs accumulation and WGT-derived gene loss between this *B. oleracea* pan-genome and the published *B. rapa* pan-genome.

In **Chapter 4**, we investigated meiotic homologous recombination in *B. oleracea*. We constructed mapping populations from four-way crosses (FwC populations) between the five parental *B. oleracea* morphotypes, the genomes of which were *de novo* assembled in **Chapter 3**, with each two F1s being reciprocally crossed per population. We re-sequenced these mapping populations using Illumina technology and hereafter we detected meiotic COs. We then compared recombination rates and landscapes among different crosses/genetic backgrounds. Moreover, combined with the *de novo* genome sequences, we investigated the effect of various genomic features on meiotic CO formation. In addition, we studied sex difference of CO rates and landscapes for each genetic background.

In **Chapter 5**, we profiled diverse GSL compounds in both roots, leaves and the edible parts of the five *B. oleracea* morphotypes (the same accessions that were used in **Chapter 3**) to evaluate the GSL variation among different tissues and different morphotypes. We also generated mRNA-Seq data for the same plant materials to study gene expression levels of all GSL related genes, which were identified in the corresponding five *B. oleracea* genomes based on homology to Arabidopsis GSL genes. The correlation between gene expression level and GSL relative abundance was analysed. Moreover, we investigated genomic variations among the five genomes for two important GSL genes (*AOP2* and *MAM3*), which probably explain some of the observed GSL differences.

In **Chapter 6**, I further discuss the results obtained in this thesis in a broader context.

# Chapter 2

## Evidence for two domestication lineages supporting a middle-eastern origin for *Brassica oleracea* crops from diversified kale populations

**Chengcheng Cai[1,2], Johan Bucher[1], Freek T. Bakker[3] and Guusje Bonnema[1,2,*]**

[1] Plant Breeding, Wageningen University and Research, 6708 PB, Wageningen, The Netherlands

[2] Graduate School Experimental Plant Sciences, Wageningen University and Research, 6708 PB, Wageningen, The Netherlands

[3] Biosystematics Group, Wageningen University and Research, 6708 PB, Wageningen, The Netherlands

[*] Corresponding author

**Published in *Horticulture Research* (2022).**

**DOI: https://doi.org/10.1093/hr/uhac033**

## Abstract

*Brassica oleracea* displays enormous phenotypic variation, including vegetables like cabbage, broccoli, cauliflower, kohlrabi, kales etc. Its domestication has not been clarified, despite several genetic studies and investigations of ancient literature. We used 14,152 high-quality SNP markers for population genetic studies and species-tree estimation (treating morphotypes as 'species') using SVD-quartets coalescent-modelling of a collection of 912 globally distributed accessions representing ten morphotypes of *B. oleracea*, wild *B. oleracea* accessions and nine related C9 Brassica species. Our genealogical tree provided evidence for two domestication lineages, the 'leafy head' lineage (LHL) and the 'arrested inflorescence' lineage (AIL). It also showed that kales are polyphyletic with regards to *B. oleracea* morphotypes, which fits ancient literature describing highly diverse kale types at around 400BC. The SVD-quartets species tree topology showed that different kale clades are sister to either the LHL or the AIL. Cabbages from the middle-east formed the first-branching cabbage-clade, supporting the hypothesis that cabbage domestication started in the middle-east, which is confirmed by archeological evidence and historic writings. We hypothesize that cabbages and cauliflowers stem from kales introduced from Western Europe to the middle-east, possibly transported with the tin-trade routes in the Bronze age, to be re-introduced later into Europe. Cauliflower is the least diverse morphotype showing strong genetic differentiation with other morphotypes except broccoli, suggesting a strong genetic bottleneck. Genetic diversity reduced from landraces to modern hybrids for almost all morphotypes. This comprehensive Brassica C-group germplasm collection provides valuable genetic resources and a sound basis for *B. oleracea* breeding.

**Key words**: *Brassica oleracea*, arrested inflorescence lineage, leafy head lineage, diversified kales, domestication, genetic diversity, population genomics, tin-trade route

## Introduction

*B. oleracea* is an important vegetable and fodder crop species and exhibits enormous diversity in its appearances, including the leafy heading morphotypes var. *capitata* (cabbage), the typical curd subspecies with large arrested inflorescences, including var. *botrytis* (cauliflower) and var. *italica* (broccoli), the kohlrabi's with their tuberous stems (var. *gongylodes*), the Brussels sprouts with their axillary buds (var. *gemmifera*), and other leafy vegetable types with different shapes, such as var. *viridis* (Collard green), var. *alboglabra* (Chinese kale) (Kole and Henry, 2010, Bonnema *et al.*, 2011), var. *costata* (Tronchuda kale) (Dias, 2012) as well as other types of kale, such as *var. acephala* (bore and curly kale, marrow stem kale, etc). In addition, var. *acephela* (ornamental kale) has been bred by intercrossing different morphotypes (Guo *et al.*, 2019). Different *B. oleracea* morphotypes are extensively cultivated, with several crops cultivated almost worldwide (like cabbages and cauliflowers), while others in specific countries only depending on local preferences (like Tronchuda's in Portugal and Collards in the USA). Despite this enormous diversity, *B. oleracea* is still considered as one species and morphotypes can be easily interbred.

Wild *B. oleracea* can be found along the Atlantic and Mediterranean coasts, all the way from Norway to England and France down to Greece. Wild *B. oleracea* plants form a strong stem with large waxy leaves, and only flower after several winters (Crozier, 1891, Kole and Henry, 2010, Bonnema *et al.*, 2011). *B. oleracea* and its closely-related wild relatives belong to the so-called 'C-genome group', which is characterized by the possession of nine chromosome pairs (Maggioni, 2015). These wild 'C9 species', which probably form a monophyletic group, mainly occur in the Mediterranean region, particularly in Sicily (Snogerup *et al.*, 1990, Tribulato *et al.*, 2017). In his doctoral thesis, Maggioni summarized the above-mentioned origin and morphological characteristics of these wild species (Maggioni, 2015). In short, they are characterized by hairy leaves, petiolated or with wings, and/or yellow flowers, such as in *Brassica incana*, *Brassica villosa*, *Brassica macrocarpa* and *Brassica rupestris*. Phenotypic diversity of wild relatives of *B. oleracea* has been assessed within the European Cooperative Program for Plant genetic Resources (ECPGR). A few hybridization experiments have been performed to investigate relationships between wild 'C9 species' and cultivated forms of *B. oleracea* (Kianian and Quiros, 1992, Bothmer *et al.*, 1995). Kianian and Quiros (1992) (Kianian and Quiros, 1992) analysed fertility and meiotic chromosome behaviour from 172 intra- and interspecific hybrids of *B. oleracea* crops and wild 'C9 species'. Based on their experiments, fertile hybrids can be produced through crosses of *Brassica alboglabra*, *Brassica bourgeaui*, *Brassica cretica*, *Brassica montana* to *B. oleracea* crops. Crosses of *Brassica incana*, *Brassica insularis* and *Brassica rupestris* to *B. oleracea* crops produced semi-sterile progenies, which was associated with abnormal meiotic behaviour. Crossing experiments from von Bothmer et al. (1995) (Bothmer *et al.*, 1995) also indicated possibly frequent introgression between wild 'C9 species' and *B. oleracea* crops. In a recent study, Mabry *et al.* (2021) investigated the evolutionary history of wild, feral and domesticated *B. oleracea*, including nine C9 species which they refer to as 'progenitor species'. The authors identify the Aegean endemic *B. cretica* as closest living relative of *B. oleracea* (Mabry *et al.*).

Domestication of *B. oleracea* dates back to as early as around 400 BC, based on for example descriptions of kales by the Greek scholar Theophrastrus (370-285 BC). In older writings (800-600 BC) the kales were not yet mentioned (Maggioni *et al.*, 2014, Maggioni *et al.*, 2018). Northwestern Europe and the Mediterranean/Middle east regions are hypothetical ancestral areas for domesticated *B. oleracea*, of which the latter obtains more weight from ancient literature (Kole and Henry, 2010, Maggioni *et al.*, 2018). Also Mabry *et al.* (2021) see the identification of  the Aegean endemic *B. cretica* as closest living relative of *B. oleracea* as support for a Eastern Mediterranean domestication origin for *B. oleracea*. Gomez-Campo and Prakash (Gómez-Campo and Prakash, 1999), suggest several possible domestication routes, one of them being that kales were the earliest cultivated forms of *B. oleracea*, cultivated along the Atlantic and Mediterranean coasts by the Celts, from where they were brought to the East Mediterranean region (first and second millenia BC) to become domesticated. John Gerard (1597) describes in his herbal that Theophrastus, and the Romans Cato (234-149 BC) and Pliny the Elder (23-79 AD) already described "wild and tame coles", and distinguished several "coleworts" (kales). These included the smooth, great, broad-leaved, with a big stalk type, the ruffed type and types with little stalks that are tender and "very much biting". One hypothesis is that these diverse kales ("coleworts")  were transported along tin trade routes (Bronze Age, around 3300-1000 BC) from Spain and the British islands to the Phoenicians (present-day Libanon), who referred to *B. oleracea* crops as 'Krambe' (Penhallurick, 2008, Berger *et al.*, 2019). The name Krambe was later used by Linnaeus for the genus *Crambe*, which is closely-related to Brassica (Couvreur *et al.*, 2010), and is known for its coastal distribution and fleshy cabbage-like leaves. If indeed kales were introduced to the middle-east, it is also possible that domestication of cabbages and cauliflowers initiated there. In addition to kales, cabbages ("coles") are also first mentioned around the first century AD. In Maggioni (2018) we can read that Pliny the Elder (23-79 AD) already describes coles that "grow so big that a poor man's table would not be large enough to hold it". Lucius Junius Moderatus Columella (4–ca. 70 AD), born in Spain, but a military officer in the Roman Legion and later settled as landowner in Italy, described many different "caules" based on where they were grown. He also mentions the sprouting types (cymae), which were mentioned by Pliny the Elder too. These refer to axillary buds that are tender, and indicate that apical dominance was less pronounced as it is in modern cultivars. Around the same times, cauliflowers are described, as *Brassica cypria* in Latin, and as 'cauliflore' in Italian, which seems to agree with *Brassica pompeiana* described by Pliny the Elder (Maggioni *et al.*, 2018). The Spanish Arabian author Ibn-Al-Awan (c.1140), was the first to distinguish heading and sprouting cauliflowers. He named cauliflower 'quarnabit', the present Arabic name for cauliflower, suggesting Syria as center of origin, while the herbalist Dodonaeus (1578) suggested Cyprus. Crisp (1982) hypothesized that cauliflowers are evolved from broccoli's based on crossing experiments. In a recent study by Guo et al. (2021), this was also suggested based on their inflorescence phenotypes and causal mutations (Guo *et al.*, 2021). Dale-Champ (1586) writes that cauliflowers have evolved between 400-600 BC and are believed to have diversified in the Eastern Mediterranean, and from there introduced to Italy (Gómez-Campo and Prakash, 1999). Maggioni et al (2018) however did not find any mentioning of 'cauliflower' till the first century. Dodonaeus (1578) and also John Gerard (1597) describe and illustrate in detailed drawings of white, red and savoy cabbages, cauliflower and kale but kohlrabi is not mentioned (Zeven, 1996).

To date, several genetic diversity and population structure studies have been performed and published for domesticated *B. oleracea* (van Hintum *et al.*, 2007, Izzah *et al.*, 2013, Cheng *et al.*, 2016b, El-Esawi *et al.*, 2016, Stansell *et al.*, 2018, Mabry *et al.*). Several studies are limited by low numbers of markers (van Hintum *et al.*, 2007, Izzah *et al.*, 2013, El-Esawi *et al.*, 2016) and, more importantly, hardly any studies include all described *B. oleracea* morphotypes, and the ones that are included are often represented by low numbers of accessions (Farnham *et al.*, 2008, Pelc *et al.*, 2015, Stansell *et al.*, 2018). What generally lacks in these studies is data on genetic comparisons between modern hybrid and old landrace accessions. Recently, Cheng et al. resequenced 119 *B. oleracea* and 199 *Brassica rapa* accessions representing seven resp. 11 morphotypes, to study their genetic diversity and genealogical relationships (Cheng *et al.*, 2016b). They showed that the cauliflower and broccoli accessions formed a separate cluster, with a long branch length separating it from a cluster comprising cabbage, ornamental kale, Brussels sprouts and kohlrabi accessions. Only few published studies also include *B. oleracea* wild C9 relatives which can be intercrossed with domesticated *B. oleracea* crops (Warwick and Sauder, 2005, Mabry *et al.*).

Our aim was to study genealogical relationships, population structure and domestication history of *B. oleracea* morphotypes, based on dense lineage- and character-sampling. For this purpose we generated a collection of 912 accessions representing the majority of morphotypes of *B. oleracea*, including both modern hybrids and old landraces with worldwide geographical origins, and its wild relatives (C9 species). Using genotypic data (14,152 SNPs), we estimated nucleotide diversity in each group and compared the differentiation between groups. We compared genetic diversity between genebank accessions and modern hybrids as well as between different subgroups such as ecotypes within morphotypes. We estimated species tree topology using coalescent-based population genetics modelling, treating morphotypes as ‘species’. We provide evidence for two main lineages within the cultivated *B. oleracea*’s, the ‘leafy head lineage’ (LHL; cabbages, collards and ornamentals) and the ‘arrested inflorescence lineage’ (AIL; cauliflower and broccoli). We show that most cauliflower accessions form a monophyletic group, in a position, probably most-derived of all morphotypes and hypothesize that cauliflower domestication went through a strong bottleneck.

## Results

### Global geographic distribution of Brassica germplasm

A total of 912 accessions, representing 10 *B. oleracea* morphotypes, wild *B. oleracea* and nine wild C9 species, were selected for genetic diversity analysis (Table 1, Table S1). This germplasm set consists of 377 modern hybrid accessions and 531 accessions with a global geographic representation. As shown in Table S1, germplasm is selected from ~53 countries with the majority from Europe. Broccoli and cauliflower materials are mainly obtained from Italy, the UK and the Netherlands. The collection of heading cabbage has various geographical origins, such as the Netherlands, Germany, the UK, Macedonia, Russia as well as other countries, like Turkey, reflecting the fact that heading cabbages are adapted to a wide range of climatic zones. The Collard green collection includes materials from the USA and Middle-East countries (Turkey and Syria). Most of Brussels sprouts accessions are obtained from the Netherlands, Denmark, Germany and France and the majority of Tronchuda accessions are collected from Portugal and Spain. Wild *B. oleracea* accessions are mainly collected from the UK and wild C9 species from Italy.

**Table 1** Summary of all the accessions used in this study.

| Morphotype | genebank | hybrids | others | Total |
|---|---|---|---|---|
| **Broccoli** | **47** | **52** | **1** | **100** |
| *summer-autumn* | *16* | *35* | *0* | 51 |
| *unkown* | *16* | *1* | *1* | 18 |
| *winter* | *15* | *16* | *0* | 31 |
| **Cauliflower** | **84** | **138** | **1** | **223** |
| *romanesco* | *5* | *4* | *0* | 9 |
| *summer-autumn* | *28* | *84* | *0* | 112 |
| *tropical* | *9* | *13* | *0* | 22 |
| *unkown* | *24* | *2* | *1* | 27 |
| *winter* | *18* | *35* | *0* | 53 |
| Collard Green | 20 | 0 | 0 | 20 |
| **Heading cabbage** | **180** | **130** | **1** | **311** |
| *pointed* | *4* | *6* | *0* | 10 |
| *red* | *23* | *21* | *0* | 44 |
| *savoy* | *40* | *12* | *0* | 52 |
| *unkown* | *5* | *5* | *0* | 10 |
| *white* | *108* | *86* | *1* | 195 |
| Chinese Kale | 10 | 1 | 0 | 11 |
| **Kale** | **29** | **5** | **0** | **34** |
| *kale* | *8* | *0* | *0* | 8 |
| *bore-curly* | *11* | *4* | *0* | 15 |
| *marrow-stem* | *10* | *1* | *0* | 11 |
| Kohlrabi | 34 | 17 | 1 | 52 |
| Ornamental | 2 | 24 | 0 | 26 |
| Sprouts | 39 | 10 | 0 | 49 |
| Tronchuda | 33 | 0 | 0 | 33 |
| Wild oleracea | 20 | 0 | 0 | 20 |
| Wild C9 | 33 | 0 | 0 | 33 |
| **Total** | **531** | **377** | **4** | **912** |

### Pairwise genetic distance matrix reveals lower genetic distance within morphotypes and higher genetic variations in genebank accessions compared to hybrids

Overall, pairwise genetic distances between morphotypes were larger than within morphotypes (Fig. 1). Interestingly, the lowest pairwise genetic distance within morphotypes was observed in the cauliflower group, illustrating their very low genetic diversity. In addition, the genetic variation in modern hybrid accessions was much lower than that of *B. oleracea* genebank accessions for most morphotypes. However, this was not obvious for cauliflower, with overall low diversity (average normalized genetic distance of 0.37) in both genebank and hybrid accessions.

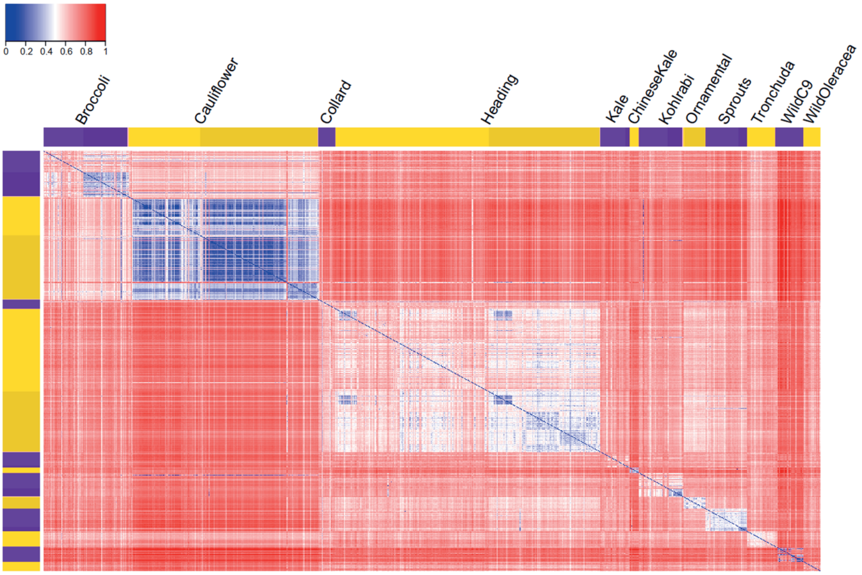


**Fig. 1 Heatmap showing genetic distance matrix between 912 accessions.** For each group, light color indicates genebank accessions, and dark color indicates modern hybrid accessions.

Based on pairwise genetic distance, we also compared genetic variation between the different ecotypes and cultivar-groups (hereafter referred to as 'varieties') in broccoli, cauliflower, heading cabbage, kale and between the different wild C9 species. Different broccoli ecotypes (summer/autumn, winter and unknown) showed similar levels of genetic variation, therefore the genetic variation between genebank and modern hybrids is larger than that between the different ecotypes (Fig. S2a). However, the cauliflower group exhibited different patterns as differences in genetic variation between ecotypes of cauliflower is greater than between genebank and modern hybrids. Romanesco cauliflower displays ample genetic variation. Variation among winter cauliflower accessions was larger  than summer/autumn and tropical types (Fig. S2b). For heading cabbage we distinguish four varieties (red, white, savoy and pointed). Red cabbage is the least diverse subgroup, with little variation in both genebank- and hybrid accessions(Fig. S2c). We observed that there are two subgroups within both the white cabbage genebank and modern hybrid accessions, which showed genetic differentiation. We further investigated pairwise genetic distance within the wild C9 species group. A few accessions behaved unexpectedly as they differed extensively from their peer accessions. This might be due to incorrect classification of genebank materials (Fig. S2e, Supplementary Notes) .

### Genome-wide diversity comparisons among groups shows that cauliflower is the least diverse morphotype

We compared genome-wide nucleotide diversity ($\pi$), reduction of diversity (ROD) and pairwise population differentiation level ($F_{ST}$) between and within morphotype groups (Table S4). Among all the morphotypes, cauliflower had the lowest mean nucleotide diversity ($\pi_{cau}$=7.13×10$^{-6}$), which is 46% lower than the highest value identified in wild *B. oleracea* ($\pi_{wbo}$=1.32×10$^{-5}$). The nucleotide diversity of kale was slightly lower (6%) than the figure of wild *B. oleracea*.$\pi$decreased from genebank accessions to hybrid accessions in broccoli, cauliflower, heading cabbage, kohlrabi and Brussels sprouts, with the highest ROD detected in broccoli (ROD$_{bro}$=2.24×10$^{-1}$). This is consistent with the finding of pairwise genetic distance even though for cauliflower the differentiation was not obviously shown in the heatmap (Fig. 1 and 2). $F_{ST}$ values between cauliflower and other morphotype groups ranged from strong (0.21 for heading cabbage) to very strong (0.34 for wild C9 species), with the exception of broccoli (0.15; moderate differentiation) (Table 2). Notably, wild *B. oleracea* had very strong differentiation with cauliflower (0.31), strong differentiation with Chinese kale (0.17) while moderate differentiation with all other *B. oleracea* morphotypes. Based on this we conclude that genetic differentiation between cauliflower and the other morphotypes including wild *B. oleracea* was larger than between wild *B. oleracea* and the rest. Wild C9 species had moderate differentiation from wild *B. oleracea*, but had strong or very strong differentiation with other *B. oleracea* morphotypes. $F_{ST}$ analyses revealed little genetic differentiation between summer/autumn and winter broccoli ($F_{ST}$=0.04) (Table S5). Cauliflower ecotypes showed little and moderate differentiation with each other ($F_{ST}$: 0.02~0.07). Red cabbages had moderate differentiation with all other heading cabbage varieties ($F_{ST}$: 0.06~0.09) whereas only little differentiation was present among pointed, savoy and white cabbages (0.03~0.04).

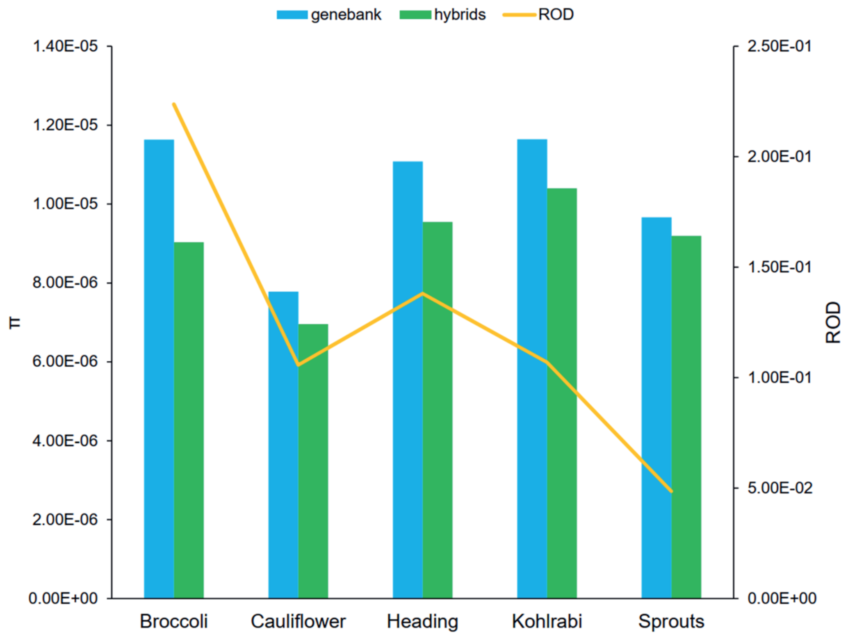


**Fig. 2 Comparison of nucleotide diversity (π) and reduction of diversity (ROD) between genebank and modern hybrid accessions in five groups.** The genebank group was used as the control for ROD calculation in this figure.

**Table 2** Pairwise comparison of $F_{ST}$ values between different morphotype groups.

| Group | Broccoli | Cauliflower | Collard | Heading | Kale | KaleChinese | Kohlrabi | Ornamental | Sprouts | Tronchuda | WildC9 | WildOleracea |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Broccoli** | - | 0.15 | 0.13 | 0.15 | 0.12 | 0.20 | 0.12 | 0.16 | 0.18 | 0.10 | 0.22 | 0.14 |
| **Cauliflower** | - | - | 0.28 | 0.21 | 0.28 | 0.28 | 0.24 | 0.29 | 0.28 | 0.23 | 0.34 | 0.31 |
| **Collard** | - | - | - | 0.05 | 0.04 | 0.17 | 0.08 | 0.09 | 0.13 | 0.07 | 0.18 | 0.04 |
| **Heading** | - | - | - | - | 0.10 | 0.19 | 0.11 | 0.08 | 0.11 | 0.12 | 0.22 | 0.11 |
| **Kale** | - | - | - | - | - | 0.18 | 0.08 | 0.10 | 0.12 | 0.06 | 0.15 | 0.02 |
| **KaleChinese** | - | - | - | - | - | - | 0.19 | 0.26 | 0.27 | 0.15 | 0.29 | 0.17 |
| **Kohlrabi** | - | - | - | - | - | - | - | 0.13 | 0.15 | 0.10 | 0.20 | 0.09 |
| **Ornamental** | - | - | - | - | - | - | - | - | 0.16 | 0.13 | 0.23 | 0.11 |
| **Sprouts** | - | - | - | - | - | - | - | - | - | 0.15 | 0.23 | 0.13 |
| **Tronchuda** | - | - | - | - | - | - | - | - | - | - | 0.19 | 0.07 |
| **WildC9** | - | - | - | - | - | - | - | - | - | - | - | 0.15 |
| **WildOleracea** | - | - | - | - | - | - | - | - | - | - | - | - |

### Genealogical relationships reveal two main lineages for cultivated *B. oleracea* morphotypes

We constructed a *B. oleracea* genealogy tree including 912 accessions (879 *B. oleracea* and 33 wild C9 species), using IQ-TREE maximum likelihood analysis which selected the GTR+F+ASC+R10 model as best-fitting based on BIC. The monophyletic group including the vast majority of wild C9 species samples was set as outgroup. Our analysis revealed that the wild *B. oleracea* and kales (WO1-3 and KA1-2) together formed a paraphyletic group, and the first node following (node A; bootstrap support 60%) represents the start of a "cultivated *B. oleracea* lineage" (named here) (Fig. 3). At node A all the cultivated morphotypes are divided over two main lineages. In one direction this includes several leafy types, that subtend at node B (bs 96%) in the Brussels Sprouts and the "leafy head lineage" (*LHL*), which includes besides cabbages, collards and ornamentals. The other direction includes few kales (KA3), tronchuda's, Chinese kale and kohlrabi, and the "arrested inflorescence lineage" (*AIL*, F, bs 100%) which includes cauliflowers and broccoli's. At node B (bs 96%) the tree subtends into the monophyletic sprouts (SP1 and SP2; bs 100%) and the LHL at node D (bs 100%). Node D subtends the ornamentals (OR, bs 100%) and at node E (bs 33%) the remaining "leafy" clades (CO, HE1-HE7; HE=heading). The Collards are sister to the cabbage clade but lacking support; the cabbage clade splits into HE1 and HE2-7 (bs 80%), and individual clades HE2-H7 are well supported. HE1 is mainly composed of savoy cabbages, HE2-HE6 of white cabbages and HE7 of red and pointed cabbages. Interestingly, HE2 is composed of germplasm accessions from Eastern Mediterranean countries, as do the few Collards that are included in HE2. In the other main lineage node F (bs 100%) demarcates the split between the cauliflowers and broccoli's on the one hand, and the kohlrabi, tronchuda's, Chinese kale and few kales on the other hand. In this very diverse group, the first branchpoint is towards several kales (no support), followed by a branchpoint towards two tronchuda clades (TR1 without support and TR2 (bs 100%)), with nested in it a small Chinese kale clade (CK; bs 100%). The kohlrabi's form a monophyletic group. In the cauliflower-broccoli clades, several broccoli clades (BR1-4) were interspersed with a Romanesco clade (CA1), but support for this placement is low (bs 59%). From here we see a branch towards the most derived clades, the cauliflowers, with CA2 mainly representing winter cauliflowers, CA3 a mixture of winter and summer/autumn cauliflowers, CA4 and CA5 the summer/autumn types, and CA6 and CA7 the tropical and summer/autumn cauliflowers.

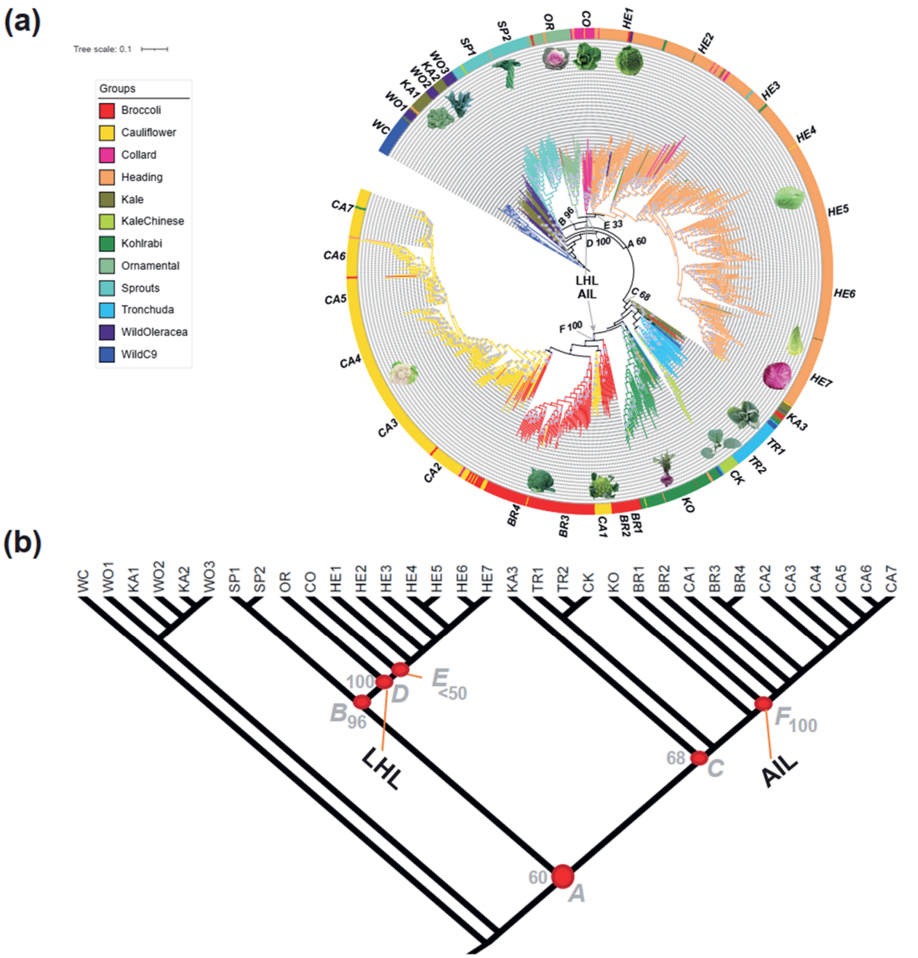


**Fig. 3 Genealogy of major *Brassica oleracea* morphotypes and wild C9 species. (a)** Maximum likelihood tree of *B. oleracea* and wild C9 species accessions inferred from 14,152 high-quality SNPs. Grey arrows indicate the two main cultivated lineages, AIL and LHL, respectively. Labels A-F show selected deep nodes with numbers representing bootstrap values. Labels outside circle refer to clade names (the abbreviations: WC: wild C9 species, WO1-WO3: wild *B. oleracea*, KA1-KA3: kale, SP1-SP2: Brussels sprouts, OR: ornamental kale, CO: collard, HE1-HE7: heading cabbage, TR1-TR2: tronchuda, CK: Chinese kale, KO: kohlrabi, BR1-BR4: broccoli, CA1-CA7: cauliflower). The circled blue dots in clades represent bootstrap values of above 80%. Pictures of morphotypes were obtained from Cheng et al. (Cheng *et al.*, 2016c) and Google. **(b)** A schematic tree depicting the ML tree structure.

### Ancestral character state reconstruction shows that in cabbages the white cabbage variety is ancestral to red, savoy and pointed, while in broccoli and cauliflower the winter ecotypes are ancestral

Ancestral character state reconstruction for heading cabbage suggests that white cabbage is the ancestral state. Both savoy and red cabbage are derived from white cabbage (Fig. 4c). Winter broccoli is suggested as the ancestral state, giving rise to summer/autumn broccoli. It appears that broccoli experiences reverse evolution from summer/autumn to winter broccoli which might be caused by breeding activities (Fig. 4a). The ancestral character state reconstruction also implies winter cauliflower as the ancestral state, with summer/autumn cauliflower derived from winter cauliflower. There is a transition from summer/autumn cauliflower to tropical cauliflower, and also in cauliflower we see this “reverse evolution” as tropical cauliflower then gives rise to summer/autumn cauliflower, followed by winter cauliflower (Fig. 4b). In general, we conclude that the predominant trend for ecotype character is towards summer/autumn from winter and that reversals are rare. We also analysed ancestral state reconstruction with states defined as the different geographical origins. For hybrids this information is difficult to interpret, as hybrids origin leads to breeding company and not necessary to growing area. For the cauliflowers, accessions from Southwest Europe (mainly from Italy, with many romanesco’s) formed the ancestral state (Fig. S3b). For broccoli’s, the ancestral state is formed by accessions from Northwest Europe, with mainly winter broccoli accessions from Great Britain (Fig. S3a). For the heading cabbages, accessions from Southeast Europe, mainly Macedonia (MKD) and Yugoslavia (YUG) formed the ancestral state (Fig. S3c), pointing to an east Mediterranean/near-east origin.

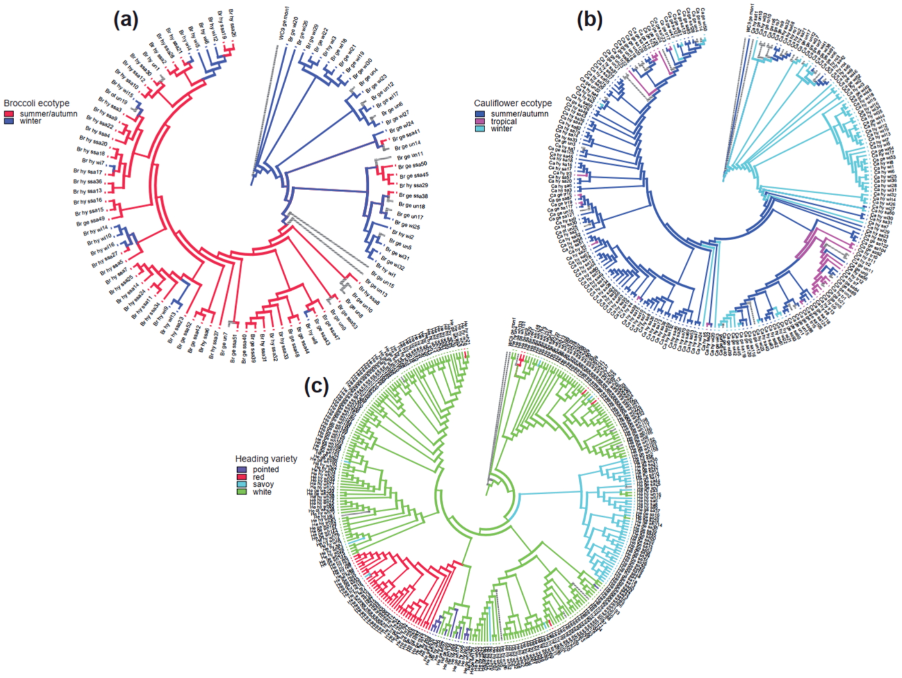


**Fig. 4 Ecotype/variety character evolution for (a) broccoli, (b) cauliflower and (c) heading cabbage.** Different varieties of heading cabbage were treated as the states of variety character, and different ecotypes of broccoli or cauliflower were treated as the states of ecotype character. Grey branches represent accessions with "unknown" state.

## Principal component analysis and population structure support genealogical relationships revealed by ML tree

The genealogical relationships between different groups were also supported by principle component analysis (PCA), with the first two principle components explaining 13.64% and 4.78% total genetic variance, respectively (Fig. 5a). In PC1 AIL with cauliflower and broccoli accessions, is clearly separated from the Brussels sprouts and the LHL (heading cabbage, ornamental and collard accessions). Chinese kale, kohlrabi and Tronchuda accessions, but also the kale accessions are located between these two groups. PC1 also separated wild C9 species into two clusters, which corresponds to the maximum likelihood tree in Fig. S4. PC2 clearly separated wild C9 species from *B. oleracea* wild type and morphotypes. The PCA analysis appeared to corroborate the genealogical tree and population differentiation analyses. Further detailed PCA analysis within morphotype was also in line with genealogical and pairwise genetic distance analyses (Supplementary Notes).

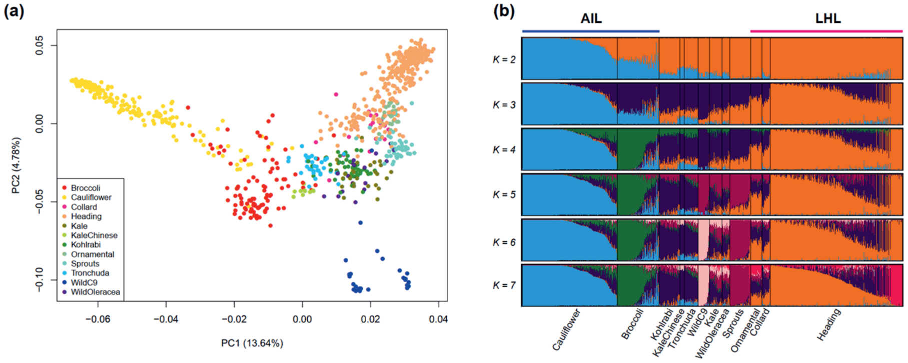


**Fig. 5 PCA and population structure. (a)** PCA plots of 875 accessions using whole genome SNP data. The first two principle components were plotted to visualize the relationships among individuals (accessions) and groups. **(b)** Population structure of major *Brassica oleracea* morphotypes and wild C9 species with different numbers of clusters (*K=2-7*). Each accession is represented by a vertical bar. The different colors represent contributions to the K-groups. The length of each colored segment in the bar quantifies cluster membership.

To further investigate population structure of the 875 *B. oleracea* and wild C9 accessions, the STRUCTURE v2.3.4 (Pritchard *et al.*, 2000) software was utilized to infer the structure of these accessions. We ran the pipeline by gradually increasing the number of clusters (*K*). The*Δk* analysis revealed that *K=4* fits the dataset best (Fig. S5). Interestingly, when *K=2*, cauliflower and not wild *B. oleracea* and wild C9 species, formed one cluster that was clearly separated from other *B. oleracea* morphotypes and these wild C9 species (Fig. 5b). All of the broccoli accessions were admixed and received genetic contributions from cauliflower and other morphotypes. When *K* increased from 2 to 7, kale and wild *B. oleracea*, ornamental kale and Collard green on the one hand  and kohlrabi, Chinese kale and tronchuda on the other hand always had similar membership and thus were clustered together, respectively. At *K=7*, heading cabbage cluster was subdivided into two clusters, with the new cluster formed by 37 red cabbage accessions. Kale and wild *B. oleracea* could not be assigned to independent clusters when *K* ranged from 8 to 12 (Fig. S6). This fits the hypothesis that wild *B. oleracea* and kale are progenitors of different morphotypes, which is also clearly shown from the ML tree (Fig. 3a). Overall, the majority of morphotypes formed a distinct group with gradual increase of the number of clusters. However, more and more accessions within each group become highly admixed, indicating both common ancestry and gene introgression among different groups.

**Species tree and divergence time estimation suggest ancient divergent kale lineages leading to the AIL and LHL lineages**

The SVD-quartets analysis on the ‘Overall’ SNP matrix yielded a tree topology that was not fully congruent with the ML tree. Although the two main lineages AIL and LHL are present (both bs 100%), sprouts are sister to Wildoleracea1 (WO1) and Kale1 (KA1) (bs 91%) and Kale2 and Kale3 (KA2 and KA3) (bs 73%) are sisters in the SVDq topology (Fig. 6a), while in the ML tree WO1 together with WO2 and WO3, and KA1 together with KA2 precede  the complete “cultivated *B. oleracea* lineage”. The position of kohlrabi (KO) is problematic as it is sister to Wildoleracea3 which contradicts its position in the overall ML genealogy.

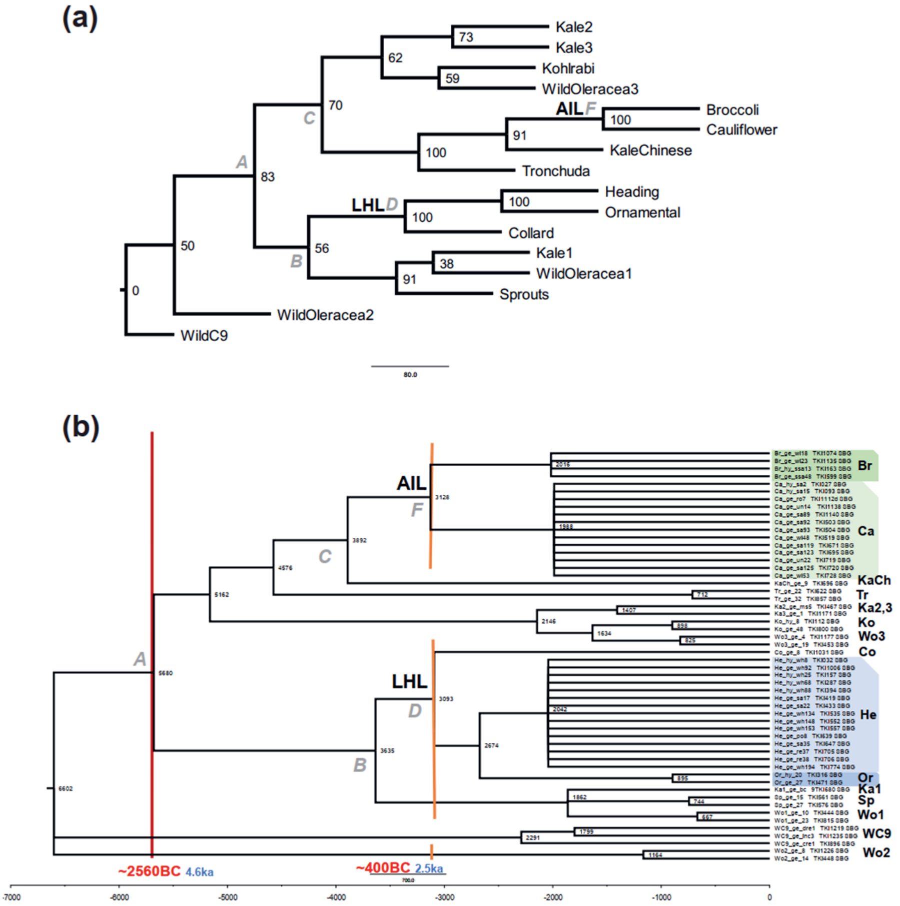


**Fig. 6 *Brassica oleracea* species tree. (a)** SVDquartet analysis of the full 875 terminal SNP matrix; bootstrap values are indicated at the nodes. Grey letters at nodes refer to main clades as distinguished in Fig. 3a – the ML Overall tree. **(b)** Tree topology of the SVDquartets analysis from (a) with nodes with bootstrap values <50% collapsed, enforced on SUB matrix, and arbitrarily ultrametricized in Mesquite (see text). Numbers at nodes and scalebar indicate relative ages. *B. oleracea* morphotypes

are indicated at terminal labels on the right; proposed dating of domestication events based on earliest reports, as well as the timing of tin trades /Bronze Age (see text) are indicated. Grey letters at nodes refer to main clades as distinguished in Fig. 3a – the ML Overall tree. The abbreviations on the right: Br: broccoli, Ca: cauliflower, KaCh: Chinese kale, Tr: tronchuda, Ka1-Ka3: kale, Ko: kohlrabi, Wo1-Wo3: wild *B. oleracea*, Co: collard, He: heading cabbage, Or: ornamental kale, Sp: Brussels sprouts, WC9: wild C9 species.

Given the accuracy of SVD-quartets analysis (Chifman and Kubatko, 2014, Leaché and Oaks, 2017) and its use of coalescent modelling (allowing gene trees to be different) (see Supplementary Notes), we consider it to be more accurate than the ML tree topology. We therefore use the SVDq topology to estimate branch lengths on, in order to enable time estimation of nodes. After ultrametricizing the SVDq-based tree with SNP nucleotide branch lengths (Fig. 6b), we see that AIL (node F) would appear of a similar age as LHL (node D). Tronchuda, Kohlrabi and Chinese kale, but also Sprouts, would appear as younger. The main divergence among Cauliflowers in AIL and among Cabbages in LHL appears 'simultaneously'. Since our approach to dating the nodes is fairly crude (it does not include molecular clock modelling, rate smoothing nor standard deviations) we interpret our relative node-ages as rough estimates and an indication of relative ancestry of lineages. Nevertheless, we did consult the literature on ancient writings for possible timing of first occurrences of, for instance, cauliflower and cabbages, which were mentioned by Theophrastus and Pliny the Elder at around 400BC (or 2.5ka; see Introduction). Applying this date to nodes F and D (which mark the onset of cauliflower/broccoli and heading cabbage diversification breeding, respectively), and assuming clock-like accumulation of SNP's, we find a SNP-rate of 3100 node height / 2500yr = 1.24 SNP yr$^{-1}$. Applying this to node A (the onset of cultivation of *B. oleracea*) we find an age of 5680 node height /1.24 = 4581yr, which is 2560 BC. As diversification within the cauliflower and heading cabbage clades is based on accessions of modern hybrids and old landraces with worldwide geographic origins, we feel the above scenario is accurate /realistic to base our time estimates on.

## Discussion

The *Brassica* C-group germplasm collection of 972 samples we generated and genotyped in this paper is the most comprehensive published so far. We included most described *B. oleracea* subspecies, herein referred to as morphotypes and wild *B. oleracea*, but also their *Brassica* C9 relatives. This allows us to 1) perform extensive genetic diversity analysis within and between morphotypes, 2) investigate 'tokogenetic' relationships between all *B. oleracea* morphotypes, their wild relatives, as well as different wild C9 species, and 3) infer population structure of this large population. As both genebank and modern hybrid accessions are included we can compare their genetic diversity which is important for breeders to evaluate their potential use for mining novel variation.

*B. oleracea* accessions are generally self-incompatible and for that reason heterogeneous. As we aimed to include a large number of these heterogeneous genebank accessions (531 in our study), we decided to represent accessions by single plants, similar to other studies (Zhao *et al.*, 2005, Mabry *et al.*, 2021). A pilot to test how this influences results showed that intra-accession variation is generally smaller than inter-accession variation, but also showed that genotypes/plants from genebank accessions that display inter-plant phenotypic variation are not always differentiated from genetically closely related accessions.

We performed several comparative analyses, and overall the results were congruent. This was especially the case for the ML tree, the PCA and the STRUCTURE analysis. These analyses showed that the cauliflowers had highest genetic differentiation from other morphotypes and the C9 species, and that the broccoli's were closest to cauliflower. In PCA analyses, cauliflowers separated from all other morphotypes in PC1 (explaining 13.64% of the variation), while C9 species only separated in PC2 (4.78% variation explained). Also in STRUCTURE analysis, at K=2, one group consisted of cauliflower accessions, and the other group included all remaining accessions. Only at K=6, the C9 species formed a separate group. Analysis of population differentiation between morphotypes again showed that $F_{ST}$ values between cauliflower and other morphotype groups was higher than between wild *B. oleracea* and other morphotypes. Results of pairwise genetic distance analyses and nucleotide diversity analyses also agreed. Genetic variation among cauliflower accessions was very low. Genetic variation within morphotypes was substantially lower between hybrids than between genebank accessions. This was to a lesser extend the case for cauliflowers. The strong genetic differentiation of cauliflower combined with a reduced genetic variation among cauliflower accessions points to a very strong genetic bottleneck. In a recent publication by Guo et al. (2021) a *de novo* genome assembly of a cauliflower and a cabbage was generated and structural variations shared among cauliflower resp cabbage accessions were exploited to identify possible selection signals. This revealed many "cauliflower" selection signals in diverse molecular pathways (flowering time, floral identity, meristem proliferation, organ size and spirality), suggesting indeed a strong genetic bottleneck.

The SVD-quartets analysis yielded a tree topology that was not fully congruent with the ML tree, which is not unexpected given the coalescent-based nature of the SVDq approach. Both trees revealed two clearly separated cultivated lineages, the LHL and AIL. In any case, we use the SVDq topology in our further discussions given the accuracy of SVD-quartets analysis (Chifman and Kubatko, 2014, Leaché and Oaks, 2017). This tree shows that the kales form different clades that may represent progenitors of the LHL and AIL. Our separate kale lineages (KA1 vs KA2,3) that diverged before the emergence of LHL and AIL, are consistent with a scenario involving 'ancient' divergent kale lineages leading to LHL and AIL. This illustrates that the kales were already diversified at the time when *B. oleracea* domestication started and that they are likely to be ancestral to the other cultivated morphotypes. For possible timing of these domestications, we did consult the literature on ancient writings, to search for first occurrences (Maggioni *et al.*, 2018). For instance, cauliflower and cabbages were mentioned by Theophrastus and Pliny the Elder at around 400BC (or 2.5ka; see Introduction); also ancient literature mentions wild and tame 'coles', and distinguished several 'coleworts' (kales) based on leaf, stalk type, sprouts and taste. Applying 400BC to nodes F (AIL) and D (LHL), we find an age of 4581yr (2560 BC) for node A (i.e., the onset of cultivation of *B. oleracea*). As wild *B. oleracea* grows along the coasts of England, France and Spain, and both our evidence of cabbages (Mesquite analysis based on origin, Fig. S3), and several publications (see Introduction) points to domestication in the middle east, one possible explanation is that ships transporting tin ores from Cornwall and the Iberian island, also carried 'coleworts' along. The 2560 BC date for the origin of *B. oleraceae* breeding would be consistent with a time frame of emerging tin-trade between Britain, France, Spain and Portugal to the Mediterranean (Berger *et al.*, 2019). Berger et al. (2019) proof that tin ingots excavated around the eastern Mediterranean sea are from Cornish tin mines, based on radiogenic composition that can be linked to geological age of the tin ores. They state that this trade shifted from the Near East to Europe and Cornwall in particular, at the demise of the Minoans and during the rise of the Mycenaeans ca. 1430 BCE, which is slightly later than our estimated time of 2560 BC.

Ancestral state analysis using Mesquite was performed for the cauliflowers, cabbages and broccoli's separately to understand their breeding history. Winter broccoli and cauliflower types were ancestral to the earlier flowering types, while for cabbages, the white cabbages were inferred to be ancestral. We also reconstructed ancestral areas and this revealed that for cauliflowers, the ancestral types came from Italy, with several Romanesco types at their origin, and for heading cabbages, ancestral types originated from eastern Europe, with many from Macedonia (MKD) and Yugoslavia (YUG), supporting an east Mediterranean/near east origin. For the broccoli's results were different, as the ancestral types were winter broccoli's from Great Britain. This fact conflicts with the genealogic tree and previous studies that suggest that cauliflower was derived from broccoli. We included all broccoli accessions from the involved genebanks so cannot expand the data to include accessions of different origins.

SVD-quartets topology also places the Brussels sprouts, sister to kale1 and wild oleracea1, as a separate lineage that emerged later than the cabbages. Former studies also indicated different ancestors for sprouts and cabbages, and often also mentioned wild C9 species in their ancestries (Mabry *et al.* 2021 and references there in (Schulz 1936; Neufchatel 1927; Helm 1963)). The position of sprouts sister to kales fits their botanical characteristics, with a long stalk and absence of apical dominance ("sprouting coleworths" were mentioned by Cato and Pliny the Elder at around 200BC (see Introduction)). Kohlrabi's form a separate lineage sister to kale 2 and 3 clades and wild oleracea 3. Little is known about their domestication. Dodenaeus (1554) mentions many cole crops, but not kohlrabi. In a  study by Zeven (1996) where they searched for 16th and 18th century pictures of colecrops, only one painting was revealed that illustrated an intermediate morphology between narrow stem kale and kohlrabi (German painter Jacob Samuel Beck (1715-1778) (Zeven, 1996). This painting however fits with the classification of kohlrabi sister to kale 2 and 3 that consist of several marrow stem kales showing its botanical proximity. The collards form an interesting morphotype, as they are cultivated since the 1800's in the southern and eastern US, mainly by home gardeners and seed savers (Pelc *et al.*, 2015). Farnham et al. (2008) mention that their origin is likely from the old world, where they were introduced as early as 1500-1600 by Spanish/Portugese and English settlers (Farnham *et al.*, 2008). The name collard is likely derived from 'colewort', non-heading kale types. Half of collard accessions are semi-heading, and a hypothesis is that as they were cultivated near cabbages, intermating between collards and cabbage was common. In both our study and the study of Pelc *et al.* (2015), collards cluster with the cabbages. Chinese kale is an interesting morphotype, as it flowers very early, which is very different from most biannual *B. oleracea*'s  that need strong vernalization for flowering.  Both summer/autumn and tropical broccoli and cauliflower also have only a weak vernalization requirement. In the SVD-quartets tree, Chinese kale is closest related to the cauliflowers and broccoli clades (AIL), which share the early flowering phenotype due to FLC mutations (Guo *et al.*, 2021) but are characterized by more mutations (e.g. curd proliferation arrested inflorescence meristems).

Both our study and the study of Mabry et al. (2021) (Mabry *et al.*, 2021) integrated phylogenetic and population genetics techniques with archaeological and historic writing evidence to investigate the evolutionary history of *B. oleracea* crops and its wild relatives. The results in these two studies are largely in agreement with each other. We provided evidence for two domestication lineages (LHL and AIL), which were also present in their individual level phylogeny. We showed that kales are polyphyletic with regard to *B. oleracea* morphotypes, while Mabry et al. also provided evidence for this observation, suggesting highly diverse kale types. Mabry et al. however specifically indicated *B. cretica* as the closest living wild relative of *B. oleracea*, pointing to an Eastern Mediterranean origin. They especially mentioned that five admixture events were identified using TreeMix and highlighted that many admixture events and lineages of exoferal origins characterized the evolutionary history of *B. oleracea*. Among the five admixture events, one is from *B. cretica* accession (198) to a clade of Chinese kale and Tronchuda, a second one from Kohlrabi to another *B. cretica* accession (199), and a third one from Chinese kale to the *B. cretica* accession (199).  Interestingly, in our genealogical tree the LHL and AIL formed clear monophyletic groups, while the kohlrabi's and Chinese kale and Tronchuda's together were situated at the junction of the AIL and the KA3 clade, which may indicate admixture between these accessions and kale, AIL and possibly wild C9 species like *B. cretica*. Mabry *et al.* mentioned that *B. cretica* is likely ancestor, however they didn't find evidence if this is the case for the two main lineages (AIL and LHL). We found evidence for far east origin of the AIL and LHL clades in both the hypothesis of diversified kales and in the first-branching cabbage clade with accessions from the middle-east that seems ancestral. In our study, we propose a scenario in which ancient divergent kale lineages have led to AIL and LHL. This scenario also leads us to support a middle-eastern origin, which corroborates the conclusion of Mabry et al., which is based on identification and inclusion of the closest living relative species, *B. cretica*. Specifically, together with archaeological and literature evidence, we hypothesized that cabbages and cauliflowers stem from kales introduced from Western Europe to the middle-east, possibly transported with the tin-trade routes in the Bronze age, to be re-introduced later into Europe. We estimated the possible timing of the domestication event for *B. oleracea* at ~2560 BC (~4.6 ka). Likewise, the data from Mabry et al. not only supported an origin of cultivation in the Eastern Mediterranean region, but also pointed to a Late-Holocene domestication, a time frame whose boundary (~4.2 ka) is close to our estimation.

Additionally, we investigated genetic diversity between different groups (i.e. between genebank and modern hybrids) that reflect the history of plant breeding. Overall, allelic variation in genebank accessions was much larger compared to that of modern hybrids, which is important in guiding plant breeders' strategies. As this study used a very large collection (almost 1000 accessions) we could also compare variation between different morphotypes, different varieties and ecotypes.

In summary, we analysed the genetic diversity, genealogical relationship and population structure among 912 *B. oleracea*- and their wild relative accessions. Our data illustrate that genetic diversity reduces from genebank accessions to modern hybrid accessions. Species tree analysis showed evidence for two lineages, LHL and AIL, with onset of diversification and breeding of cabbages and cauliflowers around 400 BC in the middle east. Different kale and wild *B. oleracea* were likely progenitors of the diverse lineages (besides the AIL and LHL also the sprouts and kohlrabi's). Cauliflower is the least diverse morphotype and has the strongest genetic differentiation with other morphotypes, which points to a very strong genetic bottleneck.

## Materials and methods

### Plant materials

A set of 912 accessions were collected from both germplasm collections (referred to here as 'genebanks') and modern hybrid cultivars from breeding companies. An accession was defined as one entry in the germplasm/cultivar collection, and represented by a single plant in this study. For details on genebank origin, breeding companies hybrids and "wild C9 species" (non *B. oleracea* species, with the same chromosome number as *B. oleracea*), see Supplementary Notes, Table 1, Table S1-S2.

*B. oleracea* is a self-incompatible species and landraces are thus heterogeneous. To assess this heterogeneity, six accessions, including two cauliflower and four heading cabbage, were selected to study this intra-accession variation. For this purpose, we selected accessions that were either phenotypically uniform or variable when grown in the field. We re-sowed these and randomly genotyped ten individual plants per accession. For cauliflower we included one phenotypically uniform accession (TKI504) and one phenotypically non-uniform accession (TKI506). We also included two uniform heading cabbage accessions (TKI424 and TKI531) and two non-uniform accessions (TKI529 and TKI541) (Table S3). Our analysis indicated that the decision to represent accessions by single plants doesn't bias the diversity analysis, as generally intra-accession variation is smaller than inter-accession variations, even though the variations are underrepresented by single plants (Supplementary Notes, Fig. S1).

### DNA extraction, sequencing and variant calling

Total genomic DNA was extracted from fresh leaves using an optimized CTAB method (Chen and Ronald, 1999). For hybrid accessions (being the result from a cross between two homozygous inbred lines, thus plants from the same accession are genetically identical), we isolated DNA from batches of around 100 seedlings to facilitate DNA isolation, while for genebank accessions, one plant per accession was genotyped; for two cauliflower and four heading cabbage accessions (see Plant materials), DNA was isolated from ten individual plants independently to investigate intra-accession variation. The DNA quality and quantity were measured with Nanodrop 2000. All the DNA samples were genotyped by Sequence-Based Genotyping (SBG) method (Truong *et al.*, 2012) at Keygene N.V., Wageningen, the Netherlands. The genomic DNA was digested by a combination of PstI (rare cutting) and MseI (frequent cutting) restriction enzymes to reduce genome complexity. After that, primers were annealed and the products were amplified with a 2 basepair (GG) extension to again reduce genome complexity. Constructed SBG libraries were then amplified and sequenced on Illumina Hiseq platform. The generated raw reads were first split for each particular sample according to the 5-bp sequence barcode. No mismatches in the barcode and PstI footprint were allowed. Reads with mismatches in these first 11 nucleotides were left unassigned. After splitting files, barcode and PstI restriction footprint sequences were removed and replaced by the proper PstI restriction sequence. Reads mapping, variant calling and filtering were performed using the method described in Supplementary Notes. An imputed dataset of 14,152 SNPs was utilized for the population genetics analysis.

### Pairwise genetic distance analysis

The R package poppr (Kamvar *et al.*, 2014) was used to create a pairwise genetic distance matrix for the 912 accessions with bitwise.dist function. All the genetic distance values were normalized to 0-1 after obtaining the matrix.

### Maximum likelihood (ML) tree construction and population structure analyses

IQ-TREE v1.6.10 (Nguyen *et al.*, 2015) was used to construct ML trees for our *B. oleracea* nucleotide version SNP matrix, with the following parameters "-m MFP+ASC -bb 1000 -bnni". The best model was automatically determined based on the Bayesian Information Criterion (BIC) in the IQ-TREE pipeline, as well as 1000 replicates of ultrafast bootstrapping (UFboot) to estimate node support. The ML tree with bootstrap group frequencies was visualized with interactive Tree of Life (iTOL) (https://itol.embl.de/) (Letunic and Bork, 2016). The tree was manually inspected to identify accessions which were clearly mis-named or -identified, clustering in unexpected morphotype groups. Those accessions were not included in the subsequent PCA, STRUCTURE, π and $F_{ST}$ analyses. As reconstructing 'phylogenetic' relationships within a species (*B. oleracea*) is actually not possible (only between species), we refer to this tree as a genealogical tree.

Principal component analysis (PCA) was conducted using EIGENSOFT v6.1.4 (Patterson *et al.*, 2006, Price *et al.*, 2006) software packages on individual genotypes with default parameters.

Population structure was inferred using STRUCTURE v2.3.4 (Pritchard *et al.*, 2000) on the genome-wide SNPs. The *K* value, which was defined as a putative number of ancestral populations, was set from 1 to 15. For each *K* value, STRUCTURE was run 10 times with 10,000 burn-in cycles and 10,000 Markov chain Monte Carlo (MCMC) replicates. The CLUMPAK (Kopelman *et al.*, 2015) software was used to estimate optimum number of sub-groups with an ad hoc statistics *Δk* method (Evanno *et al.*, 2005) and to visualize the sub-population membership for each accession.

**Ancestral type reconstruction**

To reconstruct ancestral states of the 'ecotype' (broccoli, cauliflower) and 'variety' (heading cabbage), we recompiled SNP matrices for each morphotype as found in the overall ML analysis and reconstructed a morphotype-specific ML tree using IQ-TREE. Ancestral states were reconstructed treating 'ecotype/variety' as a character. Character states we used were: winter, summer/autumn, tropical for cauliflower, winter, summer/autumn for broccoli and red, white, savoy, pointed for cabbages. The character state for Romanesco cauliflower's was set to 'unknown' since it is not clear which ecotype these accessions belong to. Character evolution was reconstructed onto the ML trees using Mesquite v3.61 (Maddison and Maddison, 2019) with the 'Trace Character History' option, and using unordered parsimony as criterion.

**Genetic diversity analysis**

The average difference per locus over each pair of accessions, π, estimates the level of genomic diversity in a group of accessions (Nei and Li, 1979). VCF-tools v0.1.15 (Danecek *et al.*, 2011) was used to calculate π for our data with a sliding window size of 100kb and step size of 10kb. Reduction of diversity (ROD) metrics was also calculated based on π-value (Xu *et al.*, 2012, Cheng *et al.*, 2016b). When calculating ROD, wild *B. oleracea* or genebank group was used as the control group. $F_{ST}$ is the population fixation statistics to calculate the pairwise genomic differentiation between two groups of samples (Weir and Cockerham, 1984). Pairwise $F_{ST}$ between two different groups was estimated using VCF tools v0.1.15 with a sliding window size of 100kb and step size of 10kb. $F_{ST}$ results were interpreted using the same methods described by Del Carpio *et al.* (Del Carpio *et al.*, 2011), where the $F_{ST}$ value of 0 denotes no differentiation and 1 denotes complete differentiation between populations. Little differentiation is considered when $F_{ST} < 0.05$, moderate differentiation when $0.05 \leq F_{ST} < 0.15$, strong differentiation when $0.15 \leq F_{ST} < 0.25$, and very strong differentiation when $F_{ST} \geq 0.25$ (Hartl, 1980, Mohammadi and Prasanna, 2003, Bird *et al.*, 2017).

**Species tree reconstruction**

We used SVD-quartets as implemented in PAUP* version 4.0a (build 168) with standard settings, for analyzing a 875 x 14,152 SNP matrix ('Overall') from which all constant sites had been removed, and in which each terminal had been assigned to one of the 12 morphotypes. Given the overall IQ-TREE genealogy in which Kales and wild *B. oleracea* were on three separate branches respectively, we decided to allow three Kale lineages and three wild *B. oleracea* lineages in the SVDq analysis, each assigned multiple members. After nodes with bootstrap values <50% were collapsed, the resulting overall SVDq morphotype tree topology was then used as a constraint to estimate branch lengths based on the nucleotide version of a subsampled 57 terminal SNP matrix ('SUB'). This matrix, representing all 12 morphotypes, and using wild C9 species as outgroup, was compiled in such a way as to represent (deep) nodes in the Overall ML tree. The 'autoModel' option in PAUP* was used to find the best-fitting model and parameter values. Using 'saveTrees', the ML based branch lengths were computed and the tree saved. The next step was to import the tree in Mesquite and use the 'arbitrarily make ultrametric' command to produce an ultrametric version of the tree.

**Data Availability**

The data for this article have been deposited in Figshare (https://doi.org/10.6084/m9.figshare.16338855).

**Acknowledgements**

This research was funded by the TKI project " A *de novo* sequencing catalogue of *B. oleracea*" , project 1309-011 (https://topsectortu.nl/nl/innovation-accelerator-genetic-variation-brassica-oleracea), and was co-supported by six breeding companies (Bejo, ENZA, Rijk Zwaan, Syngenta, Takii, Hazera) and KeyGene. C.C. is supported by China Scholarship Council (No. 201809110159). We like to thank Noortje Bas from

CGN (https://www.wur.nl/nl/Onderzoek-Resultaten/Wettelijke-Onderzoekstaken/Centrum-voor-Genetische-Bronnen-Nederland-1.htm) for her support in collecting the Brassica accessions and the staff of Unifarm for taking care of the field experiments.

## Conflict of interests

The authors declare that they have no conflict of interest.

## Author contributions

C.C. analyzed and interpreted the data, drafted and revised the manuscript. J.B. collected plant materials and extracted DNA. FT. B. contributed to genealogical data analysis and dating and to valuable discussions and revision of the manuscript. G.B. designed the research, supervised the experiment and data analysis, and partly wrote and revised the manuscript. All authors read and approved the final manuscript.

## Supporting information

Supplementary files are available at https://academic.oup.com/hr/article/doi/10.1093/hr/uhac033/6532230

**Fig. S1** Intra-accession variation analysis based on maximum likelihood tree of heading cabbage and cauliflower accessions. For the selected accessions each (TKI424, TKI531, TKI541, TKI529, TKI504 and TKI506), ten individual plants were genotyped. Other accessions, are represented by one individual plant. TKI870 (*B. montana*, wild C9 species) was used as an outgroup.

**Fig. S2** Heatmap showing the genetic distance matrix between accessions within given *B. oleracea* morphotypes or within C9 species. Different colors in the vertical bar represent different ecotypes, varieties or species. (a) broccoli (sa: summer/autumn, un: unknown, wi: winter). (b) cauliflower (ro: romanesco, sa: summer/autumn, tr: tropical, un: unknown, wi: winter). (c) heading cabbage (po: pointed, re: red, sa: savoy, un: unknown, wh: white). (d) kale (kale-bc: bore and curly kale, kale-ms: marrow stem kale). (e) wild C9 species (bou: *B. bourgeaui*, cre: *B. cretica*, dre: *B. drepanensis*, inc: *B. incana*, ins: *B. insularis*, mac: *B. macrocarpa*, mon: *B. montana*, rup: *B. rupestris*, vil: *B. villosa*).

**Fig. S3** Geographical origin character evolution for (a) broccoli, (b) cauliflower and (c) heading cabbage. Different geographical origins were treated as the states of this character. Grey branches represent accessions with "unknown" state.

**Fig. S4** Maximum likelihood tree of wild *B. oleracea* and wild C9 species accessions. The abbreviations: WC9 denotes wild C9 species, Wo denotes wild *B. oleracea*, ge denotes genebank. (bou: *B. bourgeaui*, cre: *B. cretica*, dre: *B. drepanensis*, inc: *B. incana*, ins: *B. insularis*, mac: *B. macrocarpa*, mon: *B. montana*, rup: *B. rupestris*, vil: *B. villosa*).

**Fig. S5** *Δk* analysis for the different number of clusters for the *B. oleracea* and wild C9 species accessions.

**Fig. S6** Population structure of major *B. oleracea* morphotypes and wild C9 species with different numbers of clusters (*K=8-12*). Each accession is represented by a vertical bar with different colors. The length of each colored segment in the bar quantifies cluster membership.

**Fig. S7** Boxplot of genotype missing rate for 912 accessions (a) before and (b) after imputation.

**Fig. S8** Distribution of high-quality SNP markers on the nine chromosomes of *B. oleracea*.

**Fig. S9** Tree topology of the SVDquartet analysis from Figure 7a, with nodes with bootstrap values <50% collapsed, enforced on the SUB matrix (see text). Branch lengths are according to nucleotide SNP changes.

**Fig. S10** BEAST SNAPP analysis. Preliminary results of a 146k generation Markov Chain with (a) Tracer output of the posterior showing the jump-like improvements in posterior after 50k and 80k generations; (b) Consensus tree of trees sampled for the last 40k generations; and (c) Consensus tree with effective population size $N_e$ estimates (θ) indicated at the nodes.

**Fig. S11** PCA plots of broccoli accessions. (a) Accessions were classified according to different broccoli ecotypes. (b) Accessions were classified according to the collection of materials (genebank or modern hybrids). The first two principle components were plotted to visualize the relationships among individuals and groups.

**Fig. S12** PCA plots of cauliflower accessions. (a) Accessions were classified according to different cauliflower ecotypes. (b) Accessions were classified according to the collection of materials (genebank or modern hybrids). The first two principle components were plotted to visualize the relationships among individuals and groups.

**Fig. S13** PCA plots of heading cabbage accessions. (a) Accessions were classified according to different varieties. (b) Accessions were classified according to the collection of materials (genebank or modern hybrids). The first two principle components were plotted to visualize the relationships among individuals and groups.

**Fig. S14** PCA plots of kale and Chinese kale accessions. (a) Accessions were classified according to different kale types. (b) Accessions were classified according to the collection of materials (genebank or modern hybrids). The first two principle components were plotted to visualize the relationships among individuals and groups.

**Fig. S15** PCA plots of wild C9 species accessions. The first two principle components were plotted to visualize the relationships among individuals and groups.

**Table S1** The information of all accessions used in this study. (Excel spreadsheet)

**Table S2** Summary of wild C9 species accessions.

**Table S3** The number of SNPs that vary between 10 plants within each accession.

**Table S4** Genome-wide nucleotide diversity (π) and reduction of diversity (ROD) for each group.

**Table S5** Pairwise comparison of $F_{ST}$ values between different ecotypes/varieties.

**Table S6** Summary of SNPs on each chromosome of *B. oleracea*.

# Chapter 3

## Chromosome-scale genome assemblies of five different *Brassica oleracea* morphotypes provide insights in intraspecific diversification

**Chengcheng Cai[1,2], Johan Bucher[1], Richard Finkers[1,3] and Guusje Bonnema[1,2,*]**

[1] Plant Breeding, Wageningen University and Research, 6708 PB, Wageningen, The Netherlands

[2] Graduate School Experimental Plant Sciences, Wageningen University and Research, 6708 PB, Wageningen, The Netherlands

[3] Gennovation B.V., Agro Business Park 10, 6708 PW, Wageningen, The Netherlands

[*] Corresponding author

**Preprint in *bioRxiv* (2022).**

**DOI: https://doi.org/10.1101/2022.10.27.514037**

## Abstract

*Brassica oleracea* is an economically important vegetable and fodder crop species that includes many morphotypes exhibiting enormous phenotypic variations. Previously, a pan-genome study based on short reads mapping approach has shown extensive structural variations between *B. oleracea* morphotypes. Here, to capture more complete genome sequences of *B. oleracea*, we report new chromosome-scale genome assemblies for five different morphotypes, namely broccoli, cauliflower, kale, kohlrabi and white cabbage, which were created by combining long-read sequencing data and Bionano DLS optical maps. The five assemblies are the most continuous and complete *B. oleracea* genomes to date (contig N50 > 10 Mb). Comparative analysis revealed both highly syntenic relationships and extensive structural variants among the five genomes. Dispensable and specific gene clusters accounted for ~38.19% of total gene clusters based on a pan-genome analysis including our five newly assembled genomes and four previously reported genomes. Using the pan-genome of *B. oleracea* and *B. rapa*, we revealed their different evolutionary dynamics of LTR-RTs. Furthermore, we inferred the ancestral genome of *B. oleracea* and the common ancestral genome of *B. oleracea* and *B. rapa* via a pan-genome approach. We observed faster WGT-derived gene loss in *B. rapa* than in *B. oleracea* before intraspecific diversification. We also revealed continuing gene loss bias during intraspecific diversification of the two species and a strong bias towards losing only one copy among the three paralogous genes. This study provides valuable genomic resources for *B. oleracea* improvement and insights towards understanding genome evolution during the intraspecific diversification of *B. oleracea* and *B. rapa*.

## Introduction

Long-read sequencing and long-range scaffolding technologies have significantly improved the quality of genome assemblies. Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) are two commercial high-throughput long-read sequencing platforms that can generate DNA fragments ranging from kilobases to megabases (Rousseau-Gueutin *et al.*, 2020). Although the raw reads of these technologies have error rates of up to 15%, the accuracy of assembled sequences can reach as high as 99.999% via error correction strategies (Jiao *et al.*, 2017, Jiao and Schneeberger, 2017, Belser *et al.*, 2018). More importantly, long-read sequencing data enables genome assemblies with high contiguity and completeness whereas short-read data (Illumina, Roche 454) often results in highly fragmented assemblies as it is not suitable to span repetitive regions. The last few years have seen highly continuous genomes assembled by long reads for a wide range of species (Schmidt *et al.*, 2017, Belser *et al.*, 2018, Deschamps *et al.*, 2018, Ou *et al.*, 2020). Even though long-read sequencing technologies have enabled highly continuous assemblies, they are solely not sufficient to completely assemble complex eukaryotic genomes, especially those of plants with high levels of repeat sequences and large genome size. The resulting contigs from sequencing reads require further scaffolding to eventually achieve chromosome-scale assemblies. One such long-range scaffolding technology is Bionano Genomics optical mapping with its new Direct Label and Stain (DLS) technology. This technology uses Direct Labeling Enzyme 1 (DLE-1) to attach a single fluorophore to specific sequence motifs to produce fingerprints of DNA fragments, thus not damaging DNA molecules at specific sites (Deschamps *et al.*, 2018). The DLS-labeled molecules are much longer than those labelled via the endonuclease approach, with the longest ones becoming larger than 2 Mbp. Chromosome-scale assemblies can often be generated in single maps by using DLS molecules (Formenti *et al.*, 2019).

*Brassica oleracea* (CC, 2n=18) is an economically important vegetable and fodder crop species cultivated worldwide. It consists of many morphotypes which exhibit enormous phenotypic variations, such as the leafy heading morphotype *var. capitata* (cabbage), the typical curd morphotypes with large arrested inflorescences, including *var. botrytis* (cauliflower), *var. italica* (broccoli), the kohlrabi's with their tuberous stems (*var. gongylodes*), the kales (*var. acephala*) with different leaf types and etc (Kole and Henry, 2010, Bonnema *et al.*, 2011, Dias, 2012, Cai *et al.*, 2022a). Despite this enormous diversity, *B. oleracea* truly remains one species and morphotypes can be easily interbred. *B. oleracea* is an important diploid member of the "triangle of U" model, which includes the other two diploid species, *Brassica rapa* (AA, 2n=20) and *Brassica nigra* (BB, 2n=16), and three allotetraploid species generated through pairwise crosses between the diploids, *Brassica juncea* (AABB, 2n=36), *Brassica napus* (AACC, 2n=38) and *Brassica carinata* (BBCC, 2n=34) (Nagaharu, 1935a). Similar to *B. oleracea*, *B. rapa*, *B. juncea* and *B. napus* all include many diverse morphotypes showing extreme phenotypes.

To date, several genome sequences of *B. oleracea* have been created either using short-read or long-read sequencing technology. Two *B. oleracea* reference genomes were firstly released in 2014, one from cabbage line 02-12 and the other from the doubled haploid *B. oleracea* annual kale-like type TO1000DH (Liu *et al.*, 2014b, Parkin *et al.*, 2014b). Both these genome sequences were generated on the basis of Sanger sequencing and deep short-read sequencing data, and anchored to pseudo-chromosomes by genetic maps. These two genomes have been used as the references for many years for comparative and functional genomics studies in *B. oleracea* crops (Zhang *et al.*, 2015a, Cheng *et al.*, 2016b, Golicz *et al.*, 2016a). Recently, several other reference genomes, which were primarily created by combining high-coverage long-read sequencing data with long-range scaffolding information, such as optical maps and/or Hi-C, have become available. This includes a re-assembly of cabbage line 02-12 (Cai *et al.*, 2020), a new broccoli line HDEM (Belser *et al.*, 2018), two new cabbage lines (OX-heart and D134) (Guo *et al.*, 2020, Lv *et al.*, 2020) and two new cauliflower lines (Korso and C-8) (Sun *et al.*, 2019a, Guo *et al.*, 2020). Increasing numbers of pan-genome studies based on *de novo* assembly approaches in multiple crops, such as rice (Zhao *et al.*, 2018, Qin *et al.*, 2021), soybean (Liu *et al.*, 2020), maize (Hufford *et al.*, 2021) and rapeseed (Song *et al.*, 2020), have shown that structural variations widely exist within a species. Also in *B. oleracea,* a pan-genome study based on a reference-guided approach using short-read sequencing technology shows such high level of variations between different morphotypes, with nearly 20% of genes affected by presence/absence variation (Golicz *et al.*, 2016b). These studies illustrate that a single reference genome is not sufficient to cover the genome sequences of a species. For these reasons, it is necessary to generate more *B. oleracea* reference genomes from highly diverse morphotypes, especially high quality sequences, to fully resolve structural variations that affect gene function and expression and influence agriculturally important traits.

Polyploidization is prevalent and recurrent in the plant kingdom and plays a crucial role in speciation and species diversification (Cheng *et al.*, 2014, Zhang *et al.*, 2019a, Cai *et al.*, 2021). *Brassica* species are ideal models for polyploidy and evolutionary studies in plants partly because they have experienced the common Brassiceae-specific whole genome triplication (WGT) event (Wang *et al.*, 2011b, Liu *et al.*, 2014a, Parkin *et al.*, 2014a). This event has been confirmed by extensive comparative analyses between genome sequences of *Brassica* species and *Arabidopsis thaliana* (Wang *et al.*, 2011b, Cheng *et al.*, 2014, Liu *et al.*, 2014a). In *B. rapa*, the three subgenomes: least fractionated (LF), medium fractionated 1 (MF1) and most fractionated 2 (MF2), were demonstrated to have been evolved from a common translocation Proto-Calepineae Karyotype (tPCK) ancestral diploid genome (Cheng *et al.*, 2013, Cheng *et al.*, 2014). The extant genome structures of diploid *Brassica* species were shaped during the rediploidization process that followed the WGT, which involves extensive gene fractionation, genomic reshuffling and chromosome reduction (Cheng *et al.*, 2014). To illustrate the WGT process from a genome evolution perspective in *Brassica* plants, Cheng et al proposed a “two-step theory” that suggests a tetraploidization event between the tPCK genomes of MF1 and MF2, followed by fractionation and a subsequent hybridization event between the diploidized genome and a third tPCK genome LF (Cheng *et al.*, 2012a, Cheng *et al.*, 2014, Cai *et al.*, 2021). This theory also explains the subgenome dominance observed in *Brassica* plants. Like Brassica’s, many other species including maize, wheat, cotton, grasses, and *A. thaliana* (Senchina *et al.*, 2003, Wang *et al.*, 2006, Buggs *et al.*, 2010, Schnable *et al.*, 2011, Cheng *et al.*, 2012a, Pont *et al.*, 2013, Akama *et al.*, 2014, Li *et al.*, 2014a, Murat *et al.*, 2014, Renny-Byfield *et al.*, 2015, Cheng *et al.*, 2016a) display subgenome dominance, with one subgenome retaining more genes, contributing more highly expressed genes and accumulating fewer non-synonymous mutations than other subgenomes (Cheng *et al.*, 2014).

As a pervasive source of genetic change, gene loss is prevalent in all life kingdoms and has great potential to result in adaptive phenotypic diversity (De Smet *et al.*, 2013, Albalat and Cañestro, 2016). It has been found that gene loss is coupled with extensive polyploidy events that create redundant genes, the loss of which usually does not result in apparent functional consequences (Albalat and Cañestro, 2016). In addition to biased gene loss between different subgenomes following polyploidization, gene loss is also biased towards gene function (Albalat and Cañestro, 2016). In *B. rapa* and *B. oleracea*, genes involved in the response to phytohormone signalling were found to be significantly over-retained during gene loss after WGT (Wang *et al.*, 2011b, Cheng *et al.*, 2014). Moreover, a functional bias of gene loss can often be observed in species that suffer relaxation of a given biological or environmental constraint. This functionally biased gene loss is caused by the ‘co-elimination’ of genes that are functionally linked in distinct pathways or complexes associated with relaxed constraint (Aravind *et al.*, 2000, Koonin *et al.*, 2004, Albalat and Cañestro, 2016). In a recent research, Cai et al (Cai *et al.*, 2021) inferred the *B. rapa* ancestral genome using a pan-genome based approach, which provides an essential reference to investigate gene loss during intraspecific diversification. They further illustrated the impacts of WGT event and subgenome dominance on intraspecific diversification of *B. rapa*. *B. oleracea* and *B. rapa* are sister species and their divergence occurred at about 4.6 Mya (Cheng *et al.*, 2017). Due to the lack of high quality reference genome sequences, individual genome evolution regarding gene loss during intraspecific diversification of *B. oleracea* is still unexplored. In addition, studies focussing on comparative genome evolution between sister species after speciation from their common ancestor are sparse.

Here, we release chromosome-scale genome assemblies of five *B. oleracea* morphotypes, including broccoli, cauliflower, kale, kohlrabi and white cabbage, all of which were generated by integrating long reads, optical mapping molecules (BioNano Genomics DLS technology) and Illumina short reads. The final assemblies showed extremely high contiguity, with contigs having N50 values between 11.4 Mb and 16.3 Mb and scaffolds having N50 values between 30.5 Mb and 34.1 Mb. Comparative analysis among the five new assemblies demonstrates high degrees of synteny among *B. oleracea* genomes as well as extensive structural variations. Together with four previously published high-quality genomes, we revealed the composition and features of a *B. oleracea* pan-genome via a *de novo* assembly approach. Additionally, we investigated intact LTR-RTs in the pan-genome of *B. rapa* and *B. oleracea*, and compared their evolutionary dynamics. Furthermore, using a pan-genome approach, we studied the impacts of WGT event and subgenome dominance on intraspecific diversification of *B. oleracea*. We also compared evolutionary patterns regarding biased WGT-derived gene loss between *B. rapa* and *B. oleracea* after the speciation from their common ancestor. Together, our work provides valuable resources for genomic-assisted breeding of *B. oleracea* and sheds lights on understanding intraspecific diversification of *B. oleracea* and *B. rapa*.

## Results

### Contig assembly of five *B. oleracea* morphotypes

Five *B. oleracea* accessions (DH lines) representing five different morphotypes, broccoli, cauliflower, kale, kohlrabi and white cabbage, were selected for genome sequencing. Libraries generated from HMW DNA from fresh leaf tissues were used as input to generate sequences for the five genomes on Oxford Nanopore GridION platform. We produced 1.5 to 7.3 million raw ONT long reads for the five morphotypes, totalling 22.5 to 42.7 Gb of data, with N50 values ranging from 13.04 Kb to 30.22 Kb (Table S1). Assuming a 630 Mb *B. oleracea* genome size (Liu *et al.*, 2014a), these nanopore long reads represented 36~68-fold coverage, and sequences longer than 50 Kb amounted to as high as 9.1~11.6-fold coverage. Besides Nanopore long-reads, we also produced 4.2-8.3 Gb (6.7-13.2-fold coverage) PacBio sequences for each morphotype, with N50 values ranging from 17.5 Kb to 20.9 Kb (Table S2). In addition, more than 126.25Gb Illumina reads were generated for each morphotype, covering more than 200-fold coverage of the five genomes (Table S3).

The ONT reads were assembled using SMARTdenovo (Liu *et al.*, 2021b) and the resulting five assemblies all showed high contiguity. They consisted of 315-426 total contigs and featured N50 values from 6.3 to 13.1 Mb (Table S4-S8). Our broccoli raw assembly had a contig N50 size of 9.3 Mb, which is similar to the value of the final HDEM assembly (9.5 Mb, most contiguous released *B. oleracea* reference genome so far) (Belser *et al.*, 2018). The total contig size ranged from 536.6 to 562.7 Mb and the largest contig varied from 20.1 to 35.7 Mb. Due to the absence of an error correction stage in the algorithm of SMARTdenovo (Liu *et al.*, 2021b), the consensus sequences required further polishing to improve base accuracy. We polished raw contigs using both Nanopore and Illumina reads by running two rounds of Racon (Vaser *et al.*, 2017), followed by three rounds of Pilon (Walker *et al.*, 2014). Generally, the polishing process slightly increased N50 values but greatly improved complete BUSCO values (Table S4-S8). The complete BUSCO values were 82.2%-86.4% for raw contigs in the five assemblies. The scores increased to 87.6%-90.4% after Racon polishing and to greater than 97% after Pilon polishing. To further evaluate base accuracy of our assemblies, we mapped Illumina reads to the corresponding genome to identify genomic variations. In total, we identified 3,085,054-3,593,799 variations (SNPs and small InDels) in each of five raw assemblies, however, after the polishing process, only 44,806-57,274 variations were identified, indicating remarkable base quality improvement. The polished assemblies reached high QVs at around 40 and high identities at around 99.99% (see Methods) (Table S9).

### Bionano DLS genome maps generation and hybrid scaffolding

The combination of Bionano Saphyr system and DLS technology yielded 2,767,569-20,443,958 DNA molecules with lengths longer than 20 Kb for each of the five genomes. After filtering out molecules smaller than 150 Kb and molecules with < 9 labeling sites, a total of 96.47-155.61 Gb Bionano molecules, with N50 values ranging from 241.84 to 381.83 Kb, were assembled into genome maps. For each genome, the final *de novo* assembly yielded only 63-93 maps, with N50 values reaching 29.02-

33.77 Mb. The resulting total genome map length was 575.94-645.47 Mb, with the largest Bionano map being 47.51-51.00 Mb (Table S10).

To further improve genome assemblies, contigs for each genome were scaffolded with corresponding DLS maps. The resulting hybrid scaffolds showed strong improvements in contiguity, compared to the ONT contig assemblies alone. More than 97.5% of the original ONT contig sequences were anchored to only ~40 scaffolds in each genome, totalling 541.31-557.60 Mb. Scaffold N50 values for the five hybrid assemblies were ~30 Mb and the largest sequence was longer than 46 Mb (Table S11). After resolving 13-bp gaps and gap-filling, we generated the final assemblies for the five genomes, which comprised only 150-249 scaffolds and had contig N50s between 11.43 and 16.28 Mb (Table 1). The remaining gaps accounted for only 1.04-1.49% of the total scaffold size. The long terminal repeat (LTR) assembly index (LAI) (Ou *et al.*, 2018) was 15.56-16.83 in each of the five genomes, indicating high quality of the genome assemblies. We also evaluated the quality by mapping mRNA-seq reads to the corresponding assembly. Up to 98% of mRNA-seq reads can be aligned to the genomes, among which ~94% reads were aligned concordantly exactly one time (Table S23). Together, these results illustrated the high quality of the five *B. oleracea* assemblies. To construct chromosome-level pseudomolecules, we used a homolog-based approach and mapped super-scaffolds to the HDEM reference genome (Belser *et al.*, 2018, Jiao and Schneeberger, 2020). More than 94% scaffold sequences were anchored to the nine pseudochromosomes (Table 1).

**Table 1** Statistics of final genome assemblies.

| | Broccoli | Cauliflower | Kale | Kohlrabi | White Cabbage |
|---|---|---|---|---|---|
| **Contigs** | | | | | |
| Total length (Mb) | 552.24 | 540.33 | 560.01 | 550.61 | 557.03 |
| Total number | 316 | 270 | 364 | 283 | 316 |
| N50 (Mb) | 14.46 | 11.43 | 16.28 | 11.74 | 12.17 |
| Longest (Mb) | 38.59 | 23.43 | 46.44 | 28.36 | 43.79 |
| **Scaffolds** | | | | | |
| Total length (Mb) | 558.04 | 547.89 | 568.26 | 557.26 | 565.47 |
| Total number | 226 | 150 | 249 | 162 | 177 |
| N50 (Mb) | 30.57 | 34.06 | 30.34 | 30.5 | 31.59 |
| L50 | 8 | 7 | 8 | 8 | 8 |
| N90 (Mb) | 12.41 | 13.93 | 13.21 | 11.74 | 12.82 |
| L90 | 18 | 16 | 18 | 18 | 17 |
| Longest (Mb) | 47 | 47.41 | 48.64 | 46 | 48.23 |
| Gap size (Mb) | 5.8 | 7.56 | 8.25 | 6.65 | 8.44 |
| GC | 36.72 | 36.65 | 36.71 | 36.71 | 36.74 |
| LAI | 15.56 | 16.83 | 16.07 | 16.83 | 16.31 |
| Chr. length (Mb) | 526.03 | 524.69 | 535.23 | 531.79 | 533.55 |
| Anchor ratio | 94.26% | 95.77% | 94.19% | 95.43% | 94.36% |

## Genome annotation and comparative genomics

We annotated repetitive elements (TEs) for the five genomes using the EDTA pipeline (Ou *et al.*, 2019). Approximately 51.92-53.74% (280.54-299.05 Mb) of the assembled sequences of each genome were composed of TEs (Fig. 1A and 1B, Table S12-S16), similar to both the JZS v2 and HDEM genomes (Belser *et al.*, 2018, Cai *et al.*, 2020). The most abundant TEs were the LTR-RTs, representing 26.13-28.97% (141.21-161.35 Mb) sequences in each of the five genomes. Using an integrated strategy combining *ab initio*, homolog-based and transcript-based prediction, we identified a total of 60,644-61,995 protein-coding genes in each of the five genomes (Table S17). Nearly 99% (98.16-99.25%) of the genes were located on chromosomes in our five assemblies. BUSCO assessment based on 1,440 conserved plant genes showed that 96.80-97.80% complete genes were present in each genome. Gene functional annotation demonstrated that more than 97.30% of the annotated genes in each genome are supported by homology to known proteins or functional domains in other species (Table S18). Taken together, these results showed the near-complete gene models in our five assemblies.

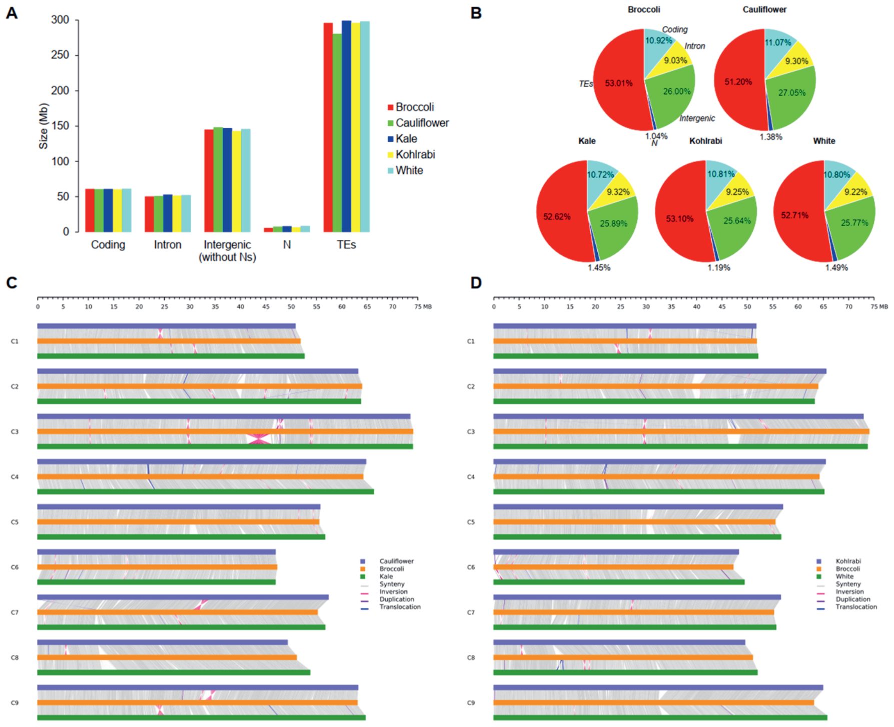


**Fig. 1 Comparisons of genomic components and whole genomic sequences among the five new *Brassica oleracea* assemblies.** (**A**) Size of genomic components in the five assemblies. (**B**) Percentage of genomic components in the five assemblies. As indicated in (**B**) for the broccoli genome, different colours in each of the other four genomes represent different genomic components (red: TEs, aqua: coding sequence, yellow: intron sequence, green: intergenic sequences, blue: Ns). (**C**) Whole-genome comparison among cauliflower, broccoli and kale. (**D**) Whole-genome comparison among kohlrabi, broccoli and white cabbage. Note: in (**C**) and (**D**), only genomic rearrangements with length > 50Kb are shown.

On a genome-wide scale, protein-coding genes were enriched in chromosome arms while TEs tended to be enriched in (peri-)centromeric regions (Fig. S1). Overall, the five assemblies showed similar genomic components, with approximately 11% (60.25-61.09 Mb) coding sequences, 9% (50.42-52.14 Mb) intron sequences, 26% (142.89-148.19 Mb) intergenic sequences and 1% (5.80-8.44 Mb) gaps (N's). However, the size of TEs were more variable among the five assemblies. In cauliflower genome, 280.53 Mb sequences were annotated as TEs, which was 15.33-18.52 Mb less than the figures of the other four genomes (Fig. 1A and 1B).

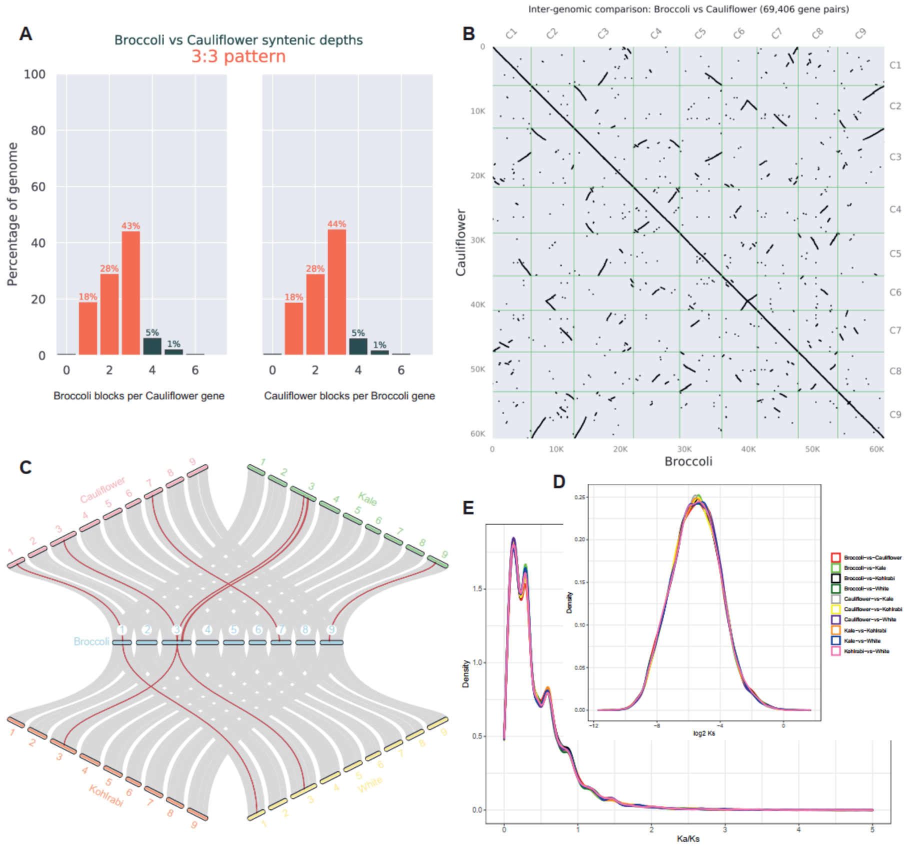


**Fig. 2** (**A**) Ratio of syntenic depth between broccoli and cauliflower genome. (**B**) Homologous dot plot between broccoli and cauliflower genome. (**C**) Macrosynteny among the five genomes with grey links connecting blocks of >30 one-to-one gene pairs. Red lines indicate inversions. (**D**) Ks distribution of each two of the five genomes. (**E**) Ka/Ks distribution of each pair of the five genomes.

Comparative analyses among the five genomes show that they are highly collinear, which was demonstrated by their highly conserved genomic sequences, as well as gene content and order along each chromosome (Fig. 1C and 1D, Fig. 2, Fig. S2 and S3). Approximately 80% of each of the other four genome sequences matched in one-to-one syntenic blocks with the broccoli genome (Table S32). Substantial collinearity notwithstanding, we did detect 6.52-13.96 Mb inversions, 25.79-34.88 Mb translocations, 7.97-14.22 Mb insertions and 20.80-32.33 deletions (size ≥ 30bp) in whole genome comparisons of each of the other four genomes with the broccoli genome (Fig. 1C and 1D, Table S33-S44). Notably, we identified eight large inversions with sizes ranging from 134 Kb to 4.88 Mb, which were verified by

Bionano maps (Table S24, Fig. S16). Macrosynteny analysis also revealed five large inversions (size: 0.63 to 4.75 Mb) that were already identified based on whole genome comparison among the five genomes (Fig. 2C). Each pair of the five genomes shows clear 3:3 synteny patterns, with each region in one genome having up to three syntenic regions in another genome, indicating that *B. oleracea* genomes share a whole genome triplication (WGT) event (Fig. 2A and 2B, Fig. S2, Fig. S3). To detect genes that might be under selection, we calculated the rate of synonymous mutations per synonymous locus (Ks) between 1:1 orthologous genes for each pair of the five genomes. Comparable Ks distributions were identified between the ten comparisons with only one peak being found, suggesting that these genes may have diverged from the common ancestors of *B. oleracea* (Ks: 0.008~0.0111) (Fig. 2D). We then calculated the Ka/Ks ratios to find genes with evidence for selection. Because most nonsynonymous mutations are deleterious and experienced strong purifying selection (Sun *et al.*, 2018), the Ka/Ks ratios for most orthologous genes are expectedly close to zero. Approximately 13,726~17,003 genes in each of the five genomes were identified likely to be under strong purifying selection (Ka/Ks<0.6). In contrast, relatively few genes (1,737-2,038) were detected to be under positive selection (Ka/Ks>1) (Fig. 2E).

## Composition and features of *B. oleracea* pan-genome

We constructed a *B. oleracea* pan-genome comprising our five new assemblies and four published genomes (Belser *et al.*, 2018, Cai *et al.*, 2020, Guo *et al.*, 2021), all of which were assembled using long-read sequencing technology. Pairwise whole genome and gene set comparisons, and modelling of the pan-genome suggested a pan-genome size of ~653 Mb with ~84,000 genes and a core-genome size of ~458 Mb including ~33,000 genes. Both pan-genome size and pan-gene number increased when adding additional genomes, however both the core-genome size and core-gene number decreased (Fig. 3A and 3B). Using OrthoFinder, we detected 60,940 gene clusters in the *B. oleracea* pan-genome, consisting of 535,182 annotated genes in the nine genomes. Of these total gene clusters, 37,669 (~61.81%), 21,891 (~35.92%) and 1,380 (~2.27%) were considered as core, dispensable and specific gene clusters, respectively (Fig. 3C and 3D). The proportion of core gene clusters was much higher than the percentages for dispensable and specific gene clusters. We found 48,543-49,303 core genes representing 79.53%-80.05% of total gene models in our five newly generated genomes. A total of 62-117 specific genes and 530-707 orphan genes (genes with no paralogues and no orthologues in other genomes) were detected in the five genomes.

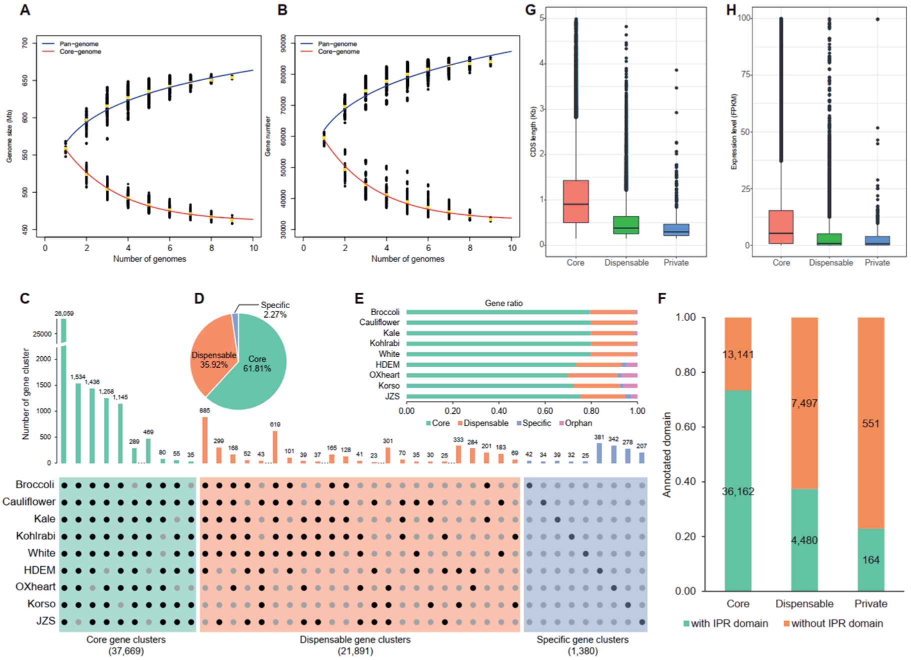


**Fig. 3 Pan-genome analyses of nine *Brassica oleracea* genomes.** (**A**) Pan-genome and core-genome modelling for sequences, which were based on pairwise whole-genome comparisons across nine genomes. (**B**) Pan-genome and core-genome modelling for gene space, which was based on gene set comparisons across nine genomes. (**C**) Compositions of the *B. oleracea* pan-genome. The histograms show subsets of core, dispensable and specific gene clusters. The dots in between the histogram bars refer to other combinations that are not displayed (see Table S45). (**D**) Ratio of gene clusters marked by each composition. (**E**) Ratio of classified genes in each genome. Orphan genes are those with no paralogues and no orthologues in other genomes. (**F**) Proportion of genes with InterPro (IPR) domains in the Broccoli core genes, dispensable genes and private genes. Private genes include genes within specific gene clusters and orphan genes. (**G**) CDS length of core, dispensable and private genes in Broccoli genome. (**H**) Expression level of core, dispensable and private genes in Broccoli genome.

We found that 73.35% of the core genes in the broccoli genome were annotated with InterPro domains. This proportion was much higher than the percentages for the dispensable and private genes (including specific genes and orphan genes), which accounted to 37.41% and 22.94% resp. (Fig. 3F). The average CDS length of core genes are significantly longer than those of less conserved genes (Student-Newman-Keuls test with *a* = 0.01) (Fig. 3G). Additionally, the average gene expression level of the core genes was significantly higher than those of dispensable and private genes (Student-Newman-Keuls test with *a* = 0.01) (Fig. 3H). Gene Ontology (GO) enrichment analysis showed that core genes were mainly enriched in essential biological processes, including regulation of metabolic processes, biosynthetic process and transcription, host cellular component and nucleotide binding process (Fig. S4A). However, dispensable genes were mainly enriched in terms of cellular respiration, translation, organelle component and ribonuclease activity. Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analyses showed that core genes were enriched in pathways related to transcription factors, specific metabolism, transporters, replication and repair, whereas dispensable genes were mainly enriched in pathways related to ribosome, translation, photosynthesis, oxidative phosphorylation (Fig. S4B).

### Different evolutionary dynamics of LTR-RTs in *B. oleracea* and *B. rapa*

We detected 4,739-5,995 intact LTR-RTs, totalling 32.04-40.57 Mb in each of the nine *B. oleracea* long-read assemblies, however, only 690 and 1,482 (3.42 and 8.13 Mb) intact LTR-RTs were found in two short-read assemblies (JZSv1 and TO1000) (Table S19). The remarkably less intact LTR-RTs in short-read assemblies could be due to the collapse of short reads from these regions. Copia-like LTR-RTs are more abundant than Gypsy-like LTR-RTs in all *B. oleracea* genomes (Fig. 4A). The nine long-read assemblies showed similar LTR-RTs length distribution. Generally, Copia-like LTR-RTs peaked at around 5-6Kb, while the lengths of Gypsy-like LTR-RTs were more variable (Fig. S5). On average, the lengths of Gypsy-like LTR-RTs are longer than those of Copia-like LTR-RTs. In the published 18 *B. rapa* genomes, we only identified 2,573-3,958 intact LTR-RTs, totalling 17.19-26.33 Mb (Fig. 4A and 4B, Table S20). Statistically, *B. oleracea* accumulated significantly more LTR-RTs than *B. rapa* (Fig. 4A and 4B), consistent to a previous study (Liu *et al.*, 2014a). The length distributions of LTR-RTs among the 18 *B. rapa* genomes are similar (Fig. S6), which are also comparable to those of *B. oleracea*.

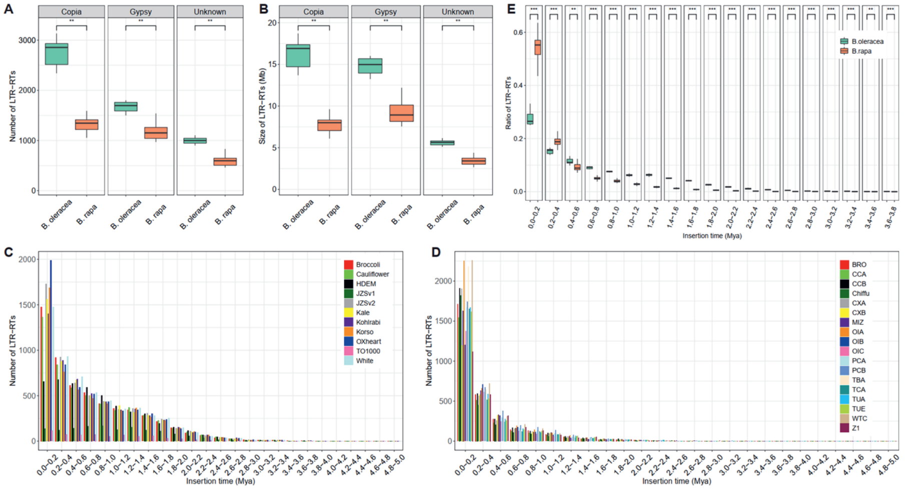


**Fig. 4 Evolutionary dynamics of full-length LTR-RTs in *Brassica oleracea* and *Brassica rapa*.** The number (**A**) and total genome size (**B**) of full-length LTR-RTs in *B. oleracea* and *B. rapa* pan-genome. Distributions of insertion time dated by divergence rate of full-length LTR-RTs (the rate of neutral mutation sites accumulated in the two terminal repeat sequences of LTR-RTs). This was calculated in the (**C**) 11 genome assemblies of *B. oleracea* and (**D**) 18 genome assemblies of *B. rapa*. (**E**) Ratio of LTR-RTs that were accumulated in each time range in *B. oleracea* and *B. rapa* genomes. JZSv1 and TO1000 were not included in this calculation. Significance is determined by a two-sided Wilcoxon rank-sum test.

Analysis of LTR insertion time revealed continuous LTR-RTs expansion in all nine long-reads assembled *B. oleracea* genomes since ~3 Mya (Fig. 4C), after the speciation with its sister species of *B. rapa* (Cheng *et al.*, 2017). We identified two LTR-RTs burst events in *B. oleracea* genomes, a “young” event that has taken place in eight long-read assembled genomes (not present in HDEM) around 0-0.2 Mya, and an “old” event that occurred in all nine genomes around 1.2-1.4 Mya, with 25.07%-33.14% and 5.58%-6.90% LTR-RTs being formed at these two time ranges, respectively (Fig. 4E). Generally, the evolutionary dynamics of LTR-RTs in *B. oleracea* genomes are similar. Nevertheless, the remarkable LTR-RTs number variation, especially for the newly inserted LTR-RTs, among *B. oleracea* genomes may indicate different changes in LTR-RTs patterns during the intraspecific diversification. As an illustration, we found less than twice as many newly inserted LTRs in HDEM than in the other eight genomes. In *B. rapa*, continuous LTR-RTs expansion initiated later, at ~2 Mya (Fig. 4D). Only one LTR-RTs burst event was found in each of the 18 *B. rapa* genomes, which is shared in time with the “young” LTR-RTs burst event in *B. oleracea*, with significantly higher percentage of LTR-RTs

(43.53%-63.55%) being accumulated during 0-0.2 Mya than in *B. oleracea* (Fig. 4E). In *B. rapa*, the vast majority of LTR-RTs (66.30%-79.22%) were formed at the time range of 0-0.4 Mya, whereas only 28.23%-47.16% LTRs were accumulated in *B. oleracea* at the same time range (Fig. 4E). *B. oleracea* accumulated significantly higher percentage of LTRs than *B. rapa* from ~3.8 Mya until ~0.4 Mya (Fig. 4E). Together, these data suggested that LTRs accumulated faster in the recent ~0.6 Mya and later in *B. rapa* than in *B. oleracea*, indicating different evolutionary dynamics of LTR-RTs in the two sister species.

### Faster gene loss in *B. rapa* than in *B. oleracea* before intraspecific diversification

Using a pan-genome strategy (Cai *et al.*, 2021), we inferred the ancestral genome of *B. oleracea* ($\mathrm{A_{Bol}}$). A total of 33,287 WGT-derived genes (14,153, 10,192, and 89,42 genes in LF, MF1 and MF2 subgenomes, respectively) formed $\mathrm{A_{Bol}}$, 3,121 more than in the inferred *B. rapa* ancestral genome ($\mathrm{A_{Bra}}$) (Cai *et al.*, 2021). Sliding windows with 500 genes and an increment of two genes were used to calculate gene densities in the three subgenomes. Similar to its sister genome $\mathrm{A_{Bra}}$, we discovered the phenomenon of subgenome dominance in $\mathrm{A_{Bol}}$, with average gene densities of 0.745, 0.537 and 0.470 in LF, MF1 and MF2 subgenomes respectively (Fig. S7A). Additionally, using the broccoli genome as a representative of the diverse morphotypes, we found significantly lower gene densities in all its three subgenomes compared to $\mathrm{A_{Bol}}$. On average, 0.075, 0.091 and 0.110 genes in broccoli LF, MF1 and MF2 subgenomes were fractionated (Fig. S7B). The gene density and gene fractionation distributions along the seven AKBr chromosomes among the nine *B. oleracea* extant genomes were similar (Fig. S12 and S13), consistent with the broccoli representative genome. Our results suggest that genes were extensively fractionated in all the individual genomes of *B. oleracea* during the intraspecific diversification (Fig. S13). Gene fractionation patterns along 24 ancestral karyotype (AK) blocks for nine *B. oleracea* genomes representing diverse morphotypes were also comparable in all three subgenomes, with blocks D and G of the LF subgenome showing much higher levels of gene loss (Fig. S8). Most *B. rapa* genomes, also representing diverse morphotypes, also showed similar gene fractionation patterns along the seven AKBr chromosomes as well as the 24 AK blocks (Fig. S9, Fig. S14, Fig. S15), which were also generally in agreement with those of *B. oleracea*. However, gene fractionation ratios in specific blocks for several *B. rapa* genomes differed remarkably from the other *B. rapa* genomes, such as blocks G, N, T and V of OIC's (*ssp. oleifera*, Rapid cycling's) LF subgenome, block I of TUE's (*ssp. rapa*, European Turnip's ) MF1 subgenome, block V of CXA's (*ssp. parachinensis*, Caixin's) MF2 subgenome, and etc (Fig. S9). Interestingly, the blocks D and G showed increased gene fractionation in the LF subgenome of almost all *B. rapa* and *B. oleracea* morphotypes.

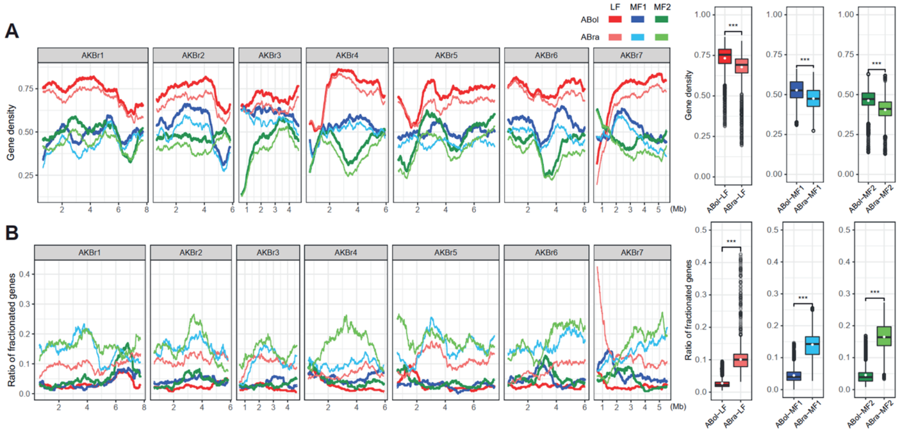


**Fig. 5 Less gene fractionation in ancestral genome of *Brassica oleracea* ($A_{Bol}$) than in ancestral genome of *Brassica rapa* ($A_{Bra}$).** (**A**) Gene density distribution on the seven inferred chromosomes of AKBr in the three subgenomes of $A_{Bol}$ and $A_{Bra}$. The figure on the right shows gene density in each window. Two-tailed Student's t-test was performed to compare gene densities between $A_{Bol}$ and $A_{Bra}$ for each subgenome. (**B**) Gene fractionation distribution in the three subgenomes of $A_{Bol}$ and $A_{Bra}$. The figure on the right shows gene fractionation ratio in each window. Two-tailed Student's t-test was performed to compare gene fractionation ratio between $A_{Bol}$ and $A_{Bra}$ for each subgenome. 500-gene windows with an increment of two genes was used to calculate gene density and gene fractionation ratio in (**A**) and (**B**).

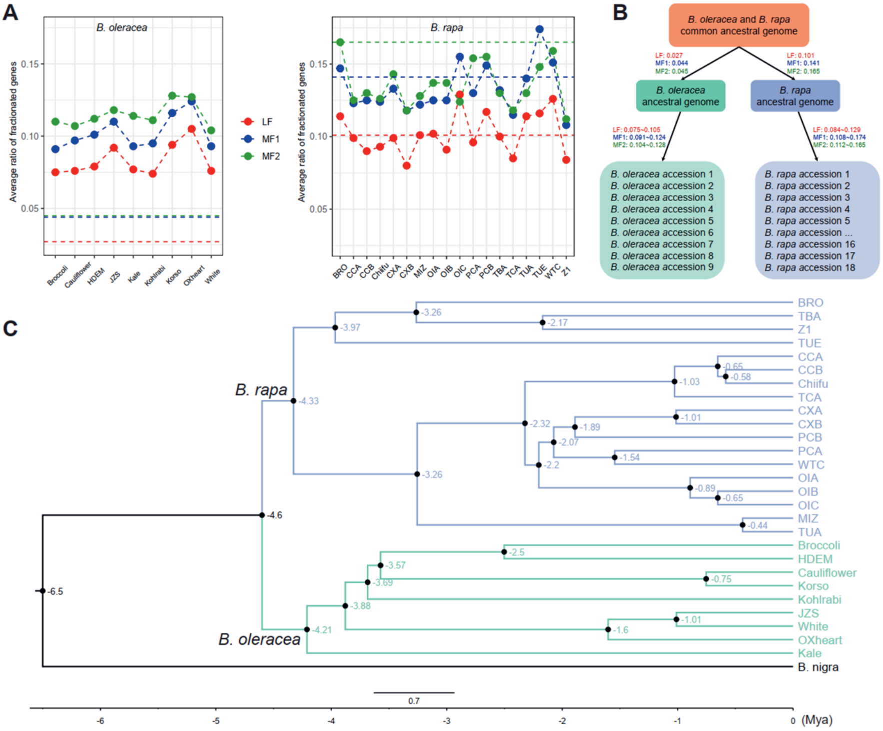


**Fig. 6 Different gene fractionation patterns between *Brassica oleracea* and *Brassica rapa*. (A)** Average gene fractionation ratio of nine *B. oleracea* and 18 *B. rapa* accessions in the three subgenomes. Red, blue and green horizontal dashed lines represent average gene fractionation ratios of ancestral genome of *B. oleracea* (left) and ancestral genome of *B. rapa* (right) relative to $A_{Bol\_Bra}$. (**B**) Gene fractionation patterns from $A_{Bol\_Bra}$ to extant genomes of *B. oleracea* and *B. rapa*. Values next to arrows represent average gene fractionation ratios from the ancestral genome to its descendant genome in the three subgenomes. (**C**) Phylogenetic tree of 28 *Brassica* genomes and their estimated divergence times (million years ago).

We further constructed the common ancestral genome of *B. oleracea* and *B. rapa* ($A_{Bol\_Bra}$), a gene repertoire consisting of 34,547 WGT-derived genes (14,539, 10,659 and 9,349 in LF, MF1 and MF2, respectively), by merging genes in $A_{Bol}$ and $A_{Bra}$ and ordering the non-redundant genes in the tPCK karyotype. The average gene densities relative to $A_{Bol\_Bra}$ for all three subgenomes of $A_{Bol}$ (0.745, 0.537 and 0.470 in LF, MF1 and MF2, respectively) were significantly higher than those of $A_{Bra}$ (0.727, 0.507 and 0.435 in LF, MF1 and MF2, respectively) (Fig. 5A). In agreement with this, the average gene fractionation ratios in three subgenomes of $A_{Bol}$ (0.027, 0.044 and 0.045 in LF, MF1 and MF2, respectively) were significantly lower than those of $A_{Bra}$ (0.101, 0.141 and 0.165 in LF, MF1 and MF2, respectively) (Fig. 5B), which was also reflected in the 24 AK blocks (Fig. S10). This suggests that after the speciation between *B. oleracea* and *B. rapa*, WGT-derived gene loss in *B. rapa* was stronger than that in *B. oleracea*. We further investigated gene fractionation in individual genomes of *B. oleracea* and *B. rapa*. In contrast to the different fractionation rates of their respective ancestral genomes, similar average gene fractionation ratios were identified between the 18 genomes of *B. rapa* (0.084-0.129 in LF, 0.108-0.174 in MF1 and 0.112-0.165 in MF2, respectively) and the nine genomes of *B. oleracea* (0.075-0.105 in LF, 0.091-0.124 in MF1 and 0.104-0.128 in MF2, respectively) (Fig. 6A and 6B).

A phylogenetic tree including the genomes of 18 *B. rapa*, nine *B. oleracea* and one *B. nigra* (outgroup) was constructed using 4,756 single-copy orthologous genes. We revealed two main lineages: one for *B. oleracea* and the other one for *B. rapa*, both of which were supported by high bootstrap values (Fig. S11). Within the *B. oleracea* lineage, we revealed the two main cultivated lineages: “Arrested Inflorescence Lineage (AIL)” and “Leafy Head Lineage (LHL)”, with kohlrabi situated at the junction of these two lineages, consistent with an earlier report (Cai *et al.*, 2022a). Our phylogenetic tree also supported the hypothesis that kale lineage leads to the “AIL” and “LHL” (Cai *et al.*, 2022a) (Fig. 6C, Fig. S11). The genealogical relationships within *B. oleracea* were consistent with morphotypes rather than geographical origin. The *B. rapa* lineage in our analysis largely corroborated the one revealed by Cai et al (Cai *et al.*, 2022c) with minor differences, both of which revealed five monophyletic groups for these diverse *B. rapa* morphotypes. Estimation of divergence times based on Bayesian inference using MCMCTree program suggested that intraspecific diversification for the sister species *B. oleracea* (~4.21 Mya) and *B. rapa* (~4.33 Mya) happened simultaneously and shortly after their speciation (~6.5 Mya) (Fig. 6C). Together with the gene fractionation ratios as observed in this study, these results suggest that *B. rapa* experienced faster WGT-derived gene loss than *B. oleracea* before their intraspecific diversification, however, during the intraspecific diversification, *B. oleracea* and *B. rapa* experienced gene loss at a comparable speed.

### Biased gene loss within and between *B. oleracea* and *B. rapa*

A total of 1,260 (386, 467 and 407 in LF, MF1 and MF2, respectively) WGT-derived genes were lost in the three subgenomes of $A_{\mathrm{Bol}}$ in relation to $A_{\mathrm{Bol\_Bra}}$ (Fig. 7A). Among these non-redundant genes, we observed that 97.2% lost one copy of their paralogues, 2.6% lost two copies and 0.2% three copies, suggesting a strong bias towards losing only one copy among the three paralogous genes. In comparison, much more genes were lost in $A_{\mathrm{Bra}}$, with 1,423 1,477 and 1,481 genes in LF, MF1 and MF2, respectively. A total of 71.1% non-redundant genes lost only one copy of the three paralogous genes in $A_{Bra}$, again suggesting biased loss of only one paralogous copy. Interestingly in $A_{Bra}$, we observed a surprisingly high ratio (26.1%) of non-redundant genes that had lost all the three copies of paralogous genes (Fig. 7B), which is a major factor contributing to the faster gene loss in $A_{Bra}$ than $A_{Bol}$. Gene ontology (GO) enrichment analysis of these genes of which all three copies were lost (738 genes) showed that the vast majority of enriched GO terms were associated with biological process, including biological regulation, response to stimulus, response to chemical, response to phytohormones (auxin, cytokinin and jasmonic acid) and etc (Table S25), which suggests a large amount of ancestral *B. rapa*-specific relaxations of biological or environmental constraints leading to co-elimination of all three copies of paralogous gene. In addition, the two species also showed biased patterns of gene loss with respect to the lost gene functions. Enriched GO terms of lost genes in $A_{Bol}$ were mainly related with 'housekeeping roles', such as DNA damage response, ATP/ADP binding, nuclease activity and other essential cellular functions (Table S26-S28). In $A_{Bra}$, besides functional GO categories with 'housekeeping roles', the vast majority of enriched GO terms of the lost genes belong to biological process including those likely related to the adaptations to changes in environmental conditions (Table S29-S31). This suggests that $A_{Bra}$ suffered relaxation of biological or environmental constraints more so than $A_{Bol}$, indicating species-specific adaptation to the environment through adaptive gene loss. During intraspecific diversification, the extant genomes of *B. oleracea* and *B. rapa* both showed clear bias to one-copy gene loss and remarkably few three-copy gene loss (Fig. 7C and Fig. S17). We then investigated gene loss bias between each two extant individual genomes of *B. oleracea* or *B. rapa*. Among all *B. oleracea* pairwise genome combinations, we found that on average only 33.42%, 37.43% and 37.81% common genes were lost in LF, MF1 and MF2, respectively (Fig. 7D, Fig. S18). These figures were relatively higher in *B. rapa*, with 38.88%, 43.38% and 44.88% in LF, MF1 and MF2, respectively (Fig. 7D, Fig. S19). These results suggest that gene loss remains biased among extant genomes during the intraspecific diversification. Among all pairwise genome comparisons between extant *B. oleracea* and *B. rapa* genomes, we found on average only 11.25%, 11.49% and 11.89% shared gene loss in LF, MF1 and MF2, respectively (Fig. 7E, Fig. S20). Taken together, our results suggest the continuing gene loss bias, both within and between species, during intraspecific diversification of *B. oleracea* and *B. rapa*.

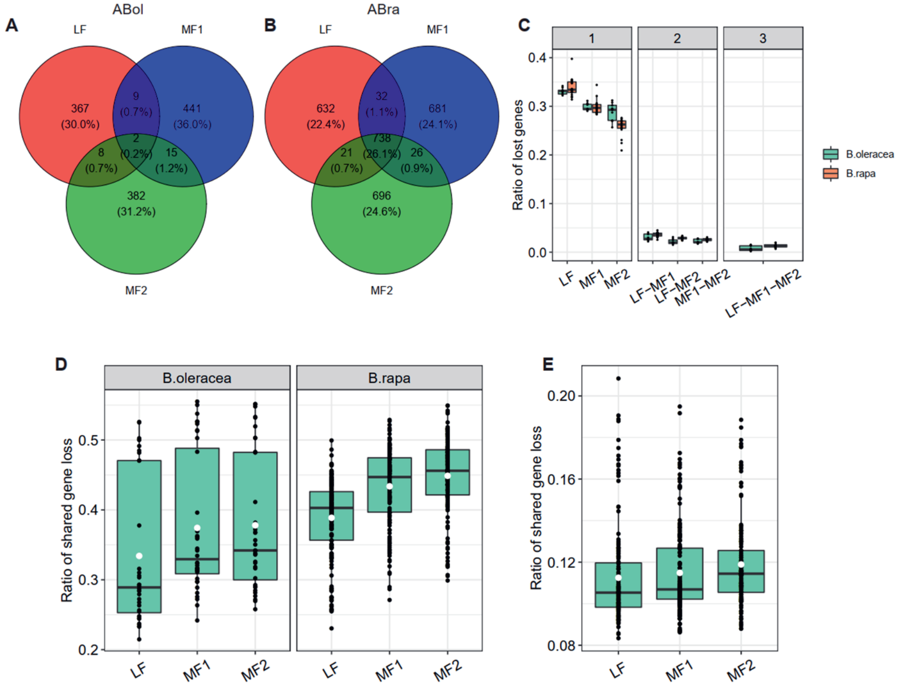


**Fig. 7 Biased gene loss in *Brassica oleracea* and *Brassica rapa*.** Common and unique gene losses among the three subgenomes of $A_{Bol}$ (**A**) and $A_{Bra}$ (**B**). (**C**) The ratio of one-copy, two-copy and three-copy gene loss in the extant genomes of *B. oleracea* or *B. rapa*. Each black dot represents one individual genome. (**D**) Distribution of shared gene loss between each two of the extant genomes of *B. oleracea* or *B. rapa* in the three subgenomes. Each black dot represents one pairwise combination of the nine *B. oleracea* or 18 *B. rapa* genomes. (**E**) Distribution of shared gene loss between extant genomes of *B. oleracea* and *B. rapa* in the three subgenomes. Each black dot represents one pairwise combination of the nine *B. oleracea* and 18 *B. rapa* genomes. White dots indicate the average value in (**D**) and (**E**).

## Discussion

In this study, we *de novo* assembled chromosome-scale reference genomes for five different *B. oleracea* morphotypes by integrating data from short-read sequencing (Illumina), long-read sequencing (Oxford Nanopore and Pacific Biosciences) and Bionano Genomics DLS optical maps. To our knowledge, these five assemblies exhibit the highest contiguity among released *B. oleracea* genomes. Comparative analysis revealed both highly syntenic relationships and extensive structural variants among the five genomes, which highlights the insufficiency of single-reference genomes to represent the sequences of a species. These five newly generated assemblies together with other published high-quality *B. oleracea* reference genomes provide an opportunity to investigate the composition and features of *B. oleracea* pan-genome via a *de novo* assembly approach (Danilevicz *et al.*, 2020). Moreover, *B. oleracea* is one of the ideal models for studying polyploidization and evolution in plants. With these genomes representing diverse morphotypes, we inferred the ancestral genome of *B. oleracea* using a pan-genome strategy (Cai *et al.*, 2021) and systematically studied the WGT-derived gene fractionation during its intraspecific diversification, as well as compared the patterns with its sister species *B. rapa*. The present work not only provides valuable genomic resources to the Brassica scientific community for *B. oleracea* improvement, but also provides insights towards understanding individual genome evolution during the intraspecific diversification of *B. oleracea* and *B. rapa*.

The completeness of a reference genome is of great importance to reliably detect LTR elements with complex structures. In the present study, we used the same approach to identify intact LTR-RTs in 11 *B. oleracea* genomes (nine long-read and two short-read assemblies, respectively) and 18 *B. rapa* genomes. By comparing the LTR-RTs content between the nine long-read and the two short-read *B. oleracea* assemblies, we showed that much more intact LTR-RTs were detected in long-read assemblies, with the sizes being 9.37-11.86 times larger than TO1000 genome and 3.94-4.99 times larger than JZS v1 genome. More complete detection of LTR-RTs also provided new insights into the evolution of LTR-RTs in *B. oleracea*. A “young” LTR outbreak event was identified in eight *B. oleracea* genomes, however this event was not observed in the two short-reads assemblies probably because these LTR-RTs were not successfully assembled in these two genomes. This “young” LTR outbreak is also not identified in the HDEM genome, however, we cannot rule out whether this is due to assembly artifacts. Transposable elements (TEs) have played an important role in the evolution of plant genomes (Lisch, 2013, Chuong *et al.*, 2017, Cai *et al.*, 2020, Liu *et al.*, 2021a). Among different categories of TEs, LTR-RTs have been shown to contribute significantly to genome size expansion in plants owing to their high copy number and large size (Rensing *et al.*, 2008, Schnable *et al.*, 2009, Nystedt *et al.*, 2013, Ming *et al.*, 2015, Ou and Jiang, 2018). It is reported that LTR-RTs are highly unstable in plant genomes (SanMiguel *et al.*, 1998, Domansky *et al.*, 2000, Devos *et al.*, 2002, Liu *et al.*, 2021a), and it is often observed that LTR-RT components vary among different subspecies (Cai *et al.*, 2020, Sun *et al.*, 2022b). Our results also showed that LTR-RTs components strongly differed in both *B. oleracea* and *B. rapa* genomes, suggesting LTR-RTs as important drivers for intraspecific diversification. Assuming that *B. oleracea* and *B. rapa* diversified at ~4.6 Mya (Cheng *et al.*, 2017), we would conclude that nearly all LTR-RTs in these two species were inserted during the period of intraspecific diversification (Fig. 4A and 4B) and no LTR-RTs were inherited from their common ancestor. A burst of transposable elements has been reported in connection with taxonomic groups and species formation as well as domestication (Belyayev, 2014). We see different patterns of both LTR-RTs and gene loss dynamics in *B. rapa* and *B. oleracea* extant genomes, which should be studied in depth to understand the diversification history of these two species.

Single-reference genomes are not sufficient to cover the whole genome sequence of a species. This has been confirmed by pan-genome studies in major crops, such as soybean (Liu *et al.*, 2020), rice (Qin *et al.*, 2021), maize (Hufford *et al.*, 2021) and rapeseed (Song *et al.*, 2020). Previously, a *B. oleracea* pan-genome including nine varieties and a wild relative was constructed using an iterative assembly approach with NGS data. Even though nearly 18.7% of the pan-genome is composed of variable genes, it is still likely that genomic variations resolved in this pan-genome are underestimated. First reason is that numerous complex variations cannot be detected by simply mapping short reads to the reference genome (Liu *et al.*, 2020). Second reason is that short reads often result in incomplete assemblies. Indeed, in our research, we found ~35.92% dispensable gene clusters and ~2.27% specific gene clusters in the pan-genome constructed based on high-quality *de novo* assembled sequences, which have the potential to resolve the vast majority of genomic variations. More than 20% of all genes in each individual genome were assigned as dispensable, specific or orphan, suggesting that SVs widely exist between different morphotypes within *B. oleracea*. Pan-genome modelling suggested a *B. oleracea* pan-genome size of ~653 Mb and a core-genome size of ~458 Mb. *B. oleracea* includes many morphotypes with enormous phenotypic diversity, however in this pan-genome analysis the nine genomes cover only five morphotypes. Inclusion of more morphotypes or increasing the number of samples likely leads to a larger estimate of the pan-genome size. Our five new assemblies are nearly identical in terms of ratio between each of the core, dispensable, specific and orphan genes (Fig. 3E). However, these figures strongly differ among the four published genomes which were generated by different labs with different approaches  (Fig. 3E). This inconsistency might have been induced by different genome assembly and gene prediction approaches. To minimize such effects in a pan-genome study, by best solution is to include a large panel of samples in a single project to capture the whole pan-genome and implement same approach for assembly and gene model prediction, which is however costly. Nevertheless, the pan-genome dataset constructed from this study is valuable for important functional genes and genetic breeding studies as it provides comprehensive genomic variations in the gene pool for *B. oleracea*.

*B. oleracea* and *B. rapa* are both mesopolyploid species that have been domesticated into a remarkable variation in morphotypes, with genomes that have experienced a triplication event, followed by extensive gene fractionation and chromosome rearrangements (Cheng *et al.*, 2012a). In *B. rapa*, gene fractionation during individual genome evolution has been investigated by constructing an inferred ancestral genome based on a pan-genome approach. Cai and colleagues observed the continuing influence of the dominant subgenome on *B. rapa* intraspecific diversification (Cai *et al.*, 2021). The *B. oleracea* pan-genome present in this study allowed us to infer the ancestral genome of *B. oleracea,* as well as a common ancestral genome of *B. oleracea* and *B. rapa*. Instead of using a single reference genome for each species, we constructed the common ancestral genome by including nine *B. oleracea* and 18 *B. rapa* genomes. This remarkably improved the common ancestral genome since gene content is added from diverse *B. oleracea* and *B. rapa* genomes, as was discussed by Cai et al (Cai *et al.*, 2021). The three inferred ancestral genomes provide the opportunity to systematically study gene fractionation patterns both before and during intraspecific diversification of the two species. Our data suggest that the ancestral genome of *B. rapa* has undergone stronger gene fractionation than the ancestral genome of *B. oleracea*. Based on divergence time estimation, the ancestors of *B. oleracea* and *B. rapa* had similar subspeciation initiation times. This brings us to conclude that the ancestral genome of *B. rapa* has experienced faster gene loss than *B. oleracea*. This largely attributes to the extensive loss of all three -copies of many genes in the *B. rapa* ancestral genome. We hypothesize that the ancestral *B. rapa* has undergone its specific relaxations of given biological or environmental constraints, which result in the ‘co-elimination’ of all the three copies of redundant genes that were triplicated by the WGT. As expected, we observed the continuing influence of the dominant subgenome on intraspecific diversification in *B. oleracea*, similar as in *B. rapa*. Different from the diversification of the two ancestral genomes, extant genomes of the two species show comparable speed of gene loss during their intraspecific diversification, which is likely driven by human domestication. Our data support the finding that gene loss is biased towards both genomic position and function. Indeed, the dominant LF subgenome displays significantly more genes than MF1 and MF2 subgenomes. The lost genes of ancestral genomes of *B. oleracea* and *B. rapa* differ in functional GO categories. Furthermore, both in *B. oleracea* and *B. rapa*, there is a strong bias towards one-copy gene losses, both before and during intraspecific diversification. The large fraction of unique gene losses between their extant genomes suggests the continuing gene loss bias during intraspecific diversification of the two species.

## Materials and Methods

### Plant materials and DNA sequencing

Five *B. oleracea* accessions (DH lines) representing five different morphotypes, broccoli, cauliflower, kale, kohlrabi and white cabbage, were used for sequencing and *de novo* assembly in this study.

Young leaves were collected and high molecular weight (HMW) DNA was extracted for each plant following a previously established protocol (Murray and Thompson, 1980). The SQK-LSK109 Ligation Sequencing kit (Oxford Nanopore Technologies; Oxford, UK) was used for library constructions according to manufacturer's instructions. Long-read sequencing data were generated using the Oxford Nanopore GridION platform and a run-time of 48 hr. Broccoli and kale samples were both sequenced using three flow cells, and cauliflower, kohlrabi and white cabbage samples were each sequenced using two flow cells. PacBio SMRTbell libraries were constructed from 5 µg HMW DNA using the SMRTbell Express Template Prep Kit v1 following the manufacturer's protocols. Sequencing was performed using diffusion loading on the "Sequel SMRT Cell 1M v2" with "Sequel Sequencing Kit 2.1" reagents. The concentration for sequencing was set at 8pM for all samples. To each SMRTcell, a Sequel "DNA Internal Control Complex 2.1" was added at a low percentage according to the protocol.

In addition, genomic DNA was extracted from these young leaves using a cetyltrimethylammonium bromide (CTAB) method (Allen *et al.*, 2006). Illumina libraries with ~450bp and ~600bp insertion sizes were constructed at GenomeScan, the Netherlands. The NEBNext® Ultra DNA Library Prep kit for Illumina (cat# NEB #E7370S/L) was used to process the DNA samples. Fragmentation of the DNA using the Biorupor Pico (Diagenode), ligation of sequencing adapters, and PCR amplification of the resulting product were performed according to the procedure described in the NEBNext Ultra DNA Library Prep kit for Illumina Instruction Manual. The quality and yield after sample preparation was measured with the Fragment Analyzer. The resulting libraries were sequenced on Illumina Hiseq 2500 (~450bp libraries) and X10 (~600bp libraries) platforms.

### RNA-seq sequencing

To aid in genome annotation, we generated mRNA-seq data for each of the five morphotypes. For each morphotype, whole young seedling and different tissues including leaves, meristems, curds, stems, and flowers were pooled in one mRNA-seq library (Table S21). We included young seedlings that were cultivated under normal condition and under heat treatment (35 °C). We also included leaves that were under different treatments, including normal condition, heat treatment (35°C) for 7 days, drought treatment (no water) for 7 days and cold treatment (10°C) for 7 days. Five mRNA-seq libraries were sequenced by the Illumina NovaSeq platform with 150bp paired-end reads. Raw reads were filtered using fastp (v0.19.5) (Chen *et al.*, 2018) with parameters "-q 15 -u 40 -n 5 -l 100 --trim_poly_x --detect_adapter_for_pe".

**Long-read genome assembly, polishing and quality assessment**

Porechop (v0.2.3_seqan2.1.1) (https://github.com/rrwick/Porechop) was used to remove adaptors from raw nanopore reads, and Filtlong (v0.2.0) (https://github.com/rrwick/Filtlong) was then used to filter sequences smaller than 1Kb. Three different assemblers: SMARTdenovo (Liu *et al.*, 2021b), Flye (v2.4.2) (Kolmogorov *et al.*, 2019) and WTDBG2 (v2.4.1) (Ruan and Li, 2020), were tested with broccoli nanopore reads that were generated from the first two flow cells. SMARTdenovo was run with parameters "-c 1" to generate consensus sequences and "-k 17" to follow developers' advices for large genomes. Flye was run with parameters "--nano-raw --genome-size 630m". WTDBG2 was run with preset2 settings "-x ont -g 630m -L 5000 -p 0 -k 15 -AS 2 -s 0.05". Statistically, SMARTdenovo yielded the most contiguous assembly, with an N50 size of 4.8 Mb and just 472 total contigs (Table S22). The largest contig in this assembly was 19.2 Mb, longer than those for the other two assemblies. Assembly completeness of the three different assemblers was assessed with BUSCO (v3.0.2) (Waterhouse *et al.*, 2018). Among the three assemblers, SMARTdenovo created an assembly with the highest complete BUSCO score (Table S22). Based on these metrics, we selected SMARTdenovo to assemble the five genomes with parameters "-c 1 -k 17".

Assembled contigs were then polished using nanopore reads for two iterations, followed by Illumina reads for three iterations. For nanopore reads polishing, Minimap2 (v2.18-r1015) (Li, 2018) was used to map raw nanopore reads to raw SMARTdenovo assembly or polished assembly after first round with parameter "-x map-on". The resulting paf file was submitted to Racon (v1.3.3) for sequence polishing using default parameters (Vaser *et al.*, 2017). For Illumina reads polishing, Illumina paired-end reads were aligned to polished contigs from previous iteration using bwa mem (v0.7.17-r1188) (Li and Durbin, 2009). The resulting bam file was sorted by SAMtools (v1.9) (Li *et al.*, 2009) and then subjected to Pilon (v1.23) (Walker *et al.*, 2014) with default parameters for assembly improvement.

Assembly completeness was assessed by BUSCO (Waterhouse *et al.*, 2018) (embryophyta_odb9 dataset, n=1,440) after each round of polishing. In addition, a variant calling approach was utilized to evaluate the base quality of our assemblies. Illumina short reads were mapped against the raw contig and polished contig sequences using bwa mem (Li and Durbin, 2009), and variations were called with FreeBayes (v1.3.1) (Garrison and Marth, 2012). Only biallelic variants were considered for quality assessment. The total length for each type of variants (SNPs, insertions and deletions) and the total number of bases covered by ≥3X reads were summed with SAMtools depth (Li *et al.*, 2009). Genome-wide quality value (QV) was calculated as $-10log10\left(\frac{length\ of\ variants}{\#bases\geq 3Xcoverage}\right)$ and identity was calculated as $100 * \left(1-\frac{length\ of\ variants}{\#bases\geq 3Xcoverage}\right)$ (Jain *et al.*, 2018, Michael *et al.*, 2018).

**DLS optical maps construction and hybrid assembly**

High Molecular Weight plant DNA was extracted from fresh tissue of the five morphotypes using the Bionano Prep™ Plant Tissue DNA Isolation Kit. The Direct Label and Stain (DLS) technology, together with Bionano Saphyr platform, were used for generation of optical mapping data. DLS labeling was performed with 750ng DNA using the Direct Labeling and Staining Kit (Bionano Genomics Catalog 80005) following manufacturer's recommendations. The loading of labeled DNA onto Saphyr chip and running of the Bionano Genomics Saphyr System were all performed according to the Saphyr System User Guide (https://bionanogenomics.com/support-page/saphyr-system/). The generated molecules were *de novo* assembled into genome maps using Bionano Solve Pipeline (version 3.4.1) and Bionano Access (version 1.3). "HybridScaffold" module in Bionano Solve Pipeline was then used to perform hybrid scaffolding between polished contig sequences and Bionano genome maps. As a default parameter, the hybrid scaffolding pipeline didn't fuse overlapped ONT contigs, which were indicated by the optical maps, but added a 13-bp gap between the two contigs. We checked all 13-bp gaps and aligned both 50-kb flanking regions with BLAT (Kent, 2002). The two flanking contigs were joined if one alignment was detected (Belser *et al.*, 2018). PBJelly (PBSuite_15.8.24) (English *et al.*, 2012) was further used for genome gap filling with the PacBio reads (Table S2).

**TE annotation and gene model prediction**

EDTA package (Ou *et al.*, 2019) was used to annotate and classify transposable elements for each assembly. This package selected and combined eight published programs based on benchmarking exercise of a collection of TE annotation programs. Raw candidates from base programs were further filtered to minimize the false discovery rate (Su *et al.*, 2021). Coding sequences from HDEM genome (Belser *et al.*, 2018) were provided for EDTA to remove potential gene-related sequences in the TE library.

Protein-coding gene models were predicted based on repeat-masked assemblies using a strategy that combined *ab initio*, homology-based and transcripts-based predictions. For *ab initio* prediction, Augustus (v3.3.3) (Stanke *et al.*, 2006), SNAP (Korf, 2004) and GlimmerHMM (Majoros *et al.*, 2004) were used to predict gene structures. For homology-based prediction, GeMoMa software (v1.6.3) (Keilwagen *et al.*, 2016) was applied to infer the annotation of protein-coding genes in each of our five assemblies based on protein sequences in previously published genomes, including *Arabidopsis thaliana* (TAIR10, https://www.arabidopsis.org/), *Brassica napus* (Darmor_V8.1), *Brassica napus* (Tapidor_V6.3), *Brassica nigra* (V1.1), *Brassica oleracea* (CAP0212), *Brassica oleracea* (HDEM), *Brassica oleracea* (TO1000), *Brassica rapa* (Chiffu_V3.0) and *Brassica rapa* (Z1). Besides homologous evidences, mRNA-seq data for each morphotype was also incorporated for splice site prediction. GeMoMa was run on each reference genome separately and the resulting gene predictions based on each reference genome were combined. The combined predictions were then filtered to only include complete gene models that were supported by ≥2 reference organisms or mRNA-seq data. For transcripts-based prediction, mRNA-seq reads were assembled into transcripts using two different approaches: *de novo* approach with Trinity (v2.9.1) (Grabherr *et al.*, 2011) and genome-guided approach with Hisat2 (v2.1.0) (Kim *et al.*, 2015) and Stringtie (v2.1.1) (Kovaka *et al.*, 2019). All the transcripts were subject to PASA (v2.4.1) (Haas *et al.*, 2008) for gene model prediction. Finally, EvidenceModeler (v1.1.1) (Haas *et al.*, 2008) was used to combine gene models that were predicted by the three approaches to a weighted consensus gene set.

The protein sequences of predicted gene models were aligned to Swiss-Prot and TrEMBL (Consortium, 2015) databases respectively using diamond (v0.9.32.133) BLASTP with *E* value 1 × $10^{-5}$. The motifs and domains of protein were predicted by using InterProScan (v5.42-78.0) (Jones *et al.*, 2014) with Pfam, PRINTS, ProSitePatterns, ProSiteProfiles and SMART databases. Gene Ontology (GO) (Ashburner *et al.*, 2000) terms for each gene were extracted from the output of InterProScan. KEGG (Kyoto Encyclopedia of Genes and Genomes) annotation was performed by KAAS (Moriya *et al.*, 2007). TBtools (version 1.09854) (Chen *et al.*, 2020a) was used to perform GO enrichment analysis.

**Identification of structural variations**

We aligned the other four assemblies to the broccoli genome using minimap2 (v2.18-r1015) (Li, 2018) with parameters "-ax asm5". The resulting alignments were subject to svim-asm (v1.0.2) (Heller and Vingron, 2020) to call SVs with parameters "haploid --min_sv_size 30 --max_sv_size 100000". In addition, we aligned ONT reads from the other four morphotypes to the broccoli genome using NGMLR (v0.2.7) (Sedlazeck *et al.*, 2018) and called SVs using Sniffles (v1.0.12) (Sedlazeck *et al.*, 2018) each with default parameters. To obtain high confidence SV datasets, we then used Jasmine (v1.1.0) (https://github.com/jasmine/jasmine) to merge SVs (insertions and deletions) that were called with different approach and only kept SV that were called by both approach. To identify inversions and translocations, we aligned the other four genomes to the broccoli reference genome using nucmer with parameters "-g 1000 -l 40 -c 90" and filtered the alignments using delta-filter with parameters "-m -i 90 -l 100". Syri (v1.2) (Goel *et al.*, 2019) was used to identify genomic translocations and inversions based on the alignments.

**Comparative genomics among five *B. oleracea* assemblies**

Homologous gene pairs and syntenic relationships between five *B. oleracea* genomes were identified using the MCSCAN toolkit implemented in python (https://github.com/tanghaibao/jcvi/wiki/MCscan-(Python-version)) with default parameters. Microsyntenic dot plots and block depths were generated in python using scripts from MSCAN. The resulting gene pairs were filtered out using a C-score cut-off of 0.99 to obtain 1:1 collinear gene pairs, which were used as input for macrosyntenic analysis with parameters "--minspan=30 --minsize=30". To calculate synonymous substitution rates (Ks) for homologous genes among the five assemblies, the identified 1:1 collinear gene pairs were used as input for sequence alignment that was performed using ParaAT_2.0 (Zhang et al., 2012) with parameters "-f axt -m muscle -g". Ks values were computed based on the alignments using KaKs_Calculator with the method of Nei and Gojobori (Zhang *et al.*, 2006).

**Pan-genome analysis**

OrthoFinder (v2.3.12) (Emms and Kelly, 2019) was used to detect orthologous gene clusters based on all protein sequences from nine *B. oleracea* genomes (HDEM, JZS v2.0, Korso, OX-heart and our five genomes) with default parameters. We defined core, dispensable and specific gene clusters as orthologous gene clusters that are present in ≥7 genomes, in 2-6 genomes and in only one genome respectively, and the remaining genes were defined as orphan genes. Pairwise whole-genome sequencing alignments of all possible pairs of the nine genomes were generated using nucmer program in MUMmer4 package (Marçais *et al.*, 2018). The outputs from OrthoFinder and nucmer were used for gene and sequence level pan-genome analyses, respectively, using the approach described in (Jiao and Schneeberger, 2020).

**LTR-RTs analysis**

Full-length LTR-RTs were identified by the parallel version of LTR_FINDER (v1.0.7) (Xu and Wang, 2007, Ou and Jiang, 2019) with default parameters, following which LTRharvest (Ellinghaus *et al.*, 2008) was applied with parameters "-minlenltr 100 -maxlenltr 7000 -mintsd 4 -maxtsd 6 -motif TGCA -motifmis 1 -similar 85 -vic 10". Raw intact LTR-RT candidates that were identified by the two programs were merged. LTR_retriever (v2.8.2) (Ou and Jiang, 2018) was then used to remove false positives and generate non-redundant LTR-RTs. The insertion time of each intact LTR-RT was extracted from the output of LTR_retriever, given the mutation rate of $1.5\times10^{-8}$ mutations per site per year.

***B. oleracea* subgenome construction and ancestral genome inference**

Syntenic gene pairs between nine *B. oleracea* genomes and *A. thaliana* genome were detected using SynOrths (Cheng *et al.*, 2012b). Three subgenomes (the least fractionated (LF), the medium fractionated (MF1) and the most fractionated (MF2) subgenome) of each of the nine *B. oleracea* genomes were constructed using the method reported by Cheng et al (Cheng *et al.*, 2012a). Additionally, a pan-genome based approach was used to infer the ancestral *B. oleracea* genome as reported by Cai et al (Cai *et al.*, 2021). Briefly, we merged all genes in the nine *B. oleracea* genomes that were syntenic to *A. thaliana* genome and ordered these non-redundant genes in the translocation Proto-Calepineae Karyotype (tPCK) (Cheng *et al.*, 2013).

**Phylogenetic tree construction and divergence time estimation**

OrthoFinder (Emms and Kelly, 2019) was used to determine single-copy genes between 18 *B. rapa* (Cai *et al.*, 2021), nine *B. oleracea* and *B. nigra* (Perumal *et al.*, 2020) genomes. This resulted in a total of 4,756 single-copy gene families within the 28 genomes. MAFFT (v7.402) (Katoh *et al.*, 2005) was then used to align coding sequences of the single-copy gene families, following which Gblock (v0.91b) (Talavera and Castresana, 2007) was used to extract the conserved sequences among the 28 genomes. IQ-TREE (v1.6.10) (Nguyen *et al.*, 2015) was used to construct Maximum Likelihood tree with the following parameters "-m MFP+ASC -bb 1000 -bnni". JTT+ASC+R4 was selected as the best model based on the Bayesian Information Criterion (BIC), and 1,000 replicates of ultrafast bootstrapping (UFboot) was used to estimate node support. *B. nigra* was designated the outgroup of the phylogenetic tree. Divergence times were estimated by the program MCMCtree in

PAML (paml4.9j) (http://abacus.gene.ucl.ac.uk/software/paml.html) based on the constructed phylogenetic tree. For calibration, we set the divergence time between *B. rapa* and *B. oleracea* at ~4.6 Mya, and between *B. nigra* and the common ancestor of *B. rapa* and *B. oleracea* at ~6.5 Mya (Cheng *et al.*, 2017).

**Acknowledgements**

We would like to thank Robin van Velzen (Biosystematics Group, Wageningen University and Research) for helpful discussions.

**Funding**

This research was funded by the TKI project KV 1605-004 “A *de novo* sequencing catalogue of *B. oleracea*” (https://topsectortu.nl/nl/de-novo-sequencing-catalogue-b-oleracea), and was co-supported by two breeding companies (Bejo and ENZA). CC is supported by China Scholarship Council (No. 201809110159).

**Authors’ contributions**

GB and CC designed the research. JB performed the experiments. CC and RF analysed data. CC drafted the manuscript. CC and GB revised the manuscript. All authors read and approved the final manuscript.

**Ethics approval and consent to participate**

Not applicable

**Consent for publication**

Not applicable

**Competing interests**

The authors declare that they have no competing interests.

**Supporting information**

Supplementary files are available at https://www.biorxiv.org/content/10.1101/2022.10.27.514037v1.abstract

**Fig. S1** Circos plot of genomic landscape in the five *B. oleracea* assemblies. Rings A and B represent transposable-element and gene density in sliding windows of 500Kb with step size of 100Kb. Ring C represents GC content in sliding windows of 2Mb with step size of 1Mb.

**Fig. S2** Ratio of syntenic depth between each two of the five genomes.

**Fig. S3** Homologous dot plot between each two of the five genomes.

**Fig. S4** Enrichment analysis for core and dispensable gene categories in broccoli genome. (**A**) Top 10 GO terms enriched in each GO domains (cellular component, biological process, and molecular function) in core and dispensable gene categories. (**B**) KEGG pathways enriched in core and dispensable gene categories.

**Fig. S5** Full-length LTR-RTs length distribution in 11 *B. oleracea* and one *B. rapa* genomes.

**Fig. S6** Full-length LTR-RTs length distribution in 18 *B. rapa* genomes.

**Fig. S7** Subgenome dominance phenomenon observed in *B. oleracea* during its intraspecific diversification. (**A**) Gene density distribution on the seven inferred chromosomes of AKBr in the three subgenomes of inferred ancestral genome of *B. oleracea* and broccoli. Broccoli genome was used as a representative to illustrate intraspecies diversification. The figure on the right shows gene density in each window. Two-tailed Student's t-test was performed to compare gene densities between inferred ancestral genome of *B. oleracea* and broccoli for each subgenome. (**B**) Gene fractionation distribution in the three subgenomes of broccoli. The figure on the right shows the distribution of ratios of fractionated genes to the genes in each window of the inferred ancestral genome, and the dotted line represents the average ratio in each subgenome. 500-gene windows with an increment of two genes was used to calculate gene density and gene fractionation ratio in (**A**) and (**B**).

**Fig. S8** Gene fractionation ratios of nine *B. oleracea* accessions in the 24 AK blocks

**Fig. S9** Gene fractionation ratios of 18 *B. rapa* accessions in the 24 AK blocks.

**Fig. S10** Gene fractionation ratios of ancestral genome of *B. oleracea* and *B. rapa* in the 24 AK blocks.

**Fig. S11** Phylogenetic relationships of nine *B. oleracea* and 18 *B. rapa* accessions using *B. nigra* as an outgroup. Numbers below each node show the bootstrap values.

**Fig. S12** Gene density distribution on the seven inferred chromosomes of AKBr in the three subgenomes of nine *B. oleracea* accessions. 500-gene windows with an increment of two genes was used to calculate gene density.

**Fig. S13** Gene fractionation distribution on the seven inferred chromosomes of AKBr in the three subgenomes of nine *B. oleracea* genomes. 500-gene windows with an increment of two genes was used to calculate gene fractionation ratio.

**Fig. S14** Gene density distribution on the seven inferred chromosomes of AKBr in the three subgenomes of 18 *B. rapa* accessions. 500-gene windows with an increment of two genes was used to calculate gene density.

**Fig. S15** Gene fractionation distribution on the seven inferred chromosomes of AKBr in the three subgenomes of 18 *B. rapa* genomes. 500-gene windows with an increment of two genes was used to calculate gene fractionation ratio.

**Fig. S16** Eight large inversions that were supported by Bionano maps.

**Fig. S17** One-copy gene loss bias during intraspecific diversification of *B. oleracea* and *B. rapa*.

**Fig. S18** Examples of common and shared gene loss between two extant *B. oleracea* genomes in the three subgenomes.

**Fig. S19** Examples of common and shared gene loss between two extant *B. rapa* genomes in the three subgenomes.

**Fig. S20** Examples of common and shared gene loss between extant *B. rapa* and *B. oleracea* genomes in the three subgenomes.

**Table S1** Statistics of Nanopore data (GridION) for the five morphotypes.

**Table S2** Statistics of PacBio subreads for the five morphotypes.

**Table S3** Summary of Illumina sequencing data for the five morphotypes.

**Table S4-S8** Statistics of raw contigs and contigs after each round of polishing for the five *B. oleracea* genomes.

**Table S9** Statistics of variant calling based approach to assess the quality of contig sequences.

**Table S10** Summary of Bionano DLS molecules and de novo assembled maps.

**Table S11** Summary of original ONT contigs and hybrid scaffolds.

**Table S12-S16** Repetitive elements in the five *B. oleracea* genomes.

**Table S17** Statistics of predicted genes in the five genome assemblies.

**Table S18** Statistics of gene functional annotation in the five genome assemblies.

**Table S19** Statistics of intact LTR-RTs in 11 *B. oleracea* genome assemblies.

**Table S20** Statistics of intact LTR-RTs in 18 *B. rapa* genome assemblies.

**Table S21** Tissues and treatments for mRNA-seq sequencing for the five morphotypes.

**Table S22** Statistics of broccoli genome assemblies using three different assemblers with two flowcells nanopore data.

**Table S23** Evaluation of genome assemblies based on RNA-Seq PE-reads.

**Table S24** Position of large inversions supported by Bionano maps.

**Table S25** GO enrichment analysis of the 738 three-copy lost genes in ABra.

**Table S26-S28** GO enrichment analysis of the lost genes in LF, MF1 and MF2 subgemones of ABol.

**Table S29-S31** GO enrichment analysis of the lost genes in LF, MF1 and MF2 subgemones of ABra.

**Table S32** Summary of one-to-one aligned sequences.

**Table S33-S36** InDels identified between broccoli and the other four genomes with broccoli as the reference genome.

**Table S37-S40** Inversions identified between broccoli and the other four genomes with broccoli as the reference genome.

**Table S41-S44** Translocations identified between broccoli and the other four genomes with broccoli as the reference genome.

**Table S45** Number of clusters and genes for different combinations of genomes. Source data for Fig. 3C.

# Chapter 4

## Fine mapping of meiotic crossovers in *Brassica oleracea* reveals patterns and variations depending on direction and combination of crosses

**Chengcheng Cai[1,2], Alexandre Pelé[3], Johan Bucher[1], Richard Finkers[1,4] and Guusje Bonnema[1,2,*]**

[1] Plant Breeding, Wageningen University and Research, 6708 PB, Wageningen, The Netherlands

[2] Graduate School Experimental Plant Sciences, Wageningen University and Research, 6708 PB, Wageningen, The Netherlands

[3] Laboratory of Genome Biology, Institute of Molecular Biology and Biotechnology, Adam Mickiewicz University in Poznan, 61-614 Poznan, Poland

[4] Gennovation B.V., Agro Business Park 10, 6708 PW, Wageningen, The Netherlands

[*] Corresponding author

**Published in *The Plant Journal* (2023).**

**DOI: https://doi.org/10.1111/tpj.16104**

## Summary

Meiotic recombination is crucial for assuring proper segregation of parental chromosomes and generation of novel allelic combinations. As this process is tightly regulated, identifying factors influencing rate and distribution of meiotic crossovers is of major importance, notably for plant breeding programs. However, high-resolution recombination maps are sparse in most crops including the *Brassica* genus and knowledge about intraspecific variation and sex differences is lacking. Here, we report fine-scale resolution recombination landscapes for ten female and ten male crosses in *B. oleracea*, by analyzing progenies of five large Four-Way-Cross populations from two reciprocally crossed F1s per population. Parents are highly diverse inbred lines representing major crops, including broccoli, cauliflower, cabbage, kohlrabi and kale. We produced ~4.56T Illumina data from 1,248 progenies and identified 15,353 crossovers across the ten reciprocal crosses, 51.13% of which being mapped to less than 10 Kb. We revealed fairly similar megabase-scale recombination landscapes among all cross combinations and between the sexes, and provided evidence that these landscapes are largely independent of sequence divergence. We evidenced strong influence of gene density and large structural variations on crossover formation in *B. oleracea*. Moreover, we found extensive variations in crossover number depending on the direction and combination of the initial parents crossed with, for the first time, a striking interdependency between these factors. These data improve our current knowledge on meiotic recombination and are important for *Brassica* breeders.

**Keywords:** *Brassica oleracea*, crossover, gene density, genetic background, meiotic recombination, sex difference, structural variation

## Introduction

Meiotic recombination is crucial in a plant's life cycle as it assures proper segregation of parental chromosomes during meiosis, thus guaranteeing genome integrity and stability. In addition, meiotic recombination is a key driving force for creating novel allelic combinations, enabling plant breeders to combine desired alleles and eliminate deleterious mutations (Wijnker and de Jong, 2008, Martin and Wagner, 2009, Li *et al.*, 2015). Meiotic recombination is initiated in prophase I of meiosis by the formation of hundreds of DNA Double Strand Breaks (DSBs) (Mercier *et al.*, 2015). However, this process is tightly controlled and only a small fraction of DSBs results in reciprocal exchanges between homologous non-sister chromatids, also referred to as crossovers (CO) (De Muyt *et al.*, 2009, Mercier *et al.*, 2015), due to anti-CO factors (Crismani *et al.*, 2012, Séguéla-Arnaud *et al.*, 2015, Mieulet *et al.*, 2018). Moreover, besides the mandatory CO formed per chromosome pair, which ensures their proper segregation, rarely more are observed due to the so-called phenomenon of CO interference that reduces the chance of two close-by COs (Sturtevant, 1915, Muller, 1916b, Mercier *et al.*, 2015). In most species, two types of meiotic COs co-exist (De Muyt *et al.*, 2009). Class I COs, contributing around 85%, notably rely on the ZMM complex and are subject to CO interference, whereas Class II COs, representing 15% of all COs, are unaffected by CO interference and rely on *MUS81* protein (Mezard *et al.*, 2007, Osman *et al.*, 2011, Mercier *et al.*, 2015).

COs are not uniformly distributed along the chromosomes in almost all studied species (Mézard *et al.*, 2015, Kianian *et al.*, 2018). Typically, they are concentrated in distal regions and always suppressed in centromeric and pericentromeric regions (Marand *et al.*, 2017, Dreissig *et al.*, 2019, Raz *et al.*, 2021). In plants, recombination landscapes with high CO resolution have been established in the model species *Arabidopsis* and a very limited number of crops with generally more complex genomes, like maize (Kianian *et al.*, 2018) and potato (Marand *et al.*, 2017), allowing investigation of the influence of associated genomic and epigenomic features. This is however lacking in many other genera such as *Brassica,* with CO maps of *B. rapa* (Pelé *et al.*, 2017) and *B. napus* (Boideau *et al.*, 2022) generated with limited numbers of SNP markers. It has been found that the occurrence of COs positively correlates with gene density that is generally high in distal regions and negatively with TE density, which is highest at and next to centromeres (Wu *et al.*, 2003, Erayman *et al.*, 2004, Anderson *et al.*, 2006, Dooner and He, 2008). Sequence divergence is another important genomic factor with a complex relationship with meiotic recombination, which varies across different scales and chromosomal contexts (Serra *et al.*, 2018a, Blackwell *et al.*, 2020). Small-scale sequence divergence, such as single nucleotide polymorphisms (SNPs) and small InDels, are associated with increased recombination frequency (Ziolkowski *et al.*, 2015, Lian *et al.*, 2022b). In natural populations of multiple species, positive correlations were observed between SNP density and the historical CO landscape, which is measured from linkage disequilibrium (Begun and Aquadro, 1992, Nordborg *et al.*, 2005, Spencer *et al.*, 2006, Gore *et al.*, 2009, Paape *et al.*, 2012, Cutter and Payseur, 2013). In *Arabidopsis*, juxtaposition of heterozygous and homozygous regions results in increased recombination frequency in heterozygous regions while decreased recombination frequency in homozygous regions (Ziolkowski *et al.*, 2015, Blackwell *et al.*, 2020). Blackwell *et al.* (Blackwell *et al.*, 2020) discovered a parabolic relationship between SNP density and recombination frequency, with initially a positive relationship and then a negative relationship along with the increase of SNP density. In tomato, Fuentes *et al.* (Fuentes *et al.*, 2022) found significantly positive overlap between short deletions (< 500 bp) and recombination hotspots, suggesting that small InDels do not suppress recombination. However, large-scale structural rearrangements have suppressive effects on recombination (Rowan *et al.*, 2019, Boideau *et al.*, 2022, Lian *et al.*, 2022b). In *Arabidopsis*, the ~1.2Mb inversion between Col and L*er* on chromosome 4 inhibits recombination in this region (Rowan *et al.*, 2019, Lian *et al.*, 2022b). A 70Kb transposition on chromosome 3 identified between *A. thaliana* accessions BG-5 and Kro-0 also displayed extreme local suppression of recombination (Alhajturki *et al.*, 2018, Rowan *et al.*, 2019). In *B. napus*, Boideau *et al.* (Boideau *et al.*, 2022) discovered large inversions (> 1Mb) in two most distal non-pericentromeric regions lacking recombination, suggesting that megabase-scale inversions prevent recombination. The causality for the strong correlation between sequence divergence and CO occurrence is still poorly understood. In a recent study, Lian *et al.* (Lian *et al.*, 2022b) hypothesized that polymorphisms are not causal for the shape of the megabase-scale recombination landscape in *Arabidopsis*, but on the contrary recombination contributes to shaping the sequence divergence across the genome. Besides genomic factors, epigenetic features, such as DNA methylation, histone modifications and nucleosome occupancy, also locally affect CO formation (Melamed-Bessudo and Levy, 2012, Mirouze *et al.*, 2012, Yelina *et al.*, 2012, Choi *et al.*, 2013, Habu *et al.*, 2015, Choi *et al.*, 2018).

Despite the strong regulation of CO number and distribution, extensive variation is observed both between and within species. Factors responsible for these variations have the potential to profoundly influence selective responses and facilitate adaptation (Nei, 1967, Feldman *et al.*, 1996, Coop and Przeworski, 2007), all the while being of interest for plant breeders (Wijnker and de Jong, 2008). On the one hand, a large range of external factors such as temperature fluctuation, nutritional status, or pathogen attack result in CO variations (Modliszewski and Copenhaver, 2017, Dreissig *et al.*, 2019, Henderson and Bomblies, 2021). On the other hand, intra-specific differences in CO rates are observed in the same environment. This is well illustrated in *A. thaliana* for which CO rates vary twofold among dozens of accessions tested in selected intervals (Ziolkowski *et al.*, 2015). In crops, similar observations were repeatedly made, as exemplified in maize from which the study of 23 doubled-haploid populations revealed intraspecific variation of recombination rates and landscapes (Bauer *et al.*, 2013). Marked differences in CO level and pattern were also observed between male- and female meiosis; a phenomenon referred to as heterochiasmy (Lenormand, 2003, Lenormand and Dutheil, 2005, Sardell and Kirkpatrick, 2020, Capilla-Pérez *et al.*, 2021). In *Arabidopsis* Col-0×L*er* populations, many more COs were observed in male than female meiosis (Drouaud *et al.*, 2007, Giraut *et al.*, 2011, Lian *et al.*, 2022b). Moreover, the megabase-scale recombination landscapes were remarkably different with distal regions displaying the highest recombination rates in male and the lowest in female meiocytes (Lian *et al.*, 2022a). However, this is not a universal feature of plants. Indeed, some species like *Brassica oleracea* showed a reverse pattern with more COs formed in female meiosis (Kearsey *et al.*, 1996), and others like *B. napus* or *Coffea canephora* exhibiting similar CO levels in male and female meiosis, thus no heterochiasmy (Kelly *et al.*, 1997, Lashermes *et al.*, 2001, Lenormand and Dutheil, 2005). Interestingly, in maize no differences between male and female CO levels were observed in the B73×Mo17 background, while the Zheng58×SK background revealed heterochiasmy with more COs generated during male meiosis (Kianian *et al.*, 2018, Luo *et al.*, 2019). This latter observation suggests that heterochiasmy could be genetic background dependent, however studies to support this suggestion, focussing on interaction of both factors on meiotic recombination, remain sparse.

In this study, we investigated recombination variation among ten different genetic backgrounds (hereafter also referred to as crosses/cross combinations) in *B. oleracea*, a diploid species displaying enormous phenotypic variation between different morphotypes, and studied the sex differences for each cross. To do so, we constructed five large Four-way-Cross (FwC) populations with each two F1s being reciprocally crossed per population. In total, we sequenced 1,248 progeny genomes and harvested ~4.56T Illumina data for the five FwC populations. From this fine mapping, we identified a total of 15,353 COs and characterized recombination landscapes for all ten female and male crosses. We revealed key genomic factors that shape the recombination landscape in *B. oleracea*, with gene density and large-scale structural variations (SVs) influencing genome-wide and local CO formation, respectively. While megabase-scale recombination landscapes were fairly similar among the ten sex-averaged as well as female-/male-specific crosse, we highlight extensive variations in CO number among different crosses and heterochiasmy in some cross combinations, revealing that CO variation in *B. oleracea* is shaped by genetic background, heterochiasmy and their interaction.

## Results

### CO identification

We previously generated five chromosome-scale genome assemblies of five homozygous *B. oleracea* morphotypes, including broccoli, cauliflower, kale, kohlrabi and white cabbage (Cai *et al.*, 2022b). To explore the recombination landscapes in *B. oleracea*, we constructed five Four-way-Cross (FwC) populations using the above five genotypes as founders, each including four different parents representing four different morphotypes (Fig. S1 and Fig. 1). For each FwC population, the two F1s were reciprocally crossed to analyse independently female and male meiosis (Fig. 1). The five FwC populations included a total of 1,248 progenies (Fig. 1) for which we obtained ~4.56T data through Illumina paired-end genome sequencing, with an average of 6.55-fold coverage per progeny (Table S1). By mapping reads to the broccoli reference genome that is one of the five parents, we called SNPs for the parental genomes and progenies. We selected 184,152-413,849 (0.35-0.79 SNP/Kb) segregating SNPs in each of the ten reciprocal crosses for CO analyses (Table 1), which were uniformly and genome-wide distributed across the nine chromosomes (Fig. S2).

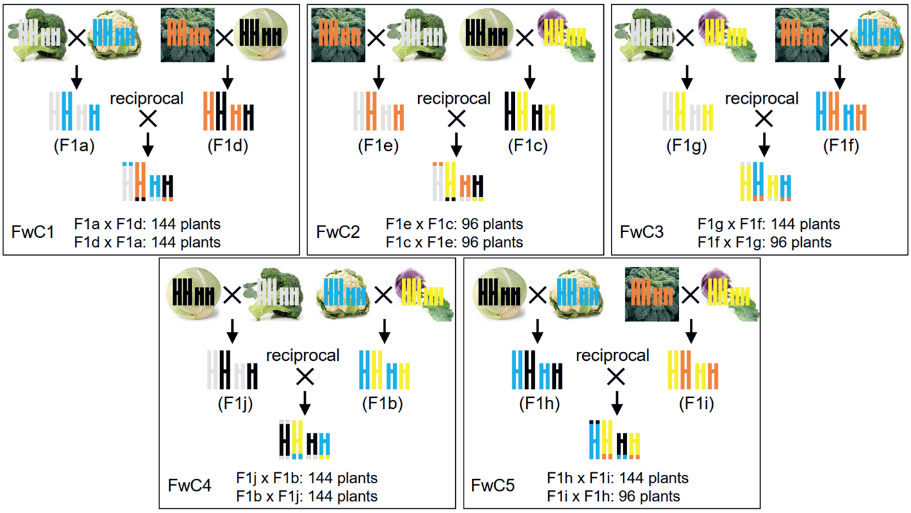


**Fig. 1 Crossing scheme with five parental essentially homozygous lines, representing five different *Brassica oleracea* morphotypes**. Five morphotypes are pairwise intercrossed to generate ten F1 hybrids (F1a-F1j). The F1s are then intercrossed in both directions to generate Four-way-Cross (FwC) populations. Meiotic recombinations of all ten combinations of the five morphotypes can be studied with these five FwC populations. The number of progenies collected for each FwC population is indicated in the figure. *B. oleracea* has nine pairs of chromosomes: here only two sets are depicted. For each FwC population, one progeny is depicted with recombined chromosomes, where colors depict parental origin.

Each set of SNPs was independently subjected to phaseLD for CO identification, which implemented sliding windows, Bayesian inference and logistic regression approaches (Marand *et al.*, 2017). In total, we identified 15,353 COs from the ten reciprocal crosses with 7,492 and 7,861 COs arising during female and male meiosis, respectively (Fig. S3 and Table 1). This translates into an average of ~6.27 and ~6.64 COs per female and male gamete, respectively. Over all ten sex-averaged crosses, we identified 5.80-7.02 COs per haploid gamete over the 2,377 (1,194 female and 1,183 male gametes) analysed (Fig. 5a). The distribution of total CO number per gamete across the ten reciprocal crosses follows a normal distribution (Fig. S4). The high SNP marker densities enabled the fine-scale identification of COs, resulting in a median resolution of 9,088 bp for the pool of ten sex-averaged crosses, with 89.16%, 80.26% and 51.13% COs having an interval resolution lower than 200Kb, 100Kb and 10Kb, respectively (Table 1). The observed distributions of CO number for all nine chromosomes in all ten sex-averaged crosses revealed that most gametes exhibit between zero and one CO per chromatid (Fig. S5a). The rare occurrence of multiple COs per chromatid suggests elevated CO interference. On average, 74.10%-96.97%

of gametes had zero or one CO in each of the nine chromosomes across the ten sex-averaged crosses, with larger chromosomes (*i.e.*, C3, C4, C5 and C9) displaying more multiple (≥2) COs (Fig. S5a). To analyse CO interference in a model-independent way, we calculated the physical distance between adjacent COs only using chromosomes of gametes having at least two CO's. In all the ten cross combinations, distributions of adjacent CO distance clearly peaked around 50-55Mb. By contrast, the corresponding "non-interference" distributions (see Methods) mainly peaked at lower values (< 5Mb), validating strong CO interference in *B. oleracea* (Fig. S6a). Accordingly, the interference strengths as indicated by Kullback-Leibler (KL) divergence among the ten sex-averaged crosses were similar (Table S2). Under the hypothesis of random CO placements and independent CO events, CO numbers per chromatid are expected to follow a Poisson distribution. Comparison between the observed CO number distributions and expected Poisson distributions revealed a deficit in gametes with 0 CO and an excess of gametes with 1 CO (Fig. S5a and S5b), fitting the occurrence of an obligate CO that ensures proper segregation of homologous chromosomes.

**Pattern of recombination landscape strongly relies on genomic features**

To compare recombination landscapes among the ten sex-averaged crosses on a genome-wide scale, we calculated the CO frequencies using 2Mb sliding windows with 50Kb steps and plotted this along *B. oleracea* chromosomes. Interestingly, we revealed remarkably similar recombination landscapes among the ten crosses deriving from different parental combinations of the five *B. oleracea* morphotypes (Fig. 2a and Fig. S7). The *B. oleracea* genome includes three acrocentric- (C6, C7 and C8) and six (sub)metacentric chromosomes (C1-C5 and C9). In all the ten crosses, highest CO frequencies were observed in distal regions of both arms in the (sub)metacentric chromosomes, whereas COs were markedly suppressed at centromeric and pericentromeric regions. For the acrocentric chromosomes, besides centromeric and pericentromeric regions, COs were also strongly suppressed in the short arm, however, CO rates were the highest in distal regions of the long arm. We found significant positive correlations between the CO landscapes of different cross combinations, in the range of 0.79-0.87 (Spearman's rank correlation, $P < 1e\text{-}4$, Fig. 2b). The conserved recombination landscapes under the given window size are independent of level of polymorphisms and structural variations between the two parental genomes.

**Table 1** Summary of crossovers (COs) identified for the five Four-way-Cross (FwC) populations.

| Population | Cross[#] | Number of gamete[$] | CO studied | Female/male | SNPs used for CO identification | Total number of CO | median resolution (bp) | No. of CO with resolution within 200Kb | No. of CO with resolution within 100Kb | No. of CO with resolution within 10Kb |
|---|---|---|---|---|---|---|---|---|---|---|
| FwC1 | F1a*F1d | 141 | Broccoli*Cauliflower | female | 263,618 | 906 | 4,463 | 859 | 790 | 541 |
| | | 141 | Kale*White | male | 298,508 | 881 | 9,684 | 803 | 725 | 445 |
| | F1d*F1a | 135 | Broccoli*Cauliflower | male | 247,307 | 974 | 6,068 | 892 | 803 | 545 |
| | | 140 | Kale*White | female | 294,283 | 862 | 5,958 | 802 | 724 | 476 |
| FwC2 | F1e*F1c | 94 | Kale*Broccoli | female | 274,274 | 679 | 8,695 | 618 | 565 | 356 |
| | | 96 | White*Kohlrabi | male | 184,152 | 573 | 30,414 | 436 | 380 | 225 |
| | F1c*F1e | 96 | Kale*Broccoli | male | 413,849 | 644 | 4,675 | 610 | 559 | 385 |
| | | 95 | White*Kohlrabi | female | 233,330 | 608 | 22,623 | 481 | 412 | 255 |
| FwC3 | F1f*F1g | 92 | Kale*Cauliflower | female | 409,758 | 482 | 10,141 | 430 | 375 | 239 |
| | | 92 | Broccoli*Kohlrabi | male | 244,283 | 643 | 11,379 | 563 | 507 | 311 |
| | F1g*F1f | 131 | Kale*Cauliflower | male | 294,161 | 811 | 11,203 | 727 | 642 | 386 |
| | | 131 | Broccoli*Kohlrabi | female | 207,906 | 922 | 12,389 | 823 | 726 | 434 |
| FwC4 | F1j*F1b | 138 | White*Broccoli | female | 331,205 | 919 | 3,923 | 861 | 795 | 555 |
| | | 133 | Cauliflower*Kohlrabi | male | 262,210 | 890 | 11,644 | 790 | 703 | 424 |
| | F1b*F1j | 135 | White*Broccoli | male | 278,731 | 918 | 8,060 | 820 | 748 | 481 |
| | | 136 | Cauliflower*Kohlrabi | female | 238,850 | 801 | 13,896 | 692 | 617 | 363 |
| FwC5 | F1i*F1h | 92 | Kale*Kohlrabi | female | 292,207 | 559 | 10,525 | 494 | 439 | 275 |
| | | 91 | White*Cauliflower | male | 352,797 | 604 | 8,018 | 529 | 488 | 326 |
| | F1h*F1i | 133 | Kale*Kohlrabi | male | 231,516 | 923 | 15,309 | 781 | 702 | 418 |
| | | 135 | White*Cauliflower | female | 317,912 | 754 | 7,028 | 677 | 622 | 412 |

[#]: F1a-F1j correspond to the ten F1 hybrids as shown in Fig. 1.

[$]: The number of gametes after removing samples based on the amount of sequencing data and total CO number per gamete (see Methods).

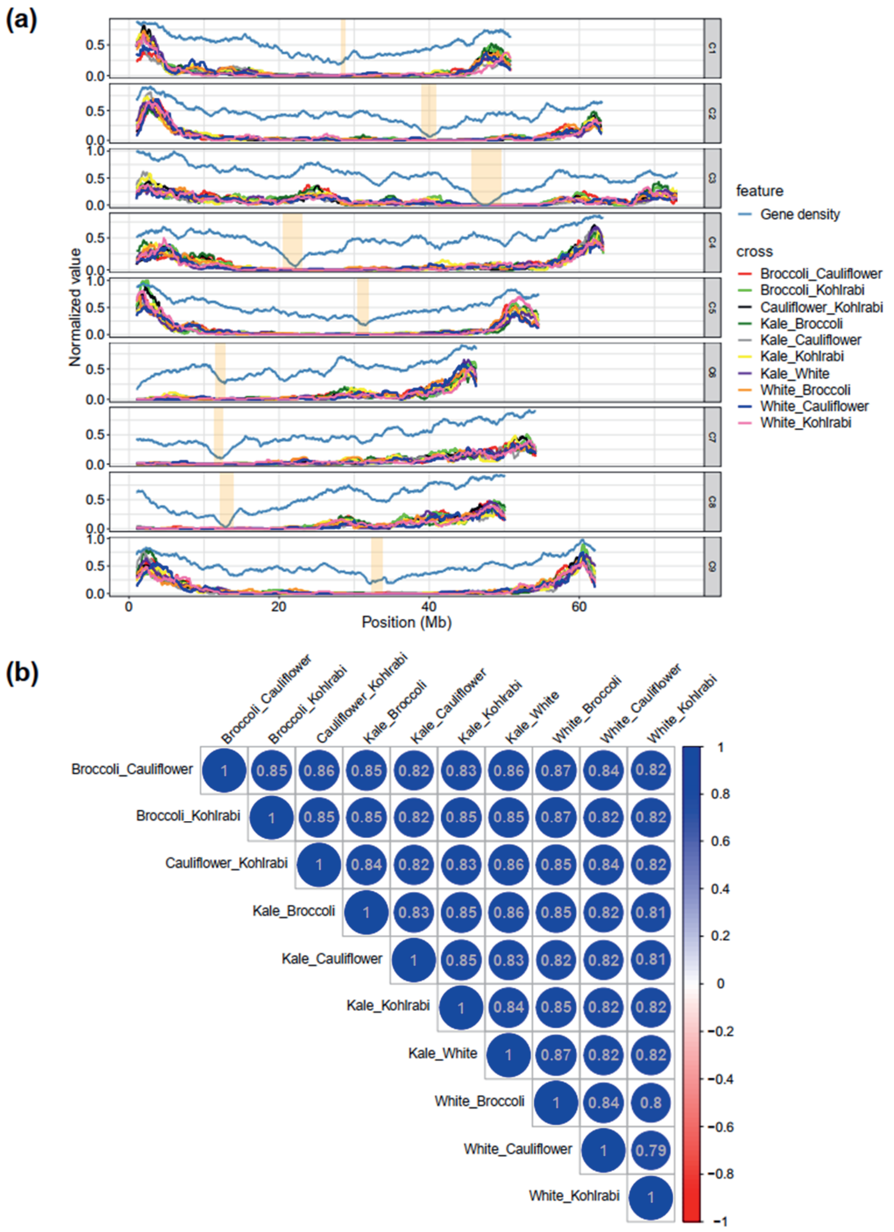


**Fig. 2 Recombination landscapes for the ten sex-averaged crosses. (a)** Crossover (CO) rate and gene density distributions along the nine chromosomes of *Brassica oleracea*. Analysis is done with 2-Mb sliding windows and 50-Kb step sizes. The centromere regions are indicated by orange shadings. CO rate and gene density values are normalized to 0 to 1. See Fig. S6 for the figures of CO

rates before normalization. **(b)** Genome-wide correlation coefficient (the Spearman's correlation) matrices among the ten sex-averaged CO distributions.

The 7,852 (480-1,086 COs in each of the ten sex-averaged crosses) fine-resolution scale COs with interval length under 10Kb were selected to analyse their potential associations with various genomic features. We found that 50.42%-57.29% of COs in each of the ten sex-averaged crosses overlapped with gene bodies (exons and introns) (Fig. 3a). Interestingly, more than half of these COs (31.54%-39.04% of total COs) overlapped with exons, while exon and intron sequences occupied only 10.92% and 9.03% space in the reference genome (Fig. 3c), respectively. In contrast, only 13.28%-18.15% of COs overlapped with TEs, which account for more than 53% of the genome. In comparison to random CO sites generated from 10,000 permutations, the observed CO sites were significantly enriched in gene bodies and their flanking 1Kb regions, but significantly depleted in TEs (empirical, $P$ < 1e-4) (Fig. 3b). Permutation tests performed by regioneR (Gel *et al.*, 2016) also suggested that genes and COs overlap significantly more than expected by chance in all crosses (Fig. S8). Together, these results suggest the regional preference of CO sites. More than 99% of fine-resolution COs in each of the ten crosses were located within 10Kb of a gene and the distributions showed a similar pattern among the ten sex-averaged crosses (Fig. 3d). Gene density distribution (2Mb sliding windows with 50Kb steps) was strongly correlated with the ten sex-averaged recombination landscapes, with high CO rates in distal gene-rich regions (Fig. 2a and Fig. S7). More interestingly, we found that "CO bumps" in regions (*i.e.*, C3, C6 and C8) that were far from distal regions in all ten crosses were also associated with increased gene density compared to that of their nearby regions (Fig. 2a).

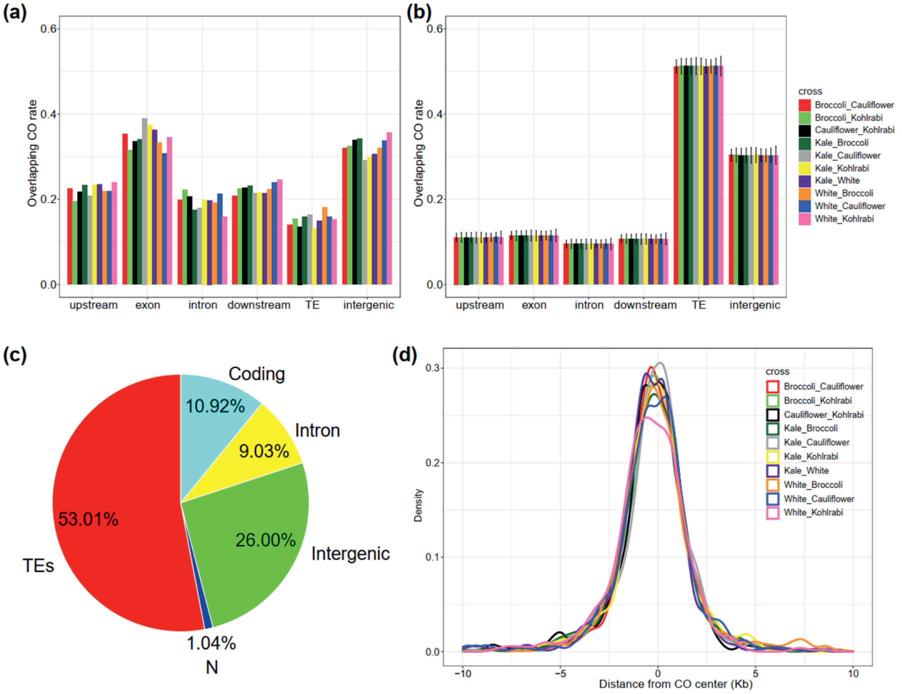


**Fig. 3 Genomic features associated with crossovers (COs) for the ten sex-averaged crosses. (a)** Overlap analysis of observed COs with different genomic features. **(b)** Overlap analysis of random regions derived from 10,000 simulations with different genomic features, with error bars denoting the standard deviation. **(c)** Percentage of genomic components in the broccoli reference genome. **(d)** Distribution of distance from each CO to the nearest gene. Note: in **(a)** and **(b)**, the CO interval was used for the overlap analysis. If an interval overlapped with multiple genomic features, the interval was counted towards each genomic feature. In **(d)**, the middle position of both CO interval and gene was used for calculating the distance.

Recombination frequency strongly correlates with sequence polymorphisms as observed in natural populations of many species, with historical recombination events positively correlated with SNP densities (Blackwell *et al.*, 2020, Lian *et al.*, 2022b). The ten genetic backgrounds in our study showed varying levels of small-scale sequence divergence (Table S3), allowing to further explore the relationship between CO occurrence and polymorphism. The distribution of parental SNPs along chromosomes showed strong local differences among the ten crosses, especially near-centromeric regions in several chromosomes (Fig. S9). Genome-wide, we observed weak but significant positive correlations between SNP density and CO rate in all the ten sex-averaged crosses (Spearman’s rank correlation, ρ = 0.15-0.30, *P* < 1e-4). Consistent with *Arabidopsis* (Blackwell *et al.*, 2020), we found a parabolic relationship between SNP density and CO rate, with high CO rate being associated with moderate SNP density (Fig. S10a). Although pericentromeric regions contribute remarkably to this relationship, a weak parabolic relationship can still be observed when excluding these regions (Fig. S10b). In our data, we did not find significantly different recombination patterns when comparing regions with most striking differences in SNP densities (Fig. S11 and S12). One reason for this is that regions with large differences in SNP densities between crosses were mainly located near centromeres where recombination is suppressed. However, even distal regions with high and different recombination rates among crosses, such as the regions of 0-2Mb and 50-54Mb on C5 of White_Cauliflower and White_Kohlrabi crosses (Fig. S12), are not associated with major differences in SNP densities.

To investigate if SVs locally affect COs, we first focused on two inversions that were validated by Bionano optical maps (Cai *et al.*, 2022b); a 4.88Mb Kale specific inversion in C3 and a 1.42Mb Cauliflower specific inversion in C7 (Fig. S13). We found only one CO inside this region in crosses involving Kale, while 13 COs were detected from crosses between parents lacking this inversion (Fig. 4a). Similarly, we did not find any CO inside the 1.42Mb Cauliflower specific inversion on C7 when crosses involved Cauliflower, whereas ten COs were observed inside this region in crosses between parents without this inversion (Fig. 4b). To examine the local suppression effect on a broader scale, we compared all the SVs reported previously (Cai *et al.*, 2022b) (Table S4) against fine-resolution COs in the corresponding four crosses (Broccoli_Cauliflower, Broccoli_Kohlrabi, Kale_Broccoli and White_Broccoli). In all cases, occurrence of large deletions (≥500bp) and inversions, translocations and transpositions independent of size translate into significantly reduced numbers of COs compared to expected by chance based on 5,000 permutations (Fig. S14-S17). However, short deletions (< 500bp) and insertions overlapped significantly more with CO sites than expected by chance.

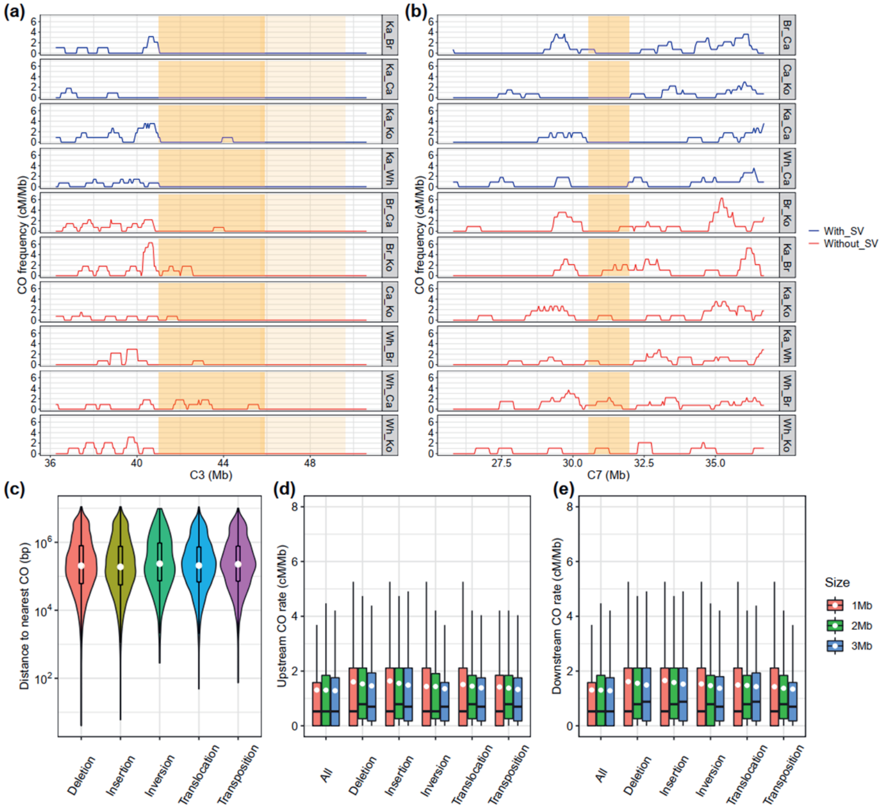


**Fig. 4 The effects of structural variations (SVs) on crossovers (COs). (a)** CO frequency in flanking regions of a 4.88-Mb kale specific inversion for the ten sex-averaged crosses. **(b)** CO frequency in flanking regions of a 1.42-Mb cauliflower specific inversion for the ten sex-averaged crosses. The inversion and centromere regions are indicated by dark and light orange shadings, respectively. **(c)** The distance of different types of SVs to their nearest CO, using the cross of Kale_Broccoli as an example. **(d, e)** The distributions of CO rates in windows of the indicated sizes in the upstream **(d)** and downstream **(e)** regions of different types of SVs, using the cross of Kale_Broccoli as an example. Note: in **(a)** and **(b)**, CO frequency was calculated using 500-Kb sliding windows with 50-Kb step sizes. Abbreviations: Br: Broccoli, Ca: Cauliflower, Ka: Kale, Ko: Kohlrabi, Wh: White Cabbage. In **(d)** and **(e)**, 'All' represents all windows genome-wide. The white dots indicate the average value in **(c)**, **(d)** and **(e)**.

While COs tend to be suppressed inside SVs, we investigated possible redistribution in flanking regions of SVs. SVs are spread all along chromosomes, and neither preferentially occur in distal arms nor in centromeric and pericentromeric regions (Fig. S18). The SV distributions, however, differ between different genetic backgrounds. From the distribution of distance to the nearest CO for each SV, we found that in all four crosses investigated, inversions displayed the largest mean distance till their nearest CO in comparison with other types of SVs (Fig. 4c and Fig. S19, Table S4). We found the largest average distance to the nearest CO for all SVs in Broccoli_Kohlrabi (906.12-1316.71Kb), followed by Kale_Broccoli (690.90-856.84Kb), Broccoli_Cauliflower (651.76Kb-749.37Kb) and White_Broccoli (575.04-668.61Kb) (Table S4), pointing to an effect of genetic background on the distance between SVs and their nearest CO. To test whether these distances are greater than expected by chance, we simulated random CO sites for the four crosses with each 10,000 times. The simulated mean distances ranged from 140.35Kb to 201.90Kb, all of which were significantly less than the observed distances (Table S4). Based on these findings, we conclude that CO suppression does not limit to SVs but also extends beyond their borders in *B. oleracea.* Interestingly, SV size has no effect on the distance to the nearest CO as revealed by correlation analyses (Spearman's rank correlation, ρ = -0.01-0.09) (Fig. S20-S23). Regarding CO rates in the flanking 1-, 2- and 3-Mb upstream- and downstream regions of SVs, inversions showed lower average CO rates than the other SV types in nearly all cases, in agreement with larger distances to the nearest CO (Fig. 4d and 4e, Fig. S24). Inversions are not more interstitially localised than other types of SVs, but randomly distributed in the genome (Fig. S18e). Thus, this lower average CO rates in flanking regions of inversions appears independent of the U-shaped CO distribution. The average CO rates in flanking regions of SVs were slightly higher than the observed genome-wide averages, implicating that loss of COs in SV regions is compensated by elevated COs in their flanking regions. Again, we found no correlation between SV size and CO rates in flanking regions (Spearman's rank correlation, ρ = -0.05-0.09), indicating that CO rates in the 1Mb regions flanking SVs didn't depend on their size (Fig. S25-S28).

### CO rate is shaped by cross combination, heterochiasmy and their interaction

Our reciprocal FwC populations allowed us to examine CO variation between different cross combinations and between female *versus* male meiosis per genetic background. We first observed strong variation in CO rate among the ten sex-averaged crosses, with the lowest average CO number per gamete in Kale_Cauliflower (5.80) and the highest in Broccoli_Kohlrabi (7.02) (Fig. 5a). Based on the CO number distribution, the ten sex-averaged crosses were classified into two groups: the "low CO rate" group including Kale_Cauliflower, White_Cauliflower, White_Kohlrabi, Kale_White and Cauliflower_Kohlrabi and the "high CO rate" group including Broccoli_Kohlrabi, Kale_Broccoli, Broccoli_Cauliflower, White_Broccoli and Kale_Kohlrabi. Within each group, no significant difference in CO number was observed between cross combinations (Student-Newman-Keuls test with a = 0.05). Four from the five crosses in the “high CO rate” group (all, except Kale_Kohlrabi) exhibited significantly more COs than four out of five crosses in the “low CO rate” group (all, except Cauliflower_Kohlrabi) (Student-Newman-Keuls test with a = 0.05). Very interestingly, we found that generally less COs were produced when Cauliflower and/or White Cabbage were present in the cross combination. By contrast, the presence of Broccoli in the background always resulted in higher CO numbers, even in crosses with Cauliflower or White Cabbage. Hierarchical clustering based on chromosome-wide sex-averaged recombination rates also revealed the above mentioned two groups, again indicating intraspecific variation of recombination rate in *B. oleracea* (Fig. 5c). Chromosome-wide sex-averaged CO rates also varied among chromosomes, with higher recombination rates in C3, C5 and C9 than in C2, C6, C7 and C8. The intraspecific and chromosome-wide variations of recombination rate were also reflected in the sex-specific recombination rates (Fig. 5d and 5e).

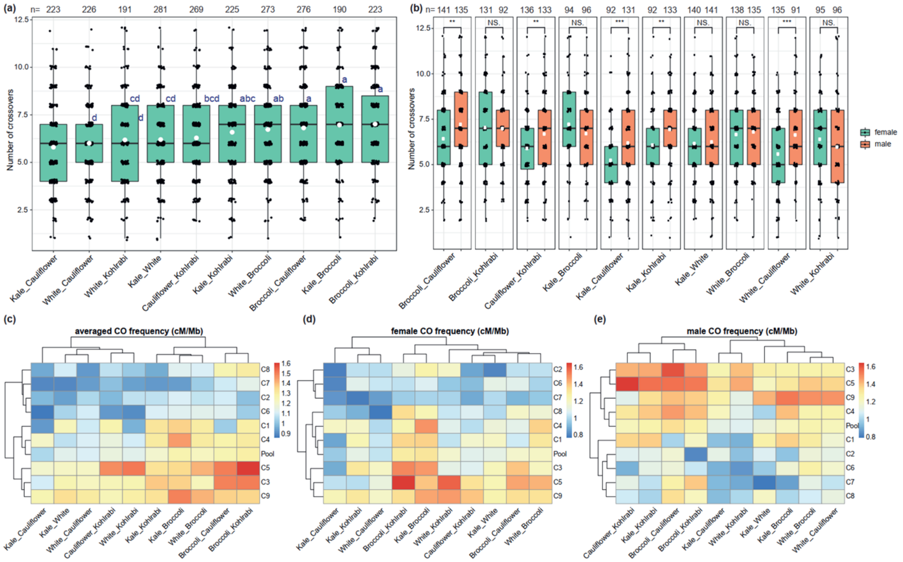


**Fig. 5 Crossover (CO) number variation and diversity of recombination frequency. (a)** Total number of COs per gamete shown for each of the ten sex-averaged crosses. Multiple comparisons were performed using Student-Newman-Keuls test with a=0.05. **(b)** The comparison of total number of COs per gamete between female and male meiosis for each cross. Asterisks indicate significant effects determined by Student’s t-test ($P$ < 0.05). **(c, d, e)** Heatmap of the chromosome-scale recombination rates measured for each chromosome in the ten sex-averaged **(c)**, female **(d)** and male **(e)** crosses. On the y-axis, ‘Pool’ represents pooled analysis of all nine chromosomes. Note: in **(a)** and **(b)**, each black dot represents a gamete of the given cross and the white dots indicate the average value.

Interestingly, when comparing CO numbers between male and female meiosis, we highlighted marked differences, and so heterochiasmy, according to the cross combinations analysed. Indeed, significant variations were observed for five out of the ten cross combinations (Broccoli_Cauliflower, Cauliflower_Kohlrabi, Kale_Cauliflower, Kale_Kohlrabi and White_Cauliflower), with always more COs formed during male than female meiosis (Fig. 5b). Strikingly, variation in CO number between female and male gametes was always significant when Cauliflower was involved in the cross combination. Together, this suggests that in *B. oleracea*, direction and combination of crosses are inter-dependent for CO variation. At chromosome scale, the average number of COs per chromatid was positively correlated with chromosome length in both sexes for all crosses (Spearman's rank correlation: 0.67-0.93, $P < 0.05$) (Fig. S29 and S30), except for male meioses of Kale_Broccoli ($P = 0.0503$) and Kale_White ($P = 0.0589$). Despite these variations, megabase-scale recombination landscapes were remarkably conserved for the ten sex-averaged and 20 sex-specific crosses, as well as between male and female meioses of each cross combination (Fig. 6a, Fig. S31-S33). Variations in CO numbers for genetic background and sex of meiosis were essentially located on chromosome extremities that always exhibited the highest CO rates. Centromeres remained deprived of COs in all cases and pericentromeric regions showed the lowest CO frequencies and variations. The female and male recombination landscapes of each cross positively correlated with each other, with correlations in the range of 0.72-0.83 (Spearman's rank correlation, $P < 1e\text{-}4$) (Fig. 6b). In both sexes, CO sites were enriched in gene bodies and their upstream and downstream 1Kb regions, whereas underrepresented in TE regions (Fig. S34a). Distribution of distances between COs and genes was also similar between female and male gametes (Fig. S34b). Moreover, the majority of COs formed per chromosome pair in both sexes were apart by large distances (50-55Mb), independent of the cross combination (Fig. S6b). We also observed rare occurrence of multiple COs per chromatid in both sexes of all ten crosses, with a deficit in gametes with 0 CO and an excess of gametes with 1 CO when comparing to the expected Poisson distributions (Fig. S35). Together with our previous comparisons between cross combinations, these results indicate that CO interference intensity remains elevated, independently of the genetic background and sex of meiosis in *B. oleracea*.

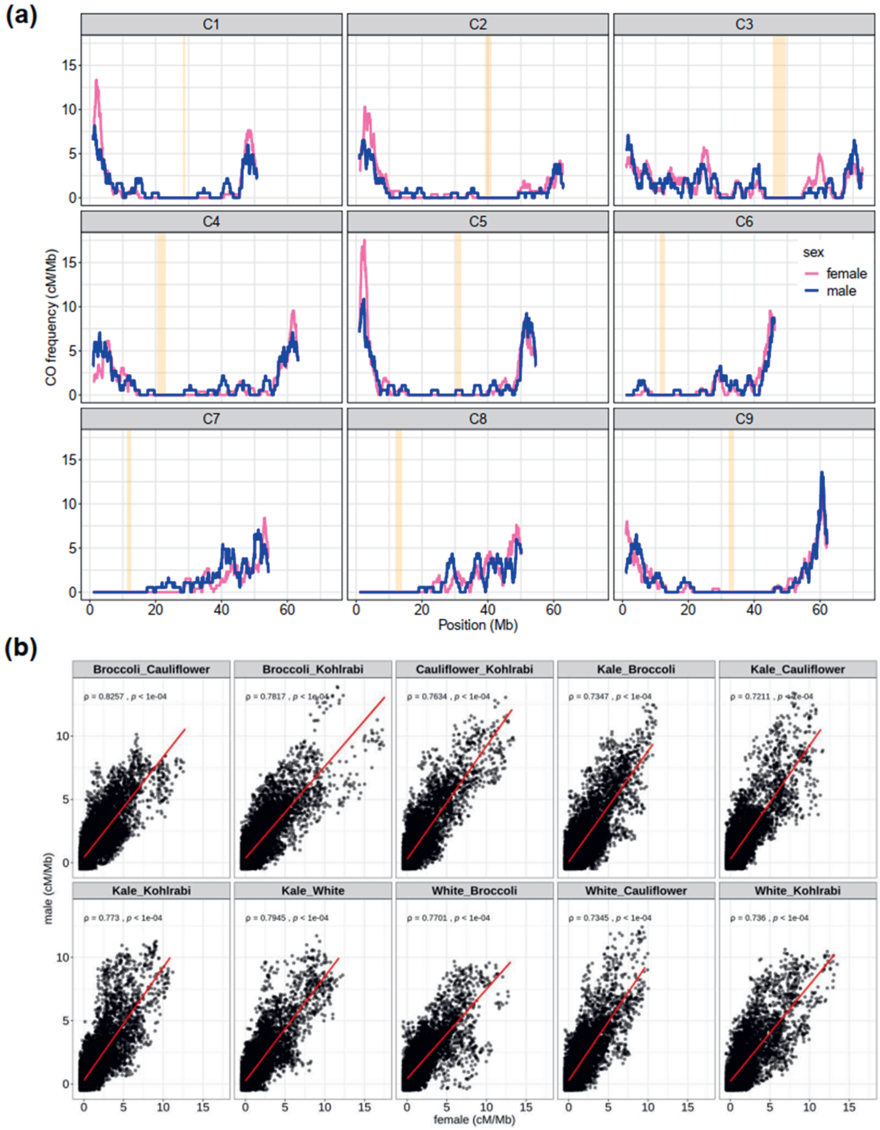


**Fig. 6 Comparison of recombination landscapes between female and male meiosis. (a)** Crossover (CO) distribution (window size 2-Mb, step size 50-Kb) along the nine chromosomes in female and male meiosis of the cross of Broccoli_Kohlrabi. **(b)** Correlation (Spearman's rank correlation) between female and male CO frequency in 2-Mb windows with 50-Kb step size. Red lines indicate best fit and the shading areas represent the 95% C.I.

## Discussion

In the present study, we generated ten female and ten male CO maps in *B. oleracea*. To our knowledge, this is the first time that CO maps reached such a fine resolution in *Brassica* species. Besides maize, tomato and potato, we expand the recombination knowledge to the economically important *B.oleracea* crops which exhibit enormous phenotypic variations. More importantly, we included very diverse genetic backgrounds and reciprocal crosses, enabling the investigation of intraspecific variation, sex of meiosis and their interaction on recombination in *B. oleracea*. To our knowledge, this is the most comprehensive study towards revealing intraspecific variation and sex differences of recombination rates and distributions. As meiotic recombination promotes genetic diversity by shuffling parental chromosomes, the present work provides insights towards improving breeding efficiency in *B. oleracea* via parental selection.

Recombination rate varies remarkably along chromosomes in *B. oleracea,* however, the megabase-scale landscapes are highly conserved between different genetic backgrounds. We observed preferential CO distribution towards distal regions in all *B. oleracea* crosses, consistent to those reported in related species, such as *B. rapa* (Pelé *et al.*, 2017) and *B. napus* (Bayer *et al.*, 2015, Boideau *et al.*, 2022), as well as distant species like potato (Marand *et al.*, 2017), tomato (Demirci *et al.*, 2017, Rommel Fuentes *et al.*, 2020), maize (Kianian *et al.*, 2018) and barley (Dreissig *et al.*, 2020). We demonstrated that the megabase-scale *B. oleracea* U-shaped landscape is highly correlated with gene density. More interestingly, we observed several "CO bumps" that co-localize with elevated gene density. These observations indicate that gene density has a major contribution in shaping the megabase-scale recombination landscape in *B. oleracea*. In *A. thaliana*, Lian *et al.* (Lian *et al.*, 2022b) also reported that gene density together with chromatin accessibility and DNA methylation could explain 85% of the megabase-scale recombination landscape using a machine-learning algorithm. This is in agreement with the strong positive correlations between CO distribution and gene density found in our study. Many epigenetic factors are reported to also affect the recombination frequency and distribution, with CO occurrence correlating with low levels of DNA methylation, low nucleosome density and enrichment in specific histone marks. In this study, these hallmarks of open chromatin were not investigated. Generally, euchromatin is present in distal gene-rich regions and heterochromatin exists in large pericentromeric regions of the chromosome which usually show a high level of DNA methylation and K3K9me2 histone marks (Choulet *et al.*, 2014, Swagatika and Tomar, 2016, Li *et al.*, 2019b, Boideau *et al.*, 2022). The U-shaped CO distribution we found in *B. oleracea* fits with the preferred occurrence of these epigenetic features. Nevertheless, the local influence of epigenetic factors on recombination in *B. oleracea* needs to be further analysed to deepen our knowledge, as realised by Boideau *et al.* (Boideau *et al.*, 2022), who investigated effects of methylation and SVs on the absence of recombination in *B. napus*. Both model-independent CO interference and Kullback-Leibler (KL) divergence analysis validate strong CO interference in *B. oleracea*. However, the similar interference strengths indicated by KL divergence suggest weak influence of genetic background on CO interference (Table S2).

Small-scale sequence divergence can be both positively and negatively associated with recombination rate. We analysed the level of SNPs from 2-by-2 comparisons between all parental genomes used in this study (Table S3). Like in *Arabidopsis* (Blackwell *et al.*, 2020), we also found parabolic relationships between SNP density and CO rate in all *B. oleracea* crosses. This is to some extent in agreement with the "juxtaposition effect" in *Arabidopsis* (Ziolkowski *et al.*, 2015). Meiotic recombination is mutagenic, and this may increase polymorphism levels in regions with high recombination rate. Aside from this, genetic hitchhiking and background selection tend to reduce genetic diversity in low recombination regions by increasing the frequency of beneficial mutations and eliminating deleterious mutations (Ziolkowski *et al.*, 2015, Lian *et al.*, 2022b). These are likely two reasons explaining the positive correlation between small-scale sequence divergence and recombination rate, given the hypothesis that polymorphisms are not causal for the shape of the megabase-scale recombination landscape but rather the consequence (Lian *et al.*, 2022b). The negative correlation is also expected because high levels of polymorphisms function like large-scale genomic rearrangements, which do have local inhibitory effects on COs, as reported in *Arabidopsis* (Rowan *et al.*, 2019, Lian *et al.*, 2022b), tomato (Fuentes *et al.*, 2022), *B. napus* (Boideau *et al.*, 2022) and this study. We systematically investigated the impact of SV size and type on local recombination. This leads to our first conclusion that all types of large-scale SVs locally suppress recombination. We cannot totally exclude that a counter-selection of gametes exhibiting such events occurs as these would alter plants' viability (Rowan *et al.*, 2019). A second conclusion is that SV size does not affect flanking CO rates and distance to nearest COs. Rowan *et al.* (Rowan *et al.*, 2019) came up with several possible mechanisms explaining the underlying suppressive effects of SVs on COs, including the prevention of synaptonemal complex (SC) establishment, or reduction of DSB formation, or reduction of repairment from DSBs into COs. It is worth noting that we observed the enrichment of COs in short deletions (< 500 bp) in our study, similar to what was observed by Fuentes *et al.* (Fuentes *et al.*, 2022). This may be due to recombination resulting in short deletions instead of short deletions inducing more

COs. Alternatively, this may suggest that small deletions do not suppress recombination. We found that large deletions (>=500 bp) overlapped significantly less with COs than expected by chance, confirming local suppression of large SVs on recombination (Fig. S14-S17). For insertions, we would have expected similar patterns as for deletions, since analysing insertions or deletions is just a matter of setting a reference. However, we can only find significantly more insertions overlapping with COs regardless of insertion size. As the sequence of an insertion is absent in the reference genome, we feel the overlap between CO interval and insertion loci is not accurate/realistic, and is also likely influenced by the CO resolution.

Among the ten cross combinations, we observed significant differences in CO number. Our data do not support that CO number and relatedness of parents are correlated. Indeed, we would expect higher numbers of COs when the two parents show higher levels of relatedness as a substrate with high homology may facilitate recombinational repair. However, this is not a general trend. For example, we observed a similar (not significantly different) average CO number in Kale_Broccoli (low level of relatedness) and Broccoli_Cauliflower (high level of relatedness). Interestingly, we found that the presence of broccoli in the parental combinations always results in higher number of COs, even in crosses with cauliflower and white cabbage that were generally associated with lower number of COs (Fig. 5a). Differences in CO rate have repeatedly been found according to the genetic background in plants. Recently, the study of loci affecting CO frequency has been made possible with the development of high-throughput technologies for measuring CO frequency from seeds in *A. thaliana* (Melamed-Bessudo *et al.*, 2005, Ziolkowski *et al.*, 2015). To date, three causal genes were identified: *HEI10*, *TAF4b* and *SNI1* (Ziolkowski *et al.*, 2017, Lawrence *et al.*, 2019, Zhu *et al.*, 2021). Accordingly, our data support a genetic control for CO rate *via* allelic variants segregating between the *B. oleracea* parental lines used in this study, with possible dominant allele(s) increasing CO rates in broccoli. While CO numbers varied, we did not identify major differences between the megabase-scale CO landscapes among our ten cross combinations. CO variations were essentially located on chromosome extremities as observed between different populations of maize (Bauer *et al.*, 2013). Accordingly, polymorphism in *HEI10*, *TAF4b* and *SNI1* essentially resulted in CO variation at chromosome extremities (Ziolkowski *et al.*, 2017, Lawrence *et al.*, 2019, Zhu *et al.*, 2021). To date, the only natural factor associated with major changes in the shape of recombination landscapes, corresponds to variation in ploidy level. The most striking example arises in *Brassica* AAC allotriploids resulting from the cross between *B. napus* and its *B. rapa* progenitor. In these plants, an unprecedented boost of CO number was observed between A genomes and associated with formation of COs in pericentromeric regions that are totally deprived of any recombination event in *B. rapa* and *B. napus* (Pelé *et al.*, 2017, Boideau *et al.*, 2021).

Comparison of female *versus* male meiosis for each of our ten cross combinations revealed interdependency between heterochiasmy and genetic background. Indeed, half of the crosses showed significant variations, with higher recombination rates in male meiocytes. Our results contrast with a previous study conducted in *B. oleracea*, showing much more COs in female than male meiocytes (Kearsey *et al.*, 1996). However, this study was based on 75 molecular markers, affecting the reliability. Importantly, our data support observations made in maize for which B73×Mo17 and Zheng58 × SK backgrounds result in absence and presence of heterochiasmy, respectively (Kianian *et al.*, 2018, Luo *et al.*, 2019). In their study, Luo *et al.* (Luo *et al.*, 2019) suggested that the occurrence of CO maturation inefficiency (CMI), which will block some designated COs developing into actual COs, differs between genetic backgrounds and between sexes in maize. CMI was indeed detected in maize in both male and female meiosis within the B73×Mo17 background and in male meiosis of the inbred line KYS, however not in the Zheng58×SK background (Luo *et al.*, 2019). This corroborates with our observations as higher CO rates were always observed in male *versus* female meiosis when cauliflower was involved in the cross combination. One possible explanation for not detecting heterochiasmy in other combinations tested is that CMI is a recessive trait that does not exist in cauliflower. Alternative explanations for observed heterochiasmy in all cross combinations with cauliflower is a dominant locus in cauliflower promoting increased CO rates exclusively during pollen grain formation but not in female meiosis. This locus may influence synapsis progression during male meiosis, the length of which is positively related to CO number in both sexes as observed in maize (Luo *et al.*, 2019). Despite the heterochiasmy, we observed fairly similar CO distribution between male and female meiocytes. This observation in *B. oleracea* is consistent with that in maize (Kianian *et al.*, 2018), however is remarkably different with that in *A. thaliana* (Lian *et al.*, 2022b).

In conclusion, we generated high resolution recombination landscapes, improving our knowledge of CO formation in *B. oleracea*, and showed remarkable CO variation depending on the direction and combination of the cross, which is highly relevant for breeders.

## Experimental procedures

### Four-way-Cross (FwC) populations construction, DNA isolation and sequencing

We previously *de novo* assembled genome sequences for five *B. oleracea* accessions, representing five diverse morphotypes, by integrating Nanopore long reads, optical mapping molecules (BioNano Genomics DLS technology) and Illumina short reads (Cai *et al.*, 2022b). We generated chromosome-scale genome assemblies, with contig N50's ranging from 11.4 Mb to 16.3 Mb and scaffold N50's ranging from 30.5 Mb to 34.1 Mb. The complete BUSCO values were greater than 97% for all the five assemblies using BUSCO (embryophyta_odb9 dataset, n=1,440) assessment. In this study, we used these five DH lines (broccoli, cauliflower, kale, kohlrabi and white cabbage) as founders to construct FwC populations to study the inter-morphotype recombination landscape (Fig. S1). These five founders were pairwise crossed to generate ten F1s, after which the ten F1s were inter-crossed to generate large FwC populations. To observe recombination between each combination of the five morphotypes, we constructed five FwC populations with each containing four different parents (Fig. 1). We reciprocally crossed the two F1 plants for each FwC population, resulting in ten populations that allow us to study female and male COs for each of the ten crosses.

Genomic DNA was isolated from young leaves of the five FwC population plants using a cetyltrimethylammonium bromide (CTAB) method (Allen *et al.*, 2006). Libraries were constructed with the RipTide DNA library prep kit (iGenomX, Carlsbad, CA), which is designed for the preparation of 96 next-generation sequencing DNA libraries at a time. We collected a total of 1,248 plants from the five FwC populations (Fig. 1). Individual samples were labelled in thirteen 96-well plates, after which samples per plate were pooled together and converted into a NGS library in one single tube. The 13 pooled libraries were respectively sequenced as 150bp paired-end reads using Illumina NovaSeq 6000.

### Read processing, SNP calling and filtering

Fgbio DemuxFastqs (v1.1.0) was used to demultiplex samples per plate according to sample barcodes provided by iGenomX. A total of 47 samples (Table S1) each with less than 100 Mb (~0.18X) sequencing data were excluded from downstream analyses. All reads from each sample were aligned to the broccoli reference genome (Cai *et al.*, 2022b) using BWA-MEM (v0.7.15) (Li and Durbin, 2009) with default parameters. SAMtools (v1.3.1) (Li *et al.*, 2009) was used to perform sorting of the alignments. The function of HaplotypeCaller in Genome Analysis Toolkit (GATK, v4.1.7.0) (McKenna *et al.*, 2010) was used to produce GVCF files on a per-sample basis, before which duplicated, secondary alignment reads and reads with low mapping quality were filtered out using default settings. We used CombineGVCFs function in GATK to combine per-sample GVCF files into a single GVCF file for each of the ten reciprocal populations, following which GenotypeGVCFs function was used to perform the joint genotyping. SelectVariants function was then used to select biallelic SNPs and further SNP filtering was performed using VariantFiltration function with parameters "--filter-expression 'QD < 2.0 || FS > 60.0 || MQ < 40.0' --cluster-window-size 5 --cluster-size 2 --filter-name LowQual". SNPs among the five parental lines were called using the high-depth Illumina sequencing data generated by (Cai *et al.*, 2022b) with similar strategies as described above for the populations. Only homozygous parental SNPs were retained.

**Detection of CO**

Given the FwC strategy in this study, we selected parental SNPs which allow us to identify COs that occurred for each combination of the five parents. To identify COs between P1 and P2, we selected parental SNPs in which we only allow genotype variations between these two parents while not between the other two parents (P3 and P4) (Fig. S36). Similarly, to detect COs between P3 and P4, we selected another group of SNPs in which we allow genotype variations between P3 and P4 while not between P1 and P2. The selected parental SNPs were intersected with SNP matrix of the corresponding population. SNP sites with more than 40% missing genotype calls were discarded. The parental origin for the allele at each SNP site was inferred based on the observed genotypes for the four parents and the FwC progeny. Chi-square test for goodness of fit was used to analyse segregation distortions. The expected segregation ratio is 1:1. SNP sites with significant segregation distortion at a significance level of $P$ = 0.001 were removed from further analysis. Since SNPs in close proximity in physical maps are supposed to be highly linked in biparental populations (Marand *et al.*, 2017, Marand *et al.*, 2019), SNPs demonstrating low levels of linkage disequilibrium (LD) with neighbouring markers are likely false-positive variants. We thus estimated local $r^2$ values for each SNP using the nearest 100 SNPs to determine associated alleles and removed SNPs with local $r^2$ values < 0.3 (mean $r^2$ values across the 100 comparisons).

PhaseLD (Marand *et al.*, 2017) (https://github.com/plantformatics/phaseLD) was then used to reconstruct haplotype phase with parameters "--quick_mode --win 100 --bwin 200 --bstep 5 --rpen 0.3". This pipeline implemented a sliding window approach to overcome sequencing and genotyping errors that could arise from assembly errors or structural variations. We applied 200-SNP sliding windows with 5-SNP steps to estimate the posterior probability of both haplotypes using Bayes theorem for each individual in a given window. The haplotype with the highest probability was called for the given window. Putative COs were determined from these overlapping adjacent haplotype bins. Precise CO breakpoints were then identified using logistic regression approach as implemented in "extract_crossovers.pl" (https://github.com/plantformatics/phaseLD/tree/master/bin), which assigns CO probabilities to each SNP. To reduce false positive CO counts, only the pair of SNPs with a CO probability greater than 0.9 were kept as the identified CO intervals. To further remove likely false positive COs, CO positions that appeared to be double COs < 2Mb apart were removed. We employed all these steps to minimize the risk of false CO detection, keeping in mind that true CO numbers should be similar to those reported in other Brassica populations (0.70-0.92 CO occurrences on average per chromatid) (Pelé *et al.*, 2017, Boideau *et al.*, 2022).

## CO number and landscape analyses

We calculated three-fold interquartile ranges for the ten reciprocal crosses using total CO number (TCN) of each gamete. Gametes with TCN outside the three-fold interquartile range (0.5-12.5) were removed, like done in other studies (Dreissig *et al.*, 2020). Compared to the population mean of 6.57 and outlier-pruned mean of 6.49, these outliers (22 out of 2,399 gametes) showed a mean TCN of 15.09 (Table S5). We calculated Poisson distributions of CO number per chromatid per gamete using the following formula: $S(k) = N\frac{e^{-m}m^{k}}{k!}$ where $S(k)$ is the number of gametes harbouring exactly $k$ CO, $N$ is the total number of gametes, $m$ is the observed mean number of CO per gamete, and $e$ is the Natural logarithm base (Drouaud *et al.*, 2007, Giraut *et al.*, 2011). Recombination landscapes of each chromosome of each cross were visualized using 2Mb sliding windows with 50Kb steps. We summarized CO frequency (cM/Mb) as C/n/(w/10^6)*100, where w is the window size, C is the number of recombinant gamete in the given window, and n is the total number of gametes for the population (Campoy *et al.*, 2020, Dreissig *et al.*, 2020).

## Inference of putative centromeric and pericentromeric regions

Centromeric regions were determined using the approach described by (Cheng *et al.*, 2013) and (Cai *et al.*, 2020). Briefly, centromere-specific repeat sequences, such as CentBr, CRB and TR238 (Koo *et al.*, 2004, Lim *et al.*, 2005, Lim *et al.*, 2007, Koo *et al.*, 2011), were aligned to the broccoli reference genome using nucmer with parameters "--maxmatch -g 500 -c 16 -l 16" (Marçais *et al.*, 2018). Centromeric regions were located based on the distribution of these elements in the reference genome (Table S6). Thereafter, the broccoli genome and *B. napus* cv. Darmor-bzh v10 C-genome (Rousseau-Gueutin *et al.*, 2020) were aligned to identify syntenic regions using SyRI (Goel *et al.*, 2019). The closest syntenic regions to each border of *B. napus* cv. Darmor-bzh v10 C-genome pericentromeres, which were defined by (Boideau *et al.*, 2022), were extracted from *B. napus* genome. The corresponding regions in broccoli genome were then defined as the borders for the putative pericentromeric regions (Table S6).

**Association analyses between genomic features and COs**

To investigate CO interference, we calculated distances between adjacent COs for each gamete per chromatid having at least two COs, using the mid-value of the positions of the two CO flanking markers. The observed interference distributions were compared to no-interference distributions, obtained using a randomly shuffling approach proposed by  (Pelé *et al.*, 2017). COs at fine-resolution (less than 10Kb) were selected to analyse the overlap (minimum 1-bp) with various genomic features (exons, introns, 1Kb upstream and -downstream gene regions, TEs and intergenic regions) (Marand *et al.*, 2017). Random genomic regions were permuted 10,000 times using BEDtools shuffle (v2.27.1) (Quinlan and Hall, 2010), and were then assessed for overlap with each genomic feature. SNP densities between each combination of the parents were calculated using 2Mb sliding windows with 50Kb steps. The positions of SVs between parental genomes were obtained from Cai *et al.* (Cai *et al.*, 2022b). Overlaps between fine-resolution CO intervals and the SV regions were studied using regioneR (Gel *et al.*, 2016).  We performed another 10,000 times of Monte Carlo simulation using BEDtools shuffle (v2.27.1) to generate random genomic sequences matched by number and length to the CO dataset from the broccoli reference genome. Thereafter, we searched for the nearest simulated genomic sequence from each simulation in the flanking regions of each SV, and calculated the distance between each of the two SV borders and the nearest simulated region. We took the smaller value as the distance to the nearest simulated region for each SV. We then compared this expected distance distribution with the observed distribution that was obtained using the real CO dataset. CO rates in 1-, 2- and 3-Mb upstream and downstream of the borders for SVs were calculated and were compared to the genome-wide level CO rates.

**Acknowledgements**

We would like to thank Theo Borm, Hans de Jong, Korbinian Schneeberger and Roven Rommel Fuentes for helpful discussions. We thank Fernanda de Alves Freitas Guesdes for the try-out data analysis. This research was funded by the TKI project

KV 1605-004 “A *de novo* sequencing catalogue *B. oleracea*” (https://topsectortu.nl/nl/de-novo-sequencing-catalogue-b-oleracea), and was co-supported by two breeding companies (Bejo and ENZA). CC is supported by China Scholarship Council (No. 201809110159).

## Author contributions

GB and CC designed the research. JB performed the wet lab experiments. CC, AP and RF analysed data. CC drafted the manuscript. CC, AP and GB revised the manuscript. All authors read and approved the final manuscript.

## Conflict of interests

The authors declare that they have no conflicts of interest.

## Data availability statement

The raw sequencing reads for the five Four-way-Cross (FwC) populations in this study have been deposited in NCBI under the accession number PRJNA847181. The list of CO positions identified for each reciprocal cross can be found in Table S7.

## Supporting information

Supplementary files are available at https://onlinelibrary.wiley.com/doi/10.1111/tpj.16104

**Fig. S1** Illustration of the five parental *Brassica oleracea* genotypes used for Four-way-Cross (FwC) population construction and the ten F1s generated by pairwise crosses.

**Fig. S2** Distribution of segregating SNPs, which were used for crossover (CO) detection, along the nine chromosomes of *Brassica oleracea* for each of the ten reciprocal crosses.

**Fig. S3** Haplotype map of the ten reciprocal crosses. Orange and blue segments reflect the two parental alleles segregating in the corresponding population.

**Fig. S4** Distribution of total crossover (CO) number per gamete across the 2,377 gametes in the ten reciprocal crosses.

**Fig. S5** Distribution of crossover (CO) numbers per chromatid in the ten sex-averaged crosses. **(a)** Observed distributions. **(b)** Expected Poisson distributions (see Methods).

**Fig. S6** Distribution of inter-crossover (CO) distance for chromatids having at least two COs. **(a)** Comparison of inter-CO distance distribution among the ten sex-averaged crosses. Solid lines indicate the observed data. Dashed lines correspond to the corresponding distributions in the shuffled data (“non-interference” situation), as was described in Methods. **(b)** Comparison of inter-CO distance distribution between female and male meiosis in each cross.

**Fig. S7** Sliding window based recombination landscapes (window size 2-Mb, step size 50-Kb) for the ten sex-averaged crosses. The centromere regions are indicated by orange shadings.

**Fig. S8** Permutation tests for evaluating overlaps between genes and crossover (CO) intervals for all the ten reciprocal crosses. Female and male meioses are indicated in the left and right column, respectively. On x-axis, the values are the total number of overlaps. On y-axis, the values are the frequency. The vertical red lines indicate the number of overlaps where $P = 0.05$. The vertical green lines indicate the observed number of overlaps. The vertical black lines indicate the mean of 5,000 permutations. The double arrow highlights the difference between the mean and the observed values.

**Fig. S9** The normalized distribution of crossovers (COs) in the ten sex-averaged crosses and of SNP density between each pair of parents along the nine chromosomes of *Brassica oleracea*. Analysis is done with 2-Mb sliding windows and 50-Kb step sizes. CO frequency was normalized to range from 0 (min) to 1 (max) and SNP density was normalized to range from  -1 (min) to 0 (max). The centromere regions are indicated by orange shadings.

**Fig. S10** Correlation (Spearman's rank correlation) between crossover (CO) frequency and SNP density in 2-Mb sliding windows with 50-Kb step sizes for each sex-averaged cross. (**a**) Correlation analysis with pericentromeric regions being included. (**b**) Correlation analysis with pericentromeric regions being excluded. Trend lines were generated using a generalized additive model (GAM) with the formula y ~ poly(x,2). Verticle blue lines represent mean SNP density (the number of SNPs per 2-Mb window).

**Fig. S11** Crossover (CO) distribution comparison between the cross of Broccoli_Cauliflower and White_Cauliflower, and the corresponding parental SNP density distribution comparison. Analysis is done with 2-Mb sliding windows and 50-Kb step sizes. The centromeric and pericentromeric regions are indicated by dark and light orange shadings, respectively. Asterisks indicate intervals (non-overlapping 2-Mb windows) with significant CO frequency difference between the two crosses ($P < 0.05$, chi square test).

**Fig. S12** Crossover (CO) distribution comparison between the cross of White_Cauliflower and White_Kohlrabi, and the corresponding parental SNP density distribution comparison. Analysis is done with 2-Mb sliding windows and 50-Kb step sizes. The centromeric and pericentromeric regions are indicated by dark and light orange shadings, respectively. Asterisks indicate intervals (non-overlapping 2-Mb windows) with significant CO frequency difference between the two crosses ($P < 0.05$, chi square test).

**Fig. S13** Bionano evidence for two large inversions. **(a)** A 4.88-Mb kale-specific inversion on chromosome C3. **(b)** A 1.42-Mb cauliflower-specific inversion on chromosome C7.

**Fig. S14** Permutation tests for evaluating overlaps between different types of structural variations (indicated in the top right corner of each figure) and crossover (CO) intervals for the cross of Broccoli_Cauliflower. On x-axis, the values are the total number of overlaps. On y-axis, the values are the frequency. The vertical red lines indicate the number of overlaps where $P = 0.05$. The vertical green lines indicate the observed number of overlaps. The vertical black lines indicate the mean of 5,000 permutations. The double arrow highlights the difference between the mean and the observed values.

**Fig. S15** Permutation tests for evaluating overlaps between different types of structural variations (indicated in the top right corner of each figure) and crossover (CO) intervals for the cross of Broccoli_Kohlrabi. On x-axis, the values are the total number of overlaps. On y-axis, the values are the frequency. The vertical red lines indicate the number of overlaps where *P* = 0.05. The vertical green lines indicate the observed number of overlaps. The vertical black lines indicate the mean of 5,000 permutations. The double arrow highlights the difference between the mean and the observed values.

**Fig. S16** Permutation tests for evaluating overlaps between different types of structural variations (indicated in the top right corner of each figure) and crossover (CO) intervals for the cross of Kale_Broccoli. On x-axis, the values are the total number of overlaps. On y-axis, the values are the frequency. The vertical red lines indicate the number of overlaps where *P* = 0.05. The vertical green lines indicate the observed number of overlaps. The vertical black lines indicate the mean of 5,000 permutations. The double arrow highlights the difference between the mean and the observed values.

**Fig. S17** Permutation tests for evaluating overlaps between different types of structural variations (indicated in the top right corner of each figure) and crossover (CO) intervals for the cross of White_Broccoli. On x-axis, the values are the total number of overlaps. On y-axis, the values are the frequency. The vertical red lines indicate the number of overlaps where *P* = 0.05. The vertical green lines indicate the observed number of overlaps. The vertical black lines indicate the mean of 5,000 permutations. The double arrow highlights the difference between the mean and the observed values.

**Fig. S18** Distribution of structural variations (SVs) for four morphotypes relative to the broccoli reference genome. **(a-d)** The number of SVs in 2-Mb windows with 50-kb steps for deletions, insertions, translocations and transpositions, respectively. **(e)** The distribution of inversions along the chromosomes. For large inversions (≥10,000bp), the size of each segment corresponds to the size of the inversion. However, to show small inversions (<10,000bp) in the figure, we reset their size as 10,000bp. The centromere regions are indicated by orange shadings.

**Fig. S19** Violin plot of distance to the nearest crossover (CO) for different types of structural variations (SVs). The white dots indicate average values. The vertical rectangles indicate the interquartile ranges.

**Fig. S20** Correlation (Spearman's rank correlation) between the size of structural variations (SVs) (Broccoli *vs* Cauliflower genome) and the distance to their nearest crossover (CO) (Broccoli_Cauliflower cross). The type of SVs is indicated in the top right corner of each figure.

**Fig. S21** Correlation (Spearman's rank correlation) between the size of structural variations (SVs) (Broccoli *vs* Kohlrabi genome) and the distance to their nearest crossover (CO) (Broccoli_Kohlrabi cross). The type of SVs is indicated in the top right corner of each figure.

**Fig. S22** Correlation (Spearman's rank correlation) between the size of structural variations (SVs) (Broccoli *vs* Kale genome) and the distance to their nearest crossover (CO) (Kale_Broccoli cross). The type of SVs is indicated in the top right corner of each figure.

**Fig. S23** Correlation (Spearman's rank correlation) between the size of structural variations (SVs) (Broccoli *vs* White Cabbage genome) and the distance to their nearest crossover (CO) (White_Broccoli cross). The type of SVs is indicated in the top right corner of each figure.

**Fig. S24** The distribution of crossover (CO) rates in windows of the indicated sizes in the upstream **(a)** and downstream **(b)** regions of different types of structural variations (SVs). 'All' represents all windows genome-wide. The white dots indicate the average values.

**Fig. S25** Correlation (Spearman's rank correlation) between the size of structural variations (SVs) (Broccoli *vs* Cauliflower genome) and the crossover (CO) rates (Broccoli_Cauliflower cross) in their flanking 1-Mb regions.

**Fig. S26** Correlation (Spearman's rank correlation) between the size of structural variations (SVs) (Broccoli *vs* Kohlrabi genome) and the crossover (CO) rates (Broccoli_Kohlrabi cross) in their flanking 1-Mb regions.

**Fig. S27** Correlation (Spearman's rank correlation) between the size of structural variations (SVs) (Broccoli *vs* Kale genome) and the crossover (CO) rates (Kale_Broccoli cross) in their flanking 1-Mb regions.

**Fig. S28** Correlation (Spearman's rank correlation) between the size of structural variations (SVs) (Broccoli *vs* White Cabbage genome) and the crossover (CO) rates (White_Broccoli cross) in their flanking 1-Mb regions.

**Fig. S29** The mean number of crossovers (COs) per chromatid along the nine chromosomes of *Brassica oleracea* for the sex-averaged, female and male crosses. On x-axis, chromosomes were ordered according to their lengths.

**Fig. S30** Correlation (Spearman's rank correlation) between chromosome length and mean crossover (CO) number per chromatid. **(a)** Correlation analysis for each of the ten sex-averaged, female or male crosses. **(b)** Correlation analysis based on the pool of ten sex-averaged, female or male crosses.

**Fig. S31** Recombination landscapes for the ten female crosses. **(a)** Crossover (CO) rate distributions along the nine chromosomes of *Brassica oleracea*. Analysis is done with 2-Mb sliding windows and 50-Kb step sizes. The centromere regions are indicated by orange shadings. **(b)** Genome-wide correlation coefficient (Spearman's rank correlation) matrices among the ten female CO distributions.

**Fig. S32** Recombination landscapes for the ten male crosses. **(a)** Crossover (CO) rate distributions along the nine chromosomes of *Brassica oleracea*. Analysis is done with 2-Mb sliding windows and 50-Kb step sizes. The centromere regions are indicated by orange shadings. **(b)** Genome-wide correlation coefficient (Spearman's rank correlation) matrices among the ten male CO distributions.

**Fig. S33** Crossover (CO) distribution (window size 2-Mb, step size 50-Kb) along the nine chromosomes in female and male meiosis of nine crosses. For the cross of Broccoli_Kohlrabi, see Fig. 6. The centromere regions are indicated by orange shadings.

**Fig. S34** Genomic features associated with female and male meiotic crossovers (COs). **(a)** Overlap analysis of female and male COs with different genomic features. **(b)** Distribution of distance from each CO to the nearest gene. Note: in **(a)**, the CO interval was used for the overlap analysis. If an interval overlapped with multiple genomic features, the interval was counted towards each genomic feature. In **(b)**, the middle position of both CO interval and gene was used for calculating the distance.

**Fig. S35** The comparison of crossover (CO) number distributions per chromatid between female and male meiosis for all ten cross combinations. The obseved and expected CO number distributions were also compared. ‘Expected distribution’ denotes Poisson distribution (see Methods).

**Fig. S36** Parental SNPs selection for crossover (CO) identification (see Methods). This example shows how parental SNPs were selected to identify COs that occur between P1 and P2. Only genotype variations between P1 and P2 are allowed. Parental origin can be inferred based on the genotype of parents and FwC progenies.

**Table S1** Illumina sequencing data for each progeny of the five Four-way-Cross (FwC) populations.

**Table S2** Quantitative measurement of interference strength based on the the Kullback-Leibler (KL) divergence from the observed to the "no-interference" distribution (see Methods).

**Table S3** Number of SNPs between each pair of the five parental genomes.

**Table S4** Statistics of structural variations (SVs) and the distance to their nearest crossover (CO).

**Table S5** Three-fold interquartile range of total crossover (CO) number per gamete.

**Table S6** Inferred positions of the centromeric and pericentromeric regions for the nine broccoli chromosomes.

**Table S7** The list of crossover (CO) positions identified for each of the ten reciprocal crosses.

# Chapter 5

## Metabolomic and transcriptomic profiles in diverse *Brassica oleracea* crops provide insights into genetic regulation of glucosinolate variation

**Chengcheng Cai[1,2], Ric C.H. de Vos[3], Hao Qian[1], Johan Bucher[1] and Guusje Bonnema[1,2,*]**

[1] Plant Breeding, Wageningen University and Research, 6708 PB, Wageningen, The Netherlands

[2] Graduate School Experimental Plant Sciences, Wageningen University and Research, 6708 PB, Wageningen, The Netherlands

[3] Bioscience, Wageningen University and Research, 6708 PB, Wageningen, The Netherlands

[*] Corresponding author

## Abstract

Glucosinolates (GSLs) are plant secondary metabolites commonly found in the cruciferous vegetables of the Brassicaceae family. GSLs play important roles in defense against pathogens and pests and several GSLs have health benefits to humans. In this study, the relative abundance levels of 23 GSL compounds were determined in both roots, stems and the edible parts of five different *Brassica oleracea* morphotypes, including broccoli, cauliflower, kale, kohlrabi and white cabbage. A total of 183 GSL related genes were identified in the five corresponding high-quality *B. oleracea* genome assemblies based on the experimentally characterized GSL genes in Arabidopsis. Expression of these GSL related genes were then analyzed using mRNA-Seq data. We revealed strong variation in terms of relative abundance and composition of GSLs among different tissues and different morphotypes. Accordingly, GSL related genes were differently expressed between different tissues and different morphotypes. We found a total of 289 GSL-gene combinations with significant correlation between GSL and gene expression level, which involved all the 23 GSLs and 109 genes. Interestingly, we observed a non-functional *AOP2* in broccoli, which is related to the loss of a conserved 2OG-FeII_Oxy domain, explaining the accumulation of health-promoting 4-methylsulfinylbutyl and 5-methylsulfinylpentyl GSLs in broccoli. Additionally, we found many transposable element (TE) insertions in one paralog of the *MAM3* gene in three genomes via detailed gene structure and sequence analyses, resulting in long and repeat-rich intronic sequences. The present work increases our understanding of GSL variation and its genetic regulation in *B. oleracea* morphotypes and provides insights for breeding with tailored GSL profiles in these crops.

## Introduction

Glucosinolates (GSLs) are a group of secondary plant metabolites almost exclusively found in the order Brassicales (Halkier and Gershenzon, 2006). These GSL compounds, derived from glucose and amino acids, are rich in nitrogen and sulfur and are water-soluble (Petersen *et al.*, 2018, Bonnema *et al.*, 2019). According to their precursor amino acid, GSLs can be classified as aliphatic GSLs (derived from alanine, isoleucine, leucine, methionine and valine), aromatic GSLs (derived from phenylalanine and tyrosine) or indolic GSLs (derived from tryptophan) (Halkier and Gershenzon, 2006, Petersen *et al.*, 2018). Briefly, the biosynthesis of GSLs includes three independent stages: side-chain elongation of the precursor amino acid, formation of the core structure, and side-chain modification (Sønderby *et al.*, 2010). All GSLs share a common core structure, which is linked to an amino acid derived side-chain, with thioglucose and sulphate groups (Fahey *et al.*, 2001, Agerbirk and Olsen, 2012). GSLs are extremely variable due to the differences in side-chains, chain lengths and additional side-chain modifications. To date, more than 130 GSL structures are scientifically documented in *A. thaliana* and other plants (Fahey *et al.*, 2001, Agerbirk and Olsen, 2012, Petersen *et al.*, 2018, Harun *et al.*, 2020). In plants, the hydrolysis products of GSLs play important roles in defense against pathogens and pests (Wittstock *et al.*, 2004, Beekwilder *et al.*, 2008, Van Dam *et al.*, 2009, Bruce, 2014). In vegetables, they provide diverse tastes like bitterness and pungency (Bell *et al.*, 2018). In addition, GSLs have been reported to be implicated in both antinutritional and health-promoting effects (Wattenberg, 1977, Tripathi and Mishra, 2007, Petersen *et al.*, 2018). For example, increasing evidences point to a cancer prevention and anti-inflammatory effect of isothiocyanates (ITCs) that are produced from GSLs upon cell damage (Higdon *et al.*, 2007, Hayes *et al.*, 2008, Verkerk *et al.*, 2009, Wu *et al.*, 2013). However, some GSLs are antinutritional, such as progoitrin which promotes goitre disease (Voorrips *et al.*, 2000).

*Brassica oleracea* is an economically important vegetable and fodder crop species cultivated worldwide. It consists of many morphotypes which exhibit an enormous diversity in their appearance. For example, cabbages (*var. capitata*) form leafy heads, with different varieties differing in leaf colour and texture and/or head shape; broccoli (*var. italica*) and cauliflower (*var. botrytis*) are characterized by their typical curd with large arrested inflorescences; kohlrabi's (*var. gongylodes*) form enlarged tuberous stems; kales (*var. acephala*) are characterized by their variation in leaf shapes, color and structure, including bore and curly kale, and marrow stem kale, etc (Kole and Henry, 2010, Bonnema *et al.*, 2011, Dias, 2012, Cai *et al.*, 2022a). Despite this enormous diversity, *B. oleracea* truly is a single species and morphotypes can be easily interbred. Several studies clearly showed that the genomes of the different morphotypes contain numerous structural variations and that their gene contents can vary extensively (Golicz *et al.*, 2016b, Cai *et al.*, 2022b). Besides diversity in appearance and genome sequence, *B. oleracea* crops also vary remarkably in their GSLs content and composition. For example, Yi and colleagues (Yi *et al.*, 2015) determined the content of 16 different types of GSLs in edible organs of 12 *B. oleracea* genotypes, including four different morphotypes. Hahn and colleagues (Hahn *et al.*, 2016) estimated the content of five GSLs in 25 kale varieties and 11 non-kale *B. oleracea* cultivars. Bhandari and colleagues (Bhandari *et al.*, 2020) assessed the content of 12 types of GSLs in the head of 146 cabbage genotypes. All these studies showed that GSL composition and levels differed markedly among different *B. oleracea* genotypes.

To date, most of the knowledge with regard to biosynthesis, degradation, transport and regulation mechanisms as well as the function of GSLs is based on extensive studies performed in *A. thaliana*, including mutant screens, quantitative trait loci (QTL) mapping and genome-wide association studies (GWAS) (Sønderby *et al.*, 2010, Jensen *et al.*, 2014). Indeed, a total of 113 genes controlling GSL biosynthesis, degradation, transport and storage have been experimentally characterized in this model species, including 85 enzyme-encoding genes, 23 transcription factors and five transporter proteins (Harun *et al.*, 2020). Comparative genomic analyses between Arabidopsis and other *Brassica* crops can provide comprehensive information for the GSL biosynthetic pathway in these non-model crops (Wang *et al.*, 2011a, Yi *et al.*, 2015). Recently, we *de novo* assembled five high-quality chromosome-scale *B. oleracea* genomes (Cai *et al.*, 2022b), which provide valuable genomic resources for identifying GSL related genes in *B. oleracea* and studying their sequence divergence. From the aspect of breeding, an important goal is to generate optimal GSL profiles with regard to health and taste in the edible organs of *B. oleracea*, and retain the plant growth protective effects such as the pest insect damage protection and the inhibition of weed growth in surrounding areas (Macias *et al.*, 2007, Feng *et al.*, 2012). To do so, it is key to better understand the genetic regulation of GSL variation between genotypes and tissues in *B. oleracea*.

Here, we analyzed the relative abundance of 23 different GSL structures in four different plant tissues, including roots, stems and the edible parts of five different *B. oleracea* morphotypes. Based on all the experimentally characterized GSL genes in Arabidopsis, we identified their homologs in the corresponding five high-quality *B.*

*oleracea* genome assemblies. Moreover, mRNA-Seq was generated for the same samples that were used for GSL profiling to study gene expression. We revealed strong variation in terms of composition and relative abundance of GSLs among different tissues and different morphotypes. The identified GSL related genes were differently expressed between different tissues and different morphotypes. We found significant correlations between the abundance of 23 GSLs and expression level of 109 related genes. We also present interesting observations in this study, including a non-functional *AOP2* in broccoli related to the loss of the conserved 2OG-FeII_Oxy domain, explaining the specific accumulation of desired 4-methylsulfinylbutyl and 5-methylsulfinylpentyl GSLs in broccoli, and transposable element (TE) insertion activities in one paralog of the *MAM3* gene in three out of the five genomes causing long and repeat-rich introns, respectively.

## Results

### GSL profile variations among *B. oleracea* morphotypes and tissues

Intact GSLs were analyzed in five *B. oleracea* morphotypes (broccoli, cauliflower, kale, kohlrabi and white cabbage), with each morphotype including four different tissues from plants grown in three biological replicates (Fig. S1). We identified a total of 23 different GSLs using LCMS, consisting of 17 aliphatic, two aromatic and four indolic GSLs. Analytical variation was determined for each GSL by analyzing so-called quality control samples (QCs), which consisted of five independent extractions from a large pooled sample prepared by mixing the same small amount of each biological sample. These QCs were similarly and jointly prepared with the biological samples. One QC was analyzed before and one QC after the entire series, and the remaining three QCs were evenly distributed between the 60 real samples. Based on the chromatographic GSL peak areas obtained with these five QCs (data not shown), the analytical variation of the GSLs in the *B. oleracea* samples was on average 11.2%, ranging from 1.1% for 4-hydroxy-3-indolylmethyl-GSL to 28.9% for 8-methylsulfinyloctyl-GSL.

Principal component analysis (PCA) based on GSL data showed close positions between the  three biological replicates, indicating very low biological variations between the replicates (Fig. S2). We found extensive variation in GSL profiles among different morphotypes as well as between different tissues by comparing the relative abundances of each compound (Fig. 1). Overall, kohlrabi showed relatively low abundance of nearly all detected GSLs, whereas kale and white cabbage exhibited high levels of most GSLs. We detected morphotype signature GSLs of which the abundance in one morphotype was significantly higher than that in any other morphotype (Student-Newman-Keuls test with a=0.05). Accordingly, kale and white cabbage contain nine and eight signature GSLs, respectively (Table S1). In contrast, we did not find any kohlrabi signature GSL among these 23 compounds. Three compounds (1-methoxy-3-indolylmethyl, 8-methylsulfinyloctyl_II and 4-methoxy-3-indolylmethyl) were not signature for any morphotype. Two broccoli signature GSLs (4-methylsulfinylbutyl and 5-methylsulfinylpentyl) showed relatively high levels in all its tissues, while their levels were remarkably lower in any tissue of the other morphotypes. This was also the case for two of the white cabbage signature GSL (3-butenyl and 2-hydroxy-3-butenyl) (Fig. 1 and Fig. S3). In the profiled 23 GSLs, we observed that broccoli lacked C3 aliphatic GSLs (Fig. S3a). Similarly, we found either remarkably low abundance or no accumulation of C3 aliphatic GSLs in kohlrabi tissues (Fig. S3a). Cauliflower did not accumulate C4 and C5 aliphatic GSLs in any tissue (Fig. S3b and S3c). Generally, most detected GSLs tend to be accumulated at a higher level in roots than in other tissues (Fig. 1 and Fig. S3). For example, the aromatic 2-phenylethyl GSL was detected at relatively high levels only in roots of the three morphotypes broccoli, cauliflower and kale.

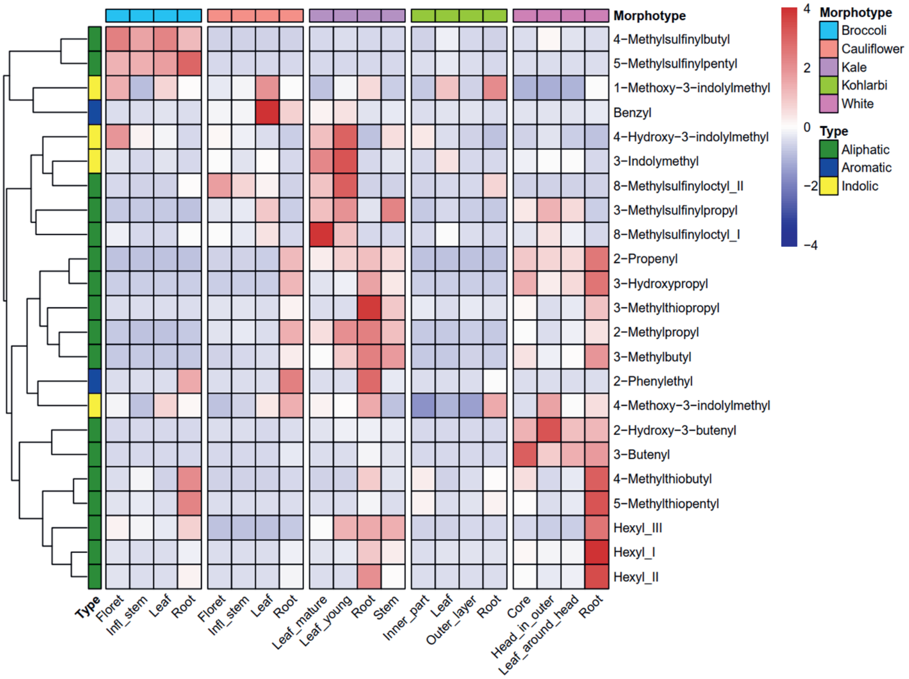


**Fig. 1 Variations in glucosinolate (GSL) levels**. For each tissue, the average relative intensity value of the three biological replicate samples was considered as the GSL concentration (note: each sample

consists of 200 mg FW extracted with 600 µl solvent). Per GSL the relative intensity values were normalized across samples using Z-score standardization.

## GSL related gene identification in *B. oleracea* genomes and their overall expression

In Arabidopsis, a total of 113 GSL related genes have been identified, including 85 GSL biosynthesis genes, 23 transcriptional components and five transporters (Harun *et al.*, 2020). Using these genes as queries, we identified a total of 183 non-redundant orthologous genes in the five high-quality *B. oleracea* genome assemblies (Cai *et al.*, 2022b), including 167, 165, 170, 166 and 170 genes in broccoli, cauliflower, kale, kohlrabi and white cabbage, respectively, which are distributed all along the nine *B. oleracea* chromosomes (Table S2, Table S3 and Fig. S4). The vast majority (153 genes) of these genes showed one-to-one syntenic relationship among the five genomes (Table S2). However, we also found a total of 30 presence and absence genes among the five genomes. Across the 113 GSL related genes in Arabidopsis, we observed that 50 of them had multiple orthologous gene copies (⩾2) in at least one of the five *B. oleracea* genomes, with the copy number differing among the five morphotypes for several genes. As an example, one and four homologous copies of *MYB51* were found in broccoli and kale, respectively. Also, 42 out of the 113 Arabidopsis GSL genes had maximum one copy in each of the five *B. oleracea* genomes. However, for some genes, such as *NSPs*, *FMOGS-OXs*, and *NITs*, less paralogous copy numbers were found in *B. oleracea* than in Arabidopsis. Additionally, two Arabidopsis genes (*CCA1* and *MYB115*) had no orthologues in any of the five *B. oleracea* genomes. The expanded number of GSL related genes in *B. oleracea* is partly attributed to the whole genome triplication (WGT) event (Liu *et al.*, 2014a, Parkin *et al.*, 2014a). However, due to the extensive gene fractionation that occurred following the WGT (Liu *et al.*, 2014a, Parkin *et al.*, 2014a), less than three copies and even no orthologues in *B. oleracea* were also found for some GSL homologs (Table S2).

Paired-end mRNA-seq was performed for the same samples that were used for GSL profiling. We generated a total of $3.08\times10^9$ clean reads (~460.53 Gb) from the 60 samples (20 tissues × 3 biological repeats), averaging $5.13\times10^7$ clean reads (7.68 Gb) per sample (Table S4). On average, 91.33% of these clean reads were concordantly and uniquely mapped to the five corresponding reference genomes. Quality control by PCA and hierarchical clustering analysis showed that the three biological replicates closely clustered together, indicating good repeatability of gene expression data between biological replicates (Fig. S5). Different tissues of each morphotype were clearly separated by PC1, PC2 or PC3. Accordingly, the hierarchical clustering analysis divided the samples of each morphotype into four groups, representing four different tissues (Fig. S5). Generally, the roots are the most derived tissues that are separated from the remaining tissues. We estimated gene expression levels by transcripts per million (TPM) based on the alignments of mRNA-Seq reads.

We then investigated the overall expression pattern of the 153 one-to-one syntenic GSL genes among different *B. oleracea* morphotypes and tissues by constructing heatmaps combined with a hierarchical clustering analysis based on gene expression profiles (log2 transformed and z-scored TPM values separately). Three clusters were revealed based on the log2 transformed TPM values, with genes in cluster I displaying the highest, while in cluster III the lowest expression level across the 20 tissues (Fig. 2a and Table S8). We did not observe a consistent clustering of genes involved in the same process. Cluster I consists of seven genes (*BoCYP83A1*, *BoESP.2*, *BoGSTF9.1*, *BoGSTF9.2*, *BoGSTU20*, *BoGSTU13.1* and *BoGSTU13.2*) involved in core structure synthesis, three genes (*BoGSH1.1*, *BoASA1.1* and *BoASA1.2*) involved in co-substrate pathways, two genes (*BoIIL1.1* and *BoIIL1.3*) involved in side-chain elongation and one gene (*BoESP.2*) involved in GSL degradation. Both cluster II and cluster III include a large number of genes involved in diverse phases/pathways, which showed extensive gene expression variation among different morphotypes and different tissues. For example, two genes (*BoESM1* and *BoPYK10.1*) involved in GSL degradation in cluster III were extremely highly expressed in roots in all the five morphotypes but remarkably lowly expressed in all other tissues. Only six genes (*BoCYP79C1*, *BoCYP79C2.1*, *BoMAM3.1*, *BoMYB118*, *BoNIT1;2;3* and *BoSD1.3*) were not expressed in any tissue. The remaining 147 genes were all differentially expressed either between tissues or between morphotypes. This is well demonstrated by the heatmap constructed using z-scored TPM values (Fig. 2b). It is also clearly shown that many GSL genes were expressed at a higher level in root than in other tissues for broccoli, kohlrabi and white cabbage. Interestingly, those are generally not the same sets of genes and differ between the three mentioned morphotypes. In cauliflower, many genes were highly expressed in both roots and florets. For kale, and unlike the other four morphotypes, only a few genes were highly expressed in roots as compared to the other tissues.

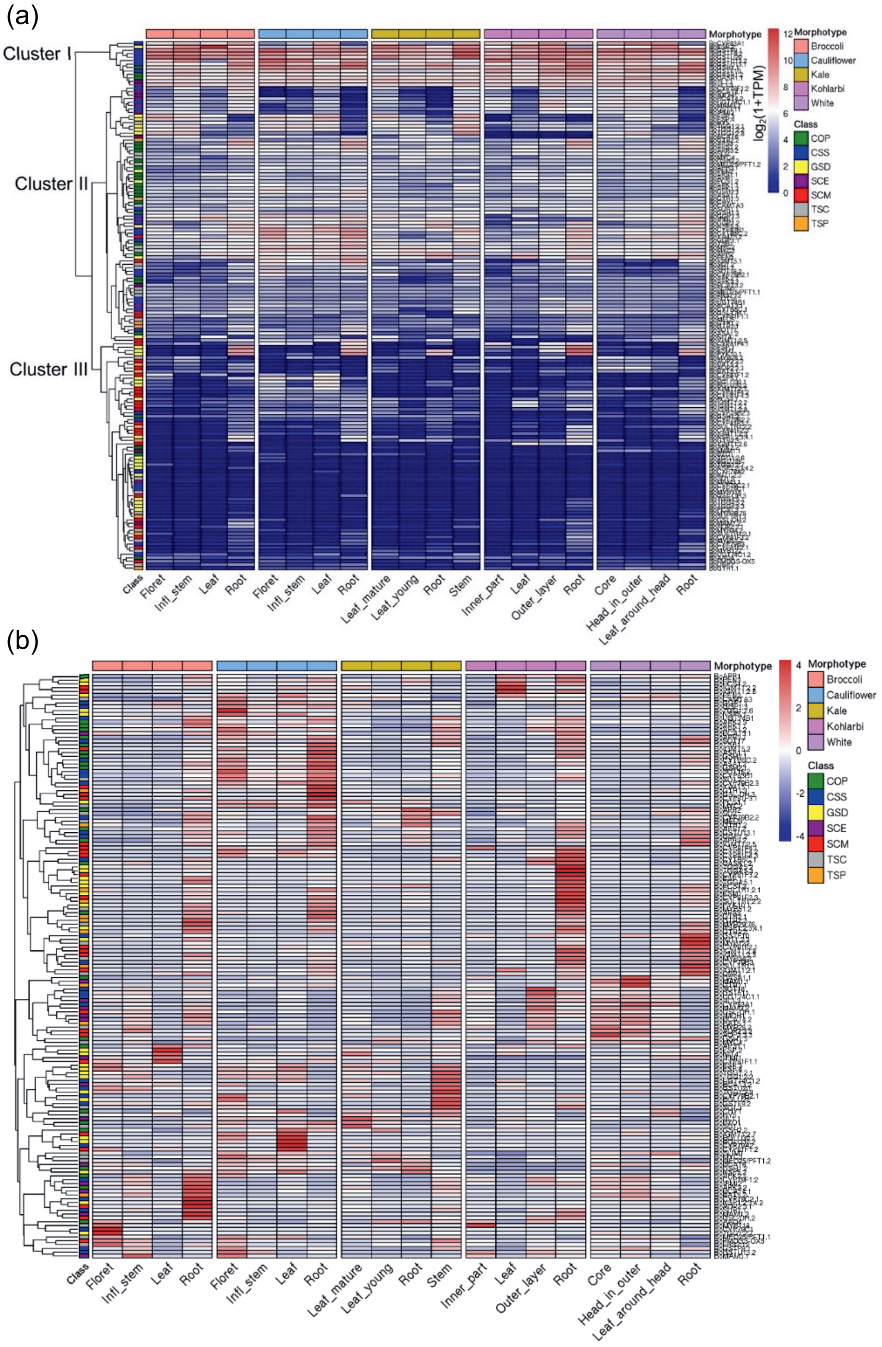

(a)
Cluster I
Cluster II
Cluster III
Morphotype
log2(1+TPM)
12
10
8
6
4
2
0
Morphotype
Broccoli
Cauliflower
Kale
Kohlarbi
White
Class
COP
CSS
GSD
SCE
SCM
TSC
TSP
Class
Floret
Infl_stem
Leaf
Root
Floret
Infl_stem
Leaf
Root
Leaf_mature
Leaf_young
Root
Stem
Inner_part
Leaf
Outer_layer
Root
Core
Head_in_outer
Leaf_around_head
Root
(b)
Morphotype
4
2
0
-2
-4
Morphotype
Broccoli
Cauliflower
Kale
Kohlarbi
White
Class
COP
CSS
GSD
SCE
SCM
TSC
TSP
Class
Floret
Infl_stem
Leaf
Root
Floret
Infl_stem
Leaf
Root
Leaf_mature
Leaf_young
Root
Stem
Inner_part
Leaf
Outer_layer
Root
Core
Head_in_outer
Leaf_around_head
Root

**Fig. 2 Expression profiles for GSL related genes in four tissues of five *B. oleracea* morphotypes**. Heatmaps were constructed using (a) log2 transformed and (b) z-scored TPM values. Blue and red colors are used to represent low to high expression levels, respectively. Genes are classified based on their involvement in different processes/phases (The abbreviations: COP: Cosubstrate Pathways, CSS: Core Structure Synthesis, GSD: GSL Degradation, SCE: Side-Chain Elongation, SCM: Side-Chain Modification, TSC: Transcriptional Components, TSP: Transporters).

## Correlation between gene expression and GSL levels

To examine the correlation between the accumulation of GSLs and the expression of GSL related genes in *B. oleracea*, we performed a Pearson correlation analysis across all the 20 tissues. Out of the total 3,381 combinations (147 expressed genes × 23 accumulating GSLs), 289 (8.55%) of them showed a significant positive (251 combinations, r = 0.44~0.95, $P < 0.05$) or negative (38 combinations, r = -0.61~-0.45, $P < 0.05$) correlation, which involve all the 23 GSLs and 109 genes (Fig. 3a and 3b). We found that two white cabbage signature GSLs, 5-methylthiopentyl and 4-methylthiobutyl, were significantly correlated with the highest (28 genes) and second highest (26 genes) number of genes, respectively. The kale signature GSL 2-phenylethyl was significantly correlated with 23 genes (Table S5 and S6). Forty-five out of the 109 genes were significantly correlated with a single GSL (Table S6). We also identified some genes that were strongly correlated with diverse GSLs (Fig. 3b, Table S5 and S6). For example, *BoAPR2*, a homolog of *AtAPR2* that is assumed to be involved in GSL cosubstrate pathways, was strongly correlated with eight aliphatic, one aromatic and one indolic GSLs. *MYB122* is identified as a transcription factor that is needed for indolic GSL biosynthesis in Arabidopsis (Frerigmann and Gigolashvili, 2014). Interestingly, in our morphotypes/tissues samples we discovered significant positive correlations between *BoMYB122* and eight aliphatic GSLs, rather than indolic GSLs. We also performed above correlation analysis after pooling the TPM values for paralogous GSL genes in *B. oleracea* (Fig. S6), from which a total of 167 gene-GSL combinations showed significant correlation involving all the 23 GSLs and 62 pools of paralogous genes. Together, these genes that show significant correlation with GSLs may play an important role in regulating GSL biochemical pathways.

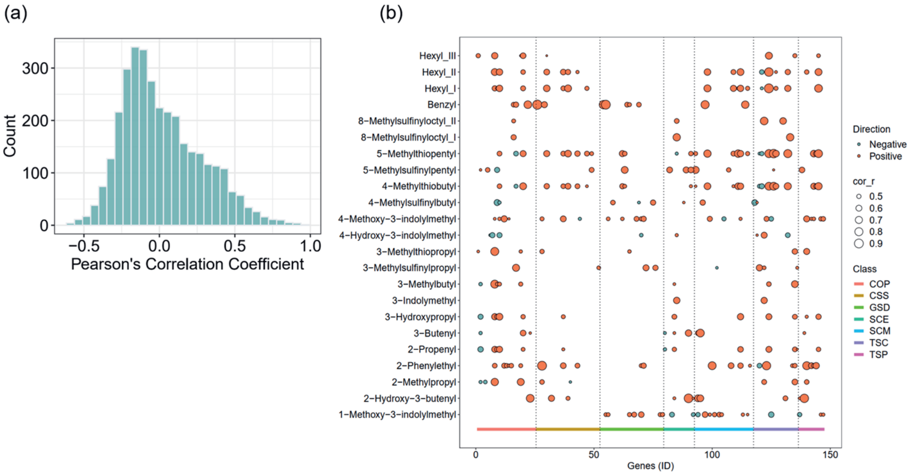


**Fig. 3 Pearson's correlation analysis between GSLs and their related genes**. (a) Distribution of Pearson's correlation coefficient. (b) Significantly ($P < 0.05$) correlated GSLs and genes. Genes are classified based on their involvement in different processes/phases as shown in Fig. 2. See Table S5 for the source data. (The abbreviations: COP: Cosubstrate Pathways, CSS: Core Structure Synthesis, GSD: GSL Degradation, SCE: Side-Chain Elongation, SCM: Side-Chain Modification, TSC: Transcriptional Components, TSP: Transporters).

We then focused on specific genes putatively involved in side-chain modification processes in aliphatic GSL biosynthesis to investigate the correlation between gene expression and GSL levels (Fig. 4a), namely *FMO*$_{GS\text{-}OX}$, *AOP* and *GSL-OH* locus and the methionine derived C3, C4 and C5 aliphatic GSLs. We identified two paralogs of *FMO*$_{GS\text{-}OX}$, four paralogs of *AOP* and four paralogs of *GSL-OH* in the *B. oleracea* genome (Fig. 4b and Fig. S7) and detected four C3, four C4 and two C5 methionine derived aliphatic GSLs (Fig. S3). In the C3-GSL biosynthesis pathway, the enzyme encoded by the *FMO*$_{GS\text{-}OX}$ gene converts 3-methylthiopropyl into 3-methylsulfinylpropyl (Fig. 4a). We observed the highest abundance of 3-methylthiopropyl and the lowest abundance of 3-methylsulfinylpropyl in roots of cauliflower, kale and white cabbage (Fig. S7a); in the same samples the gene expression levels of the *FMO*$_{GS\text{-}OX}$ paralogs were relatively low (though one paralog was not identified in kale) (Fig. S8), which pattern may thus explain the corresponding relative high and low levels of 3-methylthiopropyl and 3-methylsulfinylpropyl, respectively. This result suggests a direct relation between expression levels of *FMO*$_{GS\text{-}OX}$ paralogs and the relative abundance of 3-methylthiopropyl and 3-methylsulfinylpropyl. In the next step of this pathway, *AOP2* and *AOP3* convert 3-methylsulfinylpropyl into 2-propenyl and 3-hydroxypropyl, respectively. The relative levels of both GSLs are highest in roots of cauliflower, kale and white cabbage (Fig. S7a). However, we did not observe any *AOP* paralogue displaying a higher expression level in roots than in other tissues of these three morphotypes (Fig. S9). On the contrary, their expression was almost absent or relatively low in all cauliflower and kale tissues, and relatively high for two out of three paralogues in the non-root tissues of white cabbage. With regard to C4-GSLs, we observed a high accumulation of 4-methylsulfinylbutyl while 3-butenyl was undetectable in all broccoli tissues (Fig. 4b). However, *AOP* paralogues were expressed in several broccoli tissues including roots (Fig. S9). 3-butenyl was also low in cauliflower, kohlrabi and kale, while it highly accumulated in all white cabbage tissues tested. This was related to a high expression of *AOP's* in white cabbage, compared to lower levels in all other morphotypes. In the next step of this pathway, *GSL-OH* converts 3-butenyl into 2-hydroxy-3-butenyl (Fig. 4). Both these two compounds highly accumulated in white cabbage in all tissues, and were not detectable in the other four morphotypes (Fig. 4b). Interestingly, one paralog of *BoGSL-OH* (*BoGSL-OH.1*) had a high expression level in all the five morphotypes (Fig. S10). Besides the C4-GLS 4-methylsulfinylbutyl, broccoli also accumulated a relative high level of the C5-GSL 5-methylsulfinylpentyl in all its tissues, while this compound was hardly detectable in any tissue of the four other morphotypes (Fig. S7b); its conversion product 4-pentenyl GSL was not detectable in any of the five morphotypes. Together, these observations suggest that expression levels of *AOP* and *GSL-OH* genes cannot explain the relevant GSL variation.

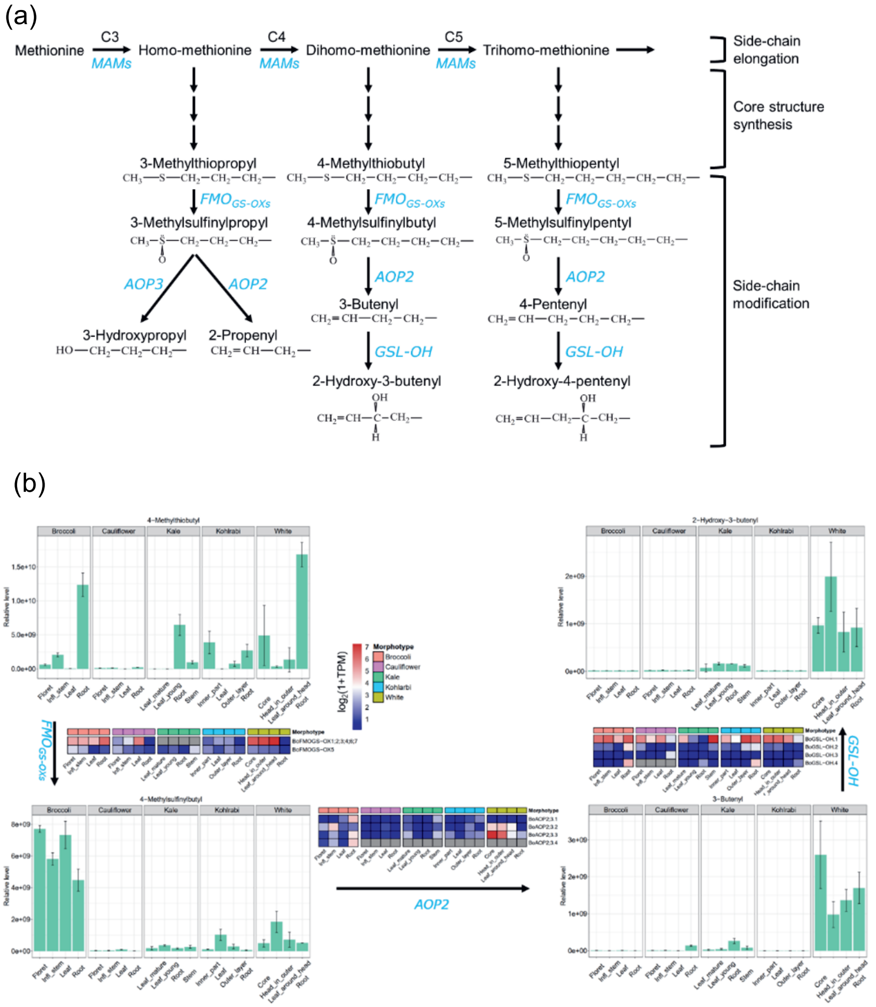


**Fig. 4 Aliphatic GSL biosynthesis pathway**. (a) A genetic model of the biosynthesis of aliphatic GSLs with different chain length. Figure was adapted from G. Padilla *et al.* (Padilla *et al.*, 2007). (b) C4 aliphatic GSL profiles and expression levels of related genes in different tissues and morphotypes. The bar charts show relative abundance of individual C4 GSLs in respective tissues and morphotypes. Error bars indicate standard deviation (n = 3 biological replicates). Heatmaps show gene expression levels. Blue and red colors are used to represent low to high expression levels, respectively. Gray color denotes that the gene is not identified in the corresponding morphotype. Note: *BoAOP2;3.3* is *BoAOP2*.

### Non-functional *AOP2* in broccoli

*AOP2* catalyzes the conversion of 3-methylsulfinylpropyl, 4-methylsulfinylbutyl and 5-methylsulfinylpentyl GSLs to the corresponding alkenyl GSLs of 2-propenyl, 3-butenyl and 4-pentenyl, respectively (Kliebenstein *et al.*, 2001b, Zhang *et al.*, 2015b) (Fig. 4a). It has been reported that broccoli harbors a non-functional *AOP2* allele, which results in accumulation of 4-methylsulfinylbutyl (Li and Quiros, 2003). Also in our study, both 4-methylsulfinylbutyl and 5-methylsulfinylpentyl levels were relatively high in broccoli (Fig. 4b and Fig. S7b), while their conversion products, 3-butenyl and 4-pentenyl GSL, respectively, were not detectable in broccoli (Fig. 4b and Fig. S7b). Based on sequence homology with Arabidopsis *AOP*s, we found three copies of *AOP* that are present in each of the five *B. oleracea* genomes; as *AOP2* and *AOP3* are highly homologous, it is difficult to define the function of the three *AOP* paralogues. Since white cabbage accumulated large amounts of 3-butenyl GSL in all its tissues (Fig. 4b), and *BoAOP2;3.2* and *BoAOP2;3.3* are the only two *AOP*s that were expressed in white cabbage (Fig. S9), it is suggested that *BoAOP2;3.2* or *BoAOP2;3.3* represents *BoAOP2.* While *BoAOP2;3.2* and *BoAOP2;3.3* are clearly expressed in both broccoli and white cabbage (Fig. S9), this does not lead to the expected accumulation of the enzyme product 3-butenyl GSL in broccoli, suggesting that *BoAOP2* is not functional in broccoli while it is so in white cabbage.

To better understand the underlying genetic factor that may cause the non-functional *AOP2* specifically in broccoli, we compared the gene structure, motifs and domains of 15 *BoAOPs*, *i.e.* the three *AOP* paralogs that were present in each of the five morphotypes. The gene lengths of both *BoAOP2;3.1* (2,330-2,553 bp) and *BoAOP2;3.3* (1,883-1,898 bp) homologs were similar among the five *B. oleracea* morphotypes, but more variable for the *BoAOP2;3.2* (660-4,479 bp) homologs (Fig. 5a). The gene structure varied among the 15 *BoAOPs*, with most genes having 3-4 exons (Fig. 5a). Accordingly, the motif compositions also varied among the 15 *BoAOPs*. All their encoded proteins contain three conserved motifs (motif 1, 2 and 6), while other motifs varied between the 15 genes. We identified a total of five conserved domains across the 15 encoded proteins and each *BoAOP* contained up to three domains (Fig. 5a). Interestingly, the conserved 2OG-FeII_Oxy domain at the C-terminal was present in all *BoAOPs* except for two, *i.e.* broccoli *BoAOP2;3.3* (*BolC9g002510.Br*) and cauliflower *BoAOP2;3.2* (*BolC3g035820.Ca*). This protein domain is known to be essential for 2-oxoglutarate/Fe(II)-dependent dioxygenase activity, which is associated with an important class of enzymes that mediate a variety of oxidative reactions (Prescott and Lloyd, 2000, Zhang *et al.*, 2015b). The absence of 2OG-FeII_Oxy domain in *BoAOP2;3.3* specifically in broccoli further suggests that *BoAOP2;3.3* represents *BoAOP2*. The gene structure comparison and multiple alignments of the amino acid sequences of the five *BoAOP2's* clearly showed a novel intron in broccoli between exon 2 and exon 3 (Fig. 5a and Fig. 5b), which possibly results in the absence of the 2OG-FeII_Oxy domain of its encoded enzyme. The amino acid sequence in this region is identical in the other four morphotypes, suggesting that *BoAOP2* is functional in these four morphotypes. While the relative level of 3-butenyl GSL is high in white cabbage, it is very low in any tissue of cauliflower, kale and kohlrabi; this can be due to either the low expression of *BoAOP2* or a lack of its precursor 4-methylsulfinylbutyl GSL.

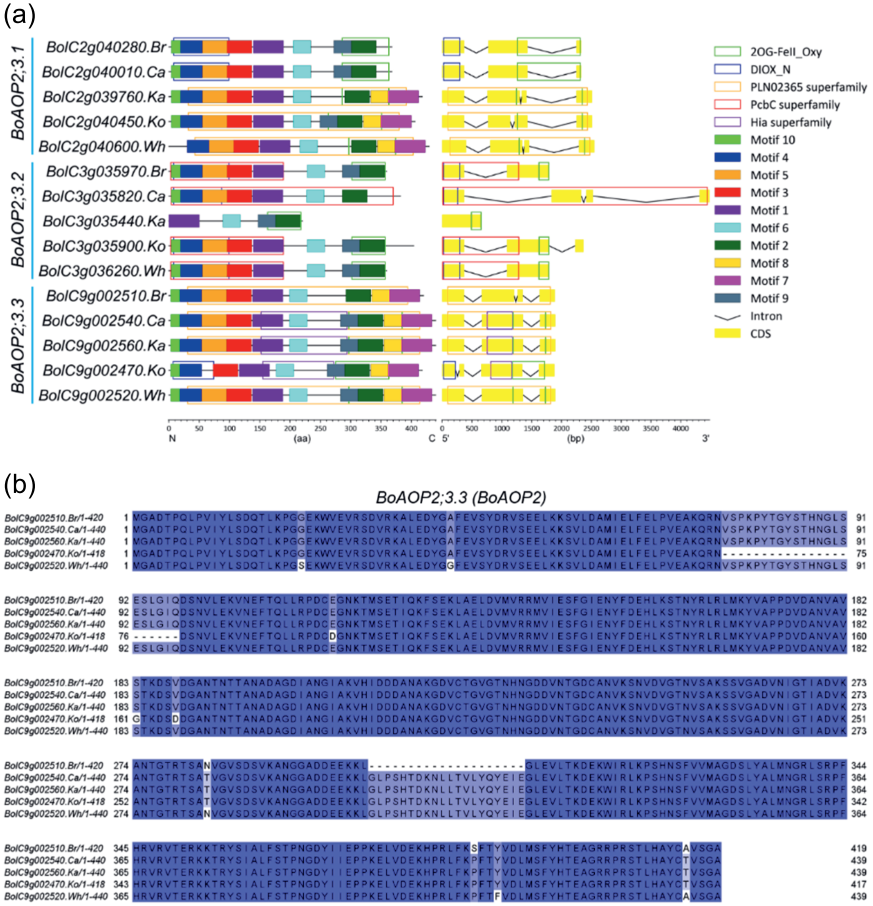


**Fig. 5** ***AOPs*** **in five** ***B. oleracea*** **genomes**. (a) Diagram of motif, domain and gene structure of *AOP* genes in *B. oleracea*. Different colored squares and boxes represent different motifs and domains, respectively. (b) Multiple alignments of the *BoAOP2* (*BoAOP2;3.3*) amino acid sequences.

## Transposable element insertions result in long and repeat-rich intronic sequences in *BoMAM3.2*

In *A. thaliana*, the *MAM* genes encode enzymes involved in chain elongation and produce GSLs with diverse chain-lengths during the biosynthesis of methionine derived GSLs (Kliebenstein *et al.*, 2001a). *MAM1* catalyzes the condensation reaction of the first two elongation cycles, while *MAM3* is considered to contribute to the generation of all GSL chain lengths (Textor *et al.*, 2007, Benderoth *et al.*, 2009, Liu *et al.*, 2014a). We identified two paralogs of *MAM1* and two paralogs of *MAM3* in each of the five *B. oleracea* morphotypes. Structures and motifs of *MAM* genes in these genomes were then analyzed. Most *MAM* genes shared conserved gene structures (Fig. 6a). From these 20 *MAMs*, three (white cabbage *MAM1.1* and *MAM1.2*, and kale *MAM3.1*) lost five to six conserved motifs and were thus heavily differentiated from the other homologs. We only identified two conserved domains across the 20 *BoMAMs* using MEME tool (Bailey *et al.*, 2009). Interestingly, nine out of ten MAM1 proteins contain TIM superfamily domain and nine out of ten MAM3 proteins contain PLN03228 domain, with each MAM1 and MAM3 having one protein displaying the opposite pattern (a cabbage *MAM1.1* and a kale *MAM3.1*) (Fig. 6a).

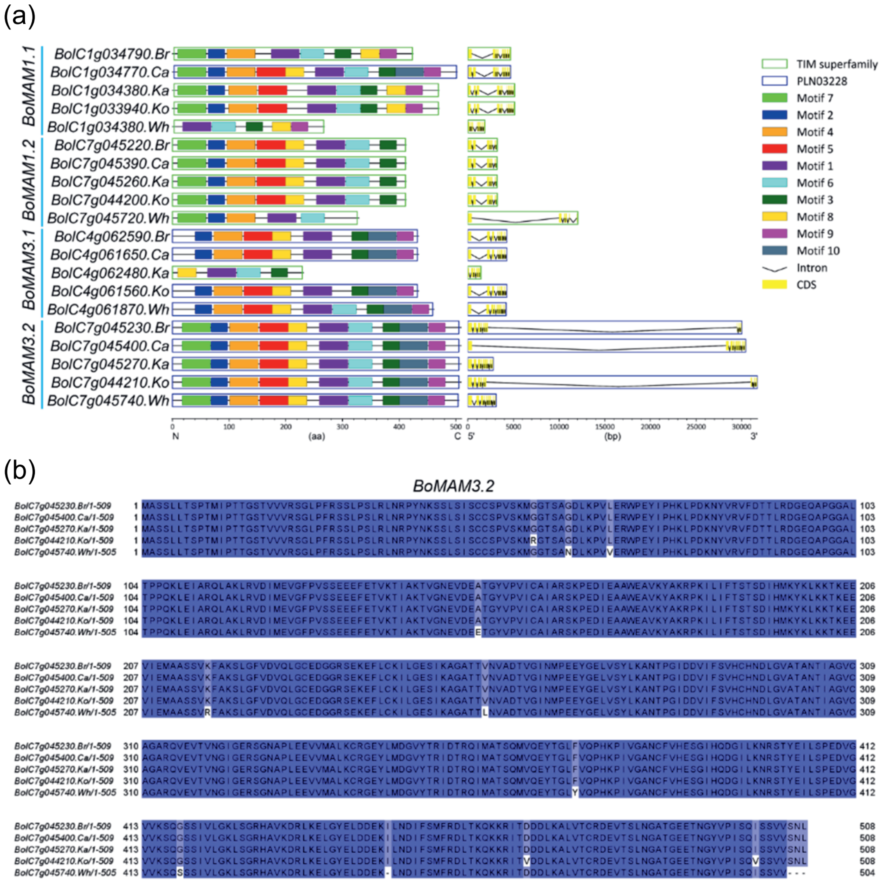


**Fig. 6** ***MAMs* in five *B. oleracea* genomes**. (a) Diagram of motif, domain and gene structure of *MAM* genes in *B. oleracea*. Different colored squares and boxes represent different motifs and domains, respectively. (b) Multiple alignments of the *BoMAM3.2* amino acid sequences.

The five proteins of *BoMAM3.2* have identical motif compositions, with each containing all ten motifs. Accordingly, multiple alignments of the amino acid sequences of five *BoMAM3.2* homologs displayed remarkably high sequence similarity (Fig. 6b). However, these genes strikingly varied in length, ranging from 2,795 to 31,732 bp. By comparing structure of the five *BoMAM3.2* homologs, we found three extremely long introns (27,317-29,020 bp) in broccoli, cauliflower and kohlrabi. The predicted structures of these five *BoMAM3.2* genes are well supported by mRNA-seq evidences (Fig. S12). Interestingly, we observed that a few mRNA-Seq reads could still be mapped to the three large intronic regions (Fig. S12). A closer inspection of the three introns indicates that they were composed of many transposable elements (TEs), especially DNA transposons (Table S7). For example, we found a 2.8Kb DTC in broccoli and cauliflower, and a 3.6Kb DTC in cauliflower and kohlrabi. Instead of introducing new introns, these TEs were inserted in different existing introns in the three *BoMAM3.2* genes. Gene expression profiles showed that *BoMAM3.2* was highly expressed in all five genomes in varying degrees, while the expression of its paralog *BoMAM3.1* was hardly detectable in any morphotype (Fig. S11). Also, most *BoMAM1* (*BoMAM1.1* and *BoMAM1.2*) genes were hardly expressed, except for low expression levels in broccoli and kohlrabi roots and in all white cabbage tissues. While broccoli plants did not accumulate any C3 aliphatic GSL, they accumulated 4-methylsulfinylbutyl and 5-methylsulfinylpentyl GSLs (Fig. 4b and Fig. S7), which points to a functional *BoMAM3.2* (as *BoMAM3.1*, *BoMAM1.1* and *BoMAM1.2*'s are not or very low expressed) gene converting all C3 GSLs to longer-chain GSLs. We also observed several long-chain aliphatic GSLs accumulating in the other morphotypes, such as 8-methylsulfinyloctyl GSL in both cauliflower and kale and possibly the less well described hexyl_III GSL in kale (Fig. S3), suggesting that *BoMAM3.2* is also active in the other morphotypes. These TE insertion activities in *BoMAM3.2* may modify the expression level of the encoded gene.

## Discussion

In the present study, we generated comprehensive GSL profiles (23 compounds) of four selected tissues in five different *B. oleracea* morphotypes, revealing extensive variation in GSL composition and their relative levels between both tissues and morphotypes. Taking advantage of the high-quality genome assemblies of these five morphotypes (Cai *et al.*, 2022b), we generated a comprehensive list of GSL related genes in each of the five genomes based on comparative genomics with *A. thaliana*. Expression of these genes was then studied using mRNA-Seq data generated from the same tissues that were used for GSL profiling. We found a total of 289 GSL-gene combinations with significant correlation between GSL abundance and gene expression level, which involve all the 23 GSLs and 109 genes. These genes may play an important role in regulating GSL biochemical pathways. Our results also confirm the proposed non-functional *AOP2* in broccoli and suggest active *AOP2* homologs in other morphotypes, shedding light on effects of allelic composition on metabolite production. Moreover, we revealed extremely long intronic sequences in one paralog of *MAM3* in three *B. oleracea* genomes, which are caused by many TE insertion activities. Our work provides a molecular basis for understanding and further elucidating the genetic control of GSL profiles and for breeding with tailored GSL content in *B. oleracea*.

Our results indicate a large variation in both the abundance and the composition of GSLs between plant tissues and morphotypes of *B. oleracea*. Many GSL forms were found to accumulate at a higher level in roots than in other tissues (Fig. 1 and Fig. S3), such as 3-hydroxypropyl, 3-methylthiopropyl, hexyls and 2-phenylethyl GSLs. Also, the composition of GSLs differs between morphotypes. For example, broccoli and cauliflower lack C3 and C5 aliphatic GSLs, respectively. Signature GSLs were identified for four of the five morphotypes (Table S1), suggesting significant GSL relative abundance differences between the studied morphotypes. Glucoraphanin (4-methylsulfinylbutyl GSL) and glucoiberin (3-methylsulfinylpropyl GSL) are the two most desirable GSLs, in view of the nutritional value of their breakdown products (Wang *et al.*, 2012, Yi *et al.*, 2015). Among the five *B. oleracea* morphotypes, only broccoli accumulated relative high levels of glucoraphanin in all its tissues (Fig. S3b), while glucoiberin was relative high in leaves and stems of both cauliflower, kale and white cabbage, but undetectable in broccoli (Fig. S3a). In contrast to these two wanted GSLs, progoitrin (2-hydroxy-3-butenyl GSL) is unwanted since it can be hydrolyzed into oxazolidine-2-thione, which causes goiter and other harmful effects in mammals (Voorrips *et al.*, 2000, Tripathi and Mishra, 2007). Interestingly, progoitrin was relatively high in all tissues of white cabbage, while it was not detectable in any of the sampled tissues of the other four morphotypes (Fig. S3b). Previously, Yi *et al.* (Yi *et al.*, 2015) showed that progoitrin was absent in florets in one of the three cauliflower genotypes investigated. Wang *et al.* (Wang *et al.*, 2012) found comparatively higher progoitrin in commercial broccoli genotypes than in inbred lines. These suggest that GSL content also varies between genotypes of the same morphotype in *B. oleracea*. To breed varieties with tailored GSL content, the information with regard to which genotype and organ these desired and unwanted compounds are accumulated is vital for parental selection.

Several reasons could possibly explain why the vast majority (91.45%) of GSL-gene combinations (3,092 out of 3,381 combinations) do not show correlation between GSL relative abundance and gene expression level (Fig. 3). First, the GLS biosynthesis pathway is very complex, with many compounds being produced that can either accumulate or convert into derived compounds (like the side-chain modification steps), or even convert into a breakdown product (like the ITCs, nitriles and indoles). The relative abundance of an intermediate GSL does not necessarily show a correlation with the expression level of its regulating genes. As an example, our data suggest a relation between the relative abundances of intermediate GSLs 3-methylthiopropyl and 3-methylsulfinylpropyl and expression level of *FMO*$_{GS\text{-}OX}$ paralogs. However, we did not detect the correlation. A direct correlation between activity of a structural enzyme (presuming this activity is fully regulated by its gene expression level) and its product can however only be expected if the next step in a pathway is absent or highly limiting. Second, GSL transporters have been identified and experimentally verified (Nour-Eldin *et al.*, 2012, Jensen *et al.*, 2014), which establishes dynamic GSL patterns between source and sink tissues in Arabidopsis. In our study, we observed that *AOP2* was not expressed in roots of cauliflower and kale whereas the derived products 2-propenyl and 3-butenyl accumulated in roots of these two morphotypes, suggesting that these GSLs are likely transported from source tissues (*i.e.* leaves) to roots. Interestingly, significant correlations were identified between these compounds and several GSL transporters (*i.e. BoGTR1.1* and *BoSULTR1;1*), suggesting active transport. Expression of several transport genes (*i.e. GTR*s and *SULTR*s) was indeed high in roots (Fig. 2b). The function of transporters is to import GSL from the apoplastic apace to the symplast (Nour-Eldin *et al.*, 2012, Jensen *et al.*, 2014). They should be present in both sink and source tissues. The long distance transport is via the vascular system. Thus, GSL compounds accumulated in a specific tissue do not necessarily mean that they are originally produced in this tissue, and so are not directly correlated to the expression level of a responsible biosynthetic gene in that specific tissue. Third, the biosynthesis and regulation of GSLs are well studied in the model plant *A. thaliana* while most studies in *Brassica* crops are based on this reference pathway (Harun *et al.*, 2020). The function of these presumed genes in *Brassica* has to be further verified. Due to the whole genome triplication event in *Brassica*, many genes are present in multiple copies, which may display different expression patterns. Due to sequence divergence along evolution, it is also difficult to know which copy is functional. In addition, non-functional genes can still be quantified as expressed by mRNA-Seq data, such as the non-functional *AOP2* in broccoli. Lastly, Yu *et al.* (Yu *et al.*, 2020) suggested that much of the regulation of metabolite levels in tea may not occur only at the transcriptional level but at multiple levels, such as transcriptional, post-transcriptional, translational, post-translational and epigenetic levels, which also seems possible for GSLs in *Brassica* crops. Despite all these challenges, we still detected 289 GSL-gene combinations with significant correlation between GSL relative abundance and gene expression level. Interestingly, several GSLs were correlated with the expression of many genes that were involved in different processes (like core structure synthesis pathway, side-chain elongation and modification, but also co-substrate and transport and regulatory pathways). In order to better understand GSL biosynthesis and identify genetic loci controlling GSL production in *Brassica* crops, more extensive and in-depth studies, such as genome-wide association analysis of genomic SNPs, transcriptomic and GSL data from a large number of samples, the construction of gene regulatory networks, and mGWAS, will be needed.

In our study, we investigated the sequence divergence of two important GSL genes (*AOP2* and *MAM3*) among the five different *B. oleracea* morphotypes. *AOP2* is not functional in broccoli, which results in the accumulation of two main health-promoting compounds in this morphotype: 4-methylsulfinylbutyl (Glucoraphanin) and 5-methylsulfinylpentyl (Glucoalyssin). However, the functional *AOP2* in white cabbage converts 4-methylsulfinylbutyl into highly accumulated 3-butenyl. The functional divergence of *AOP2* between these morphotypes is likely attributed to their sequence divergence, with *AOP2* in broccoli lacking a conserved 2OG-FeII_Oxy domain. We found many TE insertion activities in one paralog of *MAM3* (*BoMAM3.2*) in three morphotypes, which resulted in the extremely long (~29Kb) and repeat-rich intronic sequences. As these insertions happen in existing introns and we identified long-chain aliphatic GSLs for which *MAM3* is responsible, the gene function is unlikely to have been altered. Cai *et al.* (Cai *et al.*, 2022c) reported that TE insertions within introns tend to largely modify gene expression levels. We observed that this gene is differentially expressed among tissues and among morphotypes (Fig. S11) and the TE insertion activities may have contributed to the expression difference between morphotypes. In addition, extremely long introns seem to be prevalent in plant genomes. For example, they have been found in Arabidopsis, with lengths larger than 5Kb or even 10Kb (Chang *et al.*, 2017). Accordingly, Liu *et al.* (Liu *et al.*, 2021a) reported that ginkgo possesses very long introns characterized by many repeat-element insertions, with the 10% of its longest intron even greater than 100Kb. It is also reported that large introns are involved in regulating gene expression levels (Kim *et al.*, 2006, Rigal *et al.*, 2012), probably through intron DNA methylation.

In conclusion, we profiled GSLs and mRNA-Seq in both roots, leaves and the edible parts of five different *B. oleracea* morphotypes, revealing strong variations of GSL relative abundance and composition as well as GSL related gene expression. We found a total 289 GSL-gene combinations with significant correlation between GSL and gene expression level, which involve all the 23 GSLs and 109 related genes. We observed a non-functional *AOP2* in broccoli, which is related to the loss of a conserved 2OG-FeII_Oxy domain, and found many TE insertions in one paralog of *MAM3* gene in three genomes, resulting in long and repeat-rich intronic sequences.

## Materials and Methods

### Plant materials and sample collection

For the current study, we used five homozygous lines (broccoli, cauliflower, kale, kohlrabi and white cabbage) for metabolite extraction and mRNA sequencing, the genomes of which were previously *de novo* assembled (Cai *et al.*, 2022b). The seeds were sown in April 2020 in a single greenhouse compartment at Unifarm (Wageningen University and Research) and samples were harvested between July to September 2020 depending on the maturity of plants. A completely randomized block design with three blocks was used for plant growth. Over each block, three plants of each accession were randomly distributed. As shown in Fig. S1, we collected four different tissues from each accession with three biological replicates. For each biological replicate, equal weight of tissue from the three plants per accession in the same block was pooled. These samples were immediately frozen in liquid nitrogen, ground into a fine powder and stored at -80°C until further use.

## Glucosinolate profiling

Intact GSLs were determined using HPLC coupled to both UV/vis and accurate mass detection (LCMS). GSLs were extracted from 200 mg powder to which 600 µL of 99.87% methanol containing 0.13% formic acid was added, followed by 15 min sonication and then 15 min centrifugation at 16,000g. The clear supernatants were directly used for LCMS analysis conforming with Jeon *et al.* (Jeon *et al.*, 2021), using a Dionex U-HPLC sequentially coupled to a photodiode array detector and a Q-Exactive Orbitrap FTMS (Thermo Scientific). In short, five µl of each extract was injected into an Alliance 2795 HT instrument (Waters) and compounds were separated on a C18 column (Phenomenex Luna, 2.0 mm × 150 mm, 3 µm particle size) using a 45 min gradient from 5 to 35% acetonitrile acidified with 0.1% formic acid. Electrospray ionization in negative mode at a mass resolution of 60,000 FWHM was used to detect eluting compounds. GSLs were identified based on the observed accurate masses and relative retention times (Bonnema *et al.*, 2019), allowing a maximum deviation of 3 ppm from the calculated molecular ion masses. Chromatographic peaks areas of GSLs were subsequently integrated using the QualBrowser module of Xcalibur version 4.1 (Thermo Scientific).

## mRNA extraction and sequencing

Total RNA was extracted from the frozen powders using the TRIZOL reagent (Invitrogen) according to manufacturer's protocol and treated with RNase-free DNase I (Invitrogen, Carisbad, CA, USA) to remove genomic DNA contaminations. Total RNA was cleaned using the cleanup protocol of the RNeasy Mini Kit (Qiagen, the Netherlands) according to supplier's recommendations. RNA quantity and quality were assessed using a NanoDrop™ One Spectrophotometer (ThermoFischer

Scientific, USA), agarose gel electrophoresis, and a Qubit RNA BR Assay Kit (Thermo Fisher Scientific) on a Qubit 4 Fluorometer. mRNA-Seq libraries were prepared using the Illumina TruSeq RNA Sample Prep Kit and sequenced on Illumina NovaSeq platform with 150bp paired-end reads.

**Reads mapping and gene expression profiling**

Low quality reads were removed using fastp (v0.19.5) (Chen *et al.*, 2018) with parameters "-q 15 -u 40 -n 5 -l 100 --trim_poly_x --detect_adapter_for_pe". To minimize alignment errors, we mapped the clean reads to the five corresponding reference genomes (Cai *et al.*, 2022b) using Hisat2 (v2.1.0) (Kim *et al.*, 2015) with parameters "--dta". Read counts for each gene were computed using htseq-count (part of the HTSeq version 0.12.4) (Putri *et al.*, 2022) with parameters "-s no -q -f bam -r pos". Hierarchical clustering and principle component analysis (PCA) were performed using PCAPlot and clusterPlot function in SARTools (v1.8.1) package (Varet *et al.*, 2016) to check qualities for mRNA-seq replicates. Stringtie (v2.1.1) (Kovaka *et al.*, 2019) was utilised to compute expression level of genes in terms of transcripts per kilobase of exon model per million mapped reads (TPM), with parameters "-e -B". Genes with an average TPM ≥ 1 across the three biological replicates were considered as expressed. If a gene is expressed in any tissue of a morphotype, it is considered as expressed for the given morphotype.

**Identification of GSL genes in five *B. oleracea* morphotypes**

OrthoFinder (v2.3.12) (Emms and Kelly, 2019) was used to detect orthologs based on protein sequences from five *B. oleracea* (Cai *et al.*, 2022b) and one *A. thaliana* genomes (TAIR10) with default parameters. A comprehensive *A. thaliana* GSL gene list was obtained from Harun *et al.* (Harun *et al.*, 2020), which includes a total of 113 genes encoding 85 enzymes, 23 transcriptional components and five protein transporters. Sequences of these GSL genes were retrieved from the TAIR database (https://www.arabidopsis.org/). To identify GSL genes in the five *B. oleracea* genomes, all *A. thaliana* GSL genes were compared with the orthologs identified between *A. thaliana* and each of the five *B. oleracea* genomes. We extracted all the *B. oleracea* genes that are orthologous to *A. thaliana* GSL genes. Subsequently, we filtered these candidate GSL homologs based on blastp alignments between all *B. oleracea* and *A. thaliana* GSL protein sequences with a cutoff *E* value ≤ $1 \times 10^{-20}$, coverage ≥ 50% and identity ≥ 35%.

**Motif and domain analysis and multiple sequence alignment**

The MEME (https://meme-suite.org/meme/) tool was used to identify conserved motifs for selected GSL genes in *B. oleracea* genomes, with a maximum number of 10 motifs and a motif width of 6-50 (Bailey *et al.*, 2009). NCBI-CDD (https://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi) was used to search for the conserved domains. CFVisual software (https://github.com/ChenHuilong1223/CFVisual) was used to visualize gene structure and distribution of motifs and domains (Chen *et al.*, 2022). Protein sequences of interested genes were aligned using MAFFT (Katoh *et al.*, 2005) with default parameters and were then visualized with Jalview 2 (v2.11.2.4) (Waterhouse *et al.*, 2009).

## Supporting information

Supplementary files are available at https://github.com/cai1991/GSL

**Fig. S1** The five *B. oleracea* morphotypes and collected tissues for GSL extraction and RNA sequencing.

**Fig. S2** Principle component analysis (PCA) based on GSL data showing overall variation between the three biological replicates (20 samples × 3 biological replicates).

**Fig. S3** Relative quantity of individual GSL in four tissues of five *B. oleracea* morphotypes. (a) Methionine derived aliphatic C3 GSLs. (b) Methionine derived aliphatic C4 GSLs. (c) Methionine derived aliphatic C5 GSLs. (d) Methionine derived aliphatic C8 GSLs. (e) Branched-chain amino acid derived aliphatic GSLs. (f) Aliphatic hexyl GSLs. (g) Indolic GSLs. (h) Aromatic GSLs. The Y-axis shows the peak surface area measured in LCMS for the indicated compound. Error bars indicate standard deviation (n = 3).

**Fig. S4** Distribution of GSL related genes in the five *B. oleracea* genomes. See Table S3 for source data.

**Fig. S5** PCA plots and cluster dendrogram based on mRNA-Seq data. In PCA plots, morphotypes from the top to bottom are: broccoli, cauliflower, kale, kohlrabi and white cabbage. Sample IDs in cluster dendrogram correspond to those used in PCA plots. Abbreviations in the PCA legends: BrLe: Broccoli Leaf, BrRo: Broccoli Root, BrIs: Broccoli infl_stem, BrFl: Broccoli Floret, CaLe: Cauliflower Leaf, CaRo: Cauliflower Root, CaIs: Cauliflower Infl_stem, CaFl: Cauliflower Floret, KaLm: Kale Leaf_mature, KaLy: Kale Leaf_young, KaSt: Kale Stem, KaRo: sssKale Root, KoLe: Kohlrabi Leaf, KoOl: Kohlrabi Outer_layer, KoIn: Kohlrabi Inner_part, KoRo: Kohlrabi Root, WhCo: White Core, WhLh: White Leaf_around_head, WhHo: White Head_in_outer, WhRo: White Root.

**Fig. S6** Pearson's correlation analysis between GSLs and their related genes. Gene expression profiles (TPM values) are pooled for different copies of paralogous genes in *B. oleracea*. (a) Distribution of Pearson's correlation coefficient. (b) Significantly ($P$ < 0.05) correlated GSLs and genes. Genes are classified based on their involvement in different processes/phases as shown in Fig. 2. (The abbreviations: COP: Cosubstrate Pathways, CSS: Core Structure Synthesis, GSD: GSL

Degradation, SCE: Side-Chain Elongation, SCM: Side-Chain Modification, TSC: Transcriptional Components, TSP: Transporters).

**Fig. S7** C3 (a) and C5 (b) aliphatic GSL profiles and expression levels of related genes in different tissues and morphotypes. The bar charts show relative quantity of individual GSLs in respective tissues and morphotypes. Error bars indicate standard deviation (n = 3). Heatmaps show gene expression levels. Blue and red colors are used to represent low to high expression levels, respectively. Gray color denotes that the gene is not identified in the corresponding morphotype. Note: *BoAOP2;3.3* is *BoAOP2.*

**Fig. S8** Gene expression analysis of *FMO*$_{GS\text{-}OX}$ paralogues in four tissues in five *B. oleracea* morphotypes. The expression level was estimated using TPM values based on mRNA-Seq data. Error bars indicate standard deviation (n = 3).

**Fig. S9** Gene expression analysis of *AOP* paralogues in four tissues in five *B. oleracea* morphotypes. The expression level was estimated using TPM values based on mRNA-Seq data. Error bars indicate standard deviation (n = 3). Note: *BoAOP2;3.3* is *BoAOP2.*

**Fig. S10** Gene expression analysis of *GSL-OH* paralogues in four tissues in five *B. oleracea* morphotypes. The expression level was estimated using TPM values based on mRNA-Seq data. Error bars indicate standard deviation (n = 3).

**Fig. S11** Gene expression analysis of *MAM* paralogues in four tissues in five *B. oleracea* morphotypes. The expression level was estimated using TPM values based on mRNA-Seq data. Error bars indicate standard deviation (n = 3).

**Fig. S12** IGV snapshots showing mRNA-Seq alignments in *BoMAM3.2* genes from five *B. oleracea* morphotypes (Top to bottom: broccoli, cauliflower, kohlrabi, kale and white cabbage). In each snapshot, four tracks from top to bottom represent alignments in four different tissues.

**Table S1** Signature GSLs for the five *B. oleracea* morphotypes.

**Table S2** GSL related genes identified in the five *B. oleracea* genomes.

**Table S3** Position of GSL related genes identified in the five *B. oleracea* genomes.

**Table S4** Summary of RNA-Seq data and statistics for read mapping.

**Table S5** List of significantly correlated GSLs and related genes.

**Table S6** The number of GSLs/Genes that are significantly correlated with the given Gene/GSL.

**Table S7** TE annotations in the long intron of *MAM3* gene in three *B. oleracea* genomes.

**Table S8** Gene expression (TPM values) of GSL related genes in sampled tissues.

# Chapter 6

## General Discussion

*Brassica oleracea* includes economically important crops that exhibit enormous diversity in their appearance (ranging from cauliflower to kohlrabi and from cabbage to kale) and uses (ranging from fodder to vegetables to ornamentals). Artificial selection during domestication and breeding activities resulted in these highly diverse *B. oleracea* crops that are associated with different genomic architectures. *B. oleracea* is an ideal model for evolutionary and polyploidy studies in plants because it has undergone the Brassiceae-specific whole genome triplication (WGT) event, followed by rediploidization processes to form its extant genome architecture. In this thesis, we performed comprehensive genomics and genetics studies in *B. oleracea* crops, such as investigation of genetic diversity and domestication history, *de novo* genome assembly of five morphotypes, structural variation (SV) identification, genome evolution analysis, characterization of recombination landscape of inter-morphotype crosses, and investigation of genetic regulation underlying the variations of glucosinolate (GSL) profiles. In this general discussion, I will discuss how the results in **Chapters 2-5** can improve our understanding of genomic and genetic features of the highly diverse *B. oleracea* species, and the potential application of our results in breeding programs. As high-quality reference genomes can significantly facilitate investigations on genomic evolution, genomic variation and genetic loci controlling important traits, I will also discuss how to generate high-quality reference genomes in plants using state-of-the-art technologies. Furthermore, I will discuss the accelerating plant pan-genome studies, which are enhanced by increasing numbers of high-quality genome assemblies.

## Reflection on chapters of this thesis

### Better understanding of genetic diversity and domestication of *B. oleracea*

Genetic diversity is essential for a species to survive as it contributes to the ability to adapt to changing environments (Ellegren and Galtier, 2016, Maia and de Araújo Campos, 2021). In general, three different sources can contribute to the generation of genetic diversity: mutation, recombination and immigration (Maia and de Araújo Campos, 2021). To measure genetic diversity of a population, various molecular markers can be used, such as the traditional molecular markers including random amplified DNAs (RAPDs) (Williams *et al.*, 1990), restriction fragment length polymorphisms (RFLPs) (Botstein *et al.*, 1980), amplified fragment length polymorphic DNAs (AFLPs) (Vos *et al.*, 1995), simple sequence repeats (SSRs) (Litt and Luty, 1989) and etc (Fu *et al.*, 2016, Singh *et al.*, 2022). Compared with these traditional markers, single nucleotide polymorphisms (SNPs) markers are more widely distributed across the whole genome at remarkably higher abundance, which are also easy to be detected in a high throughput manner. An intuitive and simple measure of genetic diversity is to calculate the nucleotide diversity (π) based on molecular markers, which describes the average pairwise difference between all possible pairs of individuals in a population. Although the nucleotide diversity can be estimated at fine accuracy even with few samples, inclusion of more samples and more high-quality molecular markers will likely lead to more accurate estimation.

In *B. oleracea* crops, several previous studies have assessed the genetic diversity (van Hintum *et al.*, 2007, Farnham *et al.*, 2008, Izzah *et al.*, 2013, Pelc *et al.*, 2015, El-Esawi *et al.*, 2016). However, these studies are usually limited by at least one of the following factors: 1) very low numbers of molecular markers, 2) very few accessions, 3) not including the majority of described *B. oleracea* morphotypes, 4) lacking wild relatives that can be intercrossed with domesticated *B. oleracea* crops. For genetic diversity assessment, it is important to also include accessions from crossable wild species in the diversity panel. One reason is to understand whether these wild species already contribute to the present variation in domesticated plants. Another reason is that these wild species can be valuable sources to provide alleles for abiotic or biotic resistance. In **Chapter 2**, we generated an unprecedented collection of 912 globally distributed accessions representing ten *B. oleracea* morphotypes, wild *B. oleracea* and nine wild C9 species. To our knowledge, this is the most comprehensive *Brassica* C-group germplasm collection published so far. We obtained large amounts of high-quality SNPs (14,152 SNPs) for the estimation of π. Therefore, our study can provide a much more accurate estimate for the overall *B. oleracea* genetic diversity. By including both modern hybrids and old landraces accessions, we revealed that breeding activities in *B. oleracea* have reduced genetic diversity from old landraces to modern hybrids. Usually, the reduced genetic diversity can result in the loss of some natural defence mechanisms (*i.e.* disease, insect and abiotic defence), making the plants less resilient to environmental changes (Jump *et al.*, 2009). Our results highlight the importance of screening genebank germplasms in breeding programs to bring back, among others, the lost defence mechanisms in plants. For example, germplasms from regions with warmer and drier climates (*i.e.* Turkey, Syria and Lebanon in our collection) can be used to create sustainable and resilient varieties that can better withstand drought and heat in hot summers. An European project (BrasExplor, https://www6.inrae.fr/brasexplor/) aims to explore more genetic variation by sampling *B. oleracea* and *B. rapa* accessions of wild populations and local landraces from a large climatic gradient around the Mediterranean sea. With these plant materials, the genetic basis of adaptive traits to climate change will be dissected and new varieties with relevant desirable traits will be developed. It is important to note that most *B. oleracea* genebank accessions are heterogeneous. For this reason, representation by only one plant per genebank accession will result in underrepresentation of allelic variations. This is one of the limitation in our study (**Chapter 2**) and other relevant studies (Zhao *et al.*, 2005, Mabry *et al.*, 2021). Our data suggest a strong genetic bottleneck of cauliflower, which resulted in very low genetic variation. Indeed, a recent study revealed many selection signals in cauliflower in diverse molecular pathways (*i.e.* flowering time, floral identity, meristem proliferation, organ size and spirality), also suggesting a strong genetic bottleneck (Guo *et al.*, 2021).

We propose a two-step domestication scenario of *B. oleracea* crops based on the findings in **Chapter 2**, with the first step taking place in Western Europe where wild *B. oleracea* was domesticated into highly diverse kale populations and the second step occurring in the Middle East where these divergent kale populations were cultivated into modern *B. oleracea* crops. These diversified kale populations are likely transported from Western Europe to Middle East with the tin-trade routes in the Bronze age (around 3300-1000 BC). Our hypothesis is supported by various archaeological and literature evidences: 1) wild *B. oleracea* grows along the coasts of England, France and Spain, 2) highly diverse kale types are already described in ancient literatures, 3) cabbages and cauliflowers were mentioned in ancient literatures at around 400 BC, 4) tin was mined in Cornwall and Galicia and brought to the Middle East by ship around 2500 BC, 5) the boatmen took vegetables and seeds with them for the journey. In **Chapter 2**, we provided evidence for the highly diverse kale lineages and their potential role as progenitors of two domesticated lineages, the "leafy head" lineage (LHL) and the "arrested inflorescence" lineage (AIL), based on genealogical analysis and SVD-quarts species tree estimation. Moreover, we found that cabbages from the Middle East formed the first-branching cabbage-clade, supporting the hypothesis that cabbage domestication started in the Middle East. Also, we estimated a date of 2560 BC for the origin of *B. oleracea* breeding, consistent with a time frame of the emergence of tin-trade between the Middle East and Cornwall and Galicia (Berger *et al.*, 2019).

### High-quality genome assembly of five different *B. oleracea* morphotypes

Although SNP numbers in **Chapter 2** are abundant enough to perform genetic diversity and mapping studies, they only represent a small portion of genomic variations in the population. Using deep Illumina whole genome re-sequencing instead of the Sequence-Based Genotyping (SBG, a method that reduces genome complexity) approach used in this chapter, one can reveal much more small-scale genomic variations. However, numerous large and complex SVs cannot be revealed based on simply mapping short-reads or even long-reads to a reference genome (Liu *et al.*, 2020). The most direct approach to generate a comprehensive SV catalogue is to compare genome sequences of high-quality *de novo* assemblies. Many studies in major crops indeed have shown that a single reference genome is not sufficient to cover the genome sequences of a species. Besides, in *B. oleracea*, only two short-read assemblies, which are not of sufficient quality, were available at the moment when we initiated this research. The pan-genome study in *B. oleracea* indeed illustrated that these two reference genomes only captured part of the genome sequences of this species (Golicz *et al.*, 2016b). In addition to this, the pan-genome tool can shed novel light on the role of polyploidization in speciation and species diversification in the mesopolyploid *Brassica* species. Those are the reasons that made us to decide to generate several high-quality *B. oleracea* genome assemblies representing different morphotypes in **Chapter 3**.

We used double haploid *B. oleracea* genotypes for the *de novo* genome assembly, which avoids problems arising from heterozygous regions. In diploid genome assemblies, differences between the homologous chromosomes are usually ignored and only one consensus haplotype is used to represent the diploid genome (Campoy *et al.*, 2020). Basically, accuracy, completeness, contiguity and resolved haplotype are the factors determining the quality of a genome assembly. Our five *B. oleracea* genome assemblies meet high criteria of all these factors. In our study, we coupled long-read sequencing and long-range scaffolding technology, which is currently an efficient approach to enable high contiguity and completeness of a plant genome assembly. As we used long noisy Nanopore reads to construct the contigs, we then generated Illumina data to polish the sequences, resulting in genome assemblies with high base accuracy. For the long-range scaffolding information, we used Bionano Genomics' new DLS technology to generate optical molecules, which can produce substantially long molecules. After hybrid scaffolding with Bionano optical maps, we harvested genome sequences that include scaffolds representing chromosome arms. The five final assemblies are the most continuous and complete *B. oleracea* genomes to date (contig N50 > 11 Mb, scaffold N50 > 30Mb). The quality of these assemblies was assessed using various approaches including BUSCO (Waterhouse *et al.*, 2018), LTR Assembly Index (LAI) (Ou *et al.*, 2018), and DNA and mRNA short-reads mapping. All these results support the high-quality of the five *B. oleracea* assemblies.

High-quality reference genomes facilitate investigations of SVs and genome evolution. We identified extensive SVs (**Chapter 3**) based on these five high-quality genome assemblies, which were further used to investigate the influence of genomic features on meiotic crossover formation in **Chapter 4**. In particular, direct comparison between high-quality genome assemblies allows identification of very large SVs that cannot be detected by read mapping approaches, such as the 4.88Mb kale specific inversion identified in our study. We also investigated genome evolution patterns of *B. oleracea* (five genomes in our study and four published genomes) and *B. rapa* (18 published genomes) using a pan-genome approach (Belser *et al.*, 2018, Cai *et al.*, 2020, Cai *et al.*, 2021, Guo *et al.*, 2021), which revealed different evolutionary dynamics of LTR-RTs and WGT-derived gene loss between the two sister species (**Chapter 3**).

### Investigation of genetic regulation of GSL variation in *B. oleracea*

In **Chapter 5**, we also showcased how these high-quality assembled genomes (**Chapter 3**) could be used to investigate genetic loci regulating important traits. We particularly focused on the trait of GSLs abundance because producing varieties with enhanced beneficial GSLs composition for health-promoting nutrition is an important aim of current *B. oleracea* breeding programs. Besides, some GSLs play a role in defense against pathogens and pests; the optimal GSL composition is not yet clear. Taking advantage of the five high-quality genome assemblies (**Chapter 3**), we generated a comprehensive list of genes that are potentially involved in GSL biosynthesis, degradation and transport in *B. oleracea* based on homology to the experimentally characterized GSL genes in Arabidopsis. We also produced GSLs profiles and transcriptomic data for 20 samples to explore variation of gene expression and relative GSL levels and composition between different tissues and morphotypes. A correlation analysis between gene expression level and GSL abundance revealed a group of genes that may play a role in controlling the GSL biochemical pathways. Most of the GSL-gene combinations did not show a direct correlation between gene expression level and relative GSL abundance. This is actually not unexpected due to transporters that establish dynamic GSL patterns between source and sink tissues (Nour-Eldin *et al.*, 2012, Jensen *et al.*, 2014) and intermediate GSL products that are not at the end of a pathway. We also analysed structure of several genes in detail to explain some of the observed GLS abundance variation. These combined analyses of transcriptome, metabolome and allelic variation has the potential to increase our understanding of the genetic control of GSL profiles in *B. oleracea* and lead to breed varieties with optimal GSL composition. Nevertheless, more in-depth analyses are required to fully dissect the genetic basis underlying the accumulation of GSLs in *B. oleracea*. This is because on the one hand, the function of many GSL related genes identified in *B. oleracea* still needs to be further verified to see whether they are functionally diverged with Arabidopsis or which homologous copy is active due to the WGT event in *Brassica*. In addition, a metabolome-GWAS (mGWAS) analysis of a large panel of accessions, with a long history of recombination, can reveal candidate genes and loci involved with the accumulation of metabolites. However, we could not perform mGWAS analysis because we only have 20 samples, which are insufficient to achieve an adequate statistical power in association studies.

### Patterns and variations of meiotic crossovers and their affecting factors in *B. oleracea*

As meiotic recombination generates genetic diversity by shuffling parental chromosomes, deciphering recombination patterns, variations and their affecting factors is of great importance to improve breeding efficiency via parental selection and defining population size. In plants, although meiotic recombination is well studied in several species, such as Arabidopsis and crops like maize, tomato and potato (Demirci *et al.*, 2017, Marand *et al.*, 2017, Kianian *et al.*, 2018, Fuentes *et al.*, 2022, Lian *et al.*, 2022a), it still remains largely unexplored in species like *B. oleracea* and *B. rapa*, which display extreme phenotypic variations among morphotypes. To close this gap in *Brassica* genera, we characterized recombination landscapes for ten reciprocal crosses in *B. oleracea* in **Chapter 4**. To our knowledge, this is the most comprehensive study towards revealing intraspecific variation and sex difference of recombination rates and distributions. Besides, our data give more precision about CO distribution in *B. oleracea* genotypes, which allows further investigation of genomic features that influence recombination. Because of the inclusion of very diverse genetic backgrounds and reciprocal crosses, we revealed the very original finding that heterochiasmy is dependent on genetic background. This message is missing in most previous and recent studies in Arabidopsis and crops like maize (Bauer *et al.*, 2013, Kianian *et al.*, 2018, Blackwell *et al.*, 2020), since usually only one factor, either cross combination or cross direction, is focused on. Although the mechanism of heterochiasmy remains elusive, multiple evidences have indicated the important role of the synaptonemal complex (SC), a structure that zips homologous chromosomes together during meiosis, in causing sex difference during CO formation. In maize, Luo *et al.* observed that the length of SC is positively related to CO number in both sexes, suggesting that the sex with longer SC has more COs (Luo *et al.*, 2019). In *Arabidopsis*, Capilla-Pérez *et al.* showed that the SC is essential for CO interference and suggests that heterochiasmy is due to variation of CO interference imposed by the SC (Capilla-Pérez *et al.*, 2021). With regard to the recombination landscape, we observed an uneven U-shaped CO distribution in *B. oleracea*, with typically high CO frequency in distal regions and low frequency in centromeric and pericentromeric regions, which is consistent with those reported in other plant species (Demirci *et al.*, 2017, Marand *et al.*, 2017, Pelé *et al.*, 2017, Rommel Fuentes *et al.*, 2020, Boideau *et al.*, 2022, Fuentes *et al.*, 2022). We showed that megabase-scale recombination landscapes are similar among different genetic backgrounds and between sexes. The SV catalogue (**Chapter 3**) and fine-scale resolution CO location allowed to investigate the effect of SVs on CO formation, with large-scale SVs locally suppressing recombination. Rowan *et al.* (Rowan *et al.*, 2019) came up with five possible mechanisms explaining the underlying suppressive effects of SVs on COs: (1) reduced DSBs formation in SVs regions; (2) lack of template for recombination repair; (3) COs in SVs producing inviable gametes; (4) prevented interaction with homologous chromosome and/or the central element; (5) elevated DNA methylation in SVs. Some of these hypotheses can be further tested. For example, SPO11 complexes produce the programmed DSBs during prophase I of meiosis (Szostak *et al.*, 1983, Keeney *et al.*, 1997, de Massy, 2013). Overlap analyses between SVs and SPO11-1-oligo hotspots generated by purification and sequencing of SPO11-1-oligonucleotides could be sufficient to check whether DSBs are reduced in SVs. Similarly, DNA methylation levels in SVs and non-SVs regions can be compared using whole genome bisulfite sequencing (BS-seq).

In our study, we investigated the influence of various genomic features on local CO formation, such as genes, TEs, SNPs and SVs. Allelic variation in genes essential for CO formation can also influence CO frequency, however, the identification of causal genes for CO frequency is unexplored in *B. oleracea*. In Arabidopsis, three causal genes were identified to date, including *HEI10*, *TAF4b* and *SNI1* (Ziolkowski *et al.*, 2017, Lawrence *et al.*, 2019, Zhu *et al.*, 2021). Ziolkowski *et al.* identified *HEI10* meiotic E3 ligase gene, with natural genetic polymorphisms in this gene being associated with quantitative variation in CO frequency between Arabidopsis accessions (Ziolkowski *et al.*, 2017). They also demonstrated that *HEI10* is a limiting factor for interference-sensitive CO formation in Arabidopsis. Lawrence *et al.* identified the *TAF4b* gene, which encodes a subunit of the RNA polymerase II general transcription factor TFIID, by screening natural Arabidopsis populations for genetic modifiers for meiotic CO frequency. In the *taf4b* mutant, widespread transcriptional changes occurred, including in regulators of meiosis. They also showed that *taf4b* mutants display a genome-wide decrease of COs (Lawrence *et al.*, 2019). Zhu *et al.* identified the *SNI1* gene, which encodes a component of the SMC5/6 complex, as the causal gene underlying a major modifier locus (Zhu *et al.*, 2021). They also showed that COs are elevated in distal regions but reduced in pericentromeric regions in the *sni1* mutant, and mutations in *SNI1* result in reduced CO interference. However, these genes have never been tested in crops and the knowledge is too limited in other plant species. In each of our five *B. oleracea* assemblies, we found two copies of *HEI10*, one copy of *TAF4b* and one copy of *SNI1* gene based on sequence homology with Arabidopsis. We identified some allelic variations in each gene among the five morphotypes, including a total of ten SNPs (eight non-synonymous SNPs) in *BoHEI10.1*, four SNPs (three non-synonymous SNPs) and three InDels in *BoHEI10.2*, 23 SNPs (ten non-synonymous SNPs) and two InDels in *BoSNI1*, and 21 SNPs (17 non-synonymous SNPs) and two InDels in *BoTAF4b*. These variations may play a role in influencing meiotic CO frequency in *B. oleracea*. However, these causal genes reported in Arabidopsis may also have diverged functions in *Brassica* species. Therefore, extensive in-depth analyses, such as recombination quantitative trait loci (rQTL) mapping, CO mapping, meiotic cytology, immunostaining and functional validation, are required to better understand the mechanisms behind the CO frequency variation in *B. oleracea*.

Besides genomic factors, epigenetic features also play a role in affecting the recombination frequency and distribution, such as DNA methylation, histone modifications and nucleosome occupancy (Melamed-Bessudo and Levy, 2012, Mirouze *et al.*, 2012, Yelina *et al.*, 2012, Choi *et al.*, 2013, Habu *et al.*, 2015, Choi *et al.*, 2018). Boideau *et al.* discovered that high DNA methylation levels can explain the lack of recombination in some large non-pericentromeric regions in *B. napus* (Boideau *et al.*, 2022). Choi *et al.* found that CO incidence is significantly associated with reduced nucleosome occupancy in Arabidopsis (Choi *et al.*, 2018). Previous studies have shown that meiotic recombination is largely suppressed by some histone modifications, such as H3K27me3, H3K9me3, H3K27me1, and H3K9me2 (Aliyeva-Schnorr *et al.*, 2015, Baker *et al.*, 2015, Dreissig *et al.*, 2019). However, COs are also positively associated with H3K4me3 and histone variant H2A.Z in plant genomes (Liu *et al.*, 2009, Choi *et al.*, 2013, Drouaud *et al.*, 2013, Wijnker *et al.*, 2013, Shilo *et al.*, 2015, Choi *et al.*, 2018). Moreover, environmental conditions, such as temperature, nutritional status, pathogen attack, also influence rates and patterns of recombination (Dreissig *et al.*, 2019). Environmental factors may influence those via changing the epigenetic features. However, the effect of these epigenetic features on CO formation in *B. oleracea* was not investigated in our study due to lack of epigenetic data. To deepen our knowledge regarding how these epigenetic factors influence CO formation in *B. oleracea*, data need to be generated and analysed in detail in future studies.

Machine learning (ML) involves a series of computational approaches that aim to find predictive patterns in data (van Dijk *et al.*, 2021), which has been applied in meiotic recombination studies in plants. For example, Demirci *et al.* built models to predict where COs are likely to occur in the genome and learn about genomic features underlying CO formation in four different plant species (Demirci *et al.*, 2018). Lian *et al.* developed models to predict the occurrence of meiotic COs in Arabidopsis for a given interval with chromatin-related features and analysed how the model learned to perform the prediction (Lian *et al.*, 2022a). These applications allow us to understand genomic and epigenomic features related to CO frequency in different plants. Currently, enormous data related to meiotic recombination, including reference genomes, SVs catalogues, allelic variations in known candidate genes influencing CO formation, population resequencing reads, CO profiles, various epigenomics profiles and etc, have been released in many plant species. Using ML models to further analyse these data in detail will not only increase our understanding towards genomic and epigenomic features affecting meiotic CO formation in different plant species, but also reveal common and specific patterns across species.

### Manipulating crossovers for plant breeding

Manipulating CO formation to increase the frequency and modify the distribution is vital for improving the efficiency of plant breeding. The reason is that on the one hand, the number of COs is constrained, which limits novel allelic combinations and genetic diversity, thus affecting favourable alleles to be combined into elite varieties (Mieulet *et al.*, 2018). On the other hand, the distribution of COs is uneven, which limits the power in genetic mapping studies, as some regions tend to be entirely devoid of COs (*i.e.* centromeric and pericentromeric regions) (Fernandes *et al.*, 2018). One of the strategies to increase CO frequency in plants is to knock-out anti-CO factors (*i.e.* RECQ4, FANCM and FIGL1) (Blary and Jenczewski, 2019). In Arabidopsis, several studies have demonstrated that mutations in anti-CO genes can increase CO frequency. For example, the *recq4a* and *recq4b* mutants lead to approximately four-fold increase in CO frequency in Arabidopsis hybrids (Fernandes *et al.*, 2018, Mieulet *et al.*, 2018, Serra *et al.*, 2018b). The *fancm* mutation results in three-fold increase in CO frequency in pure lines but not in hybrids (Crismani *et al.*, 2012, Girard *et al.*, 2015, Ziolkowski *et al.*, 2015, Fernandes *et al.*, 2018). Also, the *figl1* mutation alone results in over 25% increase in CO frequency in Arabidopsis hybrids whereas the combination of *recq4* and *figl1* mutations leads to approximately eight-fold CO increase (Fernandes *et al.*, 2018, Mieulet *et al.*, 2018). Accordingly, overexpressing or increasing dosage of pro-CO factors (*i.e.* ZMM proteins) is another strategy to increase CO frequency. For example, Durand *et al.* showed that overexpression of HEI10 doubled the Class I COs in Arabidopsis (Durand *et al.*, 2022). Interestingly, combining the *zyp1* mutation, which causes the absence of SC, and overexpression of HEI10 leads to a massive and unprecedented increase in the number of Class I COs (Durand *et al.*, 2022). Besides in the model species *A. thaliana*, COs could also be increased by mutation of some genes in crops. Indeed, previous work showed that single *recq4* mutation leads to about three-fold COs increase in rice, pea and tomato (Mieulet *et al.*, 2018). However, sterility issues need to be taken into account when manipulating these genes in crops for breeding purpose, as after generating the optimal genotype, efficient seed production is a must. An example is the *figl1* mutation, which results in fully sterile plants in rice (Zhang *et al.*, 2017). Compared with CO frequency, altering the distribution by manipulating anti-CO or pro-CO genes seems less feasible (Blary and Jenczewski, 2019). Studies in Arabidopsis and rice showed that extra COs generated by anti-CO factor mutants preferentially occur in regions where wild-type COs already formed (Fernandes *et al.*, 2018, Mieulet *et al.*, 2018, Blary and Jenczewski, 2019). Unlocking pericentromeric CO formation by modifying epigenetic patterns and fine-tuning CO numbers locally via inducing DSB formation at specific sites can be potential approaches to alter CO distribution (Blary and Jenczewski, 2019). Besides, manipulating ploidy level is another potential approach to alter the U-shaped recombination landscape, which can boost CO formation in pericentromeric regions (Pelé *et al.*, 2017, Boideau *et al.*, 2021).

How to generate mutations in crops? One approach is to generate TILLING populations, which rely on chemical mutagenesis followed by DNA isolation and pooling of individuals and high-throughput screening for point mutations (McCallum *et al.*, 2000, Blary and Jenczewski, 2019). This approach can be applied in many plant species. However, it requires tremendous crossing work, especially for self-incompatibility species like *B. oleracea* and *B. rapa* (Himelblau *et al.*, 2009, Stephenson *et al.*, 2010). Also, many mutations that are not desirable will be generated. Another approach is through the targeted gene editing tool CRISPR-Cas9 that relies on transgenesis. Although this approach can avoid generation of unwanted mutations, transgenesis is not always feasible in many plant species. Besides, strict regulations with genetically modified organism (GMO) restrict the application of CRISPR-Cas9 (Bakhsh *et al.*, 2023).

## Future perspectives

### How to generate high-quality genome assemblies in plants

Currently, combining long-read sequencing technology and long-range information is the key to produce high-quality genome assemblies in plants. To simplify the assembly procedure and maximize the sequence contiguity and accuracy (Shi *et al.*, 2022), a variety of tools have been developed to generate the primary assembly (contigs) using long-reads (PacBio and Nanopore), such as Canu (Koren *et al.*, 2017), FALCON (Carvalho *et al.*, 2016), MECAT2 (Xiao *et al.*, 2017), Flye (Kolmogorov *et al.*, 2019), SMARTdenovo (Liu *et al.*, 2021b), Ra (https://github.com/rvaser/ra), Wtdbg2 (Ruan and Li, 2020), NextDenovo (https://github.com/Nextomics/NextDenovo), Miniasm (Li, 2016) and Shasta (Shafin *et al.*, 2020). Although these assemblers broadly follow the same Overlap-Layout-Consensus (OLC) paradigm, they could perform remarkably differently with regard to the required central processing unit (CPU) hours and the quality of output. Therefore, it is advisable to try several tools, among which one can select the assembly with the highest quality. In our study, we tested five assemblers, including Canu, Ra, Flye, SMARTdenovo and Wtdbg2. Canu and Ra could not produce a final assembly due to the high computational requirements. Indeed, the availability of computational resources is an important factor to consider before starting a genome assembly task. The other three assemblers are fast and do not require massive computational resources. Interestingly, SMARTdenovo generates the best results with our data, which far surpasses its successor Wtdbg2. Another wise strategy to generate a superior consensus assembly is to merge assemblies obtained from different tools, as they have complementary effects. Moreover, sequence depth and length are important factors that can affect the quality of a long-reads genome assembly. Ou *et al.* assessed this effect using PacBio datasets of maize inbred line NC358 (Ou *et al.*, 2020). They demonstrated that assemblies with ≤ 30× depth and N50 subread length of 11 kb are highly fragmented, with higher depth and longer reads resulting in more contiguous contigs. In **Chapter 3**, we showed that 35× Nanopore long-reads with N50 length of 26kb can produce highly contiguous *B. oleracea* assemblies. Tang *et al.* also used an average of 30× PacBio high-fidelity (HiFi) reads to generate contig sequences of 44 diploid potato genomes with very high contiguity (Tang *et al.*, 2022). Indeed, 30× long-reads are sufficient to generate a medium-size plant genome assembly with high accuracy and contiguity. However, for most large and complex plant genomes, long-reads alone are not sufficient to assemble a chromosome into a single contig. Long-range information, such as Bionano, chromosome conformation capture (Hi-C) reads, 10X linked reads and linkage map, can be used to order and orient the primary contigs to obtain the chromosome architecture. Also, as the primary long-reads assembled contigs are currently large in general, reference-guided approaches can also be used to arrange contigs to chromosome-level pseudomolecules (Belser *et al.*, 2018, Alonge *et al.*, 2019, Jiao and Schneeberger, 2020, Alonge *et al.*, 2022).

It has long been challenging to construct haplotype-resolved assemblies for highly heterozygous plant genomes (Shi *et al.*, 2022). Recently, several such heterozygous genomes have been successfully assembled in plants including diploids of potato (Zhou *et al.*, 2020b), apple (Sun *et al.*, 2020) and apricot (Campoy *et al.*, 2020) as well as tetraploids of potato (Bao *et al.*, 2022, Hoopes *et al.*, 2022, Sun *et al.*, 2022a), highbush blueberry and alfalfa (Colle *et al.*, 2019, Chen *et al.*, 2020b, Shen *et al.*, 2020). To enable successful assemblies of these complex genomes, many efforts have been made, such as development of advanced assemblers, use of state-of-the-art sequencing technologies and development of haplotype phasing strategies. For example, hifiasm (Cheng *et al.*, 2021) was a recently developed assembler, which aims at generating haplotype-resolved primary contig sequences using the latest PacBio HiFi reads. Unlike most other assemblers that collapse different homologous haplotypes into a single consensus representation or produce all haplotypes but only aim to maintain the contiguity of one haplotype, hifiasm strives to preserve the contiguity of all haplotypes (Cheng *et al.*, 2021). Other efforts made on haplotype phasing to assist assembly of heterozygous genomes include separation of sequencing reads into haplotype-specific read sets before assembly based on the genomic differences between parental genomes (trio binning) (Koren *et al.*, 2018) or on genetic mapping information derived from single-cell sequencing of gamete genomes (Shi *et al.*, 2019, Campoy *et al.*, 2020, Li *et al.*, 2020b) or Illumina sequencing of progeny genomes (Zhou *et al.*, 2020a, Zhou *et al.*, 2020b). Also, high throughput/resolution Hi-C technology can help to resolve haplotypes during or before genome assembly (Zhang *et al.*, 2018, Linsmith *et al.*, 2019, Zhang *et al.*, 2019b, Chen *et al.*, 2020b, Garg *et al.*, 2021). With the rapid advances of these sequencing technologies and assembling strategies, we would expect a surge of heterozygous genomes with different ploidy levels to be deciphered in the coming future. In **Chapter 2**, we observed high genetic diversity in genebank accessions that are heterogeneous populations of highly heterozygous *B. oleracea* plants. To date, there is still no haplotype-resolved genome assembly for such heterozygous *B. oleracea* plants. It is therefore needed to generate haplotype-resolved reference genomes for these plants, which will greatly facilitate genomic-assisted breeding strategies in *B. oleracea*. Indeed, haplotype-resolved genomes allow comprehensive dissection of their genome organization and haplotype divergence, thus potentially guiding the design of optimal haplotypes by avoiding combination of known incompatibility alleles (Sun *et al.*, 2022a).

It has also long been difficult to assemble some complex regions in plant genomes, with the vast majority of published ones still containing gaps to date. These gaps are mainly attributed to long arrays of DNA simple repeats, ribosomal DNAs, tandem duplications or complex transposable elements (Navrátilová *et al.*, 2022, Shi *et al.*, 2022). Nevertheless, a series of Telomere-to-Telomere (T2T) or gapless genomes are recently reported in plants (Belser *et al.*, 2021, Song *et al.*, 2021a, Deng *et al.*, 2022, Yang *et al.*, 2022), thanks to the rapid advances in sequencing technologies, assembling strategies and gap closure strategies. For instance, Song *et al.* took advantage of complementary effects of different latest sequencing technologies (PacBio Hifi and CLR reads, Bionano optical mapping) and seven different assembling tools to successfully produce two T2T rice genomes (Song *et al.*, 2021a). Yang *et al.* constructed the potato T2T genome (DM8.1) by combining ONT ultra-long reads, Hi-C sequencing and multiple gap closing strategies including targeted sequencing of gap regions using Hifi technology (Yang *et al.*, 2022). Indeed, difficult-to-assemble regions can be enriched by amplification experiments. Subsequently, DNA from these regions can be sequenced and assembled, resulting in sequences that can be used as patches to close the gaps (Shi *et al.*, 2022). T2T genomes allow decoding of the genomic architecture and biological function of complex regions that were previously inaccessible. Besides, there is no need to regularly upgrade T2T genomes. Therefore, T2T genomes are expected to become new references for fundamental studies and breeding applications in the future (Shi *et al.*, 2022). Although it is still not easy and there is no standard pipeline to construct T2T plant genomes, I suggest more efforts to be made to generate T2T assemblies for future genome sequencing projects.

**The surging pan-genomes in plants**

The concept of pan-genome is first mentioned in bacteria in 2005 (Tettelin *et al.*, 2005), which represents a collection of all the DNA sequences that occur in a species (Sherman and Salzberg, 2020). A pan-genome is made up of core genome and dispensable genome, which contain DNA sequences shared between all individuals and between only a subset of individuals of a species, respectively (Sherman and Salzberg, 2020). Over the past few years, pan-genome studies have surged in plants because of dramatic improvements in sequencing technologies and better understanding of genomic variations. The first plant pan-genome was reported in soybean, which revealed novel genes in seven wild *Glycine soja* accessions but not in the domesticated *Glycine max* (Li *et al.*, 2014b). Subsequently, the pan-genome study has been performed in many other plant species, including the model species

*Arabidopsis* and major crops, rice, maize, barley, wheat, cucumber, tomato, potato, citrus, rapeseed, *B. rapa* and *B. oleracea* (Golicz *et al.*, 2016b, Gao *et al.*, 2019, Jayakodi *et al.*, 2020, Jiao and Schneeberger, 2020, Liu *et al.*, 2020, Song *et al.*, 2020, Cai *et al.*, 2021, Hufford *et al.*, 2021, Li *et al.*, 2021, Qin *et al.*, 2021, Hoopes *et al.*, 2022, Li *et al.*, 2022). One of the major findings of these studies is the unexpectedly extensive intraspecific genomic variations within a species, which highlight the insufficiency of single reference genomes to represent genome sequences of a species.

There are three general approaches to construct a pan-genome: 1) iterative mapping and assembly, 2) sequence comparison of high-quality *de novo* assembled genomes, 3) graph-based pan-genome (Bayer *et al.*, 2020, Shi *et al.*, 2022). For the iterative mapping and assembly approach, genomic reads from a panel of individuals are aligned to a reference genome, followed by assembling unaligned reads into novel contigs and linearly appending novel contigs to the backbone genome sequences. One advantage of this approach is that hundreds or even thousands of individuals can be extended to the constructed pan-genome, which allows representation of a more complete genomic diversity in a species and identification of rare genes. However, genomic locations of newly assembled contigs remain unknown in the given individual without further analysis (Della Coletta *et al.*, 2021). Moreover, this approach is weak in characterising large SVs or assembling highly repetitive sequences, especially when using short-reads in a project that aims at including individuals from a large population (Danilevicz *et al.*, 2020). In comparison, pan-genomes constructed via *de novo* genome assembly can characterise large SVs at megabase level and provide important physical position information of genes and other genomic components for each individual. Indeed, we identified eight large inversions with sizes ranging from 134Kb to 4.88Mb in our *B. oleracea* pan-genome (**Chapter 3**), which are almost impossible to be detected in the previous *B. oleracea* pan-genome (Golicz *et al.*, 2016b). However, this approach is very costly when including a large number of individuals and really depends on the quality of the genome assembly. Additional evidence may be required to distinguish real genomic differences from assembly errors or from artifacts derived from different assembly methods. Storing genomic information in a graph seems to be the most promising approach to represent a pan-genome and is now becoming increasingly popular in plants. In this approach, the genomic variation information of dispensable regions are stored along the linear reference genome as alternative paths through a graph. Unlike linearly stacked pan-genomes, the graph-based pan-genome contains multiple haplotypes for each SV locus (Garrison *et al.*, 2018, Sherman and Salzberg, 2020, Shi *et al.*, 2022), which can improve the accuracy of read mapping and variant calling.

Graph-based pan-genomes also allow continuous optimization by incorporating more newly identified SVs (Shi *et al.*, 2022, Wang *et al.*, 2022, Zhou *et al.*, 2022). However, this approach requires massive computational resources. Several graph-based pan-genome construction tools are under active development, such as Vg (Garrison *et al.*, 2018), SevenBridges (Rakocevic *et al.*, 2019), GraphAligner (Rautiainen *et al.*, 2019), minigraph (Li *et al.*, 2020a) and PanTools (Sheikhizadeh *et al.*, 2016, Jonkheer *et al.*, 2022). Moreover, to standardize plant pan-genomic analyses, Shi *et al.* proposed a three-step pipeline: 1) construction of non-redundant pan-genome sequences and pan-genes, 2) deep annotations of variable and core sequences/genes by applying multi-omics analysis, 3) SVs identification and construction of graph pan-genomes using ultra-high quality backbone genomes (Shi *et al.*, 2022). We believe that these efforts will greatly promote plant pan-genomics studies and thus facilitate breeding.

Plant pan-genomes can advance QTL mapping, GWAS, population genomics and domestication studies (Della Coletta *et al.*, 2021), providing broad insights into fundamental researches and breeding programs. Compared to single references, pan-genome references improve the accuracy of short-reads mapping, resulting in higher quality variant calls that are critically important to the success of QTL mapping and GWAS analysis (Bayer *et al.*, 2020). Besides, as a causal gene may not be present in the given single reference genome, using a pan-genome reference can avoid these type of biases. This has been illustrated in both maize and potato. The gene conferring resistance to sugarcane mosaic virus could be identified using GWAS analysis based on the B73 reference genome while cannot be identified based on the PH207 reference genome (Gage *et al.*, 2019, Della Coletta *et al.*, 2021). This is because the gene is absent in PH207. Similarly, the *StOFP20* gene, which is associated with regulating tuber shape, is found to be absent in DM reference genome in potato but present in M6 reference genome (Wu *et al.*, 2018, van Eck *et al.*, 2022). Moreover, previous phenotype-genotype association studies have almost exclusively relied on SNPs or small InDels, with the contribution of SVs to trait variation being largely unexplored. The availability of high-quality pan-genome sequences enables accurate detection of SVs and thus benefits association studies between SVs and agronomically important traits. In *B. napus*, Song *et al.* performed GWAS using presence and absence variations (PAVs) from eight high-quality *de novo* genome assemblies and identified causal SVs for silique length, seed weight, and flowering time that were not detected by SNP-GWAS (Song *et al.*, 2020). In this thesis, we revealed extensive genomic variations (**Chapter 3**) using the new *B. oleracea* pan-genome, especially for the dispensable genes and large SVs, which are fundamentally important to guide *B. oleracea* breeding. We also investigated diversification mechanisms (**Chapter 3**) by focusing on studying gene fractionation patterns both before and during intraspecific diversification of *B. rapa* and *B. oleracea* using a pan-genome approach. This however cannot be done without the high-quality *de novo* assembled sequences of many accessions of the two species, as the complete gene repertoire for each extant individual accession needs to be captured. In the coming future, we expect that pan-genomes will become new references for comparative genomics studies and plant breeding programs, especially the graph-based pan-genome in combination with T2T backbone sequences.

**Concluding remarks**

In this thesis, I first investigated genetic diversity, genealogical relationship and domestication history of *B. oleracea* morphotypes using SBG data. I provided evidence for two domestication lineages, which support a middle-eastern origin for *B. oleracea* crops from diversified kale populations. Then, I generated high-quality genome assemblies for five different *B. oleracea* morphotypes by combining long-read sequencing technologies and long-range scaffolding information. These genomes allowed identification of extensive intraspecific SVs and comparison of genomic evolution patterns between *B. oleracea* and *B. rapa* (published data) using a pan-genome approach. Next, I characterized fine-scale resolution recombination landscapes for ten reciprocal crosses deriving from different parental combinations of these five *B. oleracea* morphotypes. I revealed genomic features that affect CO formation and highlighted extensive variations for CO number relying on striking interdependency of genetic background and sex of meiosis. Finally, I investigated GSL profile variations between different *B. oleracea* morphotypes and tissues, and provided insights into understanding genetic regulation of GSL profile variations, taking advantage of the five high-quality genome assemblies. These studies comprehensively dissect genomic architecture of highly diverse *B. oleracea* species and provide valuable genomic resources for *B. oleracea* improvement. As constructing a T2T plant genome is becoming possible, I suggest future *B. oleracea* sequencing projects also focus on making complete genome sequences, which will remarkably benefit fundamental studies and breeding applications in *Brassica*.

# References

## References

**Agerbirk, N. and Olsen, C.E.** (2012) Glucosinolate structures in evolution. *Phytochemistry*, **77**, 16-45.

**Akama, S., Shimizu-Inatsugi, R., Shimizu, K.K. and Sese, J.** (2014) Genome-wide quantification of homeolog expression ratio revealed nonstochastic gene regulation in synthetic allopolyploid Arabidopsis. *Nucleic acids research*, **42**, e46-e46.

**Albalat, R. and Cañestro, C.** (2016) Evolution by gene loss. *Nature Reviews Genetics*, **17**, 379-391.

**Alhajturki, D., Muralidharan, S., Nurmi, M., Rowan, B.A., Lunn, J.E., Boldt, H., Salem, M.A., Alseekh, S., Jorzig, C. and Feil, R.** (2018) Dose-dependent interactions between two loci trigger altered shoot growth in BG-5× Krotzenburg-0 (Kro-0) hybrids of Arabidopsis thaliana. *New Phytologist*, **217**, 392-406.

**Aliyeva-Schnorr, L., Beier, S., Karafiátová, M., Schmutzer, T., Scholz, U., Doležel, J., Stein, N. and Houben, A.** (2015) Cytogenetic mapping with centromeric bacterial artificial chromosomes contigs shows that this recombination-poor region comprises more than half of barley chromosome 3 H. *The Plant Journal*, **84**, 385-394.

**Allen, G.C., Flores-Vergara, M., Krasynanski, S., Kumar, S. and Thompson, W.** (2006) A modified protocol for rapid DNA isolation from plant tissues using cetyltrimethylammonium bromide. *Nature protocols*, **1**, 2320-2325.

**Alonge, M., Lebeigle, L., Kirsche, M., Jenike, K., Ou, S., Aganezov, S., Wang, X., Lippman, Z.B., Schatz, M.C. and Soyk, S.** (2022) Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biology*, **23**, 1-19.

**Alonge, M., Soyk, S., Ramakrishnan, S., Wang, X., Goodwin, S., Sedlazeck, F.J., Lippman, Z.B. and Schatz, M.C.** (2019) RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome biology*, **20**, 1-17.

**Amarasinghe, S.L., Su, S., Dong, X., Zappia, L., Ritchie, M.E. and Gouil, Q.** (2020) Opportunities and challenges in long-read sequencing data analysis. *Genome biology*, **21**, 1-16.

**Anderson, L.K., Lai, A., Stack, S.M., Rizzon, C. and Gaut, B.S.** (2006) Uneven distribution of expressed sequence tag loci on maize pachytene chromosomes. *Genome research*, **16**, 115-122.

**Arabidopsis Genome Initiative** (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *nature*, **408**, 796-815.

**Aravind, L., Watanabe, H., Lipman, D.J. and Koonin, E.V.** (2000) Lineage-specific loss and divergence of functionally linked genes in eukaryotes. *Proceedings of the National Academy of Sciences*, **97**, 11319-11324.

**Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S. and Eppig, J.T.** (2000) Gene ontology: tool for the unification of biology. *Nature genetics*, **25**, 25-29.

**Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W. and Noble, W.S.** (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic acids research*, **37**, W202-W208.

**Baker, K., Dhillon, T., Colas, I., Cook, N., Milne, I., Milne, L., Bayer, M. and Flavell, A.J.** (2015) Chromatin state analysis of the barley epigenome reveals a higher-order structure defined by H3K27me1 and H3K27me3 abundance. *The Plant Journal*, **84**, 111-124.

**Bakhsh, A., Zainab, R., Ali, M.A., Chung, G., Golokhvast, K.S. and Nawaz, M.A.** (2023) Genetically modified organisms in Europe: state of affairs, birth, research, and the regulatory process (es). In *GMOs and Political Stance*: Elsevier, pp. 165-172.

**Bao, Z., Li, C., Li, G., Wang, P., Peng, Z., Cheng, L., Li, H., Zhang, Z., Li, Y. and Huang, W.** (2022) Genome architecture and tetrasomic inheritance of autotetraploid potato. *Molecular Plant*, **15**, 1211-1226.

**Bauer, E., Falque, M., Walter, H., Bauland, C., Camisan, C., Campo, L., Meyer, N., Ranc, N., Rincent, R. and Schipprack, W.** (2013) Intraspecific variation of recombination rate in maize. *Genome biology*, **14**, 1-17.

**Bayer, P.E., Golicz, A.A., Scheben, A., Batley, J. and Edwards, D.** (2020) Plant pan-genomes are the new reference. *Nature plants*, **6**, 914-920.

**Bayer, P.E., Ruperao, P., Mason, A.S., Stiller, J., Chan, C.-K.K., Hayashi, S., Long, Y., Meng, J., Sutton, T. and Visendi, P.** (2015) High-resolution skim genotyping by sequencing reveals the distribution of crossovers and gene conversions in Cicer arietinum and Brassica napus. *Theoretical and Applied Genetics*, **128**, 1039-1047.

**Beekwilder, J., Van Leeuwen, W., Van Dam, N.M., Bertossi, M., Grandi, V., Mizzi, L., Soloviev, M., Szabados, L., Molthoff, J.W. and Schipper, B.** (2008) The impact of the absence of aliphatic glucosinolates on insect herbivory in Arabidopsis. *PLoS One*, **3**, e2068.

**Begun, D.J. and Aquadro, C.F.** (1992) Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. *Nature*, **356**, 519-520.

**Beilstein, M.A., Nagalingum, N.S., Clements, M.D., Manchester, S.R. and Mathews, S.** (2010) Dated molecular phylogenies indicate a Miocene origin for Arabidopsis thaliana. *Proceedings of the National Academy of Sciences*, **107**, 18724-18728.

**Bell, L., Oloyede, O.O., Lignou, S., Wagstaff, C. and Methven, L.** (2018) Taste and flavor perceptions of glucosinolates, isothiocyanates, and related compounds. *Molecular Nutrition & Food Research*, **62**, 1700990.

**Belser, C., Baurens, F.-C., Noel, B., Martin, G., Cruaud, C., Istace, B., Yahiaoui, N., Labadie, K., Hřibová, E. and Doležel, J.** (2021) Telomere-to-telomere gapless chromosomes of banana using nanopore sequencing. *Communications biology*, **4**, 1-12.

**Belser, C., Istace, B., Denis, E., Dubarry, M., Baurens, F.-C., Falentin, C., Genete, M., Berrabah, W., Chèvre, A.-M. and Delourme, R.** (2018) Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. *Nature plants*, **4**, 879-887.

**Belyayev, A.** (2014) Bursts of transposable elements as an evolutionary driving force. *Journal of evolutionary biology*, **27**, 2573-2584.

**Benderoth, M., Pfalz, M. and Kroymann, J.** (2009) Methylthioalkylmalate synthases: genetics, ecology and evolution. *Phytochemistry Reviews*, **8**, 255-268.

**Berchowitz, L.E. and Copenhaver, G.P.** (2008) Fluorescent Arabidopsis tetrads: a visual assay for quickly developing large crossover and crossover interference data sets. *Nature protocols*, **3**, 41-50.

**Berger, D., Soles, J.S., Giumlia-Mair, A.R., Brügmann, G., Galili, E., Lockhoff, N. and Pernicka, E.** (2019) Isotope systematics and chemical composition of tin ingots from Mochlos (Crete) and other Late Bronze Age sites in the eastern Mediterranean Sea: An ultimate key to tin provenance? *PloS one*, **14**, e0218326.

**Bhandari, S.R., Rhee, J., Choi, C.S., Jo, J.S., Shin, Y.K. and Lee, J.G.** (2020) Profiling of individual desulfo-glucosinolate content in cabbage head (Brassica oleracea var. capitata) germplasm. *Molecules*, **25**, 1860.

**Bird, K.A., An, H., Gazave, E., Gore, M.A., Pires, J.C., Robertson, L.D. and Labate, J.A.** (2017) Population structure and phylogenetic relationships in a diverse panel of Brassica rapa L. *Frontiers in plant science*, **8**, 321.

**Blackwell, A.R., Dluzewska, J., Szymanska-Lejman, M., Desjardins, S., Tock, A.J., Kbiri, N., Lambing, C., Lawrence, E.J., Bieluszewski, T. and Rowan, B.** (2020) MSH 2 shapes the meiotic crossover landscape in relation to interhomolog polymorphism in Arabidopsis. *The EMBO journal*, **39**, e104858.

**Blary, A. and Jenczewski, E.** (2019) Manipulation of crossover frequency and distribution for plant breeding. *Theoretical and Applied Genetics*, **132**, 575-592.

**Boideau, F., Pelé, A., Tanguy, C., Trotoux, G., Eber, F., Maillet, L., Gilet, M., Lodé-Taburel, M., Huteau, V. and Morice, J.** (2021) A Modified Meiotic Recombination in Brassica napus Largely Improves Its Breeding Efficiency. *Biology*, **10**, 771.

**Boideau, F., Richard, G., Coriton, O., Huteau, V., Belser, C., Deniot, G., Eber, F., Falentin, C., Ferreira de Carvalho, J. and Gilet, M.** (2022) Epigenomic and structural events preclude recombination in Brassica napus. *New Phytologist*, **234**, 545-559.

**Bonnema, G., Del Carpio, D.P. and Zhao, J.J.** (2011) Diversity analysis and molecular taxonomy of Brassica vegetable crops. *Genetics, Genomics and Breeding of Vegetable Brassicas*, 81-124.

**Bonnema, G., Lee, J.G., Shuhang, W., Lagarrigue, D., Bucher, J., Wehrens, R., De Vos, R. and Beekwilder, J.** (2019) Glucosinolate variability between turnip organs during development. *PloS one*, **14**, e0217862.

**Bothmer, R.v., Gustafsson, M. and Snogerup, S.** (1995) Brassica sect. Brassica (Brassicaceae). II. Inter-and intraspecific crosses with cultivars of B. oleracea.

**Botstein, D., White, R.L., Skolnick, M. and Davis, R.W.** (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American journal of human genetics*, **32**, 314.

**Bruce, T.** (2014) Glucosinolates in oilseed rape: secondary metabolites that influence interactions with herbivores and their natural enemies. *Annals of Applied Biology*, **164**, 348-353.

**Buggs, R.J., Chamala, S., Wu, W., Gao, L., May, G.D., Schnable, P.S., Soltis, D.E., Soltis, P.S. and Barbazuk, W.B.** (2010) Characterization of duplicate gene evolution in the recent natural allopolyploid Tragopogon miscellus by next-generation sequencing and Sequenom iPLEX MassARRAY genotyping. *Molecular ecology*, **19**, 132-146.

**Cai, C., Bucher, J., Bakker, F.T. and Bonnema, G.** (2022a) Evidence for two domestication lineages supporting a middle-eastern origin for Brassica oleracea crops from diversified kale populations. *Horticulture research*, **9**.

**Cai, C., Bucher, J., Finkers, R. and Bonnema, G.** (2022b) Chromosome-scale genome assemblies of five different Brassica oleracea morphotypes provide insights in intraspecific diversification. *bioRxiv*.

**Cai, X., Chang, L., Zhang, T., Chen, H., Zhang, L., Lin, R., Liang, J., Wu, J., Freeling, M. and Wang, X.** (2021) Impacts of allopolyploidization and structural variation on intraspecific diversification in Brassica rapa. *Genome biology*, **22**, 1-24.

**Cai, X., Lin, R., Liang, J., King, G.J., Wu, J. and Wang, X.** (2022c) Transposable element insertion: a hidden major source of domesticated phenotypic variation in Brassica rapa. *Plant Biotechnology Journal*.

**Cai, X., Wu, J., Liang, J., Lin, R., Zhang, K., Cheng, F. and Wang, X.** (2020) Improved Brassica oleracea JZS assembly reveals significant changing of LTR-RT dynamics in different morphotypes. *Theoretical and Applied Genetics*, **133**, 3187-3199.

**Campoy, J.A., Sun, H., Goel, M., Jiao, W.-B., Folz-Donahue, K., Wang, N., Rubio, M., Liu, C., Kukat, C. and Ruiz, D.** (2020) Gamete binning: chromosome-level and haplotype-resolved genome assembly enabled by high-throughput single-cell sequencing of gamete genomes. *Genome biology*, **21**, 1-20.

**Capilla-Pérez, L., Durand, S., Hurel, A., Lian, Q., Chambon, A., Taochy, C., Solier, V., Grelon, M. and Mercier, R.** (2021) The synaptonemal complex imposes crossover interference and heterochiasmy in Arabidopsis. *Proceedings of the National Academy of Sciences*, **118**.

**Carvalho, A.B., Dupim, E.G. and Goldstein, G.** (2016) Improved assembly of noisy long reads by k-mer validation. *Genome research*, **26**, 1710-1720.

**Chalhoub, B., Denoeud, F., Liu, S., Parkin, I.A., Tang, H., Wang, X., Chiquet, J., Belcram, H., Tong, C. and Samans, B.** (2014) Early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome. *science*, **345**, 950-953.

**Chang, N., Sun, Q., Hu, J., An, C. and Gao, H.** (2017) Large introns of 5 to 10 kilo base pairs can be spliced out in Arabidopsis. *Genes*, **8**, 200.

**Chen, C., Chen, H., Zhang, Y., Thomas, H.R., Frank, M.H., He, Y. and Xia, R.** (2020a) TBtools: an integrative toolkit developed for interactive analyses of big biological data. *Molecular plant*, **13**, 1194-1202.

**Chen, D.-H. and Ronald, P.** (1999) A rapid DNA minipreparation method suitable for AFLP and other PCR applications. *Plant Molecular Biology Reporter*, **17**, 53-57.

**Chen, H., Song, X., Shang, Q., Feng, S. and Ge, W.** (2022) CFVisual: an interactive desktop platform for drawing gene structure and protein architecture. *BMC bioinformatics*, **23**, 1-8.

**Chen, H., Zeng, Y., Yang, Y., Huang, L., Tang, B., Zhang, H., Hao, F., Liu, W., Li, Y. and Liu, Y.** (2020b) Allele-aware chromosome-level genome assembly and efficient transgene-free genome editing for the autotetraploid cultivated alfalfa. *Nature communications*, **11**, 1-11.

**Chen, S., Zhou, Y., Chen, Y. and Gu, J.** (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, **34**, i884-i890.

**Cheng, F., Liang, J., Cai, C., Cai, X., Wu, J. and Wang, X.** (2017) Genome sequencing supports a multi-vertex model for Brassiceae species. *Current opinion in plant biology*, **36**, 79-87.

**Cheng, F., Mandáková, T., Wu, J., Xie, Q., Lysak, M.A. and Wang, X.** (2013) Deciphering the diploid ancestral genome of the mesohexaploid Brassica rapa. *The Plant Cell*, **25**, 1541-1554.

**Cheng, F., Sun, C., Wu, J., Schnable, J., Woodhouse, M.R., Liang, J., Cai, C., Freeling, M. and Wang, X.** (2016a) Epigenetic regulation of subgenome dominance following whole genome triplication in Brassica rapa. *New Phytol*, **211**, 288-299.

**Cheng, F., Sun, R., Hou, X., Zheng, H., Zhang, F., Zhang, Y., Liu, B., Liang, J., Zhuang, M., Liu, Y., Liu, D., Wang, X., Li, P., Liu, Y., Lin, K., Bucher, J., Zhang, N., Wang, Y., Wang, H., Deng, J., Liao, Y., Wei, K., Zhang, X., Fu, L., Hu, Y., Liu, J., Cai, C., Zhang, S., Zhang, S., Li, F., Zhang, H., Zhang, J., Guo, N., Liu, Z., Liu, J., Sun, C., Ma, Y., Zhang, H., Cui, Y., Freeling, M.R., Borm, T., Bonnema, G., Wu, J. and Wang, X.** (2016b) Subgenome parallel selection is associated with morphotype diversification and convergent crop domestication in Brassica rapa and Brassica oleracea. *Nat Genet*, **48**, 1218-1224.

**Cheng, F., Wu, J., Cai, C., Fu, L., Liang, J., Borm, T., Zhuang, M., Zhang, Y., Zhang, F. and Bonnema, G.** (2016c) Genome resequencing and comparative variome analysis in a Brassica rapa and Brassica oleracea collection. *Scientific data*, **3**, 1-9.

**Cheng, F., Wu, J., Fang, L., Sun, S., Liu, B., Lin, K., Bonnema, G. and Wang, X.** (2012a) Biased gene fractionation and dominant gene expression among the subgenomes of Brassica rapa. *PloS one*, **7**, e36442.

**Cheng, F., Wu, J., Fang, L. and Wang, X.** (2012b) Syntenic gene analysis between Brassica rapa and other Brassicaceae species. *Frontiers in plant science*, **3**, 198.

**Cheng, F., Wu, J. and Wang, X.** (2014) Genome triplication drove the diversification of Brassica plants. *Horticulture research*, **1**.

**Cheng, H., Concepcion, G.T., Feng, X., Zhang, H. and Li, H.** (2021) Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature methods*, **18**, 170-175.

**Chifman, J. and Kubatko, L.** (2014) Quartet inference from SNP data under the coalescent model. *Bioinformatics*, **30**, 3317-3324.

**Choi, K., Zhao, X., Kelly, K.A., Venn, O., Higgins, J.D., Yelina, N.E., Hardcastle, T.J., Ziolkowski, P.A., Copenhaver, G.P. and Franklin, F.C.H.** (2013) Arabidopsis meiotic crossover hot spots overlap with H2A. Z nucleosomes at gene promoters. *Nature genetics*, **45**, 1327-1336.

**Choi, K., Zhao, X., Tock, A.J., Lambing, C., Underwood, C.J., Hardcastle, T.J., Serra, H., Kim, J., Cho, H.S. and Kim, J.** (2018) Nucleosomes and DNA methylation shape meiotic DSB frequency in Arabidopsis thaliana transposons and gene regulatory regions. *Genome research*, **28**, 532-546.

**Choulet, F., Alberti, A., Theil, S., Glover, N., Barbe, V., Daron, J., Pingault, L., Sourdille, P., Couloux, A. and Paux, E.** (2014) Structural and functional partitioning of bread wheat chromosome 3B. *Science*, **345**, 1249721.

**Chuong, E.B., Elde, N.C. and Feschotte, C.** (2017) Regulatory activities of transposable elements: from conflicts to benefits. *Nature Reviews Genetics*, **18**, 71-86.

**Colle, M., Leisner, C.P., Wai, C.M., Ou, S., Bird, K.A., Wang, J., Wisecaver, J.H., Yocca, A.E., Alger, E.I. and Tang, H.** (2019) Haplotype-phased genome and evolution of phytonutrient pathways of tetraploid blueberry. *GigaScience*, **8**, giz012.

**Compeau, P.E., Pevzner, P.A. and Tesler, G.** (2011) How to apply de Bruijn graphs to genome assembly. *Nature biotechnology*, **29**, 987-991.

**Consortium, U.** (2015) UniProt: a hub for protein information. *Nucleic acids research*, **43**, D204-D212.

**Coop, G. and Przeworski, M.** (2007) An evolutionary view of human recombination. *Nature Reviews Genetics*, **8**, 23-34.

**Couvreur, T.L., Franzke, A., Al-Shehbaz, I.A., Bakker, F.T., Koch, M.A. and Mummenhoff, K.** (2010) Molecular phylogenetics, temporal diversification, and principles of evolution in the mustard family (Brassicaceae). *Molecular Biology and Evolution*, **27**, 55-71.

**Crismani, W., Girard, C., Froger, N., Pradillo, M., Santos, J.L., Chelysheva, L., Copenhaver, G.P., Horlow, C. and Mercier, R.** (2012) FANCM limits meiotic crossovers. *Science*, **336**, 1588-1590.

**Crozier, A.A.** (1891) *The Cauliflower*: Register Publishing Company.

**Cutter, A.D. and Payseur, B.A.** (2013) Genomic signatures of selection at linked sites: unifying the disparity among species. *Nature Reviews Genetics*, **14**, 262-274.

**Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T. and Sherry, S.T.** (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156-2158.

**Danilevicz, M.F., Fernandez, C.G.T., Marsh, J.I., Bayer, P.E. and Edwards, D.** (2020) Plant pangenomics: approaches, applications and advancements. *Current Opinion in Plant Biology*, **54**, 18-25.

**de Massy, B.** (2013) Initiation of meiotic recombination: how and where? Conservation and specificities among eukaryotes. *Annu Rev Genet*, **47**, 563-599.

**De Muyt, A., Mercier, R., Mezard, C. and Grelon, M.** (2009) Meiotic recombination and crossovers in plants. *Meiosis*, **5**, 14-25.

**De Smet, R., Adams, K.L., Vandepoele, K., Van Montagu, M.C., Maere, S. and Van de Peer, Y.** (2013) Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proceedings of the National Academy of Sciences*, **110**, 2898-2903.

**Del Carpio, D.P., Basnet, R.K., De Vos, R.C., Maliepaard, C., Visser, R. and Bonnema, G.** (2011) The patterns of population differentiation in a Brassica rapa core collection. *Theoretical and applied genetics*, **122**, 1105-1118.

**Della Coletta, R., Qiu, Y., Ou, S., Hufford, M.B. and Hirsch, C.N.** (2021) How the pan-genome is changing crop genomics and improvement. *Genome biology*, **22**, 1-19.

**Demirci, S.** (2021) A computational study of genomic rearrangements in plants: Wageningen University and Research.

**Demirci, S., Peters, S.A., de Ridder, D. and van Dijk, A.D.** (2018) DNA sequence and shape are predictive for meiotic crossovers throughout the plant kingdom. *The Plant Journal*, **95**, 686-699.

**Demirci, S., van Dijk, A.D., Sanchez Perez, G., Aflitos, S.A., de Ridder, D. and Peters, S.A.** (2017) Distribution, position and genomic characteristics of crossovers in tomato recombinant inbred lines derived from an interspecific cross between Solanum lycopersicum and Solanum pimpinellifolium. *The Plant Journal*, **89**, 554-564.

**Deng, Y., Liu, S., Zhang, Y., Tan, J., Li, X., Chu, X., Xu, B., Tian, Y., Sun, Y. and Li, B.** (2022) A telomere-to-telomere gap-free reference genome of watermelon and its mutation library provide important resources for gene discovery and breeding. *Molecular plant*, **15**, 1268-1284.

**Deschamps, S., Zhang, Y., Llaca, V., Ye, L., Sanyal, A., King, M., May, G. and Lin, H.** (2018) A chromosome-scale assembly of the sorghum genome using nanopore sequencing and optical mapping. *Nature Communications*, **9**, 1-10.

**Devos, K.M., Brown, J.K. and Bennetzen, J.L.** (2002) Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. *Genome research*, **12**, 1075-1079.

**Dias, J.S.** (2012) Portuguese perennial kale: a relic leafy vegetable crop. *Genetic resources and crop evolution*, **59**, 1201-1206.

**Dluzewska, J., Szymanska, M. and Ziolkowski, P.A.** (2018) Where to cross over? Defining crossover sites in plants. *Frontiers in genetics*, **9**, 609.

**Domansky, A.N., Kopantzev, E.P., Snezhkov, E.V., Lebedev, Y.B., Leib-Mosch, C. and Sverdlov, E.D.** (2000) Solitary HERV-K LTRs possess bi-directional promoter activity and contain a negative regulatory element in the U5 region. *FEBS letters*, **472**, 191-195.

**Dooner, H.K. and He, L.** (2008) Maize genome structure variation: interplay between retrotransposon polymorphisms and genic recombination. *The Plant Cell*, **20**, 249-258.

**Dréau, A., Venu, V., Avdievich, E., Gaspar, L. and Jones, F.C.** (2019) Genome-wide recombination map construction from single individuals using linked-read sequencing. *Nature Communications*, **10**, 1-11.

**Dreissig, S., Mascher, M. and Heckmann, S.** (2019) Variation in recombination rate is shaped by domestication and environmental conditions in barley. *Molecular biology and evolution*, **36**, 2029-2039.

**Dreissig, S., Maurer, A., Sharma, R., Milne, L., Flavell, A.J., Schmutzer, T. and Pillen, K.** (2020) Natural variation in meiotic recombination rate shapes introgression patterns in intraspecific hybrids between wild and domesticated barley. *New Phytologist*, **228**, 1852-1863.

**Drouaud, J., Khademian, H., Giraut, L., Zanni, V., Bellalou, S., Henderson, I.R., Falque, M. and Mézard, C.** (2013) Contrasted patterns of crossover and non-crossover at Arabidopsis thaliana meiotic recombination hotspots. *PLoS genetics*, **9**, e1003922.

**Drouaud, J., Mercier, R., Chelysheva, L., Bérard, A., Falque, M., Martin, O., Zanni, V., Brunel, D. and Mézard, C.** (2007) Sex-specific crossover distributions and variations in interference level along Arabidopsis thaliana chromosome 4. *PLoS genetics*, **3**, e106.

**Durand, S., Lian, Q., Jing, J., Ernst, M., Grelon, M., Zwicker, D. and Mercier, R.** (2022) Joint control of meiotic crossover patterning by the synaptonemal complex and HEI10 dosage. *Nature communications*, **13**, 1-13.

**El-Esawi, M.A., Germaine, K., Bourke, P. and Malone, R.** (2016) Genetic diversity and population structure of Brassica oleracea germplasm in Ireland using SSR markers. *C R Biol*, **339**, 133-140.

**Ellegren, H. and Galtier, N.** (2016) Determinants of genetic diversity. *Nature Reviews Genetics*, **17**, 422-433.

**Ellinghaus, D., Kurtz, S. and Willhoeft, U.** (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC bioinformatics*, **9**, 1-14.

**Emms, D.M. and Kelly, S.** (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome biology*, **20**, 1-14.

**English, A.C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., Qin, X., Muzny, D.M., Reid, J.G. and Worley, K.C.** (2012) Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PloS one*, **7**, e47768.

**Erayman, M., Sandhu, D., Sidhu, D., Dilbirligi, M., Baenziger, P. and Gill, K.S.** (2004) Demarcating the gene-rich regions of the wheat genome. *Nucleic acids research*, **32**, 3546-3565.

**Evanno, G., Regnaut, S. and Goudet, J.** (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular ecology*, **14**, 2611-2620.

**Fahey, J.W., Zalcmann, A.T. and Talalay, P.** (2001) The chemical diversity and distribution of glucosinolates and isothiocyanates among plants. *Phytochemistry*, **56**, 5-51.

**Farnham, M., Davis, E., Morgan, J. and Smith, J.** (2008) Neglected landraces of collard (Brassica oleracea L. var. viridis) from the Carolinas (USA). *Genetic Resources and Crop Evolution*, **55**, 797-801.

**Feldman, M.W., Otto, S.P. and Christiansen, F.B.** (1996) Population genetic perspectives on the evolution of recombination. *Annual review of genetics*, **30**, 261-295.

**Feng, J., Long, Y., Shi, L., Shi, J., Barker, G. and Meng, J.** (2012) Characterization of metabolite quantitative trait loci and metabolic networks that control glucosinolate concentration in the seeds and leaves of Brassica napus. *New phytologist*, **193**, 96-108.

**Fernandes, J.B., Séguéla-Arnaud, M., Larchevêque, C., Lloyd, A.H. and Mercier, R.** (2018) Unleashing meiotic crossovers in hybrid plants. *Proceedings of the National Academy of Sciences*, **115**, 2431-2436.

**Formenti, G., Chiara, M., Poveda, L., Francoijs, K.-J., Bonisoli-Alquati, A., Canova, L., Gianfranceschi, L., Horner, D.S. and Saino, N.** (2019) SMRT long reads and Direct Label and Stain optical maps allow the generation of a high-quality genome assembly for the European barn swallow (Hirundo rustica rustica). *GigaScience*, **8**, giy142.

**France, M.G., Enderle, J., Röhrig, S., Puchta, H., Franklin, F.C.H. and Higgins, J.D.** (2021) ZYP1 is required for obligate cross-over formation and cross-over interference in Arabidopsis. *Proceedings of the National Academy of Sciences*, **118**, e2021671118.

**Francis, K.E., Lam, S.Y., Harrison, B.D., Bey, A.L., Berchowitz, L.E. and Copenhaver, G.P.** (2007) Pollen tetrad-based visual assay for meiotic recombination in Arabidopsis. *Proceedings of the national academy of sciences*, **104**, 3913-3918.

**Frerigmann, H. and Gigolashvili, T.** (2014) MYB34, MYB51, and MYB122 distinctly regulate indolic glucosinolate biosynthesis in Arabidopsis thaliana. *Molecular Plant*, **7**, 814-828.

**Fu, L., Cai, C., Cui, Y., Wu, J., Liang, J., Cheng, F. and Wang, X.** (2016) Pooled mapping: an efficient method of calling variations for population samples with low-depth resequencing data. *Molecular Breeding*, **36**, 1-12.

**Fuentes, R.R., de Ridder, D., van Dijk, A.D. and Peters, S.A.** (2022) Domestication shapes recombination patterns in tomato. *Molecular biology and evolution*, **39**, msab287.

**Gage, J.L., Vaillancourt, B., Hamilton, J.P., Manrique-Carpintero, N.C., Gustafson, T.J., Barry, K., Lipzen, A., Tracy, W.F., Mikel, M.A. and Kaeppler, S.M.** (2019) Multiple maize reference genomes impact the identification of variants by genome-wide association study in a diverse inbred panel. *The plant genome*, **12**, 180069.

**Gao, L., Gonda, I., Sun, H., Ma, Q., Bao, K., Tieman, D.M., Burzynski-Chang, E.A., Fish, T.L., Stromberg, K.A. and Sacks, G.L.** (2019) The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nature genetics*, **51**, 1044-1051.

**Garg, S., Fungtammasan, A., Carroll, A., Chou, M., Schmitt, A., Zhou, X., Mac, S., Peluso, P., Hatas, E. and Ghurye, J.** (2021) Chromosome-scale, haplotype-resolved assembly of human genomes. *Nature biotechnology*, **39**, 309-312.

**Garrison, E. and Marth, G.** (2012) Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907*.

**Garrison, E., Sirén, J., Novak, A.M., Hickey, G., Eizenga, J.M., Dawson, E.T., Jones, W., Garg, S., Markello, C. and Lin, M.F.** (2018) Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature biotechnology*, **36**, 875-879.

**Gel, B., Díez-Villanueva, A., Serra, E., Buschbeck, M., Peinado, M.A. and Malinverni, R.** (2016) regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. *Bioinformatics*, **32**, 289-291.

**Ghurye, J., Rhie, A., Walenz, B.P., Schmitt, A., Selvaraj, S., Pop, M., Phillippy, A.M. and Koren, S.** (2019) Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLoS computational biology*, **15**, e1007273.

**Girard, C., Chelysheva, L., Choinard, S., Froger, N., Macaisne, N., Lehmemdi, A., Mazel, J., Crismani, W. and Mercier, R.** (2015) AAA-ATPase FIDGETIN-LIKE 1 and helicase FANCM antagonize meiotic crossovers by distinct mechanisms. *PLoS genetics*, **11**, e1005369.

**Girard, C., Crismani, W., Froger, N., Mazel, J., Lemhemdi, A., Horlow, C. and Mercier, R.** (2014) FANCM-associated proteins MHF1 and MHF2, but not the other Fanconi anemia factors, limit meiotic crossovers. *Nucleic acids research*, **42**, 9087-9095.

**Giraut, L., Falque, M., Drouaud, J., Pereira, L., Martin, O.C. and Mézard, C.** (2011) Genome-wide crossover distribution in Arabidopsis thaliana meiosis reveals sex-specific patterns along chromosomes. *PLoS genetics*, **7**, e1002354.

**Goel, M., Sun, H., Jiao, W.-B. and Schneeberger, K.** (2019) SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome biology*, **20**, 1-13.

**Golicz, A.A., Bayer, P.E., Barker, G.C., Edger, P.P., Kim, H., Martinez, P.A., Chan, C.K.K., Severn-Ellis, A., McCombie, W.R. and Parkin, I.A.** (2016a) The pangenome of an agronomically important crop plant Brassica oleracea. *Nature communications*, **7**, 13390.

**Golicz, A.A., Bayer, P.E., Barker, G.C., Edger, P.P., Kim, H., Martinez, P.A., Chan, C.K.K., Severn-Ellis, A., McCombie, W.R. and Parkin, I.A.** (2016b) The pangenome of an agronomically important crop plant Brassica oleracea. *Nature communications*, **7**, 1-8.

**Gómez-Campo, C. and Prakash, S.** (1999) 2 Origin and domestication. In *Developments in plant genetics and breeding*: Elsevier, pp. 33-58.

**Gore, M.A., Chia, J.-M., Elshire, R.J., Sun, Q., Ersoz, E.S., Hurwitz, B.L., Peiffer, J.A., McMullen, M.D., Grills, G.S. and Ross-Ibarra, J.** (2009) A first-generation haplotype map of maize. *Science*, **326**, 1115-1117.

**Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R. and Zeng, Q.** (2011) Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature biotechnology*, **29**, 644.

**Grada, A. and Weinbrecht, K.** (2013) Next-generation sequencing: methodology and application. *The Journal of investigative dermatology*, **133**, e11.

**Guo, N., Han, S., Zong, M., Wang, G., Zheng, S. and Liu, F.** (2019) Identification and differential expression analysis of anthocyanin biosynthetic genes in leaf color variants of ornamental kale. *BMC genomics*, **20**, 564.

**Guo, N., Wang, S., Gao, L., Liu, Y., Duan, M., Wang, G., Li, J., Yang, M., Zong, M. and Han, S.** (2020) Genome sequencing sheds light on the contribution of structural variants to Brassica oleracea diversification. *bioRxiv*.

**Guo, N., Wang, S., Gao, L., Liu, Y., Wang, X., Lai, E., Duan, M., Wang, G., Li, J. and Yang, M.** (2021) Genome sequencing sheds light on the contribution of structural variants to Brassica oleracea diversification. *BMC biology*, **19**, 1-15.

**Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell, C.R. and Wortman, J.R.** (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome biology*, **9**, 1-22.

**Habu, Y., Ando, T., Ito, S., Nagaki, K., Kishimoto, N., Taguchi-Shiobara, F., Numa, H., Yamaguchi, K., Shigenobu, S. and Murata, M.** (2015) Epigenomic modification in rice controls meiotic recombination and segregation distortion. *Molecular Breeding*, **35**, 1-8.

**Hahn, C., Müller, A., Kuhnert, N. and Albach, D.** (2016) Diversity of kale (Brassica oleracea var. sabellica): glucosinolate content and phylogenetic relationships. *Journal of agricultural and food chemistry*, **64**, 3215-3225.

**Halkier, B.A. and Gershenzon, J.** (2006) Biology and biochemistry of glucosinolates. *Annual review of plant biology*, **57**, 303-333.

**Hartl, D.** (1980) Principles of Population Genetics. 488 pp. Sunderland, MA: Sinauer Associates.

**Harun, S., Abdullah-Zawawi, M.-R., Goh, H.-H. and Mohamed-Hussein, Z.-A.** (2020) A comprehensive gene inventory for glucosinolate biosynthetic pathway in Arabidopsis thaliana. *Journal of agricultural and food chemistry*, **68**, 7281-7297.

**Hayes, J.D., Kelleher, M.O. and Eggleston, I.M.** (2008) The cancer chemopreventive actions of phytochemicals derived from glucosinolates. *European journal of nutrition*, **47**, 73-88.

**Heller, D. and Vingron, M.** (2020) SVIM-asm: structural variant detection from haploid and diploid genome assemblies. *Bioinformatics*, **36**, 5519-5521.

**Henderson, I.R. and Bomblies, K.** (2021) Evolution and plasticity of genome-wide meiotic recombination rates. *Annual review of genetics*, **55**, 23-43.

**Higdon, J.V., Delage, B., Williams, D.E. and Dashwood, R.H.** (2007) Cruciferous vegetables and human cancer risk: epidemiologic evidence and mechanistic basis. *Pharmacological research*, **55**, 224-236.

**Higgins, J.D., Perry, R.M., Barakate, A., Ramsay, L., Waugh, R., Halpin, C., Armstrong, S.J. and Franklin, F.C.H.** (2012) Spatiotemporal asymmetry of the meiotic program underlies the predominantly distal distribution of meiotic crossovers in barley. *The Plant Cell*, **24**, 4096-4109.

**Himelblau, E., Gilchrist, E.J., Buono, K., Bizzell, C., Mentzer, L., Vogelzang, R., Osborn, T., Amasino, R.M., Parkin, I.A. and Haughn, G.W.** (2009) Forward and reverse genetics of rapid-cycling Brassica oleracea. *Theoretical and applied genetics*, **118**, 953-961.

**Hoopes, G., Meng, X., Hamilton, J.P., Achakkagari, S.R., Guesdes, F.d.A.F., Bolger, M.E., Coombs, J.J., Esselink, D., Kaiser, N.R. and Kodde, L.** (2022) Phased, chromosome-scale genome assemblies of tetraploid potato reveal a complex genome, transcriptome, and predicted proteome landscape underpinning genetic diversity. *Molecular Plant*, **15**, 520-536.

**Hufford, M.B., Seetharam, A.S., Woodhouse, M.R., Chougule, K.M., Ou, S., Liu, J., Ricci, W.A., Guo, T., Olson, A. and Qiu, Y.** (2021) De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science*, **373**, 655-662.

**Hunter, N.** (2015) Meiotic recombination: the essence of heredity. *Cold Spring Harbor perspectives in biology*, **7**, a016618.

**Izzah, N.K., Lee, J., Perumal, S., Park, J.Y., Ahn, K., Fu, D., Kim, G.-B., Nam, Y.-W. and Yang, T.-J.** (2013) Microsatellite-based analysis of genetic diversity in 91 commercial Brassica oleracea L. cultivars belonging to six varietal groups. *Genetic resources and crop evolution*, **60**, 1967-1986.

**Jain, M., Koren, S., Miga, K.H., Quick, J., Rand, A.C., Sasani, T.A., Tyson, J.R., Beggs, A.D., Dilthey, A.T. and Fiddes, I.T.** (2018) Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature biotechnology*, **36**, 338-345.

**Jain, M., Tyson, J.R., Loose, M., Ip, C.L., Eccles, D.A., O'Grady, J., Malla, S., Leggett, R.M., Wallerman, O. and Jansen, H.J.** (2017) MinION Analysis and Reference Consortium: Phase 2 data release and analysis of R9. 0 chemistry. *F1000Research*, **6**.

**Jayakodi, M., Padmarasu, S., Haberer, G., Bonthala, V.S., Gundlach, H., Monat, C., Lux, T., Kamal, N., Lang, D. and Himmelbach, A.** (2020) The barley pan-genome reveals the hidden legacy of mutation breeding. *Nature*, **588**, 284-289.

**Jensen, L.M., Halkier, B.A. and Burow, M.** (2014) How to discover a metabolic pathway? An update on gene identification in aliphatic glucosinolate biosynthesis, regulation and transport. *Biological chemistry*, **395**, 529-543.

**Jeon, J.-S., Carreno-Quintero, N., van Eekelen, H.D., De Vos, R.C., Raaijmakers, J.M. and Etalo, D.W.** (2021) Impact of root-associated strains of three Paraburkholderia species on primary and secondary metabolism of Brassica oleracea. *Scientific reports*, **11**, 1-14.

**Jiao, W.-B., Accinelli, G.G., Hartwig, B., Kiefer, C., Baker, D., Severing, E., Willing, E.-M., Piednoel, M., Woetzel, S. and Madrid-Herrero, E.** (2017) Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data. *Genome research*, **27**, 778-786.

**Jiao, W.-B. and Schneeberger, K.** (2017) The impact of third generation genomic technologies on plant genome assembly. *Current opinion in plant biology*, **36**, 64-70.

**Jiao, W.-B. and Schneeberger, K.** (2020) Chromosome-level assemblies of multiple Arabidopsis genomes reveal hotspots of rearrangements with altered evolutionary dynamics. *Nature communications*, **11**, 1-10.

**Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A. and Nuka, G.** (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**, 1236-1240.

**Jonkheer, E.M., van Workum, D.-J.M., Sheikhizadeh Anari, S., Brankovics, B., de Haan, J.R., Berke, L., van der Lee, T.A., de Ridder, D. and Smit, S.** (2022) PanTools v3: functional annotation, classification and phylogenomics. *Bioinformatics*, **38**, 4403-4405.

**Jump, A.S., Marchant, R. and Peñuelas, J.** (2009) Environmental change and the option value of genetic diversity. *Trends in plant science*, **14**, 51-58.

**Jung, H., Winefield, C., Bombarely, A., Prentis, P. and Waterhouse, P.** (2019) Tools and strategies for long-read sequencing and de novo assembly of plant genomes. *Trends in plant science*, **24**, 700-724.

**Kamvar, Z.N., Tabima, J.F. and Grünwald, N.J.** (2014) Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ*, **2**, e281.

**Katoh, K., Kuma, K.-i., Toh, H. and Miyata, T.** (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic acids research*, **33**, 511-518.

**Kearsey, M., Ramsay, L., Jennings, D., Lydiate, D., Bohuon, E. and Marshall, D.** (1996) Higher recombination frequencies in female compared to male meisoses in Brassica oleracea. *Theoretical and Applied Genetics*, **92**, 363-367.

**Keeney, S., Giroux, C.N. and Kleckner, N.** (1997) Meiosis-specific DNA double-strand breaks are catalyzed by Spo11, a member of a widely conserved protein family. *Cell*, **88**, 375-384.

**Keilwagen, J., Wenk, M., Erickson, J.L., Schattat, M.H., Grau, J. and Hartung, F.** (2016) Using intron position conservation for homology-based gene prediction. *Nucleic acids research*, **44**, e89-e89.

**Kelly, A., Sharpe, A., Nixon, J., Lydiate, D. and Evans, E.** (1997) Indistinguishable patterns of recombination resulting from male and female meioses in Brassica napus (oilseed rape). *Genome*, **40**, 49-56.

**Kent, W.J.** (2002) BLAT—the BLAST-like alignment tool. *Genome research*, **12**, 656-664.

**Kianian, P., Wang, M., Simons, K., Ghavami, F., He, Y., Dukowic-Schulze, S., Sundararajan, A., Sun, Q., Pillardy, J. and Mudge, J.** (2018) High-resolution crossover mapping reveals similarities and differences of male and female recombination in maize. *Nature communications*, **9**, 1-10.

**Kianian, S.F. and Quiros, C.F.** (1992) Trait inheritance, fertility, and genomic relationships of some n= 9 Brassica species. *Genetic Resources and Crop Evolution*, **39**, 165-175.

**Kim, D., Langmead, B. and Salzberg, S.L.** (2015) HISAT: a fast spliced aligner with low memory requirements. *Nature methods*, **12**, 357-360.

**Kim, H. and Choi, K.** (2022) Fast and Precise: How to Measure Meiotic Crossovers in Arabidopsis. *Molecules and Cells*, **45**, 273.

**Kim, M.J., Kim, H., Shin, J.S., Chung, C.-H., Ohlrogge, J.B. and Suh, M.C.** (2006) Seed-specific expression of sesame microsomal oleic acid desaturase is controlled by combinatorial properties between negative cis-regulatory elements in the SeFAD2 promoter and enhancers in the 5′-UTR intron. *Molecular Genetics and Genomics*, **276**, 351-368.

**Kliebenstein, D.J., Kroymann, J., Brown, P., Figuth, A., Pedersen, D., Gershenzon, J. and Mitchell-Olds, T.** (2001a) Genetic control of natural variation in Arabidopsis glucosinolate accumulation. *Plant physiology*, **126**, 811-825.

**Kliebenstein, D.J., Lambrix, V.M., Reichelt, M., Gershenzon, J. and Mitchell-Olds, T.** (2001b) Gene duplication in the diversification of secondary metabolism: tandem 2-oxoglutarate–dependent dioxygenases control glucosinolate biosynthesis in Arabidopsis. *The Plant Cell*, **13**, 681-693.

**Kole, C. and Henry, R.J.** (2010) *Genetics, genomics and breeding of crop plants*: Science Publishers.

**Kolmogorov, M., Yuan, J., Lin, Y. and Pevzner, P.A.** (2019) Assembly of long, error-prone reads using repeat graphs. *Nature biotechnology*, **37**, 540-546.

**Koo, D.-H., Hong, C.P., Batley, J., Chung, Y.S., Edwards, D., Bang, J.-W., Hur, Y. and Lim, Y.P.** (2011) Rapid divergence of repetitive DNAs in Brassica relatives. *Genomics*, **97**, 173-185.

**Koo, D.-H., Plaha, P., Lim, Y.P., Hur, Y. and Bang, J.-W.** (2004) A high-resolution karyotype of Brassica rapa ssp. pekinensis revealed by pachytene analysis and multicolor fluorescence in situ hybridization. *Theoretical and Applied Genetics*, **109**, 1346-1352.

**Koonin, E.V., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Krylov, D.M., Makarova, K.S., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N. and Rao, B.S.** (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome biology*, **5**, 1-28.

**Kopelman, N.M., Mayzel, J., Jakobsson, M., Rosenberg, N.A. and Mayrose, I.** (2015) Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. *Molecular ecology resources*, **15**, 1179-1191.

**Koren, S., Rhie, A., Walenz, B.P., Dilthey, A.T., Bickhart, D.M., Kingan, S.B., Hiendleder, S., Williams, J.L., Smith, T.P. and Phillippy, A.M.** (2018) De novo assembly of haplotype-resolved genomes with trio binning. *Nature biotechnology*, **36**, 1174-1182.

**Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H. and Phillippy, A.M.** (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome research*, **27**, 722-736.

**Korf, I.** (2004) Gene finding in novel genomes. *BMC bioinformatics*, **5**, 1-9.

**Kovaka, S., Zimin, A.V., Pertea, G.M., Razaghi, R., Salzberg, S.L. and Pertea, M.** (2019) Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome biology*, **20**, 1-13.

**Lashermes, P., Combes, M.-C., Prakash, N., Trouslot, P., Lorieux, M. and Charrier, A.** (2001) Genetic linkage map of Coffea canephora: effect of segregation distortion and analysis of recombination rate in male and female meioses. *Genome*, **44**, 589-595.

**Lawrence, E.J., Gao, H., Tock, A.J., Lambing, C., Blackwell, A.R., Feng, X. and Henderson, I.R.** (2019) Natural variation in TBP-ASSOCIATED FACTOR 4b controls meiotic crossover and germline transcription in Arabidopsis. *Current Biology*, **29**, 2676-2686. e2673.

**Leaché, A.D. and Oaks, J.R.** (2017) The utility of single nucleotide polymorphism (SNP) data in phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, **48**, 69-84.

**Lenormand, T.** (2003) The evolution of sex dimorphism in recombination. *Genetics*, **163**, 811-822.

**Lenormand, T. and Dutheil, J.** (2005) Recombination difference between sexes: a role for haploid selection. *PLoS biology*, **3**, e63.

**Letunic, I. and Bork, P.** (2016) Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic acids research*, **44**, W242-W245.

**Li, A., Liu, D., Wu, J., Zhao, X., Hao, M., Geng, S., Yan, J., Jiang, X., Zhang, L. and Wu, J.** (2014a) mRNA and small RNA transcriptomes reveal insights into dynamic homoeolog regulation of allopolyploid heterosis in nascent hexaploid wheat. *The Plant Cell*, **26**, 1878-1900.

**Li, G. and Quiros, C.** (2003) In planta side-chain glucosinolate modification in Arabidopsis by introduction of dioxygenase Brassica homolog BoGSL-ALK. *Theoretical and applied genetics*, **106**, 1116-1121.

**Li, H.** (2016) Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics*, **32**, 2103-2110.

**Li, H.** (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**, 3094-3100.

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

**Li, H., Feng, X. and Chu, C.** (2020a) The design and construction of reference pangenome graphs with minigraph. *Genome biology*, **21**, 1-19.

**Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R.** (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078-2079.

**Li, H., Wang, S., Chai, S., Yang, Z., Zhang, Q., Xin, H., Xu, Y., Lin, S., Chen, X. and Yao, Z.** (2022) Graph-based pan-genome reveals structural and sequence variations related to agronomic traits and domestication in cucumber. *Nature communications*, **13**, 1-14.

**Li, J., Yuan, D., Wang, P., Wang, Q., Sun, M., Liu, Z., Si, H., Xu, Z., Ma, Y. and Zhang, B.** (2021) Cotton pan-genome retrieves the lost sequences and genes during domestication and selection. *Genome biology*, **22**, 1-26.

**Li, R., Bitoun, E., Altemose, N., Davies, R.W., Davies, B. and Myers, S.R.** (2019a) A high-resolution map of non-crossover events reveals impacts of genetic diversity on mammalian meiotic recombination. *Nature communications*, **10**, 1-15.

**Li, R., Qu, H., Chen, J., Wang, S., Chater, J.M., Zhang, L., Wei, J., Zhang, Y.-M., Xu, C. and Zhong, W.-D.** (2020b) Inference of chromosome-length haplotypes using genomic data of three or a few more single gametes. *Molecular biology and evolution*, **37**, 3684-3698.

**Li, X., Li, L. and Yan, J.** (2015) Dissecting meiotic recombination based on tetrad analysis by single-microspore sequencing in maize. *Nature communications*, **6**, 1-9.

**Li, Y.-h., Zhou, G., Ma, J., Jiang, W., Jin, L.-g., Zhang, Z., Guo, Y., Zhang, J., Sui, Y. and Zheng, L.** (2014b) De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nature biotechnology*, **32**, 1045-1052.

**Li, Z., Wang, M., Lin, K., Xie, Y., Guo, J., Ye, L., Zhuang, Y., Teng, W., Ran, X. and Tong, Y.** (2019b) The bread wheat epigenomic map reveals distinct chromatin architectural and evolutionary features of functional genetic elements. *Genome biology*, **20**, 1-16.

**Lian, Q., Solier, V., Walkemeier, B., Durand, S., Huettel, B., Schneeberger, K. and Mercier, R.** (2022a) The megabase-scale crossover landscape is largely independent of sequence divergence. *Nature communications*, **13**, 1-11.

**Lian, Q., Solier, V., Walkemeier, B., Huettel, B. and Schneeberger, K.** (2022b) The megabase-scale crossover landscape is independent of sequence divergence. *bioRxiv*.

**Lieberman-Aiden, E., Van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J. and Dorschner, M.O.** (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science*, **326**, 289-293.

**Lim, K.-B., De Jong, H., Yang, T.-J., Park, J.-Y., Kwon, S.-J., Kim, J.S., Lim, M.-H., Kim, J.A., Jin, M. and Jin, Y.-M.** (2005) Characterization of rDNAs and tandem repeats in the heterochromatin of Brassica rapa. *Mol Cells*, **19**, 436-444.

**Lim, K.B., Yang, T.J., Hwang, Y.J., Kim, J.S., Park, J.Y., Kwon, S.J., Kim, J., Choi, B.S., Lim, M.H. and Jin, M.** (2007) Characterization of the centromere and peri-centromere retrotransposons in Brassica rapa and their distribution in related Brassica species. *The Plant Journal*, **49**, 173-183.

**Linsmith, G., Rombauts, S., Montanari, S., Deng, C.H., Celton, J.-M., Guérif, P., Liu, C., Lohaus, R., Zurn, J.D. and Cestaro, A.** (2019) Pseudo-chromosome–length genome assembly of a double haploid "Bartlett" pear (Pyrus communis L.). *Gigascience*, **8**, giz138.

**Lisch, D.** (2013) How important are transposons for plant evolution? *Nature Reviews Genetics*, **14**, 49-61.

**Litt, M. and Luty, J.A.** (1989) A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. *American journal of human genetics*, **44**, 397.

**Liu, H., Wang, X., Wang, G., Cui, P., Wu, S., Ai, C., Hu, N., Li, A., He, B. and Shao, X.** (2021a) The nearly complete genome of Ginkgo biloba illuminates gymnosperm evolution. *Nature Plants*, **7**, 748-756.

**Liu, H., Wu, S., Li, A. and Ruan, J.** (2021b) SMARTdenovo: a de novo assembler using long noisy reads. *Gigabyte*, **2021**, 1-9.

**Liu, S., Liu, Y., Yang, X., Tong, C., Edwards, D., Parkin, I.A., Zhao, M., Ma, J., Yu, J. and Huang, S.** (2014a) The Brassica oleracea genome reveals the asymmetrical evolution of polyploid genomes. *Nature communications*, **5**, 1-11.

**Liu, S., Liu, Y., Yang, X., Tong, C., Edwards, D., Parkin, I.A., Zhao, M., Ma, J., Yu, J. and Huang, S.** (2014b) The Brassica oleracea genome reveals the asymmetrical evolution of polyploid genomes. *Nature communications*, **5**.

**Liu, S., Yeh, C.-T., Ji, T., Ying, K., Wu, H., Tang, H.M., Fu, Y., Nettleton, D. and Schnable, P.S.** (2009) Mu transposon insertion sites and meiotic recombination events co-localize with epigenetic marks for open chromatin across the maize genome. *PLoS genetics*, **5**, e1000733.

**Liu, Y., Du, H., Li, P., Shen, Y., Peng, H., Liu, S., Zhou, G.-A., Zhang, H., Liu, Z. and Shi, M.** (2020) Pan-genome of wild and cultivated soybeans. *Cell*, **182**, 162-176. e113.

**Lloyd, A., Morgan, C., H. Franklin, F.C. and Bomblies, K.** (2018) Plasticity of meiotic recombination rates in response to temperature in Arabidopsis. *Genetics*, **208**, 1409-1420.

**Logsdon, G.A., Vollger, M.R. and Eichler, E.E.** (2020) Long-read human genome sequencing and its applications. *Nature Reviews Genetics*, **21**, 597-614.

**López, E., Pradillo, M., Oliver, C., Romero, C., Cuñado, N. and Santos, J.** (2012) Looking for natural variation in chiasma frequency in Arabidopsis thaliana. *Journal of experimental botany*, **63**, 887-894.

**Lu, K., Wei, L., Li, X., Wang, Y., Wu, J., Liu, M., Zhang, C., Chen, Z., Xiao, Z., Jian, H., Cheng, F., Zhang, K., Du, H., Cheng, X., Qu, C., Qian, W., Liu, L., Wang, R., Zou, Q., Ying, J., Xu, X., Mei, J., Liang, Y., Chai, Y.R., Tang, Z., Wan, H., Ni, Y., He, Y., Lin, N., Fan, Y., Sun, W., Li, N.N., Zhou, G., Zheng, H., Wang, X., Paterson, A.H. and Li, J.** (2019) Whole-genome resequencing reveals Brassica napus origin and genetic loci involved in its improvement. *Nat Commun*, **10**, 1154.

**Luo, C., Li, X., Zhang, Q. and Yan, J.** (2019) Single gametophyte sequencing reveals that crossover events differ between sexes in maize. *Nature communications*, **10**, 1-8.

**Lv, H., Wang, Y., Han, F., Ji, J., Fang, Z., Zhuang, M., Li, Z., Zhang, Y. and Yang, L.** (2020) A high-quality reference genome for cabbage obtained with SMRT reveals novel genomic features and evolutionary characteristics. *Scientific reports*, **10**, 1-9.

**Lysak, M.A., Koch, M.A., Pecinka, A. and Schubert, I.** (2005) Chromosome triplication found across the tribe Brassiceae. *Genome research*, **15**, 516-525.

**Mabry, M.E., Turner-Hissong, S.D., Gallagher, E.Y., McAlvay, A.C., An, H., Edger, P.P., Moore, J.D., Pink, D.A., Teakle, G.R. and Stevens, C.J.** (2021) The Evolutionary History of Wild, Domesticated, and Feral Brassica Oleracea (Brassicaceae). *Molecular Biology and Evolution*.

**Macias, F.A., Molinillo, J.M., Varela, R.M. and Galindo, J.C.** (2007) Allelopathy—a natural alternative for weed control. *Pest Management Science: Formerly Pesticide Science*, **63**, 327-348.

**Maddison, W. and Maddison, D.** (2019) Mesquite: A modular system for evolutionary analysis. Version 3.61. 2019.

**Maggioni, L.** (2015) *Domestication of Brassica oleracea L.*

**Maggioni, L., von Bothmer, R., Poulsen, G., Branca, F. and Bagger Jørgensen, R.** (2014) Genetic diversity and population structure of leafy kale and Brassica rupestris Raf. in south Italy. *Hereditas*, **151**, 145-158.

**Maggioni, L., von Bothmer, R., Poulsen, G. and Lipman, E.** (2018) Domestication, diversity and use of Brassica oleracea L., based on ancient Greek and Latin texts. *Genetic resources and crop evolution*, **65**, 137-159.

**Maia, R.T. and de Araújo Campos, M.** (2021) Introductory Chapter: Genetic Variation-The Source of Biological Diversity. In *Genetic Variation*: IntechOpen.

**Majoros, W.H., Pertea, M. and Salzberg, S.L.** (2004) TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics*, **20**, 2878-2879.

**Marand, A.P., Jansky, S.H., Gage, J.L., Hamernik, A.J., de Leon, N. and Jiang, J.** (2019) Residual heterozygosity and epistatic interactions underlie the complex genetic architecture of yield in diploid potato. *Genetics*, **212**, 317-332.

**Marand, A.P., Jansky, S.H., Zhao, H., Leisner, C.P., Zhu, X., Zeng, Z., Crisovan, E., Newton, L., Hamernik, A.J. and Veilleux, R.E.** (2017) Meiotic crossovers are associated with open chromatin and enriched with Stowaway transposons in potato. *Genome biology*, **18**, 1-16.

**Marçais, G., Delcher, A.L., Phillippy, A.M., Coston, R., Salzberg, S.L. and Zimin, A.** (2018) MUMmer4: A fast and versatile genome alignment system. *PLoS computational biology*, **14**, e1005944.

**Mardis, E.R.** (2017) DNA sequencing technologies: 2006–2016. *Nature protocols*, **12**, 213-218.

**Marks, R.A., Hotaling, S., Frandsen, P.B. and VanBuren, R.** (2021) Representation and participation across 20 years of plant genome sequencing. *Nature plants*, **7**, 1571-1578.

**Martin, O.C. and Wagner, A.** (2009) Effects of recombination on complex regulatory circuits. *Genetics*, **183**, 673-684.

**McCallum, C.M., Comai, L., Greene, E.A. and Henikoff, S.** (2000) T argeting I nduced L ocal L esions IN G enomes (TILLING) for plant functional genomics. *Plant physiology*, **123**, 439-442.

**McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S. and Daly, M.** (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, **20**, 1297-1303.

**Melamed-Bessudo, C. and Levy, A.A.** (2012) Deficiency in DNA methylation increases meiotic crossover rates in euchromatic but not in heterochromatic regions in Arabidopsis. *Proceedings of the National Academy of Sciences*, **109**, E981-E988.

**Melamed-Bessudo, C., Yehuda, E., Stuitje, A.R. and Levy, A.A.** (2005) A new seed-based assay for meiotic recombination in Arabidopsis thaliana. *The Plant Journal*, **43**, 458-466.

**Mercier, R., Mézard, C., Jenczewski, E., Macaisne, N. and Grelon, M.** (2015) The molecular biology of meiosis in plants. *Annual review of plant biology*, **66**, 297-327.

**Mézard, C., Jahns, M.T. and Grelon, M.** (2015) Where to cross? New insights into the location of meiotic crossovers. *Trends in Genetics*, **31**, 393-401.

**Mezard, C., Vignard, J., Drouaud, J. and Mercier, R.** (2007) The road to crossovers: plants have their say. *TRENDS in Genetics*, **23**, 91-99.

**Michael, T.P., Jupe, F., Bemm, F., Motley, S.T., Sandoval, J.P., Lanz, C., Loudet, O., Weigel, D. and Ecker, J.R.** (2018) High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell. *Nature communications*, **9**, 1-8.

**Mieulet, D., Aubert, G., Bres, C., Klein, A., Droc, G., Vieille, E., Rond-Coissieux, C., Sanchez, M., Dalmais, M. and Mauxion, J.-P.** (2018) Unleashing meiotic crossovers in crops. *Nature Plants*, **4**, 1010-1016.

**Miller, J.R., Koren, S. and Sutton, G.** (2010) Assembly algorithms for next-generation sequencing data. *Genomics*, **95**, 315-327.

**Ming, R., VanBuren, R., Wai, C.M., Tang, H., Schatz, M.C., Bowers, J.E., Lyons, E., Wang, M.-L., Chen, J. and Biggers, E.** (2015) The pineapple genome and the evolution of CAM photosynthesis. *Nature genetics*, **47**, 1435-1442.

**Mirouze, M., Lieberman-Lazarovich, M., Aversano, R., Bucher, E., Nicolet, J., Reinders, J. and Paszkowski, J.** (2012) Loss of DNA methylation affects the recombination landscape in Arabidopsis. *Proceedings of the National Academy of Sciences*, **109**, 5880-5885.

**Mithen, R.F., Dekker, M., Verkerk, R., Rabot, S. and Johnson, I.T.** (2000) The nutritional significance, biosynthesis and bioavailability of glucosinolates in human foods. *Journal of the Science of Food and Agriculture*, **80**, 967-984.

**Modliszewski, J.L. and Copenhaver, G.P.** (2017) Meiotic recombination gets stressed out: CO frequency is plastic under pressure. *Current Opinion in Plant Biology*, **36**, 95-102.

**Modliszewski, J.L., Wang, H., Albright, A.R., Lewis, S.M., Bennett, A.R., Huang, J., Ma, H., Wang, Y. and Copenhaver, G.P.** (2018) Elevated temperature increases meiotic crossover frequency via the interfering (Type I) pathway in Arabidopsis thaliana. *PLoS genetics*, **14**, e1007384.

**Mohammadi, S.A. and Prasanna, B.** (2003) Analysis of genetic diversity in crop plants—salient statistical tools and considerations. *Crop science*, **43**, 1235-1248.

**Moran, E.S., Armstrong, S., Santos, J., Franklin, F. and Jones, G.** (2001) Chiasma formation in Arabidopsis thaliana accession Wassileskija and in two meiotic mutants. *Chromosome Research*, **9**, 121-128.

**Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A.C. and Kanehisa, M.** (2007) KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic acids research*, **35**, W182-W185.

**Muller, H.J.** (1916a) The mechanism of crossing-over. II. IV. The manner of occurrence of crossing-over. *The American Naturalist*, **50**, 284-305.

**Muller, H.J.** (1916b) The mechanism of crossing-over. IV. *The American Naturalist*, **50**, 421-434.

**Murat, F., Zhang, R., Guizard, S., Flores, R., Armero, A., Pont, C., Steinbach, D., Quesneville, H., Cooke, R. and Salse, J.** (2014) Shared subgenome dominance following polyploidization explains grass genome evolutionary plasticity from a seven protochromosome ancestor with 16K protogenes. *Genome biology and evolution*, **6**, 12-33.

**Murray, M. and Thompson, W.** (1980) Rapid isolation of high molecular weight plant DNA. *Nucleic acids research*, **8**, 4321-4326.

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

**Nagaharu, U.** (1935b) Genome analysis in Brassica with special reference to the experimental formation of B. napus and peculiar mode of fertilization. *Japanese Journal of Botany*, 389-452.

**Nagarajan, N. and Pop, M.** (2013) Sequence assembly demystified. *Nature Reviews Genetics*, **14**, 157-167.

**Nageswaran, D.C., Kim, J., Lambing, C., Kim, J., Park, J., Kim, E.-J., Cho, H.S., Kim, H., Byun, D. and Park, Y.M.** (2021) HIGH CROSSOVER RATE1 encodes PROTEIN

PHOSPHATASE X1 and restricts meiotic crossovers in Arabidopsis. *Nature plants*, **7**, 452-467.

**Navrátilová, P., Toegelová, H., Tulpová, Z., Kuo, Y.T., Stein, N., Doležel, J., Houben, A., Šimková, H. and Mascher, M.** (2022) Prospects of telomere-to-telomere assembly in barley: Analysis of sequence gaps in the MorexV3 reference genome. *Plant Biotechnology Journal*.

**Nei, M.** (1967) Modification of linkage intensity by natural selection. *Genetics*, **57**, 625.

**Nei, M. and Li, W.-H.** (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences*, **76**, 5269-5273.

**Nguyen, L.-T., Schmidt, H.A., Von Haeseler, A. and Minh, B.Q.** (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution*, **32**, 268-274.

**Nordborg, M., Hu, T.T., Ishino, Y., Jhaveri, J., Toomajian, C., Zheng, H., Bakker, E., Calabrese, P., Gladstone, J. and Goyal, R.** (2005) The pattern of polymorphism in Arabidopsis thaliana. *PLoS biology*, **3**, e196.

**Nour-Eldin, H.H., Andersen, T.G., Burow, M., Madsen, S.R., Jørgensen, M.E., Olsen, C.E., Dreyer, I., Hedrich, R., Geiger, D. and Halkier, B.A.** (2012) NRT/PTR transporters are essential for translocation of glucosinolate defence compounds to seeds. *Nature*, **488**, 531-534.

**Nystedt, B., Street, N.R., Wetterbom, A., Zuccolo, A., Lin, Y.-C., Scofield, D.G., Vezzi, F., Delhomme, N., Giacomello, S. and Alexeyenko, A.** (2013) The Norway spruce genome sequence and conifer genome evolution. *Nature*, **497**, 579-584.

**Olsen, U.D.J.G.I.H.T.B.E.P.P.R.P.W.S.S.T.D.N.C.J.-F., 9, R.G.S.C.S.Y.F.A.H.M.Y.T.T.A.I.T.K.C.W.H.T.Y., Genoscope, 10, C.U.-W.J.H.R.S.W.A.F.B.P.B.T.P.E.R.C.W.P., Department of Genome Analysis, I.o.M.B.R.A.P.M.N.G.T.S.R.A., 11, G.S.C.S.D.R.D.-S.L.R.M.W.K.L.H.M.D.J. and 15, B.G.I.H.G.C.Y.H.Y.J.W.J.H.G.G.J.** (2001) Initial sequencing and analysis of the human genome. *nature*, **409**, 860-921.

**Osman, K., Higgins, J.D., Sanchez-Moran, E., Armstrong, S.J. and Franklin, F.C.H.** (2011) Pathways to meiotic recombination in Arabidopsis thaliana. *New Phytologist*, **190**, 523-544.

**Ou, S., Chen, J. and Jiang, N.** (2018) Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic acids research*, **46**, e126-e126.

**Ou, S. and Jiang, N.** (2018) LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant physiology*, **176**, 1410-1422.

**Ou, S. and Jiang, N.** (2019) LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. *Mobile DNA*, **10**, 1-3.

**Ou, S., Liu, J., Chougule, K.M., Fungtammasan, A., Seetharam, A.S., Stein, J.C., Llaca, V., Manchanda, N., Gilbert, A.M. and Wei, S.** (2020) Effect of sequence depth and length in long-read assembly of the maize inbred NC358. *Nature communications*, **11**, 1-10.

**Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J.R., Hellinga, A.J., Lugo, C.S.B., Elliott, T.A., Ware, D. and Peterson, T.** (2019) Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome biology*, **20**, 1-18.

**Paape, T., Zhou, P., Branca, A., Briskine, R., Young, N. and Tiffin, P.** (2012) Fine-scale population recombination rates, hotspots, and correlates of recombination in the Medicago truncatula genome. *Genome biology and evolution*, **4**, 726-737.

**Padilla, G., Cartea, M.E., Velasco, P., de Haro, A. and Ordás, A.** (2007) Variation of glucosinolates in vegetable crops of Brassica rapa. *Phytochemistry*, **68**, 536-545.

**Parkin, I.A., Koh, C., Tang, H., Robinson, S.J., Kagale, S., Clarke, W.E., Town, C.D., Nixon, J., Krishnakumar, V. and Bidwell, S.L.** (2014a) Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid Brassica oleracea. *Genome biology*, **15**, 1-18.

**Parkin, I.A., Koh, C., Tang, H., Robinson, S.J., Kagale, S., Clarke, W.E., Town, C.D., Nixon, J., Krishnakumar, V. and Bidwell, S.L.** (2014b) Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid Brassica oleracea. *Genome biology*, **15**, R77.

**Patterson, N., Price, A.L. and Reich, D.** (2006) Population structure and eigenanalysis. *PLoS genet*, **2**, e190.

**Pelc, S.E., Couillard, D.M., Stansell, Z.J. and Farnham, M.W.** (2015) Genetic diversity and population structure of collard landraces and their relationship to other Brassica oleracea Crops. *The Plant Genome*, **8**, plantgenome2015.2004.0023.

**Pelé, A., Falque, M., Trotoux, G., Eber, F., Negre, S., Gilet, M., Huteau, V., Lodé, M., Jousseaume, T. and Dechaumet, S.** (2017) Amplifying recombination genome-wide and reshaping crossover landscapes in Brassicas. *PLoS genetics*, **13**, e1006794.

**Pelé, A., Rousseau-Gueutin, M. and Chèvre, A.-M.** (2018) Speciation success of polyploid plants closely relates to the regulation of meiotic recombination. *Frontiers in plant science*, **9**, 907.

**Penhallurick, R.D.** (2008) *Tin in Antiquity: Its Mining and Trade Throughout the Ancient World with Particular Reference to Cornwall*: Maney for the Institute of Materials, Minerals and Mining.

**Perumal, S., Koh, C.S., Jin, L., Buchwaldt, M., Higgins, E.E., Zheng, C., Sankoff, D., Robinson, S.J., Kagale, S. and Navabi, Z.-K.** (2020) A high-contiguity Brassica nigra genome localizes active centromeres and defines the ancestral Brassica genome. *Nature plants*, **6**, 929-941.

**Petersen, A., Wang, C., Crocoll, C. and Halkier, B.A.** (2018) Biotechnological approaches in glucosinolate production. *Journal of integrative plant biology*, **60**, 1231-1248.

**Petes, T.D.** (2001) Meiotic recombination hot spots and cold spots. *Nature Reviews Genetics*, **2**, 360-369.

**Pont, C., Murat, F., Guizard, S., Flores, R., Foucrier, S., Bidet, Y., Quraishi, U.M., Alaux, M., Doležel, J. and Fahima, T.** (2013) Wheat syntenome unveils new evidences of contrasted evolutionary plasticity between paleo-and neoduplicated subgenomes. *The Plant Journal*, **76**, 1030-1044.

**Prescott, A.G. and Lloyd, M.D.** (2000) The iron (II) and 2-oxoacid-dependent dioxygenases and their role in metabolism. *Natural product reports*, **17**, 367-383.

**Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. and Reich, D.** (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, **38**, 904-909.

**Pritchard, J.K., Stephens, M. and Donnelly, P.** (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945-959.

**Putri, G.H., Anders, S., Pyl, P.T., Pimanda, J.E. and Zanini, F.** (2022) Analysing high-throughput sequencing data in Python with HTSeq 2.0. *Bioinformatics*, **38**, 2943-2945.

**Qin, P., Lu, H., Du, H., Wang, H., Chen, W., Chen, Z., He, Q., Ou, S., Zhang, H. and Li, X.** (2021) Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell*, **184**, 3542-3558. e3516.

**Quail, M.A., Smith, M., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P. and Gu, Y.** (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC genomics*, **13**, 1-13.

**Quinlan, A.R. and Hall, I.M.** (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841-842.

**Rakocevic, G., Semenyuk, V., Lee, W.-P., Spencer, J., Browning, J., Johnson, I.J., Arsenijevic, V., Nadj, J., Ghose, K. and Suciu, M.C.** (2019) Fast and accurate genomic analyses using genome graphs. *Nature genetics*, **51**, 354-362.

**Rautiainen, M., Mäkinen, V. and Marschall, T.** (2019) Bit-parallel sequence-to-graph alignment. *Bioinformatics*, **35**, 3599-3607.

**Raz, A., Dahan-Meir, T., Melamed-Bessudo, C., Leshkowitz, D. and Levy, A.A.** (2021) Redistribution of meiotic crossovers along wheat chromosomes by virus-induced gene silencing. *Frontiers in plant science*, **11**, 2332.

**Renny-Byfield, S., Gong, L., Gallagher, J.P. and Wendel, J.F.** (2015) Persistence of subgenomes in paleopolyploid cotton after 60 my of evolution. *Molecular biology and evolution*, **32**, 1063-1071.

**Rensing, S.A., Lang, D., Zimmer, A.D., Terry, A., Salamov, A., Shapiro, H., Nishiyama, T., Perroud, P.-F., Lindquist, E.A. and Kamisugi, Y.** (2008) The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. *Science*, **319**, 64-69.

**Rigal, M., Kevei, Z., Pélissier, T. and Mathieu, O.** (2012) DNA methylation in an intron of the IBM1 histone demethylase gene stabilizes chromatin modification patterns. *The EMBO journal*, **31**, 2981-2993.

**Rommel Fuentes, R., Hesselink, T., Nieuwenhuis, R., Bakker, L., Schijlen, E., van Dooijeweert, W., Diaz Trivino, S., de Haan, J.R., Sanchez Perez, G. and Zhang, X.** (2020) Meiotic recombination profiling of interspecific hybrid F1 tomato pollen by linked read sequencing. *The Plant Journal*, **102**, 480-492.

**Rousseau-Gueutin, M., Belser, C., Da Silva, C., Richard, G., Istace, B., Cruaud, C., Falentin, C., Boideau, F., Boutte, J. and Delourme, R.** (2020) Long-read assembly of the Brassica napus reference genome Darmor-bzh. *GigaScience*, **9**, giaa137.

**Rowan, B.A., Heavens, D., Feuerborn, T.R., Tock, A.J., Henderson, I.R. and Weigel, D.** (2019) An ultra high-density Arabidopsis thaliana crossover map that refines the influences of structural variation and epigenetic features. *Genetics*, **213**, 771-787.

**Ruan, J. and Li, H.** (2020) Fast and accurate long-read assembly with wtdbg2. *Nature methods*, **17**, 155-158.

**Saintenac, C., Faure, S., Remay, A., Choulet, F., Ravel, C., Paux, E., Balfourier, F., Feuillet, C. and Sourdille, P.** (2011) Variation in crossover rates across a 3-Mb contig of bread wheat (Triticum aestivum) reveals the presence of a meiotic recombination hotspot. *Chromosoma*, **120**, 185-198.

**Sanchez-Moran, E., Armstrong, S., Santos, J., Franklin, F. and Jones, G.** (2002) Variation in chiasma frequency among eight accessions of Arabidopsis thaliana. *Genetics*, **162**, 1415-1422.

**Sanger, F., Nicklen, S. and Coulson, A.R.** (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences*, **74**, 5463-5467.

**SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y. and Bennetzen, J.L.** (1998) The paleontology of intergene retrotransposons of maize. *Nature genetics*, **20**, 43-45.

**Sardell, J.M. and Kirkpatrick, M.** (2020) Sex differences in the recombination landscape. *The American Naturalist*, **195**, 361-379.

**Schadt, E.E., Turner, S. and Kasarskis, A.** (2010) A window into third-generation sequencing. *Human molecular genetics*, **19**, R227-R240.

**Schmidt, M.H.-W., Vogel, A., Denton, A.K., Istace, B., Wormit, A., van de Geest, H., Bolger, M.E., Alseekh, S., Maß, J. and Pfaff, C.** (2017) De novo assembly of a new Solanum pennellii accession using nanopore sequencing. *The Plant Cell*, **29**, 2336-2348.

**Schnable, J.C., Springer, N.M. and Freeling, M.** (2011) Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proceedings of the National Academy of Sciences*, **108**, 4069-4074.

**Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L. and Graves, T.A.** (2009) The B73 maize genome: complexity, diversity, and dynamics. *science*, **326**, 1112-1115.

**Sedlazeck, F.J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., Von Haeseler, A. and Schatz, M.C.** (2018) Accurate detection of complex structural variations using single-molecule sequencing. *Nature methods*, **15**, 461-468.

**Séguéla-Arnaud, M., Choinard, S., Larchevêque, C., Girard, C., Froger, N., Crismani, W. and Mercier, R.** (2017) RMI1 and TOP3α limit meiotic CO formation through their C-terminal domains. *Nucleic acids research*, **45**, 1860-1871.

**Séguéla-Arnaud, M., Crismani, W., Larchevêque, C., Mazel, J., Froger, N., Choinard, S., Lemhemdi, A., Macaisne, N., Van Leene, J. and Gevaert, K.** (2015) Multiple mechanisms limit meiotic crossovers: TOP3α and two BLM homologs antagonize crossovers in parallel to FANCM. *Proceedings of the National Academy of Sciences*, **112**, 4713-4718.

**Senchina, D.S., Alvarez, I., Cronn, R.C., Liu, B., Rong, J., Noyes, R.D., Paterson, A.H., Wing, R.A., Wilkins, T.A. and Wendel, J.F.** (2003) Rate variation among nuclear genes and the age of polyploidy in Gossypium. *Molecular biology and evolution*, **20**, 633-643.

**Serra, H., Choi, K., Zhao, X., Blackwell, A.R., Kim, J. and Henderson, I.R.** (2018a) Interhomolog polymorphism shapes meiotic crossover within the Arabidopsis RAC1 and RPP13 disease resistance genes. *PLoS genetics*, **14**, e1007843.

**Serra, H., Lambing, C., Griffin, C.H., Topp, S.D., Nageswaran, D.C., Underwood, C.J., Ziolkowski, P.A., Séguéla-Arnaud, M., Fernandes, J.B. and Mercier, R.** (2018b) Massive crossover elevation via combination of HEI10 and recq4a recq4b during Arabidopsis meiosis. *Proceedings of the National Academy of Sciences*, **115**, 2437-2442.

**Shafin, K., Pesout, T., Lorig-Roach, R., Haukness, M., Olsen, H.E., Bosworth, C., Armstrong, J., Tigyi, K., Maurer, N. and Koren, S.** (2020) Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nature biotechnology*, **38**, 1044-1053.

**Sheikhizadeh, S., Schranz, M.E., Akdel, M., de Ridder, D. and Smit, S.** (2016) PanTools: representation, storage and exploration of pan-genomic data. *Bioinformatics*, **32**, i487-i493.

**Shen, C., Du, H., Chen, Z., Lu, H., Zhu, F., Chen, H., Meng, X., Liu, Q., Liu, P. and Zheng, L.** (2020) The chromosome-level genome sequence of the autotetraploid alfalfa and resequencing of core germplasms provide genomic resources for alfalfa research. *Molecular plant*, **13**, 1250-1261.

**Shendure, J., Balasubramanian, S., Church, G.M., Gilbert, W., Rogers, J., Schloss, J.A. and Waterston, R.H.** (2017) DNA sequencing at 40: past, present and future. *Nature*, **550**, 345-353.

**Sherman, R.M. and Salzberg, S.L.** (2020) Pan-genomics in the human genome era. *Nature Reviews Genetics*, **21**, 243-254.

**Shi, D., Wu, J., Tang, H., Yin, H., Wang, H., Wang, R., Wang, R., Qian, M., Wu, J. and Qi, K.** (2019) Single-pollen-cell sequencing for gamete-based phased diploid genome assembly in plants. *Genome research*, **29**, 1889-1899.

**Shi, J., Tian, Z., Lai, J. and Huang, X.** (2022) Plant pan-genomics and its applications. *Molecular Plant*.

**Shilo, S., Melamed-Bessudo, C., Dorone, Y., Barkai, N. and Levy, A.A.** (2015) DNA crossover motifs associated with epigenetic modifications delineate open chromatin regions in Arabidopsis. *The Plant Cell*, **27**, 2427-2436.

**Singh, H.P., Raigar, O.P. and Chahota, R.K.** (2022) Estimation of genetic diversity and its exploitation in plant breeding. *The Botanical Review*, **88**, 413-435.

**Snogerup, S., Gustafsson, M. and Von Bothmer, R.** (1990) Brassica sect. Brassica (Brassicaceae) I. Taxonomy and variation. *Willdenowia*, 271-365.

**Sohn, J.-i. and Nam, J.-W.** (2018) The present and future of de novo whole-genome assembly. *Briefings in bioinformatics*, **19**, 23-40.

**Sønderby, I.E., Geu-Flores, F. and Halkier, B.A.** (2010) Biosynthesis of glucosinolates–gene discovery and beyond. *Trends in plant science*, **15**, 283-290.

**Song, J.-M., Guan, Z., Hu, J., Guo, C., Yang, Z., Wang, S., Liu, D., Wang, B., Lu, S. and Zhou, R.** (2020) Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of Brassica napus. *Nature Plants*, **6**, 34-45.

**Song, J.-M., Xie, W.-Z., Wang, S., Guo, Y.-X., Koo, D.-H., Kudrna, D., Gong, C., Huang, Y., Feng, J.-W. and Zhang, W.** (2021a) Two gap-free reference genomes and a global view of the centromere architecture in rice. *Molecular Plant*, **14**, 1757-1767.

**Song, X., Wei, Y., Xiao, D., Gong, K., Sun, P., Ren, Y., Yuan, J., Wu, T., Yang, Q. and Li, X.** (2021b) Brassica carinata genome characterization clarifies U's triangle model of evolution and polyploidy in Brassica. *Plant Physiology*, **186**, 388-406.

**Spencer, C.C.A., Deloukas, P., Hunt, S., Mullikin, J., Myers, S., Silverman, B., Donnelly, P., Bentley, D. and McVean, G.** (2006) The influence of recombination on human genetic diversity. *PLoS genetics*, **2**, e148.

**Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S. and Morgenstern, B.** (2006) AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic acids research*, **34**, W435-W439.

**Stansell, Z., Hyma, K., Fresnedo-Ramírez, J., Sun, Q., Mitchell, S., Björkman, T. and Hua, J.** (2018) Genotyping-by-sequencing of Brassica oleracea vegetables reveals unique phylogenetic patterns, population structure and domestication footprints. *Horticulture research*, **5**, 1-10.

**Stephenson, P., Baker, D., Girin, T., Perez, A., Amoah, S., King, G.J. and Østergaard, L.** (2010) A rich TILLING resource for studying gene function in Brassica rapa. *BMC plant biology*, **10**, 1-10.

**Stevens, K.A., Wegrzyn, J.L., Zimin, A., Puiu, D., Crepeau, M., Cardeno, C., Paul, R., Gonzalez-Ibeas, D., Koriabine, M. and Holtz-Morris, A.E.** (2016) Sequence of the sugar pine megagenome. *Genetics*, **204**, 1613-1626.

**Sturtevant, A.H.** (1915) The behavior of the chromosomes as studied through linkage. *Zeitschrift für induktive Abstammungs-und Vererbungslehre*, **13**, 234-287.

**Su, W., Ou, S., Hufford, M.B. and Peterson, T.** (2021) A Tutorial of EDTA: Extensive De Novo TE Annotator. *Plant Transposable Elements*, 55-67.

**Sun, D., Wang, C., Zhang, X., Zhang, W., Jiang, H., Yao, X., Liu, L., Wen, Z., Niu, G. and Shan, X.** (2019a) Draft genome sequence of cauliflower (Brassica oleracea L. var. botrytis) provides new insights into the C genome in Brassica species. *Horticulture research*, **6**, 1-11.

**Sun, H., Jiao, W.-B., Krause, K., Campoy, J.A., Goel, M., Folz-Donahue, K., Kukat, C., Huettel, B. and Schneeberger, K.** (2022a) Chromosome-scale and haplotype-resolved genome assembly of a tetraploid potato cultivar. *Nature genetics*, **54**, 342-348.

**Sun, H., Rowan, B.A., Flood, P.J., Brandt, R., Fuss, J., Hancock, A.M., Michelmore, R.W., Huettel, B. and Schneeberger, K.** (2019b) Linked-read sequencing of gametes allows efficient genome-wide analysis of meiotic recombination. *Nature communications*, **10**, 1-9.

**Sun, S., Zhou, Y., Chen, J., Shi, J., Zhao, H., Zhao, H., Song, W., Zhang, M., Cui, Y. and Dong, X.** (2018) Extensive intraspecific gene order and gene structural variations between Mo17 and other maize genomes. *Nature genetics*, **50**, 1289-1295.

**Sun, X., Jiao, C., Schwaninger, H., Chao, C.T., Ma, Y., Duan, N., Khan, A., Ban, S., Xu, K. and Cheng, L.** (2020) Phased diploid genome assemblies and pan-genomes provide insights into the genetic history of apple domestication. *Nature genetics*, **52**, 1423-1432.

**Sun, X., Li, X., Lu, Y., Wang, S., Zhang, X., Zhang, K., Su, X., Liu, M., Feng, D. and Luo, S.** (2022b) Construction of a high-density mutant population of Chinese cabbage facilitates the genetic dissection of agronomic traits. *Molecular Plant*, **15**, 913-924.

**Swagatika, S. and Tomar, R.** (2016) Modulation of epigenetics by environmental toxic molecules. *Advances in Molecular Toxicology*, **10**, 361-389.

**Szostak, J.W., Orr-Weaver, T.L., Rothstein, R.J. and Stahl, F.W.** (1983) The double-strand-break repair model for recombination. *Cell*, **33**, 25-35.

**Talavera, G. and Castresana, J.** (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic biology*, **56**, 564-577.

**Tang, D., Jia, Y., Zhang, J., Li, H., Cheng, L., Wang, P., Bao, Z., Liu, Z., Feng, S. and Zhu, X.** (2022) Genome evolution and diversity of wild and cultivated potatoes. *Nature*, 1-7.

**Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., Angiuoli, S.V., Crabtree, J., Jones, A.L. and Durkin, A.S.** (2005) Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences*, **102**, 13950-13955.

**Textor, S., De Kraker, J.-W., Hause, B., Gershenzon, J. and Tokuhisa, J.G.** (2007) MAM3 catalyzes the formation of all aliphatic glucosinolate chain lengths in Arabidopsis. *Plant Physiology*, **144**, 60-71.

**Town, C.D., Cheung, F., Maiti, R., Crabtree, J., Haas, B.J., Wortman, J.R., Hine, E.E., Althoff, R., Arbogast, T.S. and Tallon, L.J.** (2006) Comparative genomics of Brassica oleracea and Arabidopsis thaliana reveal gene loss, fragmentation, and dispersal after polyploidy. *The Plant Cell*, **18**, 1348-1359.

**Tribulato, A., Donzella, E., Sdouga, D., Lopes, V. and Branca, F.** (2017) Bio-morphological characterization of Mediterranean wild and cultivated Brassica species. In *VII International Symposium on Brassicas 1202*, pp. 9-16.

**Tripathi, M. and Mishra, A.** (2007) Glucosinolates in animal nutrition: A review. *Animal feed science and technology*, **132**, 1-27.

**Truong, H.T., Ramos, A.M., Yalcin, F., de Ruiter, M., van der Poel, H.J., Huvenaars, K.H., Hogers, R.C., van Enckevort, L.J., Janssen, A. and van Orsouw, N.J.** (2012) Sequence-based genotyping for marker discovery and co-dominant scoring in germplasm and populations. *PloS one*, **7**, e37565.

**Van Dam, N.M., Tytgat, T.O. and Kirkegaard, J.A.** (2009) Root and shoot glucosinolates: a comparison of their diversity, function and interactions in natural and managed ecosystems. *Phytochemistry Reviews*, **8**, 171-186.

**van Dijk, A.D.J., Kootstra, G., Kruijer, W. and de Ridder, D.** (2021) Machine learning in plant science and plant breeding. *Iscience*, **24**, 101890.

**Van Dijk, E.L., Jaszczyszyn, Y., Naquin, D. and Thermes, C.** (2018) The third revolution in sequencing technology. *Trends in Genetics*, **34**, 666-681.

**van Eck, H.J., Oortwijn, M.E., Terpstra, I.R., van Lieshout, N.H., van der Knaap, E., Willemsen, J.H. and Bachem, C.W.** (2022) Engineering of tuber shape in potato (Solanum tuberosum) with marker assisted breeding or genetic modification using StOFP20.

**van Hintum, T.J., van de Wiel, C.C., Visser, D.L., van Treuren, R. and Vosman, B.** (2007) The distribution of genetic diversity in a Brassica oleracea gene bank collection related to the effects on diversity of regeneration, as measured with AFLPs. *Theor Appl Genet*, **114**, 777-786.

**Varet, H., Brillet-Guéguen, L., Coppée, J.-Y. and Dillies, M.-A.** (2016) SARTools: a DESeq2- and EdgeR-based R pipeline for comprehensive differential analysis of RNA-Seq data. *PloS one*, **11**, e0157022.

**Vaser, R., Sović, I., Nagarajan, N. and Šikić, M.** (2017) Fast and accurate de novo genome assembly from long uncorrected reads. *Genome research*, **27**, 737-746.

**Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A. and Holt, R.A.** (2001) The sequence of the human genome. *science*, **291**, 1304-1351.

**Verkerk, R., Schreiner, M., Krumbein, A., Ciska, E., Holst, B., Rowland, I., De Schrijver, R., Hansen, M., Gerhäuser, C. and Mithen, R.** (2009) Glucosinolates in Brassica vegetables: the influence of the food supply chain on intake, bioavailability and human health. *Molecular nutrition & food research*, **53**, S219-S219.

**Voorrips, L., Goldbohm, R., van Poppel, G., Sturmans, F., Hermus, R. and Van Den Brandt, P.** (2000) Vegetable and fruit consumption and risks of colon and rectal cancer in a prospective cohort study The Netherlands Cohort Study on Diet and Cancer. *American Journal of Epidemiology*, **152**, 1081-1092.

**Vos, P., Hogers, R., Bleeker, M., Reijans, M., Lee, T.v.d., Hornes, M., Friters, A., Pot, J., Paleman, J. and Kuiper, M.** (1995) AFLP: a new technique for DNA fingerprinting. *Nucleic acids research*, **23**, 4407-4414.

**Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C.A., Zeng, Q., Wortman, J. and Young, S.K.** (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS one*, **9**, e112963.

**Wang, H., Wu, J., Sun, S., Liu, B., Cheng, F., Sun, R. and Wang, X.** (2011a) Glucosinolate biosynthetic genes in Brassica rapa. *Gene*, **487**, 135-142.

**Wang, J., Gu, H., Yu, H., Zhao, Z., Sheng, X. and Zhang, X.** (2012) Genotypic variation of glucosinolates in broccoli (Brassica oleracea var. italica) florets from China. *Food chemistry*, **133**, 735-741.

**Wang, J., Tian, L., Lee, H.-S., Wei, N.E., Jiang, H., Watson, B., Madlung, A., Osborn, T.C., Doerge, R. and Comai, L.** (2006) Genomewide nonadditive gene regulation in Arabidopsis allotetraploids. *Genetics*, **172**, 507-517.

**Wang, T., Antonacci-Fulton, L., Howe, K., Lawson, H.A., Lucas, J.K., Phillippy, A.M., Popejoy, A.B., Asri, M., Carson, C. and Chaisson, M.J.** (2022) The Human Pangenome Project: a global resource to map genomic diversity. *Nature*, **604**, 437-446.

**Wang, X., Wang, H., Wang, J., Sun, R., Wu, J., Liu, S., Bai, Y., Mun, J.-H., Bancroft, I. and Cheng, F.** (2011b) The genome of the mesopolyploid crop species Brassica rapa. *Nature genetics*, **43**, 1035-1039.

**Wang, Y. and Copenhaver, G.P.** (2018) Meiotic recombination: mixing it up in plants. *Annual Review of Plant Biology*, **69**, 577-609.

**Warwick, S.I. and Sauder, C.A.** (2005) Phylogeny of tribe Brassiceae (Brassicaceae) based on chloroplast restriction site polymorphisms and nuclear ribosomal internal transcribed spacer and chloroplast trn L intron sequences. *Canadian Journal of Botany*, **83**, 467-483.

**Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M. and Barton, G.J.** (2009) Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189-1191.

**Waterhouse, R.M., Seppey, M., Simão, F.A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E.V. and Zdobnov, E.M.** (2018) BUSCO applications from quality assessments to gene prediction and phylogenomics. *Molecular biology and evolution*, **35**, 543-548.

**Wattenberg, L.W.** (1977) Inhibition of carcinogenic effects of polycyclic hydrocarbons by benzyl isothiocyanate and related compounds. *Journal of the National Cancer Institute*, **58**, 395-398.

**Weir, B.S. and Cockerham, C.C.** (1984) Estimating F-statistics for the analysis of population structure. *evolution*, 1358-1370.

**Wijnker, E. and de Jong, H.** (2008) Managing meiotic recombination in plant breeding. *Trends in plant science*, **13**, 640-646.

**Wijnker, E., James, G.V., Ding, J., Becker, F., Klasen, J.R., Rawat, V., Rowan, B.A., de Jong, D.F., de Snoo, C.B. and Zapata, L.** (2013) The genomic landscape of meiotic crossovers and gene conversions in Arabidopsis thaliana. *Elife*, **2**, e01426.

**Williams, J.G., Kubelik, A.R., Livak, K.J., Rafalski, J.A. and Tingey, S.V.** (1990) DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic acids research*, **18**, 6531-6535.

**Wittstock, U., Agerbirk, N., Stauber, E.J., Olsen, C.E., Hippler, M., Mitchell-Olds, T., Gershenzon, J. and Vogel, H.** (2004) Successful herbivore attack due to metabolic diversion of a plant chemical defense. *Proceedings of the National Academy of Sciences*, **101**, 4859-4864.

**Wu, G., Rossidivito, G., Hu, T., Berlyand, Y. and Poethig, R.S.** (2015) Traffic lines: new tools for genetic analysis in Arabidopsis thaliana. *Genetics*, **200**, 35-45.

**Wu, J., Mizuno, H., Hayashi-Tsugane, M., Ito, Y., Chiden, Y., Fujisawa, M., Katagiri, S., Saji, S., Yoshiki, S. and Karasawa, W.** (2003) Physical maps and recombination frequency of six rice chromosomes. *The Plant Journal*, **36**, 720-730.

**Wu, Q., Yang, Y., Vogtmann, E., Wang, J., Han, L., Li, H. and Xiang, Y.** (2013) Cruciferous vegetables intake and the risk of colorectal cancer: a meta-analysis of observational studies. *Annals of oncology*, **24**, 1079-1087.

**Wu, S., Zhang, B., Keyhaninejad, N., Rodríguez, G.R., Kim, H.J., Chakrabarti, M., Illa-Berenguer, E., Taitano, N.K., Gonzalo, M.J. and Díaz, A.** (2018) A common genetic mechanism underlies morphological diversity in fruits and other plant organs. *Nature communications*, **9**, 4734.

**Xiao, C.-L., Chen, Y., Xie, S.-Q., Chen, K.-N., Wang, Y., Han, Y., Luo, F. and Xie, Z.** (2017) MECAT: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. *nature methods*, **14**, 1072-1074.

**Xu, X., Liu, X., Ge, S., Jensen, J.D., Hu, F., Li, X., Dong, Y., Gutenkunst, R.N., Fang, L. and Huang, L.** (2012) Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nature biotechnology*, **30**, 105-111.

**Xu, Z. and Wang, H.** (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic acids research*, **35**, W265-W268.

**Yang, J., Liu, D., Wang, X., Ji, C., Cheng, F., Liu, B., Hu, Z., Chen, S., Pental, D. and Ju, Y.** (2016) The genome sequence of allopolyploid Brassica juncea and analysis of differential homoeolog gene expression influencing selection. *Nature genetics*, **48**, 1225-1232.

**Yang, X., Zhang, L., Guo, X., Xu, J., Zhang, K., Yang, Y., Yang, Y., Jian, Y., Dong, D. and Huang, S.** (2022) The gap-free potato genome assembly reveals large tandem gene clusters of agronomical importance in highly repeated genomic regions. *Molecular plant*, S1674-2052 (1622) 00445-00442.

**Yang, Y.-W., Lai, K.-N., Tai, P.-Y. and Li, W.-H.** (1999) Rates of nucleotide substitution in angiosperm mitochondrial DNA sequences and dates of divergence between Brassica and other angiosperm lineages. *Journal of Molecular Evolution*, **48**, 597-604.

**Yelina, N.E., Choi, K., Chelysheva, L., Macaulay, M., De Snoo, B., Wijnker, E., Miller, N., Drouaud, J., Grelon, M. and Copenhaver, G.P.** (2012) Epigenetic remodeling of meiotic crossover frequency in Arabidopsis thaliana DNA methyltransferase mutants. *PLoS genetics*, **8**, e1002844.

**Yi, G.-E., Robin, A.H.K., Yang, K., Park, J.-I., Kang, J.-G., Yang, T.-J. and Nou, I.-S.** (2015) Identification and expression analysis of glucosinolate biosynthetic genes and estimation of glucosinolate contents in edible organs of Brassica oleracea subspecies. *Molecules*, **20**, 13089-13111.

**Yu, X., Xiao, J., Chen, S., Yu, Y., Ma, J., Lin, Y., Li, R., Lin, J., Fu, Z. and Zhou, Q.** (2020) Metabolite signatures of diverse Camellia sinensis tea populations. *Nature communications*, **11**, 1-14.

**Zeven, A.** (1996) Sixteenth to eighteenth century depictions of cole crops,(Brassica oleracea L.), turnips (B. Rapa L. cultivar group vegetable turnip) and radish (Raphanus sativus L.) from Flan Ders and the present-day Netherlands. In *ISHS Brassica Symposium-IX Crucifer Genetics Workshop 407*, pp. 29-34.

**Zhang, B., Liu, C., Wang, Y., Yao, X., Wang, F., Wu, J., King, G.J. and Liu, K.** (2015a) Disruption of a CAROTENOID CLEAVAGE DIOXYGENASE 4 gene converts flower colour from white to yellow in Brassica species. *New Phytologist*, **206**, 1513-1526.

**Zhang, J., Liu, Z., Liang, J., Wu, J., Cheng, F. and Wang, X.** (2015b) Three genes encoding AOP2, a protein involved in aliphatic glucosinolate biosynthesis, are differentially expressed in Brassica rapa. *Journal of experimental botany*, **66**, 6205-6218.

**Zhang, J., Zhang, X., Tang, H., Zhang, Q., Hua, X., Ma, X., Zhu, F., Jones, T., Zhu, X. and Bowers, J.** (2018) Allele-defined genome of the autopolyploid sugarcane Saccharum spontaneum L. *Nature genetics*, **50**, 1565-1573.

**Zhang, K., Wang, X. and Cheng, F.** (2019a) Plant polyploidy: origin, evolution, and its influence on crop domestication. *Horticultural Plant Journal*, **5**, 231-239.

**Zhang, P., Zhang, Y., Sun, L., Sinumporn, S., Yang, Z., Sun, B., Xuan, D., Li, Z., Yu, P. and Wu, W.** (2017) The rice AAA-ATPase OsFIGNL1 is essential for male meiosis. *Frontiers in plant science*, 1639.

**Zhang, X., Zhang, S., Zhao, Q., Ming, R. and Tang, H.** (2019b) Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nature plants*, **5**, 833-845.

**Zhang, Z., Li, J., Zhao, X.-Q., Wang, J., Wong, G.K.-S. and Yu, J.** (2006) KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. *Genomics, proteomics & bioinformatics*, **4**, 259-263.

**Zhao, J., Wang, X., Deng, B., Lou, P., Wu, J., Sun, R., Xu, Z., Vromans, J., Koornneef, M. and Bonnema, G.** (2005) Genetic relationships within Brassica rapa as inferred from AFLP fingerprints. *Theoretical and applied genetics*, **110**, 1301-1314.

**Zhao, Q., Feng, Q., Lu, H., Li, Y., Wang, A., Tian, Q., Zhan, Q., Lu, Y., Zhang, L. and Huang, T.** (2018) Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nature genetics*, **50**, 278-284.

**Zhou, C., Olukolu, B., Gemenet, D.C., Wu, S., Gruneberg, W., Cao, M.D., Fei, Z., Zeng, Z.-B., George, A.W. and Khan, A.** (2020a) Assembly of whole-chromosome pseudomolecules for polyploid plant genomes using outbred mapping populations. *Nature Genetics*, **52**, 1256-1264.

**Zhou, Q., Tang, D., Huang, W., Yang, Z., Zhang, Y., Hamilton, J.P., Visser, R.G., Bachem, C.W., Robin Buell, C. and Zhang, Z.** (2020b) Haplotype-resolved genome analyses of a heterozygous diploid potato. *Nature genetics*, **52**, 1018-1023.

**Zhou, Y., Zhang, Z., Bao, Z., Li, H., Lyu, Y., Zan, Y., Wu, Y., Cheng, L., Fang, Y. and Wu, K.** (2022) Graph pangenome captures missing heritability and empowers tomato breeding. *Nature*, 1-8.

**Zhu, L., Fernández-Jiménez, N., Szymanska-Lejman, M., Pelé, A., Underwood, C.J., Serra, H., Lambing, C., Dluzewska, J., Bieluszewski, T. and Pradillo, M.** (2021) Natural variation identifies SNI1, the SMC5/6 component, as a modifier of meiotic crossover in Arabidopsis. *Proceedings of the National Academy of Sciences*, **118**.

**Zickler, D. and Kleckner, N.** (2016) A few of our favorite things: Pairing, the bouquet, crossover interference and evolution of meiosis. In *Seminars in cell & developmental biology*: Elsevier, pp. 135-148.

**Ziolkowski, P.A., Berchowitz, L.E., Lambing, C., Yelina, N.E., Zhao, X., Kelly, K.A., Choi, K., Ziolkowska, L., June, V. and Sanchez-Moran, E.** (2015) Juxtaposition of heterozygous and homozygous regions causes reciprocal crossover remodelling via interference during Arabidopsis meiosis. *Elife*, **4**, e03708.

**Ziolkowski, P.A., Underwood, C.J., Lambing, C., Martinez-Garcia, M., Lawrence, E.J., Ziolkowska, L., Griffin, C., Choi, K., Franklin, F.C.H. and Martienssen, R.A.** (2017) Natural variation and dosage of the HEI10 meiotic E3 ligase control Arabidopsis crossover recombination. *Genes & development*, **31**, 306-317.

# Summary

*Brassica oleracea* includes economically important crops that display enormous phenotypic variation, including vegetables like cabbage, broccoli, cauliflower, kohlrabi, kales and etc. Artificial selection during domestication and breeding activities resulted in these highly diverse crops. The domestication history of this species has not been clarified, despite several genetic studies and investigations of ancient literature. Like other *Brassica* crops, *B. oleracea* also varies extensively in content and composition of glucosinolates (GSLs), a class of secondary plant metabolites, which play important roles in defense against pathogens and pests. They may affect taste and have both anti-nutritional properties and health benefits for consumers. However, the genetic regulation of GSL variation remains elusive in *B. oleracea*. A high-quality reference genome can provide a solid basis for comparative and functional genomic studies and breeding. The first two draft *B. oleracea* genomes released in 2014 were not yet of sufficient quality. Moreover, resequencing data analysis and a pan-genome study based on short reads mapping and assembly approach provided ample evidence for structural variations (SVs) between genomes of different *B. oleracea* morphotypes. Therefore, the need for more high-quality *de novo* reference genomes representing diverse *B. oleracea* morphotypes became evident. As one of the main sources for generating genetic diversity in sexual organisms, meiotic recombination is essential in plant breeding to combine beneficial traits. This process also assures proper segregation of parental chromosomes during meiosis, thus guaranteeing genome integrity and stability. As meiotic recombination is tightly regulated, identifying factors influencing rate and distribution of meiotic crossovers is of major importance. However, high-resolution recombination maps are sparse in the *Brassica* genus and knowledge about intraspecific variation and sex difference is lacking. The aim of this thesis is to explore genomic and genetic features of the very diverse *B. oleracea* species and to provide insights into the research questions as mentioned.

In **Chapter 2**, we generated a germplasm collection of 912 globally distributed accessions representing ten morphotypes of *B. oleracea*, wild *B. oleracea* accessions and nine related C9 *Brassica* species. We used 14,152 high-quality SNP markers to study the genetic diversity, genealogical relationships, population structure and domestication history of *B. oleracea* morphotypes. We showed that genetic diversity reduced from old landraces to modern hybrids for almost all morphotypes. Cauliflower is the least diverse morphotype showing strong genetic differentiation with other morphotypes except broccoli. We provided evidence for two domestication

lineages, the ‘leafy head’ lineage (LHL; cabbages, collards and ornamentals) and the ‘arrested inflorescence’ lineage (AIL; cauliflower and broccoli). We found evidence for a far eastern origin of the AIL and LHL clades in both the presence of diversified kales and in the first-branching cabbage clade with accessions from the middle-east that seems ancestral. We proposed a scenario in which ancient divergent kale lineages have led to AIL and LHL. Together with archaeological and literature evidence, we also hypothesized that cabbages and cauliflowers stem from kales introduced from Western Europe to the middle-east, possibly transported with the tin-trade routes in the Bronze age, to be re-introduced later into Europe.

In **Chapter 3**, we *de novo* assembled chromosome-scale reference genomes for five different *B. oleracea* morphotypes, namely broccoli, cauliflower, kale, kohlrabi and white cabbage, by combining long reads (ONT), Bionano DLS optical maps and Illumina short reads. These high-quality genome assemblies provide valuable genomic resources for fundamental genomic studies and *B. oleracea* breeding programs. We revealed both highly syntenic relationships and extensive SVs among the five genomes through comparative analyses. Together with four previously published high-quality genomes, we revealed the composition and features of a *B. oleracea* pan-genome via a *de novo* assembly approach. Additionally, we investigated intact LTR-RTs in the pan-genome of *B. rapa* and *B. oleracea*, and revealed their different evolutionary dynamics. Furthermore, using a pan-genome approach, we studied the impacts of the whole genome triplication (WGT) event and subgenome dominance on intraspecific diversification of *B. oleracea*. We also compared evolutionary patterns regarding biased WGT-derived gene loss between *B. rapa* and *B. oleracea* and observed faster WGT-derived gene loss in *B. rapa* than in *B. oleracea* before intraspecific diversification. In addition, we revealed continuing gene loss bias during intraspecific diversification for both species.

In **Chapter 4**, we investigated recombination variation among ten different genetic backgrounds (crosses/cross combinations) in *B. oleracea* and studied the sex differences for each cross. To do so, we constructed five large Four-way-Cross (FwC) populations with each two F1s being reciprocally crossed per population. Parents are highly diverse inbred lines representing major crops and are the same plant materials used for genome sequencing in **Chapter 3**. We sequenced a total of 1,248 progeny genomes and harvested ~4.56T Illumina data for the five FwC populations. From this fine mapping, we identified a total of 15,353 crossovers (COs) and characterized fine-scale resolution recombination landscapes for all ten female and male crosses. We revealed fairly similar megabase-scale recombination landscapes among all cross

combinations and between the sexes, and provided evidence that these landscapes are largely independent of sequence divergence. We evidenced strong influence of gene density and large SVs (generated in **Chapter 3**) on CO formation in *B. oleracea*. Moreover, we found extensive variations in CO number depending on the direction and combination of the initial parents crossed with, for the first time, a striking interdependency between these factors.

In **Chapter 5**, we analyzed the relative abundance of 23 different GSL structures in both roots, stems and the edible parts of five different *B. oleracea* morphotypes, the same plant materials used for genome sequencing in **Chapter 3**. We generated a comprehensive list of GSL related genes in the five corresponding high-quality *B. oleracea* genome assemblies (**Chapter 3**) based on sequence homology with Arabidopsis. We also analyzed expression levels of these GSL related genes using mRNA-Seq data. We revealed strong variation in terms of relative abundance and composition of GSLs, and GSL related gene expression levels among different tissues and different morphotypes. We found significant correlations between the abundance of 23 GSLs and expression level of 109 related genes. We also present interesting observations, including a non-functional *AOP2* in broccoli related to the loss of the conserved 2OG-FeII_Oxy domain, explaining the specific accumulation of health-promoting 4-methylsulfinylbutyl and 5-methylsulfinylpentyl GSLs in broccoli, and transposable elements (TEs) insertion activities in one paralog of the *MAM3* gene in three out of the five genomes causing long and repeat-rich introns, respectively.

In **Chapter 6**, I summarized and discussed the work of previous chapters in relation to published studies. Besides, I discussed how to generate high-quality plant genome assemblies using the state-of-the-art technologies and mentioned the surging plant pan-genomes, which are expected to be the new reference in the future.

## Acknowledgements

My PhD journey has finally come to an end. Completing a PhD has been a rollercoaster ride, full of ups and downs. As I look back, moments of excitement and breakthroughs, but also moments of frustration and setbacks flood my mind. Without the assistance, support, encouragement, and companionship of so many people, I could not have completed this long and challenging journey. I want to use this opportunity to thank all the people who helped me along the way to becoming a "Dr". Please, don't be offended if your name is not on this list. I cannot express in words how much I appreciate each and every one of you.

First of all, I would like to express my sincere gratitude to my promoter and daily supervisor, **Guusje**. Dear **Guusje**, I still remember the day we met in person for my interview in October 2017 in Beijing. I was extremely nervous at the beginning of the interview because it was my first time having an interview in English, and it was far from good. However, your kind and reassuring demeanour quickly put me at ease, and our conversation flowed smoothly. Shortly after you returned to the Netherlands, you informed me that you were positive about my performance and offered me the opportunity to join your research group to pursue a PhD at WUR. I cannot thank you enough for trusting in my abilities and providing me with this incredible opportunity. Over the past ~4.5 years, I have been incredibly fortunate to have you as my supervisor. Your guidance, support, assistance and encouragement have been invaluable, and I have learned so much from you. You are a responsible and patient supervisor, and your positive attitude towards both work and life has been an inspiration to me. Your expertise in *Brassica* is truly impressive, and I have enjoyed a lot our (bi)weekly meetings where we mainly discussed my results, and you offered insightful suggestions and ideas. Your training was very helpful in making me an independent and critical researcher. During my writing period, your timely revisions and constructive feedback have been immensely helpful. I'm aware that I may be your most "pushing" PhD, often sending you emails or WhatsApp messages in the evenings or on weekends. I'm sorry if this has caused you any inconvenience. I would also like to express my gratitude for the many times you invited us to your home for dinners and parties. These occasions were always enjoyable and provided a much-needed break from work.

I would also like to thank my co-promoter, **Richard** (Finkers). Dear **Richard**, the first time I met you was when I hurried into your office to request an account for the PBR server. I hope my approach did not annoy you. Several months later, Guusje invited you to be my co-promoter, and I had the opportunity to get to know you better. You are truly an expert in bioinformatics and genetics. I appreciated your input and suggestions on my project, particularly regarding the technical aspects of bioinformatics. You also introduced me to colleagues from the PBR bioinformatics team and invited me to attend the Tuesday seminar, where we had wonderful discussions about our work and other interesting topics. During the Covid pandemic, I was feeling a bit overwhelmed, but you took the time to invite me for a chat in front of Radix. Thanks a lot for your moral support and encouragement. I was sad when you informed me in your office that you were leaving WUR to take the new position in a company. I missed you after your departure. I'm thrilled that you can be present on stage with me during my defence as my co-promoter. Thank you for all your contributions to my PhD journey.

My appreciation also goes to my colleagues and collaborators who made remarkable contributions to my project. **Johan**, you did great job on wet-lab experiments for my project and are a co-author of all my research chapters. Without your work in taking care of the plants and extracting high-quality DNAs and RNAs, I could not have produced such nice results. I really enjoyed working with you, and I appreciate your time and patience in answering my questions or helping me whenever I came to you. Also, thank you for your generous help in improving my spoken English during my IELTs exam preparation. **Alexandre**, thank you very much for your invaluable contribution to our discussions on recombination. I'm sorry that you had to work late nights and early mornings several times to revise our manuscript. You encouraged me a lot, especially when our manuscript got rejected for the first time. Anyway, we ended up with a very nice publication. It was a pity that our time together in Wageningen was short, only two months. I could have learned so much more from you if you had stayed at Wageningen longer. **Ric**, thank you for your contribution to the metabolite chapter. Your extensive revisions and thoughtful comments greatly improved our manuscript. I hope we can publish this chapter soon. **Freek**, I'm impressed by your expertise in evolution and phylogeny, and your critical approach to research. I'm very happy that with your help, we published my first paper during my PhD. Thank you very much for your contributions. I also want to thank the members from our consortium meeting, **Gerard**, **Henk**, **Ilja**, **Jan-Dick** and **Saulo** for their in-kind contributions and suggestions regarding my project.

To my external supervisor **Francine**, thank you for caring about the progress of my research and providing valuable guidance in planning for my future career. Our conversations were always a great pleasure, and I appreciated the nice coffee and desserts you kindly provided. I also greatly appreciate your support during the early stages of my PhD when I encountered difficulties submitting a sufficient language certificate.

My gratitude goes further to both the current and previous members of the PBR Growth and Develop group: **Alejandra**, **Beiyu**, **Celia**, **Christian**, **Chunmei**, **Csaba**, **Damian**, **Ernst**, **Hao** (Qian), **Hongbo**, **Jan-Kees**, **Jorge**, **Li** (Shi), **Lorena**, **Marian**, **Natalie**, **Niccolo**, **Ning** (Guo), **Sara**, **Tianpeng**, **Wei** (Sun), **Xiaoxue**, **Xulan** and **Zihan**. Our scientific discussions, coffee breaks and shared activities have been some of the most enjoyable moments of my PhD journey, and I appreciate all the help and support I have received from this amazing group of individuals. **Celia**, thank you for trusting me and inviting me to collaborate on your project. **Damian**, I enjoyed our conversations over meals and appreciate your help in revising my draft propositions. **Ernst**, thank you very much for helping me move. I also enjoyed our walks at the university. **Jorge**, you always encouraged me to relax and do sports. Our chats were always a source of relaxation and enjoyment. **Li** (Shi), you are always very warm-hearted, offering recommendations for a wide variety of things (groceries, food, restaurants, cream, shampoo, clothes, etc). These made my life easier as a lazy men. **Marian**, thank you for organizing several nice lab trips, which were such enjoyable experiences. **Ning** (Guo), many thanks for picking me up from the bus station on the first day I arrived in the Netherlands. **Wei** (Sun), I really enjoyed our lunchtime and coffee break chats and appreciated that you are always generous in sharing your delicious snacks. **Xiaoxue**, I'm grateful for the information you provided about the PhD position in Guusje's group and for your help with my application. **Xulan**, since you moved, you have always offered your spacious apartment as a gathering place for meals and games. I'm fully aware of the amount of work you put into cleaning before and after each gathering. Thank you for all the wonderful memories we have shared in your home. **Zihan**, I met you on the first day when I arrived in the Netherlands. You helped me a lot to settle down at the very beginning. I particularly enjoyed the delicious food you cooked and our travels together. Thank you for making all these nice memories. I would like to express my sincere gratitude once more to my lovely paranymphs, **Celia** and **Xulan,** for your invaluable assistance in organizing my defense.

Thanks to current and previous colleagues from the PBR bioinformatics team: **Brian**, **Danny**, **Edouard**, **Edwin**, **Evangelia**, **Fernanda**, **Marjolein**, **Martijn**, **Matthijs**, **Natascha**, **Nathalie** and **Patrick**. I learned a lot about bioinformatics from you and really enjoyed our Tuesday meetings. **Danny**, thank you for helping me with my mRNA-seq data analysis. **Martijn**, thank you for organizing the nice BBQ at your home. I wish you success in your new position. **Matthijs**, thank you for taking care of the server and ensuring that we could work smoothly.

Also, thanks to the PBR secretaries, **Daniëlle** and **Nicole**, for your efficient and timely handling of all the administrative matters. Thank you, **Richard** (Visser) and **Rene**, for your valuable suggestions that helped us address the issues we confronted during our project. Thank you, **Susan**, for your help with the EPS administrative and educational aspects.

I would also like to thank my office mates: **Bettina**, **Corentin**, **Elisabeth**, **Gabrielle**, **Peihong** and **Xinfang**, for creating a wonderful working environment and for sharing snacks with me. **Corentin**, thank you for all the nice discussions we had about our project in the office and for recommending many informative articles to me.

Thanks to all my fellow PhDs at PBR: **Alejandro**, **Anouk**, **Antonino**, **Aviv**, **Dong** (Zhang), **Eleni**, **Francesco**, **Huayi**, **Jaae**, **Jingjing** (Mao), **Kasper**, **Lampros**, **Laura**, **Marcella**, **Mengjing**, **Sri**, **Tim**, **Xuexue**, **Yerisf**, **Yi** (Wu), **Ying** (Liu), **Yingjie** and **Yanlin**. Your company throughout this long journey has made it much less lonely. **Dong** (Zhang), we started our PhD around the same time. Thank you for inviting me to play various sports (basketball, tennis, ping pong, volleyball, squash) and to your home for nice food you cooked. I hope to see you soon graduate as well. **Huayi**, thank you for preparing delicious meal for us on my first Chinese New Year Eve in the Netherlands.

Thanks to other friends whom I met here in the Netherlands. **Cheng** (Ge), **Si** (Qin), **Xiaotian**, **Yiyuan** and **Ziyang**, thank you for the wonderful memories we shared together, cooking up delicious food, hanging out, playing Nintendo Switch, card games, and more. **Si** (Qin), thank you for delivering Ibuprofen to my room during my Covid quarantine, which helped alleviate my pain and discomfort during those first two challenging days. **Chen** (Liu), **Chunzhe**, **Jianing** (Liu), **Ke** (Chen), **Xiaofei**, **Xue** (Pan), **Yun** (Yu), **Zhuang** (Yang) and **Zulin**, thank you for the wonderful times we spent playing badminton or basketball together. It was always very relaxing to do sports and chat with you guys. **Zulin**, I enjoyed the many times we had our “high calorie dinner” (like pizzas and colas) after our sports activities. And thanks for organizing the self-driving trip to Belgium Luxemburg after obtaining your Dutch driver’s license. **Anna** (Huang), **Chenglong**, **Han** (Wang), **Jiyao**, **Qi** (Wang), **Wei** (Xiong), **Weixuan**, **Wenyu**, **Xiaoning**, **Min** (Xu), **Panpan** (Xu), and **Xiyu**, thank you for the wonderful times we spent together on campus. It was always a great pleasure to see and talk to each of you. **Han** (Wang), thank you for inviting me to your home and treating me to your self-made desserts, especially the delicious Tiramisu.

Thanks to the China Scholarship Council (CSC) for providing me with a grant to pursue my studies in Wageningen. I'm also grateful for my master's supervisors, Prof. **Xiaowu Wang** and Prof. **Feng Cheng**, who have imparted a wealth of knowledge to me in the field of genomics and bioinformatics, laying a solid foundation for my future PhD work. Furthermore, I want to thank **Silong**, **Xiaobo**, **Xu** (Cai) and **Yayan** for their invaluable assistance and engaging discussions during my bioinformatic analysis.

Special thanks to **Qianru**. Thank you for all your encouragement and support, especially during my writing period. I wish you great success with your experiments, and I look forward to celebrating with you next year when you also earn your well-deserved doctorate.

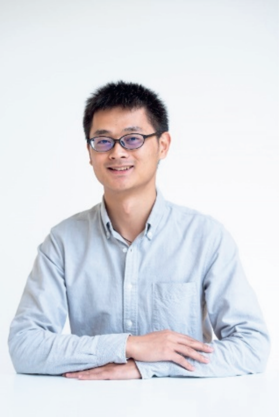
感谢我的父母，这些年一直在外求学，无法在身边陪伴你们，感谢你们的理解和支持。希望你们多注意休息，身体健康，开心快乐。哥哥嫂子，感谢你们这些年陪伴在父母身边，帮忙照顾父母，也感谢你们对我的理解。感谢我所有关心和支持我的亲人和朋友们!

Chengcheng Cai

蔡成成

10 April 2023, Wageningen

## About the author



Chengcheng Cai was born on December 31, 1991, in Suizhou, Hubei province, China. In 2010, he began his undergraduate studies in horticulture at the College of Horticulture, Shenyang Agriculture University, Liaoning, China. After obtaining his BSc degree in 2014, he pursued a master's study in vegetable science at the Institute of Vegetables and Flowers (IVF), Chinese Academy of Agricultural Sciences (CAAS), Beijing, China. He completed his master's thesis entitled "*Brassica rapa* reference genome upgrade and genome evolution analysis" under the supervision of Prof. Xiaowu Wang and Prof. Feng Cheng. In 2017, he obtained his MSc degree and joined Berry Genomics Co. Ltd (Beijing, China) as an bioinformatician. In 2018, he received a fellowship from the China Scholarship Council and moved to the Netherlands to study genomics in crops as a PhD student in Wageningen University & Research. This thesis presents the outcome of his PhD research entitled "Unlocking the secrets of *Brassica oleracea* crops: a genomic journey", which was supervised by Dr Guusje Bonnema (promoter, daily supervisor) and Dr Richard Finkers (co-promoter).

## List of publications

- **Chengcheng Cai#,** Xiaobo Wang#, Bo Liu, Jian Wu, Jianli Liang, Yinan Cui, Feng Cheng, Xiaowu Wang. *Brassica rapa* Genome 2.0: a reference upgrade through sequence re-assembly and gene re-annotation. ***Molecular Plant***, doi: 10.1016/j.molp.2016.11.008, 2017.

- **Chengcheng Cai,** Johan Bucher, Freek T. Bakker and Guusje Bonnema. Evidence for two domestication lineages supporting a middle-eastern origin for *Brassica oleracea* crops from diversified kale populations. ***Horticulture Research***, doi: 10.1093/hr/uhac033, 2022.

- **Chengcheng Cai**, Alexandre Pelé, Johan Bucher, Richard Finkers and Guusje Bonnema. Fine mapping of meiotic crossovers in *Brassica oleracea* reveals patterns and variations depending on direction and combination of crosses. ***The Plant Journal***, doi: 10.1111/tpj.16104, 2023.

- **Chengcheng Cai**, Johan Bucher, Richard Finkers and Guusje Bonnema. Chromosome-scale genome assemblies of five different *Brassica oleracea* morphotypes and the insights on intraspecific diversification. ***bioRxiv***, doi: 10.1101/2022.10.27.514037, 2022.

- **Chengcheng Cai**, Ric C.H. de Vos, Hao Qian, Johan Bucher and Guusje Bonnema. Metabolomic and transcriptomic profiles in diverse *Brassica oleracea* crops provide insights into genetic regulation of glucosinolate variation. (Submission in preparation)

- Lixia Fu#, **Chengcheng Cai#,** Yinan Cui, Jian Wu, Jianli Liang, Feng Cheng, Xiaowu Wang. Pooled Mapping: An Efficient Method of Calling Variations for Population Samples with Low-depth Resequencing Data. ***Molecular Breeding***, doi:10.1007/s11032-016-0476-9, 2016. (#co-first author)

## Education Statement of the Graduate School
## Experimental Plant Sciences

| | |
|---|---|
| **Issued to:** | **Chengcheng Cai** |
| **Date:** | **31 May 2023** |
| **Group:** | **Plant Breeding** |
| **University:** | **Wageningen University & Research** |

| **1) Start-Up Phase** | date | cp |
|---|---|---|
| ► **First presentation of your project** | | |
| A *de novo* sequencing catalogue of structural variations in different *B. oleracea* morphotypes | 03 Jun 2019 | 1.5 |
| ► **Writing or rewriting a project proposal** | | |
| A *de novo* sequencing catalogue of structural variations in different *B. oleracea* morphotypes | Jan-Mar 2019 | 6.0 |
| ► **MSc courses** | | |
| *Subtotal Start-Up Phase* | | 7.5 |

| **2) Scientific Exposure** | date | cp |
|---|---|---|
| ► **EPS PhD days** | | |
| EPS PhD days 'Get2Gether', Soest, NL | 11-12 Feb 2019 | 0.6 |
| EPS PhD days 'Get2Gether', Online | 01-02 Feb 2021 | 0.4 |
| ► **EPS theme symposia** | | |
| EPS Theme 1 Symposium 'Developmental Biology of Plants', Leiden, NL | 31 Jan 2019 | 0.3 |
| EPS Theme 3 Symposium 'Metabolism and Adaptation', Online | 30 Oct 2020 | 0.2 |
| EPS Theme 4 Symposium 'Genome Biology', Online | 11 Dec 2020 | 0.2 |
| EPS Theme 1 Symposium 'Developmental Biology of Plants', Online | 28 Jan 2021 | 0.2 |
| EPS Theme 4 Symposium 'Genome Biology', Online | 17 Jan 2022 | 0.3 |
| ► **Lunteren Days and other national platforms** | | |
| Annual Meeting 'Experimental Plant Sciences', Lunteren, NL | 08-09 Apr 2019 | 0.6 |
| Annual Meeting 'Experimental Plant Sciences', Online | 12-13 Apr 2021 | 0.5 |
| Annual Meeting 'Experimental Plant Sciences', Lunteren, NL | 11-12 Apr 2022 | 0.6 |
| ► **Seminars (series), workshops and symposia** | | |
| Symposium: FAIR Data Science for Green Life Sciences, Wageningen, NL | 12 Dec 2018 | 0.3 |
| Seminar: Maheshi Dassanayake, Multi-Ion Salt Stress Adaptation Explored Using Extremophyte Genomics | 20 May 2019 | 0.1 |
| Inaugural Lecture: Yuling Bai, Plant Breeding: Art rooted in Science, Wageningen, NL | 23 May 2019 | 0.1 |
| PE&RC/EPS workshop: Breeding for Diversity Opportunities and Challenges, Wageningen, NL | 30 Oct 2019 | 0.3 |
| Seminar: Martin van Ittersum, Tripling cerealproduction with minimum emissions by 2050 | 09 Nov 2020 | 0.1 |
| EPS Flying seminars: Stefan Geisen, Challenges as a PhD and postdoc in Science and tips to overcome those | 20 Jan 2021 | 0.1 |
| Seminar: Eske Willerslev, The hunt for our molecular past | 17 Mar 2021 | 0.1 |
| Seminar: Korbinian Schneeberger, The assembly and analysis of a tetraploid potato genome | 02 Nov 2021 | 0.1 |
| Seminar: Xiaoqi Feng, Epigenetic reprogramming in plant germlines | 09 Dec 2021 | 0.1 |
| EPS Mini-symposium: Mendel - 200 years, Wageningen, NL | 08 Jun 2022 | 0.2 |
| Seminar: Sasha Zhefnzkova, Gut microbiome: added value in understanding human health | 04 Jul 2022 | 0.1 |
| UBC Symposium: AI in bioinformatics, Utrecht, NL | 10 Oct 2022 | 0.3 |
| Webinar: Applied Plant Meiosis, Session 1: Recombination and Introgression | 10 Nov 2022 | 0.2 |
| Webinar: Applied Plant Meiosis, Session 3: Polyploidy and Karyotype Stability | 24 Nov 2022 | 0.2 |
| ► **Seminar plus** | | |
| ► **International symposia and congresses** | | |
| PAG (Plant & Animal Genome) XXVIII, San Diego, USA | 11-15 Jan 2020 | 1.5 |
| 7th International Horticulture Research Conference, Online | 01-30 Jul 2020 | 1.5 |
| SOL International Online meeting 2020 | 09-11 Nov 2020 | 0.6 |
| Online 21st EUCARPIA General Congress | 23-26 Aug 2021 | 0.8 |
| ► **Presentations** | | |
| Talk: company consortium meeting | 08 Nov 2018 | 1.0 |
| Poster: 'Genome assembly of five different *Brassica oleracea* morphotypes' - PAG XXVIII | 13 Jan 2020 | 1.0 |
| Poster: 'Evidence for two domestication lineages supporting a middle-eastern origin for *Brassica oleracea* crops from diversified kale popualations' - 21st EUCARPIA | 23-26 Aug 2021 | 1.0 |
| Talk: 'Chromosome-scale genome assemblies of five different *Brassica oleracea* morphotypes and the insights on intraspecific diversification' - Lunteren meeting 2022 | 11 Apr 2022 | 1.0 |
| Talk: company consortium meeting | 07 Feb 2022 | 1.0 |
| ► **Interviews** | | |
| Annual meetings with mentor (Prof. Dr. Francine Govers) | Sep 2019, Apr 2022 | 0.2 |
| ► **Excursions** | | |
| Open Day Bejo Zaden B.V., Warmenhuizen, NL | 26 Sep 2018 | 0.2 |
| EPS Company Visit to Averis, Veendam, NL | 07 Jun 2019 | 0.2 |
| Online EPS Networking Event - Bejo Zaden B.V. | 14 Dec 2020 | 0.2 |
| Online EPS Company Visit to Genetwister Technologies | 22 Mar 2022 | 0.2 |
| Open Day Bejo Zaden B.V., Warmenhuizen, NL | 30 Sep 2022 | 0.2 |
| *Subtotal Scientific Exposure* | | 16.8 |

| **3) In-Depth Studies** | date | cp |
|---|---|---|
| ► **Advanced scientific courses & workshops** | | |
| Open Online Introduction to R Course (Wolfgang Viechtbauer) | 16-18 Sep 2020 | 0.6 |
| PE&RC/WIMEK course: Basic Statistics, Online | Nov-Dec 2020 | 1.5 |
| EPS/ELIXIR course: Gentle hands-on introduction to Python programming, Online | Mar-Apr 2021 | 0.9 |
| EPS course: Bioinformatic Introduction Course, Online | 05-09 Jul 2021 | 1.5 |
| ► **Journal club** | | |
| Participation in Plant Breeding - Literature Discussion Club | 2018-2019 | 0.5 |
| ► **Individual research training** | | |
| *Subtotal In-Depth Studies* | | 5.0 |

# Education statement

| **4) Personal Development** | *date* | *cp* |
|---|---|---|
| ► **General skill training courses** | | |
| EPS introduction course, Wageningen, NL | 11 Jun 2019 | 0.3 |
| EPS workshop: Scientific Paper Writing, Wageningen, NL | 24 Oct 2019 | 0.1 |
| WGS course: Project and Time Management, Wageningen, NL | Nov-Dec 2019 | 1.5 |
| WGS course: Scientific Publishing, Online | 05 Oct 2020 | 0.3 |
| WGS course: Scientific Writing, Online | Feb-Apr 2021 | 1.8 |
| ► **Organisation of meetings, PhD courses or outreach activities** | | |
| ► **Membership of EPS PhD Council** | | |
| *Subtotal Personal Development* | | 4.0 |

| **5) Teaching & Supervision Duties** | *date* | *cp* |
|---|---|---|
| ► **Courses** | | |
| ► **Supervision of BSc/MSc students** | | |
| Supervision MSc major thesis, Hao Qian | May-Oct 2021 | 3.0 |
| *Subtotal Teaching & Supervision Duties* | | 3.0 |

| **TOTAL NUMBER OF CREDIT POINTS*** | **36.3** |
|---|---|

Herewith the Graduate School declares that the PhD candidate has complied with the educational requirements set by the Educational Committee of EPS with a minimum total of 30 ECTS credits.

** A credit represents a normative study load of 28 hours of study.*

The research described in this thesis was financially supported by the Top Consortium for Knowledge and Innovation (TKI) (grant number: 1605-004) and ZonMw (grant number: 435005030), and was co-supported by the breeding companies Bejo and ENZA. Part of the work was supported by TKI project 1309-011 and the plant breeding companies Bejo, ENZA, Rijk Zwaan, Syngenta, Takii, Hazera, and KeyGene.

Chengcheng Cai was sponsored by the China Scholarship Council under grant number 201809110159.

Financial support from Wageningen University for printing this thesis is gratefully acknowledged.

Cover design by Chengcheng Cai and Jingjing

Layout by Chengcheng Cai

Printed by ProefschriftMaken