Extensive crop-wild hybridisation during *Brassica* evolution, and selection during the domestication and diversification of *Brassica* crops

Running title: Hybridisation and domestication in *Brassica*

Jasmine M. Saban[1,*], Anne J. Romero[1], Thomas H. G. Ezard[2] & Mark A. Chapman[1,*]

[1] Biological Sciences, University of Southampton, Life Sciences Building, Highfield Campus, Southampton, SO17 1BJ, UK

[2] Ocean and Earth Science, National Oceanography Centre Southampton, Southampton, SO14 3ZH, UK

* Correspondence: J.M.Saban@soton.ac.uk (JMS); M.Chapman@soton.ac.uk (MAC)

Abstract

Adaptive genetic diversity in crop wild relatives (CWRs) can be exploited to develop improved crops with higher yield and resilience if phylogenetic relationships between crops and their CWRs are resolved. This further allows accurate quantification of genome-wide introgression and determination of regions of the genome under selection. Using broad sampling of CWRs and whole genome sequencing we further demonstrate the relationships among two economically valuable and morphologically diverse *Brassica* crop species, their CWRs and their putative wild progenitors. Complex genetic relationships and extensive genomic introgression between CWRs and *Brassica* crops were revealed. Some wild *B. oleracea* populations have admixed feral origins, some domesticated taxa in both crop species are of hybrid origin, while wild *B. rapa* is genetically indistinct from turnips. The extensive genomic introgression we reveal could result in false identification of selection signatures during domestication using traditional comparative approaches used previously, therefore we adopted a single population approach to study selection during domestication. We used this to explore examples of parallel phenotypic selection in the two crop groups and highlight promising candidate genes for future investigation. Our analysis defines the complex genetic relationships between *Brassica* crops and their diverse CWRs, revealing extensive cross-species gene flow with implications for both crop domestication and evolutionary diversification more generally.

Key words: *Brassica*, domestication, crop wild relatives, introgression, phylogenomics

Introduction

Large crop losses are predicted under future climate change scenarios (Challinor et al., 2014; Mbow et al., 2019), presenting significant challenges to ensuring food security and human health (Mbow et al., 2019; Nelson et al., 2018). Crop domestication generally results in a reduction of genetic diversity because of strong selection and limited population sizes (Gaut, Seymour, Liu, & Zhou, 2018). Crops can therefore lack the genetic variation needed to rapidly adapt to environmental change (Zhang, Mittal, Leamy, Barazani, & Song, 2017). Crop wild relatives (CWRs) may contain adaptive variants that can be exploited for crop improvement through selective breeding and the potential for phenotypic plasticity (Bailey-Serres, Parker, Ainsworth, Oldroyd, & Schroeder, 2019). Indeed, in some crops, natural introgression of adaptive alleles from wild relatives may have already facilitated the cultivation of early domesticates in novel environments (Janzen, Wang, & Hufford, 2019). Understanding phylogenetic relationships between crops and CWRs, and the extent of hybridisation throughout domestication, is vital to determine how evolutionary potential might aid future breeding programmes.

Brassica oleracea L. (including cabbage, Brussels sprouts, Chinese kale, cauliflower, broccoli) and *Brassica rapa* L. (turnips, Chinese cabbage, pak choy, bok choy, yellow sarson among others) are popular vegetables worldwide. Consumption of *Brassica*s is also actively promoted for nutritional benefits because of their high fibre and phytonutrient content (Francisco et al., 2017; Kaur, Kumar, Anil, & Kapoor, 2007). While global consumption of *Brassica* crops is expected to increase, substantial yield losses are predicted due to climate change, pests and diseases (Phophi & Mafongoya, 2017; Rodriguez et al., 2015). *Brassica* wild relatives can provide adaptive genetic variation relevant to *Brassica* crop breeding (Branca & Cartea, 2011), but some are endangered and poorly represented in seed banks (Branca & Tribulato, 2011; Castañeda-Álvarez et al., 2016). Efforts to establish phylogenetic relationships have been challenging: wild *Brassica* species display considerable morphological diversity (Snogerup, Gustafsson, & von Bothmer, 1990; Widen, Andersson, Rao, & Widen, 2002) and combinations of *Brassica* species readily hybridise in controlled crosses (FitzJohn, Armstrong, Newstrom-Lloyd, Wilton, & Cochrane, 2007). Wild populations of *B. oleracea* and *B. rapa* have been identified throughout their predicted native ranges, but, even for these well-studied species, phylogenetic relationships to the crops are not fully understood (Maggioni et al., 2020) and the inferred relationships suggest some 'wild' populations are derived feral populations rather than wild ancestors (Mittell et al., 2020; Mabry et al., 2021; McAlvay et al. 2021).

Several recent analyses have analysed the genetic relationships between domesticated types (e.g. Cheng et al., 2016; Guo et al., 2021; Cai et al., 2022), however, only a few phylogenetic analyses have included wild *Brassica* relatives, and these have used transcriptome, reduced representation or chloroplast DNA sequencing (An et al., 2019; Arias & Pires, 2012; Mabry et al., 2021; McAlvay et al., 2021). The most recent of these analyses have suggested that *B. cretica* is likely the closest wild relative of *B. oleracea*, but samples labelled as *B. cretica* were not monophyletic and some individuals were nested in the domesticated groups (Mabry et al., 2021) raising outstanding questions for several CWRs to determine their true ancestry, hybridisation history and taxonomic groupings. Further, Mabry et al. (2021) demonstrate that putatively wild *B. oleracea* populations are instead feral crop derivatives, and not progenitors (see also Mittell et al., 2020). For *B. rapa*, some wild populations may well be true wild progenitors, while others appear to be feral escapes from cultivation (McAlvay et al. 2021). Both of these most recent analyses (Mabry et al., 2021; McAlvay et al. 2021) indicate that crop-wild hybridisation has occurred in the evolution of some domesticated groups in both species.

Since *B. oleracea* and *B. rapa* are also excellent evolutionary models of convergent evolution due to selection during domestication for parallel phenotypes, selection analyses have compared domesticated populations to identify putative targets of selection (e.g., Cheng et al., 2016). Signatures of selection within each species alongside parallel selection pressures for the same phenotype have revealed several candidate genes that may play important roles in determining these phenotypes, despite the possibility that extensive hybridisation within and between groups could mask true signatures of selection, and/or give false signals of selection.

Here we combine newly generated and existing whole genome sequencing (WGS) data to (1) provide stronger evidence for species relationships among cultivated Brassicas and their suspected CWRs, (2) determine the extent of CWR-crop introgression, (3) resolve the taxonomic status of putative progenitor taxa, and (4) explore the role of hybridisation in the emergence of domesticates. We therefore also take this opportunity to (5) further identify genomic regions of recent positive selection in domesticated varieties (and to compare selection targets across species convergently domesticated for similar morphologies), with value to crop breeding efforts.

Materials and Methods

Whole genome resequencing, data acquisition and processing

Seeds were obtained for 22 wild *Brassica* accessions: eight wild *Brassica oleracea* accessions, eight wild *B. rapa* accessions and six crop wild relatives (CWRs). These were obtained from Warwick UK Vegetable Genebank (https://warwick.ac.uk/fac/sci/lifesci/wcc/gru/genebank/seed/), the U.S. National Plant Germplasm System (https://npgsweb.ars-grin.gov/gringlobal/search), and the Leibniz Institute of Plant Genetics and Crop Plant Research Genebank (https://www.ipk-gatersleben.de/en/genebank/). Seeds were grown in the University of Southampton glasshouse and DNA was extracted from frozen leaf material using a modified CTAB protocol (Doyle & Doyle, 1990). Novogene Bioinformatics Institute (Cambridge, UK) performed library preparation and 150 bp paired-end (PE) sequencing (350 base insert size) using an Illumina 2500 platform (Illumina, USA). Genome size of the six wild *Brassica* relative species were determined using flow cytometry by Plant Cytometry Services (http://www.plantcytometry.nl/).

Additional resequencing reads were obtained for 86 diploid samples from previously published datasets (See Methods S1) and included *Raphanus raphanistrum* and *Erucastrum elatum* as outgroups. Accession information for all 108 samples is available in Supporting Information Table S1.

WGS and acquired resequencing data were quality checked using FastQC (Andrews, 2010). Sequences were trimmed and filtered with Trimmomatic v0.36 (Bolger, Lohse, & Usadel, 2014), removing adapters, the first 5 bases, leading and trailing bases with quality <5, and where average quality per base of a sliding window dropped below 15. Reads <40 bp were removed. Data obtained from An et al. (2019) and Kiefer et al. (2019) were already trimmed. Following quality control, samples had an average of 11.3x coverage ± 1.4 (95% CI).

Alignment and SNP filtering

Brassica CWRs mapped more efficiently to the *Brassica oleracea* pangenome than the *Brassica rapa ssp. pekinensis* v3.0 genome (Supporting Information Table S2). Thus, for initial phylogenetic analysis all 108 samples were aligned to *Brassica oleracea* (Golicz et al., 2016) using Bowtie2 v2.3.1 (Langmead & Salzberg, 2012). For further phylogenetic analysis of *B. rapa* and related *Brassicas*, a subset was aligned to the *B. rapa* ssp. *pekinensis* genome. Only whole chromosome alignments were subsequently analysed.

Bam files were processed with Picard v2.8.3 (picard.sourceforge.net) and variants detected using the Genome Analysis Toolkit v3.7 (GATK) (Van der Auwera et al., 2013) as detailed in Methods S1. Filtering parameters were determined following examination of their distribution in the raw SNP and indel datasets (Supporting Information Figures S10-S13). Linkage disequilibrium (LD) decay was calculated using PopLDdecay v3.40 (C. Zhang, Dong, Xu, He, & Yang, 2019). SNPs in the two datasets were annotated using SNPeff v5.0 (Cingolani et al., 2012) according to annotation files available for genomes.

SNP phylogenies

Phylogenetic trees were constructed from filtered multi-sample gVCFs using maximum likelihood (ML) in SNPhylo (Lee, Guo, Wang, Kim, & Paterson, 2014), SNPhylo identifies blocks of sequence in LD and keeps one informative SNP per block, which reduces information redundancy while increasing computational tractability. Representative SNPs were extracted with parameters; minimum coverage depth 5 and LD threshold 0.05. SNPs were then concatenated into sequences and aligned using MUSCLE (Edgar, 2004) and a phylogenetic tree was determined using DNAML in the PHYLIP package (Felsenstein, 1989) with *R. raphanistrum* as the outgroup. Bootstrap analysis was performed using PhyML v3.0 and 100 replications (Guindon et al., 2010) and phylogenies were visualised in iTOL (http://itol.embl.de). One wild *B. rapa* individual appeared mislabelled given its position in the phylogenetic tree and was removed from further analysis (black dot, Figure 1a).

Relative minimum distance (RMDmin) to wild *Brassica* relatives

This and all subsequent statistical analyses were conducted in R v3.5.2 (R Core Team, 2015).

Relative minimum distance between (1) domesticated *B. oleracea* and wild *Brassica* relatives, and (2) domesticated *B. rapa* and wild *Brassica* relatives were examined using the summary statistic RNDmin (Rosenzweig, Pease, Besansky, & Hahn, 2016). RNDmin is a measure of the minimum pairwise distance between populations relative to divergence to an outgroup and was calculated from SNPs in 50 kb windows with 50 kb step size using R package PopGenome (Pfeifer, Wittelsburger, Ramos-Onsins, & Lercher, 2014) with outgroup *R. raphanistrum*. RNDmin was plotted using smooth.spline() in R with smoothing parameter 0.4. To determine whether there were significant differences in genome-wide RNDmin averages between comparisons of *B. oleracea* with each of the CWRs and between comparisons of *B. rapa* with each of the four CWRs, a one-way ANOVA was conducted (see Methods S1 for comparisons and further details). *Post hoc* pairwise comparisons were conducted using R package *emmeans* (Lenth, Singmann, Love, Buerkner, & Herve, 2018).

Genome-wide introgression

Introgression was detected using D-statistics, using Dtrios in Dsuite (Malinsky, Matschiner, & Svardal, 2020). D-statistics were estimated from biallelic SNPs for trios of populations using *R. raphanistrum* and *E. elatum* as outgroups. A Benjamini-Hochberg multiple test adjustment (Benjamini & Hochberg, 1995) was applied (FDR-corrected $P<0.05$). Genome-wide *fd* (Martin, Davey, & Jiggins, 2015) was calculated from windows of 50 informative SNPs across the genome for combinations of taxa. *fd* identifies and estimates the degree of unidirectional introgression from P3 into P2 in four populations with the relationship (((P1,P2),P3),O).

Phylogenetic network analyses

Hybridisation in *Brassica* phylogenetic networks was inferred using PhyloNet v3.8.2 (Wen, Yu, Zhu, & Nakhleh, 2018) which accounts for incomplete lineage sorting. Since PhyloNet is computationally demanding, multisample gVCF were subsampled to 2-4 representative individuals of wild and domesticated populations of *B. oleracea* and *B. rapa*, and one or more monophyletic CWR (Supporting Information Table S1). SNP gVCF files were split into 200 kb regions and converted to PHYLIP files. Suitable nucleotide substitution models were determined using JModeltest2 (Darriba, Taboada, Doallo, & Posada, 2012). For each genome fragment, phylogenies were constructed using RaxML v8.2.9 (Stamatakis, 2014) and bootstrapped with 100 replicates. Resulting trees were converted to nexus files and used to infer phylogenetic networks with zero to five reticulations using the InferNetwork_MPL module. Optimal number of reticulations was determined where the increase in pseudolikelihood with reticulation number began to plateau (Blair & Ane, 2020).

Networks predicted with PhyloNet were evaluated using Approximate Bayesian Computation (ABC) (Beaumont, Zhang, & Balding, 2002) and used increased sample sizes of 4-8 individuals per *B. oleracea* and *B. rapa* population (Supporting Information Table S1). Subsets of unlinked SNPs with no missing data were generated and formatted for DIYABC v.2.1.0 (Cornuet et al., 2014) using a python script https://github.com/loire/vcf2DIYABC.py. In DIYABC, uniform distributions were chosen for priors, with $10\text{-}10^7$ for population size and divergence times. All available summary statistics were utilised for *B. rapa*, with a subset of 135 used for the *B. oleracea* analysis (including means of genic diversity and pairwise F_{ST}) for computational tractability. For each network scenario 10^6 simulations were conducted.

The posterior probability of each network was estimated using logistic regression with a logit transformation, based on the number of times the network appears in the top 1% of simulations

when sorted by distance to the observed dataset (Cornuet et al., 2014). Confidence in network choice was evaluated by calculating Type I and Type II error (Cornuet, Ravigné, & Estoup, 2010).

Population structure

Population structure within *B. oleracea* and *B. rapa* were analysed separately. SNPs in LD were filtered out using PLINK v1.07 (Purcell et al., 2007) with 50 kb window size, 5 kb step-size, and variant inflation factor 2, then randomly thinned to 50,000 SNPs. Population structure was analysed in STRUCTURE v2.3.4 (Pritchard, Stephens, & Donnelly, 2000) with 1-10 genetic clusters (K). Each value of K was replicated 10 times, for 20,000 runs following a 10,000 run burn in. Optimal K was estimated in STRUCTURE HARVESTER (Earl & VonHoldt, 2012) following the ΔK method (Evanno, Regnaut, and Goudet (2005). Replicates of K were aligned, merged and plotted using R package POPHELPER v2.3.1 (Francis, 2017).

Genome-wide population statistics

Nucleotide diversity, Tajima's D and SNP and indel densities across the *B. oleracea* and *B. rapa* genomes were calculated from filtered SNPs in 50 kb windows using VCFtools v0.1.15 (Danecek et al., 2011). Population statistics were plotted using Circos v0.69-6 (Krzywinski et al., 2009).

Demographic history inference

Population size changes over time were inferred for wild and domesticated *B. rapa* and *B. oleracea* using a sequentially Markovian coalescent (SMC) method in SMC++ (Terhorst, Kamm, & Song, 2017). All domesticated *B. oleracea* varieties excluding *alboglabra* (see results [Figure 1]) were combined for the *B. oleracea* domesticated population (n=24) and compared to wild *B. oleracea* (n=10). Domesticated *B. rapa* subspecies *trilocularis, chinensis, parachinensis* and *pekinensis* were combined for the *B. rapa* domesticated population (n=25), with wild *B. rapa* and *B. rapa ssp. rapa* combined for the wild population (n=15) since the latter are not reciprocally monophyletic (see results). Regions identified as under positive selection (described in the next section) were masked. In SMC++ sample frequency spectra are conditioned on a "distinguished lineage" rather than a reference genome. Five to seven "distinguished lineages" were used per population, and each chromosome was analysed separately. Models were estimated using the *estimate* function, using a mutation rate estimate of 1.5×10^{-8} synonymous mutations per generation (Kagale et al., 2014) and a generation time of one year as in other analyses (McAlvay et al., 2021, Okazaki et al., 2007).

Identification of regions affected by positive selection and targets of selection within them

Recent hard positive selective sweeps were identified by combining outputs from Sweed (Pavlidis, Zivkovic, Stamatakis, & Alachiotis, 2013) and Omegaplus (Alachiotis, Stamatakis, & Pavlidis, 2012). Sweed identifies signatures of selection in site frequency spectra using CLR tests, while Omegaplus looks for signatures of selection in LD using the ω-statistic. For analyses of domesticates, domesticated varieties of *B. oleracea* and subspecies of *B. rapa* were analysed separately (n=4-10). For comparisons of domesticated and wild populations, populations were defined as for demographic history inference.

In Sweed, likelihood ratios are reported for a specific position as well as the genomic window that maximises CLR for that position. This window is determined dynamically and is biologically relevant since strong selection generally affects large genomic regions (subject to LD decay). In Omegaplus, statistics are reported for a specific position only. To identify regions affected by positive selection supported in both analyses, first, positions of the top 1% CLR reported in Sweed were retained if the windows that maximised CLR for these positions also contained positions within the top 1% of ω-statistics. These positions are referred to as top 1% CLR;ω-statistic positions and overlapping associated CLR windows were combined to identify regions affected by positive selection. In an attempt to distinguish between the likely target of positive selection and the genomic window affected by selection we used custom R scripts to identify regions within windows maximising CLR for top 1% CLR;ω-statistic positions, starting where the CLR value for the position first crosses the top 1% CLR threshold and ending where it falls below.

Genes overlapping regions targeted by positive selection were extracted using the R package GenomicRanges (Lawrence et al., 2013). Gene sequences were compared against The *Arabidopsis* Information Resource (TAIR10) (Berardini et al., 2015) using BLASTX (Altschul, Gish, Miller, Myers, & Lipman, 1990) (e-value $<1x10^{-4}$ and >60% sequence identity).

Gene ontology (GO) enrichment analysis

GO enrichment analysis was conducted for genes targetted by selection in each variety/subspecies separately using a Fisher's exact test with Benjamini and Yekutieli multiple test adjustment (Benjamini & Yekutieli, 2001) (FDR <0.05), in agriGO v2.0 (Du, Zhou, Ling, Zhang, & Su, 2010). Venn diagrams of gene ontology categories in targets of positive selection between domesticates were drawn using eulerr (J. Larsson, 2020).

Parallel selection analysis

Genes identified in target windows of selection were compared to those in a similar analysis using reduction in diversity (ROD) metrics and population-based integrated haplotype score (PiHS) (Cheng, Sun, et al., 2016). The *B. rapa* genome (Chiifu-401-42) and *B. oleracea* var. *capitata* (line 02–12) v1.0 genome and annotation files were downloaded from BRAD (X. B. Wang, Wu, Liang, Cheng, & Wang, 2015) and Bolbase (Yu et al., 2013), respectively. Genes in regions identified as under selection in Cheng, Sun, et al. (2016) were extracted and compared to genes identified in the SFS and LD based analyses here using BLASTX (e-value $<1x10^{-4}$ and >60% sequence identity).

The genes in candidate target regions were compared for three pairs of *B. oleracea* and *B. rapa* domesticated varieties with similar phenotypes; i.e., early flowering varieties (*B. oleracea* var. *alboglabra* and *B. rapa* ssp. *parachinensis*), heading varieties (*B. oleracea* var. *capitata* and *B. rapa* ssp. *pekinensis*), and enlarged stem varieties (*B. oleracea* var. *gongylodes* and *B. rapa* ssp. *rapa*). Gene fasta files were BLAST searched between pairs to identify putative othologues (BLASTX; e-value $<1x10^{-4}$ and >60% sequence identity). Reciprocal best BLAST was used.

Fixed polymorphisms between a domesticated variety and other same species domesticates were identified within 1 kb either side of genes of interest using vcf-contrast in VCFtools (Danecek et al., 2011), and the sequence was extracted and examined in AliView (A. Larsson, 2014).

Results

Whole genome sequencing (WGS) data of 22 wild *Brassica* individuals (8 wild *B. oleracea*, 8 wild *B. rapa* and one each of six diploid CWRs; Supporting Information Table S1) yielded an average of 43.3 M PE reads (± 3.3 M; 95% CI). The six *Brassica* relatives were confirmed to be diploid (610-645 Mbp/1C; Supporting Information Table S2). WGS data acquired from a further 86 diploid individuals from six publications averaged 27.3 M PE reads (± 3.1 M).

1. Phylogenomic relationships among *Brassica* species

All 108 samples were mapped to the *B. oleracea* pangenome and then, to determine the phylogenomic relationships among *B. rapa* and CWRs, relevant samples were mapped to the *B. rapa* genome. Average mapping efficiency for these datasets were 72.8% and 75.9% respectively and calling and filtering SNPs resulted in 6.0 M and 4.1 M SNPs respectively. Phylogenomic relationships between *Brassica* species (Figure 1; Supporting Information Figures S1, S2) demonstrate that *B. rapa*

(2n=20) formed a monophyletic clade distinct from *B. oleracea* and all other wild *Brassica* species (2n=18) in both analyses.

In the *B. oleracea*-aligned analysis (Figure 1a), five of the CWRs (*B. cretica* Lam., *B. rupestris* Raf., *B. macrocarpa* Guss., *B. insularis* Moris, and *B. hilarionis* Post) formed a clade, and *B. oleracea* formed another in which wild *B. oleracea* is polyphyletic. *B. villosa* Biv. ex Spreng. was found in both the CWR and *B. oleracea* clades; *B. incana* Ten., and *B. montana* Raf. were found at multiple places in the *B. oleracea* clade. Cultivated varieties of domesticated *B. oleracea* formed monophyletic groups, however often lacked strong bootstrap support.

In the *B. rapa*-aligned analysis (Figure 1b), wild *B. rapa* and ssp. *rapa* (turnip) formed a clade distinct from the other domesticates and were not reciprocally monophyletic. Three domesticated ssp. (*trilocularis*, *parachinensis* and *pekinensis*) were monophyletic but with low bootstrap support, and ssp. *chinensis* was polyphyletic.

The summary statistic RNDmin (Rosenzweig et al., 2016) was used to calculate the minimum genetic distance between domesticates and CWRs (Figure 1c, d). After excluding windows with zero RNDmin, the smallest minimum distance between domesticated *B. oleracea* (combined) and one of the monophyletic CWRs, was with *B. cretica* (Figure 1c; Supporting Information Figure S3a). RNDmin between *B. oleracea* and *B. cretica* was significantly different than the average RNDmin for any of the other three CWRs analysed (ANOVA; $F(3,11862)=406.4$, $p<0.001$, Tukey's HSD; all $p<0.001$; Supporting Information Table S3). Although *B. cretica* and other CWRs are equidistant to *B. oleracea* based on the phylogeny, a lower RNDmin for the *B. cretica* comparison could mean some gene flow between *B. cretica* and *B. oleracea* has caused this apparent similarity. The number of zero RNDmin windows between the wild relatives and domesticated *B. oleracea* (zero RNDmin represents fully conserved or fully introgressed regions) differed (logistic regression; all coefficients $p<0.05$, Tukey's HSD; all $p<0.05$; Supporting Information Table S3) with *B. cretica* having the most, again suggesting the close relationship, potentially due to introgression, to *B. oleracea*.

RNDmin between domesticated *B. rapa* (excluding ssp. *rapa*) and wild *Brassica* species were much larger than were observed for *B. oleracea* (Figure 1d), and no zero RNDmin windows were identified, consistent with the relative phylogenetic placement of *B. rapa* and *B. oleracea* (Figure 1a). The smallest RNDmin was between *B. rapa* and *B. insularis* (Supporting Information Figure S3b), smaller than the distance to any other CWR (ANOVA; $F(3, 16742)= 195.5$, $p<0.001$, Tukey's HSD; all $p<0.001$; Supporting Information Table S4).

2. Introgression and hybridisation among *Brassica* crops and CWRs

Among *Brassica oleracea* groups and CWRs

D-statistics detected introgression between all monophyletic CWRs and wild and domesticated *B. oleracea* (Supporting Information Table S5, Figure S4). The direction and extent of introgression was further investigated using *fd* (Martin et al., 2015). No introgression was detected from CWRs into *B. oleracea* var. *alboglabra* but was found from CWRs into all other domesticated varieties, predicted to account for 0.20-3.61% of the genome (Supporting Information Table S6). No introgression from domesticated *B. oleracea* varieties into the CWRs *B. insularis, B. macrocarpa* and *B. rupestris* was detected. However, all domesticated and wild *B. oleracea* varieties exhibited introgression into *B. cretica*, accounting for 9.46-14.28% of the genome.

Phylogenetic networks (allowing one to five reticulations) were constructed from 2174 trees (see methods), for a subset of 22 representative individuals of *B. oleracea* and two monophyletic relatives, *B. cretica* and *B. rupestris* (Supporting Information Table S1). There was no clear plateau in pseudolikelihood with increasing number of reticulations (Supporting Information Figure S5), highlighting the complexity of relationships. Therefore, the networks with the highest pseudolikelihood from zero to five reticulations were compared using ABC (Supporting Information Table S1 and Figure S6). The network with three reticulations was most likely. Type I and Type II error rates for this network were low at 1.1% and 6.2%, and model checking demonstrated low discordance between simulated network–prior combinations and observed data (Supporting Information Table S7). This network (Figure 1e) included reticulations producing var. *botrytis* (cauliflower), var. *alboglabra* (Chinese kale) and a recently derived wild *B. oleracea* clade (see below). D-statistics between domesticated *B. oleracea* varieties largely support this model showing signals of introgression between *B. oleracea* var. *alboglabra* and the other varieties ($D=0.043$-0.053, $p<0.001$), and between *B. oleracea* var. *botrytis* and var. *gongylodes* ($D=0.039$, $p<0.05$; Supporting Information Table S5).

Among *Brassica rapa* groups and CWRs

D-statistics also identified introgression between CWRs and *B. rapa* subspecies (Supporting Information Table S8, Figure S7). Unidirectional introgression from CWRs into *B. rapa* subspecies was evidenced for the wild *B. rapa*/ssp. *rapa* clade (1.74-1.78% of the genome; Supporting Information Table S9), for ssp. *trilocularis* (0.00-1.76%) and to ssp. *pekinensis* (0.03-0.80%). No introgression was detected from *B. rapa* subspecies into *B. cretica* or *B. macrocarpa* but introgression was found into *B. insularis* (0.22-0.88%) and to a lesser extent into *B. rupestris* (Supporting Information Table S9).

This is consistent with the strongest signal of introgression between *B. insularis* and *B. rapa* subspecies identified by D-statistics and the increased genetic relatedness (RNDmin) compared to other CWRs. Since all subspecies are introgressed with *B. insularis* this likely represents ancient introgression.

Phylogenetic networks including *B. rapa* subspecies resulted in 1028 trees from 18 representative individuals of *B. rapa* and *B. cretica* (Supporting Information Table S1). One reticulation maximised pseudolikelihood while minimising reticulation number (Supporting Information Figure S8) and the five one-reticulation models with the highest pseudolikelihood were compared using ABC (Supporting Information Table S1). These support a hybrid origin for ssp. *trilocularis*; two near-identical networks were well-supported (Figure 1f) but differ slightly in the contributing parental populations (Supporting Information Figure S9). Type II error rates for these networks were low (7.9% and 13.8% respectively) while Type I error rates were very high (36.2% and 51.4%), potentially evidencing a lack of SNP variation to distinguish between topologies. Discordance between simulated network–prior combinations and the observed data set was evident (Supporting Information Tables S10, S11), however was less for scenario 5 (Figure 1e).

3. Population structure and domestication history of wild and domesticated *Brassica*s

Population structure and demographic history analysis of *B. oleracea* and *B. rapa*

The focused analysis of *B. oleracea* (38 *B. oleracea* samples aligned to the *B. oleracea* pangenome) resulted in 6.1 M SNPs and 1.0 M indels (<50 bp) (Table 1). SNP and indel density, nucleotide diversity (mean 2.6 ± 0.023x10^{-3} [95% CI]) and Tajima's D (mean 1.738 ± 0.021 [95% CI]) varied throughout the genome (Figure 2a). The focused analysis of *B. rapa* (41 *B. rapa* samples aligned to the *B. rapa* pangenome) resulted in 5. 8 M SNPs and 0.8 M indels (Table 1). Again, SNP and indel densities, nucleotide diversity (mean 4.5 ± 0.049x10^{-3} [95% CI]) and Tajima's D (mean 1.780 ± 0.016 [95% CI]) varied throughout the genome (Figure 2f).

LD decay calculated from genome wide SNPs dropped to half of maximum average r^2 at c. 53 kb and c. 62 kb for wild and domesticated *B. oleracea*, respectively (Figure 2b). The difference in LD decay between wild and domesticated *B. rapa* was larger, c. 43 kb and c. 70 kb for the wild *B. rapa*/ssp. *rapa* group and domesticated *B. rapa*, respectively (Figure 2g). These are comparable to values for other crops where a recent genetic bottleneck in the domesticated populations has been cited as the cause (X. Huang et al., 2012; X. H. Huang et al., 2010; Zhou et al., 2015).

Domestication history of *B. oleracea*

There was low bootstrap support for the relative positions of wild and domesticated *B. oleracea* populations in phylogenetic analyses, however network analyses support var. *alboglabra* as the earliest diverging lineage (Figure 1). This supports the ABC analyses (above) that wild *B. oleracea* accessions are feral derivatives, not wild ancestors (Figure 1a, c). In STRUCTURE analysis of *B. oleracea*, the number of underlying populations was estimated as five (Figure 2c, d). Varieties *alboglabra*, *gongylodes*, *capitata*, and *botrytis* largely formed distinct genetic clusters with a fifth cluster including wild *B. oleracea*, var. *sabellica* and var. *acephela*. Wild *B. oleracea* individuals show admixture from each of the domesticated clusters.

Domesticated *B. oleracea* (excluding var. *alboglabra*) experienced a decline in effective population size from 10 Kya to c. 300 years ago (Ne=133,000 to Ne=1,000), followed by a prominent expansion c. 40 years ago (Ne=3,333,000) and rapid decline to the present day (Figure 2e). This is consistent with a long history of cultivation, global distribution of cultivated varieties and then improvement of *B. oleracea* varieties. The effective population size of wild *B. oleracea* similarly declined, from 100 Kya (Ne=123,000) to 500 ya (Ne=2,000), which could represent shared ancestry during the cultivation of *B. oleracea* until 1 Kya, supporting STRUCTURE analysis.

Domestication history of *B. rapa*

STRUCTURE analysis of *B. rapa* identified five clusters, three were clearly delimited (wild *B. rapa*/ssp. *rapa*, ssp. *trilocularis* and ssp. *pekinensis*), ssp. *chinensis* was partially assigned to a fourth and ssp. *parachinensis* to a fifth, but with extensive admixture (Figure 2). This largely matches the phylogenetic clades identified above. Domesticated *B. rapa* subspecies (excluding ssp. *rapa*), experienced a decline in effective population size from 25 Kya to c. 1 Kya (Ne=75,000 to Ne=2,000) and the wild *B. rapa*/ssp. *rapa* population declined from 200 Kya (Ne=137,000) to 2 Kya (Ne=5,000), followed by a small increase (Figure 2j). Considering the above analyses and the extant geographical ranges, this could describe complex parallel and largely independent cultivation histories, i.e., early cultivation of ssp. *rapa* in Europe with later independent domestication in South-East Asia (ssp. *chinensis*, *pekinensis* and *parachinensis*), and with the divergence of ssp. *trilocularis* in Southern Asia (Qi et al., 2017). The lack of significant recent expansion in *B. rapa*, compared to the prominent expansion *B. oleracea*, could reflect greater recent introgression from the wild in the latter.

4. Positive selection during domestication

Shared selection in wild and domesticated populations of *B. oleracea* and *B. rapa*

Genomic regions targeted by positive selection were identified through composite likelihood ratio (CLR) tests of site frequency spectra (SFS) in Sweed (Pavlidis et al., 2013) and LD patterns (ω) using Omegaplus (Alachiotis et al., 2012) (see methods for details). Several regions targeted by positive selection were identified in wild and domesticated (excluding var. *alboglabra*; see above) *B. oleracea*, with regions on chromosomes 4 and 5 overlapping (Figure 3a). In these overlapping regions there were 38 genes (Supporting Information Table S12) but only 14 had an *Arabidopsis* BLAST hit and no biological processes were significantly enriched in GO analysis of these. Several regions of selection were identified in wild and domesticated *B. rapa*, however no regions overlapped for wild *B. rapa*/ssp. *rapa* and domesticated subspecies (Figure 3b).

Between domesticates within species

Combined analyses of SFS and LD identified regions targeted by recent positive selection in all domesticates (Figure 3c). For *B. oleracea*, each domesticate showed overlap in regions of positive selection with at least one other domesticate, but for *B. rapa*, the only overlap was limited to the three subspecies with large leaf phenotypes (Cheng, Wu, et al., 2016), overlapping geographic ranges and a shared domestication history (Figure 3c). The other two domesticates (ssp. *trilocularis* (oilseed) and ssp. *rapa* (turnip) showed no overlap.

On average, regions targeted by positive selection represented 0.76 ± 0.24% (3.65 MB) of the assembled chromosomes for *B. oleracea* domesticates and 1.29 ± 0.67% (2.49 MB) for *B. rapa* domesticates, containing an average of 355 and 230 annotated genes respectively. The smallest proportion was for ssp. *trilocularis* with only 0.06% of the genome (0.14 MB; 18 genes).

Despite the close relationship between *Arabidopsis* and *Brassica*, *Arabidopsis* orthologues were identified for only 53% of genes (Supporting Information Tables S13-S25), which may have limited detection of potentially key genes. Gene ontology (GO) analysis evidenced large overlap of functions and processes in the genes in these regions across domesticates, with few group-specific enriched GO categories (Figure 3d, Supporting Information Tables S26-S28). GO categories such as "multicellular organism development" and "response to stimulus" were enriched (FDR<0.05, Fisher's exact test) for all domesticated and wild populations, except *B. rapa* ssp. *trilocularis*, again highlighting the distinctiveness of this subspecies.

Parallel selection during domestication for similar phenotypes

To analyse parallel selection for similar phenotypes, genes in positive selection target regions were compared for (1) *B. oleracea* var. *alboglabra* and *B. rapa* ssp. *parachinensis* (early flowering/leafy varieties), (2) *B. oleracea* var. *capitata* and *B. rapa* ssp. *pekinensis* (heading varieties), and (3) *B. oleracea* var. *gongylodes* and *B. rapa* ssp. *rapa* (enlarged stem varieties).

For comparisons (2) and (3), similar comparisons have been carried out previously (Cheng, Sun, et al., 2016) using earlier genome assemblies and alternative methods for identifying positive selection. Because of the introgression and admixture we resolved, we used a single population approach to identify regions under selection rather than comparisons between domesticates or putatively wild groups used previously. For each of the four groups in these two comparisons we compared genes in candidate regions targeted by selection in our analysis with those identified as under selection in Cheng, Sun, et al. (2016). 40% of the genes identified in target regions in *B. rapa* ssp. *pekinensis,* 25% in *B. oleracea* var. *capitata,* 44% in *B. rapa* ssp. *rapa* and 31% in *B. oleracea* var. *gongylodes* were also identified in Cheng, Sun, et al. (2016) and this is significantly more than expected by chance (χ^2 test; χ^2=21.4-59.5, all p<0.01), suggesting that the approaches show broad agreement.

We then looked at shared selection between groups of the two species that share the same phenotype. For comparisons (1) and (2), 10 and 14 putative *B. rapa–B. oleracea* orthologues were identified in selection target regions, respectively (Figure 4); more than expected by chance (χ^2 test; χ^2=7.2, p<0.01 and χ^2=11.1, p<0.001 respectively). In contrast, no putative orthologues were found in comparison 3 (large stem varieties). Among parallel pairs of genes, only 27% of the 24 gene pairs received an *Arabidopsis* hit, but we still detected over-representation of GO terms involved in transport, methylation and transcription (Supporting Information Table S29).

Selection on genes involved in anatomical structure development

GO annotation of genes in candidate selection target regions highlights promising candidates for follow-up. Of these, genes annotated with the GO term “anatomical structure development” are briefly discussed.

(1) Early flowering and leafy varieties: *B. oleracea* var. *alboglabra* and *B. rapa* ssp. *parachinensis*

Two of six such genes in regions targeted by positive selection in *B. oleracea* var. *alboglabra* and one of the ten in *B. rapa* ssp. *parachinensis,* were involved in auxin response. Both gene sets also

contained putative orthologues of genes involved in floral development; *CULLIN3* encoding a positive regulator of floral development (Chahtane et al., 2018) in var. *alboglabra* and *ICMB* encoding a negative regulator of signalling pathways affecting floral development (Bracha-Drori, Shichrur, Lubetzky, & Yalovsky, 2008) in ssp. *parachinensis*.

(2) Heading varieties: *B. oleracea* var. *capitata* and *B. rapa* ssp. *pekinensis*

Regions targeted by positive selection on chromosome 3 of ssp. *pekinensis* contained three genes with this GO annotation, including *ALE1*, encoding a subtilisin protease associated with leaf development (Tanaka et al., 2001). Irregularities in the coordination of leaf polarity are thought to play a key role in the formation of the heading phenotype in ssp*. pekinensis* (Li et al., 2019) and a putative orthologue of a gene involved in this pathway, *KANADI-2* (Yamaguchi, Nukazuka, & Tsukaya, 2012), was identified on chromosome 9 of ssp. *pekinensis*.

In var. *capitata*, a putative orthologue of *ASL5*, which operates in the same adaxial-abaxial polarity pathway as KANADI-2, was one of seventeen anatomical structure genes in regions targeted by selection. This pathway therefore warrants further investigation for the heading phenotype in both species.

(3) Enlarged stem varieties: *B. oleracea* var. *gongylodes* and *B. rapa* ssp. *rapa*

In enlarged stem varieties there was no clear overlap in the pathways of genes targeted by selection, although there were potential candidates. In var. *gongylodes, WVD2* was one of twelve anatomical structure development genes. Overexpression of *WVD2* in *Arabidopsis* results in shorter, stockier roots and stems (Perrin, Wang, Yuen, Will, & Masson, 2007). In ssp. *rapa*, an orthologue of a gene encoding a transcription factor that regulates organ size when overexpressed in *Arabidopsis* (*ANT*) is in a region targeted by positive selection (Ding et al., 2018).

Causative SNPs in genes of interest

Among the genes of interest (i.e., under parallel selection or annotated as "anatomical structure development" genes; Supporting Information Tables S30-S31), the only fixed difference between a domesticated variety and other varieties occurred in *AAT* on chromosome 8 for *B. oleracea* var. *gongylodes* (kohlrabi). This gene functions in the biosynthesis of aromatic amino acids that have diverse roles as precursors to secondary metabolites such as anthocyanins (Tzin & Galili, 2010). Three fixed SNPs result in amino acid replacement and therefore potentially functional changes.

Discussion

This analysis further demonstrates the complex phylogenomic relationships between *Brassica* crops and their crop wild relatives (CWRs), quantifying for the first time the extent of introgression in their diversification and domestication. We then adopt a single population analysis strategy to identify candidate genomic regions under selection during domestication. The incorporation of adaptive genetic diversity from CWRs into crops is a key strategy to improve crop resilience to climate change to ensure future food security (Castañeda-Álvarez et al., 2016). The resolution of *Brassica* CWR and crop genetic relationships therefore has direct application to these diverse and economically important crops.

Phylogenomic relationships among *Brassica* species and the potential for *Brassica* crop improvement

Two recent phylogenies constructed to model the *B. rapa* and *B. oleracea* groups (including a small number of CWRs) within the Core Oleracea lineage used genotyping-by-sequencing (McAlvay *et al*., 2021) and RNA-seq (Mabry *et al*., 2021). We present largely concordant findings, but with additional insight using alternative outgroups and increased resolution afforded by WGS.

Mabry et al. (2021) suggest that *B. cretica* is the closest CWR to *B. oleracea*, whereas we argue that *B. cretica* is better described as a member of a cluster of CWRs distinct from *B. oleracea* (also found by Song et al., 1990 using RFLPs). The Mabry et al. (2021) analysis uses a sample of *B. villosa* as an outgroup, which both our and their study indicate is not monophyletic. Our study instead used *Raphanus* as an outgroup which helped us to highlight the relationships between the CWRs. We further identify gene flow between *B. oleracea* and *B. cretica*, which could explain the close phylogenetic position resolved by Mabry et al. (2021). This gene flow likely took place prior to domestication, given that all domesticated groups show introgression (accounting for 9.46-14.28% of the genome) with *B. cretica*.

The close relationships between several CWRs highlights that all these CWRs could be considered potential sources of adaptive genetic variation for *B. oleracea* breeding. Indeed, introgression was detected from several CWRs into crop varieties of both *B. oleracea* and *B. rapa* demonstrating that crossing is likely to be successful. It is important to note that of these CWRs, three are near threatened or critically endangered (Bilz *et al*., 2011), and are poorly represented in seed banks (Castañeda-Álvarez *et al*., 2016), highlighting an urgent need to collect and preserve their genetic diversity.

Our data also confirms that wild *B. oleracea* populations are not monophyletic and are not the ancestors of all *B. oleracea* crops, supporting conclusions that wild populations along the Atlantic coast are feral derivatives (Mabry et al., 2021; Maggioni, von Bothmer, Poulsen, & Lipman, 2018; Mittell et al., 2020). We also show that these populations possess significant admixture from domesticated varieties. Regardless, the combination of closely related CWRs that can hybridise with domesticated *B. oleracea* (FitzJohn et al., 2007), and the pool of potentially adaptive novel allele combinations in admixed wild populations, provides an extensive resource for breeding *B. oleracea* crops where adaptive genetic diversity is lacking (Katche, Quezada-Martinez, Katche, Vasquez-Teuber, & Mason, 2019).

In agreement with McAlvay *et al*. (2021), in our analyses wild populations of *B. rapa* are polyphyletic with *B. rapa* ssp. *rapa* (turnip), which raises several possibilities about the taxonomic status of ssp. *rapa*. The turnip phenotype could be a plastic response to a cultivated environment (and hence not a genetically fixed phenotype), have evolved multiple times, and/or some 'wild' *B. rapa* populations are feral ssp. *rapa*. McAlvay *et al*. (2021) assert that true wilds are present in the Caucasus and Italy, a group that we did not identify but our sampling of wild *B. rapa* was less geographically extensive. Feral *B. rapa* populations could provide potential for intraspecific breeding of *B. rapa* domesticates. To our knowledge, experimental crosses between the CWRs and *B. rapa* have not been conducted, but our evidence for introgression suggests this is possible.

Overall, the finding of polyphyly of *B. villosa*, *B. montana* and *B. incana* (see also McAlvay et al., 2021, Mabry *et al*., 2021) highlights that any putatively wild *B. rapa* and other CWRs should be re-examined alongside samples from other *Brassica* species.

The role of introgression and hybridization among *Brassica* crops and CWRs in domestication and diversification

We detected hybrid origins of domesticates in both *B. oleracea* and *B. rapa*. In *B. oleracea,* both var. *alboglabra* (Chinese kale) and var. *botrytis* (cauliflower) were identified as having hybrid origins between unknown wild *Brassica* species and a more recently derived *B. oleracea* var. *gongylodes* and var. *capitata* lineage. A previous GBS SNP analysis also suggested var. *botrytis* is derived from introgression, albeit with var. *italica* (Stansell et al., 2018) which we did not sample. The greatest proportion of introgressed genomic sites was detected from *B. cretica* (Stansell et al., 2018) into var. *botrytis* but phylogenetic network analysis highlights an unidentified wild species as the parental species. Var. *alboglabra* contrasts with all other domesticates, in that no introgression was detected

548 from the four monophyletic CWRs and could suggest this group was domesticated outside the range
549 of CWRs.

550 Based on the WGS phylogeny, var. *alboglabra* forms a unique population that diverges before other
551 *B. oleracea* varieties, which is also shown in other analyses (Cheng, Sun, et al., 2016; Izzah et al.,
552 2013; Stansell et al., 2018). *B. oleracea* is generally considered to have been domesticated in the
553 Mediterranean (Arias, Beilstein, Tang, McKain, & Pires, 2014; Maggioni et al., 2018; Mabry *et al*.,
554 2021) where the core Oleracea lineage originated c. 3 Mya (Arias et al., 2014). However,
555 phylogenetic placement of var. *alboglabra* might suggest an earlier independent domestication.
556 From an initial hybrid origin, presumably in the Mediterranean, a subsequent absence of
557 introgression from the CWRs indicates var. *alboglabra* was geographically isolated from wild
558 relatives during its domestication. Since var. *alboglabra* is widely cultivated in China and South-East
559 Asia (Dixon, 2006), the hybrid lineage could have been transported to Asia where subsequent
560 selection and domestication took place. In this way, hybridization may have provided a starting point
561 for the cultivation of var. *alboglabra* while an absence of introgression with wild relatives promoted
562 domestication.

563 Qi *et al.* (2017) evidence a stepwise eastward progression of *B. rapa* domestication over 2000-4000
564 years, with turnip and Chinese cabbage cultivation corroborated by written records, and McAlvay *et*
565 *al*. (2021) suggest introgression may have been prominent in a subset of these Central Asian oilseed
566 crops, which our data support. Network analysis suggested a hybrid origin of ssp*. trilocularis* deriving
567 from the ssp. *parachinensis*/ssp. *chinensis* lineage and an unknown, potentially wild, lineage (Figure
568 1f). Previous analyses support that ssp. *trilocularis* forms part of a genetically distinct Asian
569 population of rapid cycling domesticates selected for high seed oil content, however these did not
570 include CWRs (Bird et al., 2017; Cheng, Wu, et al., 2016). The potential hybrid origin of ssp.
571 *trilocularis* should be followed up after more wild taxa are investigated.

572 Previously, Qi *et al.* (2017) identified *B. rapa* ssp. *pekinensis* as a hybrid between ssp. *rapa* and ssp.
573 *chinensis* which is partly supported by McAlvay *et al*. (2021). Although our analysis does not support
574 this, the reduced sampling of *B. rapa* ssp. *chinensis* in our analysis, or the absence of *B. cretica*
575 sampling by Qi *et al.,* (2017) could have led to this discrepancy.

576 **Positive selection and parallel evolution during domestication**

577 The considerable phenotypic variation in domesticated *Brassicas* provides opportunities to
578 investigate parallel evolution, similarly explored in other crops (Lin et al., 2012; M. Wang et al.,
579 2018). Hybrid origins of some domesticated varieties, introgression between domesticates and

CWRs, and the emergence of domesticated groups in geographical isolation from CWRs highlight the phylogenetic complexity of this group. Consequently, our analysis employs single population approaches to identify targets of selection during domestication rather than traditional comparative approaches (Cheng, Sun, et al., 2016). This could be advantageous in other systems too, where phylogenetic analysis has identified complex histories of hybridisation between domesticated varieties and with wild relatives (Flowers et al., 2019; Page, Gibson, Meyer, & Chapman, 2019). Although using smaller sample sizes compared to previous analyses (Cheng, Sun, et al., 2016), our analysis does make use of updated genome assemblies.

Our analysis of selection in domesticated groups, and parallel evolution among crops selected for similar phenotypes, identified further potential targets with importance for breeding programmes. These may also be relevant to research in similar phenotypes for other crop species. For one such gene (*AAT*), kohlrabi exhibited fixed non-synonymous SNPs compared to other domesticates. This gene functions in the production of aromatic amino acids, and variants have been associated with flowering time and yield in lentil (Skibinski, Rasool, & Erskine, 1984). Furthermore, aromatic amino acids are precursors to anthocyanins (Winkel-Shirley, 2001) which produce the purple colour of some kohlrabi varieties (Park et al., 2017, Petropoulos et al., 2019). Consumption of these anthocyanins can have health benefits (Kim et al., 2017), thus this gene warrants further study with reference to human health and exemplifies the potential application of this positive selection analysis. Other genes worthy of further investigation include ALE1 and ASL5 both putatively involved in leaf development and identified in the selection analysis of heading varieties of both crops.

We also note that different numbers of genomic loci appear to show signatures of selection during the evolution of different domesticated groups. For example, only 0.06% of the genome (0.14 MB; 18 genes) showed evidence for selection in *B. rapa* ssp. *trilocularis*, which may suggest the evolution of yellow seeds and high seed oil content characteristic of this taxon involved few genes.

Conclusions

Our study demonstrates through a range of approaches and genome sequencing of CWRs that hybridisation and introgression have been instrumental in the evolution of *Brassica* crops as well as continuing more recently between crops and wild relatives. Our selection analysis, which should be less prone to interference from past hybridisation, identified targets of selection during *Brassica* domestication. Overall, we show that there are several CWRs with potential to hybridise with domesticated Brassica species and we identify candidate genes for adaptive phenotypes worthy of follow-up.

Acknowledgements

This work was funded by the Natural Environment Research Council Grant number NE/S002022/1 to MAC and THGE. This work was supported by the use of the IRIDIS High Performance Computing Facility at the University of Southampton, and we thank staff at the associated support services for their assistance. We are grateful to members of the Chapman lab for their comments on this manuscript, Ying Hu and Asia Hoile for preliminary analysis of seed bank material and Mike Cotton for greenhouse assistance.

Author Contribution

JMS, THGE and MAC planned the experiments, JMS, AJR and MAC carried out the lab work, JMS analysed all data with input from MAC, JMS wrote the paper, MAC edited the paper and all authors read, edited and approved the final version.

Data Availability

All raw sequencing data generated in this study have been deposited in the NCBI SRA under project number PRJNA929712.

References

Alachiotis, N., Stamatakis, A., & Pavlidis, P. (2012). OmegaPlus: a scalable tool for rapid detection of selective sweeps in whole-genome datasets. *Bioinformatics, 28*(17), 2274-2275. doi:10.1093/bioinformatics/bts419

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic Local Alignment Search Tool. *Journal of Molecular Biology, 215*(3), 403-410. doi:10.1006/jmbi.1990.9999

An, H., Qi, X. S., Gaynor, M. L., Hao, Y., Gebken, S. C., Mabry, M. E., . . . Pires, J. C. (2019). Transcriptome and organellar sequencing highlights the complex origin and diversification of allotetraploid Brassica napus. *Nature Communications, 10*. doi:10.1038/s41467-019-10757-1

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Arias, T., Beilstein, M. A., Tang, M., McKain, M. R., & Pires, J. C. (2014). Diversification times among *Brassica* (Brassicaceae) crops suggest hybrid formation after 20 million years of divergence. *American Journal of Botany, 101*(1), 86-91. doi:10.3732/ajb.1300312

Arias, T., & Pires, J. C. (2012). A fully resolved chloroplast phylogeny of the *Brassica* crops and wild relatives (Brassicaceae: Brassiceae): Novel clades and potential taxonomic implications. *Taxon, 61*(5), 980-988.

Bailey-Serres, J., Parker, J. E., Ainsworth, E. A., Oldroyd, G. E. D., & Schroeder, J. I. (2019). Genetic strategies for improving crop yields. *Nature, 575*(7781), 109-118. doi:10.1038/s41586-019-1679-0

Beaumont, M. A., Zhang, W. Y., & Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics, 162*(4), 2025-2035. Retrieved from <Go to ISI>://WOS:000180502300043

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate - a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-Statistical Methodology, 57*(1), 289-300. doi:10.1111/j.2517-6161.1995.tb02031.x

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics, 29*(4), 1165-1188. Retrieved from <Go to ISI>://WOS:000172838100012

Berardini, T. Z., Reiser, L., Li, D. H., Mezheritsky, Y., Muller, R., Strait, E., & Huala, E. (2015). The *Arabidopsis* Information Resource: Making and mining the "gold standard" annotated reference plant genome. *Genesis, 53*(8), 474-485. doi:10.1002/dvg.22877

Bird, K. A., An, H., Gazave, E., Gore, M. A., Pires, J. C., Robertson, L. D., & Labate, J. A. (2017). Population Structure and Phylogenetic Relationships in a Diverse Panel of *Brassica rapa* L. *Frontiers in Plant Science, 8*. doi:10.3389/fpls.2017.00321

Blair, C., & Ane, C. (2020). Phylogenetic Trees and Networks Can Serve as Powerful and Complementary Approaches for Analysis of Genomic Data. *Systematic Biology, 69*(3), 593-601. doi:10.1093/sysbio/syz056

Bolger, A., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics, 30*, 2114-2120.

Bracha-Drori, K., Shichrur, K., Lubetzky, T. C., & Yalovsky, S. (2008). Functional analysis of *Arabidopsis* postprenylation CaaX processing enzymes and their function in subcellular protein targeting. *Plant Physiology, 148*(1), 119-131. doi:10.1104/pp.108.120477

Branca, F., & Cartea, E. (2011). Chapter 2 - *Brassica*. In C. Kole (Ed.), *Wild crop relatives: genomic and breeding resources oilseeds* (pp. 17-36). Berlin/Heidelberg: Springer.

Branca, F., & Tribulato, A. (2011). *Brassica* species. *The IUCN Red List of Threatened Species*, 2011-2011.

Cai, C., Bucher, J., Bakker, F. T., & Bonnema, G. (2022). Evidence for two domestication lineages supporting a middle-eastern origin for *Brassica oleracea* crops from diversified kale populations. *Horticulture Research, 9*. doi:10.1093/hr/uhac033

Castañeda-Álvarez, N. P., Khoury, C. K., Achicanoy, H. A., Bernau, V., Dempewolf, H., Eastwood, R. J., . . . Maxted, N. (2016). Global conservation priorities for crop wild relatives. *Nature Plants, 2*(4), 1-6.

Chahtane, H., Zhang, B., Norberg, M., LeMasson, M., Thévenon, E., Bakó, L., . . . Vachon, G. (2018). LEAFY activity is post-transcriptionally regulated by BLADE ON PETIOLE2 and CULLIN3 in Arabidopsis. *New Phytologist, 220*(2), 579-592. doi:https://doi.org/10.1111/nph.15329

Challinor, A. J., Watson, J., Lobell, D. B., Howden, S. M., Smith, D. R., & Chhetri, N. (2014). A meta-analysis of crop yield under climate change and adaptation. *Nature Climate Change, 4*(4), 287-291. doi:10.1038/nclimate2153

Cheng, F., Sun, R. F., Hou, X. L., Zheng, H. K., Zhang, F. L., Zhang, Y. Y., . . . Wang, X. W. (2016). Subgenome parallel selection is associated with morphotype diversification and convergent crop domestication in *Brassica rapa* and *Brassica oleracea*. *Nature Genetics, 48*(10), 1218-1224. doi:10.1038/ng.3634

Cheng, F., Wu, J., Cai, C. C., Fu, L. X., Liang, J. L., Borm, T., . . . Wang, X. W. (2016). Genome resequencing and comparative variome analysis in a *Brassica rapa* and *Brassica oleracea* collection. *Scientific Data, 3*. doi:10.1038/sdata.2016.119

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., . . . Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w(1118); iso-2; iso-3. *Fly, 6*(2), 80-92. doi:10.4161/fly.19695

Cornuet, J.-M., Pudlo, P., Veyssier, J., Dehne-Garcia, A., Gautier, M., Leblois, R., . . . Estoup, A. (2014). DIYABC v2.0: a software to make approximate Bayesian computation inferences about population history using single nucleotide polymorphism, DNA sequence and microsatellite data. *Bioinformatics, 30*, 1187-1189. doi:10.1093/bioinformatics/btt763

Cornuet, J.-M., Ravigné, V., & Estoup, A. (2010). Inference on population history and model checking using DNA sequence and microsatellite data with the software DIYABC (v1.0). *BMC Bioinformatics, 11*(1), 401. doi:10.1186/1471-2105-11-401

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., . . . Genomes Project Anal, G. (2011). The variant call format and VCFtools. *Bioinformatics, 27*(15), 2156-2158. doi:10.1093/bioinformatics/btr330

Darriba, D., Taboada, G. L., Doallo, R., & Posada, D. (2012). jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods, 9*(8), 772-772. doi:10.1038/nmeth.2109

Ding, Q., Cui, B., Li, J. J., Li, H. Y., Zhang, Y. H., Lv, X. H., . . . Gao, J. W. (2018). Ectopic expression of a *Brassica rapa AINTEGUMENTA* gene (*BrANT-1*) increases organ size and stomatal density in *Arabidopsis*. *Scientific Reports, 8*. doi:10.1038/s41598-018-28606-4

Dixon, G. R. (2006). Vegetable brassicas and related crucifers. In G. R. Dixon (Ed.), *Origins and diversity of Brassica and its relatives.* (pp. 1-33). doi:10.1079/9780851993959.0001

Doyle, J. J., & Doyle, J. L. (1990). Isolation of plant DNA from fresh tissue. *Focus, 12*, 13-15.

Du, Z., Zhou, X., Ling, Y., Zhang, Z., & Su, Z. (2010). agriGO: a GO analysis toolkit for the agricultural community. *Nucleic Acids Research, 38*, W64–W70.

Earl, D. A., & VonHoldt, B. M. (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation genetics resources, 4*(2), 359-361.

Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, *5*(1), 1-19.

Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology, 14*(8), 2611-2620.

Felsenstein J (1989) PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics*, 5 (2), 163-166.

FitzJohn, R. G., Armstrong, T. T., Newstrom-Lloyd, L. E., Wilton, A. D., & Cochrane, M. (2007). Hybridisation within *Brassica* and allied genera: evaluation of potential for transgene escape. *Euphytica, 158*(1-2), 209-230. doi:10.1007/s10681-007-9444-0

Flowers, J. M., Hazzouri, K. M., Gros-Balthazard, M., Mo, Z., Koutroumpa, K., Perrakis, A., . . . Purugganan, M. D. (2019). Cross-species hybridization and the origin of North African date palms. *Proceedings of the National Academy of Sciences of the United States of America, 116*(5), 1651-1658. doi:10.1073/pnas.1817453116

Francis, R. M. (2017). pophelper: an R package and web app to analyse and visualize population structure. *Molecular Ecology Resources, 17*(1), 27-32.

Francisco, M., Tortosa, M., Martinez-Ballesta, M. D., Velasco, P., Garcia-Viguera, C., & Moreno, D. A. (2017). Nutritional and phytochemical value of *Brassica* crops from the agri-food perspective. *Annals of Applied Biology, 170*(2), 273-285. doi:10.1111/aab.12318

Gaut, B. S., Seymour, D. K., Liu, Q. P., & Zhou, Y. F. (2018). Demography and its effects on genomic variation in crop domestication. *Nature Plants, 4*(8), 512-520. doi:10.1038/s41477-018-0210-1

Golicz, A. A., Bayer, P. E., Barker, G. C., Edger, P. P., Kim, H., Martinez, P. A., . . . Edwards, D. (2016). The pangenome of an agronomically important crop plant *Brassica oleracea*. *Nature Communications, 7,* 13390. doi:10.1038/ncomms13390

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology, 59*(3), 307-321. doi:10.1093/sysbio/syq010

Guo, N., Wang, S., Gao, L., Liu, Y., Wang, X., Lai, E., . . . Liu, F. (2021). Genome sequencing sheds light on the contribution of structural variants to *Brassica oleracea* diversification. *BMC Biology, 19*(1), 93. doi:10.1186/s12915-021-01031-2

Huang, X., Kurata, N., Wei, X., Wang, Z.-X., Wang, A., Zhao, Q., . . . Han, B. (2012). A map of rice genome variation reveals the origin of cultivated rice. *Nature, 490*(7421), 497-501.

Huang, X. H., Wei, X. H., Sang, T., Zhao, Q. A., Feng, Q., Zhao, Y., . . . Han, B. (2010). Genome-wide association studies of 14 agronomic traits in rice landraces. *Nature Genetics, 42*(11), 961-U976. doi:10.1038/ng.695

Izzah, N. K., Lee, J., Perumal, S., Park, J. Y., Ahn, K., Fu, D., . . . 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*(7), 1967-1986.

Janzen, G. M., Wang, L., & Hufford, M. B. (2019). The extent of adaptive wild introgression in crops. *New Phytologist, 221*(3), 1279-1288. doi:10.1111/nph.15457

Kagale, S., Robinson, S. J., Nixon, J., Xiao, R., Huebert, T., Condie, J., . . . Parkin, I. A. P. (2014). Polyploid Evolution of the Brassicaceae during the Cenozoic Era. *Plant Cell, 26*(7), 2777-2791. doi:10.1105/tpc.114.126391

Katche, E., Quezada-Martinez, D., Katche, E. I., Vasquez-Teuber, P., & Mason, A. S. (2019). Interspecific Hybridization for *Brassica* Crop Improvement. *Crop Breeding, Genetics and Genomics, 1*(1), e190007. doi:10.20900/cbgg20190007

Kaur, C., Kumar, K., Anil, D., & Kapoor, H. C. (2007). Variations in antioxidant activity in broccoli (*Brassica oleracea* L.) cultivars. *Journal of Food Biochemistry, 31*(5), 621-638. doi:10.1111/j.1745-4514.2007.00134.x

Kiefer, C., Willing, E. M., Jiao, W. B., Sun, H. Q., Piednoel, M., Humann, U., . . . Schneeberger, K. (2019). Interspecies association mapping links reduced CG to TG substitution rates to the loss of gene-body methylation. *Nature Plants, 5*(8), 846-855. doi:10.1038/s41477-019-0486-9

Kim, D. H., Kim, M., Oh, S. B., Lee, K. M., Kim, S. M., Nho, C. W., . . . Pan, C. H. (2017). The protective effect of antioxidant enriched fractions from colored potatoes against hepatotoxic oxidative stress in cultured hepatocytes and mice. *Journal of Food Biochemistry, 41*(1). doi:10.1111/jfbc.12315

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., . . . Marra, M. A. (2009). Circos: An information aesthetic for comparative genomics. *Genome Research, 19*(9), 1639-1645. doi:10.1101/gr.092759.109

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods, 9*(4), 357-359. doi:10.1038/nmeth.1923

Larsson, A. (2014). AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics, 30*(22), 3276-3278. doi:10.1093/bioinformatics/btu531

Larsson, J. (2020). eulerr: Area-Proportional Euler and Venn Diagrams with Ellipses (Version 6.1.0). https://cran.r-project.org/package=eulerr.

Lawrence, M., Huber, W., Pages, H., Aboyoun, P., Carlson, M., Gentleman, R., . . . Carey, V. J. (2013). Software for Computing and Annotating Genomic Ranges. *PLoS Computational Biology, 9*(8). doi:10.1371/journal.pcbi.1003118

Lee, T. H., Guo, H., Wang, X. Y., Kim, C., & Paterson, A. H. (2014). SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. *BMC Genomics, 15*. doi:10.1186/1471-2164-15-162

Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2018). Emmeans: Estimated marginal means, aka least-squares means. https://CRAN.R-project.org/package=emmeans.

Li, J. R., Zhang, X. M., Lu, Y., Feng, D. X., Gu, A. X., Wang, S., . . . Zhao, J. J. (2019). Characterization of Non-heading Mutation in Heading Chinese Cabbage (*Brassica rapa* L. ssp. *pekinensis*). *Frontiers in Plant Science, 10*. doi:10.3389/fpls.2019.00112

Lin, Z., Li, X., Shannon, L. M., Yeh, C.-T., Wang, M. L., Bai, G., . . . Yu, J. (2012). Parallel domestication of the *Shattering1* genes in cereals. *Nature Genetics, 44*(6), 720-724. doi:10.1038/ng.2281

Mabry, M. E., Turner-Hissong, S. D., Gallagher, E. Y., McAlvay, A. C., An, H., Edger, P. P., . . . Pires, J. C. (2021). The Evolutionary History of Wild, Domesticated, and Feral *Brassica oleracea* (Brassicaceae). *Molecular Biology and Evolution, 38*(10), 4419-4434. doi:10.1093/molbev/msab183

Maggioni, L., von Bothmer, R., Poulsen, G., & Lipman, E. (2018). Domestication, diversity and use of *Brassica oleracea* L., based on ancient Greek and Latin texts. *Genetic resources and crop evolution, 65*(1), 137-159.

Malinsky, M., Matschiner, M., & Svardal, H. (2020). Dsuite - Fast D-statistics and related admixture evidence from VCF files. *Molecular Ecology Resources*. doi:10.1111/1755-0998.13265

Martin, S. H., Davey, J. W., & Jiggins, C. D. (2015). Evaluating the use of ABBA–BABA statistics to locate introgressed loci. *Molecular Biology and Evolution, 32*(1), 244-257.

Mbow, C., Rosenzweig, C., Barioni, L. G., Benton, T. G., Herrero, M., Krishnapillai, M., . . . Sapkota, T. (2019). Food security. In *Climate Change and Land: an IPCC special report on climate change, desertification, land degradation, sustainable land management, food security and greenhouse gas fluxes in terrestrial ecosystems* (pp. 451-458): IPCC.

McAlvay, A. C., Ragsdale, A. P., Mabry, M. E., Qi, X. S., Bird, K. A., Velasco, P., . . . Emshwiller, E. (2021). *Brassica rapa* Domestication: Untangling Wild and Feral Forms and Convergence of Crop Morphotypes. *Molecular Biology and Evolution, 38*(8), 3358-3372. doi:10.1093/molbev/msab108

Mittell, E. A., Cobbold, C. A., Ijaz, U. Z., Kilbride, E. A., Moore, K. A., & Mable, B. K. (2020). Feral populations of *Brassica oleracea* along Atlantic coasts in western Europe. *Ecology and Evolution, 10*(20), 11810-11825. doi:https://doi.org/10.1002/ece3.6821

Nelson, G., Bogard, J., Lividini, K., Arsenault, J., Riley, M., Sulser, T. B., . . . Rosegrant, M. (2018). Income growth and climate change effects on global nutrition security to mid-century. *Nature Sustainability, 1*(12), 773-781. doi:10.1038/s41893-018-0192-z

Okazaki, K., Sakamoto, K., Kikuchi, R., Saito, A., Togashi, E., Kuginuki, Y., ... & Hirai, M. (2007). Mapping and characterization of FLC homologs and QTL analysis of flowering time in *Brassica oleracea*. *Theoretical and Applied Genetics*, *114*(4), 595-608.

Page, A. M. L., Gibson, J., Meyer, R. S., & Chapman, M. A. (2019). Eggplant domestication: pervasive gene flow, feralisation and transcriptomic divergence. *Molecular Biology and Evolution, 36*(7), 1359-1372. doi:doi.org/10.1093/molbev/msz062

Park, C. H., Yeo, H. J., Kim, N. S., Eun, P. Y., Kim, S.-J., Arasu, M. V., . . . Park, S. U. (2017). Metabolic profiling of pale green and purple kohlrabi (*Brassica oleracea* var. *gongylodes*). *Applied Biological Chemistry, 60*(3), 249-257.

Pavlidis, P., Zivkovic, D., Stamatakis, A., & Alachiotis, N. (2013). SweeD: Likelihood-Based Detection of Selective Sweeps in Thousands of Genomes. *Molecular Biology and Evolution, 30*(9), 2224-2234. doi:10.1093/molbev/mst112

Perrin, R. M., Wang, Y., Yuen, C. Y. L., Will, J., & Masson, P. H. (2007). WVD2 is a novel microtubule-associated protein in *Arabidopsis thaliana*. *Plant Journal, 49*(6), 961-971. doi:10.1111/j.1365-313X.2006.03015.x

Petropoulos, S. A., Sampaio, S. L., Di Gioia, F., Tzortzakis, N., Rouphael, Y., Kyriacou, M. C., & Ferreira, I. (2019). Grown to be Blue-Antioxidant Properties and Health Effects of Colored Vegetables. Part I: Root Vegetables. *Antioxidants, 8*(12). doi:10.3390/antiox8120617

Pfeifer, B., Wittelsburger, U., Ramos-Onsins, S. E., & Lercher, M. J. (2014). PopGenome: An Efficient Swiss Army Knife for Population Genomic Analyses in R. *Molecular Biology and Evolution, 31*(7), 1929-1936. doi:10.1093/molbev/msu136

Phophi, M. M., & Mafongoya, P. L. (2017). Constraints to vegetable production resulting from pest and diseases induced by climate change and globalization: a review. *Journal of Agricultural Science, 9*(10), 11-25.

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics, 155*(2), 945-959.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., . . . Daly, M. J. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics, 81*(3), 559-575.

Qi, X., An, H., Ragsdale, A. P., Hall, T. E., Gutenkunst, R. N., Chris Pires, J., & Barker, M. S. (2017). Genomic inferences of domestication events are corroborated by written records in *Brassica rapa*. *Mol Ecol, 26*(13), 3373-3388. doi:10.1111/mec.14131

RCoreTeam. (2015). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rodriguez, V. M., Soengas, P., Alonso-Villaverde, V., Sotelo, T., Cartea, M. E., & Velasco, P. (2015). Effect of temperature stress on the early vegetative development of *Brassica oleracea* L. *BMC Plant Biology, 15*. doi:10.1186/s12870-015-0535-0

Rosenzweig, B. K., Pease, J. B., Besansky, N. J., & Hahn, M. W. (2016). Powerful methods for detecting introgressed regions from population genomic data. *Molecular Ecology, 25*(11), 2387-2397. doi:10.1111/mec.13610

Skibinski, D. O. F., Rasool, D., & Erskine, W. (1984). Aspartate-Aminotransferase Allozyme Variation in a Germplasm Collection of the Domesticated Lentil (Lens-Culinaris). *Theoretical and Applied Genetics, 68*(5), 441-448. doi:Doi 10.1007/Bf00254816

Snogerup, S., Gustafsson, M., & von Bothmer, R. (1990). *Brassica* sect. *Brassica* (Brassicaceae). *Willdenowia, 19*, 271-365.

Song, K., Osborn, T. C., & Williams, P. H. (1990). *Brassica* taxonomy based on nuclear restriction fragment length polymorphisms (RFLPs). *Theoretical and Applied Genetics, 79*(4), 497-506. doi:10.1007/BF00226159

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics, 30*(9), 1312-1313. doi:10.1093/bioinformatics/btu033

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

Tanaka, H., Onouchi, H., Kondo, M., Hara-Nishimura, I., Nishimura, M., Machida, C., & Machida, Y. (2001). A subtilisin-like serine protease is required for epidermal surface formation in *Arabidopsis* embryos and juvenile plants. *Development, 128*(23), 4681-4689. Retrieved from <Go to ISI>://WOS:000172740900002

Terhorst, J., Kamm, J. A., & Song, Y. S. (2017). Robust and scalable inference of population history froth hundreds of unphased whole genomes. *Nature Genetics, 49*(2), 303-309. doi:10.1038/ng.3748

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., . . . DePristo, M. A. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current protocols in bioinformatics, 43*(1110), 11.10.11-11.10.33. doi:10.1002/0471250953.bi1110s43

Wang, M., Li, W. Z., Fang, C., Xu, F., Liu, Y. C., Wang, Z., . . . Tian, Z. X. (2018). Parallel selection on a dormancy gene during domestication of crops from multiple families. *Nature Genetics, 50*(10), 1435-1441. doi:10.1038/s41588-018-0229-2

Wang, X. B., Wu, J., Liang, J. L., Cheng, F., & Wang, X. W. (2015). *Brassica* database (BRAD) version 2.0: integrating and mining Brassicaceae species genomic resources. *Database-the Journal of Biological Databases and Curation*. doi:10.1093/database/bav093

Wen, D. Q., Yu, Y., Zhu, J. F., & Nakhleh, L. (2018). Inferring Phylogenetic Networks Using PhyloNet. *Systematic Biology, 67*(4), 735-740. doi:10.1093/sysbio/syy015

Widen, B., Andersson, S., Rao, G. Y., & Widen, M. (2002). Population divergence of genetic (co)variance matrices in a subdivided plant species, *Brassica cretica*. *Journal of Evolutionary Biology, 15*(6), 961-970. doi:10.1046/j.1420-9101.2002.00465.x

Winkel-Shirley, B. (2001). Flavonoid biosynthesis. A colorful model for genetics, biochemistry, cell biology, and biotechnology. *Plant Physiology, 126*(2), 485-493. doi:DOI 10.1104/pp.126.2.485

Yamaguchi, T., Nukazuka, A., & Tsukaya, H. (2012). Leaf adaxial-abaxial polarity specification and lamina outgrowth: evolution and development. *Plant and Cell Physiology, 53*(7), 1180-1194. doi:10.1093/pcp/pcs074

Yu, J. Y., Zhao, M. X., Wang, X. W., Tong, C. B., Huang, S. M., Tehrim, S., . . . Liu, S. Y. (2013). Bolbase: a comprehensive genomics database for *Brassica oleracea*. *BMC Genomics, 14*. doi:10.1186/1471-2164-14-664

Zhang, C., Dong, S. S., Xu, J. Y., He, W. M., & Yang, T. L. (2019). PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics, 35*(10), 1786-1788. doi:10.1093/bioinformatics/bty875

Zhang, H. Y., Mittal, N., Leamy, L. J., Barazani, O., & Song, B. H. (2017). Back into the wild-Apply untapped genetic diversity of wild relatives for crop improvement. *Evolutionary Applications, 10*(1), 5-24. doi:10.1111/eva.12434

Zhou, Z. K., Jiang, Y., Wang, Z., Gou, Z. H., Lyu, J., Li, W. Y., . . . Tian, Z. X. (2015). Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nature Biotechnology, 33*(4), 408-U125. doi:10.1038/nbt.3096

Figures

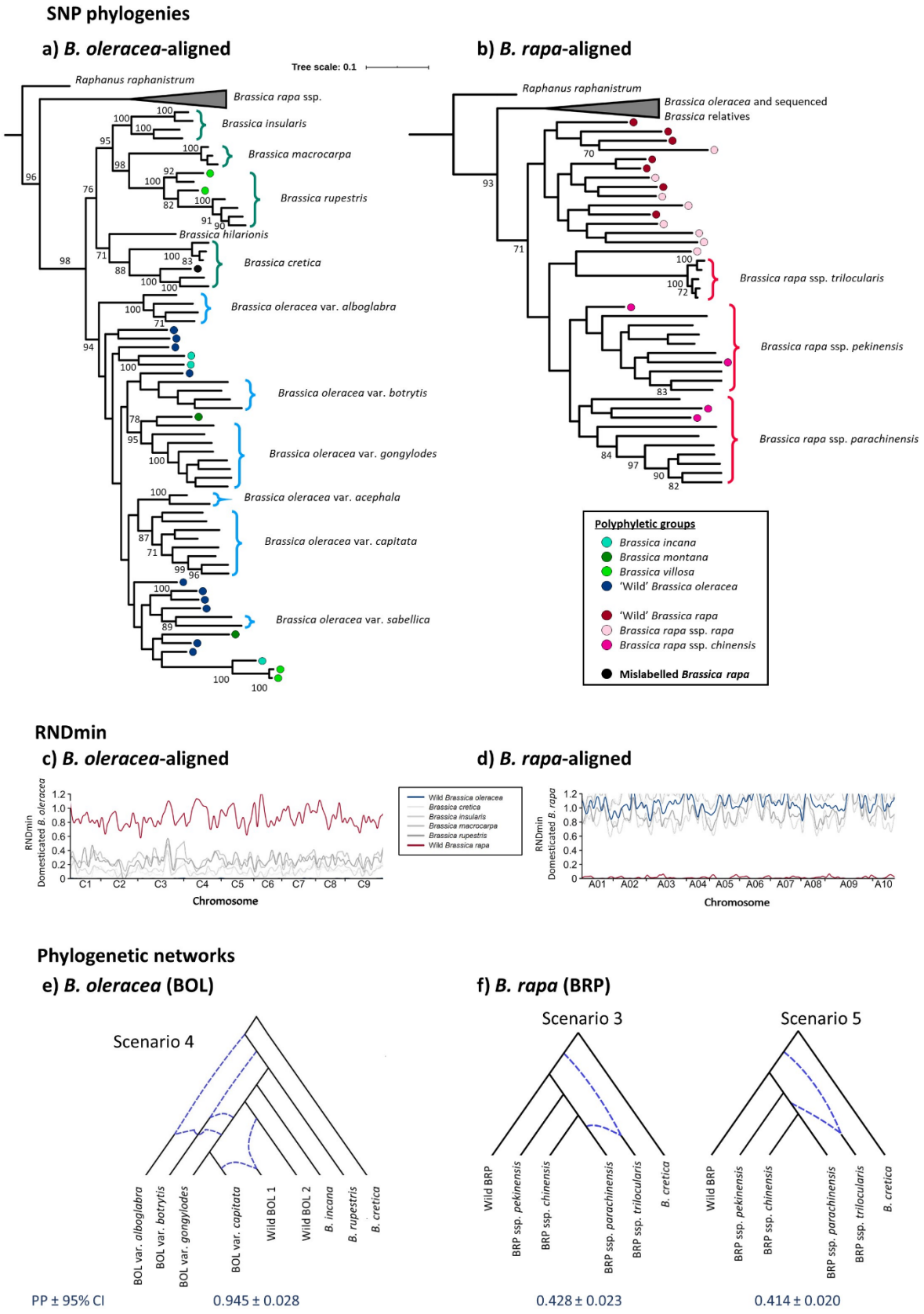
SNP phylogenies
a) B. oleracea-aligned
b) B. rapa-aligned
Tree scale: 0.1
Raphanus raphanistrum
Brassica rapa ssp.
Brassica insularis
Brassica macrocarpa
Brassica rupestris
Brassica hilarionis
Brassica cretica
Brassica oleracea var. alboglabra
Brassica oleracea var. botrytis
Brassica oleracea var. gongylodes
Brassica oleracea var. acephala
Brassica oleracea var. capitata
Brassica oleracea var. sabellica
Raphanus raphanistrum
Brassica oleracea and sequenced Brassica relatives
Brassica rapa ssp. trilocularis
Brassica rapa ssp. pekinensis
Brassica rapa ssp. parachinensis
Polyphyletic groups
Brassica incana
Brassica montana
Brassica villosa
'Wild' Brassica oleracea
'Wild' Brassica rapa
Brassica rapa ssp. rapa
Brassica rapa ssp. chinensis
Mislabelled Brassica rapa
RNDmin
c) B. oleracea-aligned
d) B. rapa-aligned
RNDmin Domesticated B. oleracea
Wild Brassica oleracea
Brassica cretica
Brassica insularis
Brassica macrocarpa
Brassica rupestris
Wild Brassica rapa
RNDmin Domesticated B. rapa
C1 C2 C3 C4 C5 C6 C7 C8 C9
A01 A02 A03 A04 A05 A06 A07 A08 A09 A10
Chromosome
Phylogenetic networks
e) B. oleracea (BOL)
f) B. rapa (BRP)
Scenario 4
BOL var. alboglabra
BOL var. botrytis
BOL var. gongylodes
BOL var. capitata
Wild BOL 1
Wild BOL 2
B. incana
B. rupestris
B. cretica
Scenario 3
Wild BRP
BRP ssp. pekinensis
BRP ssp. chinensis
BRP ssp. parachinensis
BRP ssp. trilocularis
B. cretica
Scenario 5
Wild BRP
BRP ssp. pekinensis
BRP ssp. chinensis
BRP ssp. parachinensis
BRP ssp. trilocularis
B. cretica
PP ± 95% CI
0.945 ± 0.028
0.428 ± 0.023
0.414 ± 0.020

Figure 1. Phylogenetic relationships and hybridisation within and between *Brassica oleracea* (a, c, e) and *Brassica rapa* (b, d, f), and their wild relatives.
(a, b) Maximum likelihood phylogenetic relationships based on single nucleotide polymorphisms (SNPs) filtered by linkage disequilibrium for samples mapped to (a) *B. oleracea* and (b) *B. rapa*. Polyphyletic groups are identified by coloured dots (see legend), and bootstrap values >70 are indicated. (c, d) RNDmin, a measure of the pairwise distance relative to an outgroup calculated in 50 kb windows for (c) domesticated *B. oleracea* versus wild relatives, and (d) domesticated *B. rapa* (excluding ssp. *rapa*) versus wild relatives. (e, f) Most likely phylogenetic networks identified for (e) *B. oleracea* (BOL) and (f) *B. rapa* (BRP), with dotted lines indicating admixture. Posterior probabilities (PP) with 95% confidence intervals are in blue.

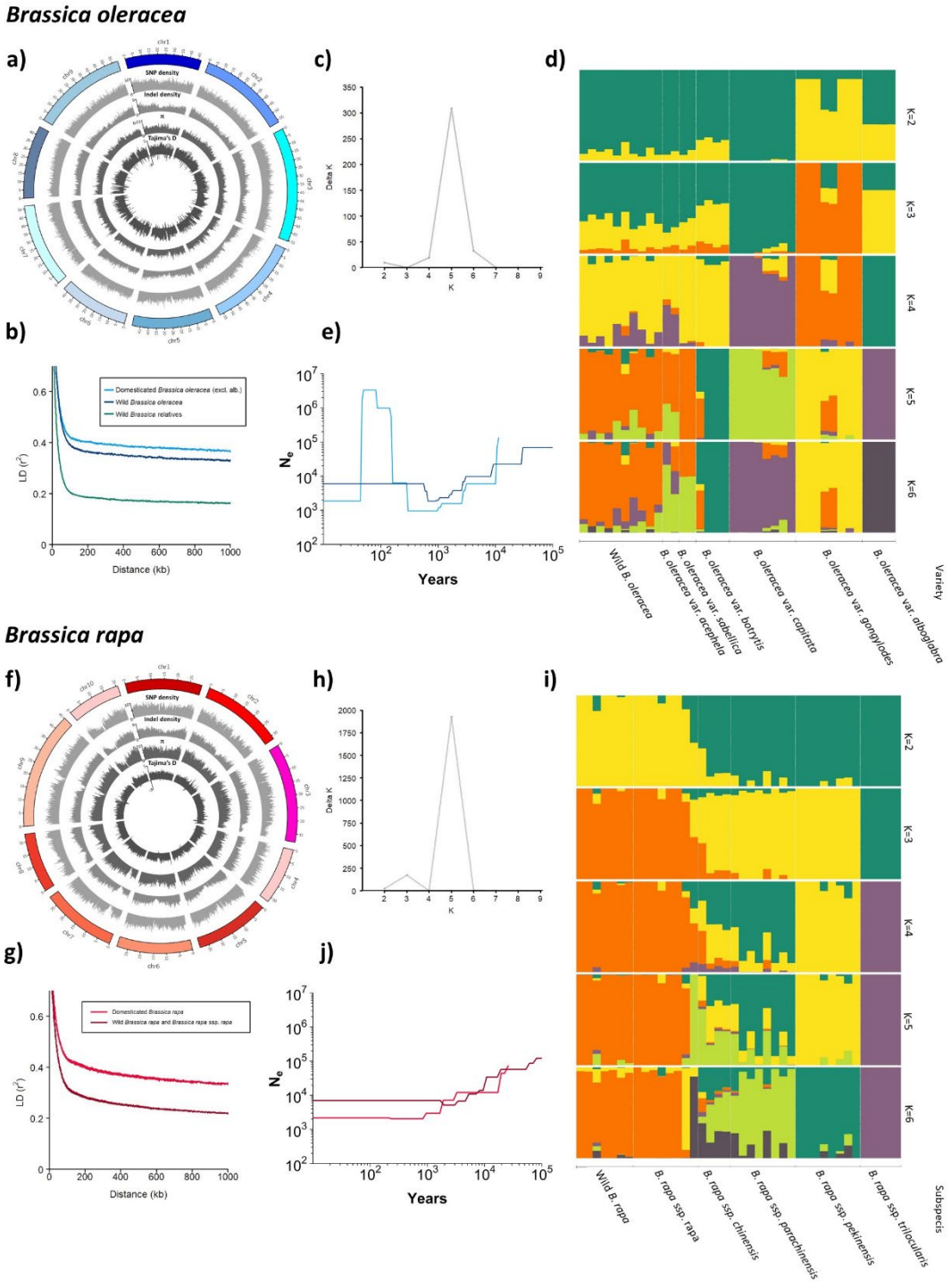

Figure 2. Population genetic statistics and population structure of wild and domesticated *Brassica oleracea* (a-e) and *Brassica rapa* (f-j).

(a, f) Distribution of population genetic statistics across the genome, (b, g) linkage disequilibrium decay, (c, h) Evanno's delta K for STRUCTURE analyses, (d, i) STRUCTURE analysis, with colours representing the proportional assignment of each individual to each of the K clusters, (e, j) demographic history inference of effective population size over time.

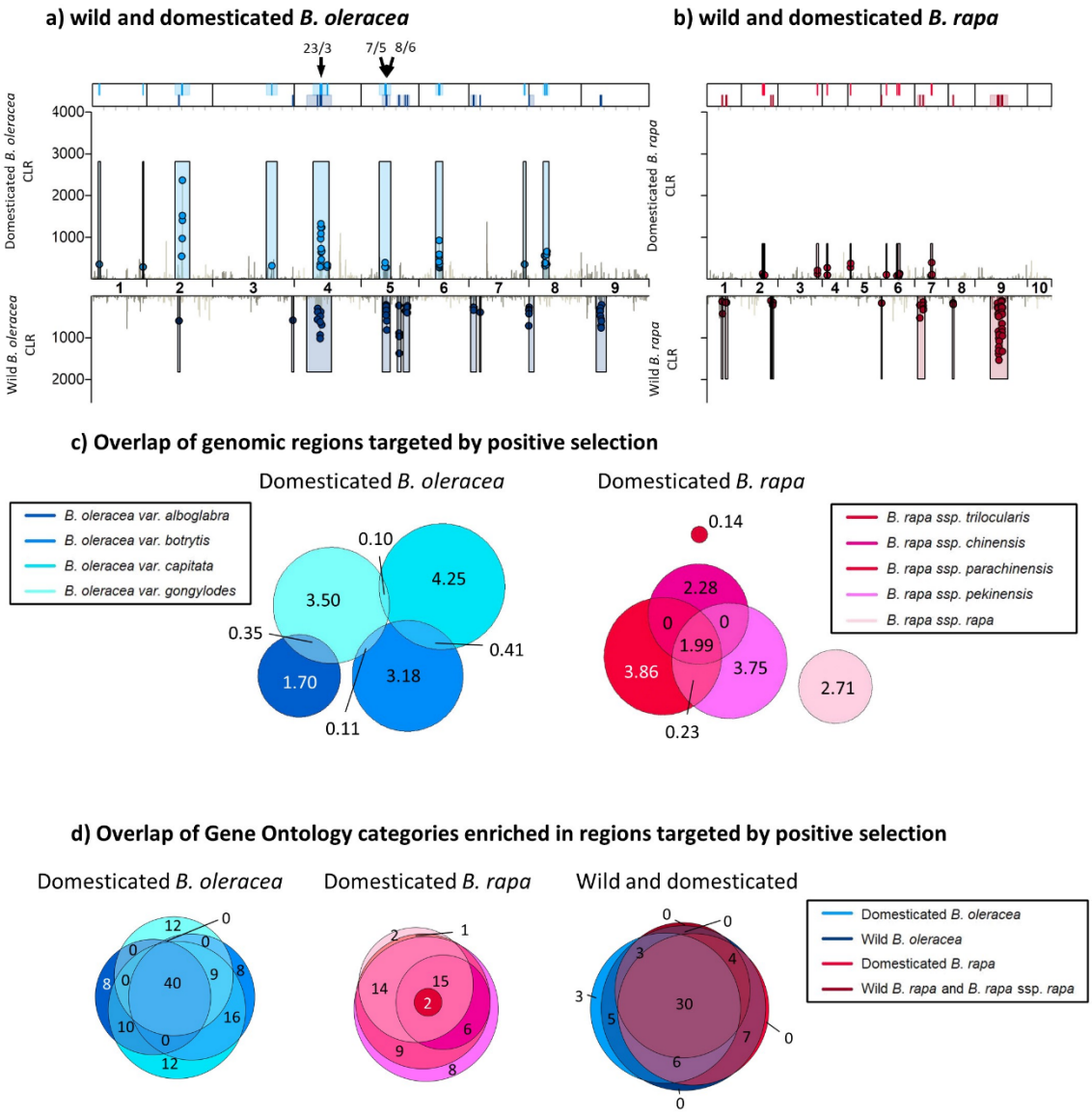

Figure 3. Signatures of selection in *Brassica oleracea* and *Brassica rapa*.
(a, b) Overlap between genomic regions targeted by positive selection in (a) wild (bottom) and domesticated (top) *B. oleracea* (excluding *B. oleracea var. alboglabra*), and (b) overlap between regions targeted by positive selection in combined wild (bottom, including *B. rapa* ssp. *rapa*) and domesticated (top) *B. rapa*. CLR values in the top 1% of both the CLR (Sweed) and ω-statistic (Omegaplus) are highlighted as red or blue points. Shaded boxes define windows around these points that maximise CLR. Bars at the top show the location of these windows affected by selection (light) and the likely targets of selection within them (dark). Overlapping regions are indicated with arrows and numbers indicate the number of genes in the overlap and the number with AT annotations. (c) Size (Mb) of genomic regions targeted by positive selection and their overlap between domesticated *B. oleracea* varieties, and between *B. rapa* subspecies. (d) Overlap in gene ontology categories that were enriched in regions targeted by positive selection.

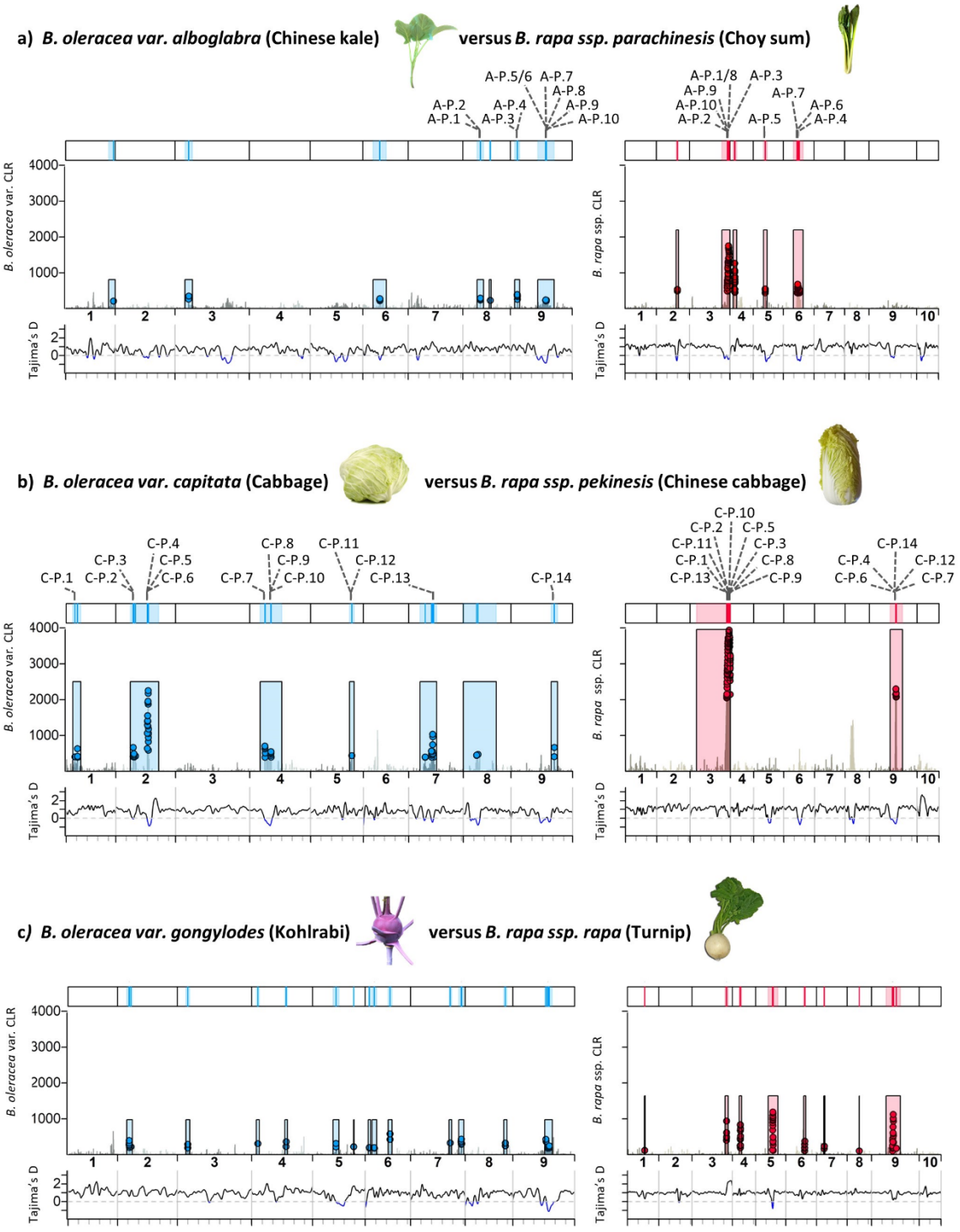

Figure 4. Evidence for parallel positive selection between pairs of *Brassica oleracea* domesticates (left) and *Brassica rapa* domesticates (right) with similar phenotypes. CLR values in the top 1% CLR values and top 1% of w-statistic values highlighted as blue and red points for *B. oleracea* and *B. rapa* respectively. Shaded boxes define windows around these points that maximise CLR. The top bars show the location of these windows affected by selection (light) and the candidate target regions of selection within them (dark). Positions of putative *rapa-oleracea* orthologues in target selection regions according to reciprocal BLAST are indicated (full gene information is given in SI Appendix, Table S11). Tajima's D is plotted below with negative values highlighted in blue.

SI Legends

Methods S1 Extended materials and methods.
Figure S1 Maximum likelihood phylogeny of *Brassica* species based on single nucleotide polymorphisms (filtered by linkage disequilibrium) identified by mapping resequencing data of 108 samples to the *Brassica oleracea* pangenome.
Figure S2 Maximum likelihood phylogeny of *Brassica* species based on single nucleotide polymorphisms (filtered by linkage disequilibrium) identified by mapping resequencing data of 77 samples to the *Brassica rapa* ssp. *pekinensis* genome.
Figure S3 Average genome-wide relative minimum distance (RNDmin) between domesticated *Brassica* crops and wild monophyletic *Brassica* species relative to outgroup *Raphanus raphanistrum*.
Figure S4 Signals of introgression between *Brassica oleracea* varieties and wild *Brassica* relatives as a heatmap of significant (P<0.05, FDR correction) D-statistics.
Figure S5 Pseudolikelihood of models inferred for zero to five reticulations in phylogenetic network analysis of *Brassica oleracea.*
Figure S6 Phylogenetic networks identified as having the highest pseudolikelihood for number of reticulations 5:0 (a-f) in analysis of *Brassica oleracea.*
Figure S7 Signals of introgression between *Brassica rapa* varieties and wild *Brassica* relatives as a heatmap of significant (P<0.05, FDR correction) D-statistics.
Figure S8 Pseudolikelihood of models inferred for zero to five reticulations in phylogenetic network analysis of *Brassica rapa.*
Figure S9 The five phylogenetic networks with highest pseudolikelihood for one reticulation in analysis of *Brassica rapa* phylogenies.
Figure S10 Density distribution of annotation values informing SNP discovery for 108 individual samples aligned to the *Brassica oleracea* pangenome assembly.
Figure S11 Density distribution of annotation values informing INDEL discovery for 108 individual samples aligned to the *Brassica oleracea* pangenome assembly.
Figure S12 Density distribution of annotation values informing SNP discovery for 77 individual samples aligned to the *Brassica rapa* v3.0 genome assembly.
Figure S13 Density distribution of annotation values informing INDEL discovery for 77 individual samples aligned to the *Brassica rapa* v3.0 genome assembly.
Table S1 Meta-data for samples used in whole genome sequencing analysis. Accession information is provided along with an outline of the samples used in each analysis.
Table S2 Genome size of wild *Brassica* relatives determined using flow cytometry.
Table S3 Statistical analysis of differences in relative minimum distances between domesticated *B. oleracea* varieties and wild *Brassica* relatives.
Table S4 Statistical analysis of differences in relative minimum distances between domesticated *B. oleracea* varieties and wild *Brassica* relatives.
Table S5 D-statistics used to test for signals of introgression between *Brassica oleracea* varieties and wild *Brassica* relatives.
Table S6 Estimated genome-wide proportion of introgressed sites using the *fd* statistic in *Brassica oleracea* analyses.
Table S7 ABC model checking for Scenario 4 in network analysis of *Brassica oleracea* phylogenies.
Table S8 D-statistics used to test for signals of introgression between *Brassica rapa* varieties and wild *Brassica* relatives.

Table S9 Estimated genome-wide proportion of introgressed sites using the *fd* statistic in *Brassica rapa* analyses.
Table S10 ABC model checking for Scenario 3 in network analysis of one reticulation *Brassica rapa* phylogenies.
Table S11 ABC model checking for Scenario 5 in network analysis of one reticulation *Brassica rapa* phylogenies.
Table S12 Genes identified as overlapping with peaks of positive selection in both domesticated *B. oleracea* (excluding var. *alboglabra*) and Wild *B. oleracea*.
Tables S13-S25 Genes identified as overlapping with peaks of positive selection, and their associated AT identifier where applicable for all analyses: Wild *B. oleracea*, *B. oleracea* var. *alboglabra*, *B. oleracea* var. *capitata*, *B. oleracea* var. *botrytis*, *B. oleracea* var. *gongylodes*, combined domesticated *B. oleracea* (excl. alboglabra), Wild *B. rapa*, *B. rapa* ssp. *chinensis, B. rapa* ssp. *parachinensis, B. rapa* ssp. *pekinensis, B. rapa* ssp. *trilocularis, B. rapa* ssp. *rapa,* combined domesticated *B. rapa* (excluding ssp. rapa).
Table S26-S28 GO categories identified as enriched for genes in positive selection peaks for comparison of: Wild and domesticated *B. oleracea* and *B.rapa*, domesticated *B.oleracea* varieties, and domesticated *B. rapa* subspecies. All enriched GO categories identified in the analysis are listed with the presence or absence of this term in the enrichment of each population indicated with 1 or 0.
Table S29 Descriptions for putative *B. oleracea – B. rapa* orthologues (identified as reciprocal best blast pairs) in peaks of positive selection of domesticates with similar phenotypes.
Table S30-S31 Genes of interest identified as putative orthologues under parallel selection (reciprocal best blast) or annotated as "anatomical structure development" genes for *B. oleracea* and *B. rapa.*

Tables

Table 1: Summary statistics for *Brassica oleracea* only and *Brassica rapa* only datasets.

Dataset	***B. oleracea*-aligned**	***B. rapa*-aligned**
Reference sequence	*B. oleracea* pangenome	*B. rapa* ssp. *pekinensis* v3.0
Number of individuals	38	41
No. SNPs post-filtering	8,113,885	5,862,399
No. indels post-filtering	1,001,661	815,569
Percentage intergenic SNPs	88.6 %	85.9 %
Percentage exonic SNPs	7.5 %	8.9 %
Percentage intronic SNPs	3.9 %	5.2 %
Mean non-synonymous to synonymous ratio	0.781	0.566