# Genes derived from ancient polyploidy have higher genetic diversity and are associated with domestication in *Brassica rapa*

Xinshuai Qi[1,4], Hong An[2,3], Tara E. Hall[1], Chenlu Di[1], Paul D. Blischak[1], J. Chris Pires[2], and Michael S. Barker[1]

1. Department of Ecology & Evolutionary Biology, University of Arizona, Tucson, Arizona, USA

2. Division of Biological Sciences, University of Missouri, Columbia, Missouri, USA

3. These authors contributed equally to this work

Correspondence should be addressed to M. S. B. (msbarker@arizona.edu)

**Abstract**

Many crops are polyploid or have a polyploid ancestry. Recent phylogenetic analyses have found that polyploidy often preceded the domestication of crop plants. One explanation for this observation is that increased genetic diversity following polyploidy may have been important during the strong artificial selection that occurs during domestication. To test the connection between domestication and polyploidy, we identified and examined candidate genes associated with the domestication of the diverse crops of *Brassica rapa*. Like all "diploid" flowering plants, *B. rapa* has a diploidized paleopolyploid genome and experienced many rounds of whole genome duplication (WGD). We analyzed transcriptome data of more than hundred cultivated *B. rapa* accessions. Using a combination of approaches, we identified more than 3,000 candidate genes associated with the domestication of four major *B. rapa* crops. Consistent with our expectation, we found that the candidate genes were significantly enriched with genes derived from the Brassiceae mesohexaploidy. We also observed that paleologs contained significantly more genetic diversity than non-paleologs, suggesting that elevated genetic variation may explain why paleologs are enriched among domestication candidate genes. Our analyses demonstrate the key role of polyploidy in the domestication of *B. rapa* and provide support for its importance in the success of modern agriculture.

## Introduction

Polyploidy, or whole genome duplication (WGD), has long been associated with crop domestication and diversity[1–8]. Many desirable crop traits such as larger seed size, greater stress tolerance, and increased disease resistance are often attributed to polyploidy[2,9]. A recent phylogenetic analysis found that domesticated plants have experienced significantly more polyploidy than their wild relatives[10]. Polyploidy often precedes domestication and crops are nearly twice as likely to be domesticated in lineages with a relatively recent WGD compared to those without[10]. Among the potential explanations for the relationship between polyploidy and domestication, the expanded genetic diversity and plasticity of polyploid plants may be especially advantageous during domestication and crop improvement[11–15]. Analyses in yeast have shown that polyploid lineages not only have higher genetic diversity but also adapt to new environments faster than their lower ploidal level relatives[16]. Similarly, the niches of polyploid plants evolve faster than their diploid relatives[17]. These features may collectively give polyploids unique advantages over diploids during domestication and the global spread of crops that occurred with human population expansion.

Although nearly 30% of plant species are recent polyploids, all flowering plants are paleopolyploids with varying histories of WGD[18]. Given that the

genetic consequences of polyploidy play out over time as genomes diploidize and paralogs fractionate[13,19,20], we may expect that the effects of polyploidy extend to diploidized species. Here, we sought to test whether past polyploidy is associated with increased diversity and domestication in the crops of *Brassica rapa*. Like all flowering plants, the genome of *B. rapa* has been multiplied and fractionated many times over. The most recent polyploidization in the ancestry of *B. rapa* was a mesohexaploidy that occurred approximately 9–28 MYA[21–27]. Further, *B. rapa* has been domesticated into many different types of crops across Europe and Asia. These include turnips, oil seeds, pak choi, Chinese cabbage, and mustard seeds. Many researchers have suggested that there is a connection between the mesohexaploidy and the diversity of *B. rapa* crops[26,28], but the relationship has never been explicitly tested. Using recently sequenced transcriptomes from a diverse array of *B. rapa* accessions[29], we tested if polyploid-derived regions of the genome are enriched with candidate genes associated with domestication. We also compared genetic variation in the polyploid vs non-polyploid derived regions of the *B. rapa* genome. Given the frequency of ancient polyploidy and its contribution to the evolution of plants, our analyses demonstrate the key role of polyploidy in the domestication of *B. rapa* and provide support for its importance in the success of modern agriculture.

## Results

### Partitioning the *B. rapa* genome into paleologs vs non-paleologs

To test the contribution of paleopolyploidy to the domestication of *Brassica rapa*, we used an integrated approach to classify genes as paleologs—genes derived from the *Brassica* mesohexaploidy—or non-paleologs. An initial list of putative paleologs was generated from pairs of genes with synonymous divergence (Ks) values that correspond to the peak of duplication associated with the Brassiceae mesohexaploidy. Using this approach, we identified 21,280 genes that are likely derived from the mesohexaploidy (Fig. 1a). This result is consistent with previous *Ks* estimates[30,31]. Ancient WGDs and their associated paralogs may also be classified by identifying syntenic blocks of duplication. As an alternative to our paralog divergence approach, we used synteny analyses in CoGe to identify putative paleologs. Syntenic analyses recovered 19,810 syntenic gene sets that contained 31,796 paleologs. Finally, we compared these two paleolog lists with a previously curated list of 23,716 paleologs reported by an independent research group[32]. Genes that appeared at least twice in these three lists were classified as paleologs and were used in further analyses (Fig. 1b). The final paleolog list included 27,919 genes (Supplementary Table 1), which represents 68.06% of the 41,020 annotated genes in the *B. rapa* reference genome (Version 1.5). We also considered the 5,424 genes that were not present in any of the above three

paleolog lists as non-paleologs (Fig. 1b), which represents 13.22% of the total *B. rapa* genes. Paleologs and non-paleologs are distributed throughout the *B. rapa* genome (Fig. 2).

**Analyses of selection identify candidate genes associated with domestication of *B. rapa* crops**

We used a diverse collection of sequenced transcriptome data from across five groups to identify candidate genes associated with the domestication and improvement of *B. rapa* crops. In total, 1.32 million SNPs from transcriptome data with 25X coverage or greater from 102 *B. rapa* accessions were analyzed (Supplementary Table 2). Based on our previous population genomic analyses of these data[29], the accessions are comprised of a European-Central Asian *B. rapa* population represented by turnip (TN, *B. rapa* subsp. *rapa*) and four derived groups represented by pak choi (PC, *B. rapa* subsp. *chinensis*), Chinese cabbage (CC, *B. rapa* subsp. *pekinensis*), Indian sarson (IS, *B. rapa* subsp. *trilocularis*) and toria (TO, *B. rapa* subsp. *dichotoma*). We analyzed these data with an ensemble of molecular evolution and population genomic approaches to identify candidate genes associated with each group.

During domestication and crop improvement, we expect genetic variation important for agricultural traits to be selected in crop populations. To identify genes or genomic regions that have experienced recent positive selection, we used two different approaches. First, we used a selective sweep test, SweeD[33], to identify regions associated with significantly reduced variation consistent with a recent selective sweep in each of the crops. We identified 3,387 unique genes within the identified selective sweep regions (Fig. 3a, Supplementary Tables 3 and 4, and Supplementary Fig. 1 and 2). This included 1,072 genes in *chinensis*, 687 genes in *pekinensis*, 1,048 genes in toria, 1,228 genes in the sarsons, and 1,713 genes in European-Central Asian group. On average 70% of these genes were found in only one crop (Fig. 3a), indicating that many of these genes may have swept during the putative independent domestication of each crop. We also used the McDonald-Kreitman test (M-K test)[34] to identify coding regions with a significant excess of fixed amino acid substitutions, a different signature of positive selection. We used *B. oleracea* as an outgroup. Analyzing each crop with the M-K test, we found 92 genes in total with a molecular evolutionary signature of positive selection(Fig. 3b and Supplementary Tables 3 and 4). All but one of these genes was uniquely identified in a different crop, similar to the selective sweep analysis and consistent with independent domestication and differentiation of these *B. rapa* crops. Overall, we identified 3,479 candidate genes

that may be associated with the domestication and improvement of the five *B. rapa* crop groups.

We may also expect significant changes in gene expression to occur during domestication and crop improvement that may not be apparent in analyses of positive selection. To identify candidate genes with changes in gene expression during domestication, we performed differential gene expression (DGE) analyses with the four derived *B. rapa* genetic groups. RNA-seq data for each individual was collected 20 days after germination and the expression levels of genes in each group were compared to gene expression levels in the TN group. We identified a total of 7,813 genes with significant differential expression when comparing each group with the TN population (Fig. 3c, Supplementary Tables 3 and 4, and Supplementary Fig. 2 and 3). Unlike the analyses of selection above, the number of unique differentially expressed genes varied widely among the different crops. The sarsons (IS) were the most differentiated from TN and the other crops with over 61% of their differentially expressed genes unique. The sarsons also had the largest number of differentially expressed genes at 5,804. In contrast, pak choi (PC) had the smallest number of differentially expressed genes, 1301, and the lowest percentage that were uniquely different at 14.2%. Other crops fell between these values. The genes identified by all three methods were largely different among all the crops (Fig. 3d) with only a small number of genes identified by one

or more of these approaches (i.e., sweep test, M-K test, or DGE analyses). Similarly, most of the genes found to be associated with recent positive selection or differential gene expression were unique to a particular crop lineage as expected with independent domestication and differentiation of these distinct crops.

Across all of the analyses, only five genes were repeatedly present in tests among one or more of the crops (Fig. 3d and Supplementary Table 5). These genes—*Bra010933*, *Bra023190*, *Bra031452, Bra003055*, and *Bra035693*—were identified in some combination of the three inference methods with evidence for both signatures of positive selection and changes in gene expression. Considering that we recovered these five genes using different approaches, they may play an important role during the differentiation and improvement of *B. rapa* crops. Searches of these genes and their *Arabidopsis* homologs in the STRING[35] and UniProt[36] databases recovered a diverse range of potential functions. *Bra023190* is uncharacterized in *B. rapa*, but is homologous with *SGR2* in *Arabidopsis thaliana*, a gene associated with negative gravitropism and leaf movement in darkness[37,38]. Notably, this gene was identified as a candidate gene in both of the cabbage crops analyzed here, *B. rapa chinensis* and *pekinensis*. Other genes are homologs with *A. thaliana* genes implicated in responses to oxidative stress from photooxidation (*Bra010933*), salt stress (*Bra031452*[39]), and a vacuolar

V-type proton ATPase (*Bra035693*). Finally, *Bra003055*, identified by all three candidate gene approaches in the sarsons, was previously found to be over-expressed in *B. rapa* in soils that are deficient in iron and with an excess of zinc[40].

**Paleologs from the *Brassica* mesohexaploidy and domestication**

To test if the candidate genes associated with domestication of the *B. rapa* crops are enriched with paleologs, we compared the number of paleologs to the expected number from our genome-wide survey. Based on our classification of *B. rapa* genes as paleologs or non-paleologs (Fig. 1), we expected that approximately 68% of the candidate genes should be paleologs by chance. We found that the candidate genes were significantly enriched for paleologs from the Brassiceae mesohexaploidy (Fig. 4). Genes from all three candidate gene approaches were enriched across all crops with paleologs comprising 78.54% of SweeD ($\chi^2 = 160.06$, $p < .001$), 83.7% of McDonald Kreitman ($\chi^2 = 10.33$, $p = .0013$), and 78.6% of differentially expressed candidate genes ($\chi^2 = 345.40$, $p < .001$). These results indicate that genes derived from the Brassiceae mesohexaploidy were preferentially selected during the domestication of the *B. rapa* crops.

To further test if domestication genes in *B. rapa* are enriched for paleologs, we also developed a list of candidate genes from the literature. We focused on studies published over the last 10 years that identified genes through other approaches, such as fine mapping or bulk segregant analysis, to better establish a causal relationship between loci and crop traits. In total, we identified 40 candidate genes that fit these criteria from the literature (Supplementary Table 5). Many of these genes are associated with leaf and seed color variation, clubroot resistance, and cuticular wax biosynthesis. Notably, 15 of these genes were recovered in our candidate gene scans (Supplementary Table 5). Four of these genes were identified in our selective sweep and differential gene expression analyses. Mapping studies previously identified these genes as being associated with leaf color variation (Bra006208)[41], cuticular wax biosynthesis (Bra011470 and Bra032670)[42,43], and clubroot resistance (Bra019410)[44]. For other loci, mapping studies have identified a small collection of candidate genes in target regions. For example, *Rcr5* is a gene of major effect for clubroot resistance in *Brassica rapa*. A recent bulk segregant analysis and fine mapping study found eight genes in the *Rcr5* target region of *B. rapa*[45]. In our analyses, three genes in the center of their region were found to be significantly differentially expressed. These results suggest that future mapping studies in *B. rapa* may be able to leverage our candidate gene lists to improve gene identification.

How many of the candidate genes from our literature survey were paleologs? Of the 40 genes identified in the literature, 36 were paleologs (Fig. 4, $\chi^2 = 8.85$, $p = .0029$). This is a similar level of paleolog enrichment as our three different candidate gene lists. Given these literature-based candidate genes are largely identified with mapping based approaches rather than genomic scans, these results suggest that the paleolog enrichment observed in our analyses is not likely an artifact of the inference methods. Overall, our results indicate that paleologs were an important source of variation for domestication in *B. rapa*.

Why did artificial selection preferentially target paleologs? One possible explanation is that these genes may harbor more genetic diversity due to their paralogous history over the past 20 million years. To test this hypothesis, we examined the nucleotide diversity per gene across the *B. rapa* genome. Only reads that were uniquely mapped were used to estimate nucleotide diversity to minimize error from incorrectly mapped reads. We found that the genes derived from the mesohexaploidy had, on average, four times the nucleotide diversity of the non-paleolog fraction of the genome (Fig. 5). The mean nucleotide diversity of paleologs (mean = $0.408 \times 10^{-3}$) is approximately four times larger than that of non-paleologs (mean = $0.111 \times 10^{-3}$). Further, we observed that both nonsynonymous and synonymous diversity of the paleologs was higher than the non-paleologs. The increased genetic diversity of paleologs in *B. rapa* may have

been important for the rapid response of these plants to artificial selection during domestication.

## Discussion

Our results establish a connection between ancient polyploidy and contemporary diversity and domestication in crops of *Brassica rapa*. Although polyploidy has long been hypothesized to be important for crop domestication, phylogenetic analyses have only recently confirmed that it is a key factor[10]. However, the mechanisms linking polyploidy and domestication itself have remained unresolved. Our results provide support for at least one genetic mechanism linking polyploidy and domestication. Paleologs, genes retained in duplicate from an ancient polyploidy, were enriched in our candidate gene lists for crops of *B. rapa*. They were enriched in all three of our lists from genome scans using different statistical approaches, as well as a list developed from fine mapping and other genetic studies in the literature. In the case of *B. rapa*, the polyploid event occurred 9–28 MYA[21–27], long before crops were domesticated by humans. Many of these crops were only domesticated in the last few thousand years[29], but we still find that paleologs are over-represented in a diverse range of

candidate gene lists. This suggests that even ancient polyploidy may contribute to domestication and potentially adaptation long after genomes have diploidized.

Why are paleologs over-represented among the candidate genes for *B. rapa* domestication and crop improvement? Our results suggest that it may be because this class of genes harbors more genetic diversity than other genes in the genome. Paleologs had nearly four times the nucleotide diversity of non-paleologs in *B. rapa*. Considering that the paleologs have been maintained in duplicate since the mesohexaploidy that occurred 9–28 MYA, relaxed selection may yield elevated diversity at these genes as observed in other studies of paralog evolution[46–56]. Consistent with relaxed selection, we found that nucleotide diversity was elevated for both synonymous and nonsynonymous substitutions in *B. rapa* paleologs compared to non-paleologs. Further research is needed to better understand why the paleologs in *B. rapa* have elevated nucleotide diversity. The elevated diversity in the paleologs does not appear to be an artifact of read mapping error given that we only used uniquely mapped reads and had relatively long Illumina read lengths (150 bp). Furthermore, the paralogs from the Brassicaeae mesohexaploidy have nearly 30% synonymous divergence. Regardless of the cause of the increased variation, it likely explains why these genes are over-represented in the candidate gene lists. Increased genetic diversity is expected to be associated with greater phenotypic variation that could be

selected during domestication. Paleologs are likely over-represented in our candidate gene lists simply because they contain more genetic diversity than non-paleologs in *B. rapa*. Although many paleopolyploid plant genomes have been analyzed, this difference in genetic diversity has not yet been observed, and to the best of our knowledge, no previous study has explored this dimension of diversity.

The presence of elevated genetic variation in *B. rapa* paleologs suggests that domestication and crop improvement may have proceeded largely from standing variation. Rapid responses to selection, such as the response to selection to domestication and crop improvement, will proceed much faster if standing genetic diversity is high enough to facilitate simultaneous selection at multiple sites[57–59]. It is an open question how much domestication has proceeded from standing genetic diversity, but a recent analysis suggests that maize could have been domesticated from standing variation in teosinte[60]. There are also examples of candidate genes associated with crops, including in *Brassica oleracea*[61], that appear to be selected from standing genetic diversity. Given the elevated diversity of paleologs in *B. rapa*, we may expect domestication in these crops to be dominated by soft sweeps[57,62]. A complicating factor in characterizing hard and soft sweeps in *B. rapa* at the moment is the significant difference in genetic variation between paleologs and non-paleologs that may skew such

diversity based analyses. More sophisticated approaches that leverage recent advances in deep learning[63–65] that are trained to account for differences in variation because of gene origins will likely overcome these issues. Ultimately, this will allow us to characterize the genetics of domestication in the diverse crops of *B. rapa* while providing new insight into how paleopolyploidy may influence the architecture of adaptation in plant genomes.

The ancient hexaploidy in the Brassiceae has been hypothesized to be the source of the outstanding diversity of *Brassica* crops[22,26,28]. Our results support this hypothesis by finding that candidate genes for domestication of *B. rapa* crops are enriched in regions of the genome duplicated in the mesohexaploidy. It remains to be seen if paleologs are also significantly enriched in the candidate domestication genes in the crops of other *Brassica* species. *Brassica* are known as the dogs of the plant world for the incredible morphological diversity and number of different domesticated crops compared to other plants[26,66,67]. Experimental evolution studies in *B. rapa* have also been able to rapidly select for new pollination syndromes[68,69]. The relatively high genetic diversity of paleologs in *Brassica* may play a significant role in this morphological diversity and the rapid responses to selection. However, paleopolyploidy is common among flowering plants with the average species experiencing nearly five rounds of WGD in its ancestry[18]. If elevated variation in paleologs is a general

phenomenon, then we would expect it to be important in the domestication of many other crops and it may not completely explain the outstanding diversity of *Brassica* crops. Further analyses of paleolog variation and enrichment in other crops are needed to understand the generality of our findings and the implications for the domestication of diverse crops like in *Brassica*. Presently, our results provide a testable hypothesis to explain the observed correlation of polyploidy and crop domestication[10].

More broadly, our results suggest that paleopolyploidy may leave behind a legacy of elevated genetic diversity across the duplicated remnants of diploidized genomes. Although most models and studies of polyploid evolution compare diploids and polyploids[11,12,14,16,17,19,70–72], our comparison of paleologs and non-paleologs within a diploidized paleopolyploid uncovered evidence for similar dynamics ongoing within plant genomes even millions of years after whole genome duplication. The extensive genome duplication history of plants may result in genomes with different levels of diversity based on the mechanisms of gene origin. Although this remains to be broadly tested, it raises a few testable predictions. If this is a general phenomenon in plants, then we may expect that diploid plants with more paleologs will have higher genetic diversity than those with fewer paleologs. As paleologs are lost over time due to fractionation and gene turnover, we would also expect genetic diversity in diploid plants to be

correlated with the time since their most recent paleopolyploid event. The long-term effects of paleopolyploidy on genetic diversity observed here in *B. rapa* may also explain a broader phenomenon, the lag time of paleopolyploidy and diversification[73]. Recent research has found that paleopolyploidy in plants is associated with diversification rate increases, but these rate increases often occur many millions of years following WGDs[74]. If relatively high genetic diversity is important for the adaptation and expansion that leads to macroevolutionary signatures of net diversification rate increases, then the observed lag between diversification and polyploidy observed in many studies could be explained by the long increase in diversity at paleologs. Although our analyses are centered on how ancient polyploidy contributed to the diversity of *B. rapa*, additional analyses in other plants will provide crucial data to test the new hypotheses described above. Given the distribution of polyploidy throughout the history of flowering plants[18], our results suggest that the genetic legacy of these WGDs likely contributes to the diversity and adaptation of plants millions of years later.

## Methods

**Data Sources.** RNA-seq data for this study were previously generated by our group[29]. Based on previous population genomics analyses(Qi et al 2017), a total of 102 *B. rapa* accessions (Supplementary Table 2) were selected from the USDA

GRIN database (http://www.ars-grin.gov/) or from the author's collection to represent the five major *B. rapa* genetic groups. The five genetic groups include an earlier derived Europe and Central Asia group, represented by turnip (TN, *B. rapa* and *B. rapa* subsp. *rapa*, 22 accessions); four derived *B. rapa* groups, represented by Pak choi (PC, *B. rapa* subsp. *chinensis*, 25 accessions), Chinese cabbage (CC, *B. rapa* subsp. *pekinensis*, 28 accessions), Indian sarson (IS, *B. rapa* subsp. *trilocularis* and *B. rapa* subsp. *dichotoma* ; 20 accessions) and toria (TO, 7 accessions). The last four genetic groups diverged from the TN group about 2400–4100 years ago after the initial *B. rapa* domestication in European-Central Asia(Qi et al., 2017). Two *B. oleracea* accessions (SRR630924 and SRR1032050) from the NCBI Sequence Read Archive (SRA) were used as outgroups for analyses.

**RNA-seq Variant Calling.** Raw reads were cleaned using Trimmomatic version 0.32 [75], and mapped to the reference *B. rapa* genome (version 1.5, http://brassicadb.org/brad/) with Tophat version 2.0.14[76]. SNP calling was performed using Samtools version 0.1.18 and bcftools version 0.1.17[77,78]. The resulting VCF files were filtered with the vcfutils.pl script and vcffilter. Only SNPs with depth greater than 10 and variant quality (QUAL) greater than 30 were

retained, which included 1.32 million SNPs. Details about sequencing and variant calling were described previously[29].

**Identifying Genes Derived from the Mesohexaploidy Event (Paleologs).** To obtain a list of genes derived from the *Brassica* mesohexaploidy event (paleologs), two approaches were applied based on the gene age distribution and gene synteny, respectively. For the gene age distribution of duplications (also called Ks plot) based approach, *B. rapa* CDS sequences obtained from the *Brassica* database (http://brassicadb.org/brad/) were used to infer orthologous gene groups with OrthoFinder (https://github.com/davidemms/OrthoFinder)[79]. The synonymous substitution rate (Ks) of each orthologous gene pair was calculated using DupPipe[80,81]. The gene age distribution was visualized in R using histograms. The boundaries of the hexaploid peak were determined using EMMIX[82] by fitting a mixture model of normal distributions to the Ks data[83]. One hundred random starting points and 10 k-means starting points were used to identify the number of normal distributions (from 1 to 10). The best fit model was selected based on the Bayesian information criterion (BIC) value. Ks nodes with >50% likelihood assignment to the mesohexaploidy peak were considered as gene pairs derived from the *Brassica* hexaploidy event. For the synteny based approach, syntenic gene sets were generated by CoGe SynMap

(genomevolution.org/CoGe/SynMap.pl) with a 2: 2 quota-align ratio and default parameters[84] using the unmasked *B. rapa* genome (Version 1.5). These two paleolog lists were integrated with an independently estimated *B. rapa* paleolog list[32]. Genes present in at least two of these three lists were extracted using the Venn online tool (http://bioinformatics.psb.ugent.be/webtools/Venn/) and were considered as high confidence paleologs. Genes only present in one of these three lists were considered low confidence paleologs and were not used in our study. Genes not present in any of the above three paleologs lists were considered non-paleologs.

**Selection analyses.** We used two different methods to identify genes with signals of positive selection in the four derived *B. rapa* genetic groups. Genomic regions with evidence consistent with selective sweeps were detected using SweeD 3.0[33] based on the composite likelihood ratio (CLR) test of SNP site frequency spectrum (SFS) patterns. A total of 1.32 million SNPs were used in the SweeD tests for the TN group and each of the four derived *B. rapa* genetic groups. These analyses were performed with the default settings except each chromosome was divided to 60,000 grids. The CLR was calculated for each equally generated relative position on each chromosome. Only the top 1% significant outlier

regions were considered. Genes within the outlier regions were then annotated based on *B. rapa* genome v1.5.

We also used the McDonald-Kreitman test (M-K test)[34] to identify genes experiencing positive selection. This method is based on the proportion of synonymous and nonsynonymous substitutions within and between species that are due to directional selection. For the M-K test, we compared SNPs specific to each of the four *B. rapa* crop groups with the two *B. oleracea* outgroup accessions. For each separate SNP dataset, we annotated synonymous and nonsynonymous SNPs using SnpEff v4.2[85]. The reference genome database was created using *B. rapa* genome v1.5. The significance of the M-K test was evaluated using Fisher's exact test.

**Differential gene expression analysis.** Differential gene expression analysis was performed following the Tuexdo protocol[86]. Briefly, cleaned reads were first mapped to the *B. rapa* genome with TopHat[76], then mapped BAM files were assembled in Cufflinks[87], merged with Cuffmerge, quantified in Cuffdiff[88] (Cufflinks version2.2.1) with multiple-testing corrected q-values as 0.05, and finally indexed and visualized in CummeRbund version 2.14[89]. Genes significantly differentially expressed between the TN group and the four derived *B. rapa* groups were identified using the getSig function in CummeRbund[89].

**Candidate genes from the literature.** To confirm our findings, we compared our observations of paleolog enrichment with a list of candidate genes from published studies of B. rapa. We surveyed the literature for fine mapping and bulk segregant analyses of B. rapa that mapped traits to one or a few candidate genes. These studies and the candidate genes are listed in Supplemental Table 6. Overall, our survey identified 40 candidate genes in *B. rapa* that have been mapped to the location of classical loci associated with crop traits.

**Statistical analyses.** To test the association of paleologs and our candidate gene lists, we aggregated the number of genes in each of the four derived *B. rapa* genetic groups and the number of total unique candidate genes present in the four groups were counted. Then we summed up the observed percentages of candidate genes (SweeD, M-K test, GDE, and candidate genes) within the paleolog and non-paleologs lists. Meanwhile, the expected numbers of candidate genes within the paleologs and non-paleologs list for each gene list were estimated based on the percentage of paleologs and non-paleologs in the full genome. The significance of the deviation of the observed ratio from the expected ratio was assessed using a chi-square test.

To demonstrate the logical relations among these lists, Venn diagrams were generated using the online Venn diagram tool (http://bioinformatics.psb.ugent.be/webtools/Venn/). The density distribution of the identified paleologs, non-paleologs, SweeD outlier gene, M-K test outlier gene and differentially expressed gene were visualized in Circos with a bin size of 100 kbp.

**Nucleotide diversity of *B. rapa* genes.** The nucleotide diversity (π) of each paleolog and non-paleolog gene was estimated using a custom script from output of VCFtools[90]. The results were summarized and visualized in R.

**Data availability.** All the data sets generated during the current study are available in the NCBI Sequence Read Archive (SRA) under accession number SRP072186 (http://www.ncbi.nlm.nih.gov/sra/SRP072186). The accession numbers are summarized in Supplementary Table 2.

**Acknowledgements**

We thank A. Baniaga, G. Finch, Z. Li, M. McKibben, and B. Sutherland for feedback and comments on drafts of this manuscript. Hosting infrastructure and services provided by the Biotechnology Computing Facility (BCF) at the

University of Arizona. This research was supported by NSF-IOS-1339156 to M.S.B. and J.C.P.

**Author Contributions**

H.A. generated the RNA-seq data. X.Q. conduced gene age distribution, candidate gene survey, and all statistics. X.Q. and T.E.H. conducted the differential gene expression analyses. X.Q. and C.D. conducted the gene diversity analysis. X.Q. and M.S.B. designed the experiments and wrote the manuscript with input from all co-authors. All authors read and approved the manuscript.

**Competing Financial Interests**

The authors declare no competing financial interests.

**References**

1. Meyer, R. S., DuVal, A. E. & Jensen, H. R. Patterns and processes in crop domestication: an historical review and quantitative analysis of 203 global food crops. *New Phytol.* **196**, 29–48 (2012).
2. Lewis, W. H. Polyploidy in Species Populations. in *Polyploidy: Biological Relevance* (ed. Lewis, W. H.) 103–144 (Springer US, 1980).
3. Hilu, K. W. Polyploidy and the Evolution of Domesticated Plants. *Am. J. Bot.*

**80**, 1494–1499 (1993).

4. Heiser, C. B. *Seed to Civilization: The Story of Food*. (Harvard University Press, 1990).
5. Anderson, E. *Plants, Man and Life*. (University of California Press, 1969).
6. Renny-Byfield, S. & Wendel, J. F. Doubling down on genomes: polyploidy and crop plants. *Am. J. Bot.* **101**, 1711–1725 (2014).
7. Udall, J. A. & Wendel, J. F. Polyploidy and crop improvement. *Crop Sci.* (2006).
8. Paterson, A. H. Polyploidy, evolutionary opportunity, and crop adaptation. *Genetica* **123**, 191–196 (2005).
9. Levin, D. A. Polyploidy and Novelty in Flowering Plants. *Am. Nat.* **122**, 1–25 (1983).
10. Salman-Minkov, A., Sabath, N. & Mayrose, I. Whole-genome duplication as a key factor in crop domestication. *Nat Plants* **2**, 16115 (2016).
11. Otto, S. P. & Whitton, J. Polyploid incidence and evolution. *Annu. Rev. Genet.* **34**, 401–437 (2000).
12. Monnahan, P. *et al.* Pervasive population genomic consequences of genome duplication in Arabidopsis arenosa. *Nat Ecol Evol* **3**, 457–468 (2019).
13. Baduel, P., Bray, S. & Vallejo-Marin, M. The‘ Polyploid Hop’: shifting challenges and opportunities over the evolutionary lifespan of genome

duplications. *Front. Ecol. Environ.* (2018).

14. Paape, T. *et al.* Patterns of polymorphism and selection in the subgenomes of the allopolyploid Arabidopsis kamchatica. *Nat. Commun.* **9**, 3909 (2018).
15. Shimizu-Inatsugi, R. *et al.* Plant adaptive radiation mediated by polyploid plasticity in transcriptomes. *Mol. Ecol.* **26**, 193–207 (2017).
16. Selmecki, A. M. *et al.* Polyploidy can drive rapid adaptation in yeast. *Nature* **519**, 349–352 (2015).
17. Baniaga, A. E., Marx, H. E., Arrigo, N. & Barker, M. S. Polyploid plants have faster rates of multivariate niche differentiation than their diploid relatives. *Ecol. Lett.* **59**, 2473 (2019).
18. One Thousand Plant Transcriptomes Initiative. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* (2019). doi:10.1038/s41586-019-1693-2
19. Otto, S. P. The evolutionary consequences of polyploidy. *Cell* **131**, 452–462 (2007).
20. Arrigo, N. & Barker, M. S. Rarely successful polyploids and their legacy in plant genomes. *Curr. Opin. Plant Biol.* **15**, 140–146 (2012).
21. Beilstein, M. A., Nagalingum, N. S., Clements, M. D., Manchester, S. R. & Mathews, S. Dated molecular phylogenies indicate a Miocene origin for Arabidopsis thaliana. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 18724–18728 (2010).

22. Lukens, L. N. *et al.* Genome redundancy and plasticity within ancient and recent *Brassica* crop species: *Brassica* genome structure and plasticity. *Biol. J. Linn. Soc. Lond.* **82**, 665–674 (2004).

23. Lysak, M. A., Koch, M. A., Pecinka, A. & Schubert, I. Chromosome triplication found across the tribe Brassiceae. *Genome Res.* **15**, 516–525 (2005).

24. Wang, X. *et al.* The genome of the mesopolyploid crop species Brassica rapa. *Nat. Genet.* **43**, 1035–1039 (2011).

25. Arias, T., Beilstein, M. A., Tang, M., McKain, M. R. & Pires, J. C. Diversification times among Brassica (Brassicaceae) crops suggest hybrid formation after 20 million years of divergence. *Am. J. Bot.* **101**, 86–91 (2014).

26. Cheng, F., Wu, J. & Wang, X. Genome triplication drove the diversification of *Brassica* plants. *Hortic Res* **1**, 14024 (2014).

27. Franzke, A., Koch, M. A. & Mummenhoff, K. Turnip Time Travels: Age Estimates in Brassicaceae. *Trends Plant Sci.* **21**, 554–561 (2016).

28. Cheng, F. *et al.* Subgenome parallel selection is associated with morphotype diversification and convergent crop domestication in *Brassica rapa* and *Brassica oleracea*. *Nat. Genet.* (2016). doi:10.1038/ng.3634

29. Qi, X. *et al.* Genomic inferences of domestication events are corroborated by written records in *Brassica rapa*. *Mol. Ecol.* (2017). doi:10.1111/mec.14131

30. Cheng, F. *et al.* Deciphering the diploid ancestral genome of the

Mesohexaploid *Brassica rapa*. *Plant Cell* **25**, 1541–1554 (2013).

31. Barker, M. S., Vogel, H. & Schranz, M. E. Paleopolyploidy in the Brassicales: analyses of the *Cleome* transcriptome elucidate the history of genome duplications in *Arabidopsis* and other Brassicales. *Genome Biol. Evol.* **1**, 391–399 (2009).
32. Cheng, F. *et al.* Biased gene fractionation and dominant gene expression among the subgenomes of *Brassica rapa*. *PLoS One* **7**, e36442 (2012).
33. Pavlidis, P., Živkovic, D., Stamatakis, A. & Alachiotis, N. SweeD: likelihood-based detection of selective sweeps in thousands of genomes. *Mol. Biol. Evol.* **30**, 2224–2234 (2013).
34. McDonald, J. H. & Kreitman, M. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351**, 652–654 (1991).
35. Szklarczyk, D. *et al.* STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).
36. UniProt Consortium, T. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **46**, 2699 (2018).
37. Mano, E., Horiguchi, G. & Tsukaya, H. Gravitropism in leaves of *Arabidopsis thaliana* (L.) Heynh. *Plant Cell Physiol.* **47**, 217–223 (2006).
38. Kato, T. *et al.* SGR2, a phospholipase-like protein, and ZIG/SGR4, a SNARE,

are involved in the shoot gravitropism of *Arabidopsis*. *Plant Cell* **14**, 33–46 (2002).

39. Tuteja, N., Singh, S. & Tuteja, R. Helicases in Improving Abiotic Stress Tolerance in Crop Plants. in *Improving Crop Resistance to Abiotic Stress* (eds. Tuteja, N., Gill, S. S., Tiburcio, A. F. & Tuteja, R.) **6**, 435–449 (Wiley-VCH Verlag GmbH & Co. KGaA, 2012).
40. Li, J. *et al.* Expression profiling reveals functionally redundant multiple-copy genes related to zinc, iron and cadmium responses in *Brassica rapa*. *New Phytol.* **203**, 182–194 (2014).
41. Fu, W. *et al.* Fine mapping of lcm1, a gene conferring chlorophyll-deficient golden leaf in Chinese cabbage (*Brassica rapa* ssp. *pekinensis*). *Mol. Breed.* **39**, 52 (2019).
42. Wang, C., Li, Y., Xie, F., Kuang, H. & Wan, Z. Cloning of the *Brcer1* gene involved in cuticular wax production in a glossy mutant of non-heading Chinese cabbage (*Brassica rapa* L. var. *communis*). *Mol. Breed.* **37**, 142 (2017).
43. Wang, C. *et al.* Genetic characterization and fine mapping *BrCER4* involved in cuticular wax formation in purple cai-tai (*Brassica rapa* L. var. *purpurea*). *Mol. Breed.* **39**, 12 (2019).
44. Yu, F. *et al.* Identification of genome-wide variants and discovery of variants associated with *Brassica rapa* clubroot resistance gene *Rcr1* through bulked

segregant RNA sequencing. *PLoS One* **11**, e0153218 (2016).

45. Huang, Z., Peng, G., Gossen, B. D. & Yu, F. Fine mapping of a clubroot resistance gene from turnip using SNP markers identified from bulked segregant RNA-Seq. *Mol. Breed.* **39**, 131 (2019).
46. Aagaard, J. E., Willis, J. H. & Phillips, P. C. Relaxed selection among duplicate floral regulatory genes in Lamiales. *J. Mol. Evol.* **63**, 493–503 (2006).
47. Kondrashov, F. A., Rogozin, I. B., Wolf, Y. I. & Koonin, E. V. Selection in the evolution of gene duplications. *Genome Biol.* **3**, RESEARCH0008 (2002).
48. Innan, H. & Kondrashov, F. The evolution of gene duplications: classifying and distinguishing between models. *Nat. Rev. Genet.* **11**, 97–108 (2010).
49. Lynch, M. & Conery, J. S. The evolutionary fate and consequences of duplicate genes. *Science* **290**, 1151–1155 (2000).
50. Shan, H. *et al.* Evolution of plant MADS box transcription factors: evidence for shifts in selection associated with early angiosperm diversification and concerted gene duplications. *Mol. Biol. Evol.* **26**, 2229–2244 (2009).
51. Lee, H.-L. & Irish, V. F. Gene duplication and loss in a MADS box gene transcription factor circuit. *Mol. Biol. Evol.* **28**, 3367–3380 (2011).
52. Viaene, T. *et al.* Pistillata--duplications as a mode for floral diversification in (Basal) asterids. *Mol. Biol. Evol.* **26**, 2627–2645 (2009).
53. Ascencio, D., Ochoa, S., Delaye, L. & DeLuna, A. Increased rates of protein

evolution and asymmetric deceleration after the whole-genome duplication in yeasts. *BMC Evol. Biol.* **17**, 40 (2017).

54. Scannell, D. R. & Wolfe, K. H. A burst of protein sequence evolution and a prolonged period of asymmetric evolution follow gene duplication in yeast. *Genome Res.* **18**, 137–147 (2008).

55. Brunet, F. G. *et al.* Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Mol. Biol. Evol.* **23**, 1808–1816 (2006).

56. Conant, G. C. & Wagner, A. Asymmetric sequence divergence of duplicate genes. *Genome Res.* **13**, 2052–2058 (2003).

57. Messer, P. W. & Petrov, D. A. Population genomics of rapid adaptation by soft selective sweeps. *Trends Ecol. Evol.* **28**, 659–669 (2013).

58. Barrett, R. D. H. & Schluter, D. Adaptation from standing genetic variation. *Trends Ecol. Evol.* **23**, 38–44 (2008).

59. Matuszewski, S., Hermisson, J. & Kopp, M. Catch Me if You Can: Adaptation from Standing Genetic Variation to a Moving Phenotypic Optimum. *Genetics* **200**, 1255–1274 (2015).

60. Yang, C. J. *et al.* The genetic architecture of teosinte catalyzed and constrained maize domestication. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 5643–5652 (2019).

61. Purugganan, M. D., Boyles, A. L. & Suddith, J. I. Variation and selection at

the CAULIFLOWER floral homeotic gene accompanying the evolution of domesticated *Brassica oleracea*. *Genetics* **155**, 855–862 (2000).

62. Hermisson, J. & Pennings, P. S. Soft sweeps and beyond: understanding the patterns and probabilities of selection footprints under rapid adaptation. *Methods Ecol. Evol.* **8**, 700–716 (2017).
63. Flagel, L., Brandvain, Y. & Schrider, D. R. The Unreasonable Effectiveness of Convolutional Neural Networks in Population Genetic Inference. *Mol. Biol. Evol.* **36**, 220–238 (2019).
64. Kern, A. D. & Schrider, D. R. diploS/HIC: An Updated Approach to Classifying Selective Sweeps. *G3* **8**, 1959–1970 (2018).
65. Schrider, D. R. & Kern, A. D. Supervised Machine Learning for Population Genetics: A New Paradigm. *Trends Genet.* **34**, 301–312 (2018).
66. Liu, S. *et al.* The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes. *Nat. Commun.* **5**, 3930 (2014).
67. An, H. *et al.* Transcriptome and organellar sequencing highlights the complex origin and diversification of allotetraploid *Brassica napus*. *Nat. Commun.* **10**, 2878 (2019).
68. Gervasi, D. D. L. & Schiestl, F. P. Real-time divergent evolution in plants driven by pollinators. *Nat. Commun.* **8**, 14691 (2017).
69. Schiestl, F. P., Balmer, A. & Gervasi, D. D. Real-time evolution supports a

unique trajectory for generalized pollination. *Evolution* (2018). doi:10.1111/evo.13611

70. Laport, R. G. & Ng, J. Out of one, many: The biodiversity considerations of polyploidy. *Am. J. Bot.* (2017). doi:10.3732/ajb.1700190
71. Han, T.-S. *et al.* Polyploidy promotes species diversification of Allium through ecological shifts. *New Phytol.* (2019). doi:10.1111/nph.16098
72. Ramsey, J. & Schemske, D. W. Neopolyploidy in Flowering Plants. *Annu. Rev. Ecol. Syst.* **33**, 589–639 (2002).
73. Schranz, M. E., Mohammadin, S. & Edger, P. P. Ancient whole genome duplications, novelty and diversification: the WGD Radiation Lag-Time Model. *Curr. Opin. Plant Biol.* **15**, 147–153 (2012).
74. Landis, J. B. *et al.* Impact of whole-genome duplication events on diversification rates in angiosperms. *Am. J. Bot.* **105**, 348–363 (2018).
75. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
76. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
77. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
78. Li, H. A statistical framework for SNP calling, mutation discovery,

association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).

79. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157 (2015).
80. Barker, M. S. *et al.* EvoPipes. net: bioinformatic tools for ecological and evolutionary genomics. *Evol. Bioinform. Online* **6**, 143 (2010).
81. Barker, M. S. *et al.* Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. *Mol. Biol. Evol.* **25**, 2445–2455 (2008).
82. McLachlan, G. J., Peel, D., Basford, K. E. & Adams, P. The EMMIX software for the fitting of mixtures of normal and t-components. *J. Stat. Softw.* **4**, 1–14 (1999).
83. Tiley, G. P., Barker, M. S. & Burleigh, J. G. Assessing the performance of Ks plots for detecting ancient whole genome duplications. *Genome Biol. Evol.* (2018). doi:10.1093/gbe/evy200
84. Tang, H. *et al.* Screening synteny blocks in pairwise genome comparisons through integer programming. *BMC Bioinformatics* **12**, 102 (2011).
85. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila*

*melanogaster* strain w1118; iso-2; iso-3. *Fly* **6**, 80–92 (2012).

86. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).

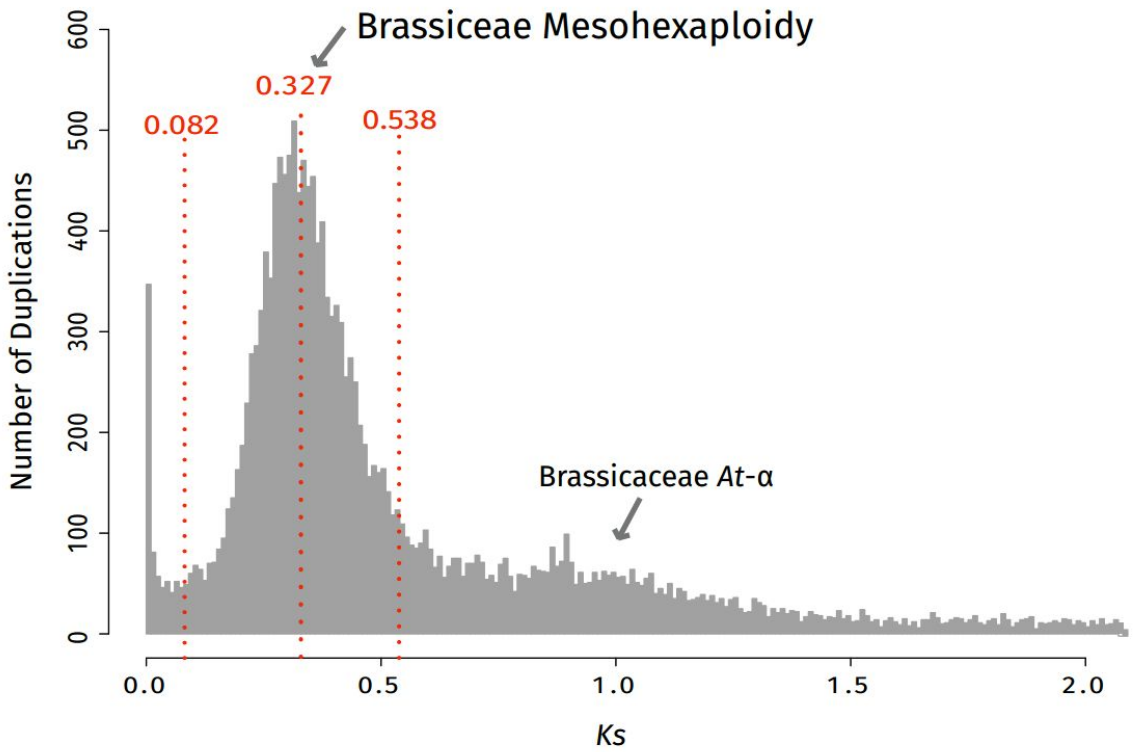
87. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).

88. Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.* **31**, 46–53 (2013).

89. Goff, L., Trapnell, C. & Kelley, D. cummeRbund: Analysis, exploration, manipulation, and visualization of Cufflinks high-throughput sequencing data. *R package version* (2012).

90. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

(a)


Brassiceae Mesohexaploidy
0.327
0.082
0.538
Brassicaceae At-α
Number of Duplications
0
100
200
300
400
500
600
0.0
0.5
1.0
1.5
2.0
Ks

(b)

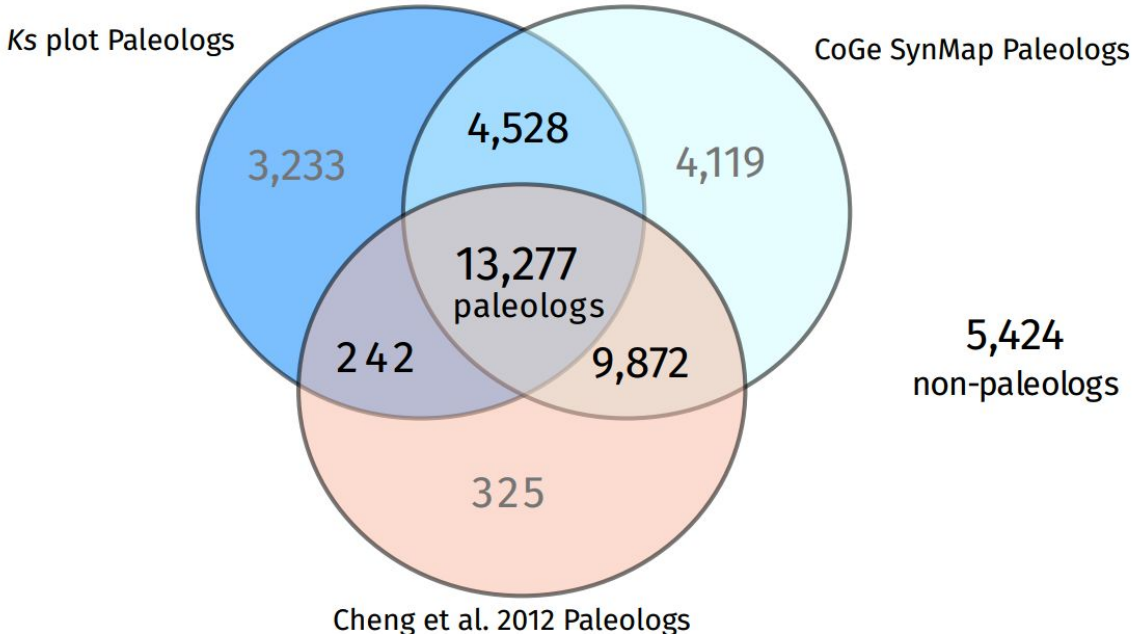

Ks plot Paleologs
CoGe SynMap Paleologs
Cheng et al. 2012 Paleologs
3,233
4,528
4,119
13,277
paleologs
242
9,872
325
5,424
non-paleologs

**Fig. 1: Identifying genes derived from the *Brassica* hexaploidy event.** (**a**) The age distribution of gene duplications from 41,020 CDSs in the *B. rapa* genome (version 1.5). The x-axis represents the synonymous divergence of duplication events (Ks value), whereas the y-axis represents the number of duplications. Gene pairs with Ks divergence 0.082–0.538 were identified as putative paleologs. (**b**) Venn diagram showing the overlap among our two *B. rapa* paleolog lists with a previously reported *B. rapa* paleolog list. Genes that appeared at least twice in these three lists were considered high confidence paleologs and were used in further analyses. Genes that were not present in any of the three paleologs lists were considered to be non-paleologs.

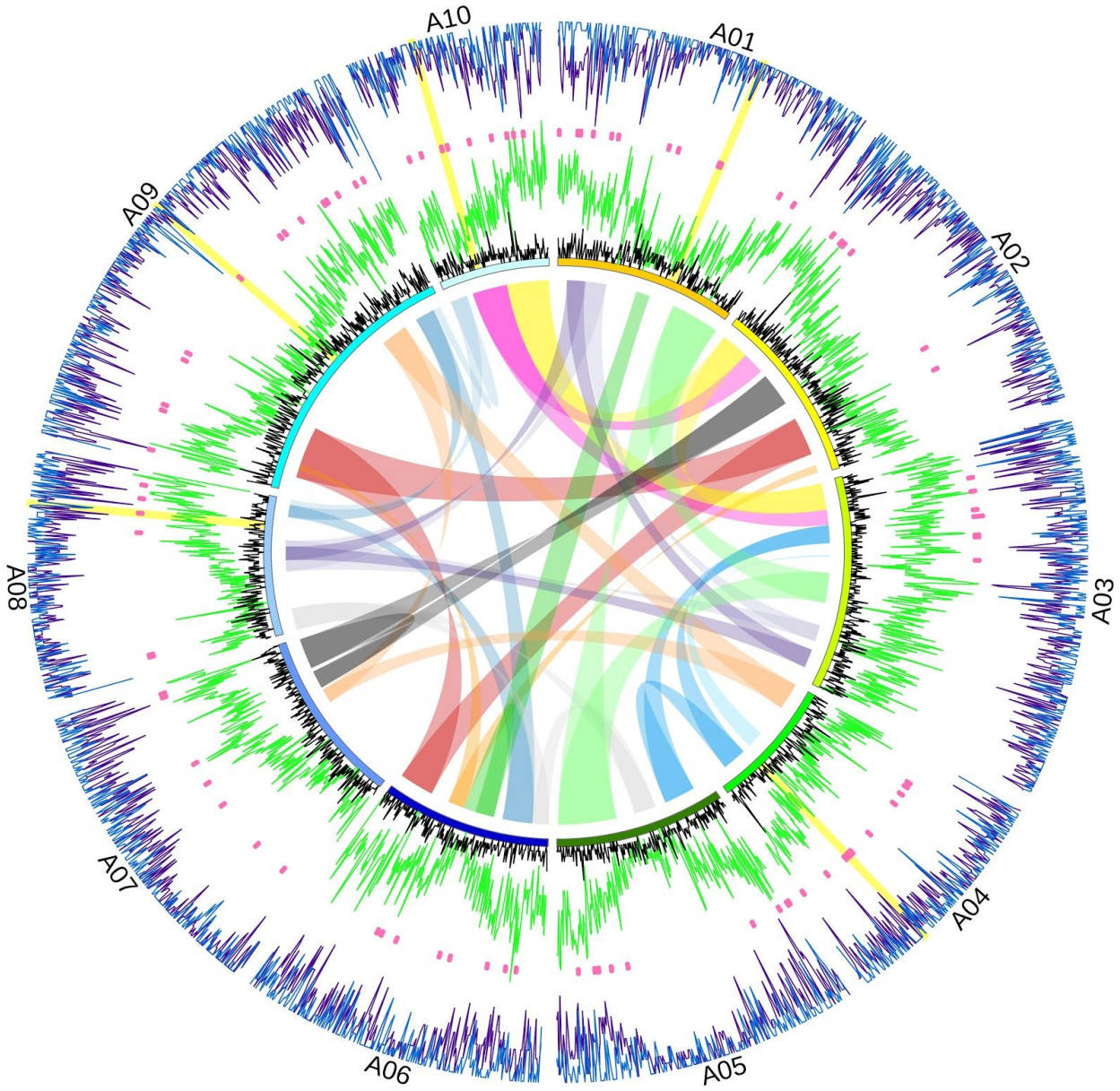



**Fig. 2: Circos plot of the distribution of surveyed genes in this study.** Highlighted on the outside of the plot are the *Brassica* hexaploidy paleologs (light green lines), non-paleologs (black lines), and candidate genes identified by SweeD (blue lines), the McDonald-Kreitman test (pink dots), and the significantly

differentially expressed genes between the EU-CA group and the four derived *B. rapa* groups (purple lines). A01-A10 represent the 10 chromosomes of *B. rapa*. The rainbow ribbons in the center represent the syntenic regions among chromosomes. The yellow bars represent the location of the five *B. rapa* domestication candidate genes listed in Table S6.

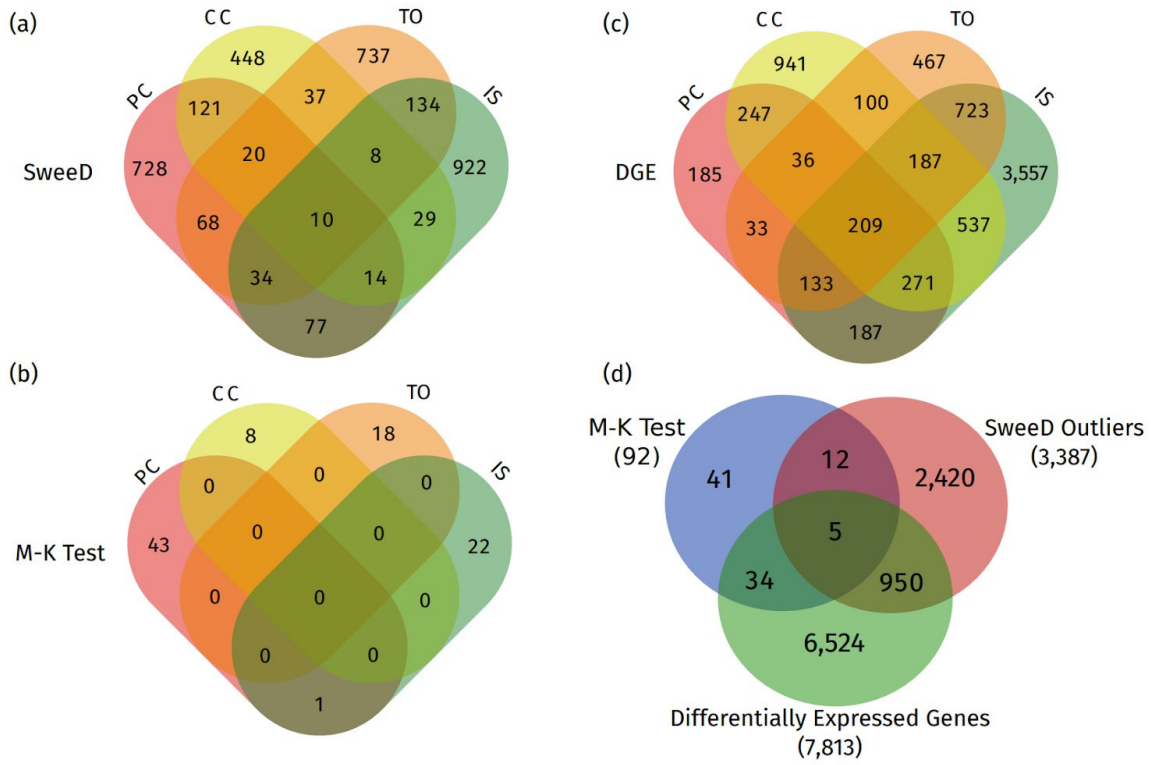



**Fig. 3: Overlap among domestication candidate genes.** PC = pak choi, CC = Chinese cabbage, TO = toria, and IS = Indian sarson. Number and overlap of candidate genes for each crop inferred by (**a**) SweeD, (**b**) the McDonald-Kreitman test, and (**c**) differential gene expression analyses. (**d**) Total number and overlap of candidate genes inferred by different methods across all crops.

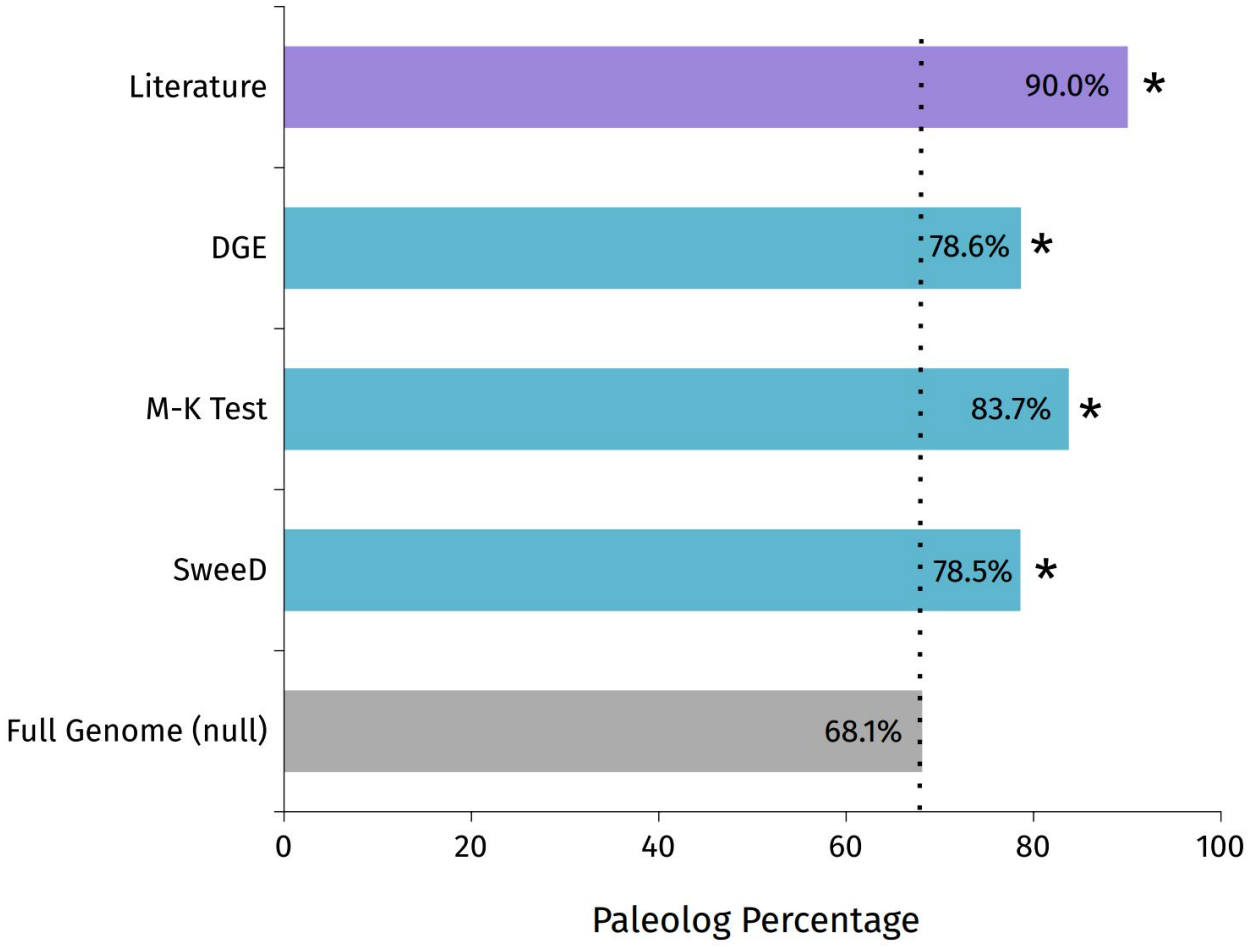



**Fig. 4: Comparison of paleolog percentages across candidate gene lists.** Percentage of paleologs in the entire *B. rapa* genome (gray), our three candidate gene analyses (blue), and the literature survey of candidate genes from mapping studies. The dashed line indicates the null expectation and asterisks indicate significant deviation from the null based on chi-square tests ($p < 0.05$).

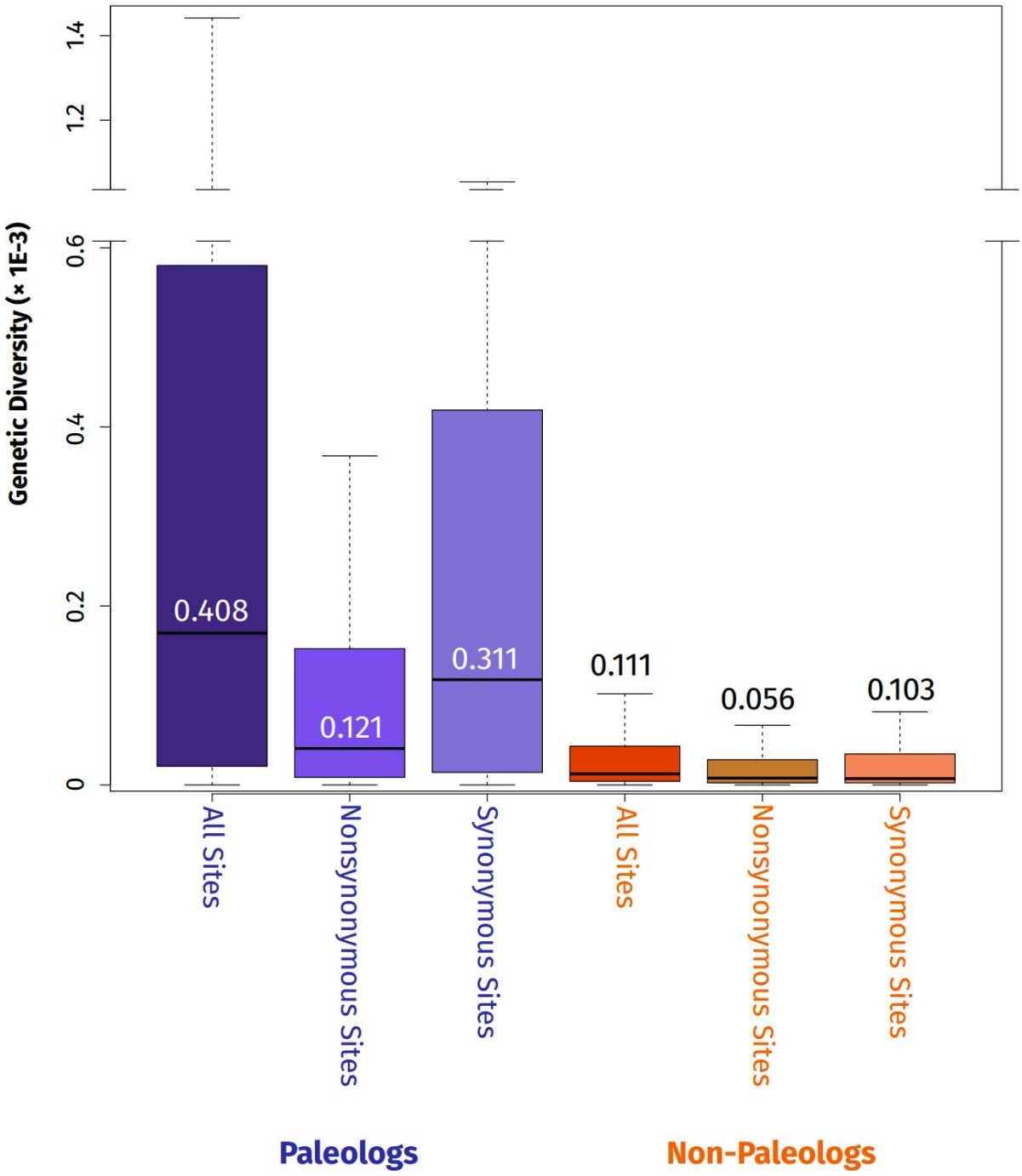



**Fig. 5: Comparison of nucleotide diversity (π) across *Brassica rapa* paleologs and non-paleologs.** The bottom and top of each box represents the first and third

quartiles, the band inside each box is the median, and the numbers represent the mean π of each category. Nucleotide diversity is shown for all sites, nonsynonymous sites only, and synonymous sites only for paleologs (purple shades) and non-paleologs (orange shades).

## Supplemental Files

**Fig S1: Manhattan plots of the SweeD analyses.** Each column represents a *B. rapa* genetic group: TO, toria; IS, Indian sarson; PC, pak choi; CC, Chinese cabbage. Each row represents a *B. rapa* chromosome (A01-A10). For each plot, the x-axis denotes the chromosome position (unit: bp), whereas the y-axis denotes the CLR value calculated in SweeD. Each dot denotes the CLR value of a genomic region. Outlier regions were indicated with red dots.

**Fig S2: Circos plot of SweeD (a) and Gene Differential Expression (GDE) gene (b) density across the *B. rapa* genome.** A01-A10 represent the 10 chromosomes of *B. rapa*. The four histogram layers denote the number of identified candidate genes. TO, toria (orange); IS, Indian sarson (light green); CC, Chinese cabbage (olive); PC, pak choi (red). The rainbow ribbons in the center represent the syntenic regions among chromosomes.

**Fig S3: Heatmaps of expression level of genes with significantly different expression in four *B. rapa* crops compared to a control group.** TO = toria, IS = Indian sarson, PC = pak choi, and CC = Chinese cabbage. The European *B. rapa* genetic group was used as control. Each column represents one *B. rapa* accession,

whereas each row represents one gene with significantly different expression between the focal group and control group.

**Table S1: *Brassica rapa* gene IDs for genes identified as paleologs and non-paleologs in our analyses.**

**Table S2: Sample information for the 102 *Brassica rapa* accessions used in this study.**

**Table S3: *Brassica rapa* gene IDs for candidate genes identified by our analyses for each crop.** TO = toria, IS = Indian sarson, PC = pak choi, and CC = Chinese cabbage.

**Table S4: Total number of candidate genes identified in our analyses and their distribution on the ten chromosomes of *Brassica rapa*.**

**Table S5: Detailed information of the five *Brassica rapa* domestication candidate genes identified in all three of our genome scan approaches.** Annotation information was obtained from the Brassica Database (brassicadb.org/brad/).

**Table S6: Summary of the *B. rapa* candidate genes identified from published mapping studies.** Brief description of functions, classical gene names, and *Brassica rapa* gene IDs are given for each candidate gene. Paleolog status is indicated as Y (yes) or N (no). If the candidate gene was also identified in our genome scans, the type of scan is indicated with S (SweeD) or D (Differential gene expression).